Investigating Social Networks with Agent Based Simulation and Link Prediction Methods

Angelico Giovanni Fetta

School of Mathematics

A thesis presented for the degree of

Doctor of Philosophy

2014

Summary

Social networks are increasingly being investigated in the context of individual behaviours. Research suggests that friendship connections have the ability to influence individual actions, change personal opinions and subsequently impact upon personal wellbeing. This thesis aims to investigate the effects of social networks, through the use of Agent Based Simulation (ABS) and Link Prediction (LP) methods. Three main investigations form this thesis, culminating in the development of a new simulation-based approach to Link Prediction (PageRank-Max) and a model of behavioural spread through a connected population (Behavioural PageRank-Max).

The first project investigates the suitability of ABS to explore a connected social system. The Peter Principle is a theory of managerial incompetence, having the potential to cause detrimental effects to system efficiency. Through the investigation of a theoretical hierarchy of workplace social contacts, it is observed that the structure of a social network has the ability to impact system efficiency, demonstrating the importance of social network structure in conjunction with individual behaviours.

The second project aims to further understand the structure of social networks, through the exploration of adolescent offline friendship data, taken from 'A Stop Smoking in Schools Trial' (ASSIST). An initial analysis of the data suggests certain factors may be pertinent in the formation of school social networks, identifying the importance of centrality measures. An ABS aiming to predict the evolution of the ASSIST social networks is created, developing an algorithm based upon the optimisation of an individual's eigen-centrality - termed PageRank-Max. This new approach to Link Prediction is found to predict ASSIST social network evolution more accurately than four existing prominent LP algorithms.

The final part of this thesis attempts to improve the PageRank-Max method, by placing particular emphasis upon specific individual attributes. Two new methods are developed, the first restricting the search space of the algorithm (Behavioural Search), while the second alters its calculation process by applying specific attribute weights (Behavioural PageRank-Max). The results demonstrate the importance of individual attributes in adolescent friendship selection. Furthermore, the Behavioural PageRank-Max offers an approach to model the spread of behaviours in conjunction with social network structure, with the value of this being evaluated against alternative models.

Declaration

This work has not been submitted in substance for any other degree or award at this or any other university or place of learning, nor is being submitted concurrently in candidature for any degree or other award.

Signed .. . . . . Date . . . . . .

STATEMENT 1

This thesis is being submitted in partial fulfilment of the requirements for the degree of PhD.

Signed .. . . . . Date . . . . . .

STATEMENT 2

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged and explicit references given. A reference section is appended.

Signed .. . . . . Date . . . . . .

STATEMENT 3

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed .. . . . . Date . . . . . .

Acknowledgements

First and foremost, I would like to thank my supervisors - Paul Harper, Janet Williams and Vincent Knight - for the time, energy and passion they have put into this research. Thank you for giving me the freedom to take us off into some (really) unexpected directions, involving jealous co-workers, smoking children and an array of specially selected snacks. I hope to be able to share a whiskey, a gin and tonic, and a cherry coke with all three of you again very soon.

I would also like to thank Israel Vieira who was integral in the early stages of my work - giving me the knowledge necessary to become an autonomous agent. Thank you also to the EPSRC and LANCS initiative, who have kept me in food, clothes and copious amounts of olive oil for the last 3.5 years.

Thank you to the DECIPHer group, namely Laurence Moore and Jo Holliday, for providing the ASSIST data. Also, a very special thank you to Jonathan Gillard, whose wise words and pastoral care have kept my head out of the shed on a number of occasions.

Truth be told, doing a PhD is soul destroying, but I was fortunate to have people around who made coming in everyday worthwhile. From muscling my way into the “Fabulous 4” all those years ago, with the group growing to incorporate special applied mathematicians, the Warwick upper classes and all manner of wonderful characters. So many memories. So much Euphoria!

Given that this research is about the effect of social networks, it is only fitting that I mention the impact of my own social network. My parents, who have always pushed me to succeed and supported me; my brothers, for keeping my feet on the ground with their pep-talks laden with sarcasm; my housemate Jo, for putting up with my lack of conversation and inability to deal with tupperware; and those special members of my family, who always welcome me with open arms and an open fridge whenever I decide to drop by. Of course, my social circle would not be complete if I didn’t mention Clare-o, always direct and always up for doing nothing - long may our ‘moments of poetry’ continue.

Finally, to my number one ‘Cardiff girl’: Leanne Smith. My biggest discovery during my research wasn’t PageRank-Max, it was the curly haired little lady sitting across from me. Thank you for being my team mate, my sous chef and my reason for bringing *two* pieces of kitchen roll to the table. Xxxx.

Publications and Presentations

Publications

- Fetta, A. G., Harper, P. R., Knight, V. A., Vieira, I. T. and Williams J. E. (2012). On the Peter Principle: An agent based investigation into the consequential effects of social networks and behavioural factors. *Physica A: Statistical Mechanics and its Applications*, 391(9): 2898 - 2910.

Awards

- Early Career Researcher Poster Prize at ORAHS Conference 2012 - First Place.

Conference Contributions & Presentations

- Fetta, A. G., Harper, P. R., Knight, V. A., Vieira, I. T. and Williams J. E. (2010). Agent Based Simulation for Complex Health Systems Interventions; *SWORDS Seminar*, Cardiff.

- Fetta, A. G., Jones, M., Knight, V. A., and Williams J. E. (2010). Workforce Planning: Signing on to system dynamics? *System Dynamic Special Interest Group Meeting* (2011), Cardiff.

- Fetta, A. G., Harper, P. R., Knight, V. A., Vieira, I. T. and Williams J. E. (2011). The Peter Principle, Agents and Beyond. *SWORDS Seminar*, Cardiff.

- Fetta, A. G., Harper, P. R., Knight, V. A., Vieira, I. T. and Williams J. E. (2011), On the Peter Principle: An Agent Based Investigation into the Consequential Effects of Social Networks and Behavioural Factors. *Winter Simulation Conference*, Pheonix, Arizona.

- Fetta, A. G., Harper, P. R., Knight, V. A., Vieira, I. T. and Williams J. E. (2012). On the Peter Principle: An Agent Based Investigation into the Consequential Effects of Social Networks and Behavioural Factors. *SCOR*, Nottingham.

- Fetta, A. G., Harper, P. R., Knight, V. A., Vieira, I. T. and Williams J. E. (2012). On the Peter Principle: An Agent Based Investigation into the Consequential Effects of Social Networks and Behavioural Factors. *EURO*, Vilnuis, Lithuania.

- Fetta, A. G., Harper, P. R., Knight, V. A. and Williams J. E. (2012). Modelling Adolescent Smoking Behaviours with Social Network Analysis. *ORAHS*, Twente, Holland.

- Fetta, A. G., Harper, P. R., Knight, V. A. and Williams J. E. (2012). Modelling Adolescent Smoking Behaviours with Social Network Analysis. *NHS Confederation Event: Change by design - systems modelling and simulation in healthcare*, Exeter.

- Fetta, A. G., Harper, P. R., Knight, V. A. and Williams J. E. (2013). Modelling Adolescent Smoking Behaviours with Social Network Analysis. *INSNA Sunbelt Conference*, Hamburg, Germany.

- Fetta, A. G., Harper, P. R., Knight, V. A. and Williams J. E. (2013). Modelling Adolescent Smoking Behaviours with Social Network Analysis. *EURO*, Rome, Italy.

List of Abbreviations

ABS	**A**gent **B**ased **S**imulation
API	**A**pplication **P**rogramming **I**nterface
ASSIST	**A** **S**top **S**moking **I**n **S**chools **T**rial
BPRM	**B**ehavioural **P**age**R**ank-**M**ax
CI	**C**onfidence **I**nterval
CS	**C**ommon **S**ense
DECIPHer	**D**evelopment and **E**valuation of **C**omplex **I**nterventions for **P**ublic **He**alth Imp**r**ovement
EOS	**E**volution of **O**rganised **S**ocieties
ESS	**E**volutionary **S**table **S**trategy
GUI	**G**raphical **U**ser **I**nterface
LP	**L**ink **P**rediction
MANTA	**M**odelling an **ANT** hill **A**ctivity
MAS	**M**ulti **A**gent **S**ystem
MCMC	**M**arkov **C**hain **M**onte **C**arlo
MLE	**M**aximum **L**ikelihood **E**stimator
MPI	**M**ass **P**sychogenic **I**llness
NBM	**N**etwork **B**ehavioural **M**odel
NE	**N**etwork **E**fficiency
OOP	**O**bject-**O**riented **P**rogramming
OR	**O**perational **R**esearch
OSS	**O**pen **S**ource **S**ystem
OOP	**O**bject-**O**riented **P**rogramming
PP	**P**eter **P**rinciple
PS	**P**roprietary **S**ystem
PSM	**P**opulation **S**tructure **M**odel
RAN	**RAN**dom Network
SAB	**S**tochastic **A**ctor **B**ased
SF	**S**cale **F**ree Network
SNS	**S**ocial **N**etwork **S**imulation
SW	**S**mall **W**orld Network
WHO	**W**orld **H**ealth **O**rganisation

Contents

List of Figures

List of Tables

-"The Isolated Node"

1

Introduction

This thesis investigates the structure of social networks, examining the role friendship connections may have upon the personal decisions of an individual. A novel approach to predicting the evolution of a social network is introduced, with the newly developed algorithm being expanded to create models of the relationship between friendship structure and individual behaviour. The specific focus of this investigation centres upon adolescent friendships and their role in influencing smoking uptake, with the research drawing upon both theoretical and empirical analyses.

This chapter serves to set the context of the subsequent research, and is structured in the following manner: the motivation for this investigation is discussed in Section 1.1; the research methods to be employed are presented in Section 1.2; an outline of the research aims is formed in Section 1.3; and an overview of the structure of the thesis is described in Section 1.4.

1.1 Motivation

1.1.1 Social Connection

Connection. The Oxford Dictionary (2010) describes the word 'connection' as :

"...a relationship in which a person or thing is linked or associated with something else."

The need for connection is at the very basis of human nature. Maslow (1943) theorises that love, affection and belonging needs are superseded solely by physiological needs (food, water etc.) and the search for physical safety. A connection may provide the recipient with a basis of support (Wasserman & Galaskiewicz, 1994) or potentially the opportunity to learn (Kashima et al., 2013), but it may also provide a great deal of influence on the attitudes and beliefs of the reciprocating party (Smith & Louis, 2008).

The body of this research focuses upon the interactions a person makes on a daily basis, extraneous to that of familial ties. Whether it be in a school environment or the workplace, often an individual may interact for extended periods of time with a group (or groups) of people with whom proximity is the primary factor in their assembly. Over such periods, connections may form and dissolve naturally - often not just as a product of the individual's personal decisions, but also as a result of group decisions (Killen, 2007; Killen & Stangor, 2001).

Regular contact with others may lead to the alteration of one's own perceptions (Campbell-Meiklejohn & Bach, 2010), performance (Sias et al., 2004) and, more gravely, their health (Cunningham & Vaquera, 2012; Salvy & Haye, 2012). Therefore, it is of great interest to investigate how an individual evaluates the prospect of a potential connection, and the effect of a connection (or a selection of connections) upon an individual's decisions and beliefs.

The topic of connections, and their subsequent implications, is an area of growing interest (Kleinberg & Easley, 2010). This is attributed to the recent growth in online social networking platforms (EMarketer, 2013), which are able to provide a vast array of relational information for analysis. Moreover, advances in technology have allowed for the processing of large amounts of data - an integral requirement for the analysis of social connections

(Scott, 2005).

In terms of online social networks, the growth in web-based communities has resulted in previously existent contact barriers being no longer applicable. For example, public figures have become accessible via Twitter (2013), and social networking sites (Facebook, 2013; Google+, 2013; Instagram, 2013) allow a user to maintain friendships without the need to actively contact individuals. Even romantic relationships are affected by an online presence, with internet dating now accounting for a third of marriages in the USA (Cacioppo et al., 2013).

While many studies discuss the topic of online interaction (Kwak et al., 2010; Mislove et al., 2008; Pollet et al., 2011b; Salter-Townshend, 2012), this research focuses upon offline connections - described as relationships which are not initiated and solely cultivated online. This is primarily motivated by research suggesting that offline connections are particularly influential (Christakis & Fowler, 2007, 2008; Potterat et al., 2002; Rankin & Philip, 1963; Raspe et al., 2008), with workplace interactions (Berman et al., 2002; Brass, 1985; Mao, 2006) and adolescent peer networks (Bearman & Moody, 2004; Jones et al., 2000; Kandel, 1978; Mercken et al., 2009, 2010) being of particular importance in this thesis.

The decision to investigate offline friendships is further advocated by research suggesting that online links lack the strength and depth of offline connections, resulting in a relationship that may be interpreted as shallow (Cocking & Matthews, 2000; Fröding & Peterson, 2012; Mesch & Talmud, 2006). However, Briggle (2008) and McKenna (2002) dispute this, claiming that the online domain offers a platform in which to portray ones true self - resulting in a more authentic connection uninhibited by social anxiety. Nevertheless, offline connections shall be the principal focus of this research, with the findings potentially being applicable in the context of online relations.

To conduct this investigation into social connections, social network simulation is identified as the primary analytical technique - the motivation for this being described in the following section (1.1.2).

1.1.2 Social Network Simulation

Simulation provides a tool to investigate the evolution of a system, increase understanding and evaluate potential outcomes. Within the domain of OR, simulation is a core tool for research, visualisation and process improvement - lending itself to applications such as Manufacturing, Defence and Healthcare (Pidd, 2004). In terms of simulation within a network structure, examples may include problems such as Vehicle Routing (Juan et al., 2013), Information Networks (Breslau et al., 2000) and Epidemiology (Huang et al., 2010).

Fewer examples may be found relating to the social applications of simulation, with Taylor et al. (2009) finding that in a review of the journals *'ACM Transactions of Modeling and Computer Simulation'*, *'Simulation: Transactions of The Society for Modeling'* and *'Simulation International and Simulation Modelling Practice and Theory'*, only 3.63% of published papers between the years of 2000-2005 may be attributed to the topic. Literature relating to social network simulation is even more scarce, with the analysis of friendship structures generally being conducted with purely statistical approaches; Wasserman & Faust being the seminal Social Network Analysis (SNA) text.

An SNA technique of particular interest is that of the Stochastic Actor-Based (SAB) approach, which seeks to generate a statistical model of network change based upon underlying investigative simulations (Snijders et al., 2010). This methodology is commonly used to explore longitudinal network evolution, often specifically examining the co-evolution of friendship and particular behaviours of interest. Literature demonstrates that a range of behaviours have been examined with SAB, including that of smoking habits (Steglich et al., 2012), alcohol consumption (Steglich et al., 2006) and substance use (Pearson et al., 2006).

Unfortunately, the SAB approach is plagued by excessive model adjustment and a requirement of multiple waves of network data to generate a model - a task which may often be costly, time-consuming and require experts in the field (Snijders et al., 2010). It would therefore be of great interest to harness the potential of simulation techniques, to produce a methodology that provides a tool for the analysis of network change, unencumbered by such extraneous needs.

The motivation for the use of network simulation is a result of the lack of studies ap-

plying simulation techniques to social networks, and the inherent limitations of the SAB approach. Furthermore, with a simulation-based approach, the evolution of a network may not only be analysed: the impact of change within the network may also be explored. The topic of simulation, and its application to networks, shall be further examined in the literature review of Chapter 2. The following section (1.1.3) discusses the specific applications of network simulation in this thesis.

1.1.3 Application

The research to be presented attempts to investigate the role of a simulation-based perspective of SNA upon two distinct environments, initially creating a theoretical model of *workplace interactions*, and then examining real world *adolescent social network* data. The interplay between social networks and *smoking behaviours* amongst adolescents shall also be explored, giving a specific application to the social network influence processes being investigated. The motivation for the selection of these three topics are as follows:

- Workplace Interaction

  To begin this investigation, a theoretical model of social connectivity in the workplace is explored. According to official figures, around 30 million people collectively form the UK workforce (ONS, 2013c). In a multitude of professions, a hierarchical corporate structure exists - emphasising promotion as a key expectation of success. Peter & Hull (1969) suggest that existing promotional strategies may be flawed in that they allow for the cultivation of incompetence at a managerial level, a concept known as 'The Peter Principle' (PP).

  Existing research into the PP uses ABS to examine the effect of alternative promotion strategies upon organisation efficiency; however, this existing research has not explored the effect of social networks and behavioural factors upon the dynamic of the system. As such, this presents an opportunity to build upon the current models of hierarchical organisations, assessing the effect of imposing theoretical social structures to overall efficiency. Additionally, this affords the ability to ascertain the suitability of using ABS to investigate the structure and influence of social networks, directing the research of this thesis.

- Adolescent Social Networks

Building upon the theoretical model of workplace interaction, an empirical analysis of adolescent social networks is conducted. The adolescent stages of life are often regarded as 'the formative years', a period of self-discovery whereby the decisions made may have a subsequent impact upon future choices (Sawyer et al., 2012). Friendships forged during this time may consequentially affect an individual's accepted social norms in adulthood. Therefore, to understand the significance of such processes, a quantitative analysis of linking behaviour is required.

Often a qualitative approach is employed to explore the friendship patterns and behaviour of the adolescent years, a demonstration of which may be found by the sheer wealth of interview based studies focusing on the topic (Dishion et al., 1995; Rudolph et al., 2013; Zimmermann, 2004). However, a quantitative approach is far more sparsely engaged, the main thrust of such research dominated by the SNA techniques of Wasserman & Faust, or SAB model construction with the RSiena software (Ripley et al., 2012).

The motivation to adopt a quantitative simulation-based research approach to adolescent friendships, is that it appears to be an unexplored niche in social network literature and has proven to be an invaluable tool in other experimentation and prediction fields. This offers the opportunity to employ techniques from other fields to understand the manner in which adolescents connect, potentially having direct consequences to the understanding of engagement with prevalent social behaviours; for example, the uptake of smoking.

- Smoking

Using the evidence-based investigation of adolescent social networks, the impact of friendship connections upon smoking behaviours is explored. Tobacco use is said to kill 5.4 million people globally per year (World Health Organisation, 2013). In the UK it is estimated that 20% of the population are regular smokers (ONS, 2013a), with the impact of smoking related diseases costing the NHS £5 billion per year (Allender et al., 2009). Given the widely known negative effects of smoking, the restriction on public tobacco advertisement, the ban on smoking in enclosed public spaces and the move to remove all branding on cigarette packets, the public are still actively engaging in this potentially damaging behaviour.

The selection of smoking as the behaviour of choice for analysis, in the context of adolescent peer relations, is motivated by a number of factors. The first is the evident conflict in terms of active intervention by public officials, and the continuation of smoking as an accepted social norm. Further to this, research has found that adolescents who smoke often find it more difficult to quit later in life (Fergusson et al., 1995; Prokhorov et al., 1996).

Additionally, smoking is said to be of key social significance. It is a behaviour often engaged within a group capacity (Lakon & Valente, 2012), whereby it is not uncommon for adolescents and adults alike to define themselves as a "social smoker" (Levinson et al., 2007). It is therefore of interest to understand the role of social networks in the uptake of smoking, investigating the inter-dependence of friendship selection and an adolescent's decision to smoke.

This section has discussed the motivation for the ensuing research, introducing social connections, social network simulation and the three specific application areas: workplace interactions, adolescent social networks and adolescent smoking. A greater review of the literature relating to simulation and networks, is presented in Chapters 2 and 3, respectively. The following section (1.2) introduces the research methods to be employed throughout this work.

1.2 Research Methods

While the key application of this research is undoubtedly the analysis and evolution of social networks, the underpinning methodology will be that of simulation. The paradigm of particular interest is Agent Based Simulation (ABS) (Macal & North, 2005), which aims to take an individualistic view of system evolution (further simulation techniques being discussed in Chapter 2). Given the individual nature of the friendship choices which form a social network, ABS would appear an appropriate method to utilise.

The specified task of investigating network dynamics falls squarely into a class of problems known as 'Link Prediction' (LP) (Liben-Nowell & Kleinberg, 2007). LP methods aim to predict, between unconnected nodes on a graph, the links that will develop at later points in time. Using simple SNA metrics, a ranking of the likelihood of potential connections may be achieved, allowing for the estimation of future instances of network structure. Existing

applications of LP methods utilise networks which primarily add new connections (such as citation networks), and exhibit very few disconnections (Liben-Nowell & Kleinberg, 2007; Lü & Zhou, 2011). As such, it is of interest to develop a new LP method to account for the dynamics of social networks and identify important aspects in their evolution.

For such a task, social network data will be required to conduct and verify the results of an LP analysis. This work will therefore utilise the data from 'A Stop Smoking in Schools Trial' (ASSIST) (Starkey et al., 2005), a peer-led intervention study into the effect of peer-nominated leaders in the prevention of smoking related behaviours. Information regarding friendship ties, and their development longitudinally across three years, along with smoking data is held for a range of secondary education schools - offering a platform from which to conduct the specified research.

The combination of LP methods and ABS techniques shall also be utilised to investigate the uptake of smoking behaviours amongst the ASSIST cohort, exploring thresholds at which smoking becomes a majority behaviour. Additionally, an Evolutionary Game Theory (EGT) (Tadelis, 2012) model and a basic compartmental model (Kermack & McKendrick, 1932) are presented, demonstrating alternative approaches to the investigation of social influence.

To summarise, this work identifies two key concepts that shall be employed in the investigation of social networks: Agent Based Simulation and The Link Prediction Problem. Following a review of the relevant literature and an analysis of the available data, the appropriateness of said techniques shall be evaluated and their application to the specified problems assessed. It is hoped that the selection of mathematical techniques presented, provide for an enriching and rigorous analysis in relation to the specified topics.

1.3 Research Aims

As outlined above, this research is concerned with the investigation of social network structure and influence, and will apply two key concepts - Agent Based Simulation and Link Prediction - in conjunction with the available data. The aims of this research may be divided and concisely represented by the following principle objectives:

- Apply Agent Based Simulation methods to investigate the effect of social network

structures in a theoretical social environment;

- Explore the social network structures of ASSIST to identify important factors in adolescent friendship selection and social influence;

- Develop a new simulation-based approach for the prediction of social network evolution, aiming to incorporate the identified important structural evolution processes of adolescent social networks;

- Evaluate the effectiveness of the developed framework in the prediction of links from the ASSIST dataset, giving particular attention to the differences between schools;

- Create a framework to investigate the interplay between social network structure and smoking behaviours.

These objectives will be revisited in the closing statements of this work, to assess their contribution and to provide an overarching point of reference for this research.

1.4 Outline

This thesis is divided into ten chapters, the first three aim to set the context for the reader - explaining both the motivation and relevance of the research amongst the currently available literature. The following six (Chapters 4 - 9) describe the actionable research and analysis itself, providing the appropriate background and contextual information where necessary. Finally, the last chapter summarises the work and forges the direction for future research.

A concise inventory of the remainder of this investigation may be detailed as follows:

- Chapter 2 provides a literature review of simulation. This chapter explores the general concepts of simulation, with a specific focus upon ABS - a key research method employed in this thesis;

- Chapter 3 presents a literature review of network science, introducing important network analysis metrics integral to this investigation. Additionally, literature relating

specifically to social network structure and influence is reviewed, with an introduction to Link Prediction problems also being provided;

- Chapter 4 initiates an investigation into the effects of social structure upon individual behaviours, through the development of a theoretical model of behaviour within a hierarchical organisation. The results from this investigation form the basis of the empirical analyses conducted in later chapters;

- Chapter 5 explores the data acquired for this investigation. An analysis of social network structure, social influence and the effects of a peer-led smoking intervention (ASSIST) is conducted. The outcomes provide a greater understanding of social connectivity, with the findings informing the new algorithms and models developed in this thesis;

- Chapter 6 introduces the new method developed in this thesis to predict social network evolution, PageRank-Max. A description of the simulation-based framework created for this investigation is provided, with an outline of existing techniques (and their implementation within the simulation) also discussed;

- Chapter 7 evaluates the performance of the developed PageRank-Max algorithm against existing Link Prediction methods. Using the network structures from the ASSIST data, link precision and network structural metrics are evaluated;

- Chapter 8 builds upon the algorithm developed in Chapter 6 and evaluated in Chapter 7, incorporating individual attribute data to inform the evolution of social network structure. This chapter aims to improve the link predictions made and develop a framework to consider the co-evolution of social networks and individual behaviours;

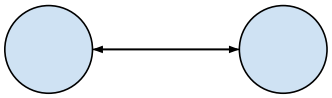
- Chapter 9 uses the framework developed in Chapter 8, to assess the role of social network structure upon the diffusion of smoking behaviours. Additional models of social influence are outlined, providing alternative directions for future research;

- Chapter 10 draws together the conclusions of this research, providing a summary of the detail covered, options for future research and a reflection on the thesis as a whole.

1.5 Chapter Review

In summary, this chapter has outlined the general context, motivation and proposed structure of the ensuing investigation. Section 1.1 described the motivation for the selected direction of research, introducing social connections, social network simulation and the three specified topics of interest: workplace interactions, adolescent social networks and adolescent smoking. Section 1.2 outlined the research methods to be employed in the analysis of the designated research, identifying Agent Based Simulation and Link Prediction as being of particular relevance.

Section 1.3 presented the key objectives of this research, which shall be addressed over the course of the remaining chapters and revisited in the closing statements of the thesis. Finally, Section 1.4 provided a breakdown of each chapter, demonstrating the overall structure of the proceeding work. Prior to discussing the research contributions of this investigation, Chapter 2 provides a review of the literature pertaining to simulation.



-"The Reciprocated Dyad"

2

Simulation Literature Review

This chapter is the first of two literature reviews to be conducted in this thesis. Chapter 1 introduced simulation as a key technique that shall be used throughout this research. As such, this chapter shall present literature relating to simulation and its applications. The main focus of this literature review is a discussion of ABS, as it is the selected simulation paradigm for this research; however, a review of the general principles of simulation is also presented, along with brief introductions to alternative simulation methods.

The following review is structured such that: an outline of simulation is provided in Section 2.1; an introduction to the different paradigms of simulation is presented in Section 2.2; literature discussing the applications of Agent Based Simulation (ABS), with particular focus upon social networks and smoking, is reviewed in Section 2.3; and a summary of this review is given in Section 2.4.

2.1 What is Simulation?

Raczynski (2006) describes *computer* simulation as the "process of making a computer behave like a cow, an airplane, a terrorist, a HIV virus... or any other thing". Simulation has become an accepted part of human conciousness, examples of its usage including: the forecasting of weather (Lynch, 2008), training pilots with flight simulators (Page, 2000) and computer games for entertainment purposes (Atkins, 2003). Of course, simulation can also refer to *physical* simulations, such as model railways and remote control boats (Robinson, 2004); however, for the purpose of this research, simulation is referred to in the context of computer simulations.

The origins of simulation are said to be based in military applications (Hill et al., 2001). It is reported that the first use of simulation occurred in 1945, using Monte Carlo methods in the design of a thermonuclear bomb (Cahn, 2001). Ever since this ground breaking study, simulation has been applied to a wealth of problems (Alam & Geller, 2012; Ashraf et al., 2011; Brooks et al., 2001), with Pidd (2004) stating that simulation is amongst the top three techniques in management science.

Within the context of Operational Research (OR), Robinson (2004) defines simulation as:

Definition 2.1.1. *Experimentation with a simplified imitation (on a computer) of an operational system as it progresses through time, for the purpose of better understanding and/or improving that system.*

Definition 2.1.1 has five key elements:

- **Operational System** - The development of a simulation requires the representation of a system of interest. Checkland (1981) states there are four types of system: a natural system (e.g. the weather, fluid dynamics), a physical system (e.g production lines, warehouses), a designed abstract system (e.g. mathematics, literature) and a human system (e.g the delivery of health services, behavioural interactions).

- **Simplified Imitation** - Once the system is identified, a simplified version of its real world counterpart must be interpreted. Box & Draper (1987) hypothesise that "all models wrong, some are useful", with Gilbert & Troitzsch (2005) stating that the

most difficult step in the design of a simulation is the decision of what to include and what to exclude. A detailed model may in fact be undesirable - due to the data and computing power required for such accuracy - or unobtainable - as often the process of creating a simulation is to gain a greater understanding of the system.

- **Time** - A simulation is generally concerned with how the simplified system develops over time, referred to as a dynamic simulation (Law & Kelton, 1999).

- **Experimentation** - The ability to model the simplified system over time, allows for experimentation with various inputs, to assess the resultant outputs. It also allows for alteration of the system, often in a 'trial and error' manner, to answer 'what if?' style questions (Pidd, 2004).

- **Understanding/Improving** - The purpose of experimenting with the dynamic model of the simplified system is to gain a greater understanding of its particular features, predict outcomes in some future time and/or improve operations.

The presented statements capture the essence of a simulation, its ideology and its purpose - to understand and improve aspects of the system. Evidently, other techniques may be used to gain such insights, for instance: direct experimentation or alternative mathematical models. The following section (2.1.1) explains the reasons a simulation model may be of particular benefit over such techniques.

2.1.1 Benefits of Simulation

Simulation is regularly used to investigate systems related to manufacturing (Negahban & Smith, 2014), healthcare (Mustafee et al., 2010) and defence (Hill et al., 2003) (amongst many others). The breadth of sectors employing simulation methods demonstrates its effectiveness in a problem solving capacity. In particular, structuring problems with a simulation framework affords the opportunity to explore system changes, or gain a greater understanding of the system being investigated. A number of benefits gained from the creation of a simulation are outlined as follows, taken from the literature of Kornbluh & Little (1976), Pidd (2004) and Robinson (1994) :

- **Cost/Savings** - Physically experimenting with a system may be costly. If changes to a factory production line required testing, this would be expensive in terms of physi-

cal modifications and the time required to conduct the alterations. Simulation allows for the testing of potential adjustments without the need to incur such expenses. Moreover, when testing specific scenarios, the optimal scenario may be identified; as such, savings can be made through the improvement of the system's current state and the avoidance of implementing sub-optimal scenarios.

- **Safety** - Simulation provides a safe environment to conduct experimentation. Investigating certain systems in reality may be illegal, impossible or dangerous - for example, aid response following a natural disaster. The consequences of changing a system may also be unknown, with its implementation posing risks to system users. Investigation through simulation allows for the reduction of such risks, improving overall safety.

- **Replication** - If investigation were to be conducted with the real world system, controlling for experimental conditions may prove difficult. For a fair assessment of the impact of implementing alternative policies, equivalent underlying conditions are required. Simulation allows for such investigation under repeatable conditions.

- **Understanding** - The construction of a simulation experiment requires an understanding of the system being modelled, which may in itself improve overall comprehension of the system. The ability to experiment, especially under extreme conditions, allows for an increased understanding of the capabilities of the system and the effects of making changes.

- **Visualisation** - Many simulations are created with a graphical representation of the system, allowing the user to visualise its evolution over time. This is a useful tool for improving understanding, and communicating ideas to interested parties.

- **Dynamic Effects** - The ability to model variability and its effects are of great benefit, especially when the system is subject to extreme conditions and transient effects. Under such circumstances, simulation may be preferable over other mathematical techniques, which may solely consider average values.

The particular benefits of simulation in relation to the context of this work are highlighted further in Section 2.2.3. While there are evidently a great deal of positive aspects to the creation of a simulation model, the limitations of the approach must also be considered.

2.1.2 Limitations

Robinson (2004) states that simulation should be used as a tool for supporting decision making, as opposed to making decisions on behalf of a user. Models may be an oversimplification, or simply inaccurate, with the main sources of inaccuracy being: the model, the data and/or the experimentation (Robinson, 1999). While the creation of a simulation may require less time and money than the real-world experimentation, and potentially offer savings in comparison to current system conditions, there are still cost and time factors to be considered:

- **Computation** - The computing resources required to develop complex models, or house the required data may be prohibitive. It may also take a considerable length of time to build a simulation, and as such, the consideration of complexity in relation to the model requirements must be considered;

- **Data** - the collection of data may also be expensive and take a considerable length of time, although simulation models can be built upon theoretical assumptions (Gilbert & Troitzsch, 2005), avoiding such costs.

Expertise is also required in the construction of a simulation, a 'simulationist' having to possess the necessary computing skills, along with abilities in conceptual modelling, validation, statistics and working as part of a team (Robinson, 2004). These limitations should be considered when opting to build a simulation.

2.1.3 Building a Simulation

Law (2009) describes the creation of simulation as a seven step procedure:

1. **Formulate the problem** - Deciding on the system to be investigated, the problems to be tackled and the scope of the model are key first steps in the development of a simulation. These preliminary measures shall decide many aspects of the created simulation, from the level of detail to be included in assumptions, to the method chosen to best represent the system (discussed further in Section 2.2). Texts specifically focused on the issue of problem structuring are Mingers & Rosenhead (2001) and Pidd (2003).

2. **Collect Information** - Once the problem has been formulated, information regarding the system must be gathered. Data relating to structure, model parameters and probability distributions is required to formulate assumptions. If data regarding specific elements of the system is unavailable or ambiguous (as is the case in many social simulations), information collection may be conducted from reviewing appropriate literature (Gilbert & Troitzsch, 2005).

3. **Assumptions** - The process of conceptualising the model in relation to the collected information, informs the assumptions. A conceptual model is the process of abstracting a model from a real system (Robinson, 2008), described as the intermediary phase between problem and assumption formulation. Further discussion regrading the model conceptualisation process may be found in Balci & Ormsby (2007), Kotiadis & Robinson (2008) and Robinson (2007). The specific assumptions made, govern the level of detail and scope of the model, with Robinson (1994) suggesting that only the minimum amount of detail to achieve the projects aims should be included.

4. **Program the Model** - The process of realising the conceptual model in a computerised framework, either through a general purpose programming language (Java, Python, C++ etc.) or with a commercial simulation software package (e.g AnyLogic, Simul8, Arena, Witness). Purpose built simulation packages are often Visual Interactive Modelling Systems (VIMS), allowing the modeller to visually interpret the system being constructed (Pidd, 2004); more information regarding the history and benefits of VIMS may be found in Bell (1991) and Au & Paul (1996) respectively.

5. **Verification and Validation** - Fishman & Kiviat (1967) state that only after "a model has been verified and validated can an experimenter justifiably use a model to probe system behaviour." *Verification* is said to be the process of determining whether the simulation is behaving as expected, and *Validation* refers to the process of assessing whether the simulation is fit for purpose (Law & Kelton, 1999). There is a great deal of literature surrounding best practices in terms of verification and validation (Balci, 1994; Gass, 1983; Kleijnen, 1995; Robinson, 1994; Sargent, 2005), but it is generally accepted that complete validation is impossible. Key aspects of the verification and validation procedure include the examination of coding logic,

distributions and random sampling; each of these aspects is discussed in more detail in Section 6.5.

6. **Experimentation** - Once the model is deemed suitable, experimentation may commence. This may be through *interactive experimentation*, whereby the model is watched and actions are assessed interactively, or *batch experimentation*, setting test parameters and allowing the model to perform multiple runs for a period of time (Robinson, 1994). In the experiments conducted, consideration must be given to warm-up period (the length of time required for the model to achieve steady-state), the number of replications required (accounting for the models inherent variability) and run length (the period of time under investigation with the simulation).

7. **Presentation/Implementation** - Following results collection with the given experimentation procedures, implementation is the final stage of the simulation process. Implementation can come in the form of a tangible outcome, whereby either the results of the model are used to alter a system, or the model itself is implemented in some capacity. Implementation may also be more abstract, the insights gained from the model changing the way a system is thought about - affecting future decisions (Robinson, 2004). If a stakeholder is involved, then this aspect of the simulation process involves presenting the results and feeding back the conclusions gained (Robinson, 2001).

The development of a simulation is not necessarily the linear process described above, with interactions occurring between many of the steps. For example, following the verification and validation procedure, the modeller may be required to return to the programming stage of model development. Figure 2.1 displays a visual representation of the simulation creation process and the interactions that may occur between the various stages. Particular phases of model development may vary based upon the chosen simulation type, which in itself should be informed by the problem formulation phase. A discussion regarding the various types of simulation are discussed in the following section (2.2).

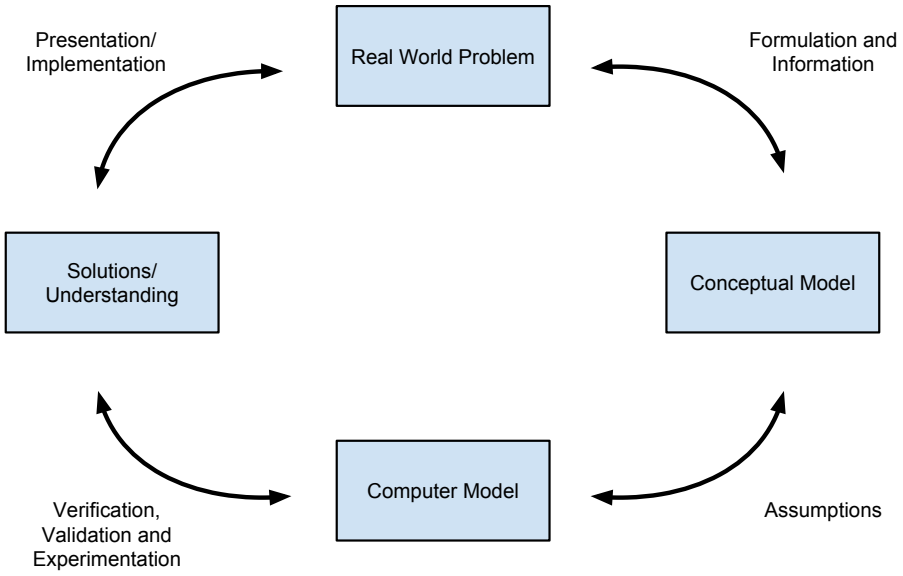


Figure 2.1: The simulation development process, adapted from Brooks et al. (2001).

2.2 Types of Simulation

There are said to be 17 different areas (types) of modelling and simulation, spanning various sectors and used for a wide range of purposes (Taylor et al., 2013). Within OR, there are three commonly used paradigms of simulation: System Dynamics (SD), Discrete Event Simulation (DES) and Agent-Based Simulation (ABS). This section provides an introduction into the principles and characteristics of each method, with particular emphasis upon ABS - this being the approach that shall be used in later sections of this thesis.

2.2.1 System Dynamics

Forrester (1958) formulated System Dynamics (SD) in the investigation of shift patterns for General Electric, examining their household electronic plants in Kentucky. Originally entitled 'Industrial Dynamics', the method was developed to incorporate qualitative aspects of factors affecting industry - such as managerial influence, leadership qualities and organisational goals (Forrester, 1995). SD takes an aggregated view of the system, with the approach of structuring a problem in an SD manner adopting the term "systems thinking" (Forrester, 1994).

Forrester (1968) defines a **system** as a "grouping of parts that operate together for a common purpose", while **dynamics** refers to "change over time" (Sweetser, 1999); therefore, System Dynamics is a method used to understand system change over time. An SD model explicitly represents a system as a composition of interconnected parts, examining both information flow and the physical flow of objects around the system. This structure allows for the investigation of interactions between various sectors of the system, offering the ability to model feedback or causal loops.

State changes in an SD model are continuous, with the objects in the system interpreted as a continuous body (Martin & Raffo, 2000) - much like water flowing through a tap. Although changes are continuous, the model is underpinned by a system of difference equations, which are solved using numerical integration methods with a discrete time slicing approach (Brailsford & Hilton, 2001). An SD model is also interpreted as deterministic, caused by the difficulty in effectively expressing variability in the models created (Doebelin, 1998).

The focus of an SD model is very much upon the system, with its structure dictating overall performance (Pidd, 2004). This macroscopic view gives little consideration to the objects within the system, Forrester (1961) justifying this view because "decisions are not entirely 'free will' but are strongly conditioned by the environment". SD is often employed to analyse large complex systems, being used across the domains of healthcare (Evenden et al., 2005b; Lane et al., 2000; Loyo et al., 2013; Royston et al., 1999), electrical power (Dastkhan & Owlia, 2014; Ford, 1997; Pruyt, 2004) and marketing (Maier, 1998; Nicholson & Kaiser, 2008; Otto, 2008). While SD shall not be explicitly employed in this thesis, the modelling paradigm provides a contrasting perspective to those of DES and ABS.

2.2.2 Discrete Event Simulation

Fishman (1978) describes a discrete event system as one in which "a phenomenon of interest changes value or state at discrete moments of time rather than continuously with time". Queuing systems are an example of a discrete system, as individuals move position in a queue at discrete moments in time (Cassandras & Lafortune, 2008). Many other systems may also be represented in a discrete event framework, with DES said to be one of the most frequently used of the classical operational research tools (Hollocks, 2006).

DES arguably originated in the late 1950's (Taylor & Robinson, 2006), its development as a core OR technique being greatly influenced by advances in computing technology (Nance & Sargent, 2002). An important characteristic of a DES is the manner in which it handles time, making use of the next-event technique to control system evolution (Pidd, 2004). As such, the simulation is only updated as and when an event occurs, with each event changing the state of the system and no state changes being exacted between consecutive events. This is opposing to SD whereby events continually take place, the simulation being updated at regular time intervals.

DES takes a process oriented view, whereby the *process* is a sequence of *events* and *activities* through which an object moves (Cuomo et al., 2012). A further characteristic of a DES is the focus upon the system at the entity level. An entity can be used to represent a person or object's movement through a system, allowing for an increased level of detail. The entities are passive, meaning that decisions regarding their progression through the structure are controlled entirely by the system - involving no decisions from the entities themselves. Therefore, the entities serve to provide detail to the process being modelled, but only a macroscopic view of the entities' actual behaviour is represented (Siebers et al., 2010).

An additional aspect of DES is its ability to incorporate variability into a model with ease. This allows for an examination of system response under variable periods, investigating the multiplicative effect of such stochastic processes. Further information regarding the DES process may be found in Pidd (2004), with historical perspectives on the evolution of DES provided in Hollocks (2006) and Robinson (2005). A comparative view of DES and SD may be found in Morecroft & Robinson (2005), Sweetser (1999) and Tako & Robinson (2009), with Banks (1998) giving three direct benefits of using a DES. While system detail may be increased over SD, the behaviour of the system ultimately remains the result of its structure as opposed to its entities. Fetta et al. (2010) describes the differences between DES and SD through the analogy of photographing a river, stating that "DES photographs a close up of the boats", while "SD takes an aerial view of the course the boats will take".

2.2.3 Agent Based Simulation

This section provides a detailed review of the ABS paradigm, as it is the selected research method for this thesis. The following discussion details the historical evolution, under-

lying theory, positive aspects, limitations and software related to ABS. This review feeds into that of Section 2.2.4 which examines present advances in simulation theory and technology.

History

ABS is said to have its origins in the work of von Neumann (1966), who wished to study artificial automata. von Neumann (1966) developed a blueprint for a self-reproducing machine, which involved a series of complex governing rules and heavy-duty machinery. On the advice of a colleague, the complexity of the machine was stripped back and represented by a Cellular Automata approach (Langton, 1997); the theory that global complexity can emerge from simple local rules later becoming an important principle of ABS (Gleick, 1997).

A Cellular Automata approach interprets a system as a grid of cells, with the "state" of each cell represented by its own variable values. The system evolves in discrete time, with the value of a variable at one cell having the potential to affect the values of adjacent cells (Wolfram, 1983). The local actions of each cell causes the emergence of global behaviour, leading to a "bottom up" approach for modelling a system (Heath, 2010); a further important aspect of modern day ABS.

Cellular Automata were also used by Conway in "the game of life" (Gardner, 1970). The game of life takes place upon a square grid, whereby each cell is either dead or alive. Every cell impacts each of its eight neighbours - those cells which are vertically, horizontally or diagonally adjacent. The following rules govern the game:

- Living cells with four or more living neighbours die of overpopulation.
- Living cells with fewer than two living neighbours die from isolation.
- Living cells with two or three living neighbours remain alive.
- Dead cells with exactly three living neighbours become a live cell.

This is a *zero player game* whereby the initial conditions dictate the evolution of the system (Björk & Juul, 2012). Figure 2.2 displays the game of life with a randomly selected initial

cell setup and its evolution to a final steady state. Various patterns can evolve from the selected initial conditions of the game (Adamatzky, 2010).

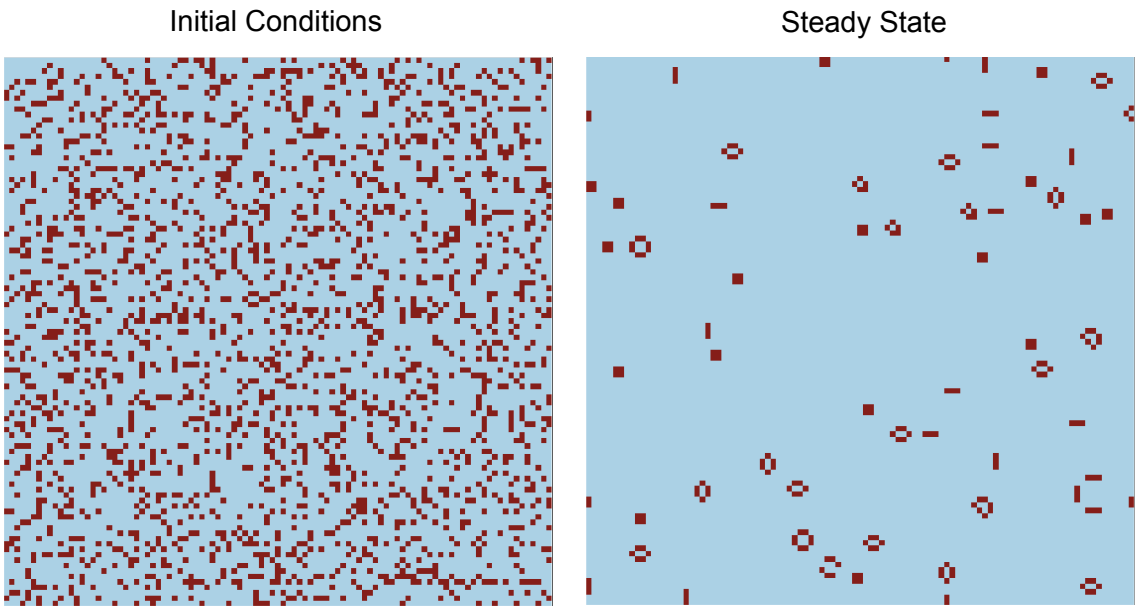


Figure 2.2: The game of life, red represents a living cell and blue represents a dead cell - images generated from NetLogo (Wilensky, 1999).

A further piece of influential work in the formulation of modern day ABS (based upon Cellular Automata) is that of the Schelling (1971) model. Schelling (1971) aimed to explore the effects of racial segregation in a community. A neighbourhood of houses was represented as a grid of cells, each cell being one of two possible types (representing two different racial backgrounds). The cells had a predefined level of tolerance to neighbourhood integration; if racial diversity in a cell's neighbours rose beyond a tolerable level, the cell could change position on the grid. The model demonstrated that even if agents had a small preference to their neighbours being of the same racial background, then segregation would occur. While initially no computer simulations were used in the creation of the segregation model, the ideas are said to be at the very foundations of ABS.

Cellular Automata have since been used to represent all manner of systems, including clouds (Nagel & Raschke, 1992), forest fires (Hernández Encinas et al., 2007) and HIV infection (Mo et al., 2014). However, the work of Reynolds (1987) marked a particular turning point in the development of ABS, through the creation of a model to represent flocking behaviour in birds. Reynolds (1987) removed the rigid cell structure of Cellular

Automata, allowing agents to inhabit 3-dimensional space - modern day ABS are able to represent agents on a grid, in 3-dimensions or in an environment where spatial capacity is irrelevant.

The historical development of ABS described herein only presents one facet of its evolution, with Epstein (1996) stating that ABS also draws upon cybernetics, connectionist cognitive science, distributed artificial intelligence, genetic algorithms and genetic programming. A wider perspective of the historical development of ABS may be found in Heath (2010). Throughout the literature presented, a resounding concept emanates: the actions of autonomous agents drives some emergent system behaviour; a discussion of how this is formulated in a simulation paradigm requires an analysis of the underlying theory of ABS.

Theory

ABS is a micro-simulation technique (Davidsson, 2001), which aims to model the *individual* behaviours of *specific* objects in a system. Agent Based Models are sometimes referred to as Individual Based Models in the study of ecology (Grimm et al., 2006), or Multi Agent Systems (MAS) - a term adopted predominantly in engineering. MAS is said to differ from ABS in that the focus is upon the development of operational agents to inform real world agents, as opposed to ABS, where the goal is to create agents which lead to an understanding of global phenomena (Niazi & Hussain, 2011).

There are three main components of ABS theory:

- **Agents** - The objects in the system being modelled are the agents. There is no commonly agreed precise definition of an agent, but Huhns & Singh (1998) state that "agents are active, persistent (software) components that perceive, reason, act and communicate". Davidsson (2001) expands this further, suggesting that agents may posses any or all of the following qualities to varying degrees:
    - *proactiveness* - reactions to the behaviour of other agents, or preventative actions to avoid certain situations;
    - *communication language* - the ability to send messages to other agents;

- *spatial explicitness* - the awareness of the spatial plane inhabited;
- *mobility* - the ability to move amongst the spatial plane;
- *adaptivity* - the ability to learn and/or change behaviours;
- *modelling concepts* - the consideration of personal beliefs, desires and intentions.

An ABS generally contains a large number of agents, contained within some environment, performing decisions or tasks predetermined by the modeller. An agent may be used to represent any object or entity, and the environment need not be spatial in nature. The specific qualities possessed by the agent depend upon the entities being modelled and the requirements of the simulation. Drogoul et al. (2003) argues that often an ABS does not possess any of the idealised properties described above; an ABS offering a convenient way to represent autonomous agents, without the agents themselves being remotely autonomous.

- **Emergence** - Emergence occurs when interactions at one level, give rise to behaviour at another level - requiring new categories of description that are not accounted for by the underlying components (Gilbert & Troitzsch, 2005). Within an ABS, this refers to the interactions and behaviours of the implemented agents causing some emergent behaviour, driving the overall system evolution (Macal & North, 2005).

  An example of emergent behaviour is a standing ovation following a theatre performance. Some individuals may have particularly enjoyed the performance, deciding to stand and show their appreciation. However, not all individuals may be compelled to stand by the performance, but do so anyway as a result of the actions of those around them. Therefore, a standing ovation emerges as a result of the actions of a number of individuals and the responses of those around them (Miller & Page, 2004).

- **Complexity** - A complex system results from the non-linear interactions of its constituent parts (Mitchell, 2009). It is a system in which there are organised but unpredictable behaviours, underpinned by dynamic networks of interaction; these be-

haviours and interactions appear greatly complex when viewed from a macro perspective (Miller & Page, 2007). The characteristics of a complex system at any one time may be directed by its history, or rather, the history of its constituent elements (Gilbert, 2004). In an ABS, the complex system is represented in terms of its agents, the agent being the primary focus of the modelling paradigm.

The perspective of an ABS is in stark contrast to SD (which takes a global view of the system) and provides a representation of entities which is not captured by DES. Further information regarding the difference between the simulation methods, may be found in Borshchev & Filippov (2004), with Siebers et al. (2010) giving a clear distinction between the entities of DES and ABS models. Simulation method selection is dependent upon the system being investigated, with each method yielding specific benefits. A discussion of the specific benefits of an ABS now follows.

Benefits of ABS

Aside from the benefits inherent with adopting a simulation approach in general (previously discussed in Section 2.1.1), Bonabeau (2002) states there are three specific advantages of selecting an ABS approach:

- **Modelling Emergence** - Many systems can be characterised by the emergent behaviour of its entities (Gilbert & Troitzsch, 2005; Regenmortel, 2004; Suweis et al., 2013). Emergent phenomena would be difficult to capture with alternative methods, making ABS the canonical approach to modelling such systems (Bonabeau, 2002). Furthermore, studying the behaviour of agents allows for an investigation of the relations between agents (O'Sullivan, 2004), which may also be an important factor of system evolution.

- **Natural Description** - ABS may provide a more natural view of a system. For example, the dynamics of pedestrians in a shopping centre may be more naturally expressed by individualistic decisions, as opposed to an overarching system dictating footfall. With regard to a population, aggregate approaches generally assume homogeneous mixing; however, there may be situations where an amalgamated view of a population is inappropriate (Axtell, 2000). For instance, in the spread of a sexually transmitted disease, modelling contact networks may be a more representative view

of disease transference. ABS is also said to be more intuitive than other approaches, especially when examining business processes. An abstract SD model of business flow may be not be intuitive for managerial staff and stakeholders, however, observing the process from an individualistic viewpoint may be simpler to conceptualise (Bonabeau, 2002).

- **Flexibility** - ABS offers a flexible framework to work with active entities. Agents can be added or removed from the system with ease, with the decision processes of agents capable of being highly simplistic or incredibly complex - as per the requirements of the model. ABS is also particularly useful when agents inhibit a geospatial platform (Axtell, 2000), with the ability to represent an environment as a discrete or continuous field (Helbing & Balietti, 2012).

ABS is also said to allow for detailed hypothesis testing, as the focus is upon specific aspects of an agent (Helbing & Balietti, 2012). As such, an ABS is most appropriate for domains characterised by a high degree of localisation (Parunak et al., 1998), whereby local actions impact the global system. While agents are a useful vehicle for understanding complex and non-linear systems (Ferber, 1999), there are also a number of limitations to consider.

Limitations

Many of the limitations discussed within the ABS literature relate to simulation methods in general, such as suitability, validity and detail. This may be particularly amplified with respect to ABS, as the individualistic processes of an agent may be difficult to quantify. Castle & Crooks (2006) argue that the surprising and counter-intuitive behaviours emerging from an ABS are rarely encountered in the real world, with Couclelis (2002) claiming that an ABS is sensitive to initial conditions and small variations in interaction rules. A further difficulty is the actual development of agents, with structural and decisional autonomy being difficult to achieve (Drogoul et al., 2003).

Axtell (2000) states that for an ABS to develop robust conclusions to theories, multiple runs are necessary - a result of the emergent behaviour potentially varying with the initial conditions selected. This may require a great deal of computational power, as agents are particularly 'memory hungry' (Siebers et al., 2010). This issue is being alleviated with

technological developments and distributed simulation approaches (discussed further in Section 2.2.4). Sophisticated ABS software packages are also available to aid in the design of a model, a brief discussion regarding ABS software may be found below.

Software

There are five main software platforms generally discussed in the development of an ABS:

- NetLogo (Wilensky, 1999);
- Repast (2013),
- SWARM (2012),
- MASON (2012),
- AnyLogic (2002).

These may be classified into two groups: Open Source Systems and Proprietary Systems. An Open Source System (OSS) is generally freely available, with access to the source code permitted. In terms of ABS, this is usually in the form of a toolkit, providing the appropriate libraries and routines to develop a model; there are said to be over one hundred toolkits available for ABS (Castle & Crooks, 2006). A Proprietary System (PS) is a software platform generally developed by an organisation who controls its licensing, with access to the source code strictly prohibited.

A brief introduction, and the positive and negative aspects of each platform, are as follows:

- **NetLogo** (OSS) - NetLogo is a high level toolkit which implements its own programming language to develop a model, whereby agents are referred to as turtles. NetLogo is programmed procedurally and does not adopt an object-oriented framework, with the software said to be highly accessible to modellers with little programming experience (Zhou et al., 2009). While a wealth of support documentation is provided, along with a vibrant online help community, NetLogo is limited in terms of functionality - although extension is possible through Application Programming Interfaces (API) (Castle & Crooks, 2006). NetLogo is used for the development of

the model discussed in Chapter 4.

- **SWARM** (OSS) - SWARM is a multi-agent platform developed predominantly for the investigation of complex biological systems (Minar et al., 1996). The platform is one of the earliest toolkits and at the time was said to be widespread and well known amongst the agent community (Hofmann & Carole, 2004). SWARM possesses moderate functionality and some demonstration models are also provided; however, Najlis et al. (2001) states that SWARM has a steep learning curve and requires an experienced programmer for effective use of the toolkit.

- **Repast** (OSS) - Recursive Porous Agent Simulation Toolkit (Repast) is available in three different programming languages: Java, Microsoft.Net and Python; however, new developments are solely released for the Java version (North et al., 2005). Repast is tailored to the development of social systems and contains a point and click GUI (Graphical User Interface) to aid model development (Railsback et al., 2006). Although the platform boasts an active online support community, accessibility for an inexperienced modeller can be problematic and documentation is often incomplete.

- **MASON** (OSS) - developed at George Mason University and based on Java. Zhou et al. (2009) states that while MASON has good extensibility, modularity and portability, its capabilities are not as comprehensive as other platforms, possessing little technical documentation and the requirement of a proficient programmer to develop a model.

- **AnyLogic** (PS) - developed by XJTechnologies, it supports the creation of ABS, DES and SD simulations. The system is based on Java, meaning that although the software and development framework cannot be shared without a licence, the self-contained simulations may be exported and demonstrated on unlicensed machines. AnyLogic benefits from powerful modelling capabilities, an intuitive interface and a professional support service (for a fee) (Zhou et al., 2009); however, specific capabilities depend on licensing agreements and only a small online community support the software in comparison to other platforms. AnyLogic is used in this thesis in the development of the simulation discussed in Chapter 6.

More detailed comparisons of each platform may be found in Castle & Crooks (2006), Nikolai & Madey (2009) and Zhou et al. (2009), assessing the software against differing criteria and from various perspectives. AnyLogic has a unique selling point in that it allows for the creation of simulations in ABS, DES and SD frameworks, and permits the development of simulations combining the methods. Such hybrid models are becoming a growing topic of interest amongst simulation literature. Further discussions regarding hybrid models, and technological advances in simulation, may be found in Section 2.2.4.

2.2.4 Advances in Simulation

While the three simulation methods discussed (SD, DES and ABS) offer unique perspectives on both the conceptualisation and design of a model, researchers have aimed to combine methods in the pursuit of more representative simulations. This allows for the investigation of larger complex systems which may be process driven in one sector, whist also demonstrating properties of emergence in another sector. Examples include the combination of: SD and DES in the investigation of Chlamydia infection (Viana et al., 2014), SD and ABS for the analysis of new healthcare technologies (Djanatliev et al., 2012) and DES and ABS applied to emergency healthcare services (Nouman et al., 2013). Additionally, Viana et al. (2012) discusses the combination of all three paradigms in the modelling of age-related macular degeneration.

Aside from the benefits of being able to model systems from varying perspectives, hybrid simulation also grants the potential for model reuse - described as a 'grand challenge' in modelling and simulation (Taylor et al., 2013). This allows pre-existing independent models of a system to be combined, generating greater overall system understanding. With the development of such models, and the design of more complex 'memory hungry' agent simulations in general, advances have also been made in dealing with the necessary computational demands. A grid or distributed approach to running simulations may be employed to expedite runtime (Kite et al., 2011; Mustafee & Taylor, 2009), although this often requires the need to overcome some technical barriers (Taylor et al., 2012). A further advantage of distributed simulation is the ability to combine models from multiple stakeholders, who may be concerned with the privacy of their data; in a distributed framework, the data need not be shared with all parties.

Discussions within Taylor et al. (2013) highlight the need for a new modelling and simu-

lation methodology to deal with complex multifaceted problems, stating that an effective approach for one problem domain may not necessarily translate to another. Therefore, a holistic approach to addressing the complexities of modern day modelling and simulation is required. Suggestions within Robinson et al. (2004) advocate the use of a web based simulation approach to address these issues, creating a domain for fast model building and easy experimentation. However, it would appear that web based simulation is very much in its infancy, with only a small number of tools supporting the platform (Byrne et al., 2010).

Overall, it would appear that technological advances in simulation are removing the rigidity of the modelling paradigm, allowing for the creation of complex models to more accurately represent all manner of systems. Perhaps as simulation literature evolves in the future, the notion of 'simulation type' will become irrelevant, echoing the sentiment of Bonabeau (2002) who states that actually, Agent Based Modelling is less of a technology and more of a mindset.

2.3 Applications of ABS

As ABS shall be the primary simulation method employed in this thesis, a review of its previous applications is required to provide context for the subsequent research. The applications of ABS are vast, set across a variety of fields and investigating all manner of problems. As such, this review of applications shall target four key areas pertinent to this investigation and is structured as follows: a broad outline of common ABS applications, demonstrating its suitability in the context of this thesis, is provided in Section 2.3.1; an introduction to applications of ABS in social theory is offered in Section 2.3.2; particular emphasis upon social networks is given in Section 2.3.3; Agent Based Simulations of smoking behaviour are presented in Section 2.3.4; finally, the conclusions drawn from the applicability of ABS are presented in Section 2.3.5.

2.3.1 Common ABS Applications

Bonabeau (2002) identifies four areas where the application of ABS may be particularly successful: flows, markets, organisation and diffusion. This is because each of these areas generally tends to foster emergent behaviours, resulting in a complex system when viewed from a macroscopic perspective. A brief outline of each specific application area is as

follows:

- **Flows** - generally relating to the flow of people in a system. The flow of individuals during an emergency evacuation is a particular topic of interest in ABS literature, due to individual behaviours that may cause panic and stampedes. Helbing et al. (2000) investigate the optimal strategy for evacuating a smoke filled room, Pan et al. (2007) examine competitive and herding behaviours in emergency evacuations, and Chen & Zhan (2008) research vehicle flow following an urban evacuation procedure. A survey of ABS in emergency response may be found in Hawe et al. (2012). ABS is also a useful tool for exploring pedestrian footfall in a system, examples including street structures (Jiang & Jia, 2011), museums (Pluchino et al., 2013) and theme parks (Huerre, 2010).

- **Markets** - ABS is used to investigate the complex adaptive systems of the financial markets (Tesfatsion, 2003). Examples include the creation of a simple ABS stock market (Palmer et al., 1994), investigating the agent processes of intraday trading (Kluger & McBride, 2011) and an ABS model of the NASDAQ stock exchange Outkin (2012). Agents have also been used to explore online auction behaviours (Mizuta & Steiglitz, 2000) and trading agreements (Bunn & Oliveira, 2001).

- **Organisation** - ABS is useful for the exploration of organisations, institutions and groups, due to the individual behaviours of entities that make up these collective systems. Organisations have been examined both in reference to their policies (Fioretti & Lomi, 2010) and particular structures (Ashraf et al., 2011; Pluchino et al., 2010, 2011), illustrating the diverse perspectives offered by ABS. Further details regarding organisational ABS may be found in Fioretti (2012), with additional discussions being presented in Section 4.1.

- **Diffusion** - the process of individuals being influenced by their social context. ABS has been applied to investigate the spread of opinions (van Eck et al., 2011) and the diffusion of new products/ideas (known as innovations) (Garcia & Jager, 2011; Zhang et al., 2011). A review of ABS applications to diffusion theory may be found in Kiesling et al. (2011). Diffusion in particular resonates strongly with the theme of this thesis - further discussions being presented in Section 3.2.2. Additionally, the data provided for this research is generated from a study based upon the principles

of innovation diffusion, considered further in Section 5.1.

Of course, ABS applications are not simply limited to the presented four topics. In many of the examples discussed, the focus is upon understanding the specific actions and responses of individuals in a given situation. This falls squarely in the domains of psychology and the social sciences, which have also adopted the techniques of ABS. As such, a brief review of social theory examined with ABS is presented in the following section.

2.3.2 Social Theory

ABS is described as a revolutionary development for social sciences, providing a natural approach to modelling social systems, free from the constraints of alternative modelling formalisms (Bankes, 2002). A particularly important application of ABS in social theory is the growing of 'artificial societies', whereby social structures and group behaviours are investigated from an individual agent level. Artificial societies may be used to explore concepts such as cultural transmission, combat and trade, which develop from basic agent-centric rules.

The 'sugarscape' is a classical example of an artificial society, developed by Epstein (1996). Agents exist in a spatial terrain called the sugarscape, with various parts of the landscape composed of high or low amounts of sugar. Each agent has a field of vision, the size of which is defined on creation of the agent. The agents must use their vision to find and eat sugar, but their travel to find sugar also burns energy - which can only be replenished through the consumption of more sugar. If an agent's energy level drops below a certain threshold, they die and become replaced by a new agent. This very simple system causes emergence both in terms of the agents' behaviour and the growth of sugar on the landscape, especially when the model is augmented with rules regarding gender, culture, conflict, trade and disease.

Further artificial societies are those of MANTA (Modelling an ANT hill Activity) and EOS (the Evolution of Organised Societies). MANTA simulates the birth of an ant colony, modelling the role of the queen and the other ants in the army. By giving the ants simple rules, cooperative behaviour emerges as the ants strive to keep the colony alive (Drogoul et al., 1995). EOS explores the growth of social complexity amongst humans in the upper Palaeolithic period in South West France. Changes during this period included hunting in

large groups, cave art and the development of status within a group. Through simulation of the environment, researchers attempted to unearth how interactions between individuals (and responses to the environment) drove the emergence of these organised societies (Doran et al., 1994).

The artificial societies investigated with ABS do not solely focus upon the vast domains of sugarscapes and the upper Palaeolithic period; further self-contained artificial societies investigated include: negotiating stakeholders in land development (Pooyandeh & Marceau, 2013), fraud in a shoe shop (Lopez-Rojas et al., 2013) and promotion in hierarchical corporate institutions (Pluchino et al., 2010). An artificial society of workplace interaction is created in Chapter 4.

The aim of creating an artificial society (with ABS) is to understand the interactions between agents; particularly in the social sciences, this is to investigate theories regarding human behaviour. Todd (1997) used ABS to investigate theories of optimal human mate selection, suggesting that taking "the next best mate" (according to some criteria) can lead to better matching than other theories of mate selection. Gotts et al. (2003) employs ABS in the study of social dilemmas, such as the 'prisoners dilemma' and the 'tragedy of commons'. Furthermore, Malleson et al. (2013) makes use of agents in the examination of policies relating to urban regeneration, exploring the resultant effect upon burglaries and home invasions.

The articles discussed demonstrate the breadth of social theory that may be examined with ABS - further examples being given within Gilbert (2007, 2008) . An area of social theory particularly suited to an ABS framework is that of social networks, due to the ability to explicitly represent network structures in a model. A brief review of ABS applied to social networks is given in Section 2.3.3.

2.3.3 Social Networks

ABS investigations related to social networks have covered a variety of topics. Epidemiology in particular has adopted ABS techniques to explore the spread of infectious diseases through networks, including HIV spread in Amsterdam (Mei et al., 2010a), Influenza in a metropolitan social network (Mao, 2014) and H1N1 on a Chinese university campus (Mei et al., 2010b). The work of Eubank et al. (2004) examines the general spread of disease

amongst urban social networks, with further information regarding the role of social networks and ABS in epidemiology found in El-Sayed et al. (2012). Aside from disease, as previously discussed, networks relating to business (Prenkert & Føgesvold, 2014) and land usage (Ronald et al., 2012) have also been explored with ABS.

ABS has also been used in the investigation of network structure, as opposed to its effects. Pujol et al. (2002) uses agents to extract *reputation* in a social network topology, Han et al. (2014) explores hierarchical geographical network structures and Bernstein & O'Brien (2013) uses ABS to generate 'realistic' social network data sets. A review of networks in ABS, particularly applied to social systems, may be found in Alam & Geller (2012) - examining areas of implementation and validity of the models. A further area of research employing simulation techniques, applied to social networks, is Stochastic Actor Based (SAB) modelling; discussions regarding SAB may be found in Section6.1.3.

Overall, it would appear that ABS allows for the investigation of social networks in all manner of applications. El-Sayed et al. (2012) states that while ABS and Social Network Analysis (SNA) techniques are becoming increasingly widespread, continued development of both approaches is required. Furthermore, Macy & Willer (2002) express the view that rich sociological research may be conduced with ABS (in conjunction with network topology), suggesting social scientists move away from examining *factors* and focus on *actors* (agents). As this thesis also explores the social aspect of smoking behaviour, a review of ABS specifically related to smoking is conducted in Section 2.3.4.

2.3.4 Smoking

Simulation in general has been applied to a number of aspects related to smoking behaviours, with examples including: spreadsheet models of smoking uptake and health effects in a population (Levy & Friend, 2002; Near et al., 2013), and an SD model of planning and evaluating healthcare intervention strategies (Homer et al., 2010). Further examples of general simulation-based smoking investigation may be found in Verzi et al. (2012).

In terms of ABS, the 'Population Structure Model' (PSM) discussed in Verzi et al. (2012) uses agents to model the health effects of changing patterns of smoking behaviour. Each agent is characterised by intrinsic values, behavioural states and health states - state changes

being established upon empirical data. As the model progresses, agents make decisions regarding smoking behaviours which impact upon their health state, generating overall predictions of smoking population size. The PSM population predictions made are said to be consistent with those of US census projections, estimating a decline in smoking rates to 12.5% in 2050. The PSM is a demonstration that by modelling the individual behaviours of agents, global conclusions may be drawn.

A further smoking related ABS employs the method to examine the adoption of anti-smoking legislation in conjunction with individual cultural norms - aiming to explore reasons for the ineffectual enforcement of smoking bans in particular EU countries (Dechesne et al., 2012). Additionally, Song (2006) investigates smoking addiction and cessation, representing adolescents as agents who gain a particular utility (based upon factors relating to physiology, psychology and genetics) in relation to smoking. More generally, Andrighetto et al. (2013) examines social norms with ABS - a social norm being an accepted behaviour within a certain group.

The selection of ABS smoking research presented does not include any explicit network structures. This is of particular interest given both the widely perceived social aspect of smoking (introduced in Section 1.1.3) and the specific aims of this thesis. Given the applicability of ABS both in the context of social networks and smoking behaviours, this offers the potential to investigate the interplay between these domains in an ABS framework - a task which does not appear greatly explored in the literature reviewed.

2.3.5 Conclusions

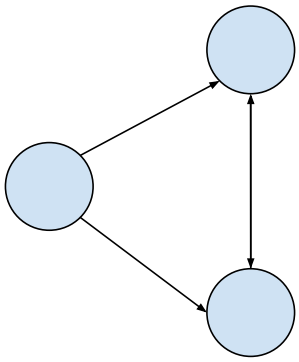
This review of applications has demonstrated the context in which ABS should be applied, highlighting systems with emergence, diffusion and network properties as being particularly well suited to the paradigm. As the systems discussed in this thesis - work place promotion dynamics and adolescent social smoking - have complex outcomes based upon the behaviours of individuals, it would appear that ABS is an appropriate choice of simulation method for this investigation. The ability to model network structure is also identified as a particular strength of the method - a review of network science literature being conducted in Chapter 3.

2.4 Chapter Summary

This chapter has provided a review of both historical and current literature relating to simulation. Section 2.1 introduced the concept of computer simulation, outlining the key elements of its theory: systems, simplified imitations, time, experimentation and understanding/improving. Additionally, the benefits and limitations of simulation were presented, culminating in a seven step guide for the creation process: problem formulation, information collection, assumption development, model programming, verification and validation, experimentation and presentation/implementation.

Section 2.2 explored three different types of simulation: System Dynamics, Discrete Event Simulation and Agent Based Simulation. SD takes a global view of system, modelling the effects of a structure upon its 'continuous' entities. DES takes a process orientated view, allowing for the explicit definition of discrete entities and the exploration of their circulation through a system. ABS does not have an overarching system representation, the model dynamics being driven solely from the emergent properties of the agents' actions. Advances in both simulation theory and technology were also discussed, presenting the steps being made to address computational issues for the ever growing complexity of models.

Section 2.3 presented a plethora of current applications of ABS. In particular, systems demonstrating properties relating to flows, markets, organisations and diffusion achieve prominence in the literature. Articles relating to social networks and smoking were also explored, highlighting the lack of investigations combining both topics in an ABS framework. This review has set the context for the simulation aspect of this research, presenting the applicability of ABS to the proposed research aims. Building upon this, Chapter 3 now introduces the essential network science literature that also underpins the ensuing investigation.



-*"A Transitive Triple"*

3

Network Literature Review

This chapter introduces the essential graph theoretic and network science literature that informs the research conducted in this thesis. As this study is concerned with the investigation of social networks, and ultimately the development of a new algorithm to predict social network evolution (PageRank-Max), the relevant metrics to analyse and interpret network structure are required; the network metrics of particular interest, along with a brief history of graph theory, are detailed in Section 3.1.

Following the introduction of the appropriate graph theory, Section 3.2 reviews the historical development of network science as a discipline. Two distinct lines of research underpin modern day network science, topology, which is concerned with the structural configuration of networks, and connection effects, which examines the impact of having connections; a brief review of both elements is conducted in Section 3.2. Modern day network science is a broad discipline, as such, Section 3.3 focuses upon literature relating to the investigation of social networks; this being of particular relevance to the investigation conduced herein.

As outlined in Chapter 1, this thesis shall employ methods relating to the Link Prediction problem. Section 3.4 outlines the concept of Link Prediction (LP) and details its development as a well defined sector of literature, giving examples of its prior applications; a further, more detailed review of LP algorithms is conducted in Chapter 6. Finally, Section 3.5 details the specific notation used to refer to graphs in this thesis, introducing the selected graph visualisations method to be used with the available network data of Chapter 5.

3.1 Graph Theory

Network science as a concept is termed to be an emergent field of study, a topic described as a "new science" with development origins rooted in the 1990's (Lewis, 2011). However, at its core are the central components of graph theory, a topic researched long before the 1990's.

Euler's work relating to the 'Seven Bridges of Kösnigsberg' problem is widely cited as the inauguration of graph theory and topology in mathematics (Boccaletti et al., 2006). The problem examines the seven bridges that cross the city of Königsberg in Prussia (now known as Kaliningran, Russia), Figure 3.1. Every bridge must be fully traversed in sequence, without retracing any path previously travelled - Euler (1736) proving a solution could not be found.

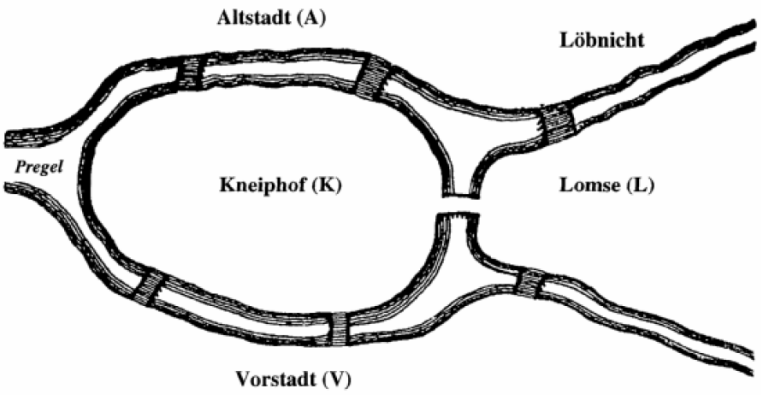


Figure 3.1: The seven bridges of Kösnigsberg as presented in Euler (1736), extracted with annotation from Gribkovskaia et al. (2007)

Nearly a century later, Francis Guthrie proposed 'The Four Colour Problem' when attempt-

ing to colour a map of English counties (Wilson, 2002). His conjecture states that given a simple planar map, four colours will suffice to ensure the unique colouring of contiguous adjacent regions. Attempts at proving what may appear to be a simplistic problem at face value, have a chequered history - the problem being first conceived in 1852, brought to the London Mathematical Society by Cayley in 1878, incorrectly proven by Kempe (1879) and eventually proven (with the aid of a computer) in Appel & Haken (1977). An example of a planar graph requiring four colours is given in Figure 3.2.

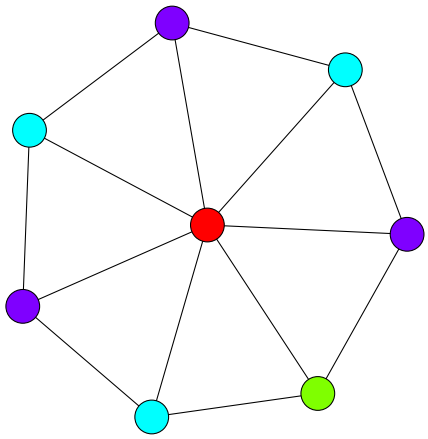


Figure 3.2: A planar graph with 8 nodes requiring 4 colours.

The examples discussed by Euler and Guthrie, demonstrate the importance of network structure; as such, a number of graph theoretical measures are regularly used to examine graphs in more detail. The following sections outline specific metrics that shall be required throughout this thesis, classified as follows: the basic concept of a graph (Section 3.1.1); network cohesion and connectivity (Section 3.1.2); the clustering of nodes (Section 3.1.3); the paths between nodes (Section 3.1.4); and nodal specific measures of centrality (Section 3.1.5).

3.1.1 The Graph

The term 'graph', in the context of a network of objects, appeared many years after the initial problems of Euler and Guthrie; 'graph' itself was coined by Sylvester (1878) in reference to molecular diagrams. A more formal definition of a graph is as follows:

Definition 3.1.1. *An undirected* ***Graph*** *is defined as a pair* $G = (V, E)$ *of sets such that* E *is a subset of the unordered pairs of* V*, where* V *is the set of vertices (or nodes) and* E

represents the set of edges (or links). A directed graph (or digraph) may be defined in the same manner, except that E is a subset of the ordered pairs of V.

The *order* of G is defined as the number of elements in the set of vertices V, denoted by $|G|$; thus $|G| = |V(G)|$ (Bollobas, 2013). For simplicity, the number of vertices for any particular graph G, shall be referred to as n.

A social network may be represented as a *directed* or *undirected* graph. A directed graph offers a rich source of information, both in terms of the qualitative implications of friendship, and the quantitative metrics of network calculation. Figure 3.3 presents the Petersen (1898) graph in a directed and undirected manner.

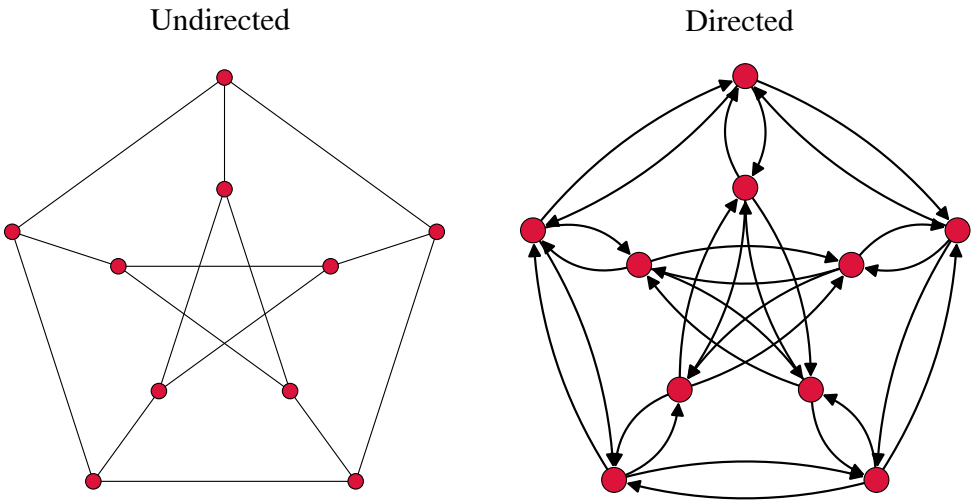


Figure 3.3: Undirected and directed Petersen graphs.

For an undirected graph, an edge $\{i, j\}$ links the vertices v_i and v_j and may be represented by ij. A directed network edge preserves the order by which a link is made, such that an edge $\{i, j\}$ implies a link from v_i to v_j is denoted by $i \rightarrow j$, therefore it cannot be assumed the link $j \rightarrow i$ exists. A number of the metrics defined in subsequent sections, require the maximum number of edges ($e_{\max}$) of a graph; for an undirected graph, $e_{\max} = \frac{n(n-1)}{2}$, and for an directed graph, $e_{\max} = n(n-1)$. With the basic elements of a graph defined, Section 3.1.2 explores the cohesive properties of a graph.

3.1.2 Network Cohesion

Network cohesion focuses upon the extent to which a graph is interlinked (Moody & White, 2003). Three graph properties are defined in this section:

- average degree;
- reciprocity;
- density.

For illustrative purposes, the example network of Figure 3.4 has been created; at the end of this (and each subsequent) section, the metrics introduced shall be calculated in reference to the example network.

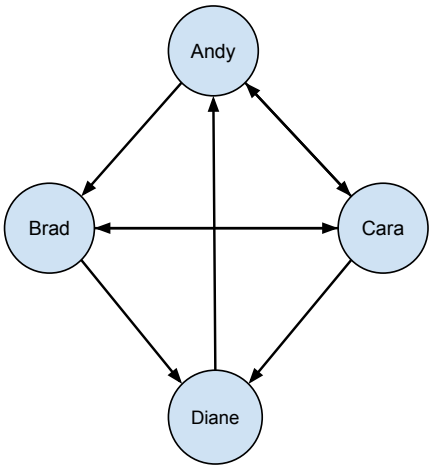


Figure 3.4: Example directed network.

Before examining the network cohesion properties of Figure 3.4, the notion of node *degree* must first be introduced. This is the simplest of nodal metrics, defined as:

Definition 3.1.2. *The **degree** of a vertex v_i is denoted as $deg(v_i)$ and represents the number of incident edges of v_i. In a directed network, these may be separated further in terms of **in-degree** $deg(v_i)_{in}$ and **out-degree** $deg(v_i)_{out}$, defined as the count of the inward links and outward links of v_i respectively (Newman, 2003).*

In terms of network cohesion, and a representation of the graph as a whole, the *average vertex degree* may therefore be calculated by:

$$\frac{\sum_{i=1}^{n} \deg(v_i)}{n} \tag{3.1}$$

where $\deg(v_i)$ is replaced by the directed network equivalent (if required). The number of out-degrees and in-degrees for each node in Figure 3.4 are presented in Table 3.1.

	Out-Degree	In-Degree
Andy	2	2
Brad	2	2
Cara	3	2
Diane	1	2

Table 3.1: The in-degree and out-degree values of the nodes from Figure 3.4.

A directed graph's in-degrees and out-degrees allows for incident edges to become unreciprocated. In terms of a social network, this could suggest the node v_i extending a link to v_j but the link $j \to i$ not being in existence. This provides a representation of network cohesion, termed reciprocity:

Definition 3.1.3. *A **reciprocated** tie is one in which for the vertices v_i and v_j, the links $i \to j$ and $j \to i$ exist. The overall **reciprocity** of the directed graph G is said to be:*

$$r = \frac{|L|}{|E|} \tag{3.2}$$

where L is the set of edges involved in reciprocal ties. As such, $r \in [0, 1]$, meaning that $r = 1$ signifies a fully reciprocated graph (Newman et al., 2002a).

An alternative calculation method has been proposed to that of Definition 3.1.3, expressing reciprocity as a correlation coefficient of the associated network adjacency matrix (Garlaschelli & Loffredo, 2004). This new index of reciprocation is said to combat the issue of relative meaning - networks often having to be compared against a random counterpart to achieve some form of benchmark (Costa & Rodrigues, 2007). As the work of this thesis aims to compare a variety of different networks against other similar networks, the issue of relative association does not factor as strongly; therefore, the basic reciprocity definition will be the adopted standard.

A further measure of graph cohesion, complementary to that of reciprocity, is the notion of network density:

Definition 3.1.4. *The **density** is regarded as the overall connectivity of a graph. It is defined as the proportion of present edges $|E|$ to the number of potential edges e_{max}:*

$$d = \frac{|E|}{e_{max}} \tag{3.3}$$

where $d \in [0, 1]$ (Wasserman & Faust).

Social networks are often reported as having low density, termed as *sparse* networks, a feature said to be one of the seven general characteristics of a social network (Bruggeman, 2013). Furthermore, density appears to be independent of size (Kunegis, 2007); social networks with varying structures may have the same density (Niemeijer, 1973).

Table 3.2 displays calculations for each of the network meaurses defined (average degree, reciprocity and density), in reference to the graph of Figure 3.4. While the definitions of 3.1.3 and 3.1.4 serve to provide an overall picture of connectivity within a graph, they do not provide a detailed account of link configuration. For a more indepth examination of the types of structures present in a network, one is required to examine vetrex grouping (Section 3.1.3).

Network Measure	Value
Average In-Degree	2.00
Average Out-Degree	2.00
Reciprocity	0.50
Density	0.67

Table 3.2: Network cohesion measures for the graph of Figure 3.4

3.1.3 Network Clustering

Network clustering examines the specific types of connections present between nodes on a graph. The most basic relation is that of the reciprocated dyad (Figure 3.5), yet copular dynamics only present one aspect of vertex link structures. Nodal groups of three or upwards of four, known as 'triadic' relations and 'cliques' respectively (Figure 3.5), may also be apparent in social networks. As such, further metrics to express the prevalence of

said grouping behaviour are also required.

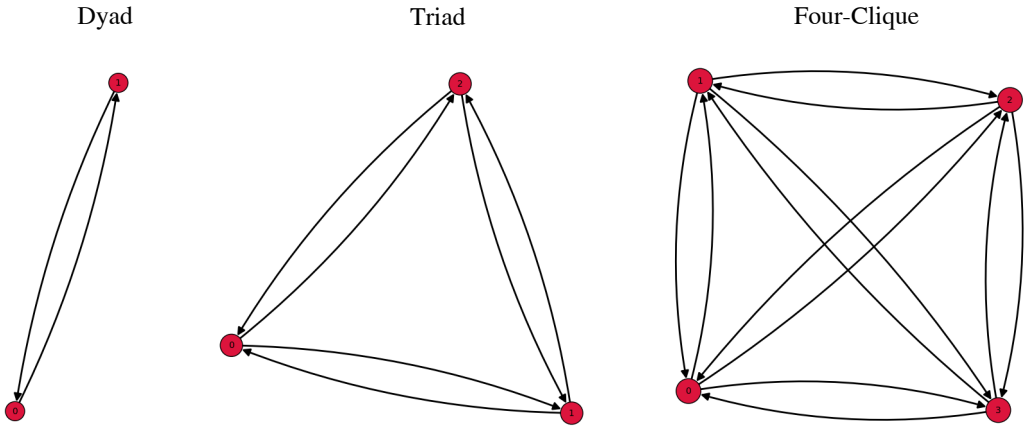


Figure 3.5: Reciprocated dyad, triad and four-clique directed graphs.

Two metrics are to be defined in the class of clustering relations, that of *transitivity ratio* and *clique number*. A transitive triple is defined as the ordering by which three elements on a graph connect with one another; an image of the four possible transitive configurations of three nodes appears in Figure 3.6. The analysis of node triples may take a variety of different forms, the *transitivity ratio* being of particular importance in the ensuing research.

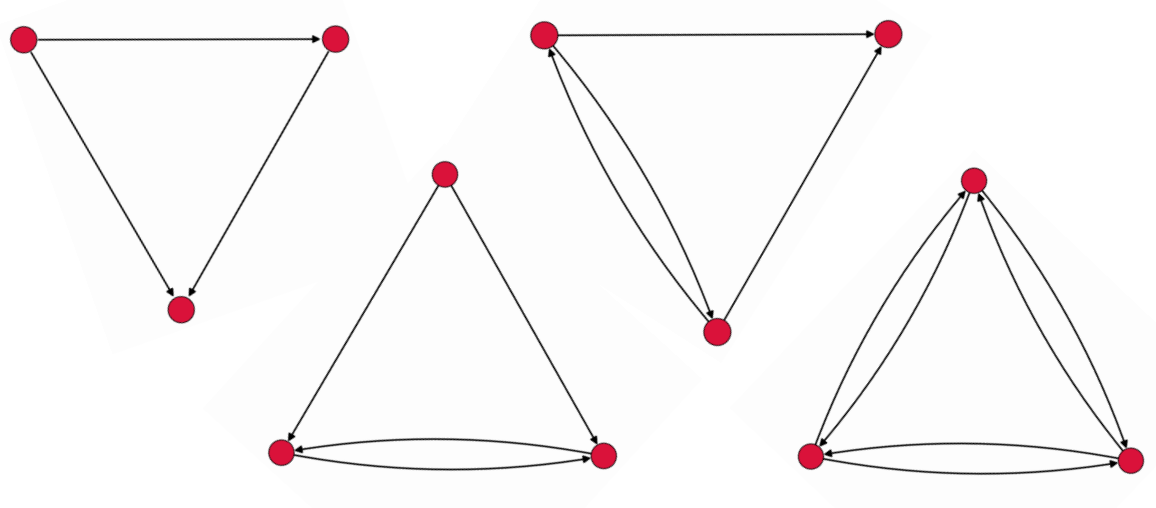


Figure 3.6: There are sixteen possible link arrangements of three nodes, presented are the four that are said to be transitive (Wasserman & Faust). A transitive triple is collection of nodes ensuring that "a friend of a friend is always a friend".

Definition 3.1.5. *For a directed graph, a* ***transitive triple*** *is defined to be a sequence of edges such that $i \to j$, $j \to k$ and $i \to k$ exist (Wasserman & Faust). A* ***subgraph*** *is defined*

*as $G' = (V', E')$ of $G(V, E)$ if $V' \subset V$ and $E' \subset E$. In an undirected graph, a **triangle** may be considered as a complete subgraph containing three nodes of G, where the number of triangles containing v_i is defined to be $\delta(v_i) = |\{\{v_i, v_j\} \in E : \{v_j, v_k\}, \{v_i, v_k\} \in E\}|$ (Schank & Wagner, 2005). The number of all possible triangles in G is denoted by $\tau(G)$, therefore the **transitivity ratio** $T(G)$ may be calculated by:*

$$T(G) = \frac{\sum_{i=1}^{n} \delta(v_i)}{\tau(G)} \tag{3.4}$$

For a directed graph, edges are converted into undirected associations (Luce & Perry, 1949).

This measurement essentially calculates the proportion of "closed triangles" of nodes, in relation to all connected triples of nodes. This gives a representation of how clustered the network is, offering an indication of mutual relations. Other interpretations of graph transitivity have been suggested; for example, the global clustering coefficient and the local clustering coefficient (Watts & Strogatz, 1998) - both of which are said to suffer from bias (Soffer & Vázquez, 2005). Given its overall simplistic and effective nature, coupled with the avoidance of inherent bias associated with other methods, the transitivity ratio has therefore been selected as the metric of choice for quantifying clustering within this research.

While a triplet of nodes may offer information regarding the commonality of ties, node n-tuples demonstrate an even broader scope of collective network clustering. In terms of a social network, such configurations present important behavioural implications regarding a vertex grouping; for example, the underlying rationale behind the formation of observed structures (Kandel, 1978; Parker & Asher, 1987; Paxton & Schutz, 1999). Prior to examining the implicit aspects, one must first define explicitly the constituents of such vertex compositions:

Definition 3.1.6. *A **clique** is a subgraph of three or more nodes, where each node is connected to all other nodes. No node extraneous to the clique may have a fully reciprocated relation with all members of the clique. The **clique number** $\omega(G)$ of a graph is defined as:*

$$\omega(G) = |H'| \tag{3.5}$$

where H' is the largest clique in G. To become a member of a clique in a directed graph,

ties must be fully reciprocated; should graph reciprocation be minimal, clique formation will also be minimal (Harary, 1994; Luce & Perry, 1949).

Following the definition of the two network clustering measures considered in this thesis (transitivity ratio and clique number), the metrics may be calculated for the example graph of Figure 3.4. The transitivity ratio is calculated as $T(G) = 1$, due to the metric ignoring directionality; as such, the number of closed nodal triangles is equal to the number of possible network triangles. The clique number $\omega(G) = 0$, as there are only four nodes in the network; all connections are not reciprocated, therefore, a clique is not present. The paths between nodes may now be considered in Section 3.1.4.

3.1.4 Paths

Cliques provide a tangible demonstration of the interconnectedness of vertices; as such, a route between clustered nodes may be forged. Travelling a concourse of nodes via a graph's incident edges is described as navigating a *path*, the definition of which is as follows:

Definition 3.1.7. *A **path** is a graph P of form $V(P) = \{v_0, v_1, ...v_l\}$, with edges $E(P) = \{v_0v_1, v_1v_2, ..., v_{l-1}v_l\}$, denoted by $v_0v_1...v_l$. The end vertices are v_0 and v_l, therefore the path may be denoted by $v_0 - v_l$. In a directed graph, the direction of the edges dictate the direction of the path (Bollobas, 2013).*

The path of a network plays an important role in the description of reachability between nodes. For example, if a path exists between the nodes v_i and v_j then these nodes are said to be *reachable* (Holme, 2005). In a fully connected graph, every node is reachable. Social Networks are unlikely to ever achieve complete reachability, even less so if the network is directed (Barabási et al., 2000). To garner an overall picture of the reachability between paths of nodes, one must consider the *geodesic* - the shortest path connecting two vertices v_i and v_j (Harary, 1994). The graph metric *average shortest path length* may then be calculated:

Definition 3.1.8. *The **average path length** (APL) l_G for G is described as the shortest distance between the nodes v_i and v_j, denoted as $d(v_i, v_j)$, divided by the maximum possible number of edges (e_{max}) Newman (2001). A disconnected APL assumes $d(v_i, v_j) = 0$ if*

$v_i = v_j$ *and* $d(v_i, v_j) = n$ *if* v_i *cannot reach* v_j *. Therefore:*

$$l_G = \frac{\sum_{i \neq j} d(v_i, v_j)}{e_{max}} \tag{3.6}$$

APL is a robust measurement of network topology, often quoted as the main factor in the classification of network type (discussed further in section 3.2.1) (Fronczak et al., 2004). For the network of Figure 3.4, $l_G = 1.33$ (2 d.p.) - meaning that the path from any v_i to any other v_j must traverse (on average) 1.33 nodes. In some instances, it may be useful to calculate the normalised APL ($\widetilde{l_G}$) :

$$\widetilde{l_G} = 1 - \frac{l_G - 1}{n - 1} \tag{3.7}$$

Therefore, as $l_G \to 1$ (where $l_G = 1$ indicates a fully connected network), $\widetilde{l_G} \to 1$; this metric is particularly important when comparing networks with different values of n. For the network of Figure 3.4, $\widetilde{l_G} = 0.89$.

The path based metrics presented, along with the measures introduced in Section 3.1.2 and Section 3.1.3, provide an analysis of the higher level elements of a graph. As the work contained within this thesis shall also investigate individuals in a network, it is of interest to examine nodal specific metrics - presented in Section 3.1.5.

3.1.5 Individual Cohesion

Individual cohesion measures focus on the representation of vertex placement in the framework of a network. There are three specific measures of *centrality* that attempt to quantify this, each offering a different perspective of nodal positioning (Wasserman & Faust). To illustrate the individual cohesion metrics, the undirected example network of Figure 3.7 shall be referred to throughout this section - with calculations being provided in the concluding remarks.

The first centrality measure presented, uses the definition of degree (Definition 3.1.2) to calculate how central a vertex may be in a graph:

Definition 3.1.9. *The* ***degree centrality*** C_B *is a nodal specific measurement, calculated as*

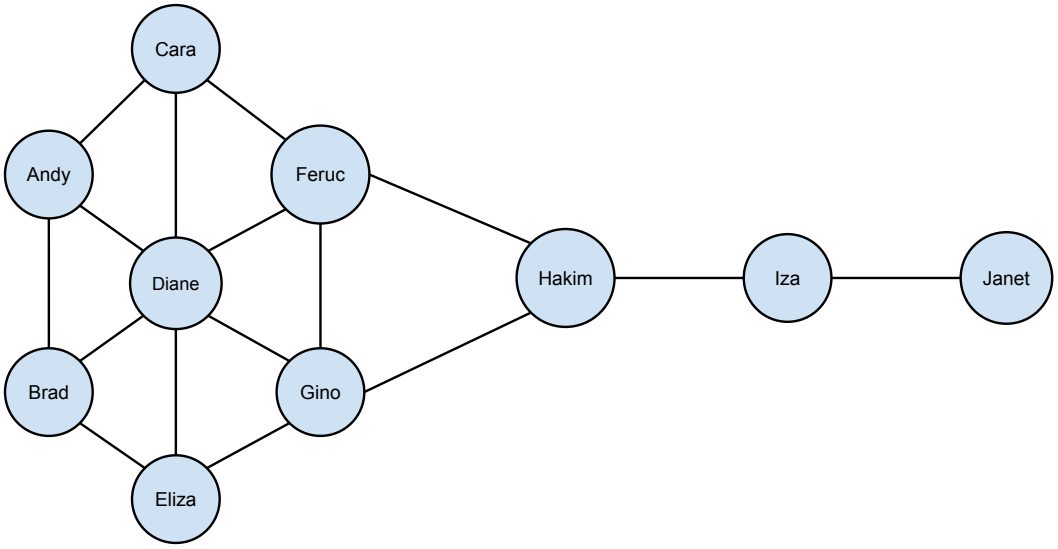


Figure 3.7: Example graph to depict differing individual cohesion measures, adapted from Krebs (2013).

the degree of the node proportional to maximum possible degree:

$$C_D(v_i) = \frac{deg(v_i)}{n-1} \tag{3.8}$$

For a directed graph $deg(v_i)$ *is replaced with either* $deg(v_i)_{in}$ *or* $deg(v_i)_{out}$ *(Proctor & Loomis, 1951; Wasserman & Faust).*

This is the simplest definition of centrality, arguing that a central node must have many edges emanating from it - Figure 3.7 demonstrates that "Diane" is the most central in the depicted graph topology. The directed properties of degree centrality have slightly adjusted meanings, in-degree presenting node *prominence* (Alexander Jr, 1963) (or popularity) and out-degree manifesting as *influence* (or the ability to diffuse information quickly) (Lin, 1976).

The second in the presented series of metrics defining centrality, explores the distances between nodes - the notion that a vertex is central if it may access all other nodes quickly (Wasserman & Faust). The concept developed by Sabidussi (1966) states:

Definition 3.1.10. *The* ***closeness centrality*** C_C *of a node, measures vetex closeness as an inverse function of its geodesics. As previously defined, the distance from* v_i *and* v_j *is*

denoted by $d(v_i, v_j)$, as such the closeness centrality for a disconnected graph is :

$$C_C(v_i) = \left[\sum_{j=1}^{n} d(v_i, v_j)\right]^{-1} \tag{3.9}$$

such that $d(v_i, v_j) = 0$ if $v_i = v_j$ and $d(v_i, v_j) = n$ if no path links v_i and v_j. This is normalised as:

$$C'_C(v_i) = (n-1)C_C(v_i) \tag{3.10}$$

where where $C'_C(v_i) \in [0, 1]$.

Figure 3.7 illustrates that the nodes "Gino" and "Feruc" are the *closest* to all other nodes, possessing the shortest overall path to all other vertices. Such a measure may be of key importance in a social network in terms of communicating information, as theorised by Bavelas (1950) and Leavitt (1951). This measure also demonstrates how a node may be central in terms of *degree*, yet not necessarily close to *all* other vertices in the graph.

The final nodal specific measurement presented in this section, is the concept of *betweenness*. This formalises a form of brokerage in the network (Friedkin, 1991), exploring how non-adjacent vertices may communicate through the vertices along the path that lies between them. Shimbel (1953) identified the importance of betweenness, a definition of which is as follows:

Definition 3.1.11. *The **betweenness centrality** C_B of a vertex is calculated as the number of shortest paths from all vertices to all other vertices in G, passing through the vertex v_i, divided by all pairs of vertices (not including i). Let g_{jk} be the number of geodesics between the vertices v_j and v_k, and $g_{jk}(v_i)$ represent the number of geodesics linking the nodes that include v_i. The betweenness centrality of a node is:*

$$C_B(v_i) = \sum_{i \neq j \neq k} \frac{g_{jk}(v_i)}{g_{jk}} \tag{3.11}$$

and normalised as:

$$C'_B(v_i) = \frac{C_B(v_i)}{e_{max}} \tag{3.12}$$

where $C'_B(v_i) \in [0, 1]$ (Freeman et al., 1977).

"Hakim" (in the example network of Figure 3.7) has the highest betweenness, linking

	Degree	Closeness	Betweenness
Andy	0.33	0.43	0.01
Brad	0.33	0.43	0.01
Cara	0.33	0.50	0.04
Diane	0.67	0.60	0.28
Eliza	0.33	0.50	0.04
Feruc	0.44	0.60	0.22
Gino	0.44	0.60	0.22
Hakim	0.33	0.53	0.39
Iza	0.22	0.39	0.22
Janet	0.11	0.29	0.00

Table 3.3: Centrality scores of the nodes in the network of Figure 3.7.

"Janet" and "Iza" to the network of "Feruc" and "Gino". Wasserman & Faust and Freeman (1979) state that nodes which regularly appear in the shortest path connecting all other nodes, are able to express more "interpersonal" influence - therefore making them more central. The individual centrality figures for all nodes in Figure 3.7 are presented in Table 3.3, illustrating the difference in centrality classifications.

Centrality may be defined in a variety of ways, with the literature expanding further to develop, new more complex concepts (Kretschmer & Kretschmer, 2007; Opsahl, 2013; Opsahl & Panzarasa, 2009). Given that the nature of this research is to explore the individual cohesion of social networks, as well as network cohesion as a whole, the defined measures will suffice to give a representation of the graph theoretic concepts pertinent to social networks. Further information regarding Social Network Analysis (SNA) metrics as a whole, may be found in Wasserman & Faust.

3.2 History of Network Science

The following sections (3.2.1 and 3.2.2) outline many of the important literary developments that form the foundations of network science. The literature may be classified into two distinct sectors, network topology and connection effects. The topological work examines the mathematical structures of networks and their construction, while the connection effects section focuses on the early work relating to the impact of having connections.

3.2.1 Topology

Three main structures are documented within network science topological literature: random, scale-free and small world graphs. This section discusses each of these structures, describing their construction and their relevance within network science literature. The structures presented are later utilised in Chapter 4 in the preliminary investigation of the effect of social structures upon individual behaviours. The first model to be considered is that of the basic random graph.

Random Graphs

The work of Euler and Guthrie demonstrates that structural configuration may be of key importance in the resolution of a graph theoretical problem; however, early topological research focused predominantly on that of random graph composition. To investigate the properties of graphs further, researches were first required to generate graph structures. As such, Erdos & Renyi (1959) and Gilbert (1959) developed random graph models.

In the random graph models of Erdos & Renyi (1959) and Gilbert (1959), the probability of a link occurring (in an undirected graph) between the vertices v_i and v_j (p_{ij}) follows a uniform distribution, such that:

$$p_{ij} = \frac{1}{n-1}. \tag{3.13}$$

However, the method by which Erdos & Renyi (1959) and Gilbert (1959) suggested generation of these graphs differed. The method proposed by Erdos & Renyi (1959) is said to be favourable, due to its ability to produce multiple graphs exhibiting an equal number of incident edges (Igor et al., 2010). Whereas the Gilbert (1959) model, in a graph with n nodes, would simply include each edge with probability p, independent from every other edge. An example random graph is present in Figure 3.8.

Barabási & Frangos (2003) argued that uniform random models do not accurately capture real world connection, with networks being goverened by "robust organising principles". Alternative graph models have thus been suggested, which are said to elicit a more demonstrative representation of connection dynamics (Newman et al., 2002b) - such as those of the scale-free and small world networks.

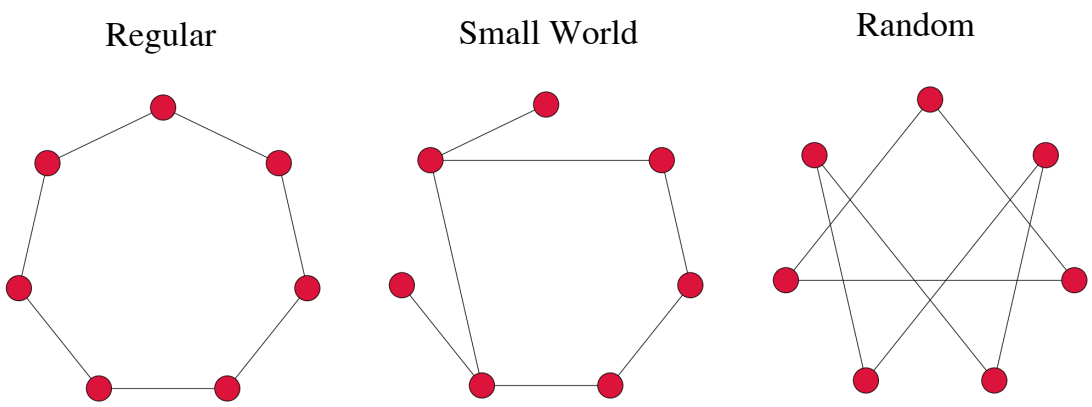


Figure 3.8: Seven vertices arranged in a regular, small world (p =0.3) and random network formations.

Scale-free

The work of Yule (1925) forms the foundations of scale-free networks. Yule (1925) found that, on examination of the evolution of flowering plants, the probability of a species generating a new offspring, was based upon the number of offspring the species already had; this came to be known as a preferential process, and can be thought of as "the rich get richer". This preferential structure implies a non-random process, and was also observed in the word frequency of documents by Simon (1955).

In terms of connection, a preferential style of attachment assumes that nodes tend to link with nodes who already posses a large number of connections. The work of de Solla Price (1965) observed this stochastic process on examination of connections between scientific literature citations, formulating the cumulative attachment model of de Solla Price (1976) - said to be characterised by a power law distribution. This emergent behaviour was later defined as a class of graphs known as scale-free networks (Barabási & Bonabeau, 2003).

In a scale-free network, the probability of a node being connected to k other nodes is described as:

$$P(k) \sim \frac{1}{k^m} \tag{3.14}$$

where typically the scaling factor $m \in [2,3]$. From this it can then be inferred that the

preferential style of link attachment is such that the probability that a randomly chosen node is connected to the i-th node is:

$$P(i) \sim \frac{k_i}{\sum_j k_j}, \tag{3.15}$$

k_i representing the degree of the node v_i (Barabási & Albert, 1999; Barabási et al., 1999).

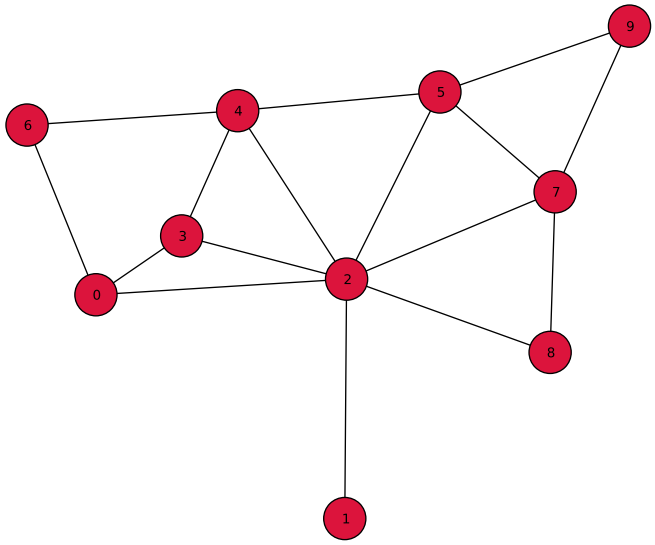


Figure 3.9: Scale-free network with ten vertices, node "2" being the largest hub.

The preferential attachment model causes certain vertices within a graph to become highly linked *hubs* (Barabási, 2009), an example given in Figure 3.9. Due to this dependence upon a selection of nodes, the removal of a hub may result in the creation of a disconnected graph. Many networks are said to exhibit scale-free properties: examples include sexual contacts (Liljeros et al., 2001), the rise of English Protestantism (Ormerod, 2008) and airline flight networks (Guimerà et al., 2005); however, scale-free properties were first observed upon the network of websites forming the internet.

Initiated in December 1970 under the acronym ARPANET (Advanced Research Projects Agency Network), the Internet's growth from its original thirteen websites (nodes) has intrigued researchers (Kleinberg & Easley, 2010). The web diagrams of Waxman (1988), followed by the observations of preferential attachment by Faloutsos et al. (1999), culminated in the observation that the Internet follows a scale-free topology by Barabási et al.

(2000); a result which has spawned a whole new arena of information network research.

The presence of scale-free structures in real world networks inspired researchers to investigate their properties further. Albert et al. (2000) found that scale-free networks (such as the Internet) are vulnerable to attack, the hubs forming a target from which to induce disconnection. Scale-free network generation algorithms have also been developed, however, these are said to exhibit significant "first mover advantage' '(Borgs et al., 2007) - the predisposition of earlier nodes to have higher degrees. A further network structure to be considered is that of the small world.

Small World

The small world graph model was formulated by Watts & Strogatz (1998). A small world network is a *regular* graph (whereby each node has the same number of links), with some randomly generated links included (Figure 3.8); these links are said to shorten the average path length, while the clustering coefficient remains unaltered (Watts & Strogatz, 1998). The model, therefore, falls between that of a random graph and a regular graph.

To generate a small world network, a regular network is first created. Each individual edge within the regular network may then be *rewired* with probability p, removing an existing edge ij and forming a new edge ik at random. This process introduces the aforementioned path shortening random links. The emergent property is such that as $p \rightarrow 1$, a random graph is approached. A further property of a small world is that the distance between two distinct vertices (L) is said to follow the proportionality:

$$L \propto log(n) \tag{3.16}$$

An example of a small world network being created is depicted in Figure 3.8. Small worlds, much like scale-free networks, are said to be exhibited in natural structures; the computational aspects have been applied to map neural architecture (Bassett & Bullmore, 2006) in a manner that may be utilised to investigate Alzheimers (Stam et al., 2007) and Schizophrenia (Liu et al., 2008).

The small world model presented, along with the random and scale-free models, demonstrate the topological aspects of network science literature. However, the topology of a network only captures the structure in which nodes are connected. Network science is also

concerned with the effect of network structure upon the nodes within it, the literary origins of which are presented in the following section (3.2.2); the particular focus being upon human behaviour, and the effect of human interaction.

3.2.2 Connection Effects

Lewis (2011) argues that the epidemiological work of Kermack & McKendrick (1927), is the first instance of literature relating to connection effects. The Kermack & McKendrick (1927) SIR models, while incorporating no elements of graph theory or connection, considered a contact rate in their differential equation model - a variable prescribed to control for human interaction (Kermack & McKendrick, 1932). This contact rate then had a bearing on the number of individuals infected with a disease, representing the probability of a susceptible individual coming into contact with an infected individual. Further details of the SIR model may be found in Section 9.3.

Graph theory and connection effects were later fully incorporated into epidemiology through the work of Solomonoff & Rapoport (1951), who examined the probabilistic transmission of disease along a graph. The considerations made by said researchers accounted for a network topology within the calculation process, although the initial work was conducted upon a random graph model (Solomonoff, 1952). Section 3.2.1 demonstrated that many real world networks do not follow a purely random process, with the social psychologist Stanley Milgram arriving at this conclusion in 1967.

During his time at Yale, Milgram attempted to quantify the connectedness of human society. His now seminal research experiment, based upon *the lost letter technique* (Milgram et al., 1965), sought to transfer a letter between seemingly unconnected persons in different regions of the USA. This ground breaking study uncovered that the documents were able to arrive at their designated destination, requiring just 5.2 intermediaries on average (Milgram, 1967).

Colloquialised as *six degrees of separation*, the work of Milgram has inspired many other researchers to replicate the study with alternative conditions. Examples include an email equivalent of the lost letter technique (Dodds et al., 2003), a current online search to find a target individual (Schuhbauer, 2012) and the investigation of 240 million Microsoft user accounts (Leskovec & Horvitz, 2008) - the results of which demonstrate that amongst

instant messenger communications, all users could be reached with an average distance estimated at 6.6 links (Kleinberg & Easley, 2010). Furthermore, the small world network of Watts & Strogatz (1998) is said to be a mathematical representation of the work initiated by Milgram (1967).

While much acclaim is given to the research of Milgram (1967), some argue that it is fundamentally flawed. Many letters in the study were not returned, therefore the proposed intermediary figure of 5.2 is based upon a subsection of the research as a whole. A further caveat identifies that the target in Boston (Massachusetts) was an affluent individual, therefore such claims may not be generalised to the population as a whole (Kleinfeld, 2002).

In an attempt to rectify the aforementioned biases, White (1970) reconfigured the study separating out the probability of discarding the letter; results demonstrated seven intermediaries were necessary. Markov models have also been used to quantify this process (Hunter & Shotland, 1974). However questionable the implementation of procedures within Milgram's research may be, its examination of the connections between individuals (and their effects) may be considered as early network science research.

It is not solely the scientific community that has been inspired by Milgram. The Broadway sensation "Six Degrees of Separation", from acclaimed writer Guare (1992), conceptualises the research for dramatic narrative purposes. Kevin Bacon also owes his synonymy with network science to Milgram; following the causal game by a group of college students to identify Bacon as the "centre of the Hollywood universe", an actor's *Bacon Number* now defines the number of interconnected actors necessary to link with Kevin Bacon (Singh, 2002). Amongst the research community, an *Erdös number* may be calculated, a similar concept to that of a Bacon Number, but searching academic citations for co-authorships that link with Paul Erdös (Odda, 1979).

Further contributions to the connection effects aspect of network science, come from economists and experts in marketing theory. Bass diffusion is a theoretical differential equation model of new product adoption in a population. Developed by Bass (1969), the likelihood of a purchase at a given time $P(T)$ conditioned on no purchase having been made up to that point is:

$$P(T) = p + qF(T) \tag{3.17}$$

$f(T)$ being the probability of a purchase at T and:

$$F(T) = \int_0^T f(t)dt \tag{3.18}$$

where p is referred to as the *coefficient of innovation* and q the *coefficient of imitation*. Adopters of a product may then be classified as innovators (those who adopt a product first) or imitators (those who adopt a product when a suitable number of other individuals have adopted), with the values of p and q controlling the overall diffusion of the product through the population. This revolutionary yet simplistic model has become the cornerstone of marketing theory, the original article achieving 4904 citations to date (Google Scholar, 2013).

The original models of Bass (1969) have no explicit concept of graph theory, however they do consider one's own opinion in the purchase of a new product, p, relative to the adoption rates of others, $qF(T)$ - the adoption of others impacting overall product adoption. This theory of adoption spawned new models, such as the representation of technological replacement (Fisher & Pry, 1972), eventually formulating the notion of "Diffusion of Innovation" - the now primary text on information diffusion (Rogers, 2003). The theory of diffusion underpins ASSIST, the data secured for the analysis of adolescent social networks in this thesis; this is discussed further in Section 5.1.

The literature presented in this section, demonstrates early research into the investigation of connection effects - highlighting their role in the formation of network science. With the benefit of hindsight, it is evident that many of the concepts - such as the epidemiological and diffusion models - would have benefited from implementation within a network structure. The following section (3.3) presents research that considers a network structure, and its resultant effects, with the specific focus being on social networks.

3.3 Social Networks

This section focuses upon literature related to social networks, classified into two distinct sectors: effect (Section 3.3.1) and construction (Section 3.3.2). Social networks are perceived to have a significant effect upon the individuals within it, such research is discussed in Section 3.3.1, with specific outcomes related to smoking behaviours being of interest.

Following the discussion of the effects of social networks, Section 3.3.2 presents research into the understanding of how social networks form, identifying specific factors said to be important in the friendship selection process.

3.3.1 Effect

Investigation into social networks (and their effects) has experienced substantial growth in recent years - 'The New York Times' highlighting social networks as *the* new idea of 2003 (Gertner, 2003). This section presents literature regarding both the positive and negative effects of social networks, with specific focus upon the behaviours of the individuals within it. Following a general review of social network effects, adolescent social smoking behaviours are discussed - as these behaviours are of key importance throughout this thesis.

Positive Effects

"No one simply goes to a party anymore. They go to network." This anecdotal quote from Kadushin (2012), while offering a misanthropic view of society, highlights the exploitation of social networks for personal gain. This view is by no means a characteristic of modern society; conferences (Matsuo et al., 2003), fundraisers (Brooks, 2005) and parent teacher association meetings (Lareau, 1987), having a long standing tradition of enterprising social contacts for personal benefit - emphasising the potential positive aspects of social networks. While social networks have potentially positive outcomes, it would appear that social isolation is rising (Hortulanus et al., 2005), with Putnam (2001) proposing the erosion of civil engagement in society. The effect of such despondence in a community is said to cause a decline in morality, an increase in crime and impact significantly upon health (Mohnen et al., 2013; Putnam, 2001; Tampubolon et al., 2013).

Marsden (1987) states that individuals have a core discussion group, a circle of individuals with whom they may converse with sensitive and personal matters. This group of individuals are reported to be instrumental in providing support, black widows being found to live longer that white widows following bereavement - a product of racial differences in support networks (Elwert & Christakis, 2006). Unfortunately, in a study of randomly chosen American adults, 12% stated that they had nobody with whom they may share personal matters or engage socially in free time (Christakis & Fowler, 2010b).

It is not only one's own core discussion group that may be influential, familial ties are also of key importance - for example positive father relations are said to reduce sexual promiscuity in female adolescents (Regnerus, 2006). Furthermore, persons with whom an individual may encounter briefly have also been regarded as significant, Weight Watchers and Alcoholics Anonymous meetings being specifically devised to support positive group outcomes (Wing & Jeffery, 1999).

An individual in a social network is said to be influenced not only by their direct connections, but also their indirect connections. Fowler & Christakis (2008) suggest that there are three degrees of influence, the notion that a single individual may affect persons up to three degrees away. Examples of such diffusion include: the spouses of Weight Watchers attendees experiencing significant weight loss (in absence of official meeting attendance) (Gorin et al., 2008); and kidney donor chains - whereby single altruistic kidney donations have sparked a chain of up to nine (previously unsuitable) donations (Rees et al., 2009).

Negative Effects

The impact of a social network is not always reported to be positive. One such instance is the substantial increase in syphilis rates amongst upper-class teenagers in Rockdale (Georgia) (Christakis & Fowler, 2010b). It would appear that the transmission of this disease, a rarity amongst wealthy communities, was being heightened by the acceptance of social norms regarding sexual acts with multiple partners - the disease becoming a product of influence spread within the network (Rothenberg et al., 1998).

Sexually transmitted diseases have become synonymous with network investigation, an outcome of their inherent methods of contraction. Multiple examples of research upon sexual contact networks in relation to HIV/Aids may be observed in the literature (Helleringer & Kohler, 2007; Liljeros et al., 2001; Potterat et al., 2002), including the employment of modelling methods to analyse contact network patterns (Anderson et al., 1991) and the targeting of subgroups to effectively design chlamydia screening programmes (Evenden et al., 2005a).

Communicable diseases (such as those described above) explicitly benefit from social network exploration, yet other more obtuse afflictions are also said to be associated with the architecture of one's own personal contacts. The rising prevalence of obesity in England

(1980: 7%, 2005: 24%) (Christakis & Fowler, 2007), back pain (Raspe et al., 2008) and binge drinking (Ormerod & Wiltshire, 2009), are all said to be exacerbated by the composition of an individual's social network.

Research suggests that adolescents appear particularly predisposed to peer influence (Brown et al., 1986; Kandel & Lazear, 1992; Sumter et al., 2009). As such, the connections made amongst groups of adolescents may be extremely salient with regard to behavioural norms - a particularly poignant example is that of *contagious suicide*. The findings of Bearman & Moody (2004) demonstrate that adolescents are twice as likely to attempt suicide if a friend had committed suicide in the previous year - the research finding a correlation between an individual's risk of suicide and low transitivity (as defined in Definition 3.1.5) in their social network.

The contagion effect of suicidal behaviour is widely reported in the literature (Brent et al., 1989; Gould et al., 1990; Wilkie et al., 1998); the research leading to the important reconfiguration in the way media outlets report an incident of suicide (Center for Disease Control, 1994), attempting to reduce the perceived contagion effect. Further demonstration of the contagious elements of behaviour are provided by mass psychogenic influence (MPI) research (Boss, 1997). Examples of MPI include the six month uncontrollable laughter outbreaks of Rankin & Philip (1963) - affected individuals suffering fear and exhaustion after the inability to cease laughter for sixteen days - and the phantom gas leaks of Jones et al. (2000) - hundreds of students hospitalised after the reported inhalation of fictitious toxic chemicals. While MPI's and contagious suicide provide extreme and reactionary examples of social network influence, network effects may be far more subtle and protracted - such as those related to adolescent smoking.

Adolescent Smoking Behaviours

This thesis is interested in the evolution of adolescent social networks, and in particular, their impact upon smoking uptake. As such, a review of literature relating to adolescent smoking behaviours is required. Adolescent smoking initiation is said to be a process involving family structure, personality and friendship selection (Arnett, 2007). An adolescent's social network is also said to be key in the the decision to smoke, with Christakis & Fowler (2010b) reporting the imitation of substance use amongst both direct and indirect friendship ties. The complex frameworks regarding the uptake of adolescent smoking,

have led researchers to explore the differentiation between cause and effect - attempting to assert whether a smoker 'infects' those around them with the need to smoke, or rather smokers organically group together based on common interest.

Christakis & Fowler (2010a) and Pearson & Michell (2000) suggest that smokers often inhabit the peripheries of an adolescent social network, marginalised into creating local pockets of smoking acceptance. This is contrary to the work of Lakon & Valente (2012), their findings arguing that smokers may often be central influential figures. Homophilly is also suggested as a crucial element in an adolescent smoker's social network (Mercken et al., 2013, 2012b; Schaefer et al., 2012), with smokers naturally gravitating toward one another based on their common interest.

The contextual elements of smoker behaviour must also be considered. Gender differences amongst adolescent smokers is well documented (Clayton, 1991; van Roosmalen & McDaniel), females demonstrating more susceptibility than males with regard to smoker influence (Mercken et al., 2010). Cultural factors have also provided an insight, the study of six European countries highlighting Dutch and Finnish adolescents as the most receptive to influential smoking (Mercken et al., 2009). Furthermore, the size of a smoker population within a network may also have profound effects - larger smoker cohorts exacting the strongest influence (Go et al., 2012).

The literature presented highlights the complex combination of elements said to be present in adolescent smoking, further compounded by the conflicting reports with regard to smoker-friendship selection. As such, there would appear to be value in the further investigation of adolescent smoker processes, providing insight though currently unutilised methods in the context of social networks - such as the simulation techniques described in Chapter 2. With the perceived effects of social networks presented, the following section (3.3.2) investigates literature related to their construction.

3.3.2 Construction

The literature of Section 3.3.1 demonstrated the reported widespread influence of a social network, but how a social network forms is also of interest; this section details general information regarding the formation of social network structures, with a view to predict the creation of new links (Link Prediction) - discussed further in Section 3.4 and Chapter 6.

Prior to having the capacity to predict future connections in a network, an understanding of friendship selection processes is required. Dunbar (1995) suggests that the neocortex size of primates is directly related to the number of stable social relations they can maintain, with an average upper limit of 150 contacts for homo-sapiens being proposed (Kudo & Dunbar, 2001; Sawaguchi & Kudo, 1990). With the increased connectivity of modern society, researchers have hypothesised that *Dunbar's number* may no longer be applicable; however, data from online social networking has only strengthened this theory (Dunbar, 2012; Gonçalves et al., 2011; Pollet et al., 2011b).

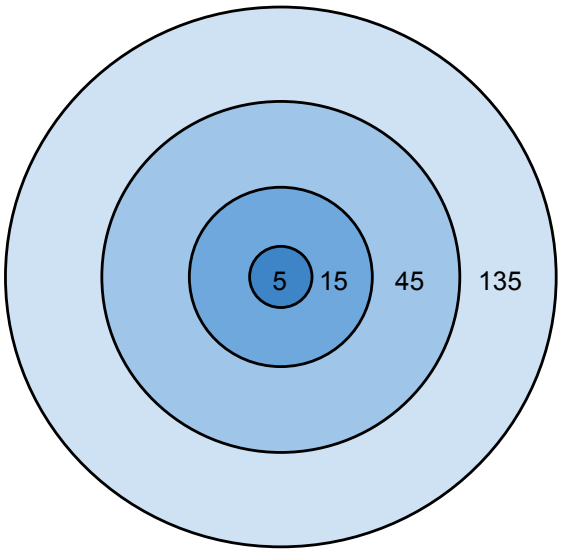


Figure 3.10: Dunbar's layers of friendship - example grouping of friendships decreasing in importance and increasing in group size Dunbar (1998).

Individuals are said to have layers of friendships, the grouping sizes of which increase by a factor of three (Figure 3.10) (Dunbar, 1998; Pollet et al., 2011a); at the centre of this hierarchical structure are those persons with whom an individual is the closest. Feld (1991) also explored the number of friends an individual may possess, leading to the paradox that, in general, "your friends have more friends than you". The logic behind this is detailed as follows: if an individual has 40 connections, they will increase the calculated "friends of friends" value of 40 individuals; yet an individual with a single connection, serves to reduce the "friends of friends" number of only one individual - hence, a heavier weighting within the calculation being given to highly connected individuals.

The friendship paradox has also been verified to be in existence online (Hodas et al., 2013;

Ugander et al., 2011), epitomising the now common practice of utilising communication data to investigate social network theory. A social network structure is said to be different to other types of network, due to the specific characteristics that personal relations embody - high degree association, transitivity and clustering (Newman & Park, 2003). While some argue that the data collected online is representative of real world connectivity (Lewis et al., 2008; Wilson et al., 2009), others demonstrate that is not always the case (Kwak et al., 2010).

Figure 3.11: Visualisation of Facebook connections across the world, taken from the Facebook profile of Mark Zuckerburg (CEO of Facebook) (Zuckerberg, 2013).

The proliferation of online data in the research community offers the ability to draw conclusions from large information repositories; Facebook being said to have 1.1 billion active users (Figure 3.11) (The Associated Press, 2013). The increase in data availability has led to the predictive modelling of human behaviour, one study finding 'curly fries', 'thunderstorms' and 'science' Facebook *likes* are the largest predictors of high intelligence (Kosinski et al., 2013).

Communication platforms are becoming regularly utilised for scientific insight: chat rooms being used to help trace syphilis outbreaks (Klausner et al., 2000); the online game World of Warcraft demonstrating human reactions to a deadly pandemic (Lofgren & Fefferman, 2007); and file sharing websites used to predict musical tastes (Lambiotte & Ausloos, 2005). This predictive trend has also branched into the domain of social contacts, attempt-

ing to quantify the processes by which individuals form friendship connections.

In the study of 3.3 million customer call logs from a Belgian phone operator, Lambiotte et al. (2008) found proximity to be a key factor in communication ties - the probability that two nodes are connected being inversely proportional to the square of the distances between them. Newton's law of gravitation specifies that two bodies attract with a force that is inversely proportional to the square of the distance between them, from Lambiotte et al. (2008) this would also appear true of individual communications. Attribute similarity has also been investigated as a predictor of friendship, with similarity in interests, university attended and location said to overlap significantly with social connections (Gong et al., 2012; Yang et al., 2011) - suggesting similarity as a good predictor of social connection.

Overall, the literature reviewed in Section 3.3 has provided a greater understanding of social connection. Section 3.3.1 described the potential effects of social network membership, while section 3.3.2 identified theories of friendship selection. With a basic understanding of how links may form, researchers have attempted to predict the formation of new links in a network; this has developed into specific area of literature known as Link Prediction problems, an outline of which is provided in Section 3.4.

3.4 Link Prediction

A variety of applications require the ability to predict new links in a network; examples include: optimisation of website navigation (Zhu et al., 2004), the recommendation of content to web users (recommender systems) (Huang et al., 2005), and the acceleration of academic collaboration (Farrell et al., 2005). Prediction of links between humans has further reaching potential implications, with investigators demonstrating the ability to map a portion of the September 11^{th} terrorist network through the use of public records (Krebs, 2002). Such information has previously been used for prosecution purposes, but researchers are moving toward the inference investigation of *dark networks* for the prevention of crime (Bakker et al., 2012; Raab & Milward, 2003) - anthropologists have even used honeybees to model the behaviour of criminal gang territories (Brantingham et al., 2012).

Link prediction is the process of attempting to foresee connections that may currently be unobserved, due to *covertness* (deliberately hidden due to criminal activity), missing

data, or links that are yet to be established (Liben-Nowell & Kleinberg, 2007). Methods employed in conjunction with the link prediction problem include machine learning (Goldenberg et al., 2003; Hasan et al., 2006), Markov methods (Domingos & Richardson, 2007; Taskar et al., 2003) and statistical inference (Popescul & Ungar, 2003). It is widely accepted that the task of accurately predicting links is difficult (Getoor, 2003; Taskar et al., 2003), in part due to the a priori probability of a link being small (Getoor & Diehl, 2005).

Many studies have contributed methods to the process of link prediction, Liben-Nowell & Kleinberg (2007) and Lü & Zhou (2011) offering reviews of the currently developed algorithms, but few have attempted the use of simulation methods in a social network context (Barabâsi et al., 2002). Researchers have analysed events within a network (O'Madadhain et al., 2005) and their impact on connectivity (Albert & Barabási, 2000; Mislove et al., 2008), but a combination of all elements encompassed within a simulation framework appears to be non-existent. The work of Rattigan & Jensen (2005) suggests that actually it is the anomalous links - links that are the most statistically unlikely - that prove to be the most interesting.

Link Prediction methods underpin the new algorithm developed in this thesis to predict social network evolution, a discussion of which is presented in Chapter 6 and Chapter 7. The specific networks of interest in this thesis are those of adolescent social connections, upon which current LP algorithms have not been tested; a review of current LP algorithms and a discussion of their applications is conducted in Chapter 6. Xu & Chen (2008) state that analysis for active preventative measures for network effects is "still missing", the research of this thesis attempting to utilise LP methods to predict adolescent smoking uptake - this investigation being presented in Chapter 9. Prior to investigating social networks, and their subsequent effects, it is of importance to understand how to best represent network data; this discussion occurring in Section 3.5.

3.5 Network Representation

As a network grows in size, keeping an accurate account of *incident edges* (as defined in Section 3.1.1) becomes a more complex task. To combat said issue, researchers have developed methods from which to interpret a graph - both mathematically and visually. The following section introduces the mathematical notation used to represent a network,

and the algorithms used to create network visualisation - detailing the standard form to be utilised in the remainder of this investigation.

The manner in which a graph is depicted may have a significant impact upon the visualisation of information. Networks are omnipresent entities, a graph image may therefore have the task of communicating required information in a concise and legible manner. Take for example the London Underground, a complex array of stations (nodes) and routes (edges). The Underground map has evolved over the years to accommodate the ever expanding service, Roberts (2013) suggesting that an orbital structure may now be a more informative representation (Figure 3.12).

To mathematically represent a network, first recall the definition of a graph $G = (V, E)$ (Section 3.1). A graph may be represented by an adjacency matrix X, such that the rows and columns of X represent the nodes in V, with the status of a link from $i \rightarrow j$ being defined as the element $x_{i,j}$. The size of X is dictated by n, the number of nodes in a network, and $x_{i,j} \in [0, 1]$ for an unweighted social network. If $x_{i,j} = 1$, there is an edge that connects the nodes v_i and v_j , otherwise $x_{i,j} = 0$. For a weighted network, $x_{i,j}$ describes the *strength* of the tie from $i \rightarrow j$ (Wasserman & Faust). In the context of social relations, it is common to refer to the adjacency matrix X as a *sociomatrix* (Wasserman & Faust), with the vertices within G referred to as *actors*.

For example, the directed network of Figure 3.14 is depicted by the sociomatrix:

$$X = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

The element $x_{1,2} = 1$ indicates a directed link from $v_1 \rightarrow v_2$. If X were to represent an undirected network, the sociomatrix would be symmetric. A sociomatrix provides a simple mathematical representation of a social network, allowing for computation of the metrics detailed in Section 3.1 in a more intuitive manner (Wasserman & Faust).

Aside from mathematical matrix notation, there are also multiple ways of visually representing a graph. Three methods of graph visualisation are to be discussed: the circle

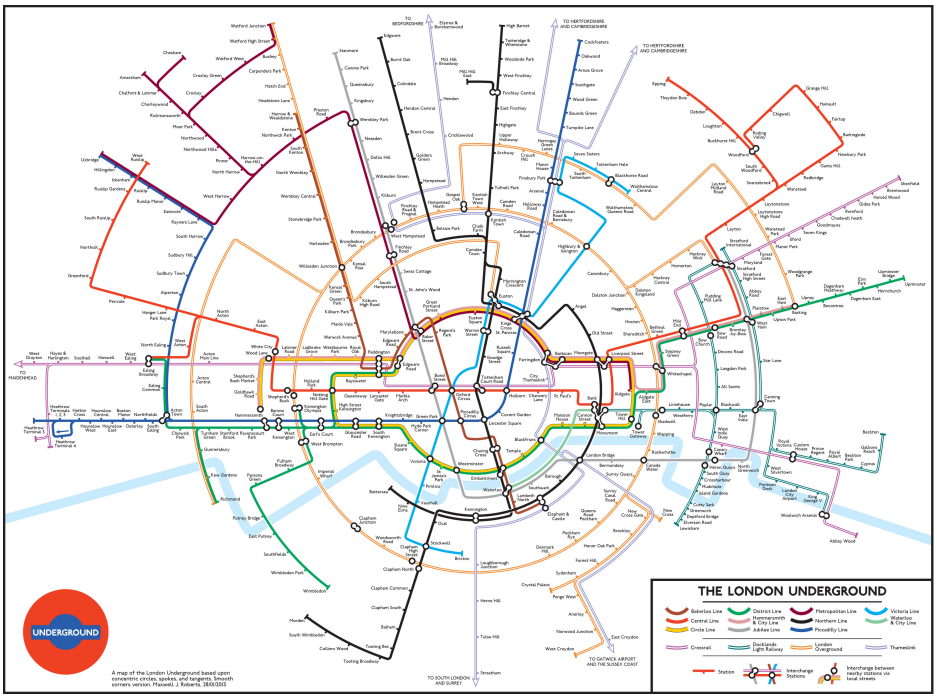


Figure 3.12: Proposed circular map of the London Underground, said to offer a more intuitive representation of network structure (Metro Reporters, 2013). Image credited to Roberts (2013).

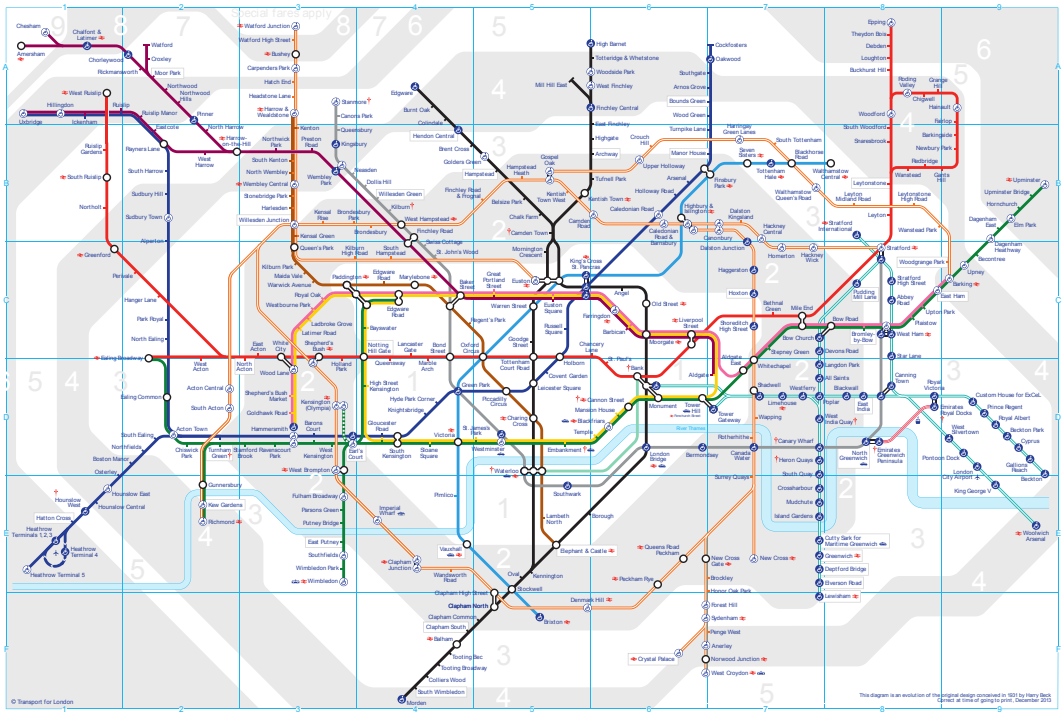


Figure 3.13: Existing London Underground map (Transport For London, 2014).

arrangement, force directed algorithms and tree structures. Each method is presented with a brief description of its origins, followed by an example image of the specified visual representation. The example images all represent the same network (the network of Figure 3.14), demonstrating the effect of different layouts upon the same graph.

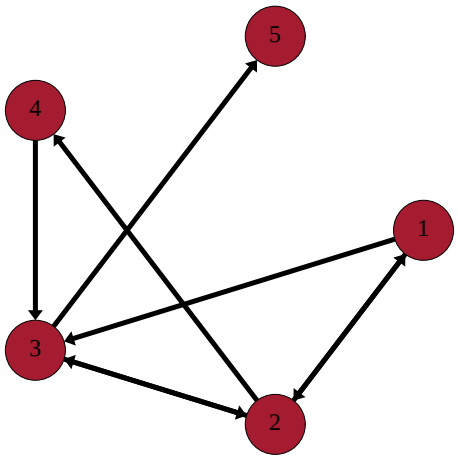


Figure 3.14: Example directed network with five nodes arranged in a circular layout.

The most simplistic is a *circle* arrangement, as depicted in Figure 3.14. Nodes are arranged in a circular manner and edges traverse the interior of the circle to arrive at their destination. A circular arrangement is often termed as *neutral*, as it does not emphasise any specific node (Iragne et al., 2005). Nodes are equally spaced, do not give an inflated view of network centrality (Huang et al., 2007) and are well positioned for the representation of star, ring and circular elements of metabolic networks (Becker & Rojas, 2001).

In certain circumstances it is of interest to emphasise specific nodes, such as the visual representation of central individuals in a social network. In a circle arrangement the number of edge crossings may be high, leading to the inability to infer important individuals in a graph - the problem of reducing the number of crossings being *NP-Complete* (Baur & Brandes, 2005). To present a more aesthetically pleasing representation of a social network, often *force directed* graphs are employed. These methods typically mimic a physical system, attaching forces to the vertices of a graph. Nodes are attracted to one another in they share a connection, with disconnected nodes repelling one another (Bannister et al., 2013); the resulting graph is centred around nodes which span a variety of connections.

The concept of *spring-like* (formally termed as spring embedder) force algorithms is credited to Eades (1984), who developed an algorithm to visualise a graph which attaches

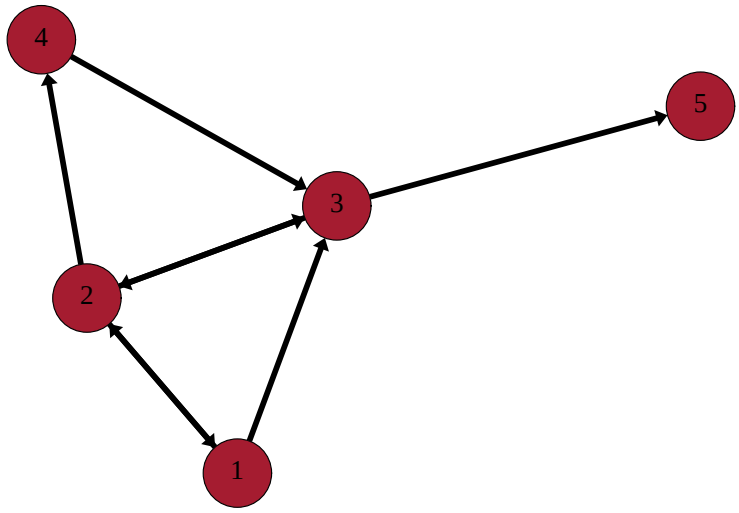


Figure 3.15: Example directed network with five nodes arranged using the Fruchterman-Reingold algorithm.

attractive forces to connected nodes - these spring-like forces being based upon Hooke's law. This was followed by the work of Kamada & Kawai (1989), who solved partial differential equations to improve layout. Kamada & Kawai (1989) built upon the work of Eades (1984), through the notion of creating distance between *all* nodes on the graph - Eades (1984) solely focusing on the distances between graph neighbours.

The work of Kamada & Kawai (1989) was later developed further by Fruchterman & Reingold (1991), offering improvements in terms of efficiency by redefining the forces relevant to node placement. Through the inclusion of "graph temperature", the Fruchterman & Reingold (1991) algorithm is often the preferred method of force directed graph layout (Figure 3.15) - receiving standard implementation in software packages such as R (R Development Core Team, 2008) and Gephi (Bastian et al., 2009).

Subsequent force directed algorithm literature included other metrics for node arrangement (Bannister et al., 2013; Brandes, 2001; Gajer et al., 2000), attempting to represent the differing forces present in a social network. However, due to its simplicity and efficiency, the Fruchterman-Reingold algorithm will be adopted as the standard visualisation method for graph images in this work. Alternative classes of layout algorithm are also available, such as the tree layout systems of Reingold & Tilford (1981) (Figure 3.16); however, these are more appropriate for structures where a clearly defined hierarchy may be observed. For additional information regarding graph visualisation, the work of Tollis et al. (1998) is

a recommended text.

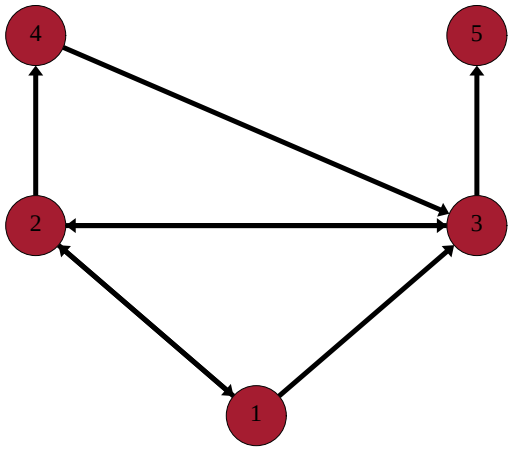


Figure 3.16: Example directed network with five nodes arranged using the tree Reingold-Tilford algorithm.

As a further note with regard to visualisation, often the attributes of an actor may be important to consider. In such circumstances, the use of heat maps may be prudent. For an unweighed social network, a heat map of the friendship connections present would essentially be a replica of the sociomatrix through the medium of colour; a heat map of the sociomatrix of Figure 3.14 is presented in Figure 3.17.

If an attribute - smoking for example - were to be considered, the similarity in levels of the attribute between individuals could be imputed into the heat map. Consider a given network of 5 actors and assume that a particular attribute s (such as smoking) can take the values $1 \leq s \leq 4$; if the actors have attributes $s = (1, 1, 2, 3, 4)$, the resultant similarity heat map is presented in Figure 3.18.

The mathematical and visual social network representations presented, offer the reader the ability to engage with high volumes of data in a more intuitive manner. The methods discussed serve as a reference point for the investigation of real social network data in Chapter 5, aiming to facilitate familiarisation with complex structures as node count increases. As such, the visualisations styles presented will regularly be revisited throughout the thesis.

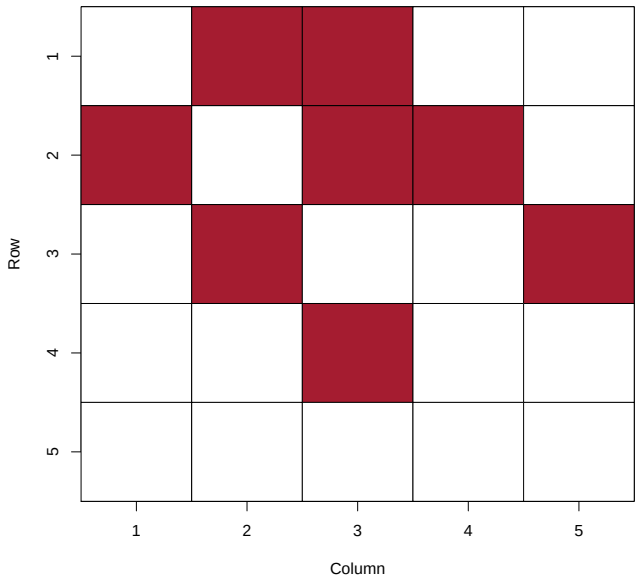


Figure 3.17: Sociomatrix heat map, coloured blocks representing a tie between vertices.

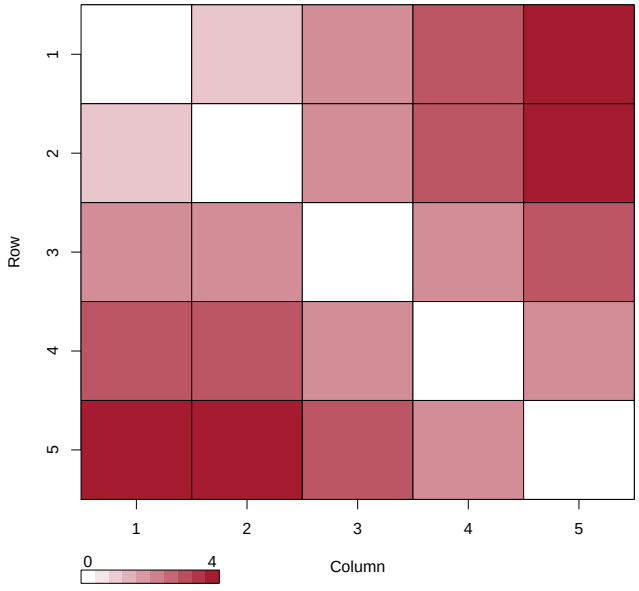


Figure 3.18: Attribute heatmap representing the adjusted differences in smoker level. The darker the block, the stronger the difference in smoking level - white blocks indicate self attribute similarity. Block differences calculated as $|a_i - a_j| + 1$ if $i \neq j$, 0 otherwise.

3.6 Overview

This chapter has introduced the key concepts of network theory that will be regularly appropriated over the course of this work. Section 3.1 highlighted the key graph theoretical concepts of social networks, proposing the metrics to be employed in the analysis of network data. These measures are as follows:

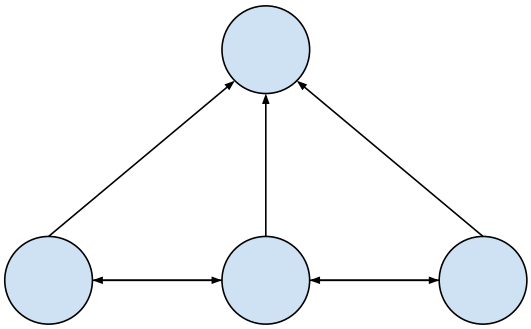
- Average Node Degree;
- Reciprocation;
- Density;
- Transitivity Ratio;
- Clique Number;
- Average Path Length;
- Degree Centrality;
- Closeness Centrality;
- Betweenness Centrality.

Section 3.2 detailed the history of network science, beginning with its graph theoretical origins. Two distinct lines of research were presented - topology and connection effects. These sections serve to emphasise the key advances in the literature, formalising many of the theoretical underpinnings regarding the formation of networks and the personal implications of connectivity.

Section 3.3 targeted social network literature, detailing the current breadth of investigatory work. This section focused on the health aspects of social network membership, questioning the process by which formation of such structures appear in society. The literature reviewed suggested influential connections may have both positive and negative effects to an individual's well-being, highlighting adolescent social networks (and their role in smoking uptake) as being of particular interest in this thesis.

Section 3.4 introduced Link Prediction, the concept of attempting to foresee future connections in a network. Link Prediction techniques are of great importance in this thesis, underpinning the investigation of adolescent social network evolution conducted in Chapter 7. Further discussion surrounding the specific algorithms that may be used for LP, and the development of a new LP algorithm (PageRank-Max), is conduced in Chapter 6.

Finally, Section 3.5 demonstrated a number of tools for social network visualisation, setting the standard for the remainder of the thesis. Many techniques have been formulated in the literature, addressing the issue of optimal link placement. Following a review of potential methods, this work has opted to make use of heat maps and the Fruchterman-Reingold force algorithm for visual interpretation. This chapter concludes, forging a path of investigatory research directed by the literature reviewed in Chapters 2 and 3.



4

-"The Influential Hierarchy"

The Peter Principle

This chapter acts as a preliminary investigation of the proposed research direction, building upon the literature set out in previous chapters. Chapter 2 has identified Agent Based Simulation (ABS) as a suitable candidate for the creation of individual-centric models, whereby an agent's actions dictate overall system outcome. This notion is synonymous with influence associated with specific individuals in a social network, a topic discussed at length in Chapter 3.

The work presented in this chapter, serves to assess the suitability of ABS as a technique to investigate social networks, and provide an exploration into the impact of theoretical network structures upon a conceptual agent population. The identification of differing outcomes between the network structures investigated, initiates the in-depth analysis of real social structures conducted in Chapter 5, which in-turn informs the development of the PageRank-Max algorithm presented in Chapter 6.

The specific social domain of this sector of research, focuses upon hierarchical organisa-

tions. Beginning with the work of Granovetter (1973), who demonstrates the ability of connections at the periphery of an individual's network to be rewarding in terms of professional mobility, social networks have increasingly become investigated within the context of the workplace. Company performance is evidently of crucial importance to any firm, with staff efficiency therefore being a principal topic in business and managerial literature.

This segment of research builds upon the foundations of the presented literature, examining social networks and corporate efficiency through ABS. The specific focus of this chapter is the 'Peter Principle' (PP), a theory of hierarchical managerial inefficiency - said to be introduced through the promotion of managers beyond their level of competency. The subsequent sections outline the problem as follows: the motivation for the work (Section 4.1); a review of previous PP ABS literature (Section 4.2); suggested expansions to previous PP ABS literature (Section 4.3); the construction of a new PP ABS (Section 4.4); the results produced (Section 4.5); and, finally, the conclusions (Section 4.6).

4.1 Motivation

The topic of managerial incompetence is often referred to anecdotally, however, Peter & Hull (1969) investigated many real world examples - attempting to understand the reasons for its occurrence. Peter & Hull (1969) theorised that the cause of managerial incompetence is not necessarily due to an individual being completely incompetent, but rather a consequence of promotional frameworks; this phenomenon was then conceptualised as the 'Peter Principle' (PP).

The PP states that all members of a hierarchical organisation are promoted to their maximum level of incompetence; once this has been achieved, career progression is halted and the employee is left stagnant in a role they can no longer effectively fulfil. The position previously inhabited by an individual may have had differing requirements to those of a higher-level managerial role; as such, a promotee's skills may not directly translate to success in their new post (Peter & Hull, 1969). Incompetence, therefore, has been shown to manifest within an organisation, causing ultimately detrimental effects to productivity (Kane, 1970) and subsequently impacts upon revenue.

The effects of the PP are said to not only occur amongst hierarchical organisations, with the theory also being used to explain the often counter-intuitive elements of everyday life.

Examples of such eventualities include the drop in quality experienced on the second visit to a restaurant and the disappointment expressed at movie sequels (Lazear, 2004); the PP is even said to be in existence in the context of NBA player performance (Dilger, 2003). Personal accounts of how to negate this seemingly unavoidable organisational incompetence have been provided by Peter (1972), but further research has also suggested that hiring external personnel (Acosta, 2010) and establishing promotional schemes (Fairburn & Malcomson, 2001) may be beneficial.

Recent research has attempted to investigate the PP though ABS methods, seeking to ascertain the effect of distinctive promotion rules on the efficiency of an organisation (Pluchino et al., 2010, 2011). The work of Pluchino et al. (2010, 2011) investigates two realms of an organisation, one in which a Common Sense (CS) mechanism applies and one where the PP is allowed to wreak havoc - CS assuming that should candidates be competent at a given level, their competence will remain relatively unaltered on promotion to a higher level.

In the context of the simulations created in Pluchino et al. (2010, 2011), the models show that: under CS circumstances, when an agent becomes promoted they retain their competency level with a chance of minor fluctuation; when the PP is assumed, the competence of an agent is randomly redistributed post-promotion. The two assumptive paradigms then assess the impact of promoting the most competent (best), the least competent (worst) or a random candidate (random) - the results of Pluchino et al. (2010, 2011) suggesting that promotion of the best candidate in a PP structure proves most destructive to overall efficiency.

The conclusions of Pluchino et al. (2010, 2011) further state that, in the absence of a clear understanding of promotional governance, decisions should be made at random or through an alternating best-worst strategic decision. Furthermore, promotion of the worst candidate proves to be more beneficial under PP circumstances, than promotion of the best candidate under CS conditions. The implications of such findings, should the constructed simulation be accurate, are evidently grand - potentially redefining the manner in which promotions are awarded.

The PP is often regarded with some scepticism however, Beeman (1981) arguing that it is just a “catchy title" coupled with “simplistic logic" that “cannot stand the light of even

the most elementary analysis or logical inquiry". Furthermore, the work of Pluchino et al. (2010, 2011) violates many of the inherent thought processes that control promotion on merit (Furnham & Petrides, 2006). However, the models proposed by Pluchino et al. (2010, 2011) pose interesting questions regarding the impact to a system as a result of individualistic behaviour - an idea in line with that of this thesis.

The primary motivation for selecting the PP as the domain of interest for this sector of research, is its prior investigation with an ABS framework. The work of Pluchino et al. (2010, 2011) has a number of limitations (discussed further in Section 4.2); in particular, the models created do not consider the behaviour of others in the heirarchy, or the effect of having social contacts. This presents an opportunity to experiment with theories of individual behaviour in an organisation, along with varying social structures, to assess the impact to system outcomes - using the original work of Pluchino et al. (2010, 2011) as a comparative baseline. The additional elements incorporated into the model created herein, have have not previously been explored through computer simulation methods in the context of the PP. Prior to discussing the created model, a review of the work of Pluchino et al. (2010, 2011) is required (Section 4.2).

4.2 Model Review

Prior to presenting the model published in Fetta et al. (2012), which builds upon the work of Pluchino et al. (2010, 2011), a detailed review of Pluchino et al. (2010, 2011) will be presented. The structure of Pluchino et al. (2010) utilises an arbitrary hierarchical organisation as a conduit for further investigation, some adjustments later being made in Pluchino et al. (2011) to investigate the effect of hierarchical tree structures upon the impact of the PP. As the original conclusions regarding random promotion were founded on the model of Pluchino et al. (2010), for comparative purposes, this basic structure will be retained as a foundation to the subsequent investigation.

4.2.1 Basic Model

The ABS model of Pluchino et al. (2010) creates a pyramidal organisation comprised of 160 agents, distributed across six tiers. The lowest tier consists of 81 agents, followed by 41, 21, 11, 5 and finally 1 agent (the boss) as the hierarchy is climbed to reach its highest

tier (see Figure 4.1). The agents are given two variables on entry into the simulation, age and competence, chosen according to a normal distribution. Age is distributed with mean 27 and standard deviation 5, while competence is distributed with mean 7 and standard deviation 2. Competence is an indicator of job performance and ranges from 0 to 10; should this drop below 4, the agent is deemed incompetent and fired.

The age of an agent is incremented each time step. If an agent eludes falling below the given competency threshold, they remain in the organisation until their retirement. An agent retires at the age of sixty, vacating their position in the organisation - allowing this to be filled by an agent residing on the tier below. Should vacancies occur on the lowest tier, a new recruit enters the system adhering to the same normally distributed principles of behaviour as previous agents. Each tier has an associated responsibility level, jobs at higher tiers demand more responsibility.

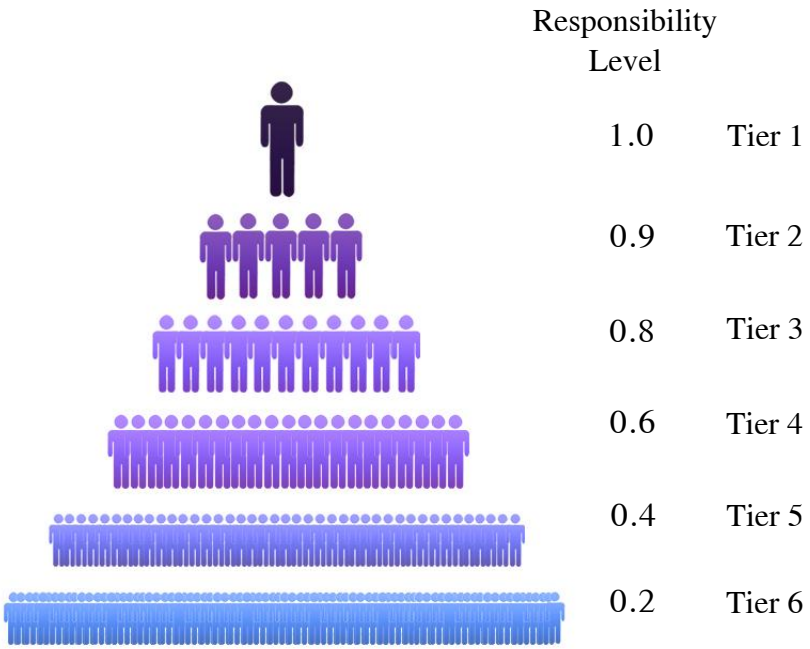


Figure 4.1: Six tier model hierarchical organisation comprised of 160 agents. Reported are the responsibility levels alongside the relevant tier.

Three methods of promotion allow either the most competent (“best"), least competent (“worst") or a random agent (“random") ascension to the next level. Furthermore, candidates may be promoted under CS principles - retaining their competence from the previous

level with a small change δ (randomly selected where $\delta \in [-1, 1]$) - or the PP, where competence becomes randomly redistributed post-promotion. System efficiency (E) may then be calculated; following each round of promotions, to assess the consequential effects on the corporation.

The manner in which E is calculated utilises the tier-based job responsibility scale (r_i), ranging from 0.2 to 1. This is then multiplied by the sum of the competencies from each tier (C_i), and divided by the maximum possible system efficiency (occurring when each agent has a competence of 10) - giving the following equation:

$$E = \frac{\sum_{i=1}^{6} C_i r_i}{10 \cdot \sum_{i=1}^{6} n_i r_i}, \tag{4.1}$$

where n_i is the number of agents on each tier. E is then evaluated over the course of the simulation (Pluchino et al., 2010), assessing the effect of promotional strategies.

4.2.2 Verification

To augment the model of Pluchino et al. (2010) with social network and behavioural effects, a recreation of the original Pluchino et al. (2010) model is necessary. To maintain consistency, no additional elements are to be incorporated at this stage. Should the constructed Verification Model provide analogous results to those of Pluchino et al. (2010), assessment of baseline procedures may be considered complete; this allows for network and behavioural factors to be included and validated accordingly in Section 4.4.

Creation of the Verification Model employs the use of Netlogo, the software package also utilised by the original authors, adhering to the specifications set out in Section 4.2.1. The overall corporation efficiency is calculated after each timestep, the simulation running for a period of 1000 timesteps. On completion of the proposed 50 model runs, the results are plotted and compared with that of Pluchino et al. (2010) - the graphs displayed in Figure 4.2.

The graphs of Figure 4.2 are organised such that steady state values of the relevant principle and promotion method are clearly displayed. The Verification Model exhibits the same counter intuitive behaviour previously discovered in reference to the PP; promotion of the worst agent being the most efficient. Differences do occur between models however; ver-

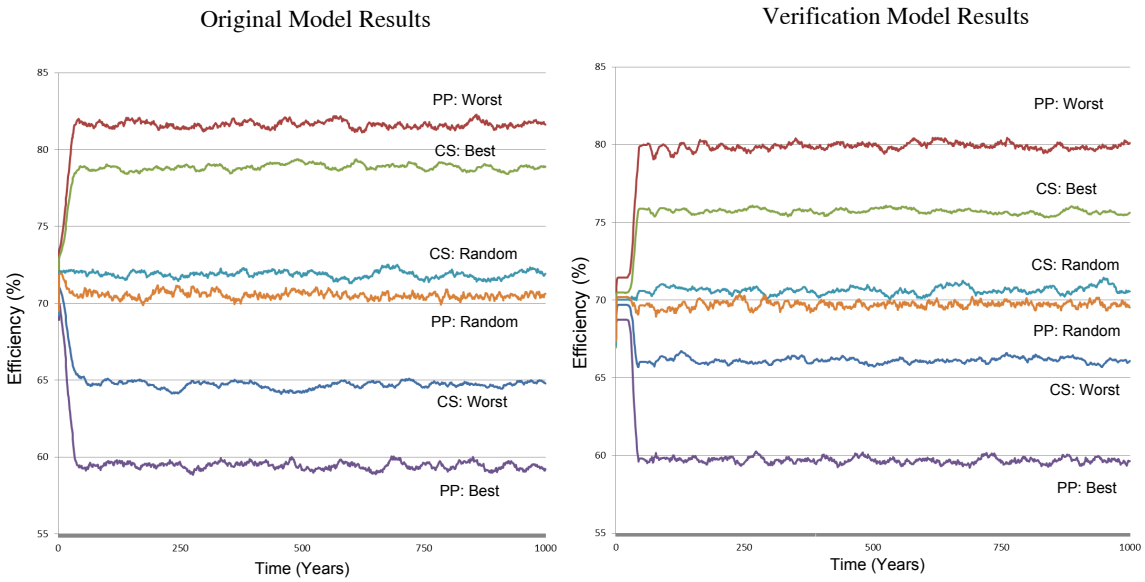


Figure 4.2: Comparison graphs of Pluchino et al. (2010) (left) and Verification Model (right), displaying average steady state efficiency values across varying promotional practices.

ification 'PP-worst' and 'CS-best' experiencing a marginal reduction to efficiency results in comparison with the original Pluchino et al. (2010) model. Methodological differences in model construction may be the cause of these inconsistencies, highlighting issues of comparability.

On further investigation, replication of the results from Pluchino et al. (2010) may be achieved on alteration of the method in which random sampling occurs. Competence (c) of an agent is said to be Normally distributed with mean 7 and standard deviation 2, evidently elements $c < 0$ and $c > 10$ become irrelevant due to $c \in [0, 10]$; agents with $c < 4$ eventually being fired. The created Verification Model discards values outside of the relevant region through resampling, thus avoiding the incorrect allocation of agent competence; Figure 4.3A demonstrating the resultant sampling distribution.

The original model of Pluchino et al. (2010), does not appear to follow the sampling method of the verification model. Following extensive testing, it would appear that Pluchino et al. (2010) create sampling boundaries at $c = 4$ and $c = 10$. The subsequent distribution

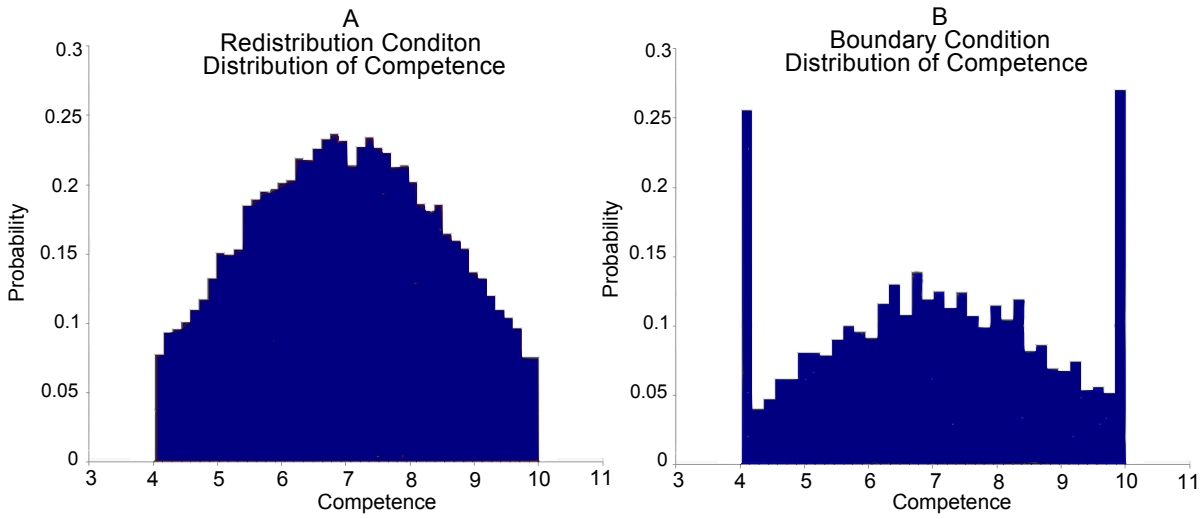


Figure 4.3: Comparison of competence distribution graphs for redistribution and boundary conditions, respectively.

($\bar{c}$) is constructed such that:

$$\bar{c} = \begin{cases} 4, & \text{if } c \leq 4 \\ c, & \text{if } 4 < c \leq 10 \\ 10, & \text{if } c > 10 \end{cases} \tag{4.2}$$

This impacts the shape of the distribution, producing a greater number of highly competent and reasonably incompetent agents (Figure 4.3B). Evidently it would appear the two methods present contrastive behaviour, suggesting a possible reason for the disparity.

Implementation of the boundary sampling method, while effective in terms of comparability, does not appear wholly accurate. In the aforementioned method, sampling distribution may not be described as Normal - violating the conditions dictated by Pluchino et al. (2010). In light of such conclusions, on creation of new ABS PP models, the sampling method employed within the Verification Model shall be retained - highlighting the first extension to the work of Pluchino et al. (2010) in an effort to give a more informed view of organisational dynamics.

Overall, it may be concluded that - following the basic principles outlined in Section 4.2.1 - a replica of the model detailed in Pluchino et al. (2010) may be created. While some

minor adjustments are accepted, the model conclusions remain synonymous - demonstrating verification of the arguments proposed by Pluchino et al. (2010). The Verification Model therefore provides a basis for further model development, laying the foundation for comparative analysis.

4.2.3 Limitations

The results of Pluchino et al. (2010) show that promotion of the best candidate under PP conditions has a detrimental impact to efficiency, concluding that random promotion should be used to limit the spread of incompetence. This evidently provides an interesting account of the PP effects and aids the discussion sparked by Peter & Hull (1969). It must be remembered, however, the aforementioned conclusions are drawn from a hierarchy of passive agents unaffected by their surroundings.

When Peter & Hull (1969) first devised their theory, many real world examples were given to justify the existence of this counter intuitive phenomenon. Discussion also touched upon the idea of organisational *pull*, the notion that "an employee's relationship by blood, marriage or acquaintance with a person above him in the hierarchy" may offer promotional gains; resulting in the *pullee* becoming unpopular (Peter & Hull, 1969). However, elements of such behaviour and social network structure are somewhat overlooked in Pluchino et al. (2010, 2011).

A further question relates to the reaction of individuals should a random promotion system be implemented, employees gaining little incentive to work hard and succeed. If promotion effectively becomes a lottery, members of the organisation may just perform the minimum tasks to avoid dismissal; such behaviour ultimately impacts upon efficiency. Pluchino et al. (2011) suggests offering prizes in reward of the most competent, but it is questionable how effective such an incentive would be in sedating employee outrage.

To give a more encompassing view of events, and to harness the power of ABS effectively, the Verification Model will be extended accordingly. While simulation models may not accurately portray the subtle nuances of human behaviour, the inclusion of social theoretical elements may provide observations regarding natural human negators of the PP. Additionally, should the created simulation models prove useful, the process of combining simulation methods and social theory will be justified as a viable investigative approach -

a key objective in the context of this thesis.

4.3 Model Augmentation

The following section aims to expand the original model of Pluchino et al. (2010), including elements of the shortcomings highlighted in Section 4.2.3. Three distinct new branches of literature are incorporated, in conjunction with the normal distribution sampling adjustment detailed in Section 4.2.2. On inclusion of these supplementary parameters in the simulation model, PP effects will be assessed in contrast with the original work of Pluchino et al. (2010); conclusions may then be drawn regarding promotion mechanisms, the validity of the statements made and the role of ABS in the investigation of social theory.

4.3.1 Workplace Social Interactions

The workplace is an environment which may potentially foster social relations. Research into the dynamic of workplace interaction has been vast (James et al., 2008); topics of discussion range from the intricacies of e-mail grouping (Skovholt & Svennevig, 2006) to the complexities surrounding office romance (Riach & Wilson, 2007). Although a selection of employees may make a conscious decision to exclude themselves from workplace relationships, those who do engage find it to be a beneficial exercise (Berman et al., 2002).

The mechanics behind workplace relationship formation is said to draw upon elements of proximity, status (Schutte & Light, 1978) and social capital (Rhee, 2007) (described further in Section 4.3.3), resulting in a diverse and vibrant network of connections. As this is a preliminary investigation, organisational network data has not been accrued - opting rather to formulate a theoretical hierarchy based on sectors of the literature detailed in Chapter 3.

Small world, scale-free and random hierarchical network topologies are to be explored, embedding the simulated agents within a variety of social network constructs - a preliminary step in creating socially aware agents. As a result, theoretical structures of workplace association may be assessed in combination with system efficiency, analysing the sensitivity of results against underlying topology.

The method in which network formation is exacted within the simulation begins upon initialisation, a basic network following the designated topology being created. As the

simulation progresses, on an agents ascension to a higher tier, the network is adjusted to reflect new connections that may form as a result of promotion. As a direct consequence of shifts in hierarchical position, existing links may be severed - the topology retaining the designated characteristics of its specified construct. Further details regarding the exact deployment of structural mechanisms within the simulation are documented in Section 4.4.

4.3.2 Office Politics

Social networks provide one facet of the workplace environment, but said relations are also defined by those individuals from whom the connections extend. The model created by Pluchino et al. (2010) assumes employees are passive individuals, remaining unaffected by colleague promotion. In reality, the application and selection process for any position is far from clinical; the literature of Morgeson & Ryan (2009) and Sieverding (2009) documenting this further. As such, the agent's own personal reactions must also be taken into consideration following colleague promotion - gauging the subsequent impact upon productivity.

Humans are predisposed to respond to changes in circumstance, the workplace being no exception. The formation of connections within an organisation fosters the investment of individuals in the success of others - the connotations of which are potentially both positive and negative. The responses exhibited by those linked to a promotee are particularly meaningful, as connected parties may be aware of the competence level of an individual; this awareness may be pertinent should a candidate be perceived as undeserving of promotion - a colleague's emotions becoming particularly heightened should they also have been considered for the role in question.

Research has shown that if an employee's application for promotion is rebuffed in place of a co-worker, it incites a tendency to work harder and succeed (Schaubroeck & Lam, 2004); the cognitive explanation to this improvement associated with envy. This primitive emotion is described as a coping mechanism when a person's self image is threatened (Salovey & Rothman, 1991), often said to occur when another's success is a threat to our own self-evaluation (Lockwood & Kunda, 1997).

Envious tendencies are said to be active under rejection conditions, especially if the subsequent successful individual is seen as a role model or to have high similarity to oneself

(Lockwood & Kunda, 1999). In the context of the workplace, Schaubroeck & Lam (2004) found that envy positively influences "post-rejection job performance" - coinciding with previous research actively discussed amongst the references therein. It may therefore be extrapolated that envy has a positive effect on competence, suggesting an impact to organisation dynamics.

While positively envious behaviour may be applicable for candidates who have been rejected in place of the "best" applicant, such conclusions may not be appropriate should the "worst" individual succeed. Failure in career progression coupled with the success of an incompetent colleague, offers a different perspective on envy and its resulting behaviour. The perceived fairness of proceedings may be called into question, a component said to be integral to work ethic, having the potential to foster counter-productive employee behaviour (Ford et al., 2009).

These counter-productive behaviours may relate to self perception, job satisfaction and stress; however, inter-personally this may have more serious ramifications (Greenberg & Colquitt, 2005). Cohen-Charash & Mueller (2007) found that unfairness coupled with episodic envy produces negative outward emotions. These emotions are not only harmful to the work ethic of the employee but, in some cases, has led to harm being inflicted on the envied other. A full scope of the literature may be found in Cohen-Charash & Mueller (2007), but implications of the research relate closely to conditions being constructed for simulation.

It would seem that during a "fair" promotion method (and when similarities can be identified between linked employees), the spirit of healthy competition is active, suggesting an increase in competence. On the other hand, when promotions are seen as "unfair", responses between linked employees become far more austere; work attitudes drop, impacting negatively upon system efficiency. Such reactive inclinations may have further significant effects to the friendship links themselves, Cohen-Charash & Mueller (2007) also finding that promotee likeability dropped by 60.7% post-promotion. This gives weight to the argument of Berman et al. (2002) who found that managers tended to have very few vertical links - links to a working level below one's own - even when connections are present prior to elevation.

4.3.3 Social Capital

Social networks, aside from their ability to evoke emotional reactions impacting upon individual efficiency, offer a vehicle for the exploitation of relational ties for informational gain. The promotion of a candidate may not always be due to a supreme demonstration of competence in their previous role. Individuals may navigate social ties in a bid to better career prospects, with such acute relational awareness referred to as social capital (Burt, 1997).

Defined as “the aggregate of the actual or potential resources which are linked to possession of a durable network of more or less institutionalized relationships of mutual acquaintance or recognition" by Bourdieu (1986), social capital has been analysed in relation to individual aspects of organisational thought for many years - the underlying influential processes of social capital often being used to define behavioural suggestion and the diffusion of innovations in the context of social networks.

A specific example is that of Brass (1985), investigating the career progression of men and women. Results of Brass (1985) confirmed workplace stereotypical notions suggesting men had more access to social capital, in turn creating more promotional momentum than women. Similar findings were also concluded when ethnicity was a considered variable; Caucasian employees promoted more often than those of Black ethnic origin due to a heightened currency of social capital (Parks-Yancy, 2006).

Investigation into the overall effects of social networks upon career mobility, as opposed to the individual variables detailed above, advanced toward the theory of social capital aided by Burt (1992). In an attempt to explain this fruitful precedent, Burt analysed “structural holes” within networks. It was identified that two unconnected nodes - mediated by one connected node - presented an opportunity to extract potentially lucrative informational rewards. It also allowed for the connected entity to broker connections between unconnected entities, leading to the inference of social capital amongst organisational networks.

This has since been refined in later works by Burt (1997, 2000) and through the inclusion of informational time sensitivity by Rhee (2007), but the initial incarnation is still applicable in the context of a hierarchical network. The structural holes theory presented by Burt infers that being a highly connected individual in an organisation will present more

opportunity for promotion, regardless of actual job competence. This definition is conditional however, with "highly connected" being a concept relative to the connectivity of other individuals in the network.

Overall the three additional elements presented - social networks, reactive emotions and social capital - present a more inclusive representation of agents within an organisational hierarchy than the original works of Pluchino et al. (2010, 2011). The interplay between agents in relation to the PP, along with the interpersonal reactions and social capital, may now be assessed in an effort to ascertain the theoretical effects of varying promotion methods. While the behavioural extensions described do not fully represent the breadth of human emotion, their inclusion may - at the very least - offer some insight into the dynamic of workplace promotion and system efficiency as a whole.

4.4 Network Behavioural Model Development

Expanding the Verification Model of Section 4.2.2, the following work describes a new model developed to incorporate aspects of the previously described literature; this model shall be referred to as the 'Network Behavioural Model' (NBM). The simulation foundations remain the same as those of Pluchino et al. (2010) (with the exception of random sampling methods), measuring the efficiency of three promotional rules (best, worst and random) upon two realms of hierarchy (PP and CS), converting agent emotions and social capital into simplified dynamic rules.

On initialisation of the simulation, a network is selected from those discussed in Section 4.3.1: scale-free (SF), small world (SW) and random (RAN). The initial structures act on a discrete tier-by-tier basis, creating six topological layers of the hierarchy functioning in an autonomous manner. The consequence of this, dependant upon the conditions selected, is to have either six small world or scale-free networks categorised by managerial level; the SF agent at the highest level being authorised to select a link from the tier directly below. However, this segregation is only active amongst the initial defined structures; elements of variability, such as the random probability associated a SW and the RAN network structure itself, allow for cross tier links to occur. Examples of network initialisation may be viewed in Figure 4.4.

The overarching topological structure is strictly defined upon initial construction, but the

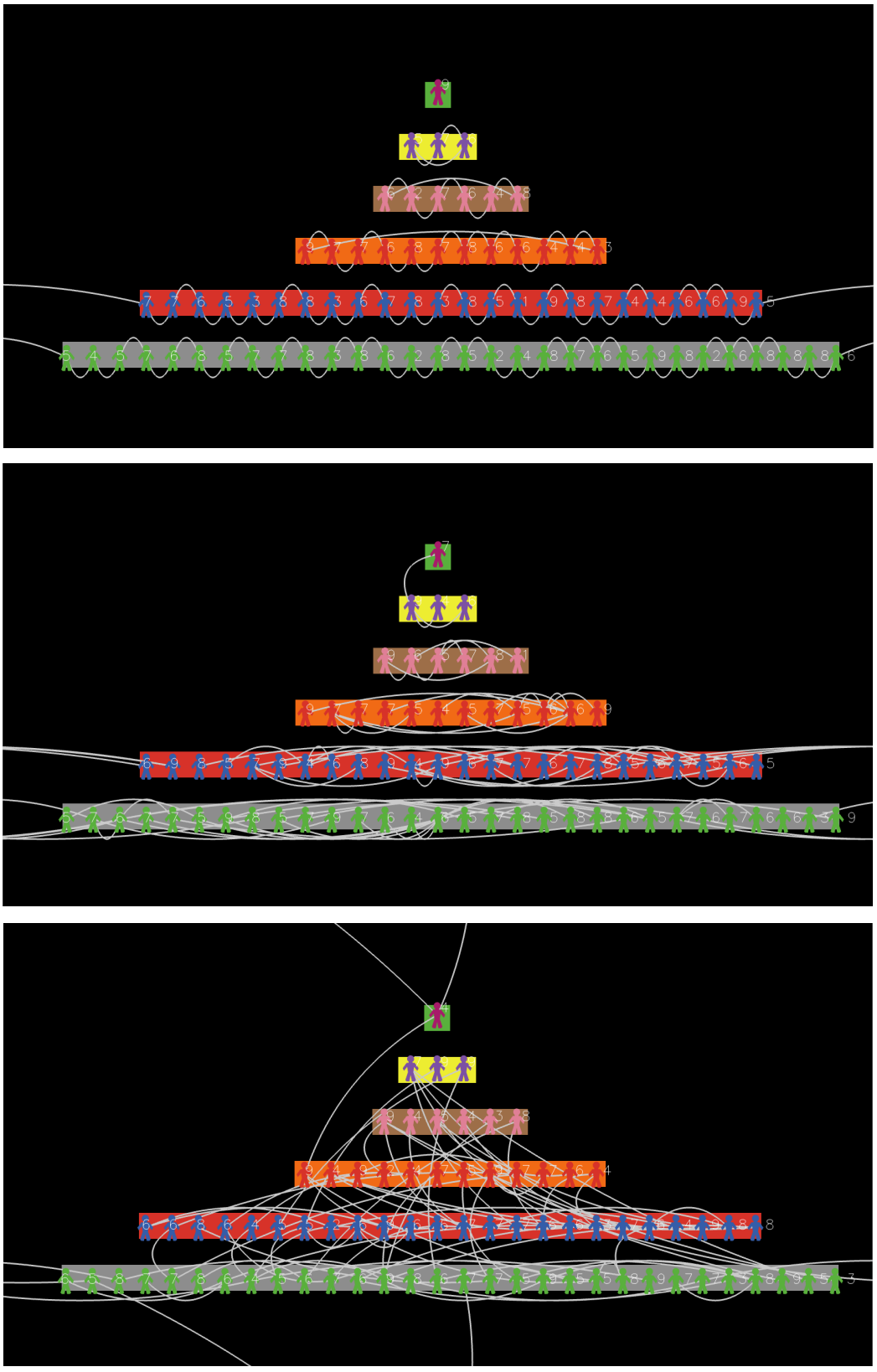


Figure 4.4: Simulation screenshots of social network structure at initialisation (from top - bottom: small world, scale-free and random). The figures adjacent to the agent indicate their associated competency level.

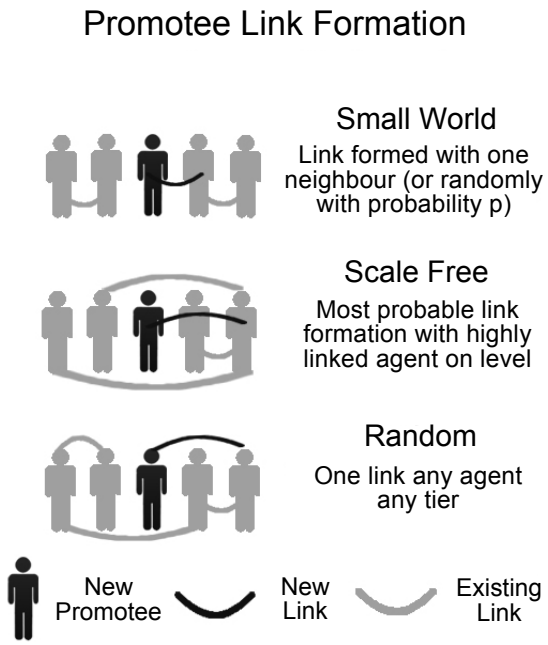


Figure 4.5: Diagram of promotee link formation. Detailed is the dynamic of each link formation on candidate promotion under the respective network topology.

framework of link formation evolves to become more complex as time progresses. Promotions begin connecting the isolated network layers, as each agent is required to form an additional link to their new level colleagues. The composition of this link is dictated by the overarching network dynamic, described graphically in Figure 4.5.

On promotion, an agent forms a new link, retaining their connections from previous hierarchical levels. Should the overarching topological structure be RAN, any agent from any tier may be selected to reciprocate the new link. If the SW structure is engaged, an agent makes a connection with the agent physically next to them (on the same tier) with probability $(1 - p)$, or with a randomly selected agent with probability p; p is selected by the investigator. For the purpose of the ensuing simulations $p = 0$, as path-shortening cross-tier links naturally occur following promotion. Finally, if the SF topology were to be selected, the promotee favours an equivalent-level agent (node) i as a connection with probability:

$$P(v_i) = \frac{\deg(v_i)}{\sum_j \deg(v_j)}, \tag{4.3}$$

following the preferential style of attachment introduced in Section 3.2.1.

Once the initial network set up procedure is complete, the behavioural elements of Section 4.3.2 may then be exacted to affect career advancement procedures. Social capital literature would suggest that highly linked agents have more promotional momentum, therefore the dynamic rules of the simulation select the most connected agent on a given tier as the first to be promoted. Following this, the remainder of the promotions at said level resume the structure of the designated precedent (best, worst or random).

The inclusion of the aforementioned assumption, effectively allows one highly linked agent to bypass all other dynamic rules on each level-specific round of promotions. The agent selected may not necessarily possess the highest/lowest competence (dependant upon the simulation conditions), achieving career progression based solely upon their exploitation of social capital. Evidently, this includes some variability in the competence of agents receiving promotion - possibly impacting upon the organisation's overall efficiency.

This leniency, to allow an agent promotion outside of the governing best/worst/random rules, is illustrated with the following example:

Tier two has six positions available for agent promotion from the tier below (tier one). If the governing promotion method is "best", then the six most competent agents will be selected to fill the available positions from tier one. However, before the promotions occur, the agent with the most social contacts from tier one will fill one of the available positions on tier two - leaving five positions available. The remaining positions on tier two will be filled by five of the six most competent agents selected. Similar rules apply for "worst" and "random" promotional rules, whereby the least competent and randomly selected agents will be chosen respectively.

The promotion of an agent based upon heightened social capital, may inadvertently alienate those agents with whom the promotee shares a link. Connections between colleagues would suggest some knowledge of an individual's job competence, subsequently one's own right to promotion may be evaluated based upon the perception of others. Should an undeserving agent advance, an individual may take umbrage - dropping their level of competence. By contrast, the colleagues of a highly competent agent may experience positive reactions to an agent's promotion.

Representation of reactive dynamics are facilitated by the simulation's constructed social network. On completion of all promotions across each tier, the promoted candidates are

given the ability to influence agents that share a social link with them. The lower level colleagues of a promoted agent will experience a representation of *envy*. If the promotee has a similar or greater level of competence to the envious agent, said agent will increase their competence (+1) in an attempt to attain the same accolades achieved by the envied agent (i.e. promotion). However, if the promotee has a level of competence below that of the envious agent, said agent will feel that the promotion criteria is *unfair* - lowering their competence (-1) in a display of counter productive work behaviour.

The influential changes that occur - termed as the γ effect - are initially dependant upon competence possessed by the promotee prior to promotion, an agent's competence (c) potentially varying post-ascension according to either CS (small change δ) or PP (random redistribution) principles. The promotee's connections will continue to experience the γ effect while the agents reside on disparate tiers, the effect in later time steps adjusting to the previously promoted agents new competence - ceasing when agents either become promoted or sever ties.

The final simulation adjustment incorporates likeability factors into the model. According to the findings of Cohen-Charash & Mueller (2007), 60% of inter-tier links become severed post-promotion. As such, this detachment process is exacted following the promotion of an agent, with the disconnections being selected uniformly at random - diminishing the protracted effects of reactive behaviour. This also results in agent popularity dwindling as their career develops, a phenomenon with which many managers may identify.

Following the completion of the additional simulation policies, and the calculation of organisational efficiency, the process repeats until the simulation stopping conditions are satisfied - a visual representation of the logic consolidated into the simulation available in Figure 4.6. The included elements are intended to provide a small representation of socially aware agents within a hierarchy, and the resultant effect upon efficiency; Table 4.1 summarises the investigated system elements.

4.4.1 Validation

For comparative purposes, the NBM retains many structural properties of the simulation produced in Pluchino et al. (2010) and the Verification Model referred to in Section 4.2.2. Social theory underpins the newly created model, due to an absence of real world data;

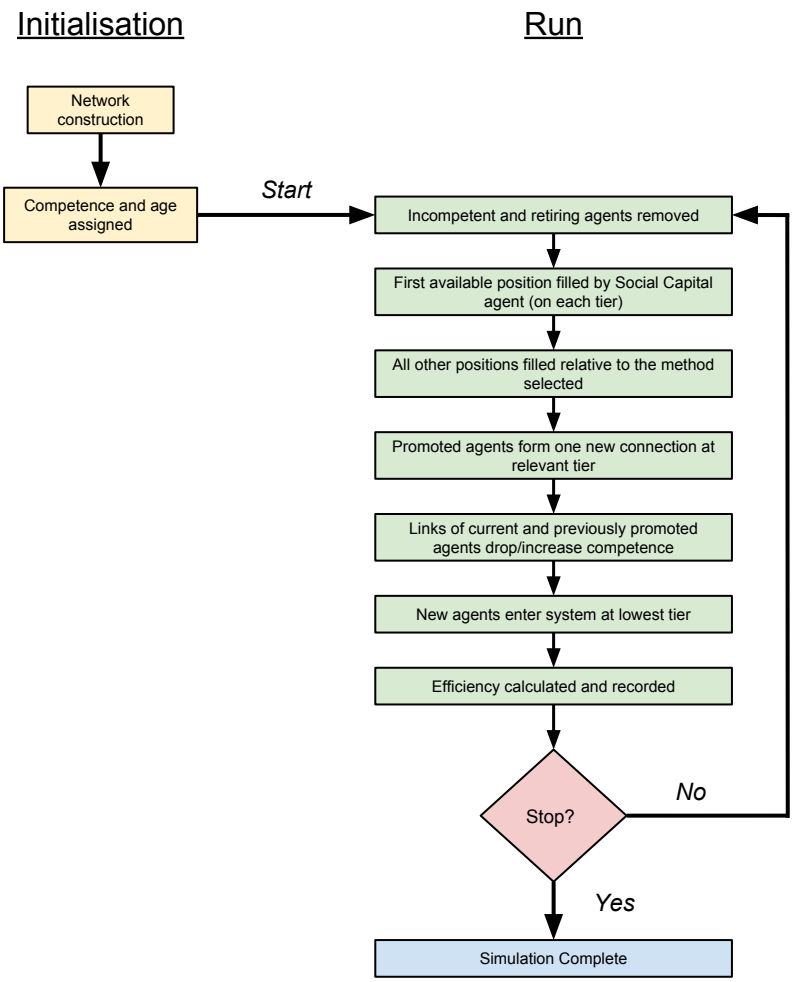


Figure 4.6: Simulation logic represented diagrammatically.

therefore, validation is based upon a detailed review of the literature. Given that the additional elements integrated into the NBM have been directly sourced from social and organisational research, it may be assumed that the notions presented for investigation are valid contextually.

Formation of the hierarchical social networks may be validated with ease, following the proposed construction algorithms detailed in Albert & Barabási (2002), Erdos & Renyi (1959) and Watts & Strogatz (1998). However, the translation of social elements (human reactions and social capital) into simulation logic are evidently open for interpretation by the creator. To streamline this process, the defining agent rules have been made as simplis-

Rule Type	Name	Description
Promotion	Best	Agent with highest competence is promoted
	Worst	Agent with lowest competence is promoted
	Random	Equal probability of any agent being promoted
Network	Small World (SW)	Distance based connections with random parameter
	Scale-Free (SF)	Highly connected “hubs" of agents
	Random (RAN)	Equal probability of one agent connecting to any other agent
Behaviour	Envy	Drives agent to work harder and succeed
	Unfairness	Disillusioned agent displays counter productive behaviour
	Social Capital	Agent exploits network to get ahead

Table 4.1: Summary of dynamic model rules.

tic as possible - both to avoid model over-complication (Pidd, 2004) and allow verification of dynamic components with ease. On assessment of their correct assimilation into the NBM, and recalling that this model acts as a *preliminary* investigation, the adjustments are proposed to be valid for the intended purpose.

As previously discussed, the Normal distribution is used within the simulation to sample the competence and age of agents, making use of internal processes within the Netlogo software; the results generated appearing to indicate the appropriate shape (Figure 4.3). However, the selection criteria for describing competence $\sim \mathcal{N}(7,4)$ and age $\sim \mathcal{N}(27,25)$ in such a manner are not documented by Pluchino et al. (2010), casting doubt upon the accuracy of said assumptions. In the absence of data regarding this issue, and for consistency with the original work, said distributions have been retained.

Further issues regarding validity of the NBM, refer to the manner in which the simulation is executed. A period of 1000 years is suggested as the model run time by Pluchino et al. (2010), a selection which would appear excessive. Time granularity is increased in later work Pluchino et al. (2011) with model “steps” representing one month in the organisation, the authors reducing run length to 1000 months accordingly. However, this simply rebrands the time frame from years to months - contributing little insight to further conclusions.

Upon analysis of the graphs in Figure 4.2, it would appear that the simulation does not arrive at steady state until around 50 years have elapsed. This preparatory period within the constructed organisation may therefore be considered as a simulation “warm-up”, a period usually disregarded from overall analysis. This issue is addressed in Pluchino et al. (2011),

removing said time frame from results collection. For consistency with the conclusions of Pluchino et al. (2010), the 1000 year collection period - warm up period incorporated - will be retained.

The original model of Pluchino et al. (2010) draws conclusions from 50 simulation replications, the reasoning behind said selection being undocumented. The work of Pidd (2004) and Robinson (1994) present a number of methods to ascertain the most appropriate number of repetitions, methods which do not appear to be utilised in Pluchino et al. (2010). For consistency with the simulation set up of Pluchino et al. (2010), 50 simulation runs have also been adopted for analysis of the NBM.

Many of the validation issues discussed may be improved upon, but one must take into consideration the aim of this work - recreation of the basic Pluchino et al. (2010) hierarchy augmented to include active social theory. Although the NBM may currently be constrained, on completion of successful assessment of social network and behavioural effects, further model extension and improvement may be sought. In light of the preceding discussion, the NBM appears verified for its predetermined context.

4.5 Results

The following section describes the results amassed from the NBM, collected across 1000 time steps with 50 replications. Table 4.2 categorises the data by structure, the PP and CS, subdivided by promotion method - best, worst and random. The steady-state values for each of the six categories are calculated, classified by network topology - SF, SW and RAN. Simulation output is also presented graphically in Figure 4.7, detailing the development of efficiency over the designated time period, each individual graph displaying the six construction categories - CS-worst, PP-worst, CS-best, PP-best, CS-random and PP-random - segmented by the three network topologies.

Large fluctuations in efficiency are observed during the initial years of the model, the aforementioned instability being attributed to the method in which competence is allocated. On simulation commencement, initialisation properties may assign a number of agents positions of power with a competence below the required threshold. The concurrent timestep eradicates these agents, granting more competent individuals the opportunity to occupy a vacant position. The high level of agent turnover experienced during this tumultuous

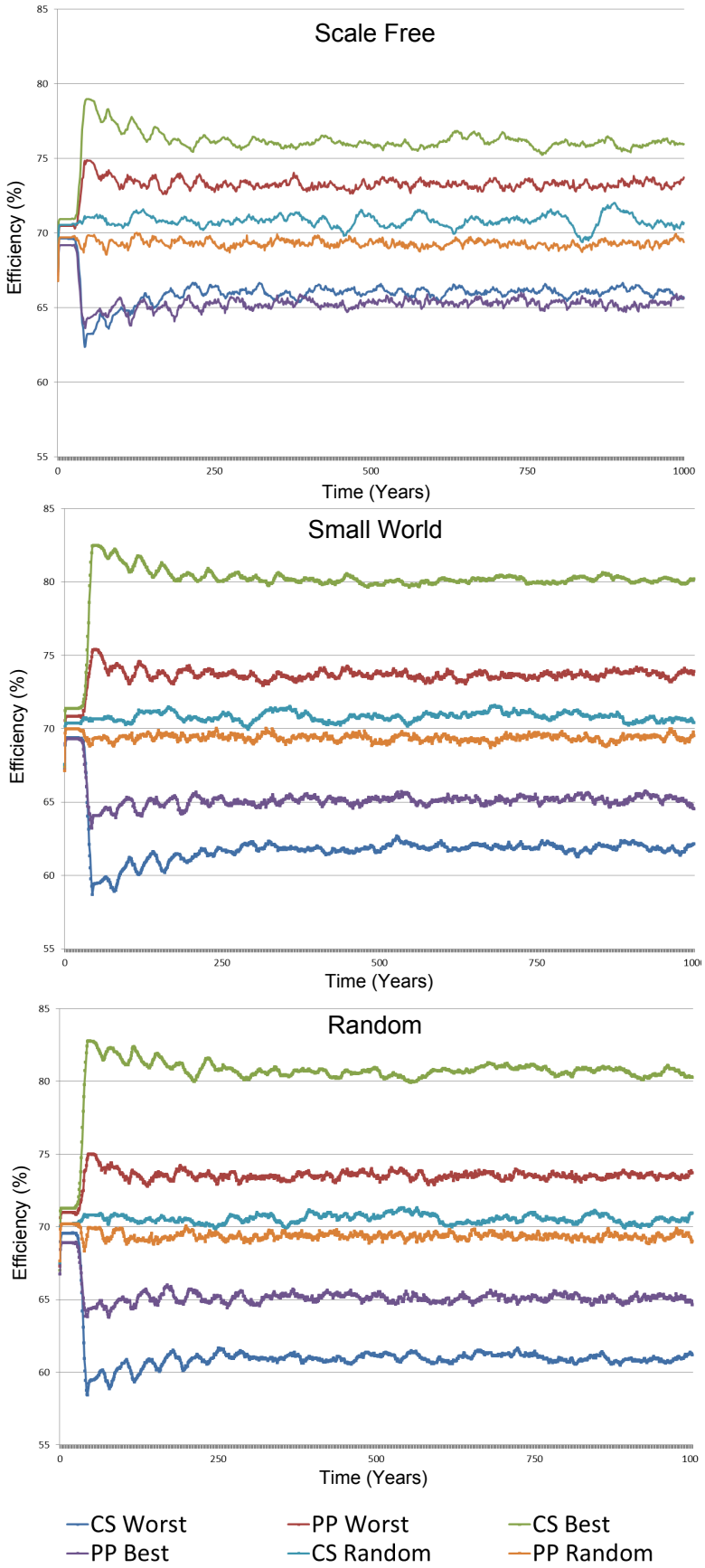


Figure 4.7: Comparison graphs for average steady state efficiency values across varying promotional practices and network topologies.

	PP-Best	PP-Worst	PP-Random	CS-Best	CS-Worst	CS-Random
Verification	60	80	70	76	66	71
SF	65	73	69	77	66	71
SW	65	74	69	81	61	71
RAN	65	74	69	80	62	71

Table 4.2: Average steady state efficiencies for each promotion method and network topology. As confidence intervals spanned a length less than one, integer values have been provided.

process eventually becomes relatively stable, as more proficient individuals are assimilated into the organisation. Once complete, the system achieves the steady state condition required for comparative analysis.

Results produced on promotion of the worst candidate indicate a difference in outcome dependent upon network structure. Analysing CS-worst, the efficiency of the system is greatly reduced for both RAN and SW, while the SF network displays minimal changes in comparison to the Verification Model (Figure 4.2). Similar reductions can be seen on analysis of PP-worst, however the effects are uniform across all three network topologies. These effects are reversed when the "best" method of promotion is assessed; both CS-best and PP-best graphs highlight substantial increases to efficiency, with the exception of CS-SF - whose results remain similar to the output of the Verification Model.

The efficiency differences between network structure indicate that behaviour is not acting autonomously, but in unison with the network construction selected - potentially attributed to the number of contacts each agent possesses. An SF network creates a small number of highly linked agents, while an SW network will contain a large number of moderately linked agents; therefore, as the first individual promoted is the most connected, an SF agent will have a greater impact upon other agents in the system than their SW counterpart.

In the context of the CS-worst results, the socially connected agent may (by chance) possess a high level of competence. When an SF network is imposed, this highly connected agent has the power to inspire positive envy in a large number of agents - increasing competence and overall system efficiency. This effect is reduced on activation of an SW network, as agents have a smaller selection of contacts to inspire - counterbalanced by the subsequent worst promotees negative influence on their social contacts. The reverse is true of CS-best, whereby incompetent socially connected SF agents may be promoted first; this

decreases the competence of their social contacts and overall system efficiency.

The SF social agent effect does not impact substantially upon PP conditions however, the reasoning behind this may be provided by the redistribution of competence that occurs post-promotion. Although a highly competent SF promotee will still affect a large number of candidates on the initial promotion step, the longer lasting effect is truncated by the possibility of competence at their new position becoming reduced. This, in turn, will reduce the number of linked agents affected by later promotional advances, minimising the overall effect to system efficiency.

Efficiency levels of the CS-best category, regardless of network topology, have increased such that promoting the best candidate (under CS conditions) is now the most efficient. This suggests an important departure from the results produced by the Verification Model, whereby PP-worst is deemed the most efficient. Random methods remain relatively unchanged however, neither network nor behaviour modifying efficiency substantially. Also remaining unchanged is the notion that, without a clear understanding of competence transmission (PP or CS), promoting at random is the most efficient compromise.

With regard to network topology development, it would appear that the defining features of each construction remain intact as the simulation progresses. To assess the development of topologies over the course of the simulation run, network statistics have been calculated over the steady state region. Table 4.3 displays the mean agent degree and network efficiency (NE) of the system under promotion of the best candidate. NE is calculated as the sum of the inverse of the shortest path length between connected nodes, over the number of connected nodes in the hierarchy (C):

$$\frac{\sum_{i,j} \frac{1}{d_{ij}}}{C}, \tag{4.4}$$

where $d_{i,j}$ is the shortest path between connected nodes i and j, and "connected" indicates a path between the nodes exists.

The mean degree statistics of the SF and SW agents demonstrate an increase over the observed RAN network values. The reasoning behind this may be attributed to the permitting of RAN networks to generate cross tier links; said vertical links then becoming severed as a result of the associated behavioural rules. This logic also provides an explanation for

	Common Sense			Peter Principle		
	SF	SW	RAN	SF	SW	RAN
Mean Degree	1.01	1.85	0.86	0.99	1.79	0.84
Network Efficiency	0.29	0.11	0.60	0.30	0.12	0.62

Table 4.3: Average steady state network statistics exclusive of warm up period. For brevity, results refer to the best method of promotion - results observed across all methods demonstrating similarities in behaviour. Full network tables may be found in Appendix A.1.

the higher mean degree SW networks experience, in comparison to SF - the highly connected SF agents becoming promoted, potentially severing a large proportion of links in the system.

Analysing the NE figures - higher values indicative of greater connectivity in the network - the RAN network appears to be the most connected. The random connections have served to shorten the path length between agents, concurrent with the literature; the regular disconnected tier-based links of the SW network substantially decreasing connectivity. The relatively low NE of the SF network may be contrary to expectation, one assuming that the highly linked hub agents would reduce APL. However, the severance of vertical links may once again be affecting dynamics - the hubs becoming promoted, severing connections and inducing a number of isolated nodes.

The ideologies implemented in the NBM exhibit efficiency shifts in comparison with the Verification Model, such observations being irrespective of network topology. Investigating said fluctuations further, supplementary results examining the scale of reaction to promoted agents are presented in Figure 4.8. During the initial instance of results collection, enactments of envy and unfairness are interpreted by an increase or decrease of one competence point respectively. To ascertain the sensitivity of values to competence adjustment (γ), a range of values are explored $\gamma \in [0, 5]$; 0 indicates no behavioural change, 5 being the maximum.

The graphs illustrated in Figure 4.8, CS-best and the PP-worst, present opposing dynamics. CS-best peaks in efficiency at $\gamma = 2$, dropping substantially as $\gamma \to 5$. The conditions at $\gamma = 2$ potentially inspire such positively envious individuals, a large number of agents achieve the maximum competence of 10, propelling highly competent agents into managerial roles. The positive impact becomes truncated once $\gamma \geq 3$ as, although the system is

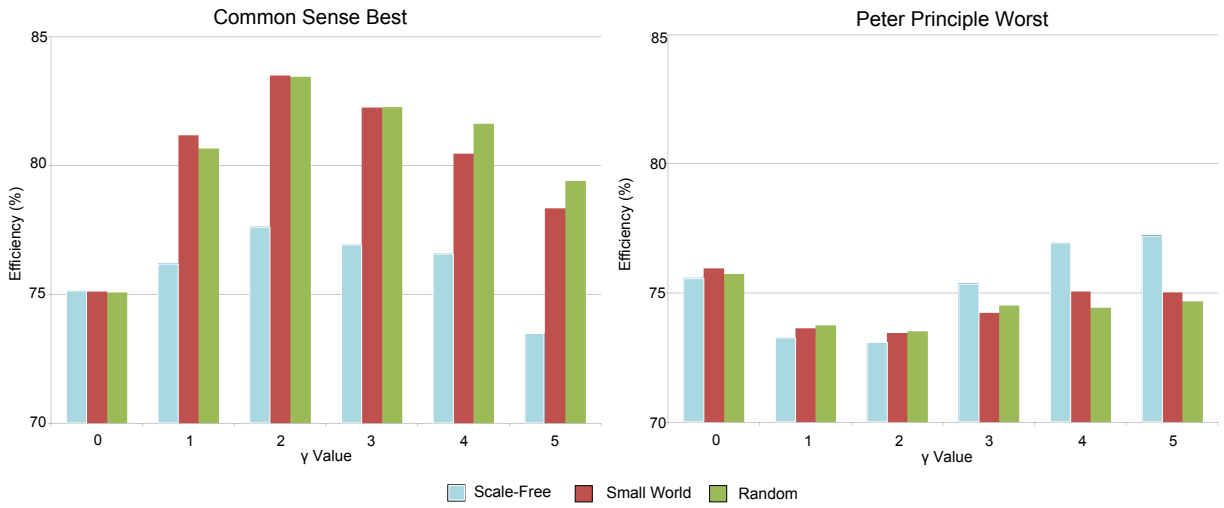


Figure 4.8: Comparison of varying γ effect upon averaged steady state efficiencies.

producing highly competent agents, managerial positions are scarce - efficient individuals becoming disillusioned with the system and dropping their competence.

The resulting behaviour is such that a decrease in efficiency is perpetrated by moderately competent individuals being passed over for promotion, becoming disillusioned, and with a high γ are forced to leave the system only to be replaced by less competent agents. The reverse is true of PP-worst, as unfairness is most influential at $\gamma = 2$, increasing the number of highly competent individuals as $\gamma \to 5$; this in turn causes efficiency to increase, producing the dip highlighted in Figure 4.8. The alternative promotion methods produce results akin to those of Figure 4.8, as such, further graphical output has been reserved for Appendix A.2.

Overall, figures produced by the NBM demonstrate a deviation in results from the Verification Model, and subsequently the results of Pluchino et al. (2010). The greatest departure centres around the reduction in efficiency of the PP-worst system, propelling CS-best methods to produce the most efficient hierarchy. Further contrastive evidence highlights the replacement of PP-best by CS-worst as the creator of the most inefficient hierarchies. The indicated results serve to emphasise the effect of network structure and behaviour upon the operations of an institution, presenting a more detailed insight into organisational thought.

4.6 NBM Discussion and Conclusions

The NBM has aimed to delve deeper into the claims of Pluchino et al. (2010, 2011), whose findings conclude that - in the absence of a clear understanding of promotional dynamics - promotion at random is the most favourable method of maximising efficiency. While the addition of social networks and behavioural minutiae have not changed this conclusion, a number of key findings in relation to this thesis are highlighted:

- **Including behaviours and social networks alters system outcomes**: PP-worst efficiency demonstrated a sizeable decrease (Verification: 80%, NBM-SF: 73%, NBM-SW: 74%, NBM-RAN: 74%), while PP-best indicates a small increase in efficiency (Verification: 60%, NBM all topologies: 65%), bringing the PP results as a whole closer to the efficiency of promotion at random (Verification PP: 70%, Verification CS: 71%). This demonstrates the value of considering behaviours and social networks in the exploration of social systems, indicating the suitability of ABS for such investigations.

- **Social network structures are important**: The imposition of a scale-free network produced differing levels of system efficiency (CS-best: 77%, CS-worst: 66%), when compared with those of small world (CS-best: 81%, CS-worst: 61%) and random (CS-best: 80%, CS-worst: 62%) networks. This indicated the specific structure of the social network is important, when considering the effect of behavioural influence in a connected system.

- **A diminished need for random promotion**: The inclusion socially active agents who can bypass normal promotional rules due to their connections, may be providing an incidental element of random promotion - potentially negating the need for explicit random-based hiring mechanisms. This further demonstrates the importance of considering social networks in conjunction with individual behaviours, and their effect upon a system.

While the ABS created cannot be lauded as capturing the full dynamic of workplace interaction, of particular interest is the differing system efficiencies observed with alternative social network structures. This would suggest that to understand the interplay between social networks and individual behaviours, it is essential to first investigate the architecture

of social network structures. To effectively explore the composition of real-world social structures, appropriate data is required.

Unfortunately, data for the analysis of workplace social networks was unavailable to the investigator, meaning that the specific application of organisational hierarchies shall not be considered further in this research. However, this thesis has secured data for an alternative social environment, that of adolescent school based networks; an analysis of their composition is conducted in Chapter 5. Prior to investigating adolescent social networks, the work of this chapter shall be concluded with a discussion of NBM limitations (Section 4.6.1) and potential further considerations (Section 4.6.2).

4.6.1 NBM Limitations

The conclusion presented may be confidently noted in relation to the work of Pluchino et al. (2010, 2011), a product of the comparative steps taken on construction of the model. However, the validity of the work must also be assessed in relation to the insight provided in conjunction with managerial processes. Evidently the model is theoretical, as such the collection of real workplace data - regarding hierarchical structure and inter-office relations - is suggested as the first progressive advancement. While the specifications of Pluchino et al. (2010, 2011) clearly state the proposed hierarchy is arbitrary, a selection of tier-based personnel capacity based upon an existing organisation may provide conclusions with more literary weight.

The simulation time frame of 1000 years has previously been highlighted as questionable, retained purely for comparative analysis. With regard to the granularity of the system proposed by Pluchino et al. (2011), a more prudent approach may be to explore turnover rates within an real world organisation. The current simulation assumes all agents retire or are dismissed due to incompetence, but it may also be conceivable that agents wish to explore alternative positions within a different organisation - albeit the dropping of competence due to unfairness may arguably control for this. The structuring of the hierarchy may also consistently evolve, the size of the firm developing as a result of its inherent success or failure - a notion also not considered by the NBM or Pluchino et al. (2010, 2011).

On reflection, the NBM indisputably cannot be considered an accurate representation of corporate institutional dynamics. However, the presented work has never intended to pro-

pose general statements regarding the improvement of organisational efficiency. Rather, the work aims to present an opposing aspect to the replicated work of Pluchino et al. (2010) - investigating the relevance of ABS in the context of modelling human behaviours in association with social networks. Given the proposed research objectives of this chapter have been achieved, the limitations do not appear to hinder the overall conclusions.

4.6.2 Further Considerations

If this work were to be continued and expanded, the limitations presented would evidently need to be rectified. Working with a target corporation, modelling a real world organisational structure may provide the essential data necessary for broader, more generalised conclusions. Aside from such considerations, the NBM could be expanded to encompass a greater depth of managerial literature. The following items explore some potential new directions:

- *Dominant Coalition* - The dominant coalition is said to be a group of high level managerial staff that predominantly control the mission and goals of an organisation (Bowler, 2006; Cyert & March, 1992). Exploration into the effect of Peter-related incompetence upon the dominant coalition, may alter effects on efficiency in conjunction with varying promotional schemes. Furthermore, the significance of being socially linked to the dominant coalition may provide more social capital; creating more opportunity for promotion.

- *Promotion Control* - As previously discussed, the Peter effect may corrupt any hierarchical system and also our everyday surroundings. It therefore stands to reason that the panel who assess a candidate for promotion may also be incompetent, the outcome is such that the candidate promoted may not be suitable for the position based upon their previous work. However, in a PP world this would translate into an incompetent candidate being removed from their position - competence being randomly redistributed - resulting in a potential higher level efficient candidate-position match.

- *External Hires* - The model created for investigation contains only one entry route for new agents, the lowest tier. In the real world, positions are often advertised outside the domain of the internal hierarchy. It may be the case that fresh agents

revitalise the system, boosting efficiency. Conversely, it may also be the case that agents transferring from other institutions may not be able to adapt to their new surroundings - reducing efficiency further.

- *Rewards and Prizes* - The topic of providing agents with an incentive scheme to counteract the negative psychological effect of the PP has been suggested in Pluchino et al. (2011) and also discussed at length by Fairburn & Malcomson (2001). Given that a basic interpretation of the psychological aspects of promotion have been included in the NBM, it offers a platform upon which to assess the impact of a reward scheme and the subsequent changes (if any) to the results profile.

- *External Systems* - This work has focused upon a closed hierarchical system, but external factors may also be affecting efficiency and the Peter Principle. Creating a number of competing organisations, and incorporating an element of the current financial climate, may offer a more contextualised insight. Making agents aware of the internal dynamics of the hierarchy has changed the efficiency results, therefore making agents perceptive to the wider social and organisational implications, may also be significant.

It is assumed in the real world, promotion on merit (best) is the factor that decides our ability to secure higher level positions; this method producing efficiency figures on average nine points higher (CS-best) and four points lower (PP-best) than random promotion. While promotion at random may offer a compromise in the absence of true understanding surrounding competence transition, the effect of the PP may be considered diminished enough that promotional governance becomes irrelevant. On balance, given the potential gains of promotion on merit outweigh the losses of a random system, it is recommended - in the context of the created hierarchy - that promoting the best candidate will maximise potential efficiency.

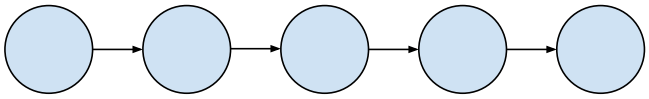
4.7 Chapter Summary

In summary, this chapter investigated the use of ABS amalgamated with theories from social literature regarding human behaviour and connection. The work examined existing publications regarding organisational inefficiency related to the PP, focusing particularly

on those studies referencing the use of simulation methods. The created NBM, as documented in Fetta et al. (2012), incorporates elements previously unexplored in conjunction with simulation methods and organisational dynamics, further developing the discussion concerning the existence of the Peter Principle.

This chapter particularly focused upon the work of Pluchino et al. (2010, 2011), who won an IG Nobel award for their findings (Improbable-Research, 2010) . The research of Pluchino et al. (2010, 2011) suggested promoting at random improves efficiency under PP conditions, with the conclusions being publicised in 'The Guardian' (Abrahams, 2010) and 'The New York Times' (Thompson, 2009). However, the conclusions from this chapter suggest that the need for random promotion (at negating the PP) may be unfounded.

The key finding of this chapter, in relation to the overall thesis, is an alteration to system outcomes based upon the network structure selected. As this thesis aims to investigate social networks and their effects, it would appear essential to first understand the composition of real world social structures, prior to fully investigating their impact. As such, the analysis of Chapter 5 conducts an in-depth examination of social connection, specifically focused upon adolescent school based networks; the findings of Chapter 5 later inform the development of a new algorithm to predict social network evolution (PageRank-Max), discussed at length in Chapter 6.



-"A Directed Line Graph"

5

Data Analysis: ASSIST

This chapter begins the investigation of social network structure, a key element in the development of a new method to predict social network evolution. Chapter 4 identified the importance of social network structure, and the influence it may have upon the outcomes of a connected social system. This chapter builds upon the work of Chapter 4, through the investigation of real-world friendship data - examining factors important in both social network construction and influence. The insights gained shall be utilised to develop a new algorithm to predict link evolution in a social network (Chapter 6), and create a model of the interplay between social structure and smoking behaviour (Chapter 9).

The data discussed in this chapter, is taken from datasets held by the "Centre for the Development and Evaluation of Complex Interventions for Public Health Improvement" (DECIPHer). The DECIPHer records contain information relating to social structures and smoking in adolescents, extracted from a study entitled "A Stop Smoking in Schools Trial" (ASSIST). Details regarding ASSIST, and the literature published from its developments, are discussed in Section 5.1; this provides context to the data, and outlines previous insights

gained into the connectivity observed in the ASSIST social networks.

On completion of this introductory overview, an in-depth analysis of the ASSIST social network structures is conducted (Section 5.2); this investigation outlines factors that appear important in both the structure of the networks and the resulting individual behaviours, directly informing the algorithm developed in Chapter 6. Additionally, the smoking outcomes of the trial are also examined, providing information related to the smoking uptake rates of the ASSIST adolescents (Section 5.3); this is further explored in the model developed in Chapter 9. An overview of the conclusions of this chapter, and their relevance to the thesis, is presented in Section 5.4.

5.1 ASSIST: The background

The formulation of ASSIST began in the mid 1990's, beginning with a small feasibility study to assess the impact of the intended research techniques (Bloor et al., 1999). The study strove to explore the effects of social networks upon attitudes toward adolescent smoking, with a view to inform potential cessation proliferation methods. Formed through a joint venture between 'Cardiff University Institute of Society, Health and Ethics' and 'The Department of Social Medicine at the University of Bristol', the project made use of informal peer education methods - grounded in *diffusion* and *social learning theory* (Holliday, 2006).

The ASSIST design was constructed as a peer-led intervention, formulated around the 'Gay Hero' work of Kelly et al. (1992). Kelly canvassed men who regularly patronised gay bars in eight US cities, identifying socially prominent individuals ('trendsetters') from whom a message could disseminate effectively (Kelly et al., 1997). Those selected were given training to diffuse safe sex practises, in a bid to encourage community level HIV prevention; findings demonstrating a significant reduction in risk behaviours following intervention. ASSIST replicated the work of Kelly with a community of adolescents, the "safer sex" message replaced with that of "stop smoking" (Audrey et al., 2004).

5.1.1 Methods

Following feasibility study success (Bloor et al., 1999), the Medical Research Council funded a large scale evaluation of the ASSIST project. 59 Schools from England and Wales were recruited to the intervention, through stratified randomisation (Starkey et al., 2005), targeting a cohort of Year 8 students (12-13 year olds) over the course of a three and a half year period. The study imposed randomised control conditions, 30 schools administered as 'intervention' - whereby students received 'hero' training - and 29 schools classified as 'control' (Audrey et al., 2004). All students were requested to undergo a "whole community" nominations procedure (Valente & Davis, 1999), attempting to identify influential students who may act as opinion leaders.

The 'Smoking Hero' (termed as a *peer supporter* within the study) selection process, resulted from the completion of a questionnaire prior to initialisation of the study. Three questions were posed, asking participants to identify:

- respected fellow students;
- leaders in sports or other group activities;
- individuals who are "looked up" to in Year 8.

The number of unique occurrences of an individual's name on a questionnaire were tallied, the sum of which constituting a nomination score for each participant. Utilising the nominations score, the top 17.5% of male and female nominated candidates (including smokers) from each intervention school were selected to be peer supporters - successful individuals receiving a two-day residential training programme to understand their roles.

Training consisted of informing peer supporters about the health, economic and environmental impacts of smoking, receiving specialised training in the subtleties of informally communicating educational material to peers. The selected participants were then requested to intervene in smoking related situations, attempting to convey the negative smoking message - peer supporters were also advised to utilise cessation material provided to each intervention school. The impact of peer supporters was assessed through questionnaires, distributed over the course of the study - the details of which are discussed in the following section (5.1.2).

5.1.2 Data Collection

ASSIST was a longitudinal study, with data collection occurring at four distinct time points. The initial collection period (T_0) examined conditions pre-intervention, the data being gathered simultaneously with the nominations process. Approximately six months later (T_1), each participant was again requested to complete a questionnaire - allowing for comparison with baseline information. The questions at T_1 also requested data previously undisclosed by study participants, that of information relating to their social network; two follow-up questionnaires also requested such information (T_2 and T_3), each distributed at one year intervals from those of the preceding time step (Figure 5.1).

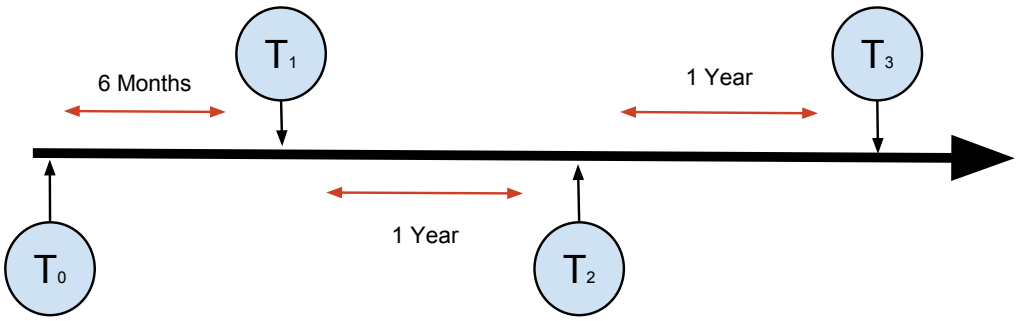


Figure 5.1: Time line of ASSIST data collection.

The questionnaires enquired about a range of attributes, from self reported smoking behaviour to family affluence, the defining feature of the study being the collection of large scale 'real world' social network data at T_1, T_2 and T_3. Each participant was asked to name up to six other students with whom they shared a friendship, the nature of which could be categorised as "best friend" or "just a friend"; information related to the type of activities the students conducted together was also collected. From such data, a school based social network may be constructed - describing friendship evolution over the course of the three year collection period. A copy of the social network data questionnaire is presented in Appendix B.

The relational data collected, offers the ability to investigate both the attributes of individual students within schools and cross-reference them against the attitudes of their connections. All attributional questionnaire data is securely contained within a Microsoft Access database, however much of the social network data remains in paper form stored

within the vaults of Cardiff University. As such, DECIPHer has provided a selection of 18 (electronically accessible) schools of network data for the purpose of this investigation; an assortment of attributional data for all schools has also been contributed, detailed in Section 5.1.3.

5.1.3 Selected Variables

Two types of data were provided by DECIPHer, student attribute data and social network data. The attributional data tables presented, contain over 120 variables and 11,454 records (for each time step). Social network data is provided in adjacency matrix form, details regarding strength of ties and friendship nature are unfortunately unavailable. Many of the variables listed in the database relate to varied responses of the same question; for brevity, the key informational areas are categorised as follows:

- *General* - Basic participant data requested relating to age, sex, ethnicity and tutor group.

- *Smoking* - Student reported smoking behaviour, classified on a scale from one to six; a smoking value of one indicating "never smoked" and six representing "more than 6 cigarettes a week". Smoking perceptions also requested: the "coolness" of smoking, the "likeability" of a smoker and the availability of cigarettes to an individual (friend, parent, corner shop etc.).

- *Familial Influence* - Data regarding parental smoking, parental expectations and wider family smoking behaviour. Responses were requested solely from participants who indicated engagement in cigarette use.

- *Socioeconomic Standing* - Questions relating to family cars, holidays and bedroom occupancy, the summary of which condensed to form a "Family Affluence Scale" (FAS) (Boyce et al., 2006). FAS ranges from zero to six, zero being the least affluent, six being the most affluent.

- *Self Image and Aspirations* - Concerns about body image, performance in school and ability to integrate in a group. Students were also requested to give expectations following school, such as: "getting a job", "becoming an apprentice" or "going on to further education".

DECIPHer provided data in an anonymised format, preserving the privacy of respondents. Each school possesses a reference number (11-87), while each student has a unique numeric identifier. Cotine saliva samples were also taken from each student at each time point, attempting to verify the self reported smoking levels detailed; this data has not been provided, therefore questionnaire responses alone are to be used for smoker analysis.

5.1.4 Previous ASSIST literature

Literature relating to ASSIST may be identified in publications of Social Sciences, Medicine and Network Science - demonstrating the breadth of disciplinary components involved in the study. Initial publications focused predominantly upon the effectiveness of a number of smaller scale trials, the research finding the proposed intervention successful in increasing the number of ex-smokers (high risk adolescents) that remain abstinent. However, preliminary findings did not demonstrate any further statistically significant differences between control and intervention schools smoking rates; to improve diffusion rates, minor changes were suggested for the larger scale intervention (Bloor et al., 1999).

Further publications relate to the feasibility, cost effectiveness and novelty of the project (Audrey et al., 2004; Starkey et al., 2005); the necessity of ASSIST based upon a lack of prior successful school-based smoking interventions. With regard to the execution of the study, the selected peer supporters were found to be effective in actioning their roles (Audrey et al., 2006a) - teachers expressing positive feedback at the possibility of implementing the intervention into the standard Year 8 curriculum (Audrey et al., 2008). Following a process evaluation (Audrey et al., 2006b), which assessed effective conduct of the methods adopted in the trial, analysis of the study outcomes were conducted by Campbell et al. (2008).

The findings of Campbell et al. (2008) suggest a reduced smoking prevalence in intervention schools at T_1 and T_2, however, a lower proportion of smokers in intervention schools was also present at baseline (T_0). The odds ratio of being a smoker in an intervention school, when compared with control, was significant at T_1 (odds: 0.77) but not at T_2 (odds: 0.85) - suggesting an attenuation of the intervention over time. No evidence was found to indicate a specific reduction in high risk smokers, however, the intervention exacts a more pronounced effect upon South Wales valley schools. Overall, Campbell et al. (2008) concluded ASSIST as a success - suggesting the adoption of the project nationwide, especially

in close-knit communities.

ASSIST was reported to have high fidelity (Holliday et al., 2009), suggesting the quality of intervention implementation was high. (Holliday et al., 2009) also found that, in the analysis of social network data, friendship and smoking behaviour was associated. Following multivariate longitudinal analyses, the friendship-smoking association was described as more complex than that of simple peer influence - with other factors also said to be of importance (Holliday et al., 2010). Researchers then employed the use of SAB methods to assess the co-evolution of friendship and smoking, the results indicating a "time heterogeneous process" whereby different elements are important at different time steps.

The SAB model findings suggest that, while initially smokers are influenced by the behaviour of their friends (T_1), as the students mature, friendship selection becomes based upon behavioural similarity (T_2) - students dropping friends embodying differing values (Mercken et al., 2012b). As such, scrutiny of the peer supporters selected to diffuse the "stop smoking" message was also conducted. The network analysis of Holliday (2006) and Starkey et al. (2009) found an appropriate number of peer supports located in segregated friendship groups, indicating positive levels of *clique embeddedness*; however, results appear derived from a small selection of intervention schools, with conclusions drawn from comparisons of nominated and non-nominated agents - differences in group size potentially affecting outcomes.

The literature presented hails ASSIST as an effective peer-led intervention study, providing a cost-effective method for increasing adolescent smoking cessation (Hollingworth et al., 2012); the study also appearing in European-wide assessments of smoking prevention methods (Mercken et al., 2012a). While the publications documented detail varying research aspects, the wealth of data provided offers the opportunity for further analysis - especially through the utilisation of social network information. The attributional and relational data acquired, provide a quantitative underpinning for the remainder of this thesis - a detailed investigation of which follows in Section 5.2 and Section 5.3.

5.2 ASSIST Network School Analysis

This section aims to analyse the available school social network data. The analysis is formed around 18 of the original 59 ASSIST participating schools, the investigation be-

ing structured in the following manner: Section 5.2.1 focuses upon context, providing a general overview of each school; Section 5.2.2 analyses the attributes of students within the schools, drawing particular attention to smoking related study outcomes; Section 5.2.3 examines the social network structure of each school, offering a network perspective of smoking uptake; and Section 5.2.4 formulates the resulting conclusions of the preceding sections.

5.2.1 Context

From the 18 electronically available "network" schools, 6 are classified as intervention, while the remaining 12 are control. Basic information regarding each network school may be found in Table 5.1, giving a brief summary of trial participant numbers, gender proportions and approximated school geographical location; additional school information is also detailed, highlighting any defining features of the cohort. For privacy protection, each school is made reference to by a unique identification number.

Each school offers a differing perspective of the trial, being distributed in a number of locations and housing varied social norms. The reasoning behind the inclusion of school specific information, is to provide context to the impending analysis. The literature of Chapter 3 suggests the consideration of social context in the interpretation of behaviour, therefore the inclusion of such information may be relevant. Further details of the specific contextual information are as follows:

- Process - Schools 12, 33, 34, 63, 71, 74 and 76 underwent evaluation procedures over the course of ASSIST, both to assess the conduct of the investigators administering the trial and study procedures as a whole. Process evaluations have the potential to cause bias in a randomised control trial (Audrey et al., 2006b), therefore the effect of the assessment measures upon analysis outcomes may be important.

- Welsh Valleys - Schools 62, 64, 71, 73 and 74 are centred in the Welsh valleys. The intervention schools residing in these areas are said to demonstrate particular success in the reduction of smoking uptake (Campbell et al., 2008), therefore valley schools may have particular properties differentiating them from other schools within the data.

- Academy - School 15 is an academy. An academy may accept external sponsorship

School Id	Type	Participants	Female	Male	Area	Additional Information
12	Intervention	164	50.61	49.39	Bristol	Process School
15	Control	188	45.74	54.26	Somerset	Academy
22	Control	149	51.01	48.99	Bristol	Comprehensive, No longer in existence
32	Intervention	229	45.41	54.59	S. Gloucestershire	Comprehensive
33	Control	153	54.25	45.75	Bath	Process School
34	Intervention	200	52.00	48.00	Bristol	Process School, No longer in existence
35	Control	158	53.80	46.20	Bristol	Catholic School
40	Control	62	100.00	0.00	Bristol	Independent Girls School
41	Control	172	45.93	54.07	Bristol	Comprehensive, No longer in existence
62	Control	144	41.67	58.33	Newport	Comprehensive, valleys
63	Control	236	52.12	47.88	Penarth	Process School
64	Control	170	44.71	55.29	Monmouthshire	Comprehensive, valleys
68	Control	164	44.51	55.49	Cardiff	Comprehensive
69	Control	80	58.75	41.25	Cardiff	Welsh Medium Comprehensive
71	Control	102	44.12	55.88	Cardiff	Process School, valleys
73	Intervention	199	49.75	50.25	Newport	Comprehensive, valleys
74	Intervention	123	47.97	52.03	Newport	Process School, valleys
76	Intervention	254	46.06	53.94	Newport	Process School

Table 5.1: Network data school information, colour coded by intervention type and country. Purple and blue cells represent intervention and control schools, respectively. Green cells indicate English schools and red cells indicate Welsh schools.

contributions to those of central government funding - sponsors being able to dictate aspects of the curriculum, governing body members and specialism. The effect of an academy ethos upon students may differ to those of standard comprehensive school practices, the results of which may naturally produce varied findings.

- Girls School - School 40 is a fully independent girls school, not associated with any form of government funding. The small homogeneous sex grouping of school 40, offers the ability to asses the impact of gender, affluence and class sizes upon smoking behaviours.

- Cultural Norms - Welsh language school 69 and Catholic school 35, offer alternate cultural environments to those of a state funded English language Comprehensive; the importance of said factors may have resultant outcomes upon adolescent behaviours.

While only basic contextual information is available within the data, the details provided may have the potential to inform the results produced in the following sections. Table 5.1 serves to provide a point of reference throughout the following analysis, and shall be regularly referred to over the course of this work.

5.2.2 Attribute Analysis

Control and intervention school attribute data are displayed in Tables 5.2 and 5.3, respectively. The tables present information regarding the general student body of each school; examples including: averaged Family Affluence Scale (FAS) values, smoking prevalence and the proportion of students possessing a parental smoker at home. Independent sample t-tests, or Mann-Whitney for non-parametric data, have been conducted to compare differences in school type; significance at the 0.05 level is indicated by (*) in the Average column of each table.

FAS values between control and intervention schools are significantly different at both T_0 and T_2, suggesting control schools contain more affluent students; FAS data being unavailable for T_1 and T_3. In a review of socio-economic impact upon adolescent smoking, Hiscock et al. (2012) states higher smoker prevalence may be observed in low affluence families. Given that no significant difference in smoking levels is apparent, said conclusions may not be drawn from the network school data. However, it must be considered that

Measure	Time	15	22	33	35	40	41	62	63	64	68	69	71	Average
FAS	T_0	3.92	3.85	4.37	4.04	4.74	3.57	3.57	4.2	4.03	3.08	4.24	3.67	***3.94**
(Average)	T_2	4.11	3.93	4.19	4.06	4.66	3.61	3.72	4.25	3.89	3.25	4.14	3.78	***3.97**
	T_0	49.45	48.99	41.06	36.08	22.58	42.44	51.39	36.86	45.29	39.63	36.25	40.20	40.85
Parental Smoking	T_1	40.11	48.32	38.41	34.81	16.13	47.09	50.69	36.86	39.41	32.32	27.50	43.14	37.90
(%)	T_2	50.00	53.69	45.03	36.71	22.58	45.93	52.08	36.44	44.71	41.46	37.50	47.06	42.77
	T_3	46.70	47.65	42.38	31.65	19.35	31.40	49.31	28.39	38.24	40.85	31.25	44.12	37.61
	T_0	11.52	7.04	14.18	13.25	1.67	9.74	9.70	7.30	14.63	7.30	7.14	29.67	11.09
Smokers	T_1	11.66	13.29	13.67	20.65	3.23	7.98	21.32	14.16	11.18	10.32	9.09	35.87	14.37
(%)	T_2	24.00	24.32	24.66	22.29	18.03	10.53	16.20	23.50	21.43	21.74	7.50	37.00	20.93
	T_3	24.26	34.97	36.30	31.51	23.21	16.08	31.85	21.27	34.62	29.25	18.84	41.67	28.65
	T_0	88.48	92.96	85.82	86.75	98.33	90.26	90.30	92.70	85.37	92.70	92.86	70.33	88.91
Non-Smokers	T_1	88.34	86.71	86.33	79.35	96.77	92.02	78.68	85.84	88.82	89.68	90.91	64.13	85.63
(%)	T_2	76.00	75.68	75.34	77.71	81.97	89.47	83.80	76.50	78.57	78.26	92.50	63.00	79.07
	T_3	75.74	65.03	63.70	68.49	76.79	83.92	68.15	78.73	65.38	70.75	81.16	58.33	71.35
	T_0-T_1	0.14	6.24	-0.52	7.40	1.56	-1.76	11.62	6.87	-3.45	3.02	1.95	6.20	3.27
Smoker Increase (%)	T_1-T_2	12.34	11.04	10.99	1.65	14.81	2.55	-5.13	9.34	10.24	11.42	-1.59	1.13	6.57
	T_2-T_3	0.26	10.64	11.64	9.21	5.18	5.56	15.65	-2.24	13.19	7.51	11.34	4.67	7.72
	T_0	9.34	4.70	6.62	4.43	3.23	10.47	6.94	1.27	3.53	16.46	12.50	10.78	7.52
Missing Smoker	T_1	10.44	4.03	7.95	1.90	0.00	5.23	5.56	1.27	10.59	23.17	31.25	9.80	9.27
Data (%)	T_2	3.85	0.67	3.31	0.63	1.61	0.58	1.39	0.85	1.18	1.83	0.00	1.96	1.49
	T_3	7.14	4.03	3.31	7.59	9.68	16.86	6.25	6.36	8.24	10.37	13.75	5.88	8.29
Missing Network	T_1	14.89	4.70	9.15	1.90	0.00	4.07	4.86	0.85	10.59	20.12	2.50	8.82	6.87
Data (%)	T_3	7.98	5.37	3.27	7.59	12.90	16.86	6.94	6.36	7.65	12.20	12.50	4.90	8.71

Table 5.2: ASSIST Control School Characteristics, * indicates a significant difference to intervention schools.

Measure	Time	12	32	34	73	74	76	Average
FAS	T_0	3.56	3.56	3.55	3.02	3.33	3.61	***3.44**
(Average)	T_2	3.44	3.70	3.38	3.27	3.10	3.56	***3.41**
	T_0	32.32	37.12	51.00	48.24	47.97	26.77	40.57
Parental Smoking	T_1	33.54	32.75	51.00	44.22	39.02	26.77	37.88
(%)	T_2	34.76	36.68	54.50	49.25	44.72	29.92	41.64
	T_3	31.10	35.81	41.50	43.22	43.09	26.77	36.91
	T_0	9.21	8.60	9.44	7.03	11.50	11.11	9.48
Smokers	T_1	10.46	15.74	12.23	4.52	8.47	17.21	11.44
(%)	T_2	20.50	24.67	33.50	13.07	18.70	16.54	21.16
	T_3	25.83	30.66	32.18	20.54	24.79	19.26	25.55
	T_0	90.79	91.40	90.56	92.97	88.50	88.89	90.52
Non-Smokers	T_1	89.54	84.26	87.77	95.48	91.53	82.79	88.56
(%)	T_2	79.50	75.33	66.50	86.93	81.30	83.46	78.84
	T_3	74.17	69.34	67.82	79.46	75.21	80.74	74.46
	T_0-T_1	1.25	7.14	2.79	-2.51	-3.03	6.10	1.96
Smoker Increase (%)	T_1-T_2	10.04	8.93	21.27	8.55	10.22	-0.68	9.72
	T_2-T_3	5.33	5.99	-1.32	7.48	6.09	2.73	4.38
	T_0	7.32	3.49	10.00	7.04	8.13	7.87	7.31
Missing Smoker	T_1	6.71	5.68	6.00	11.06	4.07	3.94	6.24
Data (%)	T_2	1.83	0.87	1.50	0.00	0.00	0.00	0.70
	T_3	7.93	7.42	13.00	7.04	1.63	3.94	6.83
Missing Network	T_1	6.71	5.24	6.00	10.05	5.69	5.12	6.47
Data(%)	T_3	9.76	8.73	15.50	5.53	1.63	4.72	7.64

Table 5.3: ASSIST Intervention School Characteristics, * indicates a significant difference to control schools.

diffusion effects at T_2 may be reducing smoking levels of the lesser affluent intervention schools - potentially bringing smoking levels closer to one another.

The lack of significance in observations related to smoking, also draws questions regarding intervention effectiveness; it suggests, within the constraints of the 18 network schools, there is little quantitative evidence of direct significant smoker reduction. Comparisons are based upon small samples sizes (6 intervention schools) and therefore the generalisability of results must be considered, with Audrey et al. (2006b) stating that simple quantitative outcomes are not appropriate for the measurement of health effects - qualitative discussions with participants also being necessary in the interpretation of alterations in opinion processes (Holliday, 2006). Further analysis of overall study outcomes shall be returned to in section 5.2.3, where a larger section of trial data is available.

Analysing school specific observations, intervention students situated in Welsh valley locations (73 & 74) appear to have a negative smoker increase (T_0 to T_1:-2.51 & -3.03 respectively); findings consistent with those of Campbell et al. (2008). The aforementioned schools appear to maintain a relatively low overall smoker population, the percentage of smokers at T_3 (73: 20.54% & 74: 24.79%) being lower than the average of both intervention schools (25.55%) and control schools (28.65%). However, valley control school 64 also demonstrates a smoker reduction from T_0 to T_1 (-3.45%) in absence of intervention conditions - indicating that taking isolated observations may not be wholly appropriate.

The overall smoker proportions of valley control schools at T_3 (62: 31.85%, 64: 34.62% & 71: 41.67%) are greater than average, indicating a difference between control and intervention measures. It would appear that in intervention schools, the "stop smoking" message has diffused with some effect over time, while in control schools, a positive smoking uptake message may be diffusing. The social structures of valley schools may therefore be naturally predisposed to message diffusion and social network influences; further analysis of network structures is provided in Section 5.2.3.

A large smoker increase is observed in school 40, rising from 1.67% (T_0) to 23.21% (T_3); this highlights a further social structure that may be conducive to message diffusion. School 40 has the highest FAS values of all the schools within the data set, smoking values therefore contradictory to the expectations within the literature of Hiscock et al. (2012) (as previously discussed). Furthermore, the school exhibits low levels of parental smok-

ing; parental smoking behaviours also said to influence that of their adolescent offspring (Bauman et al., 1990; Farkas et al., 1999; Newman & Ward, 1989). School 40 therefore does not appear to satisfy the expected conditions for high rates of adolescent smoking, suggesting other factors may also be of importance - such as social network structure.

Students of the school 15 academy display a less than average smoker population at T_3; smoker increase being small at T_0 to T_1 and T_2 to T_3, with the majority of uptake occurring between T_1 and T_2. The relatively low smoker population of 15 at T_3, is a trend also observed in schools 41 (16.08%), a low FAS English Comprehensive, and 69 (18.84%), a Welsh language high FAS Comprehensive. There would appear to be minimal similarities in attribute data between said schools, once again suggesting that factors such as FAS and parental smoking may not always be definitive variables in smoking uptake.

The school characteristics discussed appear to portray an unclear image with regard to conditions relating to smoker uptake, suggesting that global basic information alone cannot quantify an individual's decision to smoke. To investigate influential factors further, network analysis measures may also provide valuable insight. Analysis of paths, network cohesion, individual cohesion and clustering are described in the following section.

5.2.3 Network Analysis

For the 18 ASSIST network schools, tables 5.4 and 5.5 display the values of various SNA metrics. The measures selected are those detailed in section 3.1, calculated using the R software package. The appropriate statistical tests, comparing means of control and intervention schools at the 0.05 level, found one significant difference - that of closeness centrality at T_1 (indicated by (*) on the relevant table entries).

The closeness centrality of a node is calculated as the shortest path from itself to all other vertices, the average of all students (in a school) taken and standardised to compute the figures in tables 5.4 and 5.5. Closeness centrality in intervention schools appears significantly lower at T_1 than that of control schools, suggesting that (on average) individuals in intervention schools are more sparsely distributed - information potentially travelling at a slower rate around the network than in control schools. This evidently has substantial repercussions upon intervention diffusion, as peer supporters may not be able to effectively circulate the negative smoking message.

Table 5.4: Control Network Characteristics

Measure	Time	15	22	33	35	40	41	62	63	64	68	69	71	Average
Average Path Length	T_1	0.641	0.610	0.741	0.896	0.771	0.503	0.864	0.895	0.798	0.456	0.716	0.650	0.712
	T_2	0.822	0.694	0.940	0.952	0.834	0.840	0.853	0.915	0.910	0.745	0.735	0.815	0.838
	T_3	0.751	0.516	0.828	0.834	0.645	0.471	0.756	0.809	0.728	0.321	0.413	0.736	0.651
Degree (out)	T_1	3.601	3.490	4.203	4.576	3.532	3.721	4.132	4.627	3.965	2.378	3.775	3.775	3.815
	T_2	4.069	3.966	4.908	4.823	3.887	4.349	4.326	4.949	4.653	3.354	4.263	4.373	4.327
	T_3	4.043	3.383	4.621	4.456	3.306	3.506	4.014	4.203	3.782	2.707	3.675	3.980	3.806
Degree (total)	T_1	7.202	6.980	8.405	9.152	7.065	7.442	8.264	9.254	7.929	4.756	7.550	7.549	7.629
	T_2	8.138	7.933	9.817	9.646	7.774	8.698	8.653	9.898	9.306	6.707	8.525	8.745	8.653
	T_3	8.085	6.765	9.242	8.911	6.613	7.012	8.028	8.407	7.565	5.415	7.350	7.961	7.613
Reciprocity	T_1	0.549	0.665	0.641	0.617	0.676	0.673	0.615	0.584	0.555	0.549	0.510	0.561	0.600
	T_2	0.541	0.636	0.671	0.646	0.739	0.660	0.559	0.608	0.612	0.498	0.534	0.614	0.610
	T_3	0.582	0.647	0.605	0.619	0.673	0.677	0.599	0.603	0.598	0.545	0.565	0.645	0.613
Transitivity	T_1	0.402	0.437	0.436	0.371	0.452	0.481	0.421	0.311	0.390	0.362	0.458	0.372	0.408
	T_2	0.392	0.428	0.448	0.356	0.456	0.465	0.448	0.366	0.421	0.333	0.417	0.404	0.411
	T_3	0.404	0.374	0.397	0.343	0.378	0.489	0.400	0.386	0.367	0.400	0.502	0.427	0.406
Density	T_1	0.019	0.024	0.028	0.029	0.058	0.022	0.029	0.020	0.023	0.015	0.048	0.037	0.029
	T_2	0.022	0.027	0.032	0.031	0.064	0.025	0.030	0.021	0.028	0.021	0.054	0.043	0.033
	T_3	0.022	0.023	0.030	0.028	0.054	0.021	0.028	0.018	0.022	0.017	0.047	0.039	0.029
Betweenness	T_1	0.147	0.254	0.171	0.095	0.214	0.127	0.202	0.061	0.135	0.149	0.161	0.163	0.157
	T_2	0.159	0.193	0.194	0.068	0.199	0.118	0.225	0.084	0.11	0.135	0.097	0.105	0.141
	T_3	0.154	0.166	0.079	0.086	0.181	0.076	0.111	0.087	0.174	0.13	0.087	0.164	0.125
Closeness (in)	T_1	0.009	0.026	0.009	0.027	0.111	0.032	0.033	0.01	0.018	0.005	0.04	0.026	***0.029**
	T_2	0.026	0.039	0.05	0.035	0.08	0.029	0.012	0.03	0.03	0.009	0.025	0.068	0.036
	T_3	0.014	0.014	0.016	0.024	0.026	0.006	0.021	0.01	0.01	0.007	0.045	0.021	0.018
Degree Centrality	T_1	0.056	0.044	0.045	0.079	0.073	0.037	0.062	0.027	0.042	0.047	0.104	0.091	0.059
	T_2	0.037	0.054	0.04	0.071	0.084	0.045	0.047	0.026	0.055	0.041	0.073	0.095	0.056
	T_3	0.048	0.038	0.042	0.067	0.077	0.038	0.063	0.033	0.043	0.051	0.067	0.06	0.052

Table 5.5: Intervention Network Characteristics

Measure	Time	12	32	34	73	74	76	Average
Average Path Length	T_1	0.691	0.829	0.766	0.647	0.813	0.851	0.766
	T_2	0.845	0.869	0.838	0.815	0.828	0.936	0.855
	T_3	0.796	0.715	0.579	0.712	0.704	0.880	0.731
Degree (out)	T_1	4.360	4.279	4.000	4.075	4.407	4.354	4.246
	T_2	4.543	4.707	4.495	4.548	4.366	4.732	4.565
	3	4.244	4.354	3.750	4.065	4.098	4.469	4.163
Degree (total)	T_1	8.720	8.559	8.000	8.151	8.813	8.709	8.492
	T_2	9.085	9.415	8.990	9.095	8.732	9.465	9.130
	T_3	8.488	8.707	7.500	8.131	8.195	8.937	8.326
Reciprocity	T_1	0.618	0.602	0.628	0.580	0.539	0.584	0.592
	T_2	0.650	0.599	0.636	0.653	0.559	0.646	0.624
	T_3	0.586	0.602	0.637	0.586	0.624	0.641	0.613
Transitivity	T_1	0.431	0.401	0.429	0.461	0.324	0.324	0.395
	T_2	0.464	0.385	0.409	0.457	0.383	0.351	0.408
	T_3	0.403	0.405	0.433	0.389	0.539	0.347	0.419
Density	T_1	0.027	0.019	0.020	0.021	0.036	0.017	0.023
	T_2	0.028	0.021	0.023	0.023	0.036	0.019	0.025
	T_3	0.026	0.019	0.019	0.021	0.034	0.018	0.023
Betweenness	T_1	0.121	0.112	0.248	0.193	0.074	0.116	0.144
	T_2	0.144	0.082	0.212	0.107	0.145	0.072	0.127
	T_3	0.189	0.080	0.129	0.114	0.151	0.058	0.120
Closeness (in)	T_1	0.009	0.009	0.007	0.011	0.012	0.007	***0.009**
	T_2	0.011	0.021	0.008	0.010	0.027	0.027	0.017
	T_3	0.013	0.011	0.012	0.015	0.018	0.010	0.013
Degree Centrality	T_1	0.059	0.043	0.050	0.050	0.095	0.038	0.056
	T_2	0.070	0.080	0.043	0.033	0.054	0.041	0.053
	T_3	0.054	0.069	0.041	0.045	0.048	0.026	0.047

The intervention itself may be causing the low levels of closeness centrality experienced. The work of Audrey et al. (2006a) and Holliday (2006) documents a number of intervention school interview transcripts, some peer supporters feeling uncomfortable approaching students they are not friends with - smokers interpreted as disliking the authority afforded to peer supporters. The process of explicitly highlighting influential individuals in a network may actually be making peer supporters unpopular, this coupled with the unease some supporters feel in approaching smokers, degrading their ability to effectively diffuse the trial message. Furthermore, those individuals who do intervene in smoking related situations external to their friendship group, are having to consort with individuals they may not usually interact with; this may alter friendships, potentially affecting closeness centrality scores.

While the reason behind the differences in control and intervention closeness centrality at T_1 remains unclear, values do not appear significantly different at T_2 and T_3. The work of Campbell et al. (2008) suggests an attenuation of the intervention over time; discussions with analysts within the DECIPHer group suggest this may be due to peer supporters losing interest in their roles, or selected supporters no longer holding an influential position in their network. The underlying cause of intervention effectiveness reduction (Tables 5.2 and 5.3) appears to also alter closeness centrality (Tables 5.4 and 5.5), figures becoming more in line to those exhibited in control schools at T_2. It may be the case that the intervention is causing fractious groups within each school, closeness increasing as the roles of peer supporters diminish; however, this cannot be said with certainty as baseline (T_0) network data is unavailable. Further intervention school analysis follows in Section 5.2.3.1.

5.2.3.1 Intervention School Discussion

To gain a greater insight into the effect of social networks upon the success of intervention measures, a number of intervention schools are explored in detail. School 74 displays the largest *proportional* intervention success at T_1, reducing smokers by 3.03% (Table 5.3); however, this effect does not appear sustained at later timesteps. School 74 has a short Average Path Length (APL) of 0.813 (indicated by normalised disconnected APL being closer to one), and high degree statistics (out: 4.407 and total: 8.813) at T_1 compared to the respective averages for intervention schools; these figures drop below average as time progresses. The observed values suggest that, although initially the school 74 adolescents befriend a variety of individuals across the network, over time the connections become more

cliqued - indicated by an increase in reciprocation and transitivity. This behaviour can be observed in the network diagrams of Figures 5.2 and 5.3 - the Fruchterman-Reingold algorithm (discussed in Section 3.5) placing groups of individuals at greater distances from one another in T_2 than in T_1, due to the more cliqued segregated behaviour observed. Figure 5.4 is also the school 74 network at T_2, retaining the node placement of the network at T_1 from (Figure 5.2); this has been included to demostate both the change in friendships and the effect of the Fruchterman-Reingold algorithm.

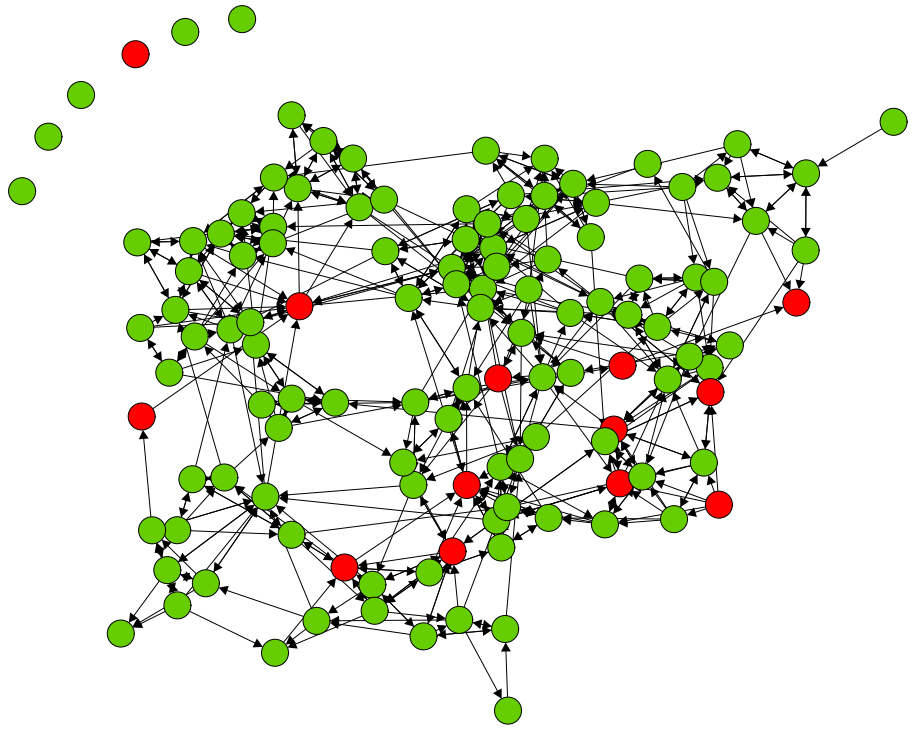


Figure 5.2: School 74 Social Network at T_1, red nodes indicate smokers.

The changing structure of school 74 over time, may also explain the low smoking uptake at T_1 (-3.03%) which increases above average at T_2 (10.22%) and T_3 (6.09%) - the nominated agents being unable to diffuse the intervention as effectively when the structure becomes more segregated. Examining the smoking characteristics of schools 73 and 76 (Table 5.3), overall smoker prevalence at T_3 appears lower than intervention average (73: 20.54%, 76: 19.26%); this would suggest some cases of peer supporter effect over the course of the trial. To further explore the differences in smoking uptake across school 73, 74 and 75, and to understand the diminishing effect of intervention in school 74, network images at T_3 highlighting the selected peer supporters are plotted in Figures 5.5, 5.6 and 5.7.

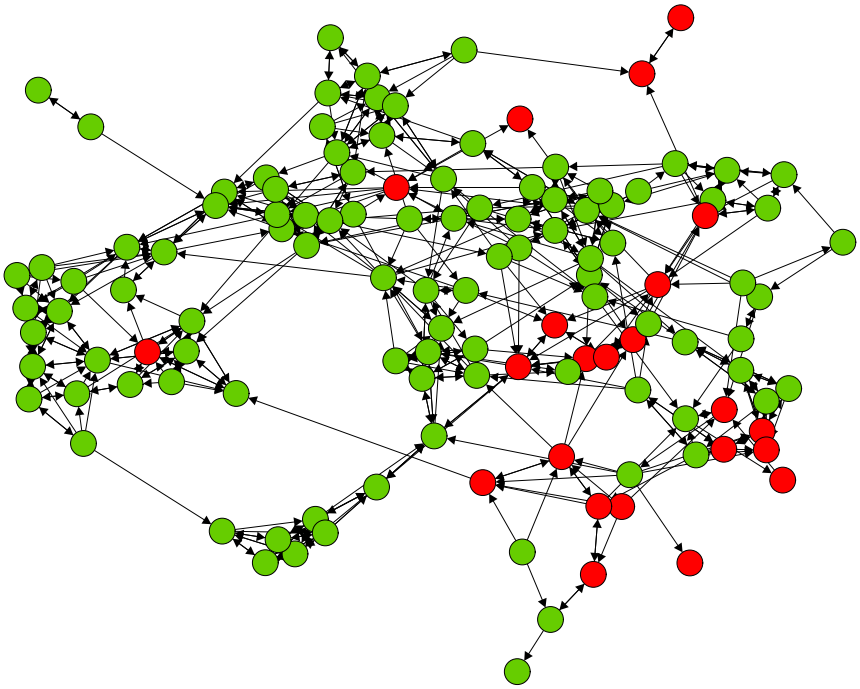


Figure 5.3: School 74 Social Network at T_2, red nodes indicate smokers.

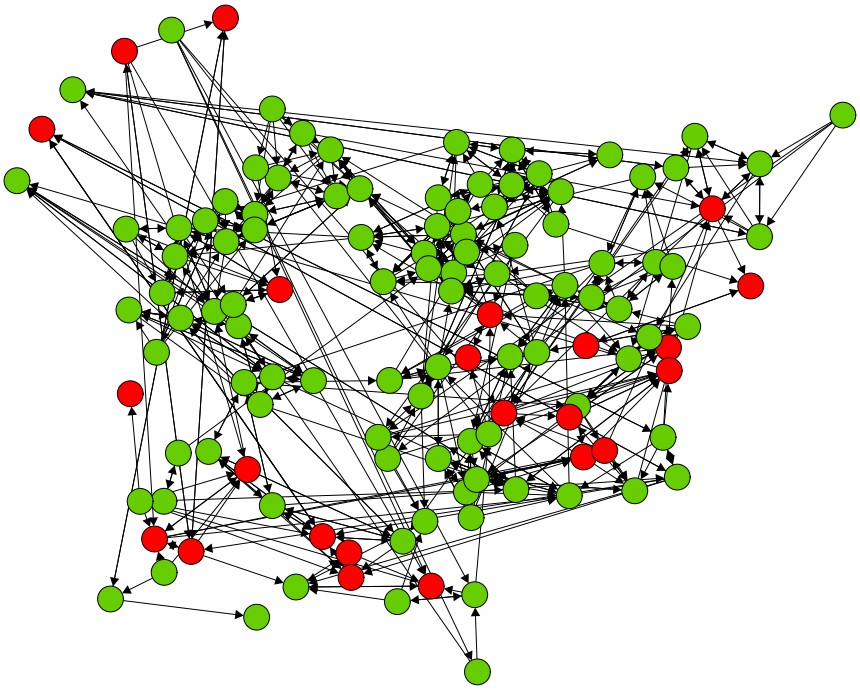


Figure 5.4: School 74 Social Network at T_2, red nodes indicate smokers. The nodes remain in the their original Fruchterman-Reingold layout from T_1 in Figure 5.2.

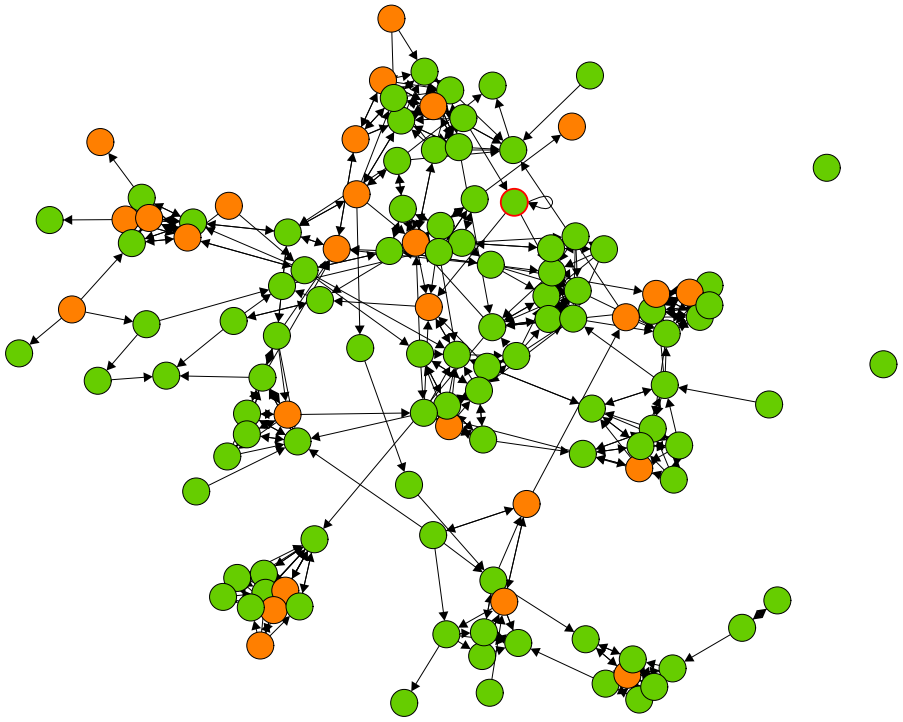


Figure 5.5: School 74 Social Network at T_3, orange nodes indicate peer supporters.

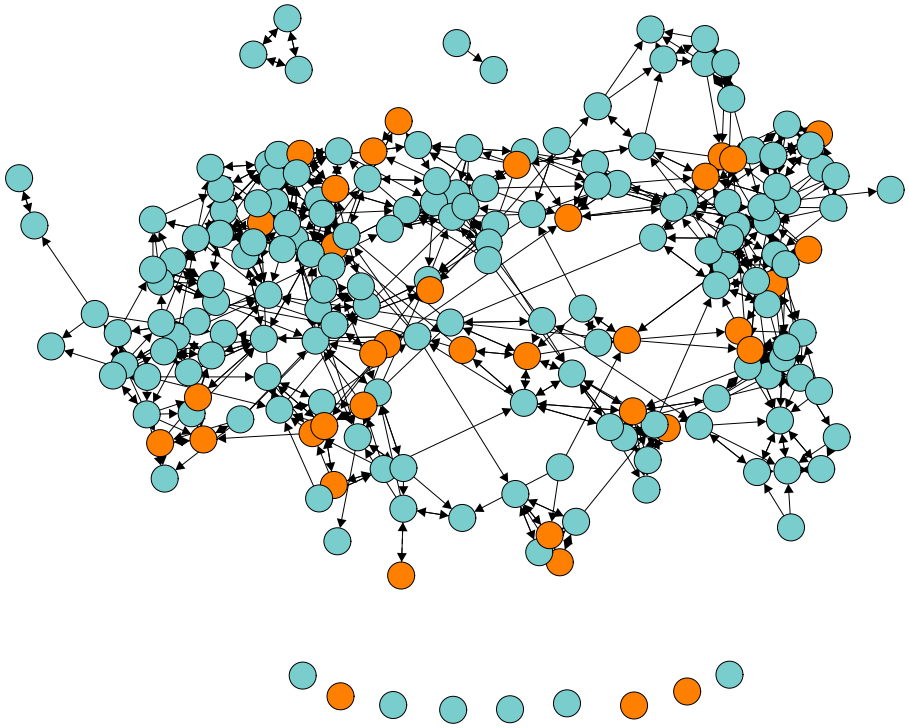


Figure 5.6: School 73 Social Network at T_3, orange nodes indicate peer supporters.

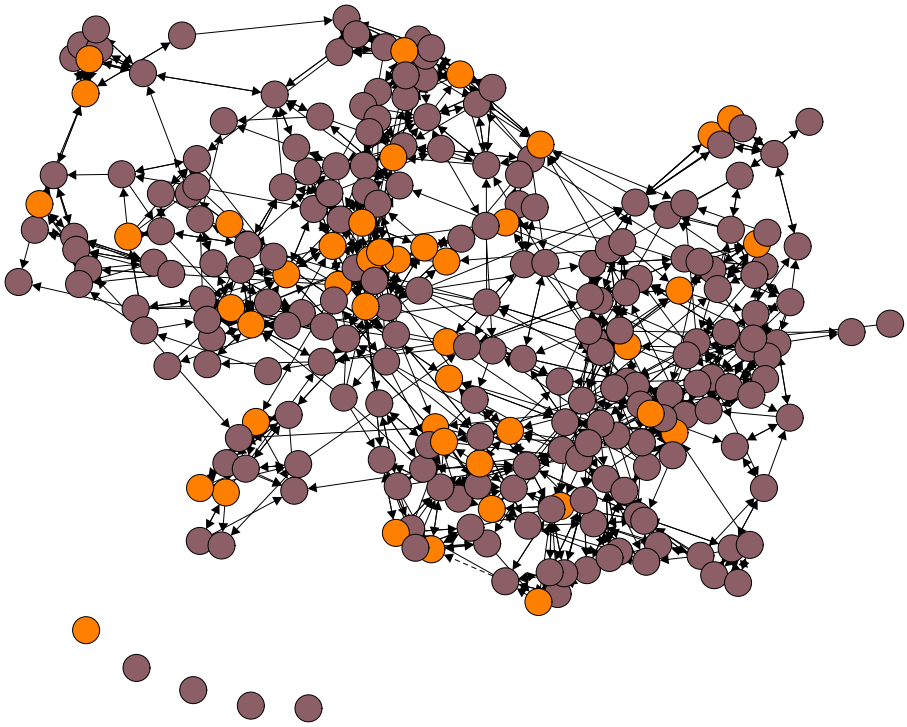


Figure 5.7: School 76 Social Network at T_3, orange nodes indicate peer supporters.

The structure of school 74 at T_3 (Figure 5.5) remains distant and cliqued, students grouping together away from the central structure of the network; findings reinforced by the longer APL (0.704) and lower degree statistics (out: 4.098 and total: 8.195) than those previously discussed at T_1. The orange nodes of Figure 5.5 highlight the peer supporters, with few appearing at the centre of the graph. The nominated students that are embedded within groups away from the centre of the network in Figure 5.5, while potentially effective at delivering the negative smoker message to their friends, may not feel comfortable approaching students external to their group (as suggested in Holliday (2006)) - reducing the overall momentum of the intervention. One student also values themselves particularly highly in school 74, said individual selecting themselves as a friend (highlighted by a red circled node in Figure 5.5).

The networks of schools 73 (Figure 5.6) and 76 (Figure 5.7) appear far more cohesive than that of school 74 at T_3, large groupings of students being evident. Nodes within the specified graphs overlap substantially, the spring like forces of the Fruchterman-Reingold algorithm drawing students together due to their clustering of friendships. It may be argued that, as schools 73 and 76 have a greater population (199 and 254 respectively, Table 5.1),

network images may visually appear more clustered due to larger node density; however, the network measures of Table 5.5 corroborate this interpretation.

School 76 has the shortest path length of all schools within the cohort at T_3, also exhibiting high levels of reciprocation and low transitivity; this means that there are a great deal of reciprocated ties that shorten path lengths, but that friends of friends are not necessarily friends. A similar structure is observed in School 73, however, according to the network statistics of Table 5.5, APL is longer (0.712) and reciprocity (0.586) is reduced. The reasoning behind the altered statistics may be the evident divide noticable through the centre of School 73 in Figure 5.6, two large defined structures appearing with a number of nodes between them. Furthermore the network of school 73 has a fully isolated triad and an unreciprocated dyad, also contributing to the longer APL.

The importance of comparing the structure of schools 73 and 76 is that, albeit slightly different in terms of overall cohesion, a large proportion of orange peer supporters in Figures 5.6 and 5.7 appear placed at the centre of the graph - few occurring at the periphery. This suggests a larger audience for peer supporters to convey the negative smoking message, with peer supporters able to interact with one another for help in actioning the intervention message to larger groups of students. The overall smoking uptake of school 73 appears higher than school 76, perhaps if the observed divide were not present, the school may have experienced a greater intervention success.

Also evident from the graphs of Figures 5.6 and 5.7, are the isolated nodes. A student may be depicted as isolated for any of the following reasons:

- A lack of connections with other students within the school (social isolation);
- The student has left the specific school and formed new connections in a new school;
- Missing data due to illness or withdrawal from the study.

The data provided by DECIPHer includes missing students in a school network at T_1 and T_3 if they are present at T_2, this means that missing data may occur at T_1 and T_3.

Table 5.6 illustrates the proportions of missing data at T_1 and T_3; no significant difference was found between control and intervention schools at either time step. Schools 73, 74

School	Type	T_1 (%)	T_3 (%)
12	Intervention	6.71	9.76
15	Control	14.89	7.98
22	Control	4.70	5.37
32	Intervention	5.24	8.73
33	Control	9.15	3.27
34	Intervention	6.00	15.50
35	Control	1.90	7.59
40	Control	0.00	12.90
41	Control	4.07	16.86
62	Control	4.86	6.94
63	Control	0.85	6.36
64	Control	10.59	7.65
68	Control	20.12	12.20
69	Control	2.50	12.50
71	Control	8.82	4.90
73	Intervention	10.05	5.53
74	Intervention	5.69	1.63
76	Intervention	5.12	4.72

Table 5.6: Proportions of missing data. No missing data for T_2 as network data sets are based on students present during T_2 data capture.

and 76 possess low amounts of missing data at T_3, meaning a greater level of accuracy in the overall network characteristics. The importance of considering missing data in the discussion of network measures, is due to the effect of isolated nodes upon relevant statistics; examples include elongating the average path length, decreasing degree statistics and reducing levels of overall centrality.

The inclusion of students with no connections because of missing data is partly due to the structure in which the data has been presented, but also to convention within the SAB software 'RSiena' manual (Ripley et al., 2012). SAB analyses regularly utilise longitudinal social network data, as such, it may not always be possible to obtain responses from all participants in the study across each timestep; to combat this, only those individuals present in the central waves of data collection (T_2) are included. This minimises the overall effect of missing data, while maximising the social network information available; those individuals available at T_2 but missing at T_1 or T_3, simply imputed as having no outward connections.

The reasoning for this thesis following SAB convention with regard to missing data, is

due to the work of Chapter 6 attempting to explore the SAB process as a Link Prediction (LP) method. LP methods are generally exacted upon complete data sets in which missing data do not factor (such as ArXiv data or airline flight networks) or upon training datasets (whereby links are strategically removed to assess the abilities of the selected algorithms). As conventions relating to LP methods are not asserted, those of the SAB method have been adopted - the discussion of implementation of missing data with the LP problem continues in Chapter 6. Given that no significant differences occur in the amount of missing data between control and intervention schools, and given the recommendations of the RSiena manual, individuals with missing responses at T_1 and T_3 shall be included as having no connections.

Returning to the analysis of intervention schools, School 34 observes the highest smoker uptake of all intervention schools. The network characteristics of Table 5.5 indicate above average levels of betweenness in School 34, especially at T_1 (0.248) and T_2 (0.212). Plotting the social network of school 34 at T_1 (Figure 5.8) and T_2 (Figure 5.9), red nodes indicating smokers, the evolution of the smoking uptake is stark. The images appear to visualise a defined divide in the school 34 social network, the segregation being navigated by a selection of individuals. High betweenness would indicate a large number of central individuals appearing in the shortest paths of other indirect connections; in terms of the visual representations, this would be those individuals navigating the network divide.

After a series of requests to DECIPHer to garner further contextual information relating to school 34, it was discovered this school is divided into two campuses - the buildings being separated by a train line. The observed separation of Figure 5.8 and T_2 Figure 5.9 may therefore be a visual representation of the physical distances between social ties, signifying the importance of proximity. Furthermore, at T_1, a number of those positioned at the centre of the graph in Figure 5.8 are smokers, their betweenness centrality potentially allowing them to diffuse a positive smoker message. This smoker centrality may account for the vast smoking increase observed at T_2.

From the intervention schools discussed, schools 34, 74 and 76 are process schools (as discussed in Section 5.2.1); as each of these schools exhibit varying reactions to the intervention, it would appear the evaluation process does not create a notable impact upon social network characteristics. The detailed examination of the selected intervention networks, in conjunction with school attributes, has highlighted possible explanations for the

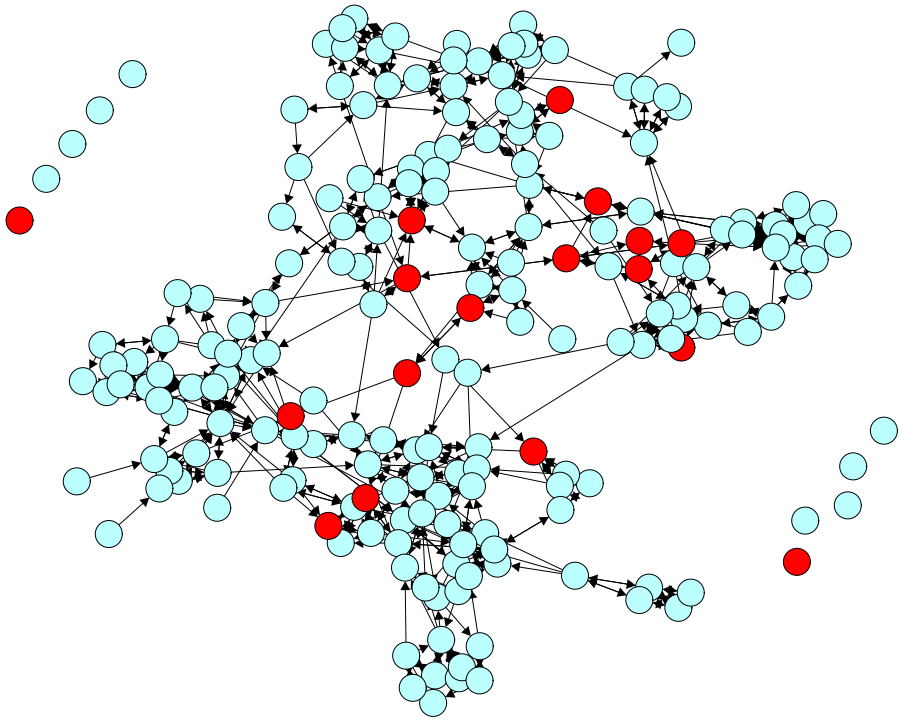


Figure 5.8: School 34 Social Network at T_1, red nodes indicate smokers.

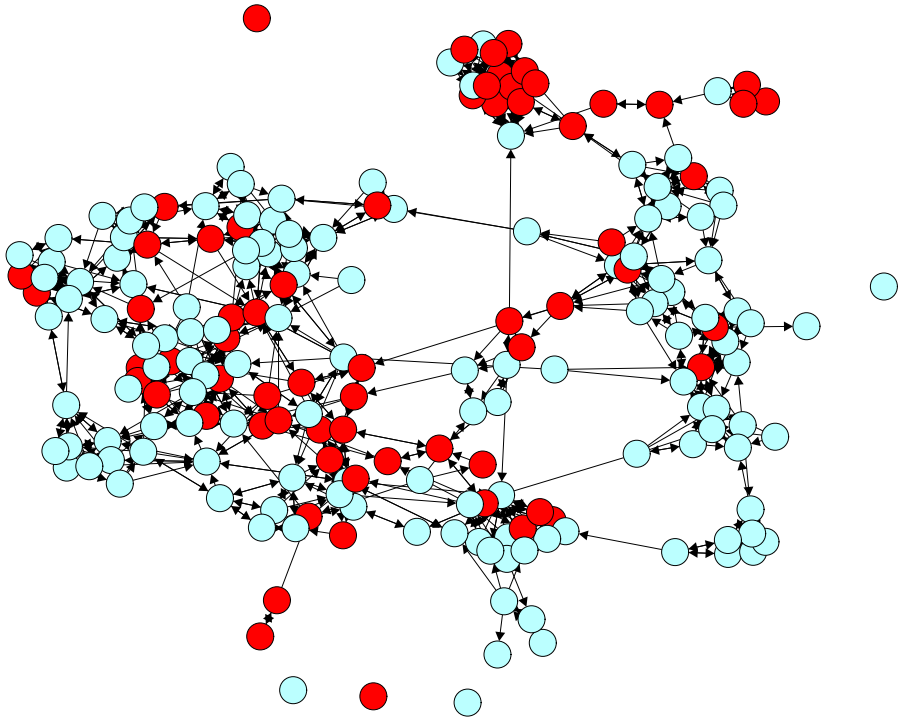


Figure 5.9: School 34 Social Network at T_2, red nodes indicate smokers.

observed intervention behaviour. Evidently there are a great number of factors that evoke the desire to smoke in adolescents, this section demonstrating the potential role of social networks. To analyse the network effects of smoking further, control schools will be examined in the following section (5.2.3.2) - providing an additional network account of smoking uptake.

5.2.3.2 Control School Discussion

This section is concerned with the natural diffusion of smoking in control schools, with particular attention focused upon the incumbent social networks. For brevity, a selection of the control schools will be explored further; the selection based upon particular outcomes of most interest to this thesis. Section 5.2.3.1 identified key areas related to SNA that may be of importance in the diffusion of a message, such as APL, closeness and betweenness; it is therefore of interest to explore smoking uptake diffusion uninhibited by the intervention.

The social network of intervention school 34 (Figure 5.9) indicated a proximity divide in friendships, however, proximity may not be the only cause of segregation when visualising the ASSIST school networks. Plotting the T_1 graph (Figure 5.10) of control school 35 indicates a social network divide based on gender, two distinct groups of male and female students visible. As the students get older (T_2, Figure 5.11), there appear to be more heterogeneous sex friendships - the divide no longer being apparent. The school 35 network statistics also imply a more cohesive group structure at T_2; path length shortens (T_1: 0.896, T_2: 0.952), while degree out (T_1: 4.576, T_2: 4.823), reciprocity (T_1: 0.617, T_2: 0.649) and density (T_1: 0.029, T_2: 0.031) all increase.

The betweenness of school 35 appears to decrease at T_2 (T_1: 0.095, T_2: 0.068), therefore students are no longer consistently appearing in the shortest paths of the network; this may be due to the sex divide decreasing, therefore paths from opposing sides of the networks no longer routing through specific central individuals. The reasons for this initial homophilous sex grouping at T_1, and why the groups become mixed at T_2, are unclear - a unique characteristic of the school being its religious teachings. The smoking uptake in school 35 between T_1 and T_2 (1.65%) is particularly low, a possible result of the divided structure of the social network; the visually central nodes (Figure 5.10) in the network being non-smokers, and therefore the positive smoking message not being diffused profusely across groups.

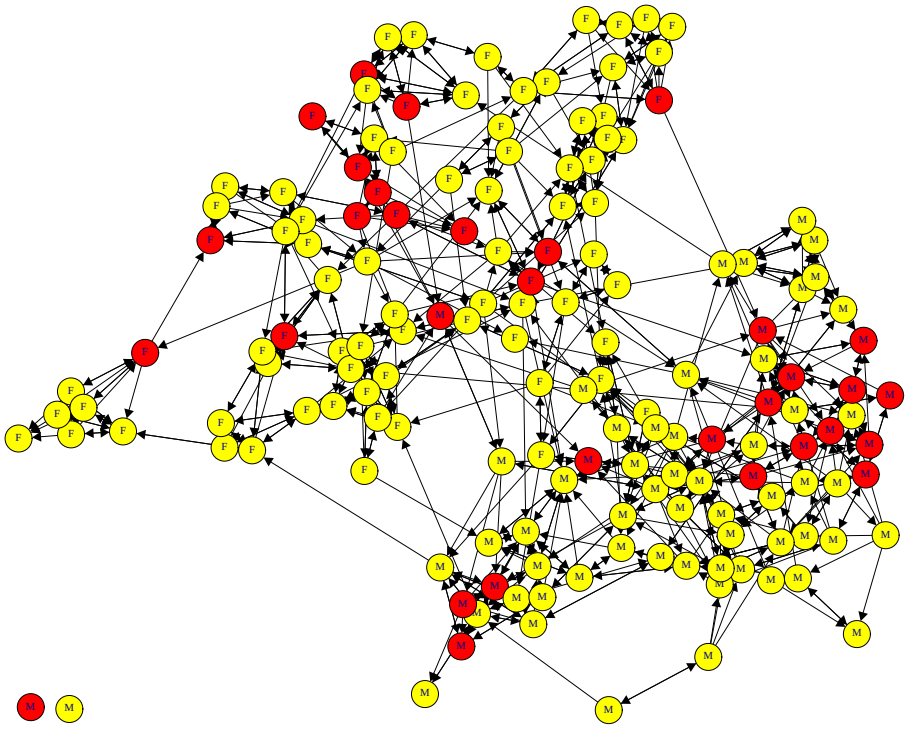


Figure 5.10: School 35 Social Network at T_1, red nodes indicate smokers. Node labels "M" represent male students, "F" identify female students.

Figure 5.11: School 35 Social Network at T_2, red nodes indicate smokers. Node labels "M" represent male students, "F" identify female students.

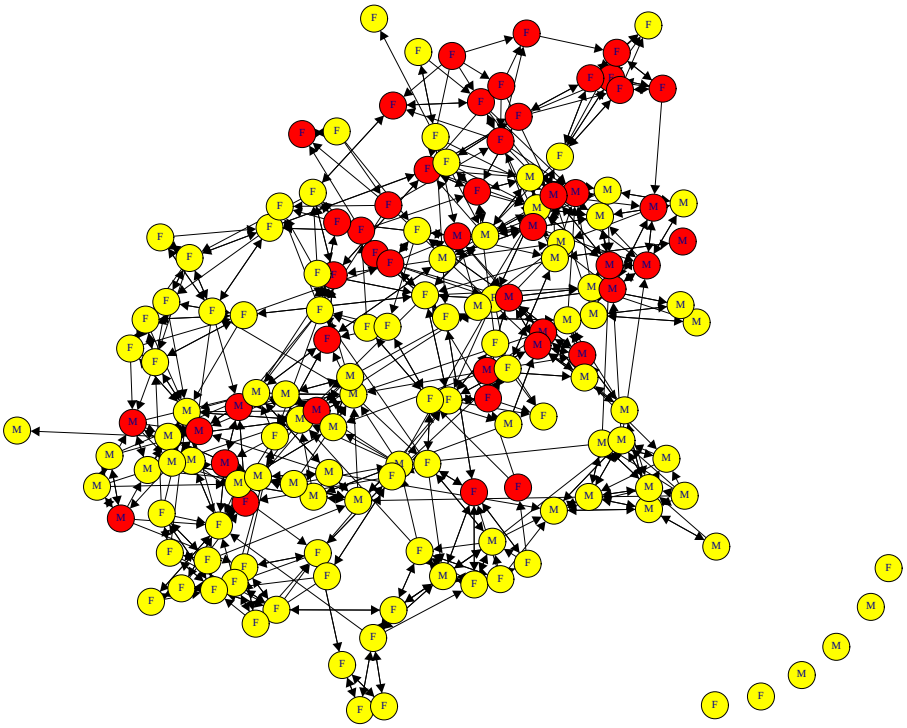


Figure 5.12: School 35 Social Network at T_3, red nodes indicate smokers. Node labels "M" represent male students, "F" identify female students.

The smoking uptake of school 35 increases between T_2 and T_3 (9.21%), following the unification of groups at T_2, the increased cohesion allowing smokers to group together and potentially recruit new smokers - a dominant smoker group being visible in Figure 5.12. School 35 initially experiences a surge in smokers between T_0 and T_1 (7.40%); unfortunately due to network data not being collected in this time period, the effect of betweenness and gender segregation between T_0 and T_1 is unclear. The importance of a cohesive structure in the uptake of smoking, however, is demonstrated effectively in school 40; this all female school experiencing a surge in smokers from 1.67% (T_0) to 23.21% (T_3) over the course of the study.

School 40 has particularly high levels of closeness (0.111), betweenness (0.214) and reciprocation (0.676) at T_1, the graph of Figure 5.13 representing the network as an amalgamation of sparsely interconnected cliques. The high betweenness may be due to paths between cliques having to traverse the same students repeatedly; for example, to connect nodes 17 and 18 in Figure 5.13 (highlighted in yellow), the geodesic (as defined in Section 3.1.4) must travel via nodes 10 and 20. Similarly, the geodesic from node 7 to 19 would

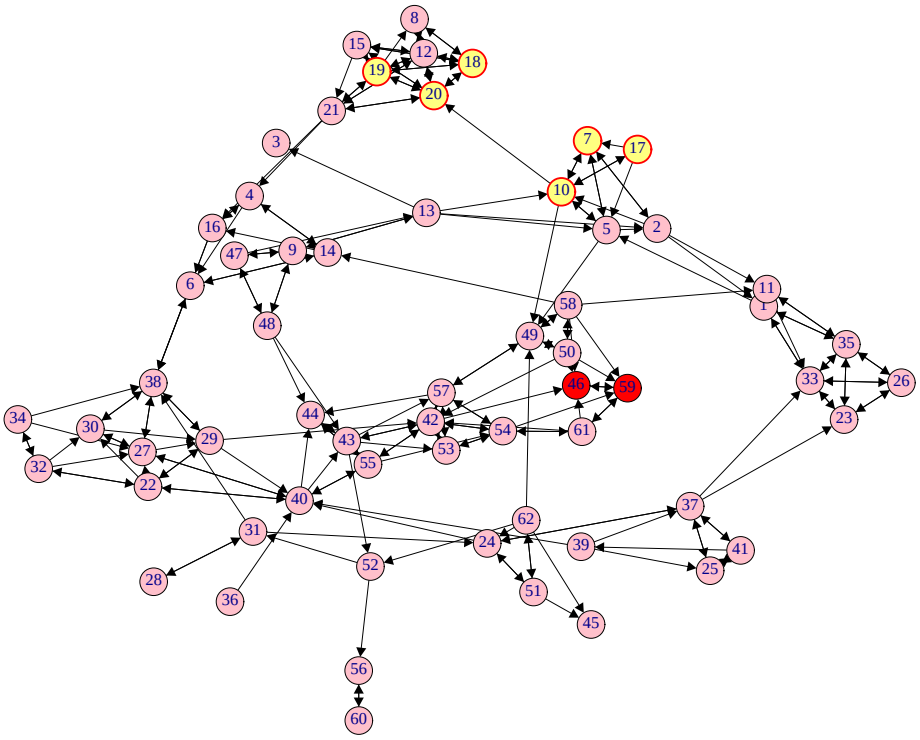


Figure 5.13: School 40 Social Network at T_1, red nodes indicate smokers. Node numerical labels represent a students school specific identification number, example nodes highlighted in yellow.

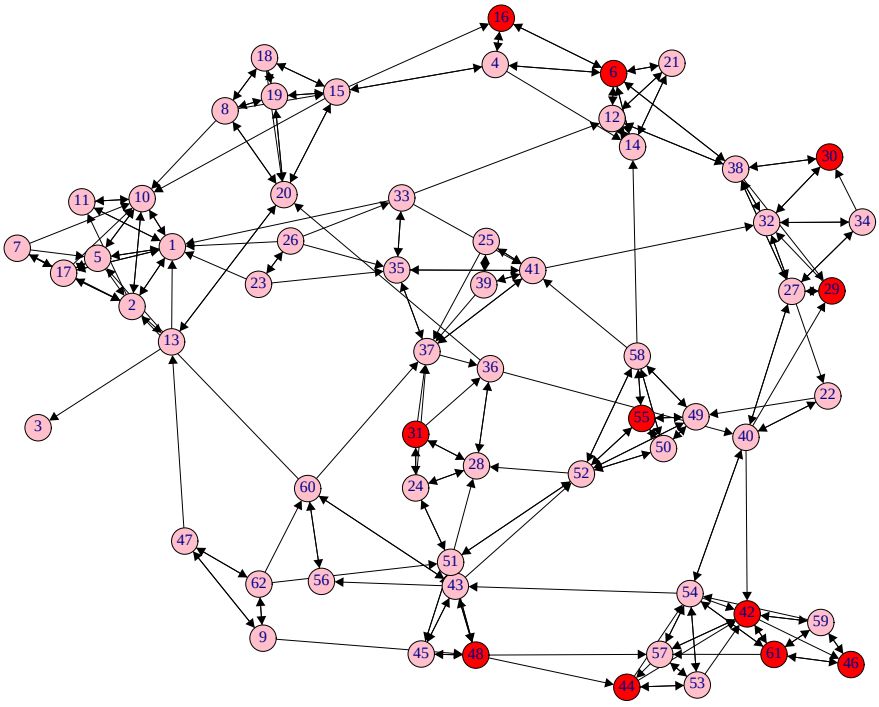


Figure 5.14: School 40 Social Network at T_2, red nodes indicate smokers. Node numerical labels represent a students school specific identification number.

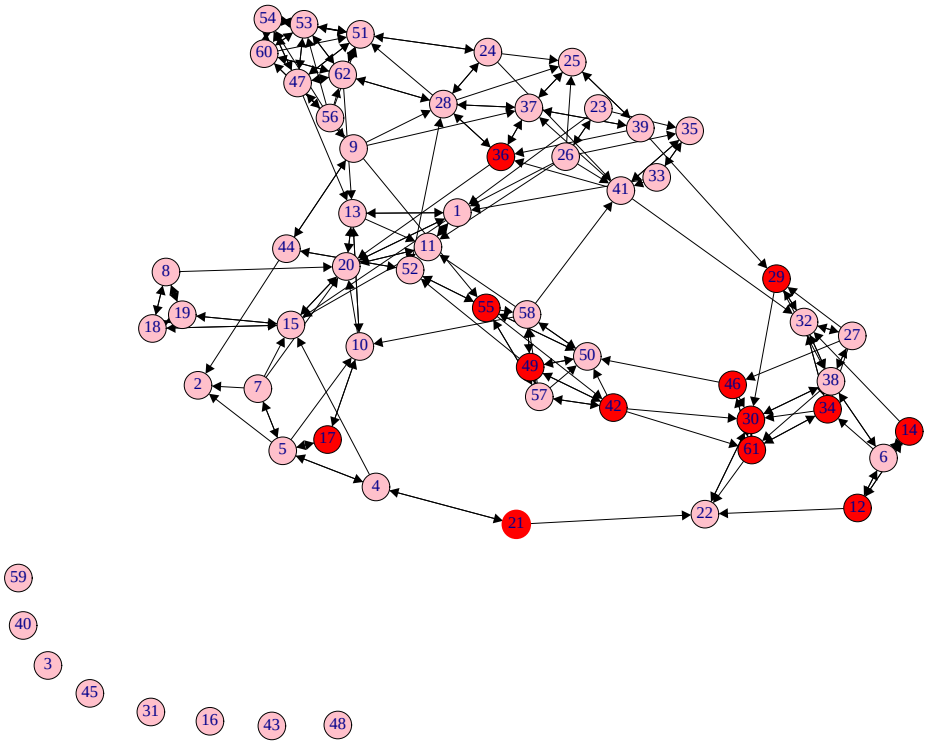


Figure 5.15: School 40 Social Network at T_3, red nodes indicate smokers. Node numerical labels represent a students school specific identification number.

also involve nodes 10 and 20; in fact, any number of geodesics originating in the clique encompassing node 10, to the clique of node 20, would involve nodes 10 and 20 in their paths - therefore the betweenness centrality of said nodes is high.

School 40 contains two smokers at T_1, students 59 and 46, this value increasing to eleven at T_2. Of the central clique surrounding nodes 59 and 46 at T_1, two new smokers are created at T_2 (42 and 61) - the clique evolving away from the centre of the network. Student 59 reports no smoking behaviours at T_2, however, the student retains three smoker connections. The network evolves further at T_3 (Figure 5.15), almost becoming divided into a densely populated smoker group and a non-smoker group. It would be expected that the betweenness of school 40 at T_3 would be high, due to the reliance upon nodes such as 21 and 29 for geodesic paths; this is not reflected in the network statistics of Table 5.4 (betweenness: 0.181), potentially due to the large number of isolated nodes.

It would appear the cliqued structure of School 40, in combination with well positioned smokers in the network, has facilitated an increased diffusion in the positive smoker mes-

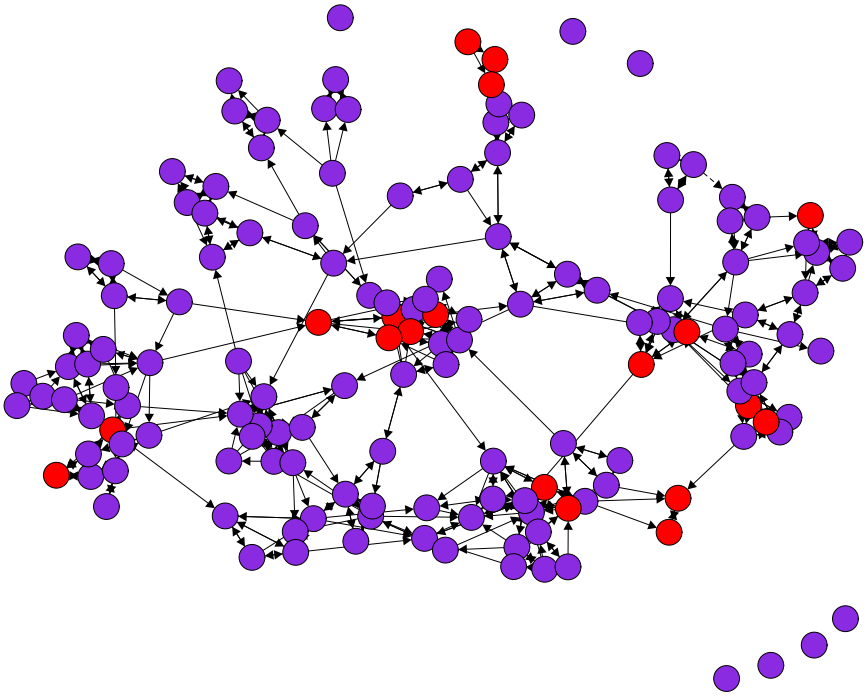


Figure 5.16: School 22 Social Network at T_1, red nodes indicate smokers.

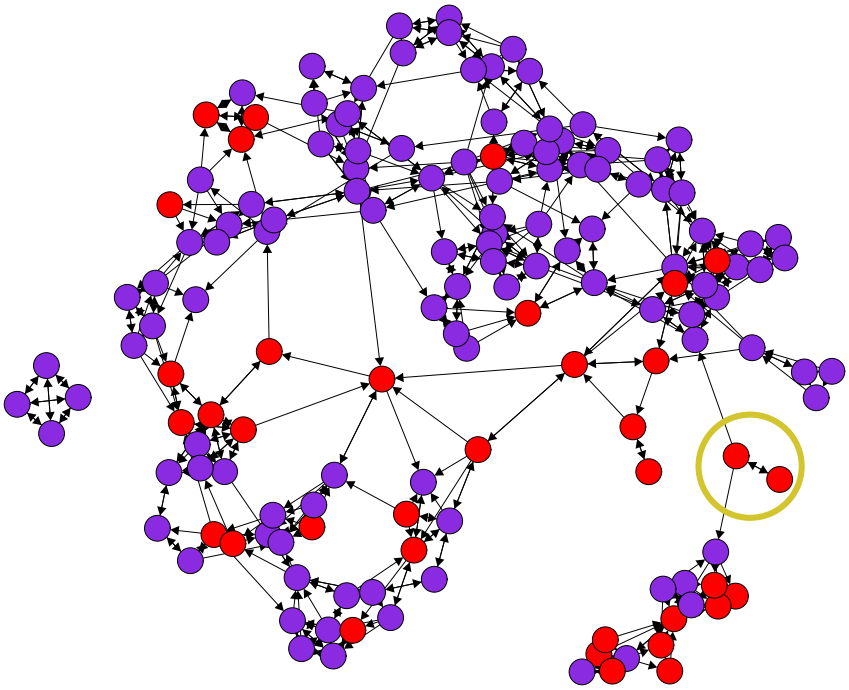


Figure 5.17: School 22 Social Network at T_2, red nodes indicate smokers. The highlighted yellow circle indicates a single node connecting a large group of vertices to the main body of the network.

sage; a similar structure also being observed in school 22. The school 22 network has a large smoking uptake rate across all timesteps ($T_0 - T_1$: 6.24%, $T_1 - T_2$: 11.04%, $T_2 - T_3$: 10.64%), the network also possessing particularly high levels of betweenness centrality (T_1: 0.254, T_2: 0.193, T_3: 0.166) - values being the largest of all schools in the data set at T_1. The network images of Figure 5.16 and 5.17 demonstrate the role of "between" students in joining the cliqued structures, the T_2 network depicting the instance of a single agent keeping a large clique of smokers connected to the main body of the network (bottom right of Figure 5.17, highlighted with a yellow circle). Figure 5.17 also indicates a large number of central nodes as smokers, once again providing evidence for the importance of smoker betweenness centrality.

While schools 22 and 40 suggest social network structures particularly conducive to smoker uptake, the network characteristics of school 41 and 69 may provide constructions for inhibiting smoker uptake - both schools maintaining a low overall smoker population. School 41 has a particularly long APL at T_1 (0.503) and T_3 (0.471), degree statistics (T_1: 3.721, T_3: 3.506) at said timesteps also being slightly below average - the networks plotted in Figures 5.18 and 5.19. A large number of isolated nodes may be the cause of the APL elongation of school 41 at T_3, however, this would not explain the long APL at T_1; furthermore, the T_3 graph has two completely disconnected cliques from the main body of the network.

Although the APL of school 41 at T_1 and T_3 is longer than average, that of T_2 (0.840) is shorter than average - the time period of $T_2 - T_3$ indicating the greatest smoking uptake in the school. The observed smoker increase may be caused by the shortening of paths between individuals, the positive smoking uptake message able to circulate more rapidly around the network; it would therefore appear that it is not only the position of smokers that is important, but also overall group cohesion. The school 69 network similarly appears to naturally inhibit smoking uptake, with APL being longer than average across time steps T_2 (0.735) and T_3 (0.413); however, this is not to the level indicated by school 41.

The analysis of school 69 demonstrates a reduction in smoking between T_1 and T_2 (-1.59%), subsequently increasing between T_2 and T_3 (11.34%). The network measures of Table 5.4 do not appear to offer any discernible features of the network, other than the marginally decreased APL previously discussed; as such, further requests for information from DECIPHer were made with regards to potential reasons for the low smoker uptake.

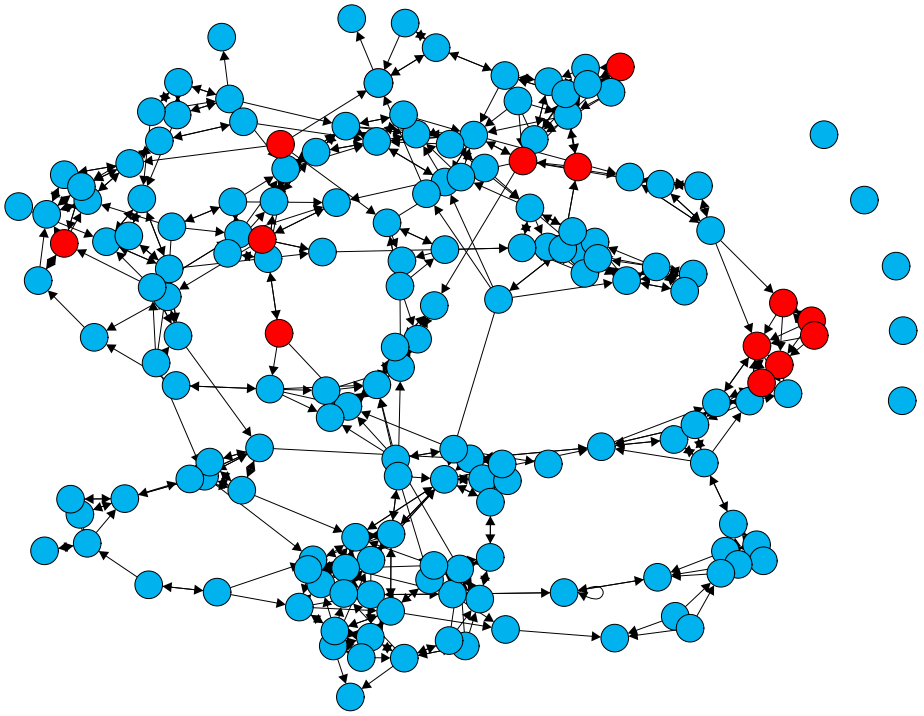


Figure 5.18: School 41 Social Network at T_1, red nodes indicate smokers.

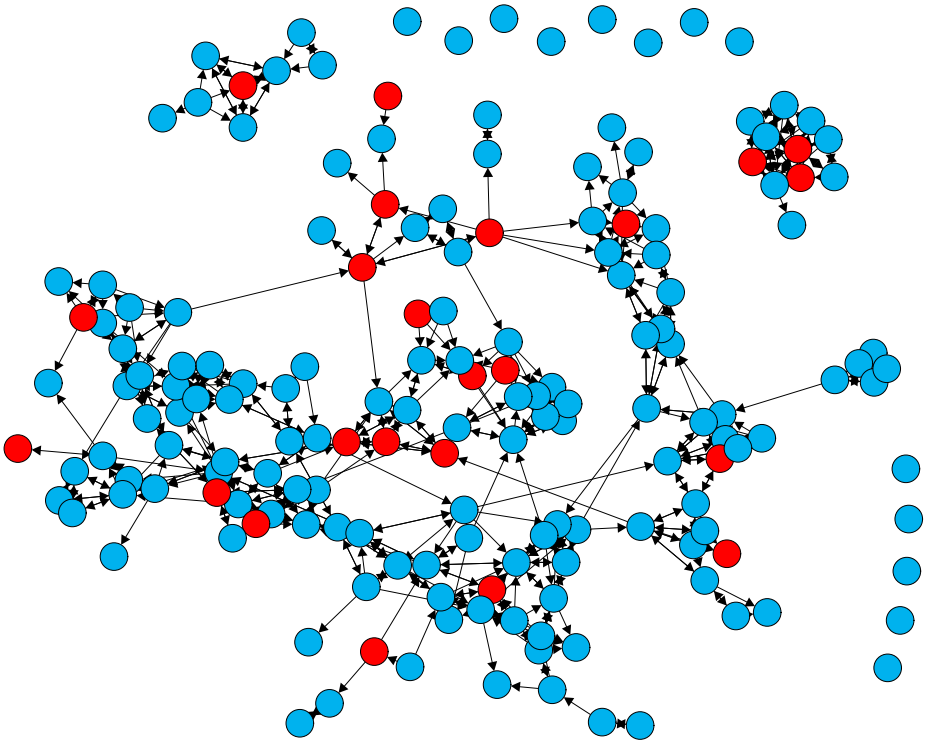


Figure 5.19: School 41 Social Network at T_3, red nodes indicate smokers.

Over the course of the study, School 69 conducted its own intervention procedures to deal with student behavioural issues - the exact details of which are unknown. This behavioural intervention, coupled with Welsh language environment, may be the cause of the reduced smoker uptake (as opposed to specific network structures), also demonstrating the importance of specific school characteristics.

School specific characteristics not only have the ability to affect smoking behaviours, but also network structure. Taking the example of school 68, the degree (out and total) network characteristics are the lowest of all schools (T_1: 2.378, T_2: 3.354, T_3: 2.707) - suggesting a particularly "unfriendly" school, whereby students are frugal in their extension of outward links. While it is evident from Figure 5.20 that the low average degree at T_1 is caused by the large amount of isolated nodes (missing data), the number of disconnected nodes decreases substantially at T_2 (Figure 5.21) yet degree statistics remain the lowest of all schools. This may be due to School 68 being situated in inner city Cardiff, therefore the large catchment area of the school, along with the possibility of a more diverse population of students, potentially accounting for this reduced degree observation.

Moreover, due to the low out-degree and lower than average reciprocation across timesteps (T_1: 0.549, T_2: 0.498, T_3: 0.545), School 68 may create more hierarchical friendships - whereby a number of students extend links to specific individuals in the network, these connections not being reciprocated. Such a structure creates individuals of influence in the network, as those with an unreciprocated tie seek the approval of said influential person - therefore potentially change their behaviour to receive a friendship connection in return (hierarchical friendship influence discussed in Christakis & Fowler (2010b)). This hierarchy may be the cause of the large smoking uptake in school 68, in the absence of a low APL and overall network cohesion; rather than a smoking message circulating, individuals may be directly influenced to smoke by their outward connections.

A further network of interest is that of School 15, which has a high smoker uptake between T_1 and T_2 (12.34%) which reduces substantially at T_2 to T_3 (0.26%) - the school attributes of Table 5.2 unable to identify specific causes for this behaviour. The network statistics at T_1 (Table 5.4) do not appear to embody the characteristics of structures conducive to smoker uptake, however, the network experiences a drop in degree centrality at T_2 (0.037) from T_1 (0.056). A student with high degree centrality possesses a large number of connections, suggesting that at T_1 there are a greater number of students (on average) with

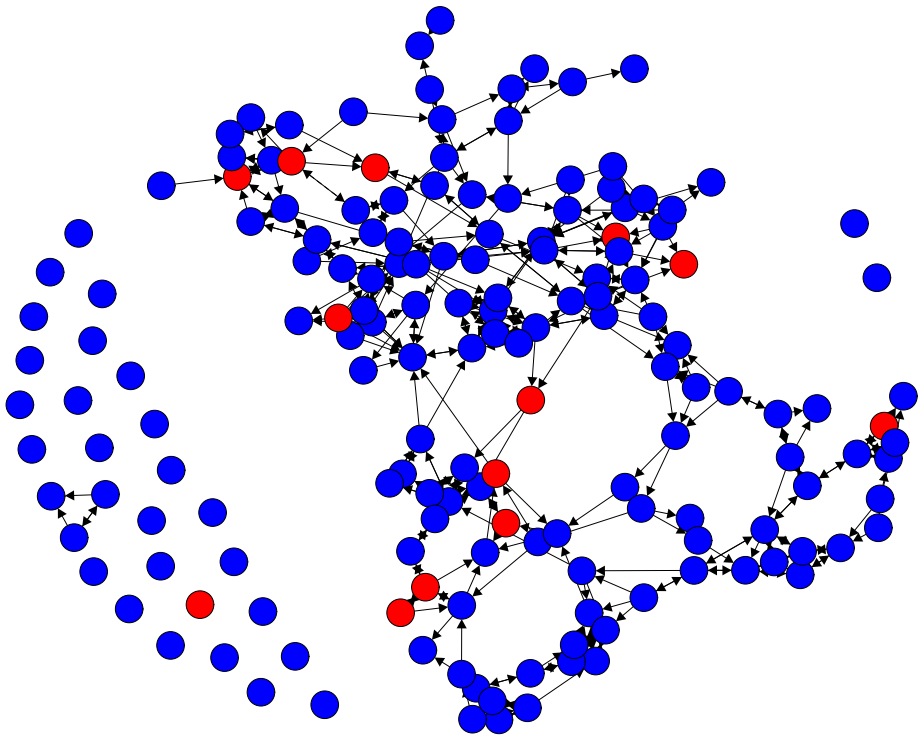


Figure 5.20: School 68 Social Network at T_1, red nodes indicate smokers.

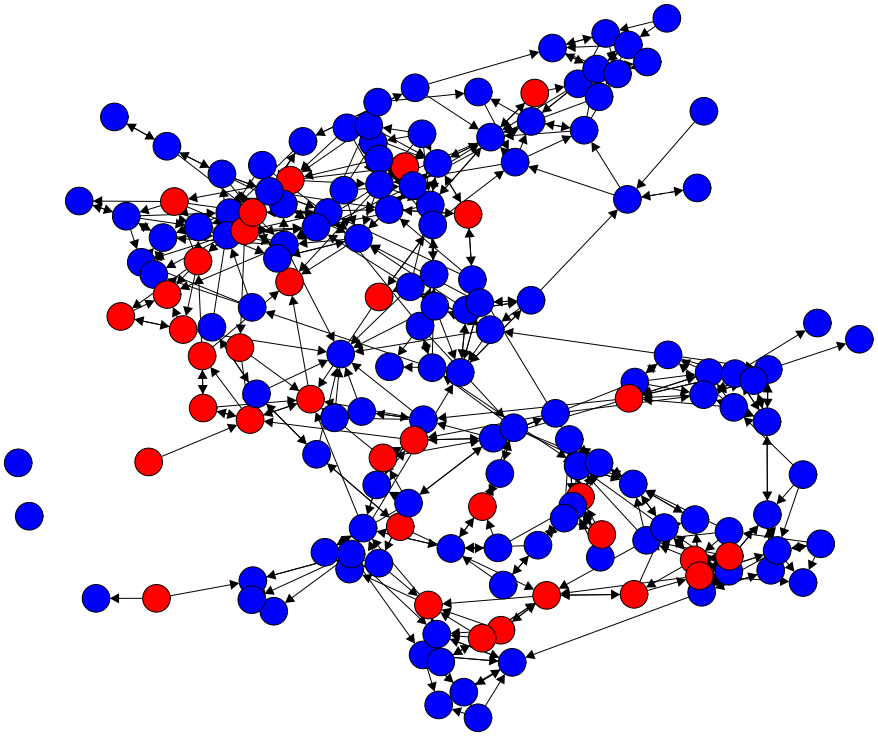


Figure 5.21: School 68 Social Network at T_2, red nodes indicate smokers.

larger degrees than at T_2. The smoker increase observed between T_1 and T_2 at school 15 may therefore be attributed to smokers having more connections, subsequently distributing a positive smoker message to a greater audience; once this decreases at T_2, so does the ability to spread the message.

While the overall process of smoking uptake in adolescents is still unclear, the network characteristics discussed offer potential factors in decisions related to smoking. The overall analysis of control schools appear to indicate two key network attributes, APL and centrality measures. It is not necessarily the structure of the network that is important in affecting smoker uptake, but rather the position of the smokers; this indicated by the differing importance of betweenness, closeness and degree centrality in smoker uptake across the investigated networks. These notions appear consistent with the network discussion of intervention schools (Section 5.2.3.1), which has also identified the importance of APL and centrality in the message diffusion by peer supporters - an overall discussion of the findings provided in section 5.2.4.

5.2.4 Network Conclusions

The analysis of network schools offer a variety of potential justifications for the behaviour exhibited over the course of ASSIST. The intervention appears to produce altered values of closeness centrality to those of control schools, indicating some differences in the friendship behaviours of students. It would seem that drawing attention to particularly popular and influential students, may actually create a negative reaction towards said individuals - hindering the intervention process. Schools in which peer supporters remain central to the network, bestowed with overall positive network cohesion, may create a greater number of opportunities for supporters to exact their roles; this leads to a successful reduction in smoking uptake.

Control schools also provide a great deal to the discussion, as particular networks appear to naturally generate fewer smokers. Once again, the position of those distributing a message is highlighted, smokers holding central positions in the network have the ability to distribute a positive smoker message further. It is suggested that those networks appearing particularly cliqued, may be more susceptible to this identified smoker diffusion - provided that the smokers appear to adopt influential network roles. Smokers embedded in intermediary positions in the network, navigating a segregated divide, also appear important -

highlighting the role of betweenness.

It must also be noted that network cohesion need not be a factor in successful smoking diffusion, providing that enough individuals view the smoker as possessing a position of influence; therefore hierarchical structures also appear important to the social influence of smoking uptake. A further analysis of the position of smokers in a social network, and their resultant impact to smoking uptake, is conducted in Chapter 9 - making use of the social network evolution algorithm developed in Chapter 6 and refined in Chapter 8.

Overall, it must be noted that the school network only provides one facet of an individual's life, and therefore their predisposition to smoke. While factors such as FAS and parental smoking have attempted to be explored, other more complex factors may not be captured by the data provided. For example, the friendship networks of students outside of school are also said to be influential to their behaviour (DuBois & Hirsch, 1990; Kiesner et al., 2004), with Figure 5.22 demonstrating the connections between adolescents that exist across schools. The behavioural norms present in other schools may transfer through connected agents, affecting individual decisions regarding smoking. Such cross-school transference is not considered in this thesis, but is an avenue for potential future research (as discussed in Chapter 10).

Furthermore, it is also unclear whether the social networks are causing the smoker uptake observed, or rather the adolescent smoking behaviour is dictating the network structure; however many observations appear consistent with discussions amongst the relevant literature. While a social network may not provide a full account of smoking behaviours, this section has identified potential areas of importance:

- The position of smokers in the network;
- The position of nominated peer supporters in intervention schools;
- The segregation of the network, and the subsequent smoking behaviours of central individuals;
- The cliquedness of the network, and the resultant ability to spread "messages".

These outcomes provide greater insight into the composition of ASSIST social network

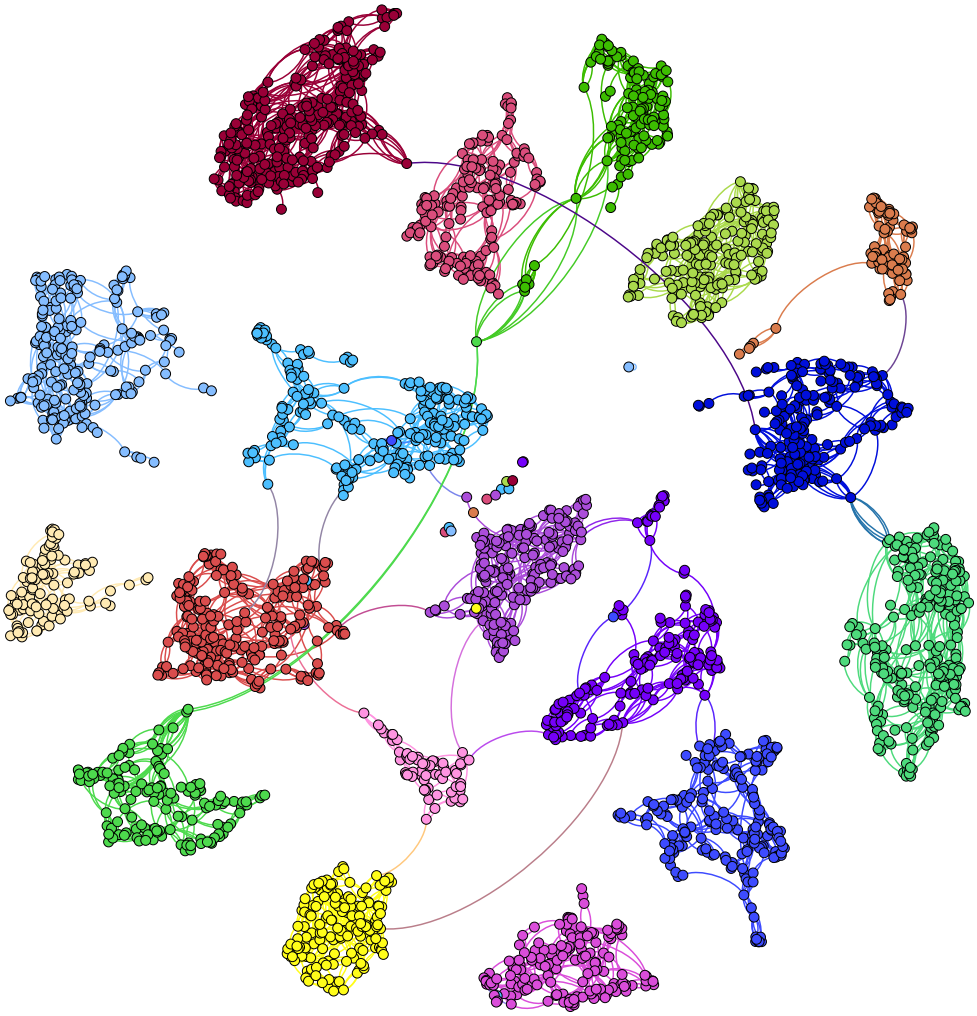


Figure 5.22: Social network of all available ASSIST schools at T_2, indicating links between schools. The colour of each school is consistent with previous figures.

structures, identifying potentially important network aspects of behavioural diffusion. To further investigate the behaviours of adolescents, Section 5.3 explores the overall trial results of all available ASSIST school data.

5.3 Smoking Data Analysis

This section focuses upon the smoking attribute data provided by DECIPHer, with information from all 59 schools in the cohort being present. Section 5.2.3 discussed smoking in the context of a social network, but a statistical analysis of smoking behaviour is also of interest. This section explores the uptake of smoking, with discussion also centring around the effectiveness of ASSIST. A more extensive analysis of smoking uptake, in relation to social network structure, is conducted in Chapter 9.

This section is structured in the following manner: to begin, a simple analysis of smoker proportions in control and intervention schools is conducted in Section 5.3.1; Section 5.3.2 examines the proportions of smokers, assessing the differences across time periods; and Section 5.3.3 draws together the conclusions of the smoker analysis - the aim being to gain an understanding of adolescent behaviours, that shall inform the proposed investigation of this thesis.

5.3.1 Smoker Proportions

For the following analysis, the 11,545 individual student records are classified into control and intervention groups. The overall smoker proportions at each time step for each school are calculated, with missing data removed from the relevant observation period. Table 5.7 displays the mean proportion of smokers and non-smoker categorised by time period and school type. Following the Kolomogorv-Smirnov test for normality, the Mann-Whitney test of two independent samples was conducted in SPSS (IBM, 2011) - the P-Values displayed in Table 5.7.

It is observed that a statistically significant difference is present at T_0, T_1 and T_3, with a larger proportion of control smokers being present. The differences at T_1 and T_3 may suggest, due to a smaller number of smokers in intervention schools, the successful application of intervention methods; however, such conclusions are not appropriate with the

	Intervention (%)	Control (%)	P-Value
T_0	8.55	10.83	<0.001
T_1	10.48	13.19	<0.001
T_2	19.20	20.74	0.059
T_3	26.16	28.65	0.006

Table 5.7: Percentage of smokers in intervention and control school, along with the P-Value of the Mann-Whitney tests comparing smokers by school type.

aforementioned figures, due to a statistical difference being present at T_0. The statistical tests are therefore unable to differentiate between the success of the intervention, with a naturally occurring reduced smoker population.

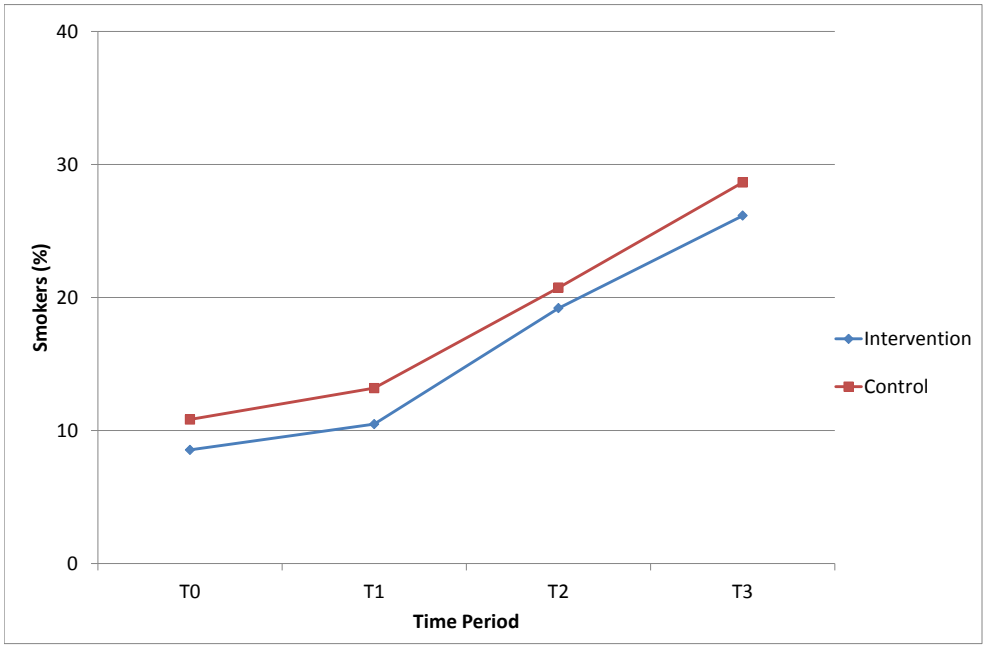


Figure 5.23: Graph depicting smoker proportions over time by school type.

To visually represent the increase in smokers, Figure 5.23 depicts the smoker proportions over time. A particularly interesting observation occurs at T_2, whereby a statistically significant difference is not present - due to a large increase in intervention school smokers. A number of reasons may cause the observed T_2 increase (said increase also documented within Campbell et al. (2008)), a viable possibility being an attenuation of the intervention

over time. As previously discussed (Section 5.2.3.1), peer supporters may lose interest in their roles over time, or may no longer be in an effective network position to exact the necessary diffusion; as such, the smoking cessation message may diminish, causing an overall smoker increase at T_2. Furthermore, the large increase in figures at T_2 may also indicate a negative intervention effect, students in intervention schools rebelling against the smoking cessation message - said rebellion abating by T_3.

The figures presented, also demonstrate alternate smoking observations to those of the 18 'network schools' discussed in Section 5.2. Statistically significant differences, between control and intervention school smoker proportions, were not found in network school data - opposing the notions presented in table 5.7. The observed deviations would suggest differing 'network data' smoking dynamics to those of the full ASSIST cohort, casting doubt upon the generalisability of the 18 network schools provided; however, given that the smoker proportions were analysed in combination with network structure, the findings of Section 5.2 still remain relevant - albeit further insights potentially gained if a larger selection of network data were available.

From the statistical tests conducted, the positive effect of the intervention is unclear. While a reduced smoker proportion in intervention schools in exhibited at T_1 and T_3, this is negated by the prior differing smoker proportions at T_0. The results serve to demonstrate a diminished effect of the intervention over time, consistent with previous findings in literature and Section 5.2, a potentially negative effect also being observed. It must also be noted that as the students mature, they arrive closer to the legal age of cigarette purchase (16 at the time of the study); therefore an increase in the number of smokers over time may be expected. Additionally, adolescence is reported to be a time of experimentation with smoking (Nichter et al., 1997), this may also be crucial in the smoking uptake observed. To investigate smoking uptake behaviour further, a statistical analysis of the difference in smoker proportions across time steps is conducted in the following section (5.3.2).

5.3.2 Smoker Difference Over Time

Section 5.3.1 indicated differences in control and intervention schools with regard to smoker proportions, however, the effects of the intervention remaining unclear. This section assesses the *differences* in smoker proportions between timesteps for control and intervention schools, avoiding the uncertainty caused by a larger smoker proportion being observed in

control schools at baseline (T_0). The differences in smokers for each school between T_0 and T_1, T_1 and T_2, T_2 and T_3, and T_0 and T_3 are calculated, categorised by school type. As normality conditions are satisfied, following the Kolomogorv-Smirnov test, a one sample t-test is conducted; this test assesses whether the smoking difference between time periods (for all schools) is significantly different from zero, the results displayed in Table 5.8.

	P-Value (t-test)	Mean	95% CI Lower	95% CI Upper	Odds Ratio
$T_0 - T_1$	<0.001	.027	.016	.038	29.89
$T_1 - T_2$	<0.001	.078	.063	.092	18.99
$T_2 - T_3$	<0.001	.072	.059	.085	24.84
$T_0 - T_3$	<0.001	.176	.161	.192	11.39

Table 5.8: One sample t-test of the proportional smoker difference between time periods displaying the P-Value, the mean difference and the 95% confidence interval. The odds of being a non-smoker and remaining a non-smoker are also included.

The null hypothesis ($H_0 : \mu = 0$) for a one sample t-test is rejected if the mean smoker uptake across a time period is not equal to zero ($\mu \neq 0$), where μ is the mean difference between the percentage of smokers between two consecutive timesteps. The results of Table 5.8 indicate a rejection of H_0 with 95% confidence for each timestep; this suggests a significant smoker increase between timesteps for all schools.

To further investigate the smoker uptake, the odds of being a non-smoker and remaining a non-smoker in the subsequent time step are calculated; these values also displayed in Table 5.8. Over the six-month period of T_0 to T_1, the odds of being a non-smoker at T_0 and staying a non-smoker at T_1, are 29.89 times greater than being a non-smoker at T_0 and becoming a smoker. The odds decrease from said initial value across all time periods, indicating individuals are less likely to remain non-smokers as time continues; the odds being particularly reduced between the one-year period of T_1 and T_2, this being the interval in which interventions schools observe a large smoker increase (visible in Figure 5.23).

To assess the effect of the intervention, an independent samples t-test is conducted to compare the smoking uptake rate of control and intervention schools (across time periods) - the results of which are displayed in Table 5.9. From the results of Table 5.9, no significant difference is reported in smoking uptake across control and intervention schools at any time period; this would suggest no significant impact on the smoking rate of students due to intervention methods, the findings consistent with the 'network data' analysis of Section

5.2.2.

	P-Value	Difference	95% CI		Odds Ratio		
			Lower	Upper	Control	Intervention	P-Value
$T_0 - T_1$	0.336	-0.011	-0.032	0.011	27.59	30.06	0.018
$T_1 - T_2$	0.616	0.007	-0.022	0.036	18.00	20.22	0.028
$T_2 - T_3$	0.708	0.005	-0.021	0.031	26.36	23.51	0.025
$T_0 - T_3$	0.916	0.002	-0.029	0.032	13.18	9.74	0.088

Table 5.9: Independent samples t-test comparing the difference in smoker uptake for control and intervention schools, with associated P-Value and confidence intervals. Also reported are the odds ratios of being a non-smoker and remaining a non-smoker in the subsequent time period, for both control and intervention schools.

Table 5.9 also displays the odds ratio of remaining a non-smoker, should an individual be a non-smoker in the proceeding time period, compared with being a smoker. The odds of remaining a non smoker in intervention schools is initially significantly higher at T_0 to T_1, and T_1 to T_2 than control schools, however the odds decreasing significantly below that of control schools at T_2 to T_3 and across the whole time period T_0 to T_3; this overall time odds ratio not being significant. The odds calculations take raw counts of data, as opposed to proportional differences in schools, hence providing a differing perspective of smoker uptake. The demonstrated odds would once again suggest a reduction in the intervention over time, but also may indicate some smoker "rebound" occurring in the time period between T_2 and T_3 - a potentially negative intervention effect occurring over time.

A further attribute of interest is that of the gender effects upon smoking uptake, as gender also appears to be an important factor in the friendship selections of adolescents - as discussed in Section 5.2.3.2. To assess gender impact, the segregation of control and intervention schools is removed, the data being recategorised by gender. The odds of being a non-smoker at a given time step and remaining a non-smoker in the next, in comparison to becoming a smoker, for male and female students is presented in Table 5.10. The odds ratios are consistently higher for male adolescents, although not significantly at T_1 to T_2 at the 0.05 level; this would indicate female adolescents have a higher odds of becoming smokers in the periods T_0 to T_1 , T_2 to T_3 and T_0 to T_3, thus indicating some gender differences. This is further qualified by the raw percentages of male and female smokers at each timestep, presented in Table 5.11.

Overall, taking into consideration the smoking uptake rate over time, significant interven-

	Male	Female	P-Value
$T_0 - T_1$	33.433	26.680	0.047
$T_1 - T_2$	25.508	14.718	0.131
$T_2 - T_3$	26.233	21.821	0.041
$T_0 - T_3$	11.814	10.937	0.022

Table 5.10: The odds ratio of being a non-smoker and remaining a non-smoker by gender, the associated P-Value of the comparison of differences also reported.

	Male Smoke (%)	Female Smoke (%)
T_0	3.18	3.17
T_1	4.26	4.55
T_2	7.72	11.30
T_3	11.91	18.98

Table 5.11: Raw smoker percentages by gender at each timestep.

tion effects are not apparent. The one sample t-test concludes a significant increase in the proportion of smokers in all schools between time points; however, no significant difference is apparent between the smoking uptake rates of control and intervention schools. Using the raw counts of smokers and non-smokers in the data at each time period, a significant difference in the odds ratio of being a non-smoker and remaining a non-smoker (at a subsequent time period) is observed; the odds initially being higher for intervention schools, but reducing below control schools at T_2 to T_3 - an overall significant difference in odds not being observed between T_0 and T_3. Finally, gender also appears to be an important factor in the odds of remaining a non-smoker; females indicating reduced odds, although not significantly between T_1 and T_2.

5.3.3 Smoking Data Conclusions

The statistical analysis of Section 5.3 has provided further insights into the ASSIST data, and the individuals documented therein. Section 5.3.1 demonstrated that simply examining the proportions of smokers in schools at each time period by intervention type, does not provide a representative account of intervention smoking differences - the results indicating a significant difference being present at baseline (T_0). Nevertheless, said results highlighted a potential period in which the smoking behaviours of intervention schools increase, such that a non significant difference is recorded (T_2) - suggesting a period of intervention reversal.

The conclusions of Section 5.3.1 prompted further investigation into the smoking behaviours of ASSIST schools, making use of the difference in smoker proportions across time steps - a more representative measure of comparison. The results of Section 5.3.2 find no significant difference in the rate of smoking uptake between control and intervention schools, the odds of remaining a non-smoker also not being significant over the full course of the intervention ($T_0 - T_3$). The analysis also indicates a significant difference in the odds of remaining a non-smoker by gender, female adolescents in ASSIST possessing significantly reduced odds over the course of the collected data T_0 to T_3.

Drawing together the conclusions of this analysis, it would appear a quantitative overall reduction in smokers (due to intervention) is not present; however, these results must be taken in context and while a high-level reduction may not be present, specific individuals may have benefited from the trial - a measure which is argued unquantifiable by such analyses. The conclusions presented have documented an account of smoker uptake in adolescents, along with time and individual characteristics that may be of key importance; said conclusions have value in gaining a greater understanding of adolescent behaviour, contributing to the aims discussed in the following section (5.4).

5.4 Informing Future Analysis

The selection of ASSIST data provided by DECIPHer has offered insights into the real-world structure of social networks, and their potential for influence upon individual behaviours. The conclusions presented suggest that each school offers its own unique perspective of social connection, smoking uptake and the ASSIST intervention as a whole - a product of the individuality of each student comprising the cohort.

Of particular interest in this thesis, is the evolution of the social networks and smoking behaviours over time, which appears to vary greatly between the schools investigated. While the analysis conducted in this chapter has suggested factors important to the evolution of the ASSIST school social systems, it would be prudent to test said factors' relevance in explaining system development. Moving forward, this thesis aims to further analyse social structure and behaviour, by creating a framework to predict their evolution over time - incorporating the insights gained from previous chapters.

Chapter 2 highlighted ABS as a simulation technique to explore the effect of individual

actions upon a system as a whole; this lending itself well for use with the ASSIST data, given the suggested effect of an individual's smoking behaviour upon the school network as a whole. Furthermore, Chapter 3 identified the importance of social networks upon behaviour, outlining a well-defined sector of literature concerning the prediction of links in a network (Link Prediction). Therefore, the ASSIST data shall be further analysed by creating simulations of adolescent connection and smoking behaviour, using LP methods to evolve the social network structures over time.

Chapter 4 identified the importance of social structure in the behavioural influence of a social network upon its members. As such, this research shall first aim to effectively represent the evolution of social connectivity, developing an algorithm informed by prior LP literature and insights gained from the analysis conduced in Section 5.2; this research is presented in Chapter 6. For this analysis, the three waves of ASSIST social network data (T_1, T_2 & T_3) shall be utilised - assessing the accuracy of the predicted network against the real network data.

For comparison purposes, the accuracy of the newly developed algorithm shall be tested against four prominent existing LP methods - each method encompassing specific factors said to be of importance in friendship selection. This assessment shall not only provide a benchmark for the accuracy of the newly developed method, but also provide an indication of the importance of the specified factors upon friendship selection. This analysis, and the conclusions drawn, are presented in Chapter 7.

On completion of the creation of a link prediction ABS, and assessment of the social network factors providing the most accurate predictions, attributional data may also be incorporated; this would include individual behavioural factors - such as sex, smoking behaviours and proximity - extraneous to those of social networks, in an attempt to understand the role of behaviour upon friendship selection. The intended outcome of such procedures is to provide a greater understanding of the ASSIST data and improve the link predictions made. Additionally, insight into adolescent social network behaviours and the effectiveness of the ASSIST intervention methods may also be gained - with a view to inform future intervention processes. The augmentation of the simulation to incorporate behavioural information is discussed in Chapter 8.

Finally, the algorithm developed in Chapter 6 and refined (to include behavioural factors)

in Chapter 8, shall be used as a method to assess the interplay between social network structure and smoking. The insights gained from this model may provide further illumination to the role of social networks upon individual behaviours, specifically focusing upon smoking and its spread through a network. Furthermore, the model created shall be discussed in conjunction with alternative models of behavioural spread, providing an additional perspective to that of ABS.

A detailed understanding of the methods by which adolescent connections evolve, along with the impact upon smoking behaviour, may have great implications to the manner in which interventions such as ASSIST are conducted. The outcomes of ASSIST are unclear, the potentially successful intervention procedures in the six-month period between T_0 to T_1 appear to reverse somewhat in later time periods. The proposed methods may provide insight into the cause of the observed intervention dynamic, the outcomes informing researchers of the factors that may obscure the diffusion of the intervention message. Selection of the nominated peer supporters is evidently crucial to the successful conduct of the intervention, said individuals being selected in the initial stages of the trial; however, through the use of LP methods, it may be possible to identify peer supporters who are important throughout the evolution of the system - potentially improving intervention outcomes.

5.5 Chapter Summary

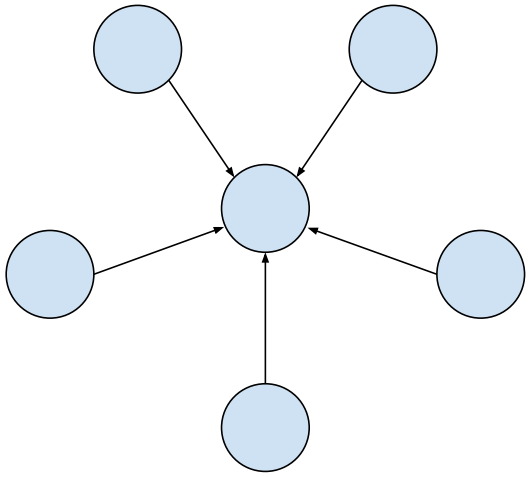
This chapter has investigated the data provided by DECIPHer, with the insights gained being used to inform the research conducted in the remainder of this thesis. Section 5.1 presented the background of the data source (ASSIST), discussing topics related to the intervention procedures, data collection and the social context of the study. Prior analyses of ASSIST were also explored, demonstrating its validity as a dataset from which to gain insights related to social networks and adolescent behaviours.

Section 5.2 explored the social network data available for analysis, 18 of the 59 ASSIST school social networks being provided. The social network data provided insights into the differences in control and intervention schools, the analysis reinforcing many of the findings of previous analyses conducted upon the data; such as, the importance of close communities (valley schools) in intervention diffusion, and an attenuation of the interven-

tion over time. Furthermore, the analysis documented a number of network measures that may be of importance, not only to the diffusion of the intervention, but of message diffusion as a whole; these metrics being average path length, degree statistics and various measures of centrality. The main conclusion of Section 5.2 highlighted that it is not necessarily the structure of the network that is important, but rather the position of influential individuals within it - a key concept which is taken forward in Chapter 6.

Section 5.3 analysed the attributional data available for all 59 ASSIST schools, assessing smoking behaviours across the whole cohort. The attributional analysis investigated the difference in proportions of control and intervention smokers at each time point, identifying a significant difference at baseline (T_0). As such, the proportions of smoker uptake within schools were assessed over time - the analysis concluding no significant differences in smoker proportions between control and intervention schools. On examination of the odds of remaining a non-smoker over time, results demonstrated a decrease as the students matured - indicating the time period/age in which smoker uptake occurs, may also be of importance. Furthermore, significant gender differences were found in the analysis of smoking uptake over the course of the study, highlighting gender as a potentially important factor in adolescent decisions to smoke.

Section 5.4 brought together the conclusions of prior sections and chapters, providing an outline of the subsequent research - the creation of an agent based simulation to predict social network evolution and investigate the role of connections in smoking uptake. Section 5.4 also highlighted the opportunity provided by the ASSIST data to conduct further exploratory analysis into: the outcomes of intervention procedures; adolescent smoking behaviours; and the factors important in the evolution of adolescent friendship. The proposed research direction outlined in Section 5.4, shall be further explored in the following chapters - attempting to satisfy the aims outlined in Chapter 1.



-"The Star Graph"

6

Link Prediction

This chapter discusses the development of a new algorithm to predict links in a social network, termed PageRank-Max, providing details of the simulation framework created for the investigation of social network evolution. Chapter 5, through the analysis of the ASSIST data, discussed the role of peer networks upon the behaviours of adolescents - identifying the location of an individual within a social network to be of potential importance. Chapter 5 also highlighted the need for analysis of the processes by which adolescent friendships evolve, should an understanding of peer influence be sought. Therefore, this chapter focuses primarily upon the development of an approach to effectively predict social network evolution, drawing upon methods from Link Prediction (LP) literature.

Prior to discussing PageRank-Max, existing LP methods, and their previous usage within the literature, are introduced in Section 6.1; this builds upon the review presented in Section 3.4. The transference of the selected existing LP methods into a simulation-based framework is discussed in Section 6.2, with the new LP algorithm developed specifically for this research being introduced in Section 6.3. Further operational details of the simu-

lation and validation procedures are presented in Section 6.4 and Section 6.5, respectively.

6.1 Link Prediction Methods

The LP problem is described as the prediction of links between entities, based upon the existence of other observed links and nodal specific attributes (Getoor & Diehl, 2005). More formally:

Definition 6.1.1. *Given a graph $G_t(V, E)$ of n nodes/vertices (V with vertices v_i) and a set of links/edges (E with edges e_i) at time t, an attempt is made to arrive at G_{t+1} through the evaluation of possible new edges, $e_{i,j}$ between vertices v_i and v_j (Liben-Nowell & Kleinberg, 2007).*

Therefore, taking Figure 6.1, a prediction of the new (red) links present at T_2 is attempted. As previously discussed (Chapter 3), LP methods have been utilised in a variety of applications, including: recommendation systems (Zhu et al., 2004), electrical power grid structures (Lü & Zhou, 2011) and academic co-authorships (Farrell et al., 2005). In the context of this thesis, LP methods shall be used to predict the development of adolescent social networks - the process being conducted upon the ASSIST social network data.

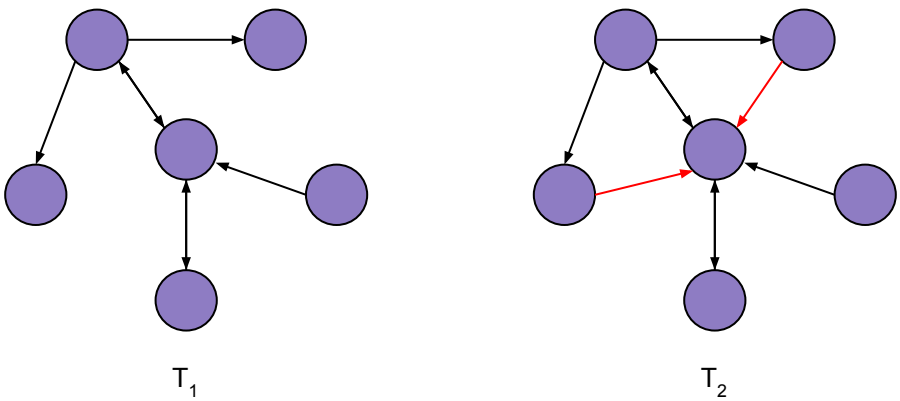


Figure 6.1: Illustration of LP, where the unobserved red links at T_2 are predicted from T_1.

Literature relating to LP methods document a large number of algorithms for the prediction of edges within a network. In their seminal paper, Liben-Nowell & Kleinberg (2007)

collated a large proportion of LP algorithms, evaluating their effectiveness at predicting academic collaborations across five different disciplines; this process was repeated by Lü & Zhou (2011), including a greater breadth of LP algorithms across a broader selection of networks. The work of Liben-Nowell & Kleinberg (2007) and Lü & Zhou (2011), while assessing the LP algorithms across a broad array of networks, did not conduct tests upon friendship networks. It would therefore be of interest to evaluate the performance of LP algorithms, in conjunction with the adolescent connections of the ASSIST data.

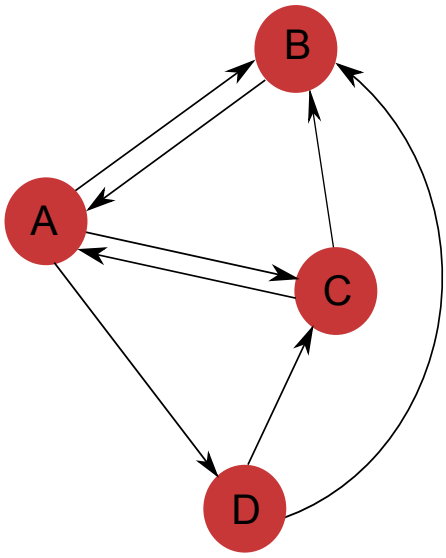


Figure 6.2: Example network for illustration of LP algorithms.

Four prediction methods have been selected for the purpose of this investigation: Adamic/Adar (Section 6.1.1), Katz (Section 6.1.2), SAB Modelling (Section 6.1.3) and PageRank (Section 6.1.4). These methods have been selected for comparison with the newly developed algorithm (PageRank-Max), each method previously demonstrating successful predictive performance within its respective field - discussed further within the context of its designated subsection. A detailed account of each method shall be provided, aided by an example based upon the illustrative network of Figure 6.2 (where appropriate).

6.1.1 Adamic/Adar

The AA method was originally developed to quantify how webpages were similar in terms of content, specifically focusing upon personal web pages; if the content between two pages is similar, Adamic & Adar (2003) theorised that a connection between them is more likely to appear. Adamic & Adar (2003) based their theory upon the notion that friends

tend to be similar to one another (Carley, 1991; Feld, 1981), therefore making connections more probable.

While the AA method originally assessed how "related" web pages were in terms of content, its implementation within Liben-Nowell & Kleinberg (2007) assessed how similar academics were in terms of their collaborators; a connection between two unconnected academics being more likely, if they shared a similar set of collaborators. To perform the AA LP method, the *neighbourhood*, $\Gamma(i)$, of each individual, i, is required; $\Gamma(i)$ being the set of individuals with whom i shares a connection. A score is calculated for each link $(i\,j)$ that is not present (unobserved) in the network, such that:

$$\text{Score[i,j]} = \sum_{z \in \Gamma(i) \cap \Gamma(j)} \frac{1}{\log |\Gamma(z)|} \tag{6.1}$$

where z is a mutual connected vertex of both i and j. The AA LP score for $i\,j$ is therefore based upon the number of connections an individual z (who is a friend of both i and j) possesses. If z has a small number of connections, then having z as a common neighbour of both i and j is rarer than if z had a high number of connections. As such, rarer common neighbours increase Score[i,j] meaning that a link between i and j is more likely.

The following example illustrates the mechanism by which AA makes a link prediction:

Example 6.1.1.

- *Taking the social network of Figure 6.2, the unobserved links are identified as:* $B \rightarrow C$, $B \rightarrow D$, $C \rightarrow D$ *and* $D \rightarrow A$.
- *Taking the unobserved link* $B \rightarrow C$, *examining the friendships of B and C gives the neighbourhoods* $\Gamma(B) = \{A\}$ *and* $\Gamma(C) = \{A, B\}$, *respectively.*
- *As both* $\Gamma(B)$ *and* $\Gamma(C)$ *contain agent A, A is identifies as the only common neighbour of agents B and C.*
- *Agent A has three outward links, as such* $|\Gamma(A)| = 3$ *and therefore the Score[B,C] = 0.910 (3 d.p.).*

- *The scores for the remaining unobserved links ($B \rightarrow D$, $C \rightarrow D$ and $D \rightarrow A$) are also calculated. The resultant scores are ranked and the links with the highest scores are most likely to develop according to the AA link prediction method.*

The example presented is conducted upon a directed network, however, the AA method does not consider the effect of reciprocation - a reciprocated tie being one in which the links $i \rightarrow j$ and $j \rightarrow i$ both exist, previously defined in Definition 3.1.3. Returning to Example 6.1.1, the calculated Score$[B, C]$ for the unobserved link $B \rightarrow C$ does not consider that the link $C \rightarrow B$ exists; this ignore the fact that agent B may wish to reciprocate the link with C, basing the strength of the "relation" purely upon the size of the neighbourhood of A.

The aforementioned reciprocation issue is not present in previous implementations of the AA method, as the method is exacted upon undirected networks (Adamic & Adar, 2003; Liben-Nowell & Kleinberg, 2007; Lü & Zhou, 2011); reciprocation being implicit in an undirected network. It is therefore of interest to investigate the success of AA in the current context of this thesis, in comparison with previous works. The investigation of Liben-Nowell & Kleinberg (2007) finds the AA method to be the most successful link prediction method amongst those it tested (on academic collaborations), indicating a minimum 16% improvement over random predictions in foreseeing future 'Astrophysics' collaborations, and a maximum 54.8% improvement within 'Condensed-Matter' collaborations. The figures quoted from Liben-Nowell & Kleinberg (2007) demonstrate the variability of the method, highlighting the importance of underlying network structure in the success of predictions.

The work of Lü & Zhou (2011) also find the AA method to be particularly successful when exacted upon the 'US Electrical Power Grid' and 'Router-Level Internet' networks, however, the method was less successful amongst the other datasets tested. A further notable issue regarding the AA method, is that it solely predicts the formation of new links and does not concern itself with the dissolution of existing links. Evidently, this is not an issue when considering networks such as those of academic collaboration or electrical power, as links are unlikely to be removed; however, the networks of adolescents may be far more volatile - potentially resulting in the AA method being unable to accurately capture the dynamics of friendship evolution.

While the discussion of the AA method highlights the issues of reciprocation and link

removal for consideration in terms of adolescent social networks, the apparent success of the method amongst the literature presented, along with its structure being based upon the theory of similarity in friendships, suggests the AA method as a suitable candidate for inclusion within the proposed ABS of adolescent connection. Further details of its inclusion within the simulation are documented in Section 6.2.2.

6.1.2 Katz

In the investigation of Liben-Nowell & Kleinberg (2007), the second highest performing LP algorithm was that of the Katz method. Developed by Katz (1953) as a method to identify individuals of status within a group "free from the deficiencies of popularity contest procedures", the method examines not only the number of "popularity votes" an agent receives, but also the popularity of the voting individuals. As such, Katz (1953) argues that a more accurate perception of high status individuals in a group may be garnered. With respect to LP, the popularity votes referred to by Katz (1953) may be considered as connections in a network.

To perform the Katz method, the sociomatrix, X, of a network is required. It is well-known that the paths between individuals in a social network may be found by exploiting the powers of the relevant adjacency matrices (Festinger, 1949). For matrices with binary entries (such as X), non-zero elements x^2_{ij} of the matrix X^2 indicate the number of paths of length two being present between agents i and j; similarly, a non-zero element x^3_{ij} of the matrix X^3, indicates the number of paths of length three between agents i and j - higher powers having corresponding interpretations. In terms of LP, a score for an unrealised link between $i \rightarrow j$ is calculated as:

$$\text{Score[i,j]} = \sum_{l=1}^{n-1} \phi^l |\text{path}^{[l]}_{i,j}| \tag{6.2}$$

whereby $|\text{path}^{[l]}_{i,j}|$ represents the number of paths of length l between i and j, and ϕ is the selected *dampening* factor. The selection of ϕ must satisfy the condition $\phi < 1$ (Phuoc et al., 2009), with $\frac{1}{\phi}$ being the smallest integer value greater than the largest eigenvalue of matrix X (Katz, 1953).

Structuring the LP Katz score in this manner allows indirect relations to be considered, with less weight being given to more distant indirect connections as $l \to \infty$. If a direct path between agents i and j exists, then $\text{path}_{i,j}^{[1]} = 1$. In terms of the popularity concept set out by Katz (1953), if individuals i and j have a large number of connections (high popularity), then the number of short paths between i and j is likely to be high, thus increasing Score[i,j]. However, i and j have low popularity, the number of short paths is likely reduced - decreasing Score[i,j]. The Katz LP method, much like the AA method, assumes undirected network connections, with the underlying concept assuming that popular individuals are more likely to connect with one another - shortening the overall average shortest path length of the network. To illustrate the calculation of the Katz method, an example using the social network of Figure 6.2 is as follows:

Example 6.1.2.

- *For the calculation of the Katz method, the* 4×4 *sociomatrix X of Fig.6.2 is required:*

$$X = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$

  elements $x_{2,3}$*,* $x_{2,4}$*,* $x_{3,4}$ *and* $x_{4,1}$ *are zero, indicating the potentially unobserved links.*

- *As the number of agents* $n = 4$*, the maximum path length for an indirect connection between agents is 3. Therefore the power of matrices to* $n - 1$ *are calculated:*

$$X^2 = \begin{pmatrix} 2 & 2 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 2 & 1 & 0 & 0 \end{pmatrix}$$

$$X^3 = \begin{pmatrix} 3 & 3 & 2 & 2 \\ 2 & 2 & 1 & 0 \\ 2 & 3 & 2 & 1 \\ 1 & 2 & 2 & 2 \end{pmatrix}$$

- *The value ϕ is selected by finding the maximal eigenvalue (λ) of X. As $\lambda = 1.950$ (3 d.p.), the value of $\frac{1}{\phi}$ is taken to be 2, allowing $\phi = 0.5$; this satisfies the requirements of $\phi < 1$ and $\frac{1}{\phi}$ being the smallest integer value greater than the characteristic root of X.*

- *Taking once again the unobserved link of $B \rightarrow C$, the Score[B,C] is calculated as:*

$$Score[B,C] = (0.5)^1 \cdot 0 + (0.5)^2 \cdot 1 + (0.5)^3 \cdot 1 = 0.375 \tag{6.3}$$

- *The remaining unobserved link scores are calculated in the same manner and ranked accordingly. The links with the highest scores, are those which are most likely to occur at a subsequent timestep.*

The results of Liben-Nowell & Kleinberg (2007) showed the Katz method performed consistently well in the prediction of collaborations, with predictions performing similarly to the AA method within 'Condensed Matter', 'General Relativity and Quantum Cosmology' and 'High Energy Physics Theory' networks. The work of Lü & Zhou (2011) also demonstrated the Katz method to be successful, excelling upon application to a number of networks. The method has also been implemented in approaches to 'collaborative filtering' (Huang et al., 2005) and recommender systems (Lü et al., 2012), with the algorithm being used as a benchmark to assess the development of new LP algorithms (Dunlavy et al., 2011; Lichtenwalter et al., 2010; Richard et al., 2010).

The prominence of the Katz method many years after its initial inception, along with the documented success of the method amongst the literature discussed, suggests it to be a suitable candidate for inclusion within the link prediction of adolescents. It must be noted, however, that much like the AA method previously discussed, the Katz method is a standard LP method; as such, only additional links to those of the initial networks may be predicted, with the method not considering links that dissipate over time. With the inclu-

sion of the Katz method within the context of the current work established, details of its implementation within the ABS are discussed further in Section 6.2.2.

6.1.3 Stochastic Actor Based Modelling

The Stochastic Actor Based (SAB) modelling approach is not a static LP method such as those of AA and Katz. Rather, Snijders (1996) defines the SAB approach to be a class of models for longitudinal network data - 'actors' (as defined in Section 3.5) within the network utilising heuristics to optimise their individual goals, subject to a selection of constraints. Discrete observations of a network are explored, with the evolution of social ties from G_t to G_{t+1} a result of many small changes occurring between the specified time periods (Carrington, 2005) - the observed networks assumed to be the result of a Markov process in continuous time.

The SAB approach takes the social network data available at distinct time points, and attempts to model the evolution of the network as a product of specific network and behavioural factors. Consider T observations of a social network, represented as the adjacency matricies X_t for $t = 1, ..., T$, each observation containing the same set of n actors. Evolution of the network is solely modelled from the point of inception X_1, with the evolution to X_1 not being considered (Snijders, 1996) .The actions of actors within the network at t are simulated, changes in friendship ties based upon actor specific personal objective functions; the process attempting to model the micro-changes necessary to arrive at the network of $t + 1$.

The personal objective function for actor (or agent) i is represented by:

$$f_i(\beta, X) \tag{6.4}$$

for a given configuration of the network $X \in \chi$, where χ denotes the class of all possible sociomatrices, and β is a vector parameters upon which f_i is dependant. Each agent holds a set of outward links, and a set of agents with whom a connection is not shared. During the SAB process, at a given time, an agent i is selected uniformly at random (from the set of all possible agents n) and given the option to make a change to their current social situation (Snijders, 2001); the agent being allowed to sever an existing connection or generate a new connection.

The decision making process that agent i follows to improve his social circle, attempts to maximise f_i. A change in the status of the link from i to j creates a new configuration of the network $X(i \rightsquigarrow j)$, the process selecting the link change which yields the maximum value of the expression:

$$\arg\max_{j \neq i}[f_i(\beta, X(i \rightsquigarrow j)) + U(j)] \tag{6.5}$$

where $U(j)$ is described as the element of enigmatic variability in the selection process (Huisman & Snijders, 2003) - potentially some unknown attraction from i to j.

The parameter $U(j)$ is selected to be the Gumbel distribution with mean 0 and scaling parameter 1, as per the conventions of random utility modelling in econometrics (Maddala, 1983). As such, the resulting probability that i changes its connection with j is given by:

$$p_{ij} = \frac{\exp(f_i(\beta, X(i \rightsquigarrow j)))}{\sum_{h=1, h \neq i}^{n} \exp(f_i(\beta, X(i \rightsquigarrow h)))} \quad (j \neq i) \tag{6.6}$$

the expression also being utilised in multinominal logistic regression (Maddala, 1983).

The actor's personal objective function may be tailored to the needs of the investigator, allowing for exploration into the effects of specific factors upon the evolution of a network. The objective function may be constructed in the following manner:

$$f_i(\beta, X) = \sum_{k=1}^{L} \beta_k S_{ik}(X) \tag{6.7}$$

where S_{ik} is the value of the k^{th} selected statistic for agent i, β_k is the coefficent of the k^{th} statistic, with L statistics being considered by the investigator. The SAB models quoted in Carrington (2005) and Snijders (1996, 2001), focus upon specific network statistics (S_k) suggested as a first basic model for use with longitudinal data. The basic model includes the following statistics:

1. Density - The number of out-degrees an agent projects: $S_{i1}(X) = \sum_j x_{ij}$;

2. Reciprocity - The number out-degrees from i that are reciprocated: $S_{i2}(X) = \sum_j x_{ij} x_{ji}$;

3. Popularity - The popularity of agent i's connections, calculated as the number of agents who have also linked to one of i's existing social contacts: $S_{i3}(X) = \sum_j x_{ij} \sum_h x_{hj}$;

4. Activity - The activity of agent i's connections, calculated as the sum of the outward links cast by agent i's outward links: $S_{i4}(X) = \sum_j x_{ij} \sum_h x_{jh}$;

5. Transitivity - The number of transitive triples within the ties of i: $S_{i5}(X) = \sum_{j,h} x_{ij} x_{ih} x_{jh}$;

6. Indirect Relations - The number of agents located at path distance 2: $S_{i6}(X) = \left| j | x_{ij} = 0,\ \max_h(x_{ih} x_{hj}) > 0 \right|$;

7. Balance - The similarity between the outward links of an agent i, compared with the outward links of agent i's connections:

$$S_{i4}(X) = \sum_{j=1}^{n} x_{ij} \sum_{\substack{h=1 \\ h \neq i,\, j}}^{n} (b_0 - |x_{ih} - x_{jh}|)$$

   with the inclusion of b_0 avoiding a balance effect too closely correlated with the density of an agent's ties. The suggested value of b_0 by Carrington (2005), Snijders (1996, 2001) and Snijders et al. (2010) is :

$$b_0 = \frac{1}{(T-1)n(n-1)(n-2)} \sum_{t=1}^{T-1} \sum_{i,j=1}^{n} \sum_{\substack{h=1 \\ h \neq i,\, j}}^{n} |x_{ih}(t) - x_{jh}(t)|$$

   the average balance of ties across the whole network at all available time points.

A further necessary requirement of the SAB process is the rate of change of the network (ρ), also described as how frequently actors make micro-changes. It is assumed that at any point in time, only one actor may make a change to their social situation - said actor only being allowed one single change to their social situation. As this process acts in continuous time, the next actor to change its outgoing ties makes use of the updated network to maximise its objective function - taking into consideration the changes made by previous actors. For simplicity the rate of change for each actor is assumed to be the same, with the time between events taken to be negative exponentially distributed with parameter ρ (Snijders, 2001).

Once the model is specified, the SAB process attempts to estimate the values of β_k - described as the importance of S_k in the model. As previously discussed, the evolution of the network is assumed to be a Markovian process in continuous time, where calculation of a likelihood function (MLE) is notoriously difficult (Snijders, 1996; Stewart, 2009); therefore the use of stochastic approximation is employed, through the Robbins & Monro (1951) algorithm - known as the Markov Chain Monte Carlo (MCMC) method. Exact details of the estimation process may be found in Snijders (2002) and Ripley et al. (2012); however, the efficiency of the MCMC approach is said to be superseded by an MLE approach detailed in Snijders et al. (2010) - the best method for the SAB process still an ongoing topic of debate within the literature.

On completion of the SAB process, estimates of the β_k parameters are produced; the value of β_k describing the importance of S_k in the evolution of X_t to X_{t+1}. Therefore, the SAB process generates a model, the β_k values interpreted in a similar manner to the coefficients generated from a regression model. Specific software for the creation of an SAB model is available, RSiena (Ripley et al., 2012) being a downloadable package for use within the R software environment. RSiena produces the required model, the user inputting the matrices desired for analysis and the specified elements for inclusion within the actor's personal objective function.

As the SAB process is complex in its execution, a simple example using the network of Figure 6.2 would be counter-intuitive. Therefore, to illustrate the process further, a step-by-step guide of the events are as follows:

- The investigator must have at their disposal, at least two sociomatrices X_t and X_{t+1} (containing the same set of actors) from which they wish to understand the social network evolution;

- The rate of network change is calculated by examining the number of changes between consecutive sociomatrices;

- The statistics selected for inclusion within the actor's personal objective function (S_k) are decided, of which the intensities (β_k) are unknown;

- The SAB process is initiated, with actors being selected at random from a uniform distribution to make a change to their social situation - subject to their personal

objective function;

- Multiple simulations are run, attempting to find the values of β_k which best represent the evolution of the network from X_t to X_{t+1};

- The values of β_k in relation to S_k are interpreted, allowing the investigator to assess, of those statistics included in the model, which are of key importance.

Further examples of the SAB method, and its output, may be found amongst Burk et al. (2007) and Snijders et al. (2007a).

SAB modelling is not a conventional LP technique such as those of AA and Katz, the procedure being used for the interpretation of important factors in social network evolution. However, underpinning the processes is a prediction of agents' links, which in turn generate the desired model. As such, the structure of SAB modelling may be presented as an LP procedure, but with a different end goal to those of standard LP mechanisms. Given that behavioural interpretations from SAB modelling rely upon the link predictions made, it would be of interest to investigate the accuracy of such predictions.

While other dynamic network model formulations have also been proposed (Bala & Goyal, 2000; Marsili, 2004; Skyrms & Pemantle, 2000), SAB is said to offer an unmatched degree of flexibility in terms of actor-driven influence investigation (Snijders et al., 2010). Furthermore, the current trend of using SAB models for behavioural analysis (Light et al., 2013; Rayner et al., 2013; Sentse et al., 2013), coupled with the prior usage of the theory upon the ASSIST data (Mercken et al., 2012b; Steglich et al., 2012), make SAB processes suitable for further investigation. The inclusion of SAB procedures, within the created ABS for the assessment of its underlying LP mechanisms, are discussed further in Section 6.2.2.

6.1.4 PageRank

The PageRank (PR) algorithm was developed by Brin & Page (1998), the founders of Google (2013) - the company now estimated to be worth $268.44 billion (Forbes, 2013). Google initially began as a "prototype large-scale search engine" from Stanford University (Brin & Page, 1998), developed to rival leading search engines AltaVista (2012), Yahoo! (2013) and Lycos Search (2013). At the time, popular search engines were said to return

many irrelevant search results, with the lack of substantial advances in this area attributed to the "closed door" policies and "advertiser driven" practices of commercial operators (Brin & Page, 1998). Google aimed to improve the web search experience by ranking the returned pages in order of importance, the rankings calculated as a result of the PR algorithm.

PR analyses the link structure of a network, taking into consideration not only the number of links to a node, but also the importance (PR) of the node sending the outward link. The PR (w_i) for each node i, is such that $w_i \geq 0$ and $w_j > w_k$ indicates j is a more important node than k. If $\bar{H}_i$ denotes the set of nodes that link to i, and H_i the set of nodes linked outwardly from i, then the PR w_i is calculated as:

$$w_i = \sum_{j \in \bar{H}_i} \frac{w_j}{|H_j|} \tag{6.8}$$

The calculation of w_i is recursive and can be initiated with any selected initial importance scores, iterating until convergence. The calculation of the PR may be interpreted as a random walk on a graph; in the context of the internet, a "random surfer" clicks on webpage links at random - the resultant probability of arriving at a page defined as its PR.

The "random surfer" calculation of PR is useful when importance scores are necessary for large graphs (such as the internet), whereby the adjacency matrix of connections X is unobtainable. However, if X is known, an adjusted matrix (M) may be calculated with $m_{ij} = \frac{1}{|H_j|}$ if the link $j \to i$ exists and $m_{ij} = 0$ otherwise. The PR calculation may then be expressed as a system of linear equations $Mw = w$, with the problem reduced to finding the principal eigenvector of the matrix M. Due to the properties of M, it is possible to find an eigenvalue $\lambda = 1$ which generates a **unique** positive eigenvector; this eigenvector being the vector of PageRanks (Page & Brin, 1999).

The matrix M is defined as column stochastic if each element $m_{ij} \geq 0$ and the sum of each column is 1, this ensures the existence of $\lambda = 1$ (Bryan, 2006). However, this does not guarantee the existence of a *unique* λ necessary for ranking, therefore other requirements of M need to be satisfied. From Perron-Frobenious theorem (Meyer, 2000), a column stochastic matrix M that is **irreducible** with $m_{ij} \geq 0$, generates:

- an eigenvalue $\lambda > 0$ with corresponding eigenvector $v > 0$.

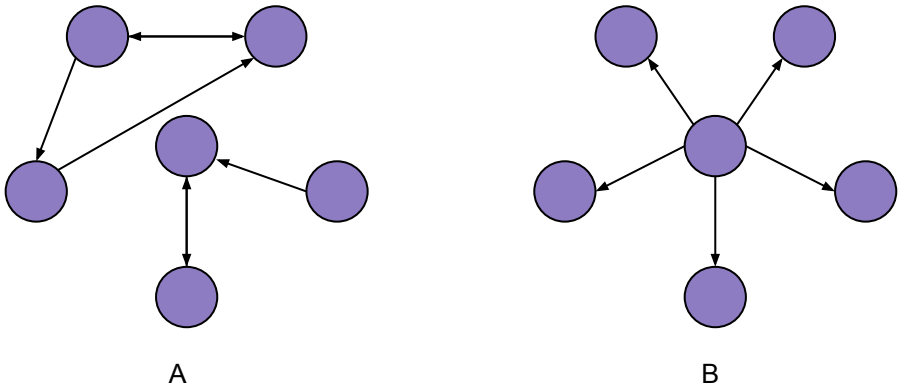


Figure 6.3: A is a representation of a network with two disconnected clusters. B is a network where the centre agent has five dangling nodes.

- the existence of a dominant eigenvalue λ_1, such that $\lambda_1 > |\lambda_2| \geq |\lambda_3| \geq ... \geq |\lambda_n|$
- all eigenvectors ≥ 0 are a multiple of w.

Therefore, M also needs to satisfy the condition of irreducibility, whereby M cannot be placed into block-upper triangular form through a series of permutations (Pillai et al., 2005). M may become reducible if disconnected clusters of nodes exist in the network (Figure 6.3A). Furthermore, nodes with an inward link but no outward links, termed as "dangling nodes" (Figure 6.3B), also affect the necessary requirements for a unique vector of PageRanks (Ipsen & Selee, 2008).

To ensure the successful calculation of the PR vector, M is required to represent a **strongly connected** graph; a graph being strongly connected if a path from any given node i to j exists (Fernández & Madrid, 2007). Performing the PR calculation upon a strongly connected graph is not always possible, as is the case for both web pages and social networks. As such, calculation of a new matrix $\bar{M}$ is required:

$$\bar{M} = (1 - d)Q + dM \tag{6.9}$$

where Q is the matrix of elements $\frac{1}{n}$ and d is the 'dampening factor', ensuring that $\bar{m}_{ij} \geq (1 - d)Q$ which satisfies the required conditions; d is generally selected to be 0.85 (Brin & Page, 1998; Bryan, 2006; Page & Brin, 1999). The principal eigenvector of $\bar{M}$ is calculated, returning the required PR.

To illustrate PR, the following example is conducted upon the network of Figure 6.2:
Example 6.1.3.

- *The sociomatrix X of the network of Figure 6.2 is :*

$$\begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$

  with the number of outward links for each agent: $|H_A| = 3$, $|H_B| = 1$, $|H_C| = 2$ *and* $|H_D| = 2$.

- *The matrix M is calculated where* $m_{ij} = \frac{1}{|H_j|}$ *if the link* $j \to i$ *exists and* $m_{ij} = 0$ *otherwise, giving:*

$$M = \begin{pmatrix} 0 & 1 & \frac{1}{2} & 0 \\ \frac{1}{3} & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \end{pmatrix}$$

- *Taking* $d = 0.85$ *with* $n = 4$*, the* $\bar{M}$ *matrix is calulated as:*

$$\bar{M} = 0.15 \cdot \begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix} + 0.85 \cdot \begin{pmatrix} 0 & 1 & \frac{1}{2} & 0 \\ \frac{1}{3} & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \end{pmatrix} \tag{6.10}$$

$$\bar{M} = \begin{pmatrix} \frac{3}{80} & \frac{71}{80} & \frac{37}{80} & \frac{3}{80} \\ \frac{77}{240} & \frac{3}{80} & \frac{37}{80} & \frac{37}{80} \\ \frac{77}{240} & \frac{3}{80} & \frac{3}{80} & \frac{37}{80} \\ \frac{77}{240} & \frac{3}{80} & \frac{3}{80} & \frac{3}{80} \end{pmatrix} \tag{6.11}$$

- *The matrix* $\bar{M}$ *is in the form that allows for the calculation of the PR vector. The*

eigenvector of $\bar{M}$ corresponding to the dominant eigenvalue is found to be:

$$W = \begin{pmatrix} 0.36816 \\ 0.28796 \\ 0.20208 \\ 0.14181 \end{pmatrix} \tag{6.12}$$

- *Hence, the PageRank of each node is found. As node A has the highest PageRank, it is therefore the most "important" node in the network.*

Although Google revolutionised web search with its PR calculations, the concept of link analysis through the calculation of the principal eigenvector appears prior to that of Brin & Page (1998). For example, the work of Pinski & Narin (1976) discuss citation based influence as an eigenvalue problem, with Marchiori (1997) formulating the problem in the context of the internet; Kleinberg (1998) also developed a web based link analysis algorithm (named 'HITS'), the same year as the publication of PR. In terms of social networks, Bonacich (1972) also formulated the eigenvector calculation problem as a measure of network centrality, termed as "eigenvector centrality", aiming to find the most central individuals.

The PR algorithm ranks the importance of nodes in a network, but does not make explicit claims about the formulation of the links between them; therefore, PR has to be interpreted specifically for inclusion within the LP process. Liben-Nowell & Kleinberg (2007) make use of the random walk formulation of the PR, calculating the expected number of steps necessary from i to arrive at j (R_{ij}), such that:

$$\text{Score}[i, j] = -(R_{ij} + R_{ji}) \tag{6.13}$$

considering also the steps taken from j to i, as this may not be symmetric. The links with maximum Score$[i, j]$ are those taken to be most likely to occur at a later timestep, as the paths between these nodes is already short. Similarly, the work of Lü & Zhou (2011) uses the transition probabilities generated by the matrix $\bar{M}$ to calculate the associated link score.

The work of Liben-Nowell & Kleinberg (2007) and Lü & Zhou (2011) do not find the PR method to be particularly successful in terms of LP, with Lü & Zhou (2011) exhibiting

only one significant observation in the prediction of 'Network Science' citations. However, PR regularly appears in link analysis literature (Ding et al., 2009; Haveliwala, 2003; Heidemann et al., 2010; Kwak et al., 2010; Ma et al., 2008), with the concept being centred upon the importance of eigen-centrality in a network; Chapter 5 identified centrality as a key factor in adolescent social network structure and behavioural influence. The existing PageRank LP method informs the newly developed PageRank-Max algorithm, discussed further in Section 6.3.

6.1.5 LP Discussion

LP literature offers a wealth of methods from which to predict the future existence of a link, as made evident by the reviews of Liben-Nowell & Kleinberg (2007) and Lü & Zhou (2011). The four methods proposed for comparison with PageRank-Max (AA, Katz, SAB and PR) have all been selected due to their importance within the literature. Each method focuses upon particular factors said to be key in the generation of new links:

- AA - Common neighbours of disconnected agents;
- Katz - Path lengths between agents;
- SAB - A variety of network statistics: density, reciprocation, popularity, activity, transitivity, number of agents at distance two and balance;
- PR - Connecting with agents of high eigen-centrality.

The concepts of each method shall be transferred into the simulation discussed in the following section (6.2), assessing the importance of said link generation factors upon the evolution of adolescent connections.

6.2 Simulation

Prior to describing PageRank-Max, this section discusses the creation of a simulation framework to predict evolving social networks - the framework being key to the development of the PageRank-Max algorithm. To begin, an outline of the general working of the social network simulation (SNS) is provided in Section 6.2.1. Section 6.1 discussed

the methods to be used within the SNS, yet the AA, Katz and PR methods proposed are conventionally exacted in a static manner, with the SAB method generally not interpreted as an LP method. As such, some conversion of the relevant processes is required to effectively exact the LP procedures in a simulation framework; this discussion occurring in Section 6.2.2. This discussion then leads to the introduction of PageRank-Max in Section 6.3.

6.2.1 Simulation Construction

The aim of the proposed simulation is to take the ASSIST data and simulate the evolution of the social networks over time, with an attempt to understand the process by which connections are modified. This section guides the reader through the creation process, each heading identifying the area of discussion. The topics of data processing, software, simulation logic and initialisation are covered, leading to a discussion of the LP implementation in Section 6.2.2.

Data

The ASSIST data provides multiple observations of a school social network, therefore, the predictions made may be assessed against real data at later time periods - gaining an insight into the accuracy of the predictions. Three waves of data are available (T_1, T_2 and T_3), as such, two predictions can be made - that of T_1 to T_2 and T_2 to T_3. The decision to segment the predictions in this manner, as opposed to a prediction of T_1 to T_3 (including T_2 data within the calculation), is due to following reasons:

- Maximum usage of the data, rather than 'blending' the data of T_1 and T_2 to make one T_3 prediction;
- The ability to assess time effects in the performance of LP methods, identifying if algorithms perform differently in their predictions of T_2 and T_3;
- The SAB work of Steglich et al. (2012) suggests analysing each observation period separately, offering an improvement to the generated model.

An Access database has been created for use with the simulation, holding information regarding friendship ties and basic student information. The database contains a separate

table relating to the adjacency matrix of social ties, for each school at each time step; this allows individual schools to be modelled separately with ease. Data tables containing basic school information are also included within the database, containing the number of tie changes (ϵ) between subsequent network observations; further details of the data contained within the database are given in Section 6.4.

Software

The software used for the SNS is AnyLogic (2002), opting to move away from Netlogo (Wilensky, 1999) - the software of choice for the PP model of Chapter 4. The reason for the selection of AnyLogic over Netlogo, is the ease in which it can connect to databases; this being a requirement when inputting the ASSIST data into the SNS. Furthermore, AnyLogic offers the user the ability to expand its basic functionality with Java, which streamlines the coding of LP methods into the simulation; a discussion of ABS software was covered in Chapter 2.

Logic

The created LP simulation logic, follows a similar process to that of the underlying SAB simulations. The following step-by-step guide describes the process:

- On initialisation, the sociomatrix (X) and number of link changes (ϵ) are read from the database, giving a network rate of change $\rho = \frac{1}{\epsilon}$;
- At time t an event occurs, with the time between events being negatively exponentially distributed with parameter ρ;
- The event signifies that an agent must make a change to their outgoing links, the agent making the change being selected uniformly at random (termed as the 'searching agent');
- The randomly selected agent i (searching agent) receives a "message" telling them they must make a change, the change made being based upon the maximisation of i's personal objective function f_i;
- Agent i iterates through the link changes offered by the selected LP method (the

'testing agents'), finding their maximum f_i.

- Agent i makes one change to their outgoing links, updating X accordingly;
- The process repeats until stopping conditions are satisfied (discussed further in Section 6.5), subsequent agents making use of the updated links from previous agents to make their decisions.

The advancement of the simulation may therefore be interpreted as having a DES structure, as the system decides when events will occur and the selection of the agent who must make a change. An important deviation from the DES structure is that the changes made to the system are agent based decisions, the agents selecting the friendship option that most suits them (through their personal objective function). As a result, agent j must consider the changes made previously by agent i; this means agent j's decisions may be affected by those of i, potentially changing j's overall decision. Modelling friendship changes in the specified manner, means that individual decisions affect the system as a whole; individual connection decisions affecting future connections the network. The simulation may therefore be thought of as an ABS, with discrete event based timing.

AnyLogic uses Java, which is an object-orientated programming (OOP) language. The simulation is structured to have a 'Main' class, where the methods necessary for running the simulation are executed, and an 'Agent' class, whereby each instance of Agent represents an individual from the ASSIST data. Each Agent object has a variable (an array list) relating to the individual's connections, with access to a global array containing the adjacency matrix of all links for the school being simulated. When an update occurs, the changing 'Agent' object (searching agent) updates its own link information variable and the global adjacency matrix; a diagram of the logic is visible in Figure 6.4. The local copy of an agent's links is used exclusively for visualisation purposes.

Initialisation

On initialisation, the simulation is required to create multiple instances of the Agent class. The user must decide the school and timestep for prediction, the simulation accessing the ASSIST database and querying the relevant tables through the use of SQL. The sociomatrix X and number of changes ϵ are saved as global variables, with the number of Agent objects created based on the information within X. A separate data table is accessed, containing

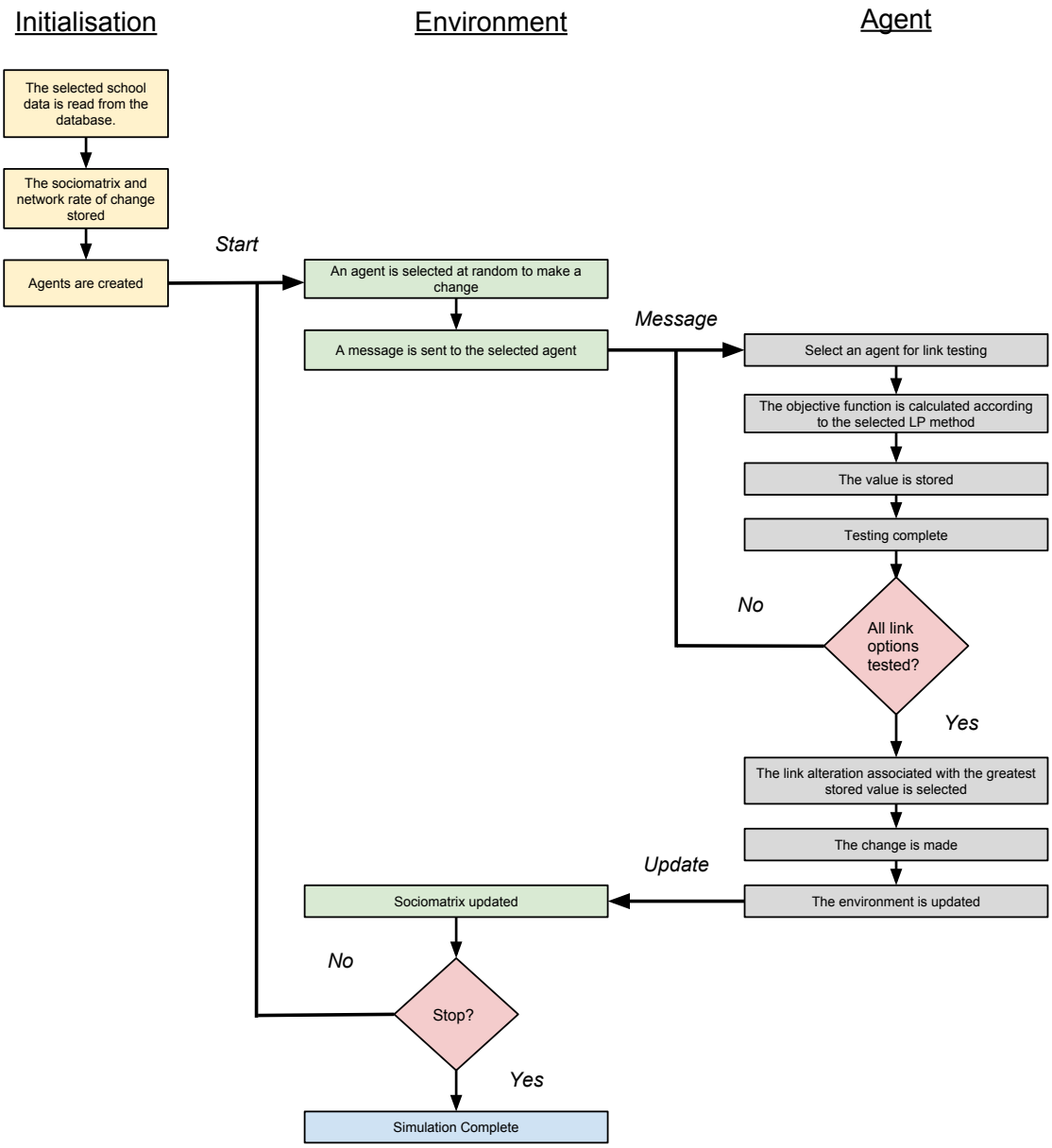


Figure 6.4: Simulation logic describing the timing and agent-based decisions.

the properties of the individuals to be simulated (such as unique id); the information is then applied, giving each agent an identity.

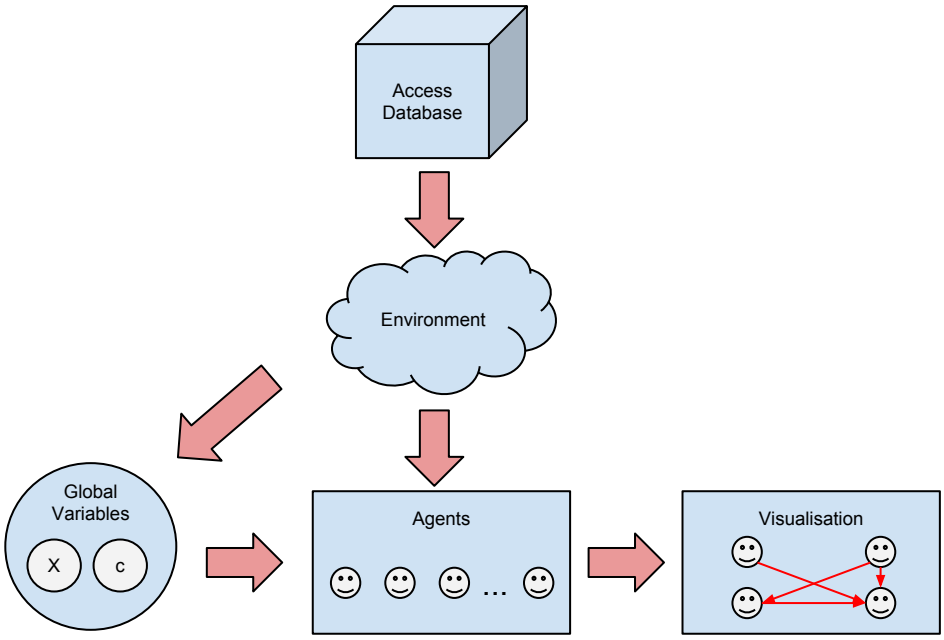


Figure 6.5: Diagram of the interaction between the elements necessary for initialisation.

Each agent then accesses the row in X that represents their connections, storing the agents to whom they send an outlink within a local variable. The network is then drawn for visualisation purposes, the graphics being able to update each time an agent makes a change; Figure 6.5 represents the interactions within the initialisation process. With the initialisation process complete, a representation of the school network (at the designated time) is present, the simulation being able to commence.

6.2.2 LP Method Implementation

Following a description of the simulation framework, and the iterative process of the agent decisions, this section focuses upon calculation of the personal objective functions. As previously discussed, the majority of LP methods presented in Section 6.1 are not necessarily implemented in a dynamic manner; therefore, some interpretation of their characteristics is necessary. The headings below describe the implementation of each of the discussed methods, AA, Katz, SAB and PR, with a discussion of the new PR based method

(PageRank-Max) occurring in Section 6.3.

Adamic Adar

The AA method is based on the commonality of neighbours, a link score being calculated to assess a connection from i to j. For simplicity, the link score is used as the agent's objective function (f_i), giving a vector of values such that:

$$f_{i,j} = \sum_{z \in \Gamma(i) \cap \Gamma(j)} \frac{1}{\log |\Gamma(z)|} \tag{6.14}$$

The AA objective function may solely be calculated to test the creation of a new link, being applicable only if the searching agent i shares a common neighbour z with the testing agent j. Therefore, to assess whether i should make a link to j, i must first check if it has any neighbours in common with j; if so, i can evaluate f_i according to equation 6.14. If i does not have any common neighbours with j, i cannot calculate f_i, as this is not a potential link that may occur (according to the AA method); this process is repeated until all agents (j) have been tested.

It may arise that after iterating through all potential links, no values of f_i were calculated - a result of i having no common neighbours with any other agents. In such circumstances, i will break one of its existing links (selected uniformly at random). The inclusion of a rule to break links, allows the simulation implementation of AA to disconnect agents - the standard AA LP algorithm not accounting for this. Furthermore, it may also be the case that i has no outward links due to isolation; agent i will then select an agent to form a connection with at random, the iterative link checking process not being completed.

An extension to Figure 6.4 describing the AA logic is displayed in Figure 6.6. It is important to note that this method does not make any changes to the network during the objective function evaluation process, only making a change when the largest value of f_i, j has been found - or when the disconnection and isolation rules are invoked. The f_i calculations therefore do not assess the effect of a change, but rather indicate which connections the AA method predicts are most likely to occur.

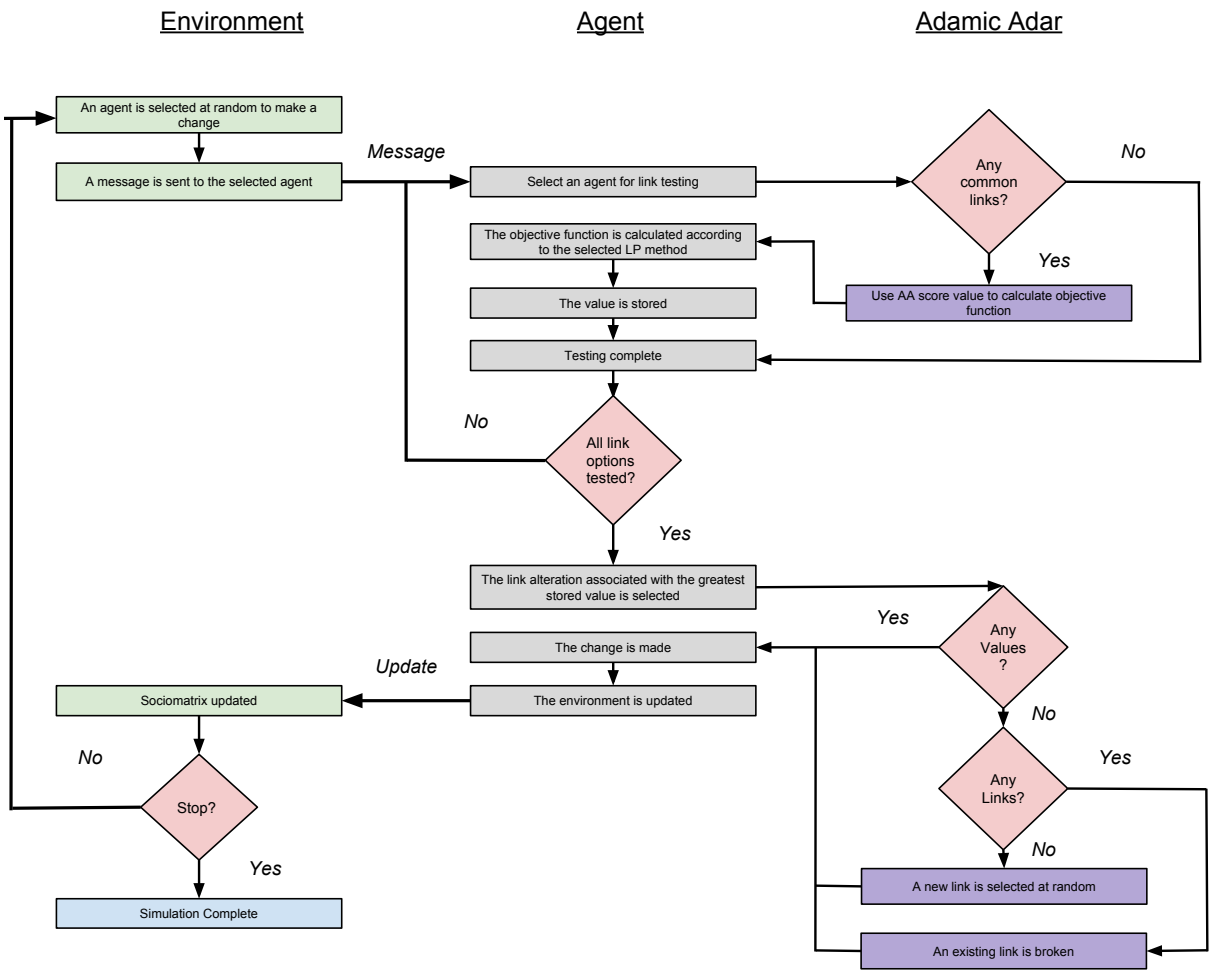


Figure 6.6: Updated simulation logic describing the process of the AA method.

Katz

The Katz implementation works in a similar manner to that of AA, using the Katz score as the value for f_i, with:

$$f_{i,j} = \sum_{l=1}^{n-1} \phi^l |\text{path}_{i,j}^{[l]}| \tag{6.15}$$

Once again, the Katz method is calculated only to check whether a new link should be generated. Therefore, if a link already exists from i to j, $f_{i,j}$ is not calculated. The method is based on the existence of paths between agents, but it may be the case that there are no paths linking the searching agent (i) and the testing agent (j); in this case, an existing link is severed. If i has no links, then the iteration is not performed - a new connection from i being generated at random.

Figure 6.7 shows the logic of the Katz method, maintaining a similar structure to that of AA. Before the simulation may commence, the dampening constant ϕ is required; this is calculated for each school sociomatrix at T_1 and T_2 in the manner described in Example 6.1.2, the values being stored in the simulation Access database. During the initialisation process, the appropriate value of ϕ is then read and stored as a global variable - the value then being used in the calculation of $f_{i,j}$.

The Katz method also requires use of the l parameter, specifying the maximum length of paths selected for inclusion within the calculation of f_i. The ASSIST school networks can contain up to 254 nodes, meaning that a path length of 253 may exist. To evaluate all possible paths would be computationally intensive, with the longer paths being discounted more heavily. Given that the ASSIST questionnaire restricts participants into naming up to six friends, and the work of Christakis & Fowler (2010b) suggesting individuals at distance greater than three do not provide significant influence to behaviour, the value of l is selected such that $l = 3$; this reduces computation time, while still capturing the influence said to be of most importance.

Just as with the AA method, the Katz implementation does not make network changes during the calculation process, with calculation of the Katz score based upon the existing set up of the network; the relevance of this is discussed further when describing the SAB and PR-Max implementations.

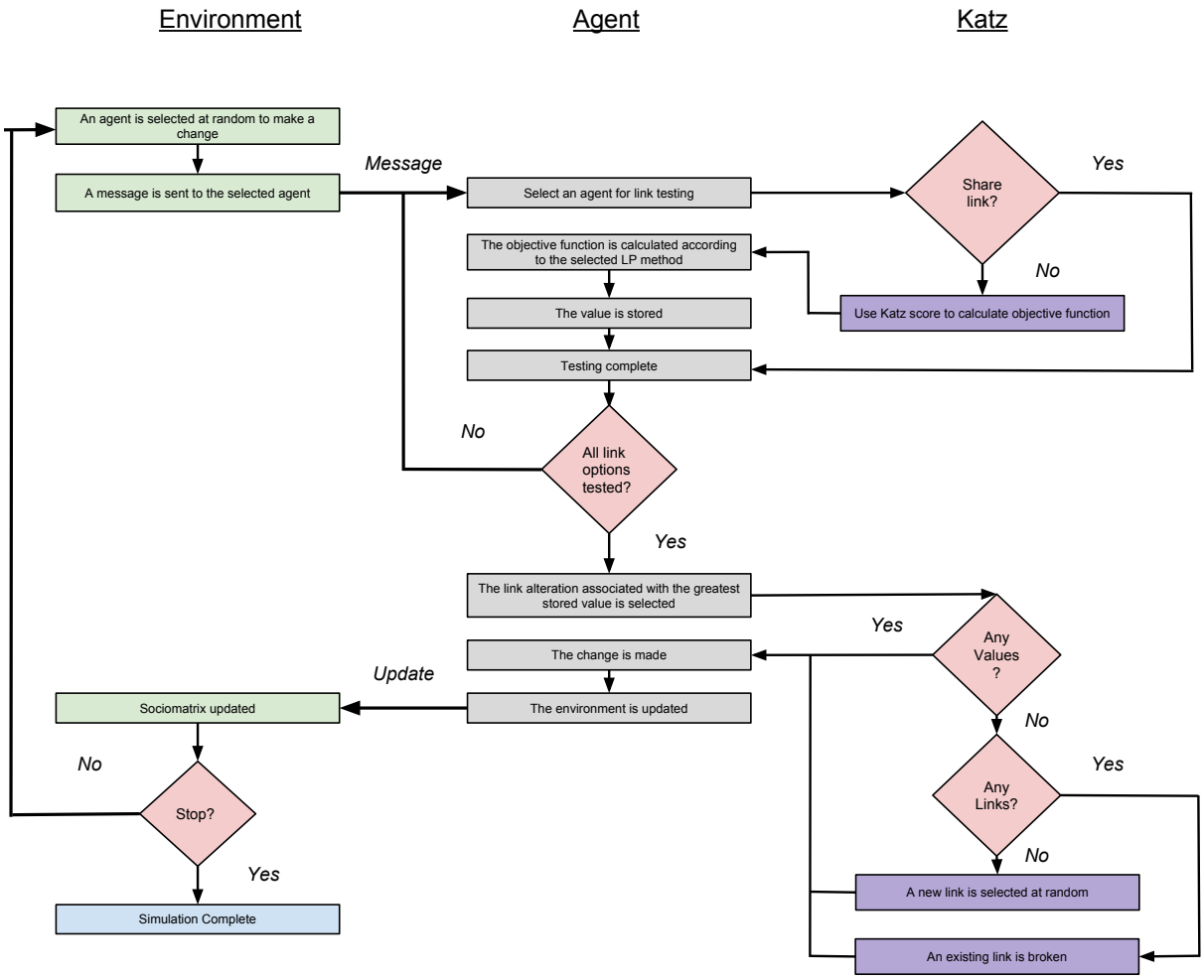


Figure 6.7: Updated simulation logic describing the process of the Katz method.

Stochastic Actor Based Method

As the simulation is based upon the framework of SAB modelling, its implementation logic is the most simplistic. The objective function f_i of agent i is selected to be:

$$f_{i,j}(\beta, X) = \sum_{k=1}^{L} \beta_k S_{ik}(X) \tag{6.16}$$

with the statistics selected to be those of density, reciprocity, popularity, activity, transitivity, balance and agents at distance two; as per the specifications of Section 6.1.3.

The conventional implementation of SAB modelling uses multiple simulations of the network to estimate the β_k values; these values describing the relative importance of each statistic in the evolution of the network. This thesis is concerned with the link predictions made, and the ability of a generated SAB model to inform the process of link placement accurately. As such, prior to running the simulation with the SAB objective function, an SAB model is generated to estimate the β_k values.

The SAB model is fitted with the use of the RSiena package, previously discussed in Section 6.1.3. A model is generated for each school, for each prediction (T_1 to T_2, T_2 to T_3), and the β_k values recorded within the database. On initialisation, the β_k values for the appropriate prediction at the specified time step are read - the figures being stored as global variables. Therefore, during the SAB simulation implementation, friendship changes are made based upon the RSiena model. The resultant network can then be assessed, giving an interpretation of the accuracy of SAB predictions and the model produced by RSiena.

A key difference with the SAB LP implementation, in comparison with that of AA and Katz, is that agents make a friendship change and *then* assess the impact to their objective function. During the simulation, the searching agent (i) iterates through each test agent as follows:

- Test agent j is selected;
- The nature of i's tie to j is changed; if no connection exists, it is created, and if a connection is present, it is removed;
- The value of $f_{i,j}$ is calculated and stored;
- The change of connection from i to j is erased, the connection reverting back to its original state;
- Searching agent i moves on to the next test agent.

During this process, agents actually evaluate a change in terms of a direct impact to themselves - assessing whether the change in connection offers more desirable network statistics (quantified by the selected options of S_k implemented). The simulation then creates a

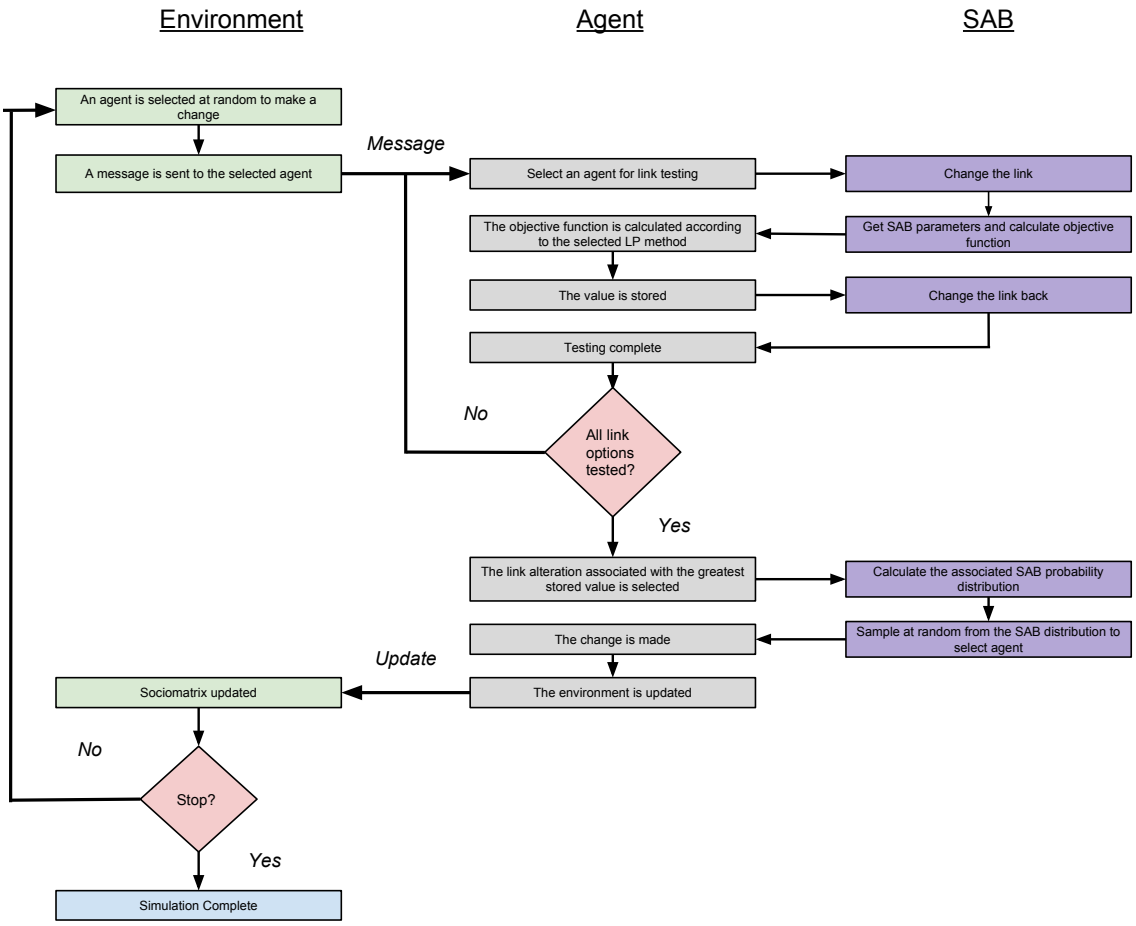


Figure 6.8: Updated simulation logic describing the process of the SAB method.

probability distribution, the probability of a change from i to j (p_{ij}) calculated as:

$$p_{ij} = \frac{\exp(f_{i,j})}{\sum_{h=1,h\neq i}^{n} \exp(f_{i,h})} \quad (j \neq i) \tag{6.17}$$

The link change made is then sampled at random from the created distribution, agent i making the selected change; a diagram of the logic is presented in Figure 6.8.

Evidently, to perform the SAB implementation, two waves of network data are required to generate the necessary β_k values; therefore, it cannot be classified as a prediction method, as prior knowledge of network evolution is required. The adoption of this method within the simulation, is to assess the performance of an SAB model in its LP capabilities; the results able to assess how well an SAB model quantifies the evolution of the network.

As previously discussed, many options are available for inclusion within the SAB objective function (Section 6.1.3). It may be the case that the selected options do not reflect the evolution process well, with other statistics potentially being more important. However, given that SAB literature classifies the selected options as a base model, the results will give an overview of the basic accuracy of SAB LP predictions.

PageRank

The PR implementation alters the agent logic of the simulation, as the searching agent does not need to iterate through each testing agent to find the PR. On receipt of the message from the environment to make a change, the agent transforms the sociomatrix into the required stochastic irreducible matrix and calculates the principle eigenvector; the PR (w_j) of each agent is found in one calculation, the value of $f_{i,j}$ set to w_j. Much like the SAB implementation, a probability distribution is created such that the probability of i selecting j:

$$p_{ij} = \frac{f_{i,j}}{\sum_{h=1,h\neq i}^{n} f_{i,h}} \quad (j \neq i) \tag{6.18}$$

The simulation then samples from the PR distribution, selecting an agent at random. If a link from i to the selected agent does not exist, then a connection is made. If i already shares a link with the selected agent, i assesses the PR for each of its existing links and disconnects from the agent with the lowest PR value.

The reason for the inclusion of a disconnection rule is that, due to the PR probability distribution, the agent with the largest PR is most likely to be selected. If the disconnection rule did not exist and a change in the relationship was simply required, should i be connected to the agent with the largest PR (and they were selected), i would have to disconnect from them; this means that the agent with the largest PR would be the most likely to be both connected to and disconnected from. Given that the PR implies status in the network, the agents with the largest status having the highest PR, it may be more realistic to consider an agent wanting to disconnect from its lowest ranked agents; hence, the inclusion of the disconnection rule. A diagram of the adjusted agent logic is visible in Figure 6.9.

The PR method does not make changes to the network before the calculation process, much

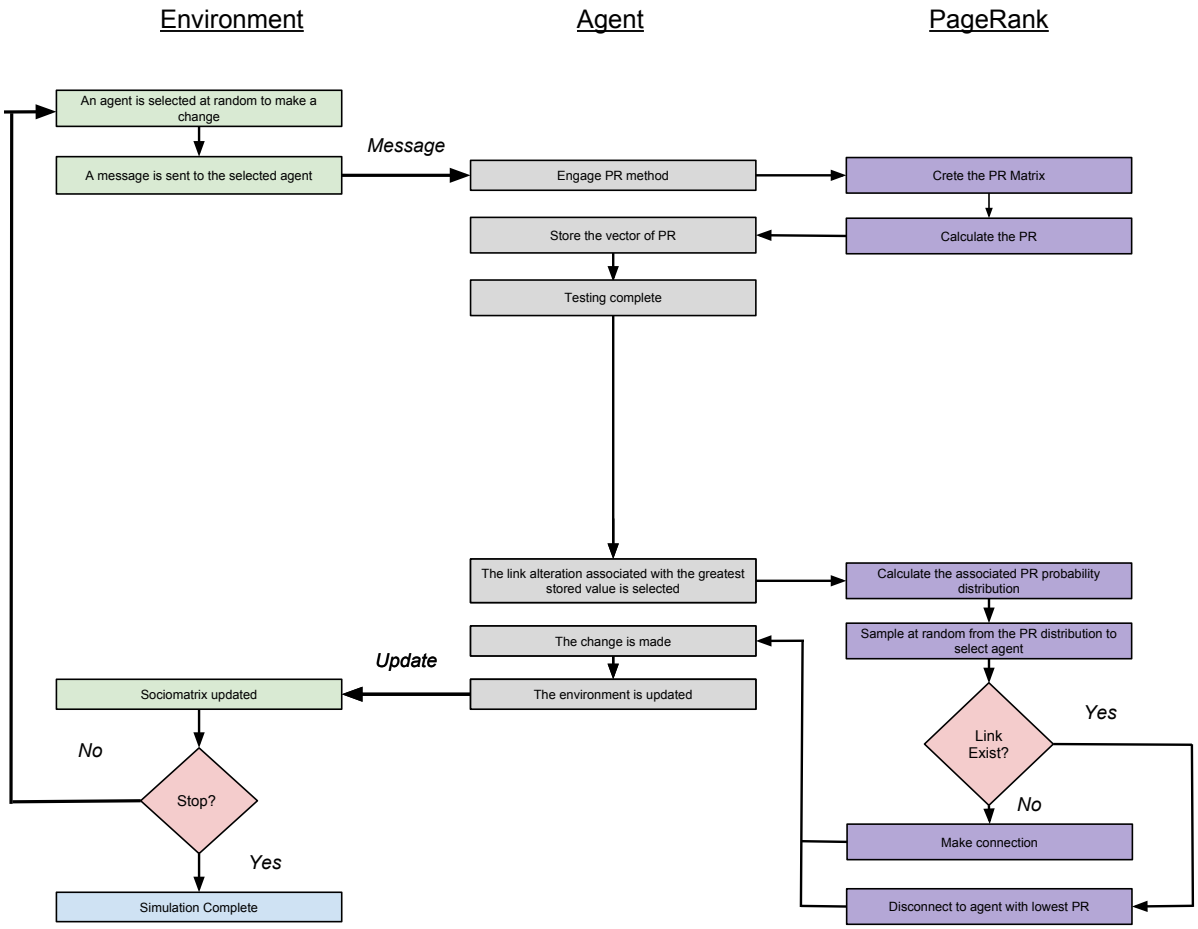


Figure 6.9: Updated simulation logic describing the process of the PR method.

like the AA and Katz methods, but uses a probability distribution similar to that of SAB method. This particular implementation of PR emphasises connections to agents with a high PR, with a high PR agent more likely selected by a searching agent for connection.

Chapter 5 identified the importance of centrality in social network structure, identifying the role of specific individuals in a network. Furthermore, the use of ABS allows for an individualistic representation of system evolution, focusing upon specific agent objectives. As such, the created simulation framework presents the opportunity to investigate the effect of an individualistic perspective of eigen-centrality. This means that rather than opting to connect with agents of high PR, a searching agent selects the connection (or disconnection) which increases their own PR. This new perspective forms the foundations of the PageRank-Max (PR-Max) algorithm, discussed further in Section 6.3.

6.3 PageRank-Max

Given the potential importance of centrality in message diffusion within a social network (previously highlighted in Chapter 5), it stands to reason that centrality may also be of importance to the individuals comprising the social network. The PR-Max method provides an individual perspective of centrality, a searching agent altering its connections based upon the personal optimisation of its own eigen-centrality. Section 6.3.1 outlines the logic of the PR-Max algorithm, with Section 6.3.2 providing an overview of all the algorithms selected for this investigation.

6.3.1 PageRank-Max Outline

The PR-Max method seeks to find the connection that may improve an agents own PR. On receipt of a message from the environment, the changing agent (i) begins iterating through all agents in the network as follows:

- Agent j is selected for testing;
- The connection from i to j is altered, either by forming a link or breaking an existing link;
- Agent i's PR is calculated and stored as $f_{i,j}$.
- The connection change is reversed;
- The process repeats.

Once all possible changes to i's connections are assessed, the greatest value of $f_{i,j}$ is selected - the associated connection change being made. The PR-Max method works much like the SAB method, testing the result of an actual change to the network; however, it does not require the creation of a model prior to use, as a transformation of the sociomatix is its sole requirement. The simplicity of the PR calculation means that PR-Max method also does not require two waves of network data, being able to predict changes in the network without prior knowledge of its evolution; a diagram of the PR-Max logic is present in Figure 6.10.

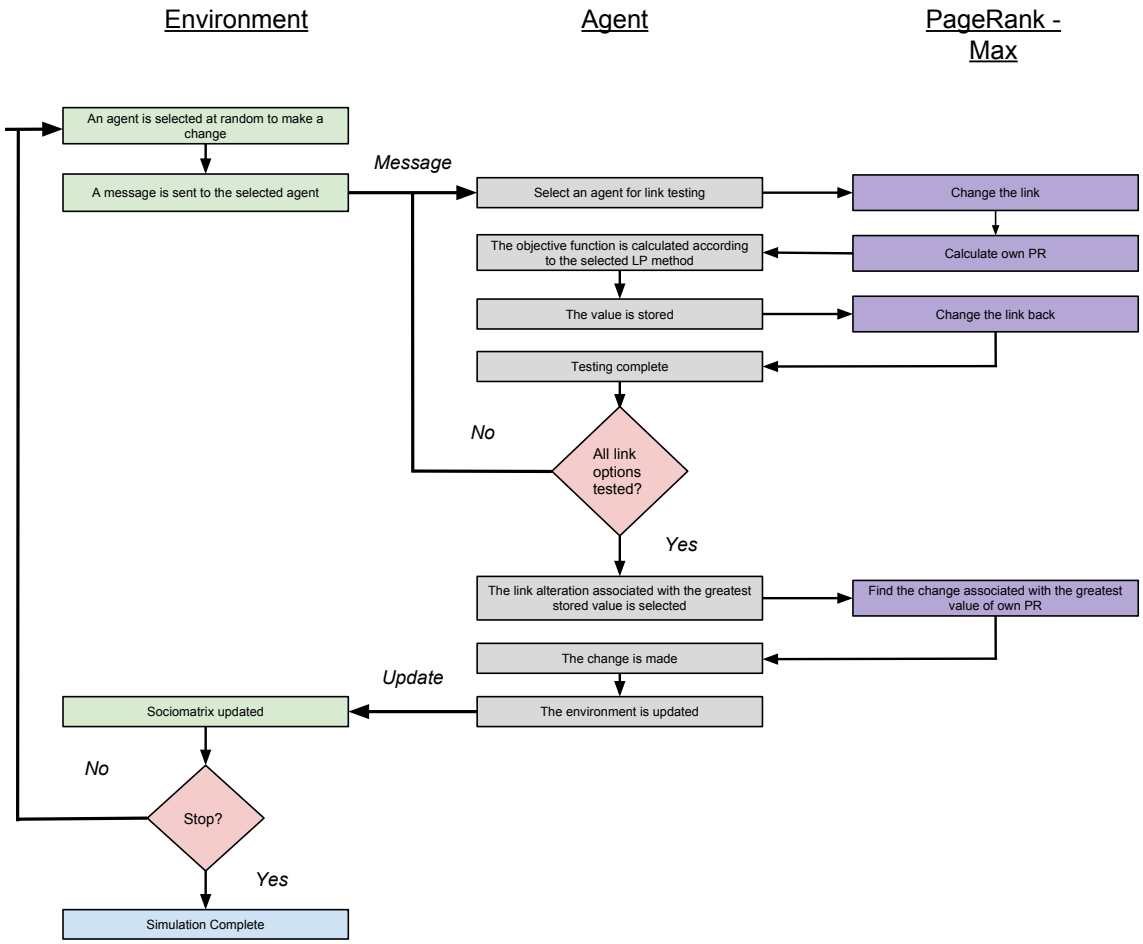


Figure 6.10: Updated simulation logic describing the process of the PR-Max method.

To calculate the PR vector, the dampening constant d is required. The value of d is generally selected to be 0.85, as this was the value originally used by Google (Brin & Page, 1998; Page & Brin, 1999). The current Google implementation of the PR algorithm (and the selected value of d) is said to have undergone many changes from the original work of Brin & Page (1998), the details of these changes have never been published (Langville & Meyer, 2004). Bressan & Peserico (2010) found that the value of d is particularly important in the calculation process, as it has the ability to alter ranking. For simplicity, this work will continue to use the original value of d - experimentation with the dampening constant occurring in Chapter 8.

As discussed, the PR of a webpage decides the ordering in which it is displayed on Google (following a search query). Users are said to be able to manipulate their webpage's PR by

making educated link choices (Avrachenkov & Litvak, 2004; Cheng & Friedman, 2006; Gyöngyi & Garcia-Molina, 2005; Malaga, 2008), with the PR-Max method aiming to demonstrate this in the context of social relations. This active improvement of PR, to the best of this author's knowledge, has not been implemented in terms of an LP method, with LP PR implementations generally following the structures of Liben-Nowell & Kleinberg (2007) and Lü & Zhou (2011).

Researchers have attempted link prediction through the use of a 'Personalised PageRank' (Chen, 2012; Yung, 2012), which orders pages differently depending on what a specific user may find more relevant (Haveliwala, 2003; Junchao et al., 2013; Walter et al., 2009). In terms of Link Prediction, this means that the PR is calculated differently depending upon the specific searching agent seeking to make a new connection; this calculation process does not consider optimising an agent's own PR. Furthermore, the 'game' of selecting links as a 'best response' to the current topology of a network are discussed by Chen et al. (2009) and Hopcroft & Sheldon (2008); however, the implications to LP are not discussed.

While the careful selection of outward links is said to be important, removal of specific links has also been shown to have an effect on PR (Bianchini et al., 2005; de Kerchove et al., 2008); this gives the PR-Max method a sensitivity to link disconnection. The AA, Katz and basic PR implementations do not demonstrate such explicit consideration of link disconnection, their focus being predominantly upon the prediction of new connections. Although the SAB method does account for disconnection, this is subject to the model generated prior to simulation. Therefore, the PR-Max method may be able to capture elements of network evolution more naturally.

6.3.2 Algorithm Overview

The implementations discussed both in Section 6.2 and Section 6.3, have provided an overview of the transference of basic LP methods into the logic of the created simulation. Some methods naturally adapt to the iterative agent objective function optimisation procedure (SAB and PR-Max), while others require a larger degree of interpretation. The inclusion of disconnection options have also been attempted in the conversion of each method, aiming to quantify the disconnective element of adolescent friendship evolution. While some original LP methods provide very little emphasis upon disconnection (Katz), PR-Max involves the process more heavily; the effect of this, and a comparison of the

methods, shall be discussed further on analysis of the SNS results (Chapter 7).

6.4 Simulation Overview

Section 6.2 described the simulation construction process and discussed the conversion of existing LP methods into the created framework, while Section 6.3 detailed the development of a new LP algorithm - created specifically for this research. This section describes the simulation procedures, giving a general overview of the prediction process. A brief discussion about the random method implemented for benchmark assessment is also included.

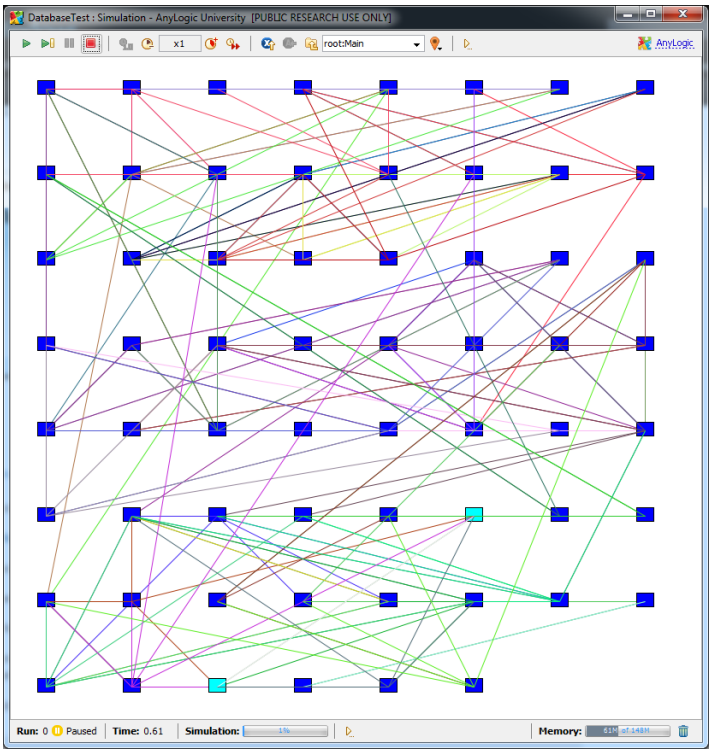


Figure 6.11: Simulation screenshot of a prediction running for School 40, with the associated network visualisation.

To make a prediction, the user must first decide upon the school number and time step required. The path to the database is already written within the code, with the initialisation process accessing the database and taking the required sociomatrix (X), inter-event time (ρ) and associated LP parameters (ϕ and β_k); the database path may be edited should an alternative database be required. The user is then required to select the prediction method

to be used; one of the five prediction options is selected and the simulation is run. A screenshot of the simulation running is visible in Figure 6.11, the connections between nodes being updated following each new link prediction. Once the process is complete, the simulation writes the resultant sociomatrix to a csv file; this can then be analysed and the predictions assessed.

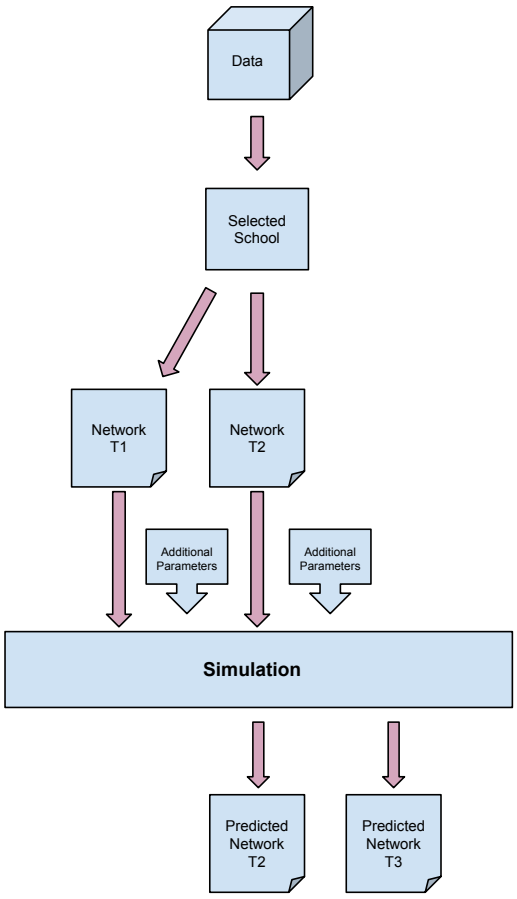


Figure 6.12: Diagram of the automated simulation process.

As a number of runs will be required, the process has been automated for simplicity. Therefore, the simulation reads in both initial sociomatricies (T_1 and T_2), along with the required additional parameters - running each prediction method in turn (for the required number of runs). Following each complete run, the simulation produces a predicted network for the relevant time step which updates the csv file accordingly; this is represented in Figure 6.12.

Finally, the simulation is also able to generate completely random predictions of each

school network. The same simulation process is adopted, with the time between network changes negatively exponentially distributed with parameter ρ; except the changing agent selects one of the testing agents to connect with uniformly at random. If a connection already exists with the randomly chosen agent, said connection is broken. The inclusion of the ability to perform random predictions is to aid the assessment of each method, giving a clear indication as to an improvement of an LP method over predictions at random; the assessment criteria are detailed further in Chapter 7. Following the completion of the simulation, validation of the model is required (Section 6.5).

6.5 Validation

Before the SNS may be used for prediction, it must undergo validation procedures; this is to ensure that the model reflects the situation being simulated, and the project requirements have been met (Robinson, 1994). The areas addressed in the validation process are: verification (Section 6.5.1), timing (Section 6.5.2), distributions and random sampling (Section 6.5.3), warm-up period (Section 6.5.4) and number of runs (Section 6.5.5). Each topic is discussed further below, informed by the validation procedures of Robinson (1994) and Pidd (2004).

6.5.1 Verification

Verification is described as a micro-check of the model, where a test of each individual element is performed. During the creation process, regular checks of the code were carried out - attempting to ensure the proper implementation of the designated logic. For each of the LP method implementations, the associated calculation of the objective function was performed in R - checking that the calculations matched. Each time a new calculation was implemented, it would be printed to screen and checked for accuracy. The network visualisation was also used to assert that the correct connections had been made, following an LP calculation.

6.5.2 Timing

A check of the 'timing', takes into consideration both the timing of events and the overall model run length. The timing of events is dictated by ρ, as discussed in Section 6.2.1. The

selection of ρ is taken directly from the network being simulated, calculated by the number of tie changes exacted between consecutive waves of network data. As this is calculated directly from the data, it can be assumed to be an accurate reflection; however, the number of changes that are made and then reversed in this period are unknown, as is the order and timing in which they occurred. The time between events is therefore assumed to be random, following a negative exponential distribution with parameter ρ - following the convention of SAB modelling.

The overall model run time is dictated by the data. There is roughly one year between consecutive waves of network data, therefore a prediction of the network one year later is required. AnyLogic requires the user to apply units to each time step of the simulation; this has been selected to be weeks, with the average number of tie changes per week calculated. Each simulation time step represents one week, with the model being run for 52 weeks.

6.5.3 Distributions and Random Sampling

A negative exponential distribution of the time between events in the network is required, the values being sampled from AnyLogic's own built in options. Sampling of random numbers in this process is from AnyLogic's default random number generator, which is an instance of the 'Random' Java class; this being a Linear Congruential Generator (AnyLogic, 2002). During the verification process, a number of runs were performed to assess the average number of changes in a selected school network; the confidence interval was calculated, and as the actual number of changes from the data fell within the bounds of the confidence interval, the distribution was said to be acting appropriately.

Other distributions within the model, are those necessary in calculations of the respective objective functions (such as the SAB probability of changing a link, Section 6.2.2). As these distributions vary based on the changing agent, and the formation of the network, the code has been written explicitly within the simulation. Once again, during the verification procedures, the construction of these distributions was checked and found to be working as expected. The random number generator used in these processes is again that of the default within AnyLogic, which has been assumed suitable for the work of this thesis.

6.5.4 Warm-Up Period

The starting conditions of the simulation (for a selected school at a given time point) are provided by the initial sociomatrix, which is read during the initialisation procedure. As such, a warm-up period is not required, as the agents begin with the required set up of connections.

6.5.5 Replications

The final validation topic, centres around the number of replications selected for simulation. As the simulation has various elements which include variability, such as the selection of the changing agent, a number of runs are required. Robinson (2004) details the confidence interval (CI) approach, which makes use of outcome-based precision criteria. Using the CI method, the required number of runs (η) is calculated as:

$$\eta = \left(\frac{100 \cdot S \cdot t_{(n-1,\alpha/2)}}{\widehat{d} \cdot \bar{x}} \right) \tag{6.19}$$

where $\bar{x}$ and S are the sample mean and standard deviation (respectively), $\widehat{d}$ the desired percentage deviation of confidence about the mean, and $t_{n-1,\alpha/2}$ from the standard t-distribution with $n-1$ degrees of freedom and significance level α (Robinson, 2004).

During the analysis of results (Chapter 7), four network measures shall be used to assess the accuracy of the produced networks: transitivity, average degree, reciprocity and average path length. As these values are the 'outcomes' of the simulation, the precision of these criteria can be can be used to calculate η; therefore, a number of preliminary test runs are required to perform the CI procedure.

The network of school 76 (at T_1) has been selected for testing, due to its large population. Given the greater number of agents in school 76, the LP methods may have a larger number of choices when selecting link changes. As such, greater variability may then be introduced into the predictions - hence its selection as the test network . The school 76 network is simulated, generating a prediction of each assessment criteria for T_2. Table 6.1 displays the required number of runs (η) to obtain 5% deviation about the mean, with a significance level $\alpha = 0.05$ and 9 degrees of freedom; the values have been generated for 10 test

	Random	AA	Katz	SAB Model	PR	PR-Max
Transitivity	1.09	0.97	1.05	1.56	1.30	4.07
Average Degree	0.64	1.00	0.80	0.88	0.91	1.15
Reciprocity	1.42	1.09	1.04	1.24	0.99	1.64
APL	9.49	6.54	0.68	8.55	4.14	1.97

Table 6.1: The required number of runs for 5% deviation, from 10 test runs.

simulation runs. From Table 6.1, the APL of the Random method requires the greatest number of runs (9.49), with Random average degree requiring the lowest (0.64).

As a 'rule of thumb', Law & Kelton (1999) suggest a minimum of around 3-5 replications are required; should too many replications be selected, this wastes valuable running time and computing resources. Given that the maximum required for school 76 is 9.49 replications, 10 replications have been selected. This is greater than the rule of thumb, but does not appear excessive.

6.5.6 Validation Overview

With each validation issue addressed in turn, the validation procedure appears complete. The verification process throughout the creation of the simulation have micro-checked the model, ensuring the distribution and coding logic reflect the intended procedures. The data has addressed the issues of run length and warm up period, while a test set of simulations has resulted in the decision of replication number. Therefore, the simulation of each school (at each prediction timestep), shall have a run length of one year, with no warm-up period and be replicated 10 times.

One further issue of validation, requires the assessment of the processes included in the model; in terms of the simulation created, this would be the processes by which friendships evolve. In the current work, this cannot be addressed as a validation issue, as this is the *reason* for the creation of the simulation - to further understand which processes accurately capture the dynamic of adolescent friendship. It is therefore the results of Chapter 7, that shall assess the accurate reflection of the model processes; the simulation predictions being validated against the real ASSIST social networks.

6.6 Chapter Summary

This chapter has introduced the social network simulation (SNS), discussed the methods implemented within it and validated the model accordingly. Section 6.1 presented the original LP methods selected for comparison with the newly developed PageRank-Max algorithm:

- Adamic Adar;
- Katz;
- Stochastic Actor Based Modelling;
- PageRank.

Each method focuses upon a different criteria in the prediction and analysis of links, borne from their respective origins in the literature. An overview of each method, an example (where applicable) and relevant literature was discussed; the section also providing justification for the inclusion of each method within the SNS.

Section 6.2 documented the creation of the SNS, describing each stage in the simulation construction process. The SNS is able to make a prediction for each school, at each prediction period, through agent-based decisions - governed by a discrete-event time structure. A framework of the agent logic was provided, expanded upon through the discussion of each individual existing LP method implementation.

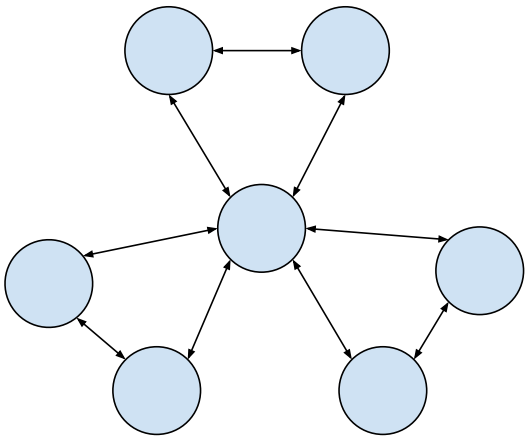
Section 6.3 provided details of the new PageRank based algorithm specifically developed for this thesis, PageRank-Max. The algorithm focuses upon the specific optimisation of an agents eigen-centrality, centrality previously identified as a key element in behavioural diffusion within a social network. An outline of the algorithm logic was presented, and an overview of all LP methods included in this work was given.

Section 6.4 described the developed simulation procedures, providing an overview of the prediction process. This section also gave information regarding the ability of the SNS to produce random predictions, necessary for the assessment process of Chapter 7. Therefore,

the list of LP methods implemented within the simulation is:

- Adamic Adar;
- Katz;
- Stochastic Actor Based Modelling;
- PageRank;
- PageRank-Max;
- Random.

Finally, Section 6.5 discussed the validation procedures necessary for appropriate use of the SNS. The issues of micro-check verification, timing, distributions, random sampling, warm-up period and replications were covered - the section discussing them accordingly. The outcome of the validation procedure is that the SNS is to adopt a granularity of weeks, with a run length of 52 weeks - 10 replications of the simulation being required. An analysis of the results produced by the SNS is discussed in Chapter 7.



-"The Friendship Graph"

7

SNS Results

The previous chapter (Chapter 6) described the creation of an ABS to predict social network evolution (termed the SNS), implementing four separate Link Prediction (LP) methods - Adamic/Adar (AA), Katz, Stochastic Actor Based (SAB) Models and PageRank (PR) - to compare with the newly developed PageRank-Max (PR-Max) algorithm. This chapter discusses the results produced from the SNS, evaluating each of the LP methods selected, across the breadth of ASSIST network school data.

For each of the 18 network schools presented in Chapter 5, a prediction is made from T_1 to T_2 and T_2 to T_3. The predicted networks of T_2 and T_3, generated from the SNS, shall be compared with the real data to evaluate their accuracy. The presentation of results is structured as follows: the **precision** of each algorithm in predicting the correct links is discussed in Section 7.1; the individual network **structures** produced by the SNS are presented in Section 7.2; and Section 7.3 focuses upon a comparison of the algorithms in reference to predicting control and intervention schools.

7.1 Precision Analysis

The first method to evaluate the T_2 and T_3 predictions made by the SNS, is that of *precision*. The precision metric was first proposed by Cleverdon (1972) and has been used in the context of both LP methods (Liben-Nowell & Kleinberg, 2007; Lü & Zhou, 2011) and recommender systems (Herlocker et al., 2004; Lü et al., 2012). In the context of LP, precision evaluates the number of *correct* predictions, y_c, relative to the number of predictions made, y_p, such that the precision is $\frac{y_c}{y_p}$.

In the following discussion of results, the precision is calculated for each of the predicted networks (for each LP method introduced in Section 6.2.2). As discussed in Section 6.4, the SNS also has the capability to generate a network based upon link predictions at random (the random method). To benchmark the precision of the predicted network, values are expressed as a percentage improvement over predictions made at random; positive values indicate an improvement in correct predictions, while negative values indicate a reduction. Ten runs of the random method for each school network are performed to generate the random predictions, this follows the suggested number of runs calculated for the other LP methods.

Also of interest is the number of *missed* predictions, which examines the number of friendship changes not made in the predicted networks of T_2 and T_3, when a friendship change has actually occurred in the real data. The missed predictions are also expressed in terms of an increase compared to the random method, negative values indicating fewer predictions missed. Therefore, two metrics are calculated for each predicted network: the percentage increase of *correct* and *missed* link predictions over the random method.

Tables 7.1 and 7.2 display control and intervention prediction values at T_2 (respectively), while Tables 7.3 and 7.4 display predictions at T_3. Values that are significantly different from random at the 0.05 level, following an independent samples t-test for parametric data or a Mann-Whitney test for non-parametric data, are highlighted and starred. Control and intervention schools shall now be examined separately at each time step, beginning with control schools at T_2.

Method	Measure	15	22	33	35	40	41	62	63	64	68	69	71
Adamic/Adar	Correct	5.54∗	8.71∗	8.51∗	7.32∗	12.85∗	7.48∗	10.54∗	8.75∗	8.24∗	4.13∗	4.69∗	5.49∗
	Missed	−4.48∗	−8.15∗	−7.35∗	−7.34∗	−13.35∗	−6.94∗	−9.98∗	−8.67∗	−7.28∗	−3.17∗	−5.08∗	−4.94∗
Katz	Correct	4.48∗	7.38∗	6.81∗	7.35∗	5.34∗	6.27∗	4.58∗	7.04∗	3.76∗	2.69∗	2.06∗	4.12∗
	Missed	−2.75∗	−5.40∗	−3.77∗	−4.86∗	−4.06∗	−4.23∗	−4.24∗	−4.57∗	−2.37∗	−1.96∗	−0.99	−2.13∗
SAB Model	Correct	17.08∗	3.94∗	3.45∗	17.29∗	13.91∗	7.07∗	24.49∗	−0.48∗	3.11∗	12.51∗	15.06∗	2.88∗
	Missed	−9.36∗	−3.20∗	−2.61∗	−10.57∗	−8.13∗	−5.85∗	−15.21∗	0.53∗	−2.22∗	−6.28∗	−8.47∗	−2.13∗
PageRank	Correct	−0.31	0.09	0.65	0.95∗	1.37∗	0.43	0.20	0.26	0.53	0.85∗	2.17∗	0.42
	Missed	0.30	−0.16	−0.44	−0.96∗	−1.07∗	−0.41	−0.09	−0.25	−0.41	−0.70∗	−1.95∗	−0.29
PageRank-Max	Correct	39.63∗	40.69∗	40.89∗	41.32∗	42.80∗	39.60∗	42.01∗	44.81∗	40.45∗	35.46∗	34.40∗	42.18∗
	Missed	−22.85∗	−23.83∗	−24.60∗	−28.08∗	−23.30∗	−23.53∗	−26.70∗	−29.19∗	−24.12∗	−16.94∗	−21.66∗	−22.63∗

Table 7.1: Control school precision at T_2, expressed as a percentage increase over random prediction. Highlighted and starred values indicate a significant difference between the associated LP method and predictions at random.

		12	32	34	73	74	76
Adamic/Adar	Correct	12.41∗	1.33∗	7.74∗	1.85∗	6.91∗	2.17∗
	Missed	−10.67∗	6.65∗	−7.96∗	−3.68∗	−6.38∗	−3.07∗
Katz	Correct	5.82∗	−0.11	2.44∗	0.94∗	0.24	−0.24∗
	Missed	−5.07∗	−0.56∗	−2.55∗	−2.09∗	−0.37	−0.14
SAB Model	Correct	21.60∗	0.47∗	17.77∗	1.26∗	35.00∗	2.96∗
	Missed	−13.56∗	0.00	−11.26∗	−1.34∗	−19.50∗	−2.18∗
PageRank	Correct	1.07∗	0.18	0.73∗	−0.02	1.66∗	0.26∗
	Missed	−0.95∗	−0.20	−0.76∗	0.01	−1.36∗	−0.11
PageRank-Max	Correct	38.53∗	9.67∗	34.52∗	9.84∗	46.59∗	24.51∗
	Missed	−25.51∗	−0.78∗	−22.10∗	−8.20∗	−29.65∗	−17.97∗

Table 7.2: Intervention school precision at T_2, expressed as a percentage increase over random prediction. Highlighted and starred values indicate a significant difference between the associated LP method and predictions at random.

Method	Measure	15	22	33	35	40	41	62	63	64	68	69	71
Adamic/Adar	Correct	7.51∗	8.17∗	0.68∗	3.87∗	1.70∗	10.05∗	8.28∗	8.44∗	4.69∗	5.66∗	6.32∗	6.14∗
	Missed	−6.80∗	−7.26∗	11.39∗	−3.37∗	−1.64∗	−7.79∗	−7.59∗	−7.70∗	−4.30∗	−4.63∗	−5.11∗	−5.93∗
Katz	Correct	8.37∗	7.87∗	−0.17	1.15∗	3.12∗	9.64∗	4.37∗	7.78∗	6.64∗	2.31∗	5.35∗	6.37∗
	Missed	−5.74∗	−5.46∗	−0.51	−1.06∗	−1.64∗	−5.69∗	−3.79∗	−4.81∗	−3.89∗	−1.83∗	−2.86∗	−3.53∗
SAB Model	Correct	2.32∗	28.10∗	5.67∗	2.18∗	16.39∗	−1.60∗	24.19∗	2.26∗	12.97∗	8.22∗	20.92∗	0.28
	Missed	−1.99∗	−15.96∗	−0.38	−1.55∗	−7.27∗	1.24∗	−15.21∗	−1.86∗	−7.38∗	−5.86∗	−11.56∗	0.11
PageRank	Correct	0.79∗	0.63	0.34∗	0.63∗	0.93	−0.35	0.39	0.35∗	0.48	0.64∗	2.67∗	0.30
	Missed	−0.66∗	−0.53	−0.57	−0.53∗	−0.76	0.26	−0.29	−0.28∗	−0.45	−0.46∗	−1.84∗	−0.28
PageRank-Max	Correct	43.04∗	46.64∗	7.19∗	47.22∗	44.67∗	46.01∗	41.94∗	47.95∗	50.69∗	46.33∗	42.50∗	50.24∗
	Missed	−26.62∗	−29.96∗	−0.88∗	−30.31∗	−23.49∗	−28.48∗	−27.51∗	−33.50∗	−34.35∗	−27.22∗	−26.70∗	−30.56∗

Table 7.3: Control school precision at T_3, expressed as a percentage increase over random prediction. Highlighted and starred values indicate a significant difference between the associated LP method and predictions at random.

		12	32	34	73	74	76
Adamic/Adar	Correct	6.29∗	8.84∗	1.43∗	1.22∗	−0.44∗	4.96∗
	Missed	−5.38∗	−7.78∗	−2.55∗	−2.06∗	0.26	−5.20∗
Katz	Correct	3.24∗	3.91∗	0.11	1.04∗	0.01	1.26∗
	Missed	−2.64∗	−3.36∗	−0.17	−1.99∗	−0.33	−1.41∗
SAB Model	Correct	8.47∗	1.21∗	2.23∗	17.14∗	9.59∗	1.08∗
	Missed	−5.93∗	−1.00∗	−1.95∗	−11.46∗	−0.44	−0.88∗
PageRank	Correct	0.39	0.99∗	0.39∗	0.23	0.38∗	0.32
	Missed	−0.31	−0.85∗	−0.31	−0.51∗	0.09	−0.38∗
PageRank-Max	Correct	49.08∗	46.30∗	20.10∗	22.93∗	7.19∗	42.54∗
	Missed	−28.99∗	−29.52∗	−18.14∗	−17.82∗	0.74	−29.03∗

Table 7.4: Intervention school precision at T_3, expressed as a percentage increase over random prediction. Highlighted and starred values indicate a significant difference between the associated LP method and predictions at random.

Control Schools at T_2

From Table 7.1 the AA method produces a significant increase in correct predictions for all control schools, and a significant reduction in missed predictions - when compared with random predictions. This demonstrates that the AA method performs significantly better at predicting the evolution of friendships, than predictions at random. Similarly, the Katz method also significantly improves in the number of correct predictions, however, the percentage of missed school 69 predictions is not a significant reduction (-0.99%) - Katz being the worst performing method in terms of school 69 network predictions.

The SAB method values are all significantly different from random (Table 7.1), with school 63 observing a reduction in the percentage of correct predictions (-0.48%); this is contrary to all other SAB 'correct' prediction values, which demonstrate a significant increase. The poor performace of the SAB method upon the school 63 network is not emulated by the PR-Max method (44.81%); the method producing the greatest improvement in correct predictions of any control network method (at T_2). The differing levels of school 63 network precision (between the PR-Max and SAB methods), indicate that particular LP methods may be more naturally suited to certain schools.

Recall that each LP method gives particular emphasis to specific linking characteristics (Section 6.1.5), with SAB focusing on optimising selected network statistics (density, reciprocity, etc.) and PR-Max seeking to optimise an agent's eigen-centrality. It would therefore appear that, with regard to the individuals in school 63, the process of improving their eigen-centrality captures friendship evolution more accurately. For example, the SAB method performs better upon schools 15 (17.08%), 35 (17.29 %) and 62 (24.49 %) than upon school 63 (-0.48%), indicating the SAB model may reflect elements of the friendship evolution present in these schools more accurately; however, the PR-Max method still offers greater increases to precision (15: 39.63%, 35: 41.32%, 62: 42.01%) overall in these schools.

While the PR-Max method appears to perform well across all schools, the standard PR precision values are considerably lower - many of the schools not exhibiting a significant improvement over the random method. The schools which display a significant increase to correct predictions for PR (35: 0.95%, 40: 1.37%, 68: 0.85% and 69: 2.17%), exhibit only small improvements - being surpassed by all other prediction methods. This indicates, in

terms of the ASSIST adolescent friendships, individuals may be interested in improving their own PageRank (PR-Max), but not necessarily concerned with linking to those with a high PageRank.

Intervention Schools at T_2

From Table 7.2, a similar trend is observed in intervention schools, to that of control schools at T_2. The PR-Max method performs significantly better than the random method upon every school network, also outperforming all other LP methods in terms of correct and missed predictions. PR-Max 'correct' values for school 32 (9.67%) and school 73 (9.84%) are lower than the remaining intervention schools, but the precision of these schools amongst other methods is also reduced. This indicates that all LP methods do not greatly outperform random link predictions in school 32 and school 73, suggesting some underlying nuances within these social networks.

It may be the case that random predictions perform particularly well in schools 32 and 73, meaning that the LP methods cannot greatly improve upon them; as such, some network features of the schools may be causing links to form naturally at random. However, a further reasoning may be the proposed LP methods do not accurately capture the linking criteria of these schools - other unexplored processes potentially being important in friendship selection. Examining the attributes presented in Table 5.1, the reduced performance may be a result of both schools having large populations (32: 229, 73: 199). A larger population gives the "changing agent" a greater pool of agents to select from when making a link prediction, potentially increasing scope for error; issues regarding network size, in reference to prediction accuracy, are discussed further in Section 7.2.3.

A further notable prediction is produced by the SAB method upon School 74, demonstrating a 35.00% significant increase in correct predictions. School 74 is a valley school (Table 5.1) which becomes more cliqued over time, with transitivity and reciprocation increasing at T_2 (as discussed in Section 5.2.3). The features of School 74 may be the reason for the SAB model's high performance, with the basic SAB model placing particular importance upon reciprocation and distance of actors; however, the SAB prediction is superseded once again by the PR-Max method (46.59%). Overall, the strength of the PR-Max method appears consistent in intervention schools at T_2.

Control Schools at T_3

Table 7.3 displays the precision values for control schools at T_3. The PR-Max method appears to perform well once again, for example the precision of school 64 rising from 40.45% at T_2 to 50.69% at T_3. Evidently, the LP algorithms exact the same logic across time steps, therefore such an increase may indicate that the PR-Max method captures the process of link evolution better at T_3 than at T_2 (for school 64). Conversely, school 33 exhibits a reduction in PR-Max precision at T_3, reducing from 40.89% (T_2) to 7.19% - suggesting that individual eigen-centrality consideration may not as appropriately capture friendship evolution in school 33 (between T_2 and T_3).

The figures highlighted, demonstrate that behavioural differences in friendship selection may be apparent between time steps - causing algorithms to perform disparately upon school networks at different time periods. From the data analysis of Chapter 5, school 33 was not highlighted as behaving substantially differently to other schools in terms of both network characteristics and smoking uptake. Therefore, this also suggests that LP algorithms may be able to detect subtleties within the network evolution, that conventional analysis has not uncovered.

The precision results of T_3 control schools also demonstrate the variability of the SAB method, with school 41 predictions being significantly worse than random (-1.60%) and school 71 predictions indicating no significant improvement (0.28%). This variability may be due to the basic SAB model selected not accurately capturing the dynamics of network evolution, with an alternative model potentially performing better; this is a weakness of SAB models, as the process can require extensive investigation into appropriate model parameters (Carrington, 2005; Lospinoso & Schweinberger, 2011). In contrast, the PR-Max method does not require extensive manipulation and still outperforms all other methods.

Intervention Schools at T_3

A difference in school predictions between timesteps is once again observed. Taking the example of school 32 PR-Max precision, this has risen from 9.67% at T_2 (Table 7.2) to 46.30% at T_3 (Table 7.4) - suggesting that friendship evolution between T_2 and T_3 is reflected better by eigen-centrality optimisation (than T_1 to T_2). School 74 also observes an alteration in precision at T_3, reducing from 46.59% at T_2 to 7.19% (PR-Max). Preci-

sion for each LP method upon school 74 is substantially reduced, such that even the AA method produces a significant reduction in precision (-0.44%) - the AA method otherwise consistently producing significantly improved predictions.

The SAB method performs better than PR-Max upon the network of School 74 (9.59%) at T_3, with the method also performing well upon school 74 at T_2 (35.00%); this indicates that the basic SAB model structure may capture the nature of friendship evolution in school 74 appropriately. The successful SAB model prediction demonstrates that, with an underlying objective function representative of network specific link evolution, SAB predictions may outperform those of PR-Max.

7.1.1 Precision Overview

To produce an overall ranking for each LP method in terms of precision measures, the average percentage increase in correct and missed predictions is calculated. Table 7.5 and Table 7.6 display the average precision of control and intervention schools (respectively), classified by LP method at each timestep. Each method is then ranked in terms of their precision performance, control ranks are displayed in Table 7.7 and intervention in Table 7.8. Finally, the harmonic mean of the ranks (in a given time period) is calculated for each method, producing an overall ranking - displayed in Table 7.9.

From Table 7.9 it is evident that PR-Max is the highest ranked method at each timestep, for both control and intervention schools. When the ranks of 'correct' and 'missed' predictions are aggregated (with equal importance), the SAB model and AA method's ranks are equivalent (for control schools at T_2). Table 7.5 indicates that, while the SAB model produces more correct predictions (over random) than AA (SAB: 10.03%, AA: 7.69%), the number of missed prediction for AA is reduced (SAB: -6.12%, AA: -7.23%); this accounts for the equivalent rankings. A similar situation occurs in intervention schools at T_3, with the AA and SAB methods obtaining the same ranking.

Examining the PR-Max values of Table 7.5, the average percentage of correct predictions in control schools (over random) increases from T_2 (40.35%) to T_3 (42.87%) - this increase being statistically significant. The decrease in missed PR-Max predictions (for control schools) from T_2 (-23.95%) to T_3 (-26.63%) is also significant. Such figures indicate that, over time, the PR-Max predictions are improving in terms of precision (for

Time	Measure	Adamic/Adar	Katz	SAB Model	PageRank	PageRank-Max
T_2	Correct	7.69	5.16	10.03	0.63	40.35*
	Missed	−7.23	−3.44	−6.12	−0.54	−23.95*
T_3	Correct	5.96	5.23	10.16	0.65	42.87*
	Missed	−4.23	−3.40	−5.64	−0.53	−26.63*

Table 7.5: Average of all control school networks at T_2 and T_3, displaying the percentage increase over random predictions. Highlighted values indicate a significant difference between time steps.

Time	Measure	Adamic/Adar	Katz	SAB Model	PageRank	PageRank-Max
T_2	Correct	5.40	1.52	13.17	0.65	27.28
	Missed	−4.18	−1.80	−7.97	−0.56	−17.37
T_3	Correct	3.71	1.60	6.62	0.45	31.36
	Missed	−3.78	−1.65	−3.61	−0.38	−20.46

Table 7.6: Average of all intervention school networks at T_2 and T_3, displaying the percentage increase over random predictions.

Time	Measure	Adamic/Adar	Katz	SAB Model	PageRank	PageRank-Max
T_2	Correct	3	4	2	5	1
	Missed	2	4	3	5	1
T_3	Correct	3	4	2	5	1
	Missed	3	4	2	5	1

Table 7.7: Ranked average precision values for control schools.

Time	Measure	Adamic/Adar	Katz	SAB Model	PageRank	PageRank-Max
T_2	Correct	3	4	2	5	1
	Missed	3	4	2	5	1
T_3	Correct	3	4	2	5	1
	Missed	2	4	3	5	1

Table 7.8: Ranked average precision values for intervention schools.

Type	Time	Adamic/Adar	Katz	SAB Model	PageRank	PageRank-Max
Control	T_2	2.4	4.0	2.4	5.0	1.0
	T_3	3.0	4.0	2.0	5.0	1.0
Intervention	T_2	3.0	4.0	2.0	5.0	1.0
	T_3	2.4	4.0	2.4	5.0	1.0

Table 7.9: Harmonic mean of ranks for each method, both control and intervention schools documented.

control schools). As the PR-Max method optimises an individual's eigen-centrality, it may be that students within the control schools become more concerned about their position in the network as they get older; thus seeking to become more central, leading to an improvement in PR-Max predictions. Table 7.6 shows that intervention school PR-Max values also increase from T_2 (27.28%) to T_3 (31.36%), however, the difference is not statistically significant.

The boxplots of Figures 7.1 and 7.2 display the raw, correct and missed prediction scores at T_2 respectively, with the plots presenting data from the control and intervention schools together. The boxplots demonstrate the higher proportion of correct predictions, and lower proportion of missed predictions, for the PR-Max method when compared with all other selected LP methods. The T_3 boxplots also demonstrate the increased precision of the PR-Max method (Figures 7.3 and 7.4), reinforcing the discussed precision accuracy of the PR-Max method (in relation to the other LP methods).

Overall, the precision analysis has highlighted a number of key outcomes with regard to the LP methods tested, summarised as follows:

- PR-Max is (in general) the LP method which performs the best in terms of increasing correct predictions, and decreasing missed predictions;

- All LP methods experience variability in their performance, with certain LP methods capturing school-specific network evolution more accurately - potentially a result of the school's underlying friendship mechanisms;

- There are a number of schools in which the LP methods perform poorly (e.g, school 32 at T_2, school 73 at T_2, school 33 at T_3 and school 74 at T_3), a result of particularly strong random predictions, or the inability of the selected LP methods to capture important aspects of the schools' linking process;

- The underlying mechanism by which links evolve, may change over time - as demonstrated by the greatly varying precision of schools 32, 33, 73 and 74 between T_2 and T_3;

- The PR-Max observes a significant increase in overall average precision in control schools at T_3 from T_2, adding further weight to the notion of time sensitivity in

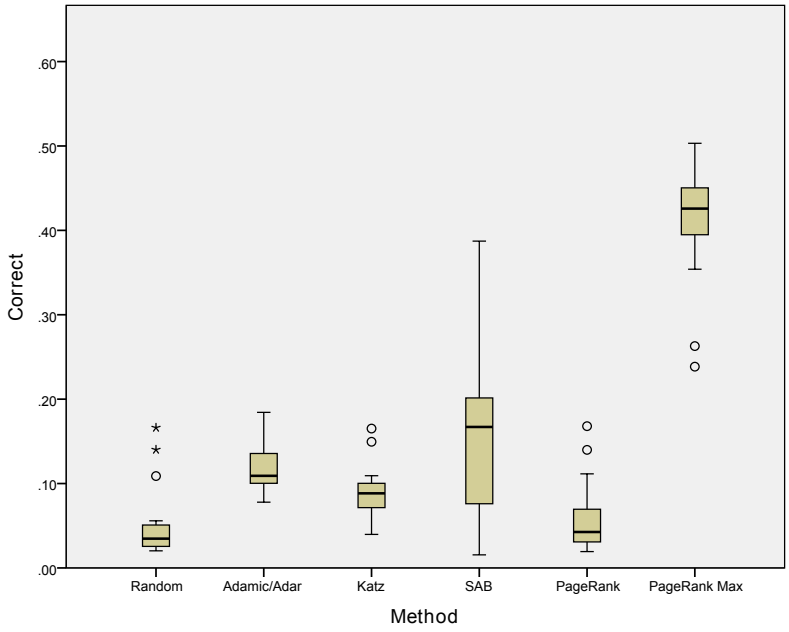


Figure 7.1: Box plot of correct prediction proportions for each method at T_2. Whiskers extend 1.5 times the height of the box, with circular points indicating outliers. Starred points indicate extreme outliers.

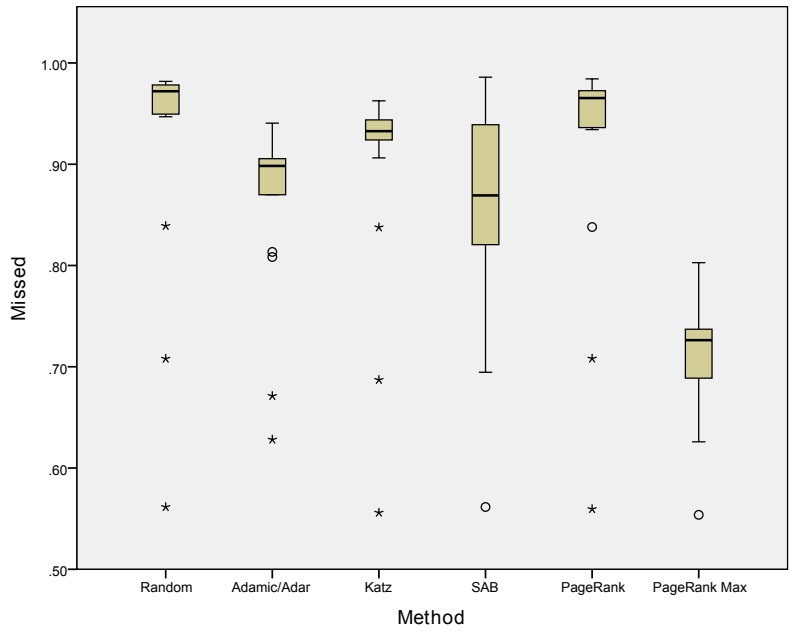


Figure 7.2: Box plot of missed prediction proportions for each method at T_2. Whiskers extend 1.5 times the height of the box, with circular points indicating outliers. Starred points indicate extreme outliers.

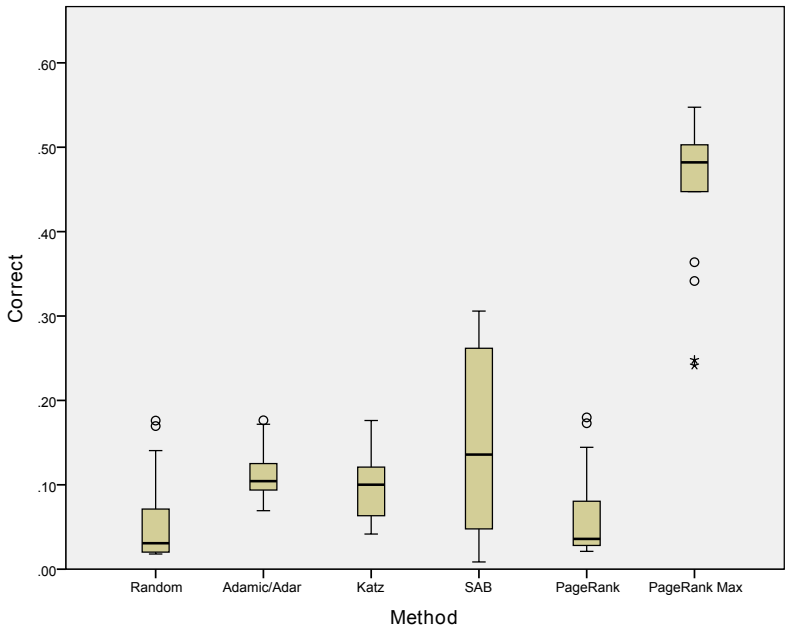


Figure 7.3: Box plot of correct prediction proportions for each method at T_3. Whiskers extend 1.5 times the height of the box, with circular points indicating outliers. Starred points indicate extreme outliers.

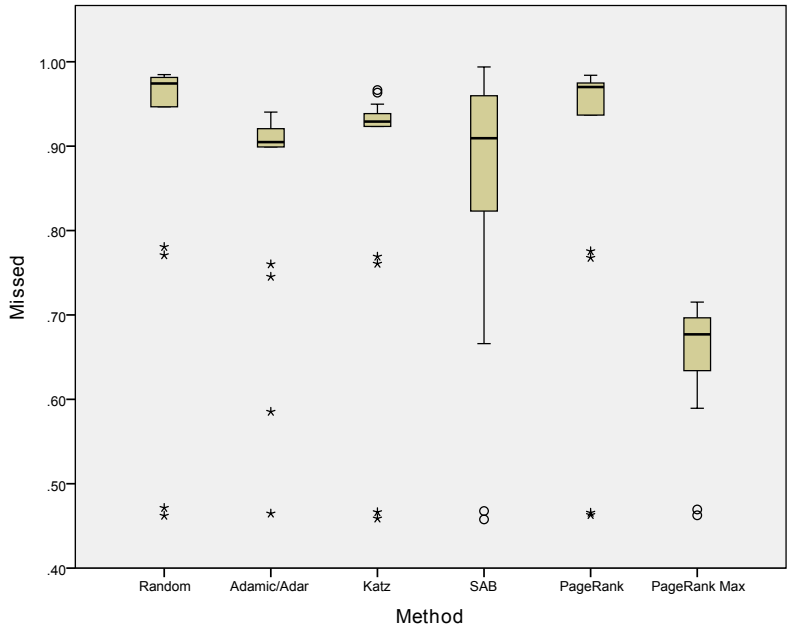


Figure 7.4: Box plot of missed prediction proportions for each method at T_3. Whiskers extend 1.5 times the height of the box, with circular points indicating outliers. Starred points indicate extreme outliers.

friendship evolution - the eigen-centrality of a student potentially becoming more important as they get older.

The conclusions presented, not only give an account of the performance of the LP methods, but also present an insight into the friendship evolution processes of adolescents. A comparison of control and intervention precision measures may also offer a further understanding of behaviour relating to the ASSIST investigation; this shall be further explored in Section 7.3. The precision metrics specified, detail the accuracy of a method in terms of forecasting individual friendship changes; however, they do not give a representation of the overall network structure. To investigate this further, the network structure analysis of Section 7.2 is conducted.

7.2 Network Structure Analysis

The analysis of Section 7.1 examined the precision of the selected LP methods, identifying the percentage improvement over the random method in predicting *specific* link evolution. However, at the current stage of investigation, individuals within the prediction networks do not have behavioural attributes - the LP methods using only information relating to social network structure to make predictions. As such, there may be agents who reside in equivalent positions within the network, with the LP methods being unable to differentiate between them. The result of this may be a predicted network with low precision accuracy, but with an overall network structure that is an appropriate representation of the real data.

To investigate the overall predicted network structure further, four network characteristics have been selected to quantify the predicted structures:

- transitivity;
- average out-degree;
- reciprocity;
- average path length (APL);

The metrics selected are those generally used to assess the structure of networks, each

being previously outlined in Chapter 3. The subsequent analysis is structured as follows: the 'effect size' is introduced in Section 7.2.1; an analysis of the average effect size (AES) of each method is presented in Section 7.2.2; finally an analysis of school specific AES is discussed in Section 7.2.3.

7.2.1 Effect Size

To analyse the predicted network structures, the output of the 10 simulation runs (for each school, at each timestep, for every LP method) are compared with the structural values from the data. The metrics selected for the analysis (transitivity, average degree, reciprocity and APL) are not on the same scale as each other; as such, meaningful comparisons between metrics is not intuitive. To rectify this issue, a new approach to network comparison in employed in this thesis - making use of 'effect size'.

The effect size is a measure that represents the magnitude of a relationship, quantifying the difference between two groups; it is the central component of a meta-analysis, which attempts to summarise the finding of multiple investigations (Hedges & Olkin, 1985). The effect size used for this analysis is Glass' Δ, calculated as:

$$\Delta = \frac{\bar{x_d} - \bar{x_p}}{s_p} \tag{7.1}$$

where $\bar{x_d}$ and $\bar{x_p}$ are the mean values of a metric from the data and predicted networks respectively, and s_p is the associated predicted network standard deviation. $\bar{x_p}$ and s_p are calculated from the 10 simulation runs, while $\bar{x_d}$ is taken directly from the data (previously calculated in Section 5.2.3). Tables 7.10 and 7.11 display the calculated effect sizes for control and intervention schools at T_2 (respectively), and Tables 7.12 and 7.13 display the control and intervention values at T_3.

A positive effect size indicates an underestimation by the LP method, as the predicted metric value is greater than that of the true network; a negative effect size represents an overestimation. An effect size close to zero indicates a small prediction "effect", with the predicted network structural value being close to that of the original data value. To determine whether the effect size is significant, the appropriate one-sample t-tests (paramteric data) or Wilcoxon signed-rank (non-parametric) tests are calculated - comparing the output of the 10 simulation runs, to the associated value from the data. In Tables 7.10, 7.11,

		15	22	33	35	40	41	62	63	64	68	69	71
Adamic/Adar	Transitivity	−7.00	−14.52	−9.90	−12.34	−7.69	−9.53	−15.25	−13.56	−9.23	−14.60	−16.59	−11.43
	Average Degree	−27.47	−24.62	−21.20	−31.34	−15.50	−19.76	−22.59	−35.21	−22.69	−9.46	−17.43	−19.32
	Reciprocity	−8.12	−8.24	−6.98	−8.30	−0.56	−5.76	−19.08	−10.71	−3.83	−5.37	−7.67	−2.73
	APL	5.30	18.93	0.74	1.89	13.90	13.73	28.92	33.66	10.52	3.75	21.83	5.28
Katz	Transitivity	4.18	−0.71	2.76	−0.99	2.30	2.76	1.59	2.58	5.87	3.59	2.70	1.71
	Average Degree	−13.98	−15.34	−7.07	−13.31	−10.71	−12.76	−23.64	−17.47	−10.66	−11.99	−12.75	−6.84
	Reciprocity	13.64	7.90	11.07	19.06	13.90	20.89	6.04	36.55	30.75	8.76	13.43	11.33
	APL	−1.10	0.61	−8.26	−5.91	0.00	−0.20	11.64	−0.16	1.32	−3.63	9.48	−1.87
SAB Model	Transitivity	13.28	59.96	54.13	28.21	11.68	0.88	13.65	75.75	79.16	11.02	7.48	9.97
	Average Degree	−2.43	−13.05	−14.53	−4.32	−2.97	−34.52	1.87	−24.40	−16.30	0.26	−3.09	−15.61
	Reciprocity	7.61	13.09	32.29	13.11	6.06	1.19	5.19	31.68	25.62	5.65	2.74	2.96
	APL	4.09	26.98	2.47	0.13	4.09	4.85	2.72	9.67	6.47	1.77	4.77	7.97
PageRank	Transitivity	67.99	64.87	64.92	43.24	22.84	54.18	47.28	119.05	73.31	35.63	16.29	29.87
	Average Degree	−23.79	−14.43	−18.49	−45.78	−11.59	−24.38	−19.04	−41.55	−22.38	−9.56	−9.20	−12.00
	Reciprocity	40.19	17.28	33.18	37.85	27.10	25.09	19.41	90.62	29.38	24.95	37.70	19.94
	APL	3.13	13.35	−1.11	−0.18	5.73	6.48	6.94	5.14	2.47	1.08	4.97	5.04
PageRank-Max	Transitivity	0.37	1.35	1.68	2.00	2.03	2.70	4.56	3.15	3.64	−0.43	3.18	2.81
	Average Degree	19.24	32.53	32.32	21.29	12.54	21.49	22.67	43.64	22.33	16.96	16.42	21.79
	Reciprocity	−2.15	−1.63	1.51	1.73	1.61	0.37	−1.71	−0.43	−0.23	−4.28	−2.20	−1.19
	APL	−17.97	−39.06	−23.38	−52.14	−11.12	−90.13	−32.46	−26.92	−14.21	−18.12	−20.67	−7.18

Table 7.10: Effect size of control schools at T_2, highlighted values indicate a predicted value not significantly different from the data.

		12	32	34	73	74	76
Adamic/Adar	Transitivity	−7.66	6.96	−15.97	20.45	−13.00	−5.37
	Average Degree	−17.14	−18.05	−22.20	−19.52	−19.23	−20.13
	Reciprocity	−3.45	15.31	−3.66	13.99	−6.60	5.77
	APL	1.34	9.93	8.87	5.66	12.70	3.96
Katz	Transitivity	1.28	66.74	1.14	32.87	15.63	52.38
	Average Degree	−18.61	−43.17	−18.44	−23.03	−18.80	−25.92
	Reciprocity	13.77	98.10	21.80	71.12	25.52	99.16
	APL	−6.15	14.26	−0.36	7.12	13.29	−17.11
SAB Model	Transitivity	20.03	98.85	33.50	111.27	6.66	82.93
	Average Degree	3.69	−22.92	2.60	−14.33	1.32	−18.09
	Reciprocity	10.83	41.19	15.71	43.66	2.98	48.17
	APL	2.61	14.86	2.91	21.27	0.48	6.22
PageRank	Transitivity	51.35	76.21	58.81	110.64	36.73	76.21
	Average Degree	−21.00	−21.83	−20.13	−17.17	−23.19	−21.27
	Reciprocity	23.56	108.62	53.09	112.71	25.59	117.07
	APL	5.42	8.93	4.37	6.61	5.01	1.29
PageRank-Max	Transitivity	3.63	40.73	3.16	23.41	3.12	10.27
	Average Degree	18.89	11.92	21.72	8.33	23.99	22.69
	Reciprocity	2.60	3.64	1.09	13.93	−2.45	4.31
	APL	−15.83	−4.79	−16.01	−3.38	−7.98	−19.90

Table 7.11: Effect size of intervention schools at T_2, highlighted values indicate a predicted value not significantly different from the data.

		15	22	33	35	40	41	62	63	64	68	69	71
Adamic/Adar	Transitivity	−8.09	−13.26	6.91	−15.06	−15.40	−10.78	−14.16	−19.38	−24.64	−6.80	−5.13	−12.48
	Average Degree	−24.34	−22.02	−15.90	−32.89	−28.77	−52.08	−34.68	−43.47	−23.07	−26.50	−28.00	−19.21
	Reciprocity	−4.46	−5.84	17.19	−11.19	−5.01	−10.18	−7.86	−8.35	−11.30	−2.46	−4.04	−5.09
	APL	24.54	3.84	656.36	451.62	40.79	21.54	47.10	59.02	18.78	53.50	16.99	274.02
Katz	Transitivity	2.92	−2.17	25.38	−0.08	−7.36	2.38	−0.36	0.65	−1.16	5.08	4.42	3.46
	Average Degree	−19.37	−25.34	−22.73	−27.23	−19.35	−21.27	−17.53	−25.26	−16.07	−17.29	−16.82	−11.88
	Reciprocity	12.63	13.36	103.38	9.13	7.00	11.69	6.33	24.60	13.00	14.51	11.85	10.74
	APL	18.03	7.82	403.68	312.47	13.94	26.57	24.30	64.83	101.60	17.94	45.00	152.45
SAB Model	Transitivity	63.76	15.10	29.34	54.90	5.88	54.82	14.76	74.39	11.46	91.86	13.05	10.82
	Average Degree	−21.22	−9.01	−4.79	−19.03	−6.28	−24.94	−2.98	−25.79	−19.44	−27.00	−9.90	−20.95
	Reciprocity	26.86	4.62	35.97	14.08	2.47	26.46	9.42	34.89	6.71	11.02	2.80	19.03
	APL	42.64	13.07	6.01	45.92	9.70	40.53	9.61	28.07	41.49	46.53	20.50	11.55
PageRank	Transitivity	70.61	54.15	51.68	53.04	11.97	73.51	64.43	100.35	51.87	65.12	26.86	28.81
	Average Degree	−25.19	−18.69	−23.37	−22.08	−13.20	−39.08	−19.54	−34.34	−33.20	−24.70	−14.43	−16.98
	Reciprocity	33.95	30.46	92.30	30.00	8.46	39.26	21.16	40.89	30.87	20.65	22.72	23.53
	APL	11.38	51.56	22.21	19.80	15.65	38.02	8.71	14.36	23.08	55.94	19.31	16.89
PageRank-Max	Transitivity	2.21	−0.50	27.04	0.33	−0.71	3.02	4.38	2.19	1.28	4.82	2.74	2.88
	Average Degree	19.02	19.12	2.52	25.32	8.87	14.48	10.61	17.60	11.82	5.38	7.64	13.90
	Reciprocity	−1.19	0.99	5.43	−0.06	0.35	0.62	1.69	−0.75	−0.43	−3.46	−1.64	0.91
	APL	−16.75	−39.21	1.94	−8.96	−20.17	−5.19	−7.39	−12.59	−15.14	−4.73	−1.20	−20.10

Table 7.12: Effect size of control schools at T_3, highlighted values indicate a predicted value not significantly different from the data.

		12	32	34	73	74	76
Adamic/Adar	Transitivity	−26.14	−15.12	6.71	−7.79	42.18	−36.76
	Average Degree	−23.94	−29.42	−29.12	−36.06	−18.76	−27.49
	Reciprocity	−11.63	−11.14	10.91	−0.67	16.53	−7.23
	APL	22.20	69.74	68.03	32.94	33.65	47.28
Katz	Transitivity	3.25	6.66	5.20	8.27	38.36	3.51
	Average Degree	−15.24	−14.54	−31.39	−38.30	−18.11	−28.10
	Reciprocity	12.47	13.99	37.49	27.21	63.86	33.16
	APL	9.73	9.48	84.79	85.61	64.45	65.44
SAB Model	Transitivity	39.98	57.55	97.44	28.04	68.40	81.35
	Average Degree	−13.55	−24.94	−25.18	−1.94	8.22	−24.58
	Reciprocity	9.97	26.19	24.21	16.93	23.36	33.13
	APL	16.51	78.59	57.85	11.51	1.08	54.38
PageRank	Transitivity	86.00	79.74	128.78	78.03	58.56	87.44
	Average Degree	−21.29	−42.91	−26.06	−36.63	−16.26	−23.29
	Reciprocity	35.22	46.24	63.40	46.40	47.58	43.91
	APL	8.93	18.53	23.07	23.90	14.63	13.12
PageRank-Max	Transitivity	−0.51	4.34	15.92	14.44	74.97	3.32
	Average Degree	19.33	32.28	4.88	7.55	0.46	25.84
	Reciprocity	−1.39	−0.25	3.75	1.19	4.32	1.24
	APL	−30.80	−13.98	−0.72	−5.17	3.72	−9.38

Table 7.13: Effect size of intervention schools at T_3, highlighted values indicate a predicted value not significantly different from the data.

7.12 and 7.13, highlighted values indicate that the null hypothesis ($H_0 : \bar{x_p} = \bar{x_d}$) is **not** rejected at the 95% level; giving a prediction value not significantly different from that of the data.

Control Schools at T_2

From Table 7.10, it is evident that many of the network structural metrics are significantly different from the true values (for control schools at T_2). First exploring the AA method, it would appear that transitivity and average degree are significantly overestimated; this indicates individuals within the control networks (at T_2) are being predicted to have too many connections, with a higher than expected number of closed friendship triangles.

The values exhibited by the AA method appear consistent with the algorithm's logic, with the method attempting to close triangles between "common neighbours" (Section 6.1.1); hence, resulting in the overestimation of transitivity. The raw transitivity of predicted networks at T_2, classified by LP method, is displayed in Figure 7.5 - plotting data from both control and intervention schools. The boxplot demonstrates the high transitivity values of networks generated by the AA method - explaining the overestimation observed.

The AA average degree values are also overestimated, suggesting that each agent has too many new links, with not enough existing links being broken. This is also consistent with the AA logic, as the method is conventionally used to predict new links; the implementation within the SNS only being required to break links under specific conditions. The reciprocity values for the AA method are also overestimated, however, the value for school 40 is not significantly different from the data (-0.56). School 40 is a girls school which exhibited a heavily cliqued network formation at T_2 (Section 5.2.3), therefore, the common neighbours structure of the AA method may capture the cliqued nature of school 40 appropriately - explaining the non significant difference.

The APL figures of the AA method appear underestimated, this suggests average path lengths are shorter than the real data. The Katz method performs better in terms of APL, schools 22 (0.61), 40 (0.00), 41 (-0.20) and 63 (-0.16) producing values not significantly different from true network values. The Katz method bases linking decisions upon indirect path length, the SNS considering individuals of distance up to three away; as such, paths between considered linking agents already exist, with the method only marginally short-

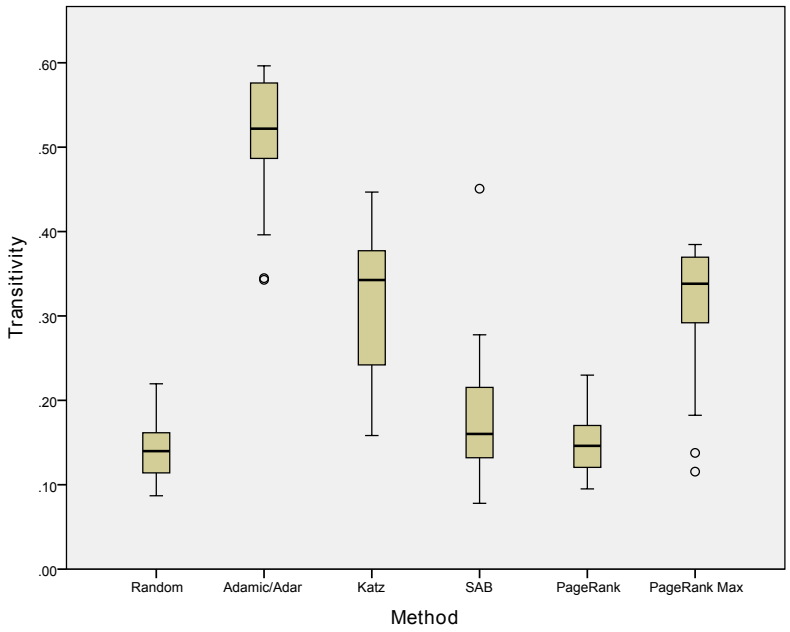


Figure 7.5: Box plot of transitivity for each method at T_2. Whiskers extend 1.5 times the height of the box, with circular points indicating outliers.

ening the APL if a link is made. It would therefore appear that the Katz method creates a representative path length in the control networks, without generating links that shorten the APL too greatly.

The transitivity effect size of the Katz method also performs well, with school 22 showing no significant difference from the real data (-0.71). This suggests that the paths approach of Katz produces an indicative number of closed triangles, avoiding the overestimation present in the AA method. However, Katz also demonstrates an overestimation in the average degree, suggesting too many new links are being formed; this is consistent with findings of the AA method, both algorithms being originally developed to predict the formation of new links.

The SAB method produces average out-degree values not significantly different from the data for School 68 (0.26) - indicating the SAB method is predicting an appropriate number of new connections. However, the equivalent figures for the remaining control school average degree predictions are variable, with school 62 being significantly underestimated (1.87) and school 41 being overestimated (-34.52).

The SAB method also produces an APL value representative of the network of school 35 (0.13), however, values for all metrics (within the SAB method) vary greatly depending upon the specific school network; for example the large transitivity effect size of school 64 (79.16), relative to that of school 41 (0.88). This highlights the model creation aspect of the SAB method, the model potentially being more representative of transitivity within school 41 than that of school 64. This provides further evidence to the conclusions of Section 7.1, which state that the accuracy of the SAB method is highly dependent upon model specification.

The precision values of the PR method (from Section 7.1), demonstrated its poor performance in terms of accuracy; it would appear that the PR method does not improve in terms of structural analysis. Transitivity and reciprocity are significantly underestimated (school 63: 119.05 and 90.62, respectively), with the PR method significantly overestimating the number of links formed (school 35: -45.78). While the APL of school 35 is not significantly different from the real network (-0.18), overall the method does not produce network structures indicative of the data.

The PR-Max method performs substantially better than PR. Transitivity effect size appears relatively low, with schools 15 (0.37) and 68 (-0.43) displaying values not significantly different from the true networks; furthermore, the method also produces indicative levels of reciprocity, schools 41 (0.37), 63 (-0.43) and 64 (-0.23) also demonstrating no significant difference. While all other methods indicate an overestimation in terms of the number of connections in the predicted network (average degree), PR-Max significantly underestimates the number of connections formed. This suggests that the PR-Max method is opting to break many of the existing friendship links, producing a sparse predicted network.

The removal of links within the PR-Max method, is consistent with the underlying PR process. Avrachenkov & Litvak (2004) and Gyöngyi & Garcia-Molina (2005) demonstrate that the removal of specific links may increase PR, as such, the agents within the predicted network are dropping links which have a negative impact upon their personal PR. The APL figures for PR-Max also reinforce these findings, with values being overestimated; this suggests an increase to APL, which would occur if path shortening links were being broken within the network. Figure 7.6 depicts the raw APL values for each method at T_2, the box plot taking into consideration both control and intervention school data. This demonstrates the variability in APL predictions for PR-Max in comparison with the other

selected LP methods.

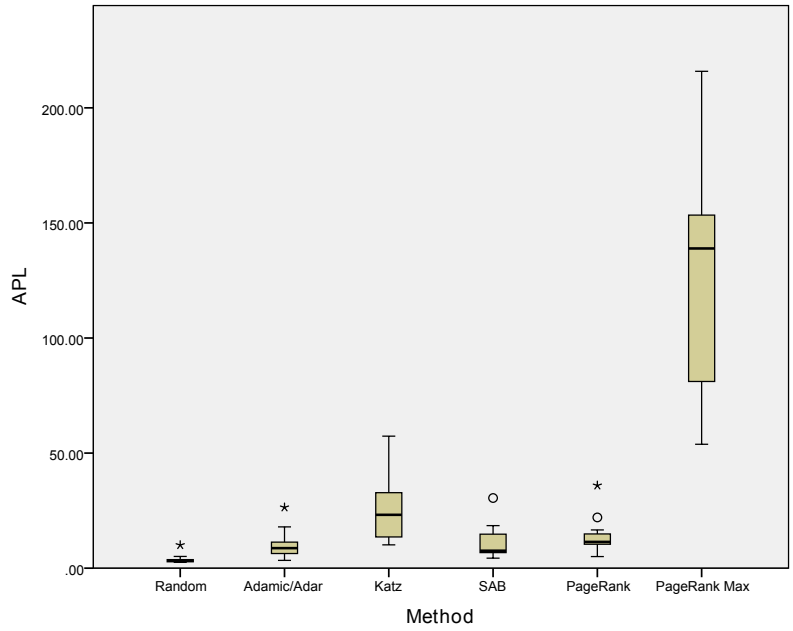


Figure 7.6: Box plot of APL for each method at T_2. Whiskers extend 1.5 times the height of the box, with circular points indicating outliers. Starred points indicate extreme outliers.

The control schools at T_2 have provided an insight into the workings of each LP method, highlighting specific structural features of the LP methods. It is of interest to investigate whether the findings specified are consistent in intervention schools at T_2.

Intervention Schools at T_2

From Table 7.11, it would appear intervention schools display some inconsistencies with the predicted network structure of control schools (at T_2). For the AA method, transitivity figures were consistently overestimated for control schools; however, intervention schools 32 (6.96) and 73 (20.45) are significantly underestimated. This is of particular interest, as the AA method seeks to generate highly transitive networks. Schools 32 and 73 also performed particularly poorly in terms of precision at T_2 (as discussed in Section 7.1), with both schools containing a larger number of individuals in their respective networks; this once again highlights the issue of network size with regard to LP, an issue examined further in Section 7.2.3.

The Katz method for school 34, produces an APL effect size not significantly different from the true network (-0.36); the Katz method also previously performing well with respect to APL in control schools. However, Katz reciprocation effect sizes are relatively large across all intervention schools; schools 32 (98.10) and 76 (99.16) demonstrating the largest figures. In contrast, the largest reciprocity value in control schools was school 63 (36.55). This suggests that the Katz method performs worse in interventions schools in terms of reciprocity, potentially a result of the alteration in friendship structures due to the nomination of peer supporters.

The SAB method also exhibits increased effect sizes, schools 32 (98.85) and 73 (111.27) producing particularly large figures relating to transitivity. However, the SAB method does indicate an APL not significantly different from the true network of school 74 (0.48); this demonstrates (once again) the variability of the SAB method, and its dependence upon the underlying SAB model generated. PR and PR-Max both generate networks significantly different from the data (across all intervention schools), with the underestimation of links in the PR-Max method being consistent.

Overall it would appear that the general trends exhibited by control schools at T_2, are carried forward by intervention schools at T_2, with many of the observations reinforcing the conclusions of the precision analysis of Section 7.1. However, many of the intervention predicted network effect sizes are greater than those of control schools, which potentially indicates less structural accuracy in intervention schools; this notion is investigated further in Section 7.3. To consider the structural performance of LP methods across time steps, an analysis of control and intervention school effect sizes at T_3 is required.

Control Schools at T_3

Table 7.12 demonstrates a marked reduction in network structural accuracy, with an increase in the number of predicted network values significantly different from the data. The AA method is no longer representative in terms of reciprocity at school 40, suggesting the structural evolution of the school 40 network may have changed between T_1 to T_2 and T_2 to T_3. Furthermore, the AA APL effect sizes for schools 33 (656.36) and 35 (451.62) are greatly increased, indicating a substantial underestimation in the APL at T_3.

The Katz method, having previously demonstrated small APL effect sizes at T_2, displays

greatly increased figures in schools 33 (403.68) and 35 (312.47) at T_3. This indicates that schools 33 and 35 may have particularly unique APL properties at T_3, with the Katz and AA methods being unable to represent these effectively. Examining the network characteristics table of Section 5.2.3 (Table 5.4), school 33 (0.828) and 35 (0.834) have shorter than average path lengths (0.651) at T_3 - potentially leading to the observed underestimation. However, the Katz method does produce transitivity effect sizes not significantly different to the real school networks of 35 (-0.08), 62 (-0.36) and 63 (0.65). This once again demonstrates a shift in network structure between timesteps, the Katz method able to capture the transitivity of control school at T_3 more appropriately than at T_2.

Exemplifying the time effects further, the SAB school 68 average degree (0.26) was not significantly different to the data at T_2 (Table 7.10); this is not the case at T_3, the SAB model subsequently underestimating the average degree (-27.00). It is of particular interest that the SAB method does not perform better in terms of network structure (across all time periods), as the method is specifically calibrated to represent structural evolution - the generated model having explicit knowledge of multiple waves of network data. This evidently questions the insights gained from an SAB model, given that its underlying link predictions are significantly different from the data; however, with alternative model objective functions, more appropriate link predictions may be produced.

The PR method retains its poor performance upon control schools at T_3, but the PR-Max method highlights a number of important structural predictions. The PR-Max transitivity effect size is low across all schools, with schools 22 (-0.50), 35 (0.33) and 40 (-0.71) indicating no significant difference from the data. Furthermore, reciprocity effect size is also low, with no significance being observed in schools 35 (-0.06), 40 (0.35), 41 (0.62) and 64 (-0.43). While the PR-Max method still severs an excessive amount of links at T_3 (as demonstrated by an underestimation in average degree), the method appears to provide a better representation of network structure than other LP methods at T_3; this being further reinforced by the high precision of control schools at T_3 (Section 7.1).

Intervention Schools at T_3

Table 7.13 demonstrates the positive performance of PR-Max upon intervention schools at T_3. Transitivity effect size for school 12 (-0.51) and reciprocity for school 32 (-0.25) are not significantly different from the data. Of particular interest is the low effect size

observed for the average degree of school 74 (0.46), as the PR-Max method significantly underestimates the number of links formed for all other school networks (at both T_2 and T_3). It would appear that the PR-Max process of removing links, represents the school 74 network evolution appropriately at T_3 - once again demonstrating both the time effects present in the investigated network structures, and the differing behaviours of individual schools. The Katz reciprocity effect size figures of school 32 (13.99) and 73 (27.21) add further weight to the discussed observations, these values being greatly reduced from T_2 (98.10 and 71.12, respectively).

To further understand the differing network structures across time periods, and the subsequent impact upon structural accuracy, the following section (7.2.2) condenses the network measures for analysis. An average effect size (AES) shall be calculated, this allowing for a ranking to be calculated for each method - giving an overall representation of structural performance.

7.2.2 Method Structural Performance

To evaluate each method's performance in terms of structural measures, a rank for each method is produced. This is calculated by taking the average absolute effect size (AES) across schools, for each structural metric. This analysis is only concerned with the magnitude of effect size, the directionality (overestimation or underestimation) being irrelevant; as such, the absolute effect size is taken in the calculation of AES.

Tables 7.14 and 7.15 display the control and intervention school AES (for each method and measure), respectively. To compare AES differences between time steps, paired sample t-tests (parametric) or paired sample Wilcoxon signed-rank (non-parametric) tests are performed - the values significantly different at the 95% level highlighted in Tables 7.14 and 7.15. Each LP method is then ranked by structural measure, values with the lowest AES achieving the highest ranks - Tables 7.16 (control) and 7.17 (intervention) displaying the individual ranks. Finally, the harmonic mean of the ranks is taken for each LP method, giving an overall structural accuracy ranking; the figures being presented in Table 7.18.

For control schools, the differences in AES between time steps is apparent from Table 7.14. The APL is predicted significantly differently across all methods, with predictions being worse for T_3 in AA (139.01), Katz (99.05), SAB (26.30) and PR (24.74) methods

Time	Measure	Adamic/Adar	Katz	SAB Model	PageRank	PageRank-Max
T_2	Transitivity	11.80	2.65	30.43	53.29	2.32
	Average Degree	22.22*	13.04*	11.11	21.01	23.60*
	Reciprocity	7.28	16.11	12.27	33.56	1.59
	Average Path Length	13.20*	3.68*	6.33*	4.64*	29.45
T_3	Transitivity	12.67	4.62	36.68	54.37	4.34
	Average Degree	29.25*	20.01*	15.94	23.73	13.02*
	Reciprocity	7.75	19.85	16.20	32.85	1.46
	Average Path Length	139.01*	99.05*	26.30*	24.74*	12.78*

Table 7.14: Control School AES for each LP method, highlighted values indicate a significant difference between time periods.

Time	Measure	Adamic/Adar	Katz	SAB Model	PageRank	PageRank-Max
T_2	Transitivity	11.57	28.34	58.87	68.32	14.05
	Average Out-Degree	19.38*	24.66	10.49	20.77	17.92
	Reciprocity	8.13	54.91	27.09	73.44	4.67
	Average Path Length	7.08*	9.71*	8.06	5.27*	11.32
T_3	Transitivity	22.45	10.87	62.13	86.43	18.92
	Average Out-Degree	27.46*	24.28	16.40	27.74	15.05
	Reciprocity	9.69	31.36	22.30	47.12	2.02
	Average Path Length	45.64*	53.25*	36.65	17.03*	10.63

Table 7.15: Intervention School AES for each LP method, highlighted values indicate a significant difference between time periods.

Time	Measure	Adamic/Adar	Katz	SAB Model	PageRank	PageRank-Max
T_2	Transitivity	3	2	4	5	1
	Average Out-Degree	4	2	1	3	5
	Reciprocity	2	4	3	5	1
	Average Path Length	4	1	3	2	5
T_3	Transitivity	3	2	4	5	1
	Average Out-Degree	5	3	2	4	1
	Reciprocity	2	4	3	5	1
	Average Path Length	5	4	3	2	1

Table 7.16: Control school AES ranks for each LP method.

Time	Measure	Adamic/Adar	Katz	SAB Model	PageRank	PageRank-Max
T_2	Transitivity	1	3	4	5	2
	Average Out-Degree	3	5	1	4	2
	Reciprocity	2	4	3	5	1
	Average Path Length	2	4	3	1	5
T_3	Transitivity	3	1	4	5	2
	Average Out-Degree	4	3	2	5	1
	Reciprocity	2	4	3	5	1
	Average Path Length	4	5	3	2	1

Table 7.17: Intervention school AES ranks for each LP method.

		Adamic/Adar	Katz	SAB Model	PageRank	PageRank-Max
Control	T_2	3.0	1.8	2.1	3.2	1.7
	T_3	3.2	3.0	2.8	3.5	1.0
Intervention	T_2	1.7	3.9	2.1	2.4	1.8
	T_3	3.0	2.2	2.8	3.6	1.1

Table 7.18: Harmonic mean of AES ranks for each LP method.

than T_2; however, AES is reduced for PR-Max at T_3 (12.78), this indicating a significant improvement in predictions. AES for average degree is also significantly different between T_2 and T_3, with AA (29.25) and Katz (20.01) increasing; once again, PR-Max values improve at T_3, with the AES value decreasing significantly.

The AES values discussed (from Table 7.14), indicate an improvement in the PR-Max structural accuracy at T_3. This is further reinforced by the ranks of Table 7.16, which demonstrate a movement of out-degree and APL predictions from last place (5) at T_2, to first place at T_3 (1). When the harmonic mean of the individual rankings is taken for each method, PR-Max is placed first across both time steps for control schools (T_2: 1.7, T_3: 1.0), however, at T_2 this is very closely followed by the Katz method (1.8).

The precision analysis of Section 7.1, placed the Katz method as fourth overall at both T_2 and T_3 (for control schools). However, it would appear that the method performs well in terms of structure at T_2, ranking first in APL AES and second for transitivity and average out-degree. This suggests that, while the specific links in the predicted networks may not be accurate, the overall network structure generated is more representative than other LP methods - only being outperformed by PR-Max in terms of transitivity and reciprocity. The findings demonstrate the importance of considering the predicted network structure

when discussing LP methods, potentially providing further insight than simply considering precision.

Intervention schools present differing outcomes to those of control schools in terms of structural accuracy. For example, in Table 7.15, PR-Max and the SAB method experience no significant difference in AES across time steps. Significant increases are still observed at T_3 however, with AA out-degree (27.46) and AA (45.64), Katz (53.25) and PR (17.03) APL predictions demonstrating less structural accuracy.

Investigating the individual rankings of Table 7.17, it would appear that the Katz method is the lowest ranked in out-degree and is fourth in terms of APL at T_2 (in intervention schools); the method previously ranking second (out-degree) and first (APL), respectively, in control school. This is further demonstrated by the low overall ranking of the Katz method in intervention schools at T_2 (3.9) from Table 7.18. With regard to the PR-Max method in intervention schools, it would appear that the high ranking of the AA method in terms of transitivity and APL (from Table 7.17), has placed PR-Max second at T_2; although, PR-Max ranking first at T_3. This highlights the differences in prediction between control and intervention schools, reasons for these difference being investigated further in Section 7.3.

Overall, the method structural performance analysis has reinforced many of the conclusions from Section 7.1. There would appear to be differences in the performance of LP methods at T_2 and T_3, suggesting an underlying change in the friendship mechanisms of adolescents within the ASSIST data. Further evidence of the strength of the PR-Max method (in predicting network evolution) is also provided, the method performing particularly well at T_3. The analysis of this section (7.2.2) has taken an average of all school effect sizes, however, there may be schools which are predicted particularly well, and those which are predicted particularly poorly. An investigation of the effect of specific schools upon the structural accuracy of predictions, is discussed in the following section (7.2.3).

7.2.3 School Structural Performance

The precision analysis of Section 7.1 and the current structural analysis highlight the importance of school attributes upon the performance of LP methods, with the friendship evolution of some schools being particularity difficult to quantify. To investigate this fur-

ther, the structural analysis effect sizes are utilised to give an overview of the structural performance of each school, across all LP methods (at each timestep).

The average structural measure effect size is calculated for each school, classified by LP method; Tables 7.19 and 7.20 displaying figures for control and intervention schools respectively. The schools are then ranked in terms of AES for each method, emphasising which schools performed particularly well in terms of structural measures - values displayed in Tables 7.21 (control) and Table 7.22 (intervention). Finally, the harmonic mean of the ranks is taken to produce an overall rank for each control (Table 7.23) and intervention school (Table 7.24).

From Table 7.19, School 40 demonstrates low AES values across all LP methods at T_2, being the best predicted school by the PR-Max method (Table 7.21). Similarly, school 71 also demonstrates low overall AES values at T_2, with the Katz and PR methods ranking the school first in terms of lowest AES (Table 7.21). This results in Schools 40 and 71 being ranked first (1.7) and second (1.8) respectively, in terms of structural predictions across all methods at T_2 - as demonstrated by Table 7.23. Thus, the LP methods are most effective in predicting the structural accuracy of schools 40 and 71 at T_2.

Examining control schools at T_3 (Table 7.19), school 40 and 69 display low AES values across the LP methods; resulting in school 40 again being ranked first (1.7) in terms of best structural prediction (by all methods) at T_3, with school 69 (2.1) being ranked second (Table 7.23). School 40 is highly ranked at both time steps, suggesting its structural properties can be represented appropriately irrespective of the LP method selected and time period being predicted.

From Table 5.1, school 40 is the smallest network in the data (62 participants), which may be the cause of its high structural performance - the searching agent having less options to select in terms testing agents. Furthermore, school 71 also has a small network (102 participants), reinforcing this notion further. The worst performing control school at T_2 is school 63, having an AES rank of 11.4 (Table 7.23). School 63 is the largest control school in the data set (Table 5.1), containing 236 individuals; this once again suggests the size of the network may factor into its effective link prediction.

Intervention schools display a similar trend, school 73 (5.0 at T_2) and 76 (3.8 at T_3) obtaining the lowest overall structural prediction ranks (Table 7.24) - both schools containing a

Time	LP Method	15	22	33	35	40	41	62	63	64	68	69	71
	Adamic/Adar	11.97	16.58	9.71	13.47	9.41	12.20	21.46	23.28	11.57	8.30	15.88	9.69
	Katz	8.23	6.14	7.29	9.82	6.73	9.15	10.73	14.19	12.15	6.99	9.59	5.44
T_2	SAB Model	6.85	28.27	25.86	11.44	6.20	10.36	5.86	35.38	31.89	4.68	4.52	9.13
	PageRank	33.78	27.48	29.42	31.76	16.82	27.53	23.17	64.09	31.89	17.80	17.04	16.71
	PageRank-Max	9.93	18.64	14.72	19.29	6.82	28.67	15.35	18.53	10.10	9.94	10.62	8.24
	Adamic/Adar	15.36	11.24	174.09	127.69	22.49	23.65	25.95	32.55	19.45	22.31	13.54	77.70
	Katz	13.24	12.17	138.79	87.23	11.91	15.48	12.13	28.84	32.96	13.70	19.52	44.63
T_3	SAB Model	38.62	10.45	19.03	33.48	6.08	36.69	9.19	40.79	19.78	44.10	11.56	15.59
	PageRank	35.28	38.71	47.39	31.23	12.32	47.47	28.46	47.49	34.76	41.60	20.83	21.55
	PageRank-Max	9.79	14.96	9.23	8.67	7.53	5.83	6.02	8.28	7.16	4.60	3.31	9.45

Table 7.19: AES values for each conrtol school and each LP method.

Time	LP Method	12	32	34	73	74	76
	Adamic/Adar	7.40	12.56	12.67	14.90	12.88	8.81
	Katz	9.95	55.56	10.43	33.53	18.31	48.64
T_2	SAB Model	9.29	44.45	13.68	47.63	2.86	38.85
	PageRank	25.33	53.90	34.10	61.78	22.63	53.96
	PageRank-Max	10.24	15.27	10.49	12.26	9.39	14.29
	Adamic/Adar	20.98	31.35	28.69	19.36	27.78	29.69
	Katz	10.17	11.17	39.72	39.85	46.19	32.55
T_3	SAB Model	20.00	46.82	51.17	14.60	25.26	48.36
	PageRank	37.86	46.86	60.33	46.24	34.26	41.94
	PageRank-Max	13.01	12.71	6.32	7.09	20.87	9.95

Table 7.20: AES values for each intervention school and each LP method.

Time	LP Method	15	22	33	35	40	41	62	63	64	68	69	71
	Adamic/Adar	6	10	4	8	2	7	11	12	5	1	9	3
	Katz	6	2	5	9	3	7	10	12	11	4	8	1
T_2	SAB Model	5	10	9	8	4	7	3	12	11	2	1	6
	PageRank	11	6	8	9	2	7	5	12	10	4	3	1
	PageRank-Max	3	10	7	11	1	12	8	9	5	4	6	2
	Adamic/Adar	3	1	12	11	6	7	8	9	4	5	2	10
	Katz	4	3	12	11	1	6	2	8	9	5	7	10
T_3	SAB Model	10	3	6	8	1	9	2	11	7	12	4	5
	PageRank	7	8	10	5	1	11	4	12	6	9	2	3
	PageRank-Max	11	12	9	8	6	3	4	7	5	2	1	10

Table 7.21: Control school AES ranks for each LP method.

Time	LP Method	12	32	34	73	74	76
	Adamic/Adar	1	3	4	6	5	2
	Katz	1	6	2	4	3	5
T_2	SAB Model	2	5	3	6	1	4
	PageRank	2	4	3	6	1	5
	PageRank-Max	2	6	3	4	1	5
	Adamic/Adar	2	6	4	1	3	5
	Katz	1	2	4	5	6	3
T_3	SAB Model	2	4	6	1	3	5
	PageRank	2	5	6	4	1	3
	PageRank-Max	5	4	1	2	6	3

Table 7.22: Intervention school AES ranks for each LP method.

Time	15	22	33	35	40	41	62	63	64	68	69	71
T_2	5.5	5.1	6.4	8.2	1.7	8.0	5.5	11.4	7.7	2.2	3.2	1.8
T_3	5.1	3.1	9.4	6.2	1.7	6.2	2.3	8.0	5.9	4.8	2.1	6.4

Table 7.23: Harmonic mean of AES ranks for each control school.

Time	12	32	34	73	74	76
T_2	1.3	4.7	2.7	5.0	1.6	3.8
T_3	1.6	3.2	2.9	1.8	2.8	3.8

Table 7.24: Harmonic mean of AES ranks for each intervention school.

large number of participants (school 73: 199, school 76: 254). School 12 is highest ranked intervention school at both T_2 and T_3 (Table 7.24), with the AA and Katz methods ranking school 12 as their best predicted school at T_2 (Table 7.22). While school 12 does not have smallest network size of all intervention schools (164 students), the school has a highly transitive network across all time periods (as demonstrated by Table 5.4); this may be the reason for the effectiveness of the AA and Katz method structural predictions.

The school-specific structural analysis has demonstrated that LP methods may be more effective in representing certain schools, with an important criteria potentially being the network size. Figures 7.7 and 7.8 display the correlation between AES magnitude and network size (classified by LP method) at T_2 and T_3 respectively; the associated correlation coefficients (r) and P-Values are presented in Table 7.25. The values are calculated amalgamating control and intervention schools, with the random method also represented for comparative purposed.

		Random	Adamic/Adar	Katz	SAB Model	PageRank	PageRank-Max
T_2	R	0.46	0.13	0.62	0.68	0.84	0.36
	P-Value	0.05	0.60	0.01	< 0.01	< 0.01	0.14
T_3	R	-0.11	-0.08	-0.02	0.75	0.79	0.05
	P-Value	0.68	0.74	0.93	< 0.01	< 0.01	0.85

Table 7.25: Correlation coefficients and associated P-Values for network size against AES magnitude, classified by LP method and time step.

From Table 7.25, there is a strong significant correlation between school network size and AES magnitude for Katz (0.62), SAB (0.68) and PR (0.84) methods at T_2. At T_3 a strong significant correlation is also observed in the PR and SAB methods, however, no correlation is observed by the Katz method. This provides evidence of the dependence of the SAB and PR methods upon network size in making valid predictions, with the Katz method perhaps demonstrating the changing network evolution processes over time.

The significant SAB correlations also suggest that the larger the network, the less structural accuracy in the underlying SAB link predictions; therefore, conclusions drawn from SAB models may be diminished as network size increases. The PR-Max AES structural values do not appear to be correlated with network size, suggesting the method is not affected by vertex count; rather, the PR-Max performs well if the linking behaviour of individuals within the network conforms to the method of optimising eigen-centrality.

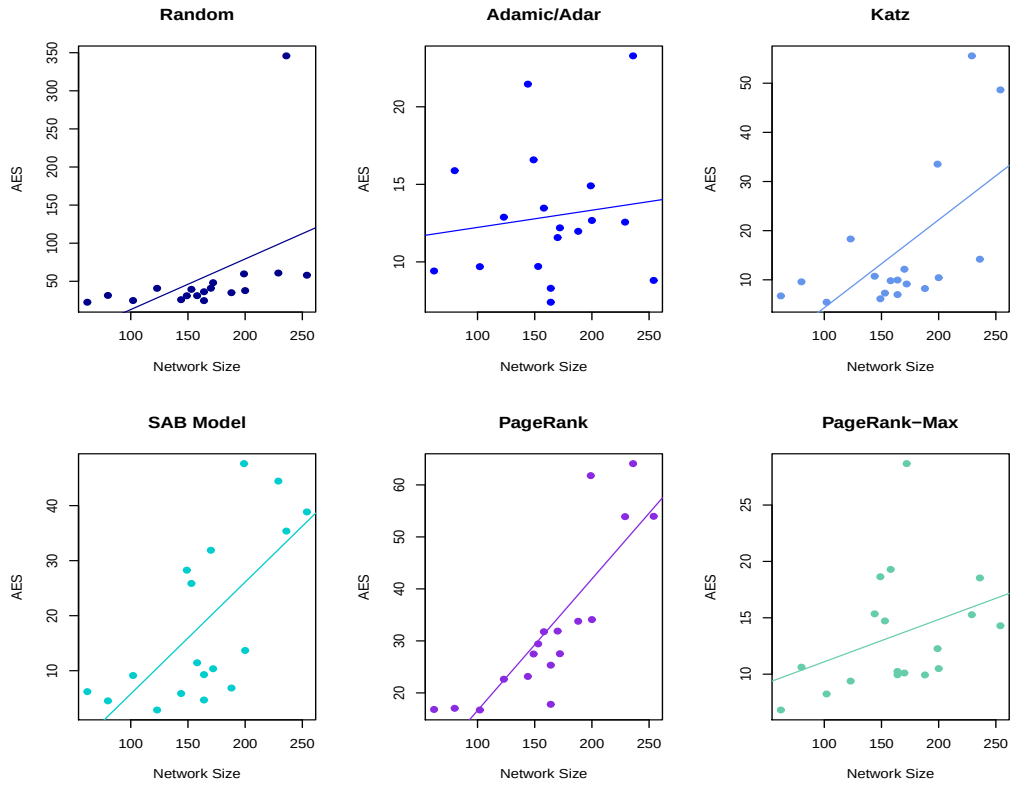


Figure 7.7: Correlation graphs for each LP method, displaying network size against AES at T_2

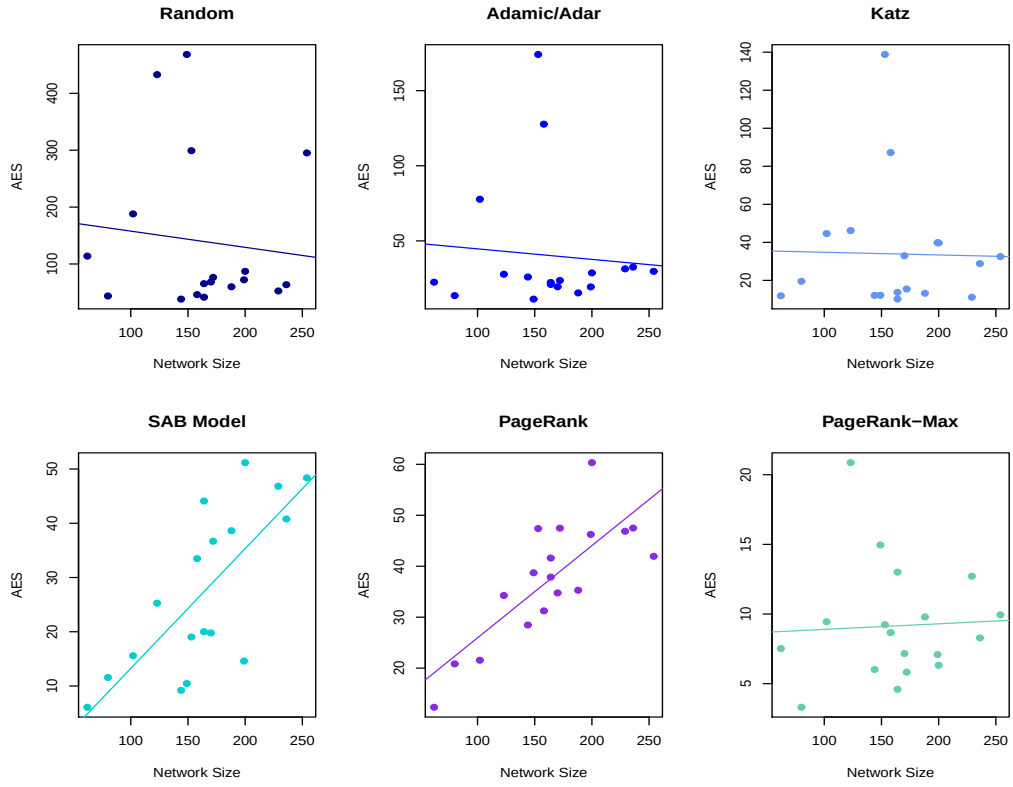


Figure 7.8: Correlation graphs for each LP method, displaying network size against AES at T_3

The effect size analysis of Section 7.2, has reinforced and expanded upon the findings of the precision analysis (Section 7.1) as follows:

- The network time period in which the LP methods are exacted is important, the network evolution process changing between subsequent time points;
- The unique characteristics by which LP methods evolve connections in a network, impact upon the predicted network - some schools being more suited to certain LP methods;
- The strength of the PR-Max method is further reinforced, attributing its success to its ability to identify links to sever;
- The standard PR method is still the lowest performing method;
- Network size is significantly correlated with the structural accuracy of predictions in Katz (at T_2), SAB and PR methods;
- The underlying SAB model link predictions decrease in accuracy as network size increases, potentially highlighting uncertainty with regard to the conclusions gained from an SAB model.

Furthermore, the analysis has also indicated that there may be differences between LP method performance upon control and intervention school predictions; the following section (7.3) investigates this further.

7.3 Control and Intervention Comparison

Section 7.1 and 7.2, indicated that there may be differences in the performance of LP methods upon control and intervention schools, which would suggest underlying differences in the real school network structures. The network data analysis of Section 5.2.3, found there to be only one significant difference in the structure of intervention and control networks - closeness (Definition 3.1.10) at T_1. Discussions within Section 5.2.3 theorised that the reduction in intervention school closeness at T_1 was due to an adjustment in the friendships of peer supporters; other students within the intervention networks being perturbed by the

heightened status afforded to peer supporters. Furthermore, peer supporters may befriend individuals external to their immediate social group, in an attempt to intervene in a smoking related situation; thus potentially befriending individuals that they generally may not consort with.

To investigate the differences further, the control and intervention analysis is constructed as follows:

- Precision - is there a significant difference in the mean 'correct' and 'missed' predictions for each LP method, in control and intervention schools? (Section 7.3.1)
- Method Structural Performance - is there a significant difference in the AES of each structural measure for each LP method, in control and intervention schools? (Section 7.3.2)
- School Structural Performance - is there a significant difference in the AES of each school for each LP method, in control and intervention schools? (Section 7.3.3)

7.3.1 Precision Comparison

Time	Measure	Adamic/Adar	Katz	SAB Model	PageRank	PageRank-Max
T_2	Correct	0.17	$< 0.01^*$	0.62	0.97	0.09
	Missed	0.15	0.05^*	0.61	0.93	0.20
T_3	Correct	0.16	0.02^*	0.45	0.40	0.19
	Missed	0.45	0.05^*	0.57	0.57	0.22

Table 7.26: P-Values for a comparison of precision measures for control and intervention schools, classified by LP method and time period. Starred values are significant at the 95% level.

Taking the mean correct and missed values for each method at each timestep, previously displayed in Table 7.5 and 7.6, independent samples t-tests (parametric) or Mann-Whitney tests (non-parametric) are conducted to compare control and intervention values at the 95% level; Table 7.26 displays the associated P-Values. Katz is the only LP method to indicate a significant difference, in terms of precision, for control and intervention schools (Table 7.26); the differences being significant across both time steps. This would suggest a difference in control and intervention school links, that results in the Katz method performing significantly better upon control schools (mean percentage increase values previously

displayed in Tables 7.5 and 7.6).

A difference in closeness centrality at T_1 has previously been observed between the school networks, closeness being lower in intervention schools (Section 5.2.3). The calculation of closeness centrality considers the paths between nodes in a network, this path structure also being an important component of the Katz methods. It would therefore appear that elements of the path structure may be different in control and intervention schools, the Katz method detecting this with reduced accuracy in intervention schools. While the closeness centrality has only been demonstrated to be significantly different in control and intervention schools at T_1, there may be unquantifiable residual effects of this reduced closeness in intervention schools - the impact of which still affecting Katz precision at T_2 to T_3.

7.3.2 Method Structural Performance Comparison

Further significant differences are investigated between school type, using the structural measures presented in Section 7.2.2. The average effect size's (AES) calculated for each structural measure are compared; the AES values used for comparison originally presented in Tables 7.14 and 7.15. The P-Values from the appropriate independent samples t-tests (parametric) or Mann-Whitney tests (non-parametric) are displayed in Table 7.27 (at the 95% level), with significant differences between control and intervention schools being starred.

The Katz method once again indicates a number of significant differences in its performance upon control and intervention schools, with significant differences being observed in average degree at T_2, transitivity at T_3 and reciprocity at T_3. This provides further evidence of underlying structural differences between the control and intervention networks, that are not necessarily highlighted by the conventional social network analysis of Section 5.2.3 - Katz being particularly responsive due to its paths based linking method.

Significant differences are also observed in the PR-Max method at T_2, with the AES of reciprocation being significantly lower (control: 1.59, intervention: 4.67), and APL AES being significantly higher (control: 29.45, intervention: 11.32) in control schools. This would suggest that the reciprocation is predicted significantly better in control schools at T_2, but APL is predicted significantly worse. Furthermore, the standard PR method also experiences significant differences in terms of transitivity (P-Value: 0.02) and APL

Time	Measure	Adamic/Adar	Katz	SAB-Model	PageRank	PageRank-Max
T_2	Transitivity	0.16	0.07	0.13	0.28	0.12
	Average Degree	0.35	$< 0.01^*$	0.63	0.64	0.17
	Reciprocity	0.04^*	0.06	0.08	0.08	0.02^*
	APL	0.20	0.76	0.85	0.63	0.03^*
T_3	Transitivity	0.66	0.02^*	0.09	0.02^*	0.11
	Average Degree	0.70	0.35	0.68	0.38	0.72
	Reciprocity	0.71	0.03^*	0.29	0.45	0.26
	APL	0.78	0.78	0.45	0.01^*	0.60

Table 7.27: P-Values for a comparison of structural measure AES for control and intervention schools, classified by LP method and time period. Starred values are significant at the 95% level.

(P-Value: 0.01) AES at T_3; this also suggests structural differences between control and intervention school networks. However, given the poor prediction performance of the PR method, the relevance of such findings should be considered.

From the data analysis of Section 5.2.3, the values of transitivity, reciprocity and APL are not significantly different between control and intervention schools; therefore, the reasons for the average measure effect size differences observed are unclear. Unique elements of the individual school structures may provide some basis for the differences observed, with an average of the social network metrics (such as those of APL, transitivity etc.) being unable to fully quantify the range of friendship behaviours in a network. Furthermore, it must also be acknowledged that only six intervention schools are available for testing - as such, the generalisability of comparing this data against the twelve control schools is questionable.

7.3.3 School Structural Performance Comparison

The final comparison of control and intervention schools, examines the school AES calculated in Section 7.2.3. An AES for each school was calculated and displayed in Tables 7.19 and 7.20, taking the average of all structural measures. The AES values are compared for control and intervention schools, utilising the appropriate independent samples test at the 95% level; Table 7.28 displays the mean control and intervention school AES.

Once again, the Katz method indicates a significant difference in its prediction of control and intervention schools, when an average of all structural measures is taken - interven-

Type	Time	Adamic/Adar	Katz	SAB Model	PageRank	PageRank-Max
Control	T_2	13.63	8.87*	15.04	28.12	14.24
	T_3	47.17	35.88	23.78	33.92	7.90
Intervention	T_2	11.54	29.40*	26.13	41.95	11.99
	T_3	26.31	29.94	34.37	44.58	11.66

Table 7.28: Mean control and intervention school AES values, classified by LP method and time period. Starred values are significant at the 95% level.

tion schools (29.40) being predicted significantly worse than control schools (8.87) at T_2. A significant difference is not observed in school AES at T_3. The discussion of school structure in Section 5.2.3.1 suggested an attenuation of the intervention over time, peer supporters' roles within the networks diminishing at later time steps. As such, the friendship processes which may be causing the observed differences in control and intervention schools, may also be diminishing - resulting in the Katz method producing no significant difference at T_3. However, this is contrary to the analysis of Section 7.3.1 and 7.3.2, which have both observed significant differences between control and intervention schools within the Katz method at T_3.

Overall, the comparison of control and intervention schools would suggest that the algorithms do perform differently dependent upon school type - LP methods generally performing better upon control schools. The Katz method in particular is highlighted as the method which appears to display the most sensitivity to school type, potentially suggesting some underlying differences in path structure between control and intervention schools. However, the generalisability of these results is unclear due to the small intervention school sample sizes.

Furthermore, the network size of intervention schools may also be a factor in the differences observed, with the six intervention schools generally having a larger number of trial participants than control schools - average number of nodes for control and intervention schools being 148.17 and 194.83 respectively. The analysis of Section 7.2.3 demonstrated a correlation between structural prediction accuracy and network size, however, no significant difference (at the 95% level) between control and intervention school network size is found (P-Value: 0.07). The interpretation of findings from this chapter, shall be discussed in the following section (7.4).

7.4 Results Interpretation

The analysis presented in this chapter has investigated the performance of the existing LP methods, against that of the PR-Max algorithm - comparing predictions based upon the ASSIST network data. In general, the performance of each LP algorithm gives an indication of their suitability in predicting the evolution of adolescent social networks - also providing insight into the important processes of friendship evolution. The variability of LP performance indicates that each school network has unique aspects that are important in an individual's friendship selections, this potentially altering over time.

This investigation proposes that, overall, the newly developed PR-Max method is the best performing algorithm in both precision and network structural accuracy. A particularly important feature of the PR-Max method is its ability to break existing friendship links, attempting to optimise specific eigen-centrality by removing detrimental connections. The effectiveness of this process demonstrates the importance of friendship dissolution in adolescent social networks. Therefore, it would appear that although new connections form over time, existing bonds can weaken - which the standard LP methods fail to capture effectively.

Furthermore, the analysis of Chapter 5 discussed the attenuation of the intervention over time, suggesting the reduction in effectiveness of peer supporters. The analysis presented in this chapter demonstrates that the adolescent social networks naturally evolve over time, with the underlying causes also changing. This suggests that individuals of status, who may have been selected as peer supporters, no longer hold the centrality and prestige of previous time steps - as such their ability to continue the intervention is reduced.

PageRank may be interpreted as a measure of status, as such, the PR-Max method may be considered as the process of improving said status. A further reason for the success of the PR-Max method, therefore, may be that it reflects the process of adolescents seeking to improve network eigen-centrality - emphasising the importance of status in an adolescent social network. This behaviour may be instrumental in the understanding of adolescent social network evolution, and the behaviours that result from their influence. Additionally, T_3 predictions exhibited an improvement in the accuracy of the PR-Max method; thus, potentially indicating an increase in importance of network status as adolescents mature.

The PR-Max method evidently does not completely capture all aspects of network evolution, with factors other than status also driving friendship selection. As discussed in Chapter 3, an individual's personal attributes may also have a pivotal role in adolescent connection. To investigate this further, the status aspect of the PR-Max algorithm shall be augmented to consider the attributional data also available within the ASSIST dataset. This allows for the consideration of an individual's personal attributes and behaviours in the friendship selection process, the 'behaviour based' friendship search (or behavioural search) being implemented within the framework of the SNS. The purpose of this process is to improve the link predictions made, and incorporate individual attributes into PageRank-Max process. A summary of this chapter follows in Section 7.5, with details of the behavioural search provided in Chapter 8.

7.5 Chapter Summary

This chapter has focused upon the results gained through the use of the Social Network Simulation (SNS) with the ASSIST data. The analysis was partitioned into three sectors: the specific accuracy of the link predictions made (precision); the accuracy of the overall network structures produced; and a comparison of control and intervention school predictions. The results produced offer insights into the suitability of the PageRank-Max algorithm, the construction processes of adolescent school-based social networks and the potential effect of intervention procedures within the ASSIST data.

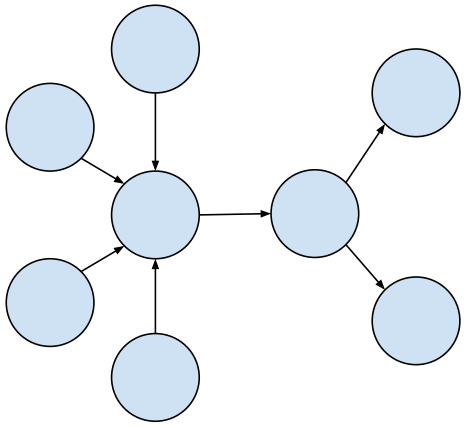
Section 7.1 investigated the precision of each algorithm, discussing the proportion of 'correct' and 'missed' predictions over a purely random linking method. All algorithms generally produced a significant improvement in the proportion of correct predictions, and a reduction in the proportion of missed predictions. However, the standard PR implementation did not consistently generate an improvement in a number of schools; thus, being classified as the poorest performing LP method. The PR-Max method was highlighted as the most precise LP method, ranking first in terms of precision for both control and intervention schools, at both predicted timesteps. The precision analysis also indicated that the accuracy of predictions in specific schools can vary with the time period being predicted, suggesting an alteration in the adolescents' friendship selection process over time.

Section 7.2 analysed the accuracy of the structures produced by the LP method. Given

that LP methods do not consider individual attributes in the evolution of networks, the overall shape of the network may be representative of the data, even if individual precision is low. Through the use of 'effect size', very few of the network structure predictions were not significantly different from the data; however, the PR-Max method was once again identified as performing well (relative to the other LP methods), especially at T_3. The Average Effect Size (AES) of each method (Section 7.2.2) reinforced the notion of a difference in adolescent network construction over time. Furthermore, the AES of each school (Section 7.2.3) demonstrated the variability of school specific network predictions, with certain networks being predicted consistently better by all LP methods. The analysis highlighted once again differences in the ASSIST networks between time periods, and the importance of network size in gaining appropriate structural predictions.

Section 7.3 examined the differences in control and intervention school predictions, with the analysis demonstrating a number of significant differences. The Katz method in particular was highlighted as performing significantly better upon control networks. This indicated that the path-based linking process of the Katz method, does not reflect intervention networks as accurately as control networks - suggesting a reduction in the number of links between individuals of paths up to three away. The analysis provided evidence of an alteration in the social network of adolescents, as a direct result of intervention methods - conventional analysis also detecting a significant difference in closeness centrality at T_1. However, with only six intervention schools available for analysis, the robustness of the performed analysis is questionable.

Finally, Section 7.4 interpreted the results of previous sections, giving an outline of elements to be further explored in this thesis. The successful performance of the PR-Max method (amongst both precision and AES metrics) was attributed to its ability to break links, suggesting that a number of existing friendships diminishing over time. The success of the PR-Max method also highlighted the importance of status in adolescent friendship selection, with status being a proxy for eigen-centrality. The importance of status was also used to explain the improvement of PR-Max predictions at T_3, with adolescents giving more credence to status in friendship selection as they mature. Section 7.4 concluded that (in their current form), LP methods only consider aspects of network structure in friendship selection; therefore, to gain more accurate link predictions, individual attributes must also be considered - this being the focus of Chapter 8.



-"A Hybrid Network"

8

Behaviour Based Link Prediction

The findings of Chapter 7 presented the results of the Social Network Simulation (SNS), identifying the PageRank-Max (PR-Max) method as the most successful in predicting the evolution of adolescent social networks (from the ASSIST data). The results also highlighted the importance of status in adolescent school-based networks, the success of the PR-Max method attributed to its individualistic optimisation of eigen-centrality. This chapter aims to investigate other individual characteristics (or behaviours) that may be important in adolescent friendship selection, based on the attributional data available from ASSIST.

The methods outlined in this chapter shall attempt to improve the PR-Max algorithm by including elements of individual behaviour, assessing the outcomes in terms of an improvement over the original PR-Max results; should the elements tested demonstrate an improvement, the criteria selected may be deemed as important in the friendship selection process. Two new PR-Max alterations are proposed. The first restricts the search space of the searching agent, thus restricting the selection of potential new connections

(Behavioural Search). The second alters the calculation of the PageRank matrix ($\bar{M}$) to consider both friendship ties and behavioural similarities (Behavioural PageRank). The outcomes of each alteration shall be presented, with accompanying discussion. Additionally, the Behavioural PageRank method provides the opportunity to assess the interplay between social structure and behaviour, this being investigated further in Chapter 9.

This chapter is structured as follows: a description of the investigation process and the schools selected for further analysis is given in Section 8.1; the restricted search (or behavioural search) algorithms are described and tested in Section 8.2; the results of the behavioural PageRank algorithms are presented in Section 8.3; finally, a summary of the findings is documented in Section 8.4.

8.1 Investigation Outline

To investigate the effects of including attributional data, the SNS logic is adjusted to incorporate the new behavioural elements for testing (discussed further in Section 8.2 and Section 8.3). Following this, the modified SNS is rerun with the ASSIST data, to generate new network predictions. The new 'Behavioural Search' and 'Behavioural PageRank' predictions, are compared with the original predictions from the PR-Max method - assessing whether the incorporated elements offer any improvement.

Before addressing the construction and results of the proposed PR-Max adjustments, procedures necessary to the investigation process are discussed. Section 8.1.1 details the school networks to be used for the behavioural PR-Max investigation, Section 8.1.2 presents the individual attributes selected for further analysis, and Section 8.1.3 introduces the Levenshtein distance. The Levenshtein distance is implemented throughout the methods to be discussed, as such, an outline of its purpose and origins is required.

8.1.1 School Network Selection

Given the number of networks available in the ASSIST data, it would appear excessive to rerun the SNS for each individual network to assess the impact of the newly proposed methods. As such, four schools have been taken for the purpose of testing; the schools selected are 12, 33, 71 and 74. These schools have been chosen to best represent the AS-

SIST data, providing a balance of networks for the investigation. The schools selected, were chosen based upon the following criteria: school matching, network size, missing data and previous PR-Max performance. This section describes the school selection process further.

During the recruitment process of ASSIST, schools with similar attributes were paired (Holliday, 2006); one school being administered the intervention, while the other acted as control. The matching procedures examined network size, region and free school meal entitlement - allowing investigators to compare the outcomes of the intervention under similar conditions. Given that the ASSIST schools had been matched in this manner, it was felt appropriate to use paired schools for the current behavioural investigation. However, only four matched pairs were available in the 18 network schools:

- Schools 33 and 12;
- Schools 41 and 34;
- Schools 63 and 76;
- Schools 71 and 74.

The matched school pair of 63 and 76, had particularly large network sizes (school 63: 236 & school 76: 254). As these were not representative of the data as a whole, they were omitted from the following investigation. The remaining matched pairs, ranged from moderately large network sizes (school 34: 200) to moderately small (school 71: 102).

Aside from network size, it is also of interest to consider the precision of the previous PR-Max predictions. From Tables 7.3 and 7.4, respectively, school 71 performed particularly well (50.24%) in terms of precision at T_3, while school 74 had the lowest precision of intervention schools (7.19%); therefore this school pair provides balance in terms of previous PR-Max performance. Moreover, school 74 experienced high prediction precision at T_2 (46.59%, Table 7.1), which subsequently dropped at T_3; therefore, it is of interest to explore whether behavioural attributes can improve its link predictions at T_3.

Similarly, PR-Max school 12 predictions have the highest precision of intervention schools at T_3 (49.08%, Table 7.4), while its matched pair (school 33) has the lowest precision of

	12	33	71	74
School Type	Intervention	Control	Control	Intervention
Network Size	164	153	102	123
Region	Bristol	Bath	Cardiff	Newport
Free School Meal (%)	6.00	6.00	26.10	25.90
PR-Max Precision T_2 (%)	38.53	40.89	42.18	46.59
PR-Max Precision T_3 (%)	49.08	7.19	50.24	7.19
Missing Data T_3 (%)	9.76	3.27	4.90	1.63

Table 8.1: Summary of school selection criteria for the four chosen networks. Free school meal entitlement figures are taken from Holliday (2006). Precision is expressed as a percentage improvement in correct predictions over the random method.

control schools at T_3 (7.19%, Table 7.3); this once again provides balance in terms of PR-Max prediction accuracy. The matched pair of school 41 and 34 were disregarded due to their high levels of missing data at T_3 (school 41: 16.86% & school 34: 15.50%), which may affect the interpretation of precision. Therefore, schools 12, 33, 71 and 74 are the four schools selected for further analysis; Table 8.1 summarises the the selection criteria for each of the chosen schools.

8.1.2 Attributes and Behaviours

As previously discussed in Chapter 5, the data provided by DECIPHer includes a number of variables. To focus the behavioural investigation upon the importance of specific attributes, a number of variables have been selected for further analysis. They are as follows:

- Gender - The analysis of Section 5.2.3 highlighted the importance of gender in the school structures of the ASSIST data; literature also suggests gender as a key factor in adolescent friendship selection (Clark, 1992; Cohen, 1977; Osgood et al., 2013; Parker & Asher, 1993). To improve the PR-Max process, consideration shall be given to gender in link decisions.

- Smoking - The main focus of ASSIST, was to investigate the impact of smoking behaviours in reference to social structure. To assess the importance of smoking, the behavioural PR-Max alterations will consider smoker similarity in network evolution. The work of Steglich et al. (2012) finds that individuals initially smoke based

on friendship influences in early adolescence, but as the adolescents mature, they select friends based on similar smoking habits.

- Nominations - Prior to commencement of the study, each individual was asked to nominate those persons who were influential in their network (as discussed in Section 5.1). This resulted in a nomination score being given to each individual, with those obtaining the highest scores being selected as peer supporters. Given the importance of status in an adolescent's social networks (as discussed in Chapter 7), consideration shall also be given to an individual's nomination score.

- Form - In the literature review of Chapter 3, proximity was highlighted as an important factor in communication. Evidently, the school networks naturally facilitate close student proximity; however, many schools selected also possess a form group structure. This subdivides the student population into classes, potentially meaning that individuals interact with certain groups more regularly. Therefore, the investigation shall also consider proximity through form group structure.

The variables presented allow for the investigation of specific aspects said to be pertinent in adolescent friendship selection. Additionally, consideration shall also be given to *all* variables available in the data set; this giving a representation of student responses to the administered questionnaires. The purpose of this, is to investigate whether similarities in questionnaire response is important in friendship selection. To consider all questionnaire responses would require a great deal of experimentation; thus, to reduce computation time, all variables shall be grouped and examined through the use of the Levenshtein distance.

8.1.3 Levenshtein Distance

The reason for investigating all variables from the ASSIST questionnaires, is that the responses represent the background and opinions of the participating individuals. This may be useful in the development of an improved LP algorithm, as individuals with similar opinions may be more likely to befriend one another. At T_1 there are 73 variables for each individual, and at T_2 there are 120 variables; free text variables have been excluded from this analysis. As previously discussed (Section 5.1.3), the variables relate to: smoking habits, availability of cigarettes, family life, personal attitudes, personal relationships, family affluence and school performance. For confidentiality reasons, the full list of vari-

ables in the data set cannot be published.

By finding individuals with similar questionnaire responses, the SNS may potentially be able to match "similar" individuals for improved linking. To conduct this matching process, literature relating to music information systems is explored. The likes of Shazam (2013) and SoundHound (2013) allow users with mobile phones (or other internet enabled devices) to find the name and artist of a song, just by recording a short audio sample of the music. When a user wishes to identify a song using Shazam, the audio clip is matched against a large database of music.

Evidently, to hold such a large database of songs would be costly, and require a large amount of storage. Therefore, to reduce the required storage, and improve the efficiency of search queries, each song is given a 'fingerprint' (Wang, 2006). The fingerprints are created by analysing a song and taking its important 'spectrogram peaks', this gives a unique arrangement of frequencies (or a 'constellation map') over the length of the song (Casey et al., 2008; Wang, 2003). When a music clip is verified against the Shazam database, the process attempts to find a fingerprint with the same distribution of spectrogram peaks over time - returning the closest song match.

Taking the concept of a fingerprint, each individual's responses to the ASSIST questionnaires are encoded as a data string. An agent's responses to each variable are coded as numeric values between 1 and 9, where 9 is reserved for a missing response; if a numeric value does not offer enough response options, a character may be used instead (9 retaining its marker as a missing value). This gives an alphanumeric string that represent an individual's questionnaire responses, for the purpose of this research, the data string created shall be referred to as an agent's fingerprint. Two fingerprints for each adolescent in ASSIST is created, the first documenting questionnaire responses at T_1, while the second details responses at T_2. An example ASSIST fingerprint, with annotations, is given in Figure 8.1.

As there may be very few people in the data set who have the exact same fingerprint, similarity between fingerprints must be assessed. If the values within the fingerprint were represented by a scale, a similarity could be quantified in terms of an increase of decrease in a specified variable field. For example, if individual i smoked 5 cigarettes per day, and student j smoked 6 cigarettes per day, then the students would only be 1 cigarette apart in terms of the number smoked per day. However, the majority of the fields in the raw data

are nominal; as such, similarity cannot be quantified in this manner.

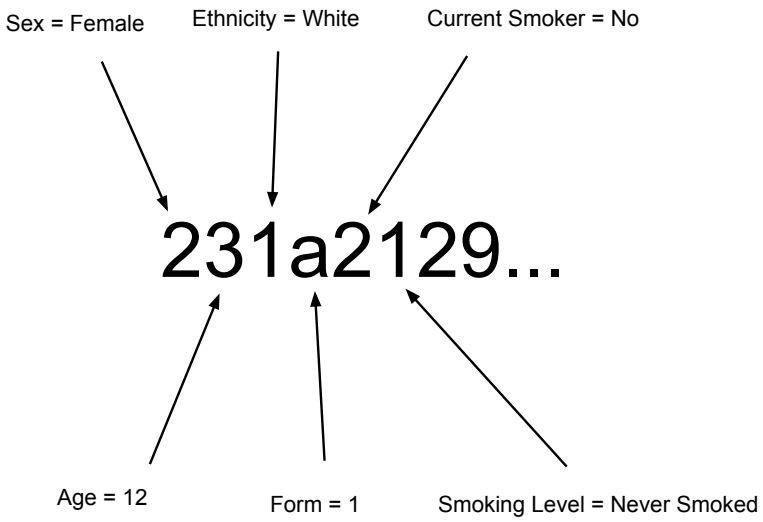


Figure 8.1: Partial fingerprint of student 15010 from the ASSIST data, annotations display the variable name and value.

15008: 231a2119...

15010: 231a2129...

Figure 8.2: Partial fingerprints of students 15008 and 15010 from the ASSIST data, requiring one edit (highlighted in red) to make them equivalent.

To measure similarity in terms of the ASSIST fingerprints, the Levenshtein distance is used. The Levenshtein distance is a metric used to quantify the distance between two strings (Levenshtein, 1966). The distance is calculated by assessing the number of single character edits needed to make two strings the same, allowing the use of insertions, deletions and substitutions. Figure 8.2 displays two partial ASSIST fingerprints for students 15008 and 15010, requiring one edit to make them equivalent; therefore, the Levenshtein distance between these two strings is one. An alternative string distance metric is the Hamming distance (Sankoff & Kruskal, 1983), however, this does not allow insertion or

deletion properties, allowing only substitutions (Navarro, 2001); the Hamming distance may be considered an upper bound of the Levenshtein distance for strings of equal length.

An explanation of the implementation of the Levenshtein distance within the newly proposed algorithms, is detailed further in Section 8.2 and Section 8.3. With an outline of the necessary elements of the investigation complete, the following section (8.2) describes the development of the new restricted search PR-Max method.

8.2 Behavioural Search

The 'behavioural search' aims to restrict the search space of the PR-Max method, reducing the number of testing agents with whom a searching agent may make a friendship alteration. This section outlines the adjustments to the PR-Max method, and presents the results gained from running the altered method upon the four selected test schools. The discussion of behavioural search is structured as follows: a general outline of the method is provided in section 8.2.1; the effect of reducing the search space by gender is presented in 8.2.2; the effect of considering smoking in the PR-Max algorithm is documented in Section 8.2.3; and the overall conclusions of the behavioural search are presented in Section 8.2.4.

8.2.1 Behavioural Search Outline

Recall that the original PR-Max method altered a friendship tie and investigated the effect to an individual's personal PR (as discussed in Section 6.2.2). The searching agent (the individual seeking to change a friendship tie), scanned all available testing agents - executing the friendship change that proved most beneficial to their eigen-centrality. During this process, all available testing agents were searched; this potentially included agents who the searching agent may never realistically consider. To attempt to improve the PR-Max method, the list of possible testing agents is reduced - the reduction based upon specific attributional or behavioural criteria.

To implement this process, the agent-specific fingerprints must be read into the SNS. Each agent has a unique identification number, taken directly from the ASSIST data. On initialisation of the simulation, the fingerprints and social network connections are matched

based on the unique identifier; this creates individual agents that represent the ASSIST adolescents at a given time period. The original process of selecting a searching agent is retained, with an agent being selected at random (with a negative exponentially distributed inter-event time) to make a friendship change.

The behavioural search process is outlined as follows:

- On initialisation of the simulation, the user selects the specific criterion (ζ), to be used for search space restriction - for example, gender;
- The simulation begins, reading in the appropriate fingerprints and connections from a purpose built database;
- The agents are created;
- The Levenshtein distance between all agents' fingerprints is calculated and stored;
- An agent is chosen at random to be the searching agent (agent i) ;
- The searching agent's fingerprint is examined to find their specified criterion value ζ_i;
- The list of testing agents is reduced to consider only those agents who match specified search criterion, therefore, agent j is only considered if $\zeta_j = \zeta_i$;
- Agents with the lowest Levenshtein distance from agent i, but who do not satisfy the previously defined search criterion, are added to the testing list;
- Agents with an existing connection to i are added to the testing list;
- Agent i iterates through the testing list to find the friendship change which provides the largest personal PR improvement;
- The selected connection is altered;
- A new agent is selected at random, and the friendship alteration process repeats.

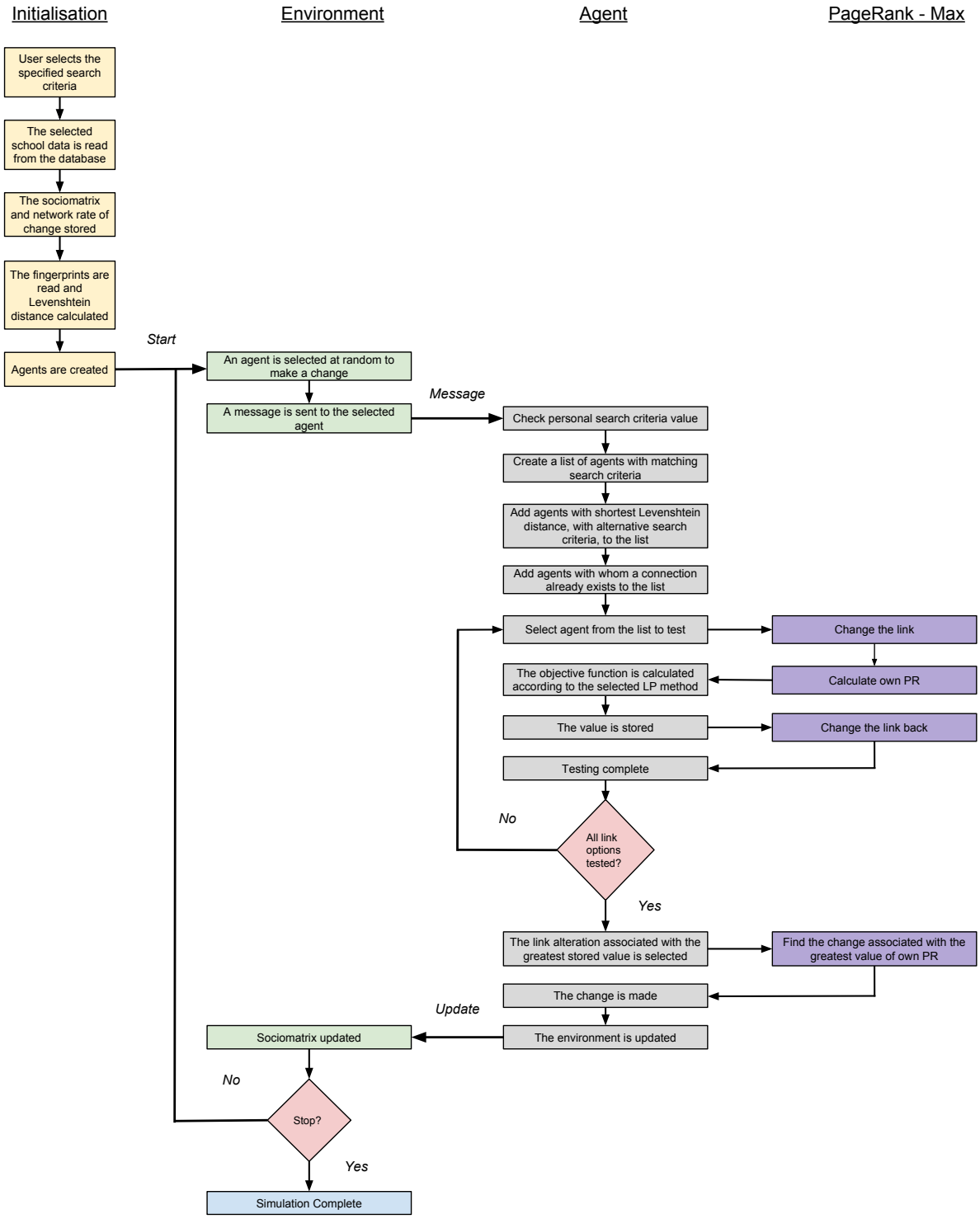


Figure 8.3: PR-Max simulation logic, including the newly created behavioural search elements.

A diagram of the behavioural search logic is presented in Figure 8.3. The first specific search criterion to be investigated is gender. Therefore, following the logic described above (in the context of gender), if the searching agent is male, only male testing agents shall be considered. Female agents shall also be considered, but only those with the lowest Levenshtein distance from the searching agent. The searching agent's own connections shall also be tested, to assess whether disconnection proves more beneficial than generating a new connection. The following section (8.2.2) presents the effects of restricting the search space by gender.

8.2.2 Gender Search

For the behavioural search analysis, only precision measures shall be used to evaluate improvement in an algorithms performance, as structural analysis becomes less meaningful. This is because agents in the SNS now embody specific attributes from the ASSIST data, with the behavioural search algorithms attempting to improve link predictions based on agent specific criteria; as such, the specific position of particular agents in the network is important. This means that the predicted network structure may be representative of the true network, but the specific agents may be in the wrong network positions.

Gender is the first behavioural search criterion to be tested. Ten runs of the SNS with a restricted gender search, are conducted for each of the four test schools (at each timestep). Table 8.2 displays the results of the gender search PR-Max method, with values expressed as a percentage improvement over the original PR-Max results. To compare the results of the restricted PR-Max and the original PR-Max method, the appropriate repeated measures statistical tests are conducted; significant differences at the 95% level are highlighted in Table 8.2.

From Table 8.2, there is a significant improvement (over the original PR-Max method) in the percentage of missed predictions at T_2 for schools 12 (-2.18%), 71 (-2.56%) and 74 (-2.23%); however, there is no significant difference in the percentage of correct predictions.

The observed decrease in missed predictions, without an increase in correct predictions, suggests an increase in the actual number link predictions made. However, the gender search predictions are generated with the same mean value and inter-event time of those from the PR-Max method. Furthermore, the PR-Max and gender search simulations have

School	Measure	T_2	T_3
12	Correct	−0.26	0.79
	Missed	−2.18	−1.43
33	Correct	0.26	1.77
	Missed	−0.81	1.18
71	Correct	1.02	−1.60
	Missed	−2.56	0.11
74	Correct	−0.09	1.41
	Missed	−2.23	−0.76

Table 8.2: Gender search 'correct' and 'missed' results, expressed as a percentage increase over the PR-Max method. Green values indicate a significant improvement over PR-Max, while red values indicate a significant deterioration. No colour indicates no significant difference from PR-Max predictions.

been run with the same random number stream for consistency. Therefore, the gender search simulations should not be producing substantially more link predictions than the PR-Max method. Rather, the increase observed may be due to agents arriving at an optimal "friendship" state later than in the original PR-Max method.

Consider the situation where a searching agent has achieved their maximum eigen-centrality, and therefore changing any of their connections causes a decrease to their PR. The PR-Max method requires the searching agent to make a change irrespective of their current PR, and as such, must accept the friendship change that causes the lowest PR decrease. If the same searching agent is selected again to make a change (at a later stage in the simulation), and the network around the agent has not been altered substantially, the optimal friendship decision is to revert the changed link back to its original state. The link state may continue to revert back and forth in this manner for the remainder of the simulation.

Due to the process in which precision is calculated, a link that is changed and then reverted back to its original state cannot be detected. This gives the impression of fewer link predictions being made. As the gender search restricts the available testing agents, it may take longer for searching agents to arrive at their optimal friendship state; thus, fewer link changes are reverted to their original state, and the number of link predictions appears increased over the PR-Max method. Therefore, in the context of the T_2 gender search results, less link reversions are made, which causes a decrease in the number of missed predictions; however, the overall proportion of correct predictions does not significantly

improve.

Examining the gender search results at T_3, presents a significant increase in correct predictions for schools 12 (0.79%), 33 (1.77%) and 74 (1.41%). A significant decrease in missed predictions is also observed in schools 12 (-1.43%) and 74 (-0.76%). This indicates that the gender restricted search, significantly improves precision at T_3 for schools 12 and 74. However, a significant increase in the percentage of missed predictions is also demonstrated for school 33. This stipulates that for school 33, although the accuracy of predictions is increased, the restricted search is causing the altered PR-Max method to miss more link changes. This may be caused by the reversion of link predictions previously discussed, but with agents arriving at their maximum eigen-centrality state sooner than the original PR-Max method (for school 33).

The gender search results highlight three key elements to friendship selection:

- Gender - making the agents link primary based upon gender, has significantly increased the overall precision in a number of schools. This highlights the importance of gender homophily in friendship selection, predictions being particularly improved at T_3 in a number of schools.

- Attribute similarity - allowing searching agents to consider testing agents with alternative gender, but who responded to questionnaires in an otherwise similar manner, has also contributed to the results presented. This would suggest some importance of opinion and attribute similarity in friendship selection.

- Link disconnection - while the friendship search space is restricted, the ability to break existing connections remains unaltered. As such, the disconnection element of the PR-Max algorithm is still aiding the improvement of predictions, allowing agents to disconnect links that reduce their own PR.

While the restricted gender search has improved link predictions in a number of schools, a significant increase to correct predictions is not observed at T_2. Furthermore, while some of the precision improvements observed are significant, they are not particularly sizeable. To investigate whether more precise link predictions may be generated, the search space is restricted by smoking (Section 8.2.3).

8.2.3 Smoker Search

The smoker search restricts the search space of a searching agent by smoking level. The method works in the same manner as the gender search, allowing testing agents with alternative smoking level to be considered, should they have a small Levenshtein distance; disconnecting an existing link is also considered. There are five possible values of an agent's smoking level:

- 1 - never smoked or currently a non-smoker;
- 2 - less than 1 cigarette per week;
- 3 - between 1 and 6 cigarettes per week;
- 4 - more than 6 cigarettes per week;
- 9 - did not answer, more than one option selected or missing from data collection.

Therefore, when a searching agent compiles a list of possible testing agents, those with an equivalent smoking level are selected. For agents with missing or incomplete data (smoking value: 9), it would appear incorrect to only consider other agents with missing data (full missing data statistics may be found in Table 5.6). As such, on initialisation of the SNS, any agent with a smoking level of 9 is assigned a new smoking level. The assigned smoking level is sampled from a distribution, based upon the existing smoking level proportions in the simulated school. This reassignment of missing data only occurs for the smoking variable, other variables with missing data in the fingerprint retain their missing marker. Table 8.3 displays the results of the smoker search.

The results of Table 8.3 at T_2 indicate that the percentage of correct predictions has significantly reduced for schools 12 (-2.41%), 33 (-0.98%) and 74 (-2.17%), with no significant difference observed for school 71. While the percentage of missed link changes is reduced significantly in school 33 (-0.94%) and 71 (-3.15%), overall, it would appear that restricting links to those of an equivalent smoking level is not appropriate at T_2.

The smoker search appears more successful at T_3, school 33 indicating a significant increase in correct predictions (2.81%), and 74 demonstrating a significant improvement in

School	Measure	T_2	T_3
12	Correct	−2.41	−1.92
	Missed	−1.43	−1.74
33	Correct	−0.98	2.81
	Missed	−0.94	0.42
71	Correct	−1.43	−1.89
	Missed	−3.15	−1.93
74	Correct	−2.17	2.36
	Missed	−0.89	−0.50

Table 8.3: Smoker search 'correct' and 'missed' results, expressed as a percentage increase over the PR-Max method. Green values indicate a significant improvement over PR-Max, while red values indicate a significant deterioration. No colour indicates no significant difference from PR-Max predictions.

both correct (2.36%) and missed predictions (-0.50%). This indicates that friendship selection based on smoker similarity may be more prominent at T_3 than at T_2, especially within schools 33 and 74.

From the data analysis of Chapter 5, school 33 has a large proportion of smokers at T_3 (36.30%); this may explain the increased percentage of correct predictions, with smoking being a prominent aspect of the school's culture, and subsequently the friendship selection process. However, the percentage of smokers in school 74 at T_3 (24.79%) is less than the intervention average (25.55%). Furthermore, School 71 has the highest percentage of smokers at T_3 (44.12%), but the smoker search performs significantly worse in terms of correct predictions (-1.89%). Therefore, it would appear that basing the majority of friendship decision solely upon smoking similarity is not wholly appropriate.

The cause of the significant precision reduction in a number of schools (especially at T_2), may be a result of the search space categories. In the gender search, there were two distinct pools of agents (male and female) from whom a searching agent could test potential connections. As there is generally an equivalent split of male and female students in the ASSIST schools (excluding girls school 40), the male and female test lists of agents would generally be of equal size. However, in many schools there is a greater proportion of non-smokers than smokers, with smokers being further subdivided into three different categories (less that 1 cigarette, between 1 and 6 cigarettes or more than 6 cigarettes per week).

Dividing the smokers across four categories, creates one large pool of non-smokers, and a number of smaller pools of tiered smoking agents. In terms of the smoker search, this means non-smoker searching agents have a large selection of testing agents, while smoker searching agents have a substantially smaller availability of testing agents. To investigate the effect of the size restriction further, an alternative search is conducted.

The smoker Levenshtein search attempts to create equivalent sized small pools of testing agents, for all searching agents. The process is described as follows:

- A searching agent is selected;
- The searching agent selects testing agents with same smoking level (only testing agents with no current connection to the searching agent are considered);
- Of the selected testing agents, only those with the lowest Levenshtein distance (to the searching agent) are considered;
- If only one agent is selected, the next group of agents with the lowest Levenshtein distance is also selected (to ensure that at least two new connections are considered);
- Current connections are also added to the testing list, to assess the effect of disconnection;
- The PR-Max process is carried out as normal.

An illustration of the new restriction procedures is displayed in Figure 8.4, with Table 8.4 displaying the precision increase of the selected test schools.

The results of the smoker Levenshtein search, do not appear to improve the correct predictions at T_2 (Table 8.4), with some minor significant improvement in the percentage of missed predictions in schools 12 (-1.03%), 33 (-0.55%) and 71 (-1.30%). However, predictions at T_3 experience greater improvement. School 33 demonstrates a significant improvement of 6.76% in correct link predictions, while school 74 experiences a 6.63% increase.

Schools 33 and 74 demonstrated moderate success at T_3 with the smoker search method (Table 8.3), and greater success with the smoker Levenshtein search. The results provide

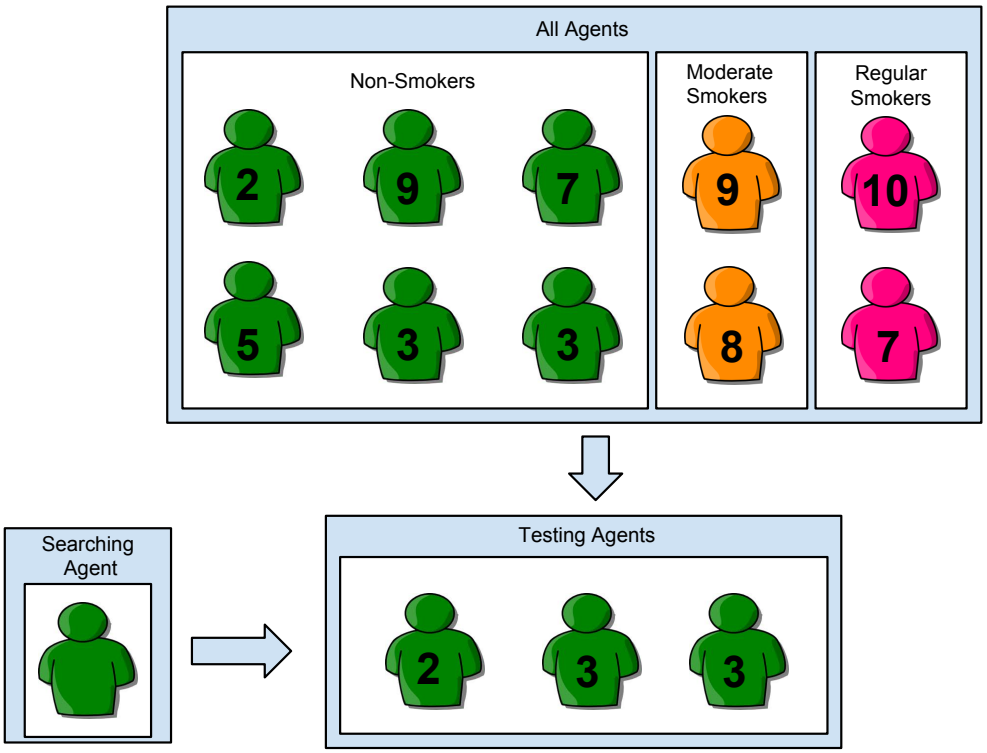


Figure 8.4: Illustration of a non-smoker searching agent, selecting non-smoker testing agents with lowest Levenshtein distance - a stipulation of the process being that more than one testing agent must be considered. The values stamped on each agent represent their Levenshtein distance from the searching agent.

School	Measure	T_2	T_3
12	Correct	−1.03	−1.54
	Missed	−1.03	−3.07
33	Correct	−0.55	6.76
	Missed	−0.55	−0.36
71	Correct	−1.30	−2.77
	Missed	−1.30	−3.10
74	Correct	0.43	6.63
	Missed	0.43	−0.92

Table 8.4: Smoker Levenshtein search ‘correct’ and ‘missed’ results, expressed as a percentage increase over the PR-Max method. Green values indicate a significant improvement over PR-Max, while red values indicate a significant deterioration. No colour indicates no significant difference from PR-Max predictions.

evidence of the importance of similar smoking behaviours in adolescent friendship selection, specifically in schools 33 and 74. Schools 33 and 74 performed poorly in terms of precision with the original PR-Max method at T_3; this demonstrates the potentially improving effect that considering personal attributes may have to link predictions - suggesting that just considering network structure may not be wholly appropriate.

Furthermore, the results indicate that the small pool of testing agents generated (for each searching agent) in the smoker Levenshtein search, are partially representative of the true friendship considerations in the networks of schools 33 and 74. However, the significant decrease in correct predictions at T_3 for schools 12 and 71, demonstrates that the smoker restrictions are not appropriate for these schools. Therefore, the smoker based searches have highlighted the following key points:

- The importance of specific attributes in the friendship selection process can vary between schools;
- Similar smoking behaviours may be more important in the friendship selection process at T_3 than at T_2;
- Restricting the search space to consider those with small Levenshtein distance, can improve predictions - suggesting that status amongst those of similar attributes and opinions is important.

A summary of the behavioural search findings is presented in the following section (8.2.4).

8.2.4 Behavioural Search Summary

The behavioural search attempted to improve link predictions by restricting the search space of the searching agent. This gave the searching agents the opportunity to optimise their eigen-centrality, amongst individuals that share similar attributes and values. Bar graphs of the increase in correct predictions for all test schools, for each method, at T_2 and T_3 are presented in Figures 8.5 and 8.6 respectively.

The results of the behavioural search have demonstrated that attributes and opinions appear to have a greater improving effect at T_3 than at T_2. The implications appear consistent with the findings of Steglich et al. (2012). On analysis of the ASSIST data, Steglich et al. (2012)

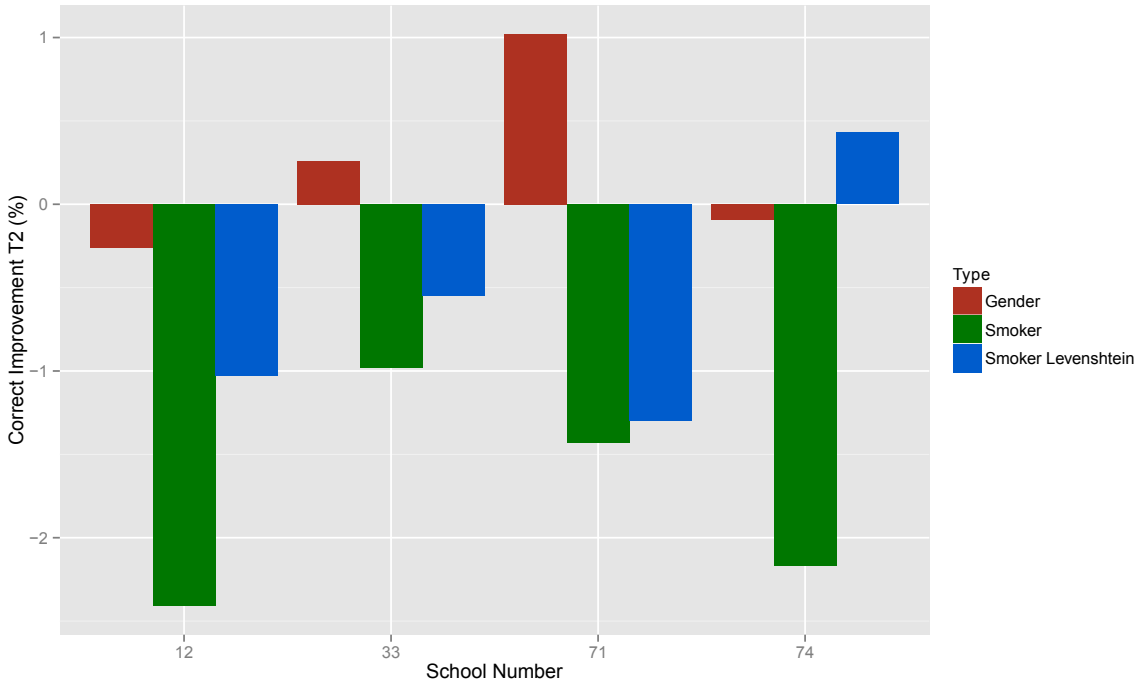


Figure 8.5: Bar chart of the percentage increase in correct link predictions over the PR-Max method at T_2.

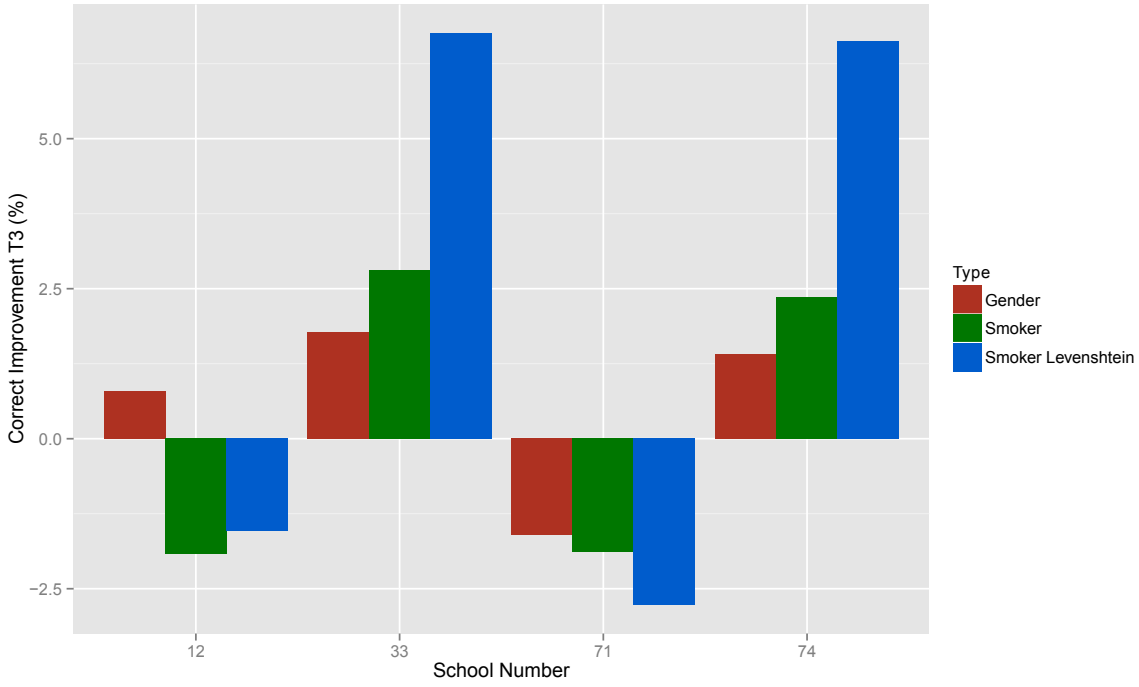


Figure 8.6: Bar chart of the percentage increase in correct link predictions over the PR-Max method at T_3.

concluded that in early adolescence, peers influence individuals to smoke; however, in later adolescence, individuals select friends who have the same smoking behaviours. This active search for friendship similarity, may be the cause of the observed increase in behavioural search precision at T_3.

From Figures 8.5 and 8.6, it would appear that the gender search produces minor increases and decreases in the percentage of correct predictions. The smoking related searches produce greater shifts in precision, relatively large increases observed with the smoker Levenshtein search at T_3 in schools 33 and 74. This suggests that basing friendship decisions on smoker similarity, and not just network stucture, has the potential to produce substantial gains in precision; however, the results are variable.

The variability in smoking related searches observed, may be a product of the shifting opinions related to smoking. Evidently, an adolescent's gender is unlikely to change over time, as such the information being given to the SNS is assumed accurate across the whole period being simulated. Smoking behaviours, however, are subject to change, meaning that the smoking information at the start of a simulation, may be obsolete by the later stages of the run. Therefore, the smoking-based link predictions may be improved if the variability of smoking behaviours could be addressed - this being the particular focus of Section 8.3.

To approach the issue of smoking behaviour changing over time, the SNS would be required to assess both potential link changes **and** potential alterations to smoking behaviour. This would suggest that restricting the search space may not be appropriate, as agents may wish to change their smoking behaviour to emulate an agent not directly included in their restricted search. As such, behavioural and friendship changes, along with their impact to PageRank, must be considered simultaneously.

The behavioural search has highlighted a number of key elements related to friendship selection, while providing further insight into the inner workings of the PR-Max method. To build upon the knowledge gained from the behavioural search, and incorporate behaviour and link changes, a new method shall be created - the Behavioural PageRank-Max (BPRM) method. Details of the BPRM process are given in Section 8.3.

8.3 Behavioural PageRank

Section 8.2 demonstrated the effect of considering individual behaviours and attributes in adolescent friendship selection, highlighting great variability upon the inclusion of changeable behaviours (such as smoking). To investigate the effect of changing individual behaviour further, the Behavioural PageRank-Max (BPRM) shall be created - attempting to include elements of similarity between agents, through the explicit calculation process of the PR. The changeable behaviour to be investigated is that of smoking, as this is the main focus of the ASSIST data; however, the framework detailed may be applied to any changeable behaviour required (such as alcohol consumption, or even levels of happiness).

The details of the BPRM investigation are outlined as follows: the alterations necessary to the SNS logic are described in Section 8.3.1; the precision of the BPRM method, prior to including changeable behaviours, is documented in Section 8.3.2; an investigation of changing smoking behaviours in conjunction with gender and ethnicity, form group, peer supporter nominations and Levenshtein distance, is then conducted in Section 8.3.3; finally, the conclusions gained from the BPRM method are discussed in section 8.3.4.

8.3.1 BPRM Calculation

Recall the calculation process of PageRank, detailed in Section 6.1.4. The original sociomatrix of links (X) is manipulated into the adjusted matrix of in-links relative to out-links (M), with the PageRank being calculated from the matrix $\bar{M}$:

$$\bar{M} = (1 - d)Q + dM \tag{8.1}$$

where Q is the matrix of fractional elements $\frac{1}{n}$, n is the number of agents in the network and d is the dampening factor. Finding the principle eigenvalue of $\bar{M}$, along with its associated eigenvector, gives the vector of PageRanks required. To generate a unique positive vector of PageRanks, $\bar{M}$ must be stochastic and irreducible.

The PR-Max method discussed in Section 6.2.2, altered the sociomatrix (X) during the course of the simulation, assessing the impact of friendship changes relative to an agent's personal PageRank. The adjusted matrix (M) changes based upon the sociomatrix updates,

but the matrix of fractional elements Q is never altered. The sole purpose of Q is to ensure that the necessary conditions for PR calculations are satisfied. As such, Q may be replaced by $\bar{Q}$, where $\bar{Q}$ incorporates attributional data, should the required conditions for the PR calculation remain satisfied.

To incorporate individual attributes into the $\bar{Q}$ matrix, similarity is once again considered. First, q_{ij} is calculated, such that:

$$q_{ij} = \begin{cases} 1 + kn, & \text{if } i \neq j \\ 1, & \text{otherwise} \end{cases} \tag{8.2}$$

where k is the number of similarities between the agents i and j. For example, if i and j possess the same smoking level, then they would have one similarity, meaning $k = 1$. The entries of the similarity matrix ($\bar{Q}$) are then given by:

$$(\bar{Q})_{ij} = \frac{q_{ij}}{\sum_{i=1}^{n} q_{ij}} \tag{8.3}$$

The resultant $\bar{Q}$ matrix is symmetric, as similarities are undirected. The reason for the calculation of $\bar{Q}$ in this manner, is due to $\bar{Q}$ being required to remain stochastic and irreducible; therefore, if there are no similarities between agents, then $\bar{Q} = Q$. Previous research has investigated the application of weights to the PR calculation process, whereby alterations to Q are made (Ding, 2011; Xing & Ghorbani, 2004; Yan & Ding, 2011); however, this has not been through the consideration of attribute similarities, or in the context of human behaviour.

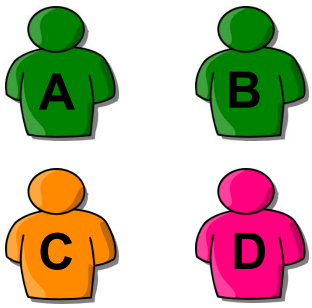


Figure 8.7: Example of agent similarity, green agents are non-smokers, orange agents are moderate smokers and red agents are regular smokers.

The decision to use n as the value to represent a similarity, is to weight the existence of a similarity highly enough to differentiate itself from no similarity. From the group of agents depicted in Figure 8.7, as agent A has the same smoking level as agent B (and since no other agent similarities exist), then:

$$\bar{Q} = \begin{pmatrix} \frac{1}{8} & \frac{5}{8} & \frac{1}{4} & \frac{1}{4} \\ \frac{5}{8} & \frac{1}{8} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{8} & \frac{1}{8} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{8} & \frac{1}{8} & \frac{1}{4} & \frac{1}{4} \end{pmatrix},$$

placing particular emphasis upon the smoking similarities observed.

Tests were conducted to obtain the appropriate weighting for a similarity in the calculation of $\bar{Q}$. If q_{ij} was calculated simply with $1+k$ (instead of $1+kn$), within the network sizes of the ASSIST data, very little impact to friendship changes was observed. By contrast, the adjusted sociomatrix (M) is particularly sensitive to a change in connections - PageRank varying greatly dependent upon the friendship selections within M. Therefore, to give an appropriate level of sensitivity to attribute similarities, $1 + kn$ was found to be a suitable selection.

The effect of changing an agent's smoking behaviour, upon both their PageRank and potential friendship connections, must be evaluated. The simulation tests the effect of increasing and decreasing an agent's smoking level, and also the effect of exacting no change to smoking level. The simulation also tests the effect of such changes to friendship selection, and consequently the agent's own PageRank. The combination of changes that provides the best improvement to an agent's personal PageRank is selected; a more detailed account of the procedure is as follows:

- Prior to initialisation, the user selects the similarities to be considered in the calculation of $\bar{Q}$ and the specific changing behaviour to be investigated - for the purposes of this study, the changing attribute is smoking;

- On initialisation, the agent fingerprints and connections are read from a database, with the specific ASSIST agents being created;

- The similarity matrix ($\bar{Q}$) is created, examining the agent fingerprints for the selected attributes to be considered;
- The simulation begins, with a searching agent being selected at random to make a friendship or behavioural alteration;
- The searching agent's current PR is calculated and stored (no network change, no smoking change);
- The searching iterates through all testing agents to assess link changes, storing the change producing the highest PR (network change, no smoker change);
- The searching agent's smoking level is increased and $\bar{Q}$ is recalculated (if possible);
- The searching agent's new PR is calculated and stored (no network change, an increase to smoking level);
- The searching agent iterates through all testing agents to assess link changes, the change producing the highest PR is stored (network change, smoker increase);
- The searching agent's previous smoker level is restored;
- The searching agent's smoking level is decreased (if possible);
- The searching agent's new PR is calculated and stored (no network change, a decrease to smoking level);
- The searching agent iterates through all testing agents to assess link changes, the change producing the highest PR is stored (network change, smoker decrease);
- The searching agent's previous smoker level is restored;
- The change producing the highest searching agent PR is selected;
- If both a smoking and link change are required, then the smoking change is exacted first - with the link change being stored, should the agent receive another opportunity to make a change;

- A new agent is selected to be the searching agent and the searching process repeats.

A diagram of the BPRM process is presented in Figure 8.8. If the smoking level of the searching agent is at a maximum, then increases shall not be considered; similarly, if the searching agent is at the minimum smoking level, decreases shall be ignored. The changeable behaviour is only increased or decreased by one level, as this is standard practice in SAB modelling. The SAB method attributes this to individuals being more likely to experience gradual changes to behaviour, as opposed to drastic changes in opinion over time (Snijders et al., 2007b; Steglich, 2013; Steglich et al., 2012).

Restricting the searching agent to perform one (behavioural or friendship) change per selected instance, is to maintain consistency with the original PR-Max algorithm. In the PR-Max method, only a single link change can occur in a designated instance, with the overarching inter-event time being based upon the number of singular link changes in the network. In the BPRM, both behavioural changes and link changes are considered. Therefore, the mean inter-event time of changes in the network (previously discussed in Section 6.2.1), is calculated to include the number of smoking changes observed. This means an increased level of searching agent changes occur in the SNS.

In the BPRM logic, a behavioural change takes precedence over a link change. In the initial exploratory tests conducted, behavioural changes occurred far less frequently than link changes. Also, within the ASSIST data, only a small number of smoker changes are observed in comparison to friendship changes. Thus, a behavioural change may be considered as more "rare" and therefore potentially more important; hence, a behavioural change is selected to occur prior to a link change. When a searching agent is reselected and allowed to make their predefined stored link change, the preselected change may no longer be the most fruitful to personal PR; this may be considered a penalty of behavioural change.

For the previous analysis of the PR-Max and behavioural search method, the dampening constant d was selected as 0.85 - as this is the value originally selected by Google. In context of the original PR calculations, this gives an 85% weighting to social network connections (M) and a 15% weighting to the fractional matrix Q. Given that the BPRM method considers individual attributes, retaining $d = 0.85$ would mean only a 15% weighting of behaviours in the BPRM calculations.

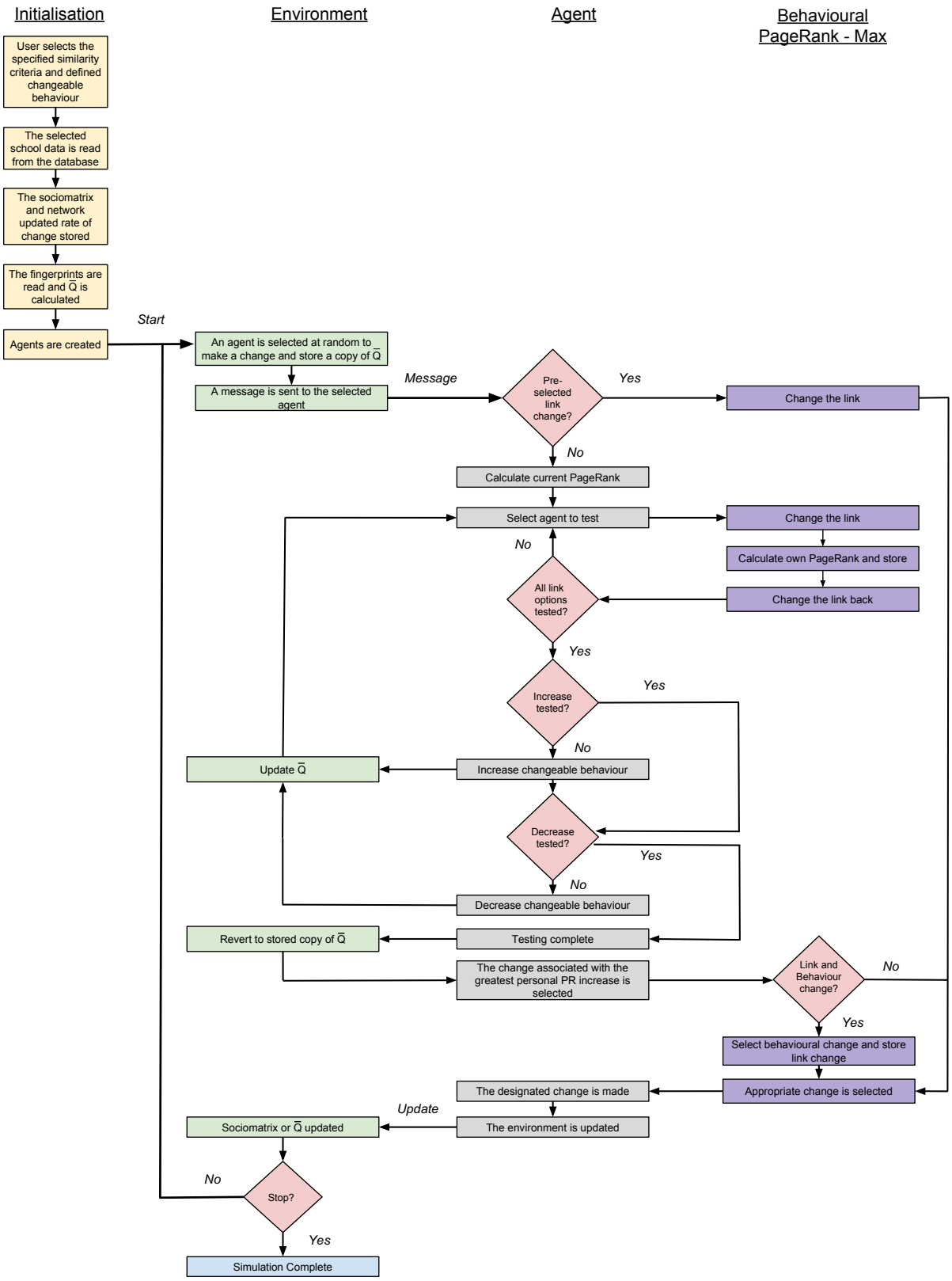


Figure 8.8: Behavioural PageRank-Max simulation logic.

The work of Bressan & Peserico (2010), suggests that the selection of d can impact the PR calculation - potentially affecting the resultant rankings. To investigate this further, d is to be altered in the following BPRM analysis. This allows for alternative weightings of behavioural and social network connections, with an investigation of the outcome of considering varying dampening constants. The values selected for investigation are $d = \{0.15, 0.5, 0.85\}$, testing the effects of the Google selected dampening on both M ($d = 0.85$) and $\bar{Q}$ ($d = 0.15$), but also investigating equal consideration ($d = 0.5$).

Additionally, the BPRM allows agents to make no change to their current social or behavioural situation, if an increase in PR is not experienced - avoiding the link reversion discussed in Section 8.2.3. This only comes into effect in the dynamic smoker consideration of Section 8.3.3. Prior to investigating the effect of altering a changeable behaviour, the BPRM method is run with static smoker behaviour. This is to test the effect of using the $\bar{Q}$ matrix in the PR-Max calculation, with respect to LP precision; a description of the process and the results produced are presented in Section 8.3.2.

8.3.2 Static BPRM Precision

To begin the precision investigation of the BPRM link predictions, the method is assessed **without** consideration to changing smoking behaviour. Essentially, the static BPRM works in the same manner as the PR-Max method, except with $\bar{Q}$ in the calculation of PR instead of Q. $\bar{Q}$ is set upon initialisation of the simulation, and remains unaltered for the remainder of the run. This gives a baseline representation of the BPRM, from which a comparison with the dynamic smoker analysis of Section 8.3.3 can be drawn.

The same test schools selected for the behavioural search of Section 8.2 are used for the static BPRM investigation (schools 12, 33, 71 and 74). For this section of analysis, smoking is the nominated attribute for assessment. Therefore, the similarities between agents in terms of smoking level is found, with the appropriate $\bar{Q}$ being constructed. The precision of the static BPRM method is displayed in Table 8.5.

From Table 8.5, with $d = 0.85$, none of the test schools indicate a significant difference from the PR-Max method in terms of correct predictions at T_2. As with the many of the selected restrictions imposed in the behavioural search of Section 8.2, missed values appear significantly better in schools 12 (-1.79%), 33 (-2.65%) and 71 (-5.76%). This

School	Measure	0.85		0.5		0.15	
		T_2	T_3	T_2	T_3	T_2	T_3
12	Correct	−0.53	−0.72	−0.44	−2.06	−1.16	−4.39
	Missed	−1.79	−3.15	−2.43	−2.95	−1.90	−2.44
33	Correct	−0.59	0.31	−2.28	0.20	−3.90	1.12
	Missed	−2.65	0.34	−1.70	0.57	−1.38	0.75
71	Correct	0.37	−0.45	−2.13	−3.75	−4.41	−9.86
	Missed	−5.76	−1.99	−6.69	−2.27	−6.60	−0.52
74	Correct	−1.21	0.56	−0.89	0.46	−1.48	0.81
	Missed	−0.19	−0.11	−1.13	0.11	−1.23	−0.24

Table 8.5: Static BPRM 'correct' and 'missed' results, expressed as a percentage increase over the PR-Max method. The values are grouped by time period and selected value of d. Green values indicate a significant improvement over PR-Max, while red values indicate a significant deterioration. No colour indicates no significant difference from PR-Max.

may once again be attributed to the conducted alterations, potentially delaying the arrival of an agent's steady "friendship state". Recall that the static BPRM does not yet employ the ability to reject worsening PR changes, friendship reversions (previously discussed in Section 8.2.2) potentially the cause of the observed values.

Schools 33 (0.31%) and 74 (0.56%), once again display a significant increase in the percentage of correct predictions at T_3 (with $d = 0.85$) when considering smoking behaviour in link predictions. These figures indicate consistency with the results of the smoker behavioural search, although, the improvements are small. Overall, the results demonstrate that giving 15% weight to smoking behaviour in the link calculations, produces some significant decreases in the percentage of missed predictions at T_2, with minor significant improvements in correct predictions at T_3 for specific schools.

Giving equal weight to static smoker similarity and network structure ($d = 0.5$), produces variable results. Schools 12 (-2.43%), 33 (-1.70%) and 71 (-6.69%) once again exhibit a significant decrease in missed predictions at T_2, but schools 33 (-2.28%) and 71 (-2.13%) indicate a significant deterioration of correct predictions. School 33 is of particular interest, as previous results have demonstrated an improvement in correct predictions on consideration of smoking behaviours; it would appear at T_2 with a 50% smoking consideration, this is not the case.

Further correct prediction deterioration is experienced at T_3, with schools 12 (-2.06%) and

71 (-3.75%) experiencing significant reductions in correct predictions. School 33 (0.20%) indicates no significant improvement in correct predictions, while school 74 experiences only a marginal significant increase (0.46%). The results would indicate giving a greater weighting to static smoker behaviour, does not offer improvements to the link prediction process.

The poor performance of a 50% weighting upon smoking behaviour, is emulated when an greater consideration is given to the behaviour. Taking $d = 0.15$, and therefore an 85% consideration to smoking similarity, results are once again variable. Particularly large significant decreases in correct predictions at T_3 are experienced for schools 12 (-4.39%) and 71 (-9.86%), demonstrating that this combination of parameters is not appropriate for predicting links in these schools. Although schools 33 (1.12%) and 74 (0.81%) observe their largest precision increase at T_3 with $d = 0.15$, overall, the large decreases in schools 12 and 71 suggest an 85% weighting to static smoking is not a representative account of adolescent linking behaviour.

The static smoking BPRM has provided a number of conclusions regarding the newly developed method:

- The BPRM method is viable as an approach to potentially produce improved link predictions (with appropriate parameters);
- Schools 33 and 74 place particular importance upon smoking similarity at T_3;
- Static smoker similarity is not necessarily the most appropriate method to include smoker dynamics in link prediction;
- The selected value of d can have an impact upon the link predictions made.

With the investigation of static smoker behaviour within the BPRM complete, an analysis of dynamic smoking behaviours in conjunction with attribute similarity is conducted in Section 8.3.3.

8.3.3 Dynamic Smoking BPRM Precision

The dynamic smoking BPRM method investigates potential changes to smoking behaviour, in an attempt to improve link predictions. The method assesses increases and decreases to smoking behaviour, but also allows agents to make no change to their social or behavioural situation, if their eigen-centrality cannot be improved. The effects of dynamic smoking behaviours are first tested in isolation, with consideration being given to gender and ethnicity, form group, peer supporter nominations and overall Levenshtein distance, as the investigation continues.

Dynamic Smoking

To first assess the improvement offered by considering dynamic smoking behaviour, the similarity matrix ($\bar{Q}$) is constructed based solely upon smoking level - being updated throughout the simulation, based upon the subsequent agent decisions. Table 8.6 displays the performance of the dynamic smoking BPRM method, expressed as a percentage increase over PR-Max predictions.

		0.85		0.5		0.15	
School	Measure	T_2	T_3	T_2	T_3	T_2	T_3
12	Correct	0.60	1.28	1.41	0.23	−0.72	−1.14
	Missed	3.89	5.22	0.51	3.58	−1.23	−1.67
33	Correct	0.92	−0.27	0.80	0.92	−1.95	1.66
	Missed	1.19	0.52	−1.44	0.08	−1.69	0.39
71	Correct	0.68	2.01	1.80	0.63	−2.17	−4.37
	Missed	−0.52	1.99	−3.92	4.09	−4.38	−0.89
74	Correct	2.30	−0.25	1.01	0.86	−1.28	1.47
	Missed	−0.02	−0.09	0.49	−0.17	−0.10	−0.44

Table 8.6: Dynamic Smoking BPRM 'correct' and 'missed' results, expressed as a percentage increase over the PR-Max method. The values are grouped by time period and selected value of d. Green values indicate a significant improvement over PR-Max, while red values indicate a significant deterioration. No colour indicates no significant difference from PR-Max.

School 74 with $d = 0.85$ at T_2, indicates a significant increase to correct predictions (2.30%), with school 74 previously experiencing no significant improvement across the behavioural alterations investigated. This would suggest the dynamic smoker evolution

has captured an important aspect of the friendship selection process in School 74.

The reason for the elevated precision experienced at T_2 for school 74 (with $d = 0.85$), may be due to the above average increase of smokers observed at this school between T_1 and T_2 (from Section 5.2.2, average: 9.72%, school 74: 10.22%). Although the smoking elevation is only 0.5% above average, it may indicate an increased importance upon smoking behaviour within the school. This may in-turn affect friendship selection, leading to the significant improvement observed by the dynamic smoker BPRM.

A significant increase in correct predictions is also observed at T_3 (with $d = 0.85$) for school 12 (1.28%), which only experienced a small significant increase during the gender based search of Section 8.2.2 (0.79%). While the increase is small, it indicates the potential of considering a changeable behaviour in a dynamic framework. However, no further significant increases are experienced with $d = 0.85$ at T_2 or T_3.

Of particular interest is the increase in correct predictions observed for schools 33 and 74 at T_3, as d decreases. For $d = 0.85$ the school 33 and 74 T_3 correct predictions are -0.27% and -0.25% (respectively), increasing to 0.92% and 0.86% when $d = 0.5$, and 1.66% and 1.47% when $d = 0.15$. Both of these schools perform well when using a static smoker consideration. This would suggest that dynamic smoking behaviour does not capture the school's friendship selection procedures as well as static smoking (for $d = 0.85$), however, the method performing better when more emphasis is given to smoking similarity.

A further notable feature of the dynamic smoking BPRM, is the absence of significantly decreased missed predictions; a previously prominent result of the attribute based methods considered. As described in Section 8.3.1, the inter-event time governing the selection of agents to make a BPRM change, considers both link and behavioural events. Evidently, the agent decides which behavioural or link change to make. This may be causing fewer link changes to be made than expected, with more behavioural changes being exacted. As a result, a number of schools indicate a significant increase in missed predictions - for example, school 12 at T_2 with $d = 0.85$ (3.89%). Discussions regarding the particular behavioural changes made in the simulation, are presented in Chapter 9.

Overall, it would appear that considering dynamic smoking in isolation has not contributed greatly to improving link predictions. However, the method has highlighted the importance of considering changing smoking behaviours in schools 74 (at T_2) and 12 (at T_3). Of

course, it is not solely smoking that may be pertinent in friendship selection, similarity in terms of gender and ethnicity may also a driving factor of link formation. To investigate this further, gender and ethnicity are included in the similarity matrix $\bar{Q}$.

Gender and Ethnicity

To investigate the effects of considering additional personal attributes, along with dynamic smoker behaviour, gender and ethnicity are included in the dynamic BPRM. To exact this, in the initialisation process of $\bar{Q}$, the number of similarities between agents in terms of smoking, gender and ethnicity are calculated. Over the course of the simulation, smoking behaviours may change based on agent decisions; however, the gender and ethnicity remain constant. This gives an underlying static component to $\bar{Q}$, in addition to the dynamic smoker behavioural link changes. Table 8.7 displays the SNS results of including gender and ethnicity in the $\bar{Q}$ matrix.

		0.85		0.5		0.15	
School	Measure	T_2	T_3	T_2	T_3	T_2	T_3
12	Correct	−0.41	0.52	1.00	−0.28	0.82	0.67
	Missed	−1.49	−2.96	−2.31	−2.87	−2.52	−3.38
33	Correct	0.74	0.49	1.30	0.60	1.85	0.92
	Missed	−3.08	0.13	−3.34	0.34	−3.49	0.51
71	Correct	0.46	0.39	0.93	0.98	1.28	0.43
	Missed	−4.97	−1.71	−5.17	−1.73	−5.28	−1.88
74	Correct	0.22	0.74	0.04	0.82	1.46	1.48
	Missed	−0.73	−0.39	−0.89	0.15	−2.10	0.07

Table 8.7: Dynamic smoking and static gender and ethnicity BPRM precision, expressed as a percentage increase over the PR-Max method. The values are grouped by time period and selected value of d. Green values indicate a significant improvement over PR-Max, while red values indicate a significant deterioration. No colour indicates no significant difference from PR-Max.

The results indicate fewer significantly poorer predictions than considering smoking in isolation; LP improvements experienced amongst both correct and missed predictions, across all values of d. However, the values do not appear greatly enhanced upon those of the behavioural search (Section 8.2), with significantly improved correct predictions indicated only in schools 33 and 74 - both schools generally performing well on consideration of smoker behaviour.

The results highlight the importance of static personal attributes, such as gender and ethnicity, in friendship selections - presenting how they may be used in conjunction with dynamic smoker behaviours. While the inclusion of gender and ethnicity has not elevated predictions substantially, the results demonstrate that significantly improved precision may be obtained with the BPRM method. It may be the case that precision may be further increased with alternative attribute consideration, with the next attribute to be considered being a representation of proximity.

Form Group

The ASSIST school friendship networks already demonstrate a great deal of consideration to proximity in friendship selection, as participants evidently encounter their selected connections regularly in the school environment. To investigate the importance of proximity further, and the effect of dynamic smoking behaviours, form group is included as a static element of the $\bar{Q}$ matrix. Therefore, agents are said to have a similarity if they are in the same form group or share the same smoking behaviour.

		0.85		0.5		0.15	
School	Measure	T_2	T_3	T_2	T_3	T_2	T_3
12	Correct	1.50	0.05	2.11	−0.38	0.89	−1.27
	Missed	−2.28	−1.24	−3.49	−2.74	−2.52	−2.23
33	Correct	0.27	0.29	1.02	0.91	−1.60	0.42
	Missed	−2.76	−0.08	−2.79	−0.23	−1.85	0.41
71	Correct	0.47	0.36	1.42	0.40	0.43	−0.88
	Missed	−4.51	−1.23	−5.49	−1.54	−5.33	−2.53
74	Correct	0.80	0.38	0.24	0.81	−1.81	0.41
	Missed	−0.89	0.00	−0.83	−0.22	−0.17	−0.20

Table 8.8: Dynamic smoking and static form group BPRM precision, expressed as a percentage increase over the PR-Max method. The values are grouped by time period and selected value of d. Green values indicate a significant improvement over PR-Max, while red values indicate a significant deterioration. No colour indicates no significant difference from PR-Max.

Table 8.8 presents the effect of including form groups in $\bar{Q}$. Form group is found to be particularly important in school 12 friendship formation at T_2, when $d = 0.85$ (1.50%) and $d = 0.5$ (2.11%). This indicates that form group may be a driving force in friendship selection at T_2 in school 12, with proximity being less important as the adolescents mature.

Form group does not appear to be an important aspect of linking in school 33, the school experiencing a significant decrease in correct predictions at T_2 with $d = 0.15$; with correct predictions at T_3 also being reduced in comparison to other attribute based methods. From this, an appreciation of the variability in elements pertinent to adolescent friendship formation may be gained - certain attributes being successful in specific schools.

The results demonstrate how the BPRM method may be used for investigative purposes, focusing upon key elements of attribute based linking to assess the outcomes to predictions made. Furthermore, the results serve to demonstrate the overall success of the PR-Max method, indicating the importance of status in adolescent friendship selection, irrespective of individual attributes. To investigate this further, the ASSIST interpretation of status (peer supporter nominations) is assessed in conjunction with link predictions.

Peer Nominations

As previously discussed (Section 5.1.2), each ASSIST individual holds a nominations number - a value indicating how many votes a study participant received to become a peer supporter. A nomination is interpreted as a vote of status in the network, with the selected peer supporters comprised of trusted individuals and 'trend setters'. Therefore, the smoking opinions of the highly nominated individuals may carry more weight in driving smoking uptake, subsequently having an impact upon link formation.

Nominations have been included in $\bar{Q}$ by an alteration to its calculation. Evidently, a nominations number is not a variable that can be matched in terms of similarity between agents. Rather, the nominations number ($\tilde{n}$) is used to weight the observed similarities, such that:

$$q_{ij} = \begin{cases} 1 + k\tilde{n}, & \text{if } i \neq j \\ 1, & \text{otherwise} \end{cases} \tag{8.4}$$

Therefore, having a similarity with an agent who possesses a high nomination value is beneficial to personal PR, with agents attempting to match their smoking behaviour with highly nominated individuals. Table 8.9 displays the results of the nomination weighted BPRM method.

School 71 with $d = 0.85$, significantly improves correct predictions at both T_2 (2.48%) and

School	Measure	0.85		0.5		0.15	
		T_2	T_3	T_2	T_3	T_2	T_3
12	Correct	1.56	0.80	0.71	−1.36	0.18	−1.13
	Missed	3.89	4.31	1.79	5.26	−1.77	1.48
33	Correct	1.54	−1.49	0.65	−0.62	−1.66	1.54
	Missed	1.63	0.42	0.07	0.18	−1.60	−0.02
71	Correct	2.48	2.91	2.31	0.99	−0.72	−2.88
	Missed	−1.45	2.51	−1.63	2.97	−4.29	−1.71
74	Correct	0.59	−0.73	0.97	0.74	−0.63	0.86
	Missed	5.81	0.00	4.10	0.11	−0.75	−0.28

Table 8.9: Dynamic smoking weighted by peer nomination BPRM precision, expressed as a percentage increase over the PR-Max method. The values are grouped by time period and selected value of d. Green values indicate a significant improvement over PR-Max, while red values indicate a significant deterioration. No colour indicates no significant difference from PR-Max.

T_3 (2.91%). School 33 also observes a significant improvement to precision at T_3, with correct predictions significantly increasing by 1.54% when $d = 0.15$. This would suggest nominated individuals may contribute somewhat to friendship selection and smoking behaviours in these schools, with the specified value of d demonstrating the appropriate level of consideration.

Both school 33 and 71 are control schools, therefore, highly nominated individuals are not selected as peer supporters - their status never being publicised. The reason for the lack of significant improvement in intervention schools, may be due to the nominated individuals losing status when being highlighted as prominent network individuals (previously discussed in Section 5.2.3); thus, the nomination figures may no longer be representative of true nominated status.

Although intervention school 74 does demonstrates a small significant increase in correct predictions at T_3 with $d = 0.5$ (0.74%) and $d = 0.15$ (0.86%), contradicting the arguments presented above, this may be due to the strong smoker similarity present in the network. Previous attribute BPRM investigations, demonstrate the positive effect of considering dynamic smoking behaviours in School 74.

While a number of small significant improvements to precision are observed, overall it would appear that weighting the smoking decisions by nominations, has not provided a

substantial increase across all schools. This is not unexpected, as the nomination values may potentially become obsolete as the students mature. However, the results serve to demonstrate the effect of weighting the BPRM with $\tilde{n}$ instead of n - providing depth to the BPRM analysis. The final element of the dynamic smoking BPRM analysis, investigates the effect of considering all available variables in the similarity matrix.

Levenshtein Distance Matrix

To consider all individual attribute data, $\bar{Q}$ is altered to consider the Levenshtein distance. Recall from Section 8.1.3, the Levenshtein distance ($\tilde{l}$) is the number of single character edits necessary to to make two agent fingerprints equivalent. Therefore, as the Levenshtein distance increases, the agents are more dissimilar. This is opposing to the previous calculations of $\bar{Q}$, which require larger values to indicate a more likely agent association. As such, q_{ij} is calculated as follows:

$$q_{ij} = \begin{cases} \frac{1}{(1+\tilde{l})}, & \text{if } i \neq j \\ 1, & \text{otherwise} \end{cases} \tag{8.5}$$

resulting in an alteration to $\bar{Q}$. The dynamic change of smoking behaviour is still considered, now having an effect of increasing or decreasing $\tilde{l}$. As such, smoking only contributes a very small part of the simulated friendship selection process, with questionnaire responses as a whole being the primary focus.

Table 8.10 displays the results of the dynamic smoker Levenshtein-based BPRM. A number of significant precision improvements are observed across all values of d, at each predicted timestep. School 33 and 74 are once again the most successful in terms of correct predictions; however, schools 12 and 71 also demonstrate selected increases, although these are not significant at the 95% level.

It would appear that all schools indicate some form of improvement, whether it be in terms of missed predictions or correct predictions. This suggests that considering Levenshtein distance has not significantly negatively impacted any of the predictions made, with potential to increase results significantly. While larger overall precision increases have been observed in the other attribute based tests, no other method has produced **only** significant

School	Measure	0.85		0.5		0.15	
		T_2	T_3	T_2	T_3	T_2	T_3
12	Correct	1.18	−0.58	0.02	−0.09	0.41	0.40
	Missed	−2.52	−2.40	−1.77	−2.75	−2.02	−3.34
33	Correct	1.57	0.48	1.53	0.36	1.16	0.39
	Missed	−3.04	−0.15	−3.43	0.21	−3.11	0.39
71	Correct	0.98	−0.82	1.56	0.12	0.74	0.04
	Missed	−4.88	−0.58	−5.44	−1.82	−4.69	−1.90
74	Correct	1.36	0.87	0.17	0.95	−0.21	1.34
	Missed	−1.27	−0.04	−0.66	−0.31	−0.96	−0.31

Table 8.10: Dynamic smoking and Levenshtein BPRM precision, expressed as a percentage increase over the PR-Max method. The values are grouped by time period and selected value of d. Green values indicate a significant improvement over PR-Max, while red values indicate a significant deterioration. No colour indicates no significant difference from PR-Max.

improvements across all values of d.

The results would suggest that employing a dynamic smoking Levenshtein based BPRM method, does not require a trade off in terms of school precision. Therefore, the method provides a manner to include agent specific attributes, along with dynamic behaviours, to inform link predictions. It may be the case that with further investigation, the method may be used to gain further improvements in precision - resulting in a greater understanding of adolescent social structures and diffused behaviour. The findings of the BPRM based methods, along with conclusions drawn across the breadth of the investigation, are summarised in the following section 8.3.4.

8.3.4 BPRM Conclusions

The BPRM investigation has reinforced a number of key findings highlighted across this thesis, presenting an alternative PageRank based method to include attribute data in the link prediction process. The conclusions of this investigation are as follows:

- Dynamic Smoking - The BPRM presents a framework to consider dynamic behavioural change, demonstrating an improvement over considering static smoker behaviour. While the alterations to the PR calculation do not provide the large increases observed in Section 8.2 with the behavioural search, the method produces

less volatility in terms of link predictions made. Although this chapter has focuses specifically upon the link precision outcomes of the BPRM, the method also produces predictions for the smoking uptake of agents in the network; the smoking predictions are examined further in Chapter 9.

- Attributes - The investigation has reinforced the findings of the behavioural search, demonstrating how specific individual attributes may inform link predictions. In particular, proximity, gender and ethnicity were highlighted as potentially important static attributes to friendship selection, with a consideration to overall questionnaire responses (through the Levenshtein distance) producing generally improved predictions. Schools 33 and 74 particularly benefited (in terms of precision) from the inclusion of personal attributes, once again suggesting the importance of giving consideration to factors other than network structure in adolescent friendship selection.

- Dampening Constant - The selection of d has been shown to affect precision, with school 74 generating increased correct predictions at T_3 with $d = 0.15$, and school 12 at T_2 generally performing better with $d = 0.85$ (across all dynamic BPRM methods). Due to this, and the small margin of increase observed with the BPRM method, it is unclear which selection is preferable. To be definitive about the selection of d, more investigation is required; however, this investigation has begun to demonstrate the effect of the d parameter with respect to link predictions. Further discussions regarding the selection of the dampening constant are presented in Chapter 9.

- PR-Max - While individual attributes have managed to capture aspects of the friendship selection process (demonstrated by small improvements to precision), the key driving force behind the link predictions is the PR-Max method. This once again highlights the strength of the process of optimising eigen-centrality, with respect to adolescent network evolution.

The BPRM investigation has discussed only one selected changeable behaviour, upon four ASSIST schools. Given this is the case, the investigation may be developed further in a number of manners:

- Alternative Fingerprints - The fingerprints used, take into consideration all the non-text based variables from the ASSIST questionnaires. However, a number of fields relate to similar outcomes - such as numerous questions relating to smoking habits.

Missing data is also an issue; if data relating to the smoking variable is missing, then the simulation selects a smoking value based upon the distribution of the existing data. However, for all other variables the data remains coded as missing - meaning agents have a similarity if both contain missing data for a specific variable field. Furthermore, parts of the questionnaire are only answered if a participant has responded in a particular manner. As such, the study related elements of the data may be introducing 'noise' into the constructed fingerprints, and affecting the resultant Levenshtein based similarities. With stronger, more defined agent fingerprints, further precision increases may potentially be obtained.

- Alternative Changeable Behaviours - The selected changeable behaviour for this chapter is smoking, however, alternative behaviours may also be selected. If data were available upon alcohol or drug use, an investigation into these behaviours could also be sought. Furthermore, the framework presented may be used for other health based investigations, such as the spread of loneliness and depression in a social network (as discussed in Fowler & Christakis (2008)).

- Alternative Data - This investigation has specifically focused upon four adolescent real social networks, gaining an understanding of the aspects important in the school friendship selection process. Evidently, considering alternative ASSIST schools may have produced differing insights. However, additional conclusions may also be drawn from different data. Examples could include a comparison with online adolescent social network data (derived from Facebook or Twitter), or focusing upon an alternative demographic altogether. It would be of interest to investigate the accuracy of the BPRM and PR-Max algorithms, in the framework of alternative social network data; this giving an understanding of the importance of status in the context of differing types of social connection.

Overall the BPRM method has investigated the contribution of behavioural attributes (both static and dynamic), and presented a framework that may be built upon to achieve superior link predictions. Chapter 7 demonstrated the ability of the PR-Max method to produce more precise link predictions than the existing methods tested, with the BPRM method having the potential to improve on this further. While the BPRM precision improvements may be relatively small, ultimately, the included attribute data only presents small aspects of an ASSIST participant's life - the friendship selection process potentially drawing on a

range of unquantifiable factors or factors that are simply not recorded in this dataset.

The focus of this chapter has been to investigate the precision of alternative, attribute based methods of link predictions, but further outcomes are also produced. During the dynamic smoking BPRM based network simulations, the process is making predictions regarding the smoking behaviour of individuals in the network. The SNS is effectively attempting to model the behaviours of the ASSIST individuals, in an effort to inform link predictions. An investigation of this process as a method to model behaviour, along with alternative behavioural models, is explored in Chapter 9.

8.4 Chapter Summary

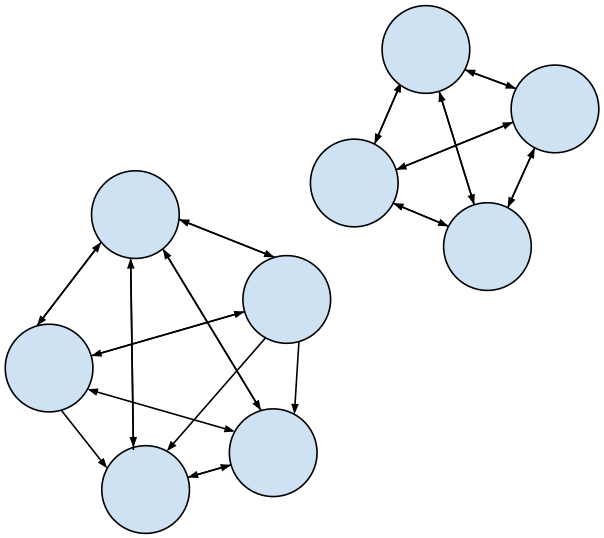
This chapter has attempted to improve the PR-Max method through the inclusion of individual attributes and behaviours. Section 8.1 outlined the investigation, detailing particular elements integral to the analysis. The schools selected for further exploration (12, 33, 71 and 74), were chosen due to their association in the original ASSIST study, previous PR-Max performance and representative ASSIST network characteristics. A number of specific elements were highlighted to direct the investigation process (gender, smoking, nominations and proximity), with consideration also given to all questionnaire data through the inclusion of the Levenshtein distance.

Section 8.2 restricted the search space of the PR-Max algorithm by specific agent attributes, while allowing agents with similar backgrounds and opinions (through short Levenshtein distance) to also be included. The behavioural search demonstrated the ability of specific attributes to improve precision, with outcomes dependent upon the schools tested and the attributes selected. Schools 33 and 74 demonstrated particular improvement at T_3 with the smoker Levenshtein search, both schools previously demonstrating low precision with the PR-Max method. However, the restrictions imposed also had the ability to significantly reduce precision, suggesting that restricting the search space may not always be appropriate. The behavioural search also highlighted the potential issue of including dynamic behaviours (such as smoking) in a static framework, inspiring the development of an evolving behaviour based link prediction process.

Section 8.3 presented the Behavioural PageRank-Max method, which altered the Q matrix in the calculation of PageRank to include personal attributes and behaviour. Six alter-

native formations of Q were investigated (referred to as $\bar{Q}$) : static smoker behaviour, dynamic smoker behaviour, gender and ethnicity, form group, nominations number and Levenshtein distance. The BPRM precision analysis demonstrated the ability to obtain precision improvements with specific consideration to static characteristics, in conjunction with dynamic smoking behaviour. The Levenshtein based $\bar{Q}$ matrix appeared to offer small significant improvements across schools, without being detrimental to overall LP precision - providing the ability to improve predictions in schools where a sole consideration of network structure may be inappropriate.

This chapter has demonstrated the ability to improve the link predictions made by the PageRank-Max algorithm, through the inclusion of attributes extraneous to social networks structure; this new algorithm termed Behavioural PageRank-Max (BPRM). The constructed BPRM framework offers the ability to make both social network *and* behavioural predictions. While the focus of Chapters 6, 7 and 8 was to gain an understanding of social network evolution, it is also of interest to investigate the interplay of social network structure and behavioural diffusion. Moving forward, the BPRM framwork shall be used to investigate smoking uptake in ASSIST schools, this being the focus of Chapter 9.



-"Disconnected Cliques"

9

Social Smoking

The focus of this chapter is to gain a greater understanding of the relationship between social network structure and behavioural influence. Chapter 5 highlighted key aspects of network structure that may contribute to the diffusion of smoking messages within the context of ASSIST, with Chapter 6 presenting PageRank-Max as a new method to predict social network evolution. These chapters centred predominantly upon social network structure, with the results of Chapter 7 highlighting the importance of personal eigen-centrality in adolescent friendship selection.

Chapter 8 demonstrated that, while network structure plays an important role in the evolution of a social network, the personal and behavioural aspects of an individual may also have an impact upon their friendship selections. This suggests smoking uptake is not simply a direct product of social networks, but rather that the process of friendship selection and smoking uptake is interconnected. Therefore, factors extraneous to immediate social networks are also important in an adolescent's decision to smoke.

This chapter uses the refined algorithm of Chapter 8, Behavioural PageRank-Max (BPRM), as a method to investigate the interplay between social network structure and smoking uptake. Using the ASSIST data, Section 9.1 introduces a new model to investigate the dynamic of adolescent smoking behaviour - assessing the role of eigen-centrality and attribute similarity in the co-evolution of friendship structure and smoking decisions.

Additionally, this chapter also presents alternative models to investigate the social factors involved in smoking uptake. An Evolutionary Game Theory (EGT) model is established in Section 9.2, and a basic compartmental model is introduced in Section 9.3. These additional models are presented as an elementary outline of their inherent methodologies, their inclusion not intended to portray a fully representative conceptualisation of adolescent smoking; rather, their incorporation into this thesis serves to demonstrate further alternative avenues of research, with the models having potential for future expansion. Section 9.4 draws together the conclusions of all three models.

9.1 BPRM Based Smoking Predictions

From Chapter 7, PageRank (PR) (as a proxy for status) is identified as a potential factor in the evolution of adolescent social networks. The Behavioural PageRank-Max (BPRM) of Chapter 8, attempted to improve link predictions by altering smoking behaviours, with an individual's behaviours and attributes being considered through the inclusion of a similarity matrix ($\bar{Q}$) - defined in Equation 8.3. As a result of altering the PR calculations in this manner, carefully selected links may increase an agent's PR, with behavioural similarities between agents also having an impact. Therefore, agents seek to be eigen-central, but can also improve their PR by increasing their similarity with other agents.

Similarity has been discussed as an important aspect in friendship selection (Section 3.3.2), with status being highlighted as a representative method to evolve social networks (Section 7.4). Therefore, it is of interest to investigate *status* (or PageRank) and *similarity*, in evolving adolescent smoking behaviours. The investigation is structured as follows: a brief discussion of BPRM smoker predictions is presented in Section 9.1.1; the results of modelling the spread of smoking with PageRank are discussed in Section 9.1.2; and Section 9.1.3 discusses the conclusions drawn from the BPRM smoker predictions.

9.1.1 BPRM Smoker Predictions

In the investigation of Chapter 8, four schools were selected for use with the BPRM method - schools 12, 33, 71 and 74. The precision of the link predictions made, using the BPRM method, was discussed in Section 8.3.3. During the BPRM investigations, smoking behaviours were also predicted as an underlying method to improve link predictions. This section discusses these smoker predictions, in an attempt to gain further insight into the workings of the BPRM method.

Figures 9.1 and 9.2 display the predicted proportion of smokers at T_2 and T_3 (respectively), for each of the test schools, with varying values of d. Recall, d is the weight given to social network structure (versus behaviour) in the PR calculations (Equation 8.1). The results are obtained following ten replications of the SNS using the BPRM method, with $\bar{Q}$ only considering smoker similarities (as in Section 8.3.3). The true proportion of smokers in each school, taken directly from the data, is also displayed in the graphs (indicated by "Data").

From Figure 9.1, the predicted proportion of smokers decreases as d decreases. Meaning that the BPRM method is predicting that agents are more likely to be non-smokers (at T_2), when more weight is given to smoker similarity. A similar trend is also observed in Figure 9.2 at T_3, however, the results of school 71 indicate an *increase* in the proportion of smokers when $d = 0.5$. It is of particular interest to understand the cause of the varying observations of school 71, and the overall trend of smoker proportions decreasing as d decreases, as the results may be key to improving link predictions. Furthermore, an understanding of the sensitivity of smoker predictions in relation to d, offers an alternative method to model behaviours.

The reason for the decreased smoker proportion as d decreases, may be due to the BPRM method giving more emphasis to matching agents in terms of similarity. To illustrate this, consider the heat map of smoker similarity in school 71 at T_1 - Figure 9.3. The SNS uses this initial smoker similarity matrix ($\bar{Q}$) and changes the smoking behaviour of agents over the course of the simulation. When the run is complete, a "predicted" similarity matrix at T_2 is created.

When $d = 0.85$, a heat map of the predicted $\bar{Q}$ at T_2 can be observed in Figure 9.4; the

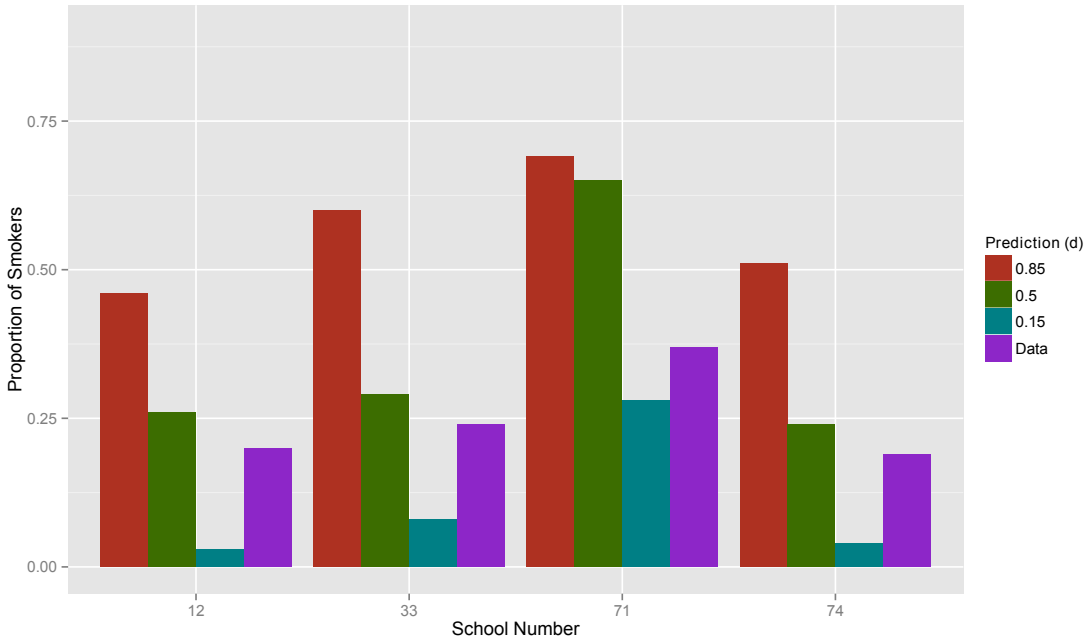


Figure 9.1: Predicted proportion of school smokers from the BPRM method at T_2, with varying values of d. The true smoking proportion is also displayed.

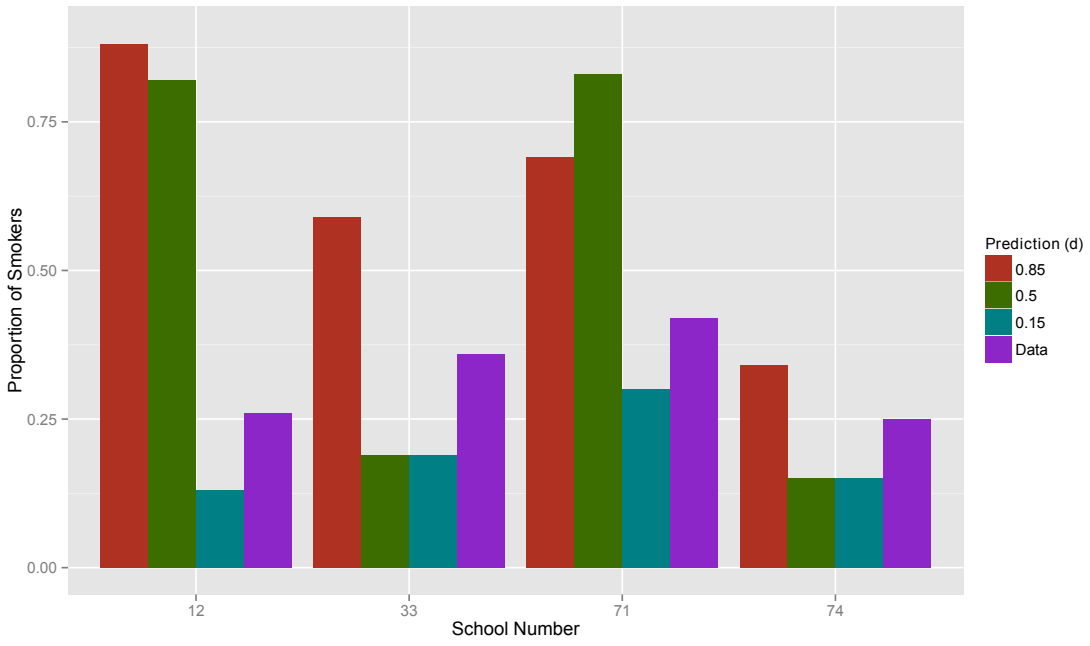


Figure 9.2: Predicted proportion of school smokers from the BPRM method at T_3, with varying values of d. The true smoking proportion is also displayed.

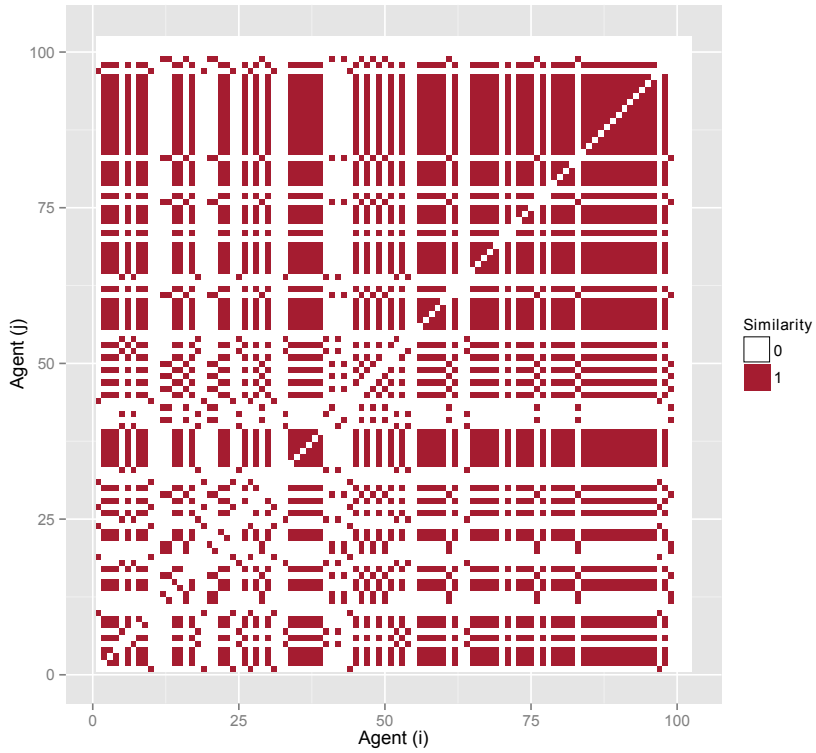


Figure 9.3: Heat map of smoking similarity from ASSIST school 71 at T_1 between agents; 0 indicates no similarity, while 1 indicates the same smoking level.

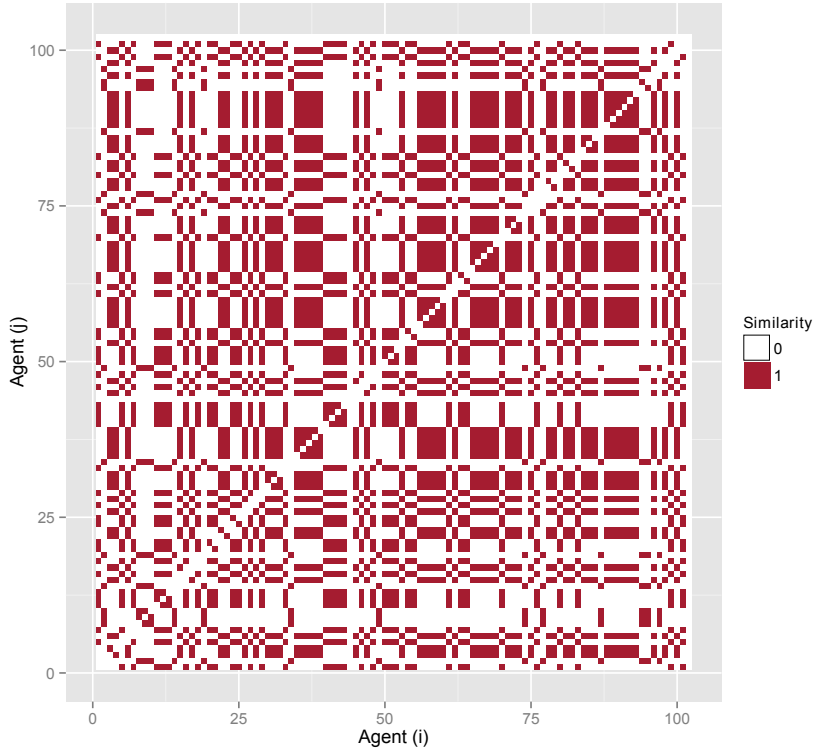


Figure 9.4: Heat map of predicted smoking similarity of school 71 at T_2 with $d = 0.85$; 0 indicates no similarity, while 1 indicates the same smoking level.

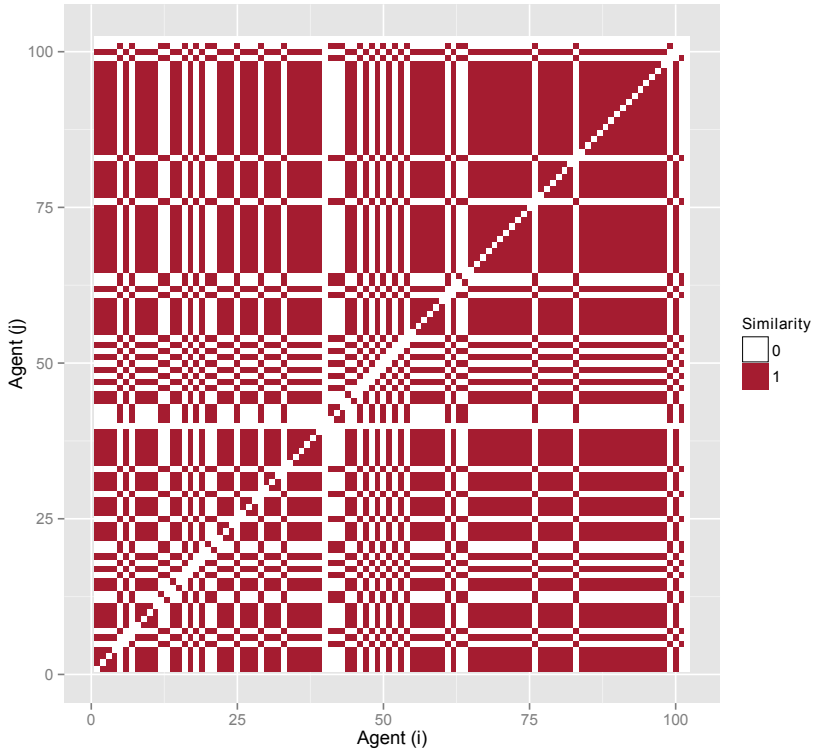


Figure 9.5: Heat map of predicted smoking similarity of school 71 at T_2 with $d = 0.15$; 0 indicates no similarity, while 1 indicates the same smoking level.

heatmap is generated from one run of the simulation, for illustrative purposes. The effect of varying d is demonstrated by Figure 9.5, which represents the predicted T_2 heatmap with $d = 0.15$. The observed increase in the number of similarities with $d = 0.15$ (Figure 9.5), compared with $d = 0.85$ (Figure 9.4), demonstrates that the BPRM method is focusing upon changing behaviours, to increase the number of similarities between agents.

The smoking proportions at T_1 for each of the test schools (found in Tables 5.2 and 5.3), are lower than the proportion of non-smokers; this means that not smoking is the majority (or dominant) behaviour in these schools. Therefore, as d decreases and similarity becomes more important, to increase an agent's PR, they must become more similar to other agents in the network. This is achieved by smokers (the minority) becoming non-smokers, resulting in the decreased proportion of predicted smokers observed at T_2 when $d = 0.5$, compared with the proportion of smokers at T_2 when $d = 0.85$.

School 71 has a large proportion of smokers at T_2 (37.00%, from Table 5.2), which may

be the reason that the T_3 smoker predictions (Figure 9.2) do not follow the same pattern displayed in other test schools. When $d = 0.5$, the proportion of smokers is higher than when $d = 0.85$, with the $d = 0.15$ predicted smoker proportion being higher than all other test schools. Due to the greater proportion of smokers in the network, the non-smokers do not hold as much of a majority in terms of smoker behaviours - potentially allowing smoking to become dominant when $d = 0.5$.

The results presented introduce the concept of a majority behaviour in the framework of the BPRM. Agents within the BPRM method are basing their similarity decision on the overall population behaviour (as d decreases), therefore, it is of interest to investigate the threshold at which a behaviour becomes dominant in the BPRM method, and illuminate potential causes of said dominance. It is also of interest to investigate if specific central individuals may affect the majority behaviour. Further exploration of these issues is provided in Section 9.1.2.

Overall, this discussion has highlighted the effect of d upon behavioural predictions, and introduced the concept of a majority behaviour in the BPRM method. The results presented specifically relate to the use of a smoker similarity matrix, however, a similar trend can be viewed across all similarity matrices investigated in Chapter 8 (gender and ethnicity, form group, nominations, and Levenshtein). The complete table of predicted smoker proportions, along with the accuracy of the predictions made, can be found in Appendix D.3. The following section (9.1.2) investigates the effect of majority behaviour, dominant individuals and the overall impact to smoking adoption.

9.1.2 Dominant Smoking Behaviours

To investigate the threshold at which smoking becomes a majority behaviour, using the BPRM method, the network of school 40 at T_1 is used. School 40 has been selected for this analysis, due to the following reasons:

- School 40 initially had only two smokers at T_1, the number increasing greatly by T_3. It of interest to explore the initial number of smokers required for a substantial increase to be observed with the BPRM method;

- The school is initially predicted well with the PR-Max method (Section 7.1). As network structure also affects the calculation of an agent's PR within the BPRM

method, it is important that network predictions are as accurate as possible;

- The school 40 social network is identified as having a cliqued structure, the analysis of Section 5.2.3 suggesting a great deal of smoker message diffusion between T_1 and T_2 - resulting in a large smoker increase. It is of interest to explore the effect of this cliqued structure upon the BPRM method;

- The network is the smallest of the ASSIST schools, meaning it will take the shortest length of time to produce multiple simulation runs.

To ascertain the behavioural impact on agents, as a result of the size of the smoker population, the BPRM method is employed to make smoking behaviour predictions. The SNS is given the network of School 40 at T_1, with T_2 smoking predictions being generated. Only smoker similarities are considered in $\bar{Q}$, with agents having the ability to dynamically update their smoking behaviours over the course of the run (following the framework discussed in Section 8.3.3). Ten replications of the simulation are used, in keeping with the required number of runs established from previous analyses within this thesis.

First, the behaviour of the agents in school 40 remains unaltered, meaning that there are two smokers in the school. On completion of the required replications, with $d = \{0.85, 0.5, 0.15\}$, the average smoker proportions are recorded. Then, the number of smokers in school 40 is increased by one, to assess the impact of a greater smoker community. The initial two smokers remain fixed, with the newly introduced additional smoker being selected uniformly at random from the remaining non-smoker population. Following the completion of the simulation with the increased smoker population, a further agent is selected to become a smoker; the process repeating until all agents in school 40 are smokers.

The results of altering the initial number of smokers in School 40 can be observed in Figure 9.6. With $d = 0.15$, not smoking remains the dominant behaviour until around half of the population smoke; then, the proportion of smokers sharply increases until smoking becomes the majority behaviour. This is logical, as $d = 0.15$ gives an 85% consideration to smoking similarity, meaning that agents focus upon increasing similarities to improve their PR - network eigen-centrality being less important.

When greater consideration is given to network structure, $d = 0.85$, the proportion of smokers gradually increases when there are between 20 and 35 additional smokers. The

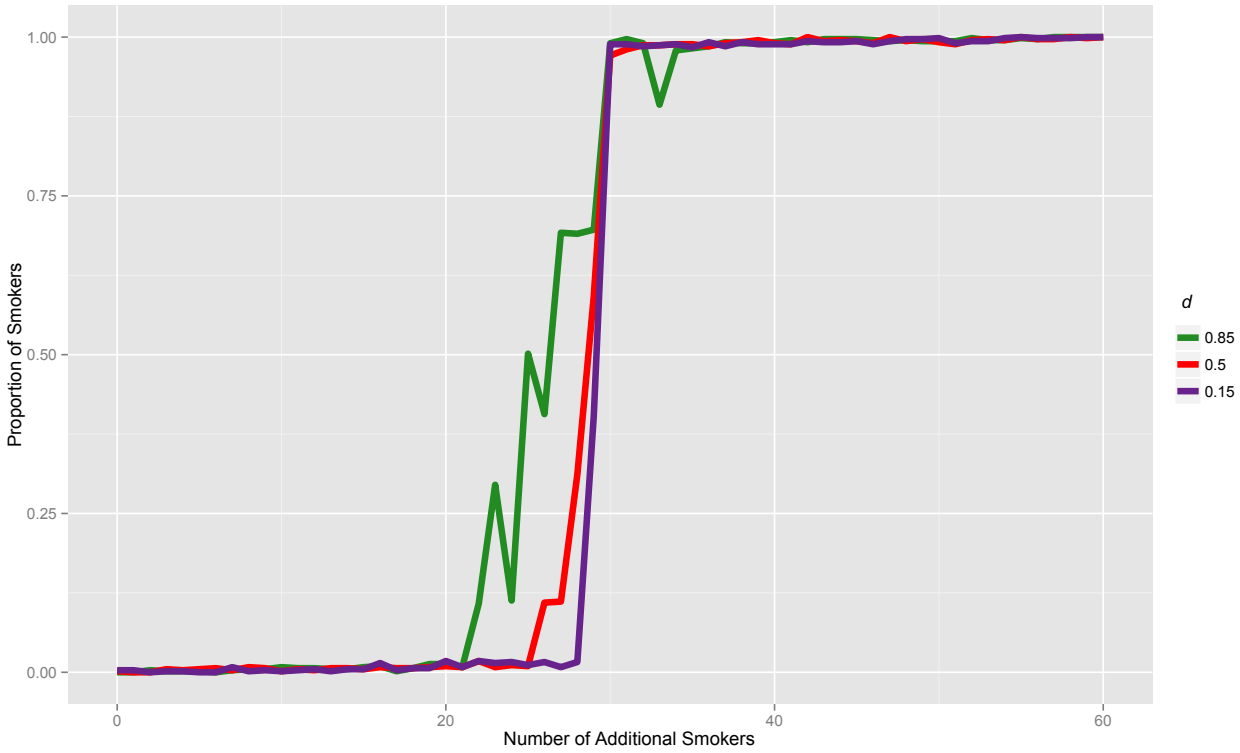


Figure 9.6: The BPRM method smoker proportion predictions with varying values of d, and increasing numbers of additional smokers in School 40.

increase is not as sharp as that exhibited when $d = 0.15$, as an agent's primary focus is to achieve an improved PR through carefully selected network connections. Of particular interest is the cause of the variability experienced within the region of gradual increase, the proportion of smokers with $d = 0.85$ not increasing smoothly as those of $d = 0.15$.

To investigate the cause of the variability further, additional experiments were conducted with $d = 0.999$ and $d = 0.95$ - assessing the effect of further reduced similarity consideration. The results of the additional parameter investigations are displayed in Figure 9.7, with the previous parameter experiments overlaid. The results demonstrate a dampening of the effect of a majority behaviour, the increase in the proportion of smokers becoming more gradual as d increases.

The results of the additional parameter tests also indicate an increase in the variability of results (as d increases). While there is a general trend of an increased smoker population as the number of additional smokers increases, the large peaks exhibited when $d = 0.999$

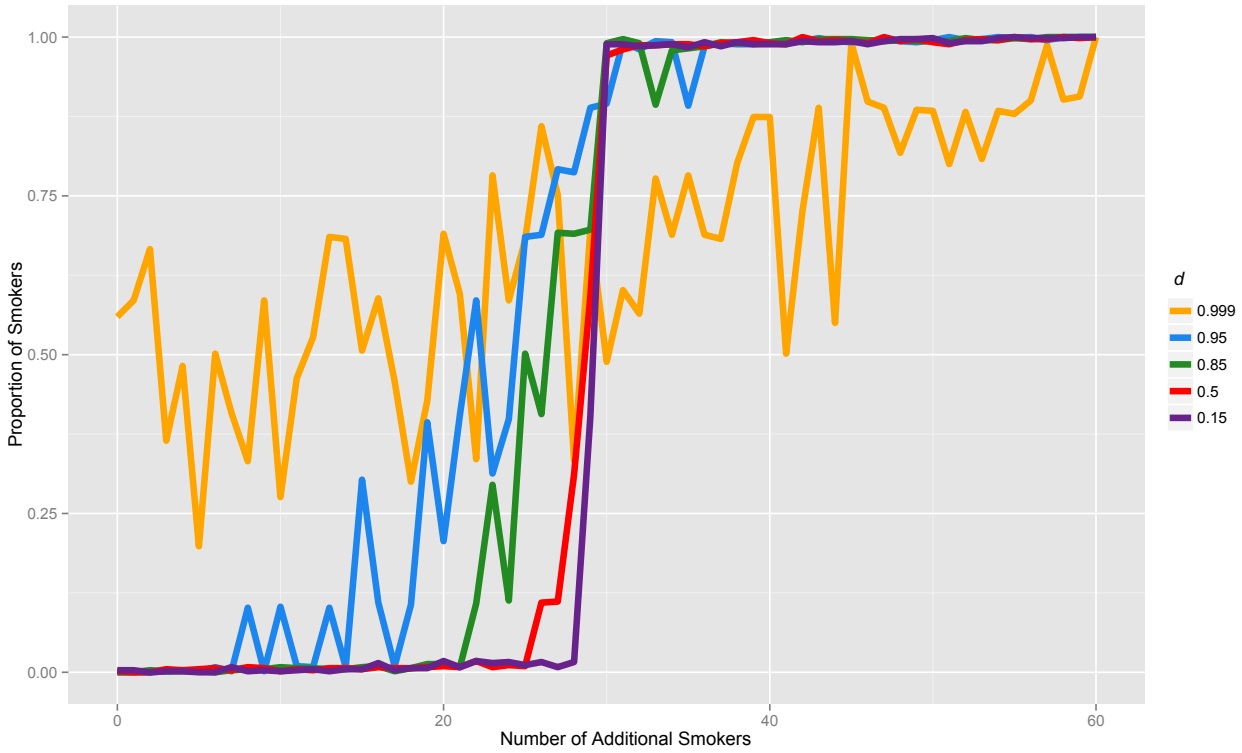


Figure 9.7: The BPRM method smoker proportion predictions with alternative values of d, and increasing numbers of additional smokers in School 40.

would indicate some inconsistency within the results. This inconsistency appears to be introduced when network centrality becomes the primary optimisation goal of agents, suggesting the structure of the network is impacting the smoking decisions of the agents.

It is hypothesised that when d is small, it is beneficial to be similar to as many agents as possible in the BPRM method; however, as d increases, it is more lucrative to be similar to agents who have the highest network eigen-centrality (or PR). To investigate this further, the selection process of the additional school 40 smokers is altered.

The network structure of school 40 is retained, however, the population is altered such that there are no smokers. This is to provide a benchmark with which to compare with as the number of smokers is increased. Following ten replications of the simulation with a non-smoker population, the average smoker proportions (for each of the previously discussed values of d) are recorded. Then, the agent with the highest PR at T_1 is selected to become a smoker - the simulations being repeated to assess the impact of a highly PageRanked

smoker. The number of smokers is increased, with agents possessing the highest PR values being selected each time, until all agents in the school are smokers.

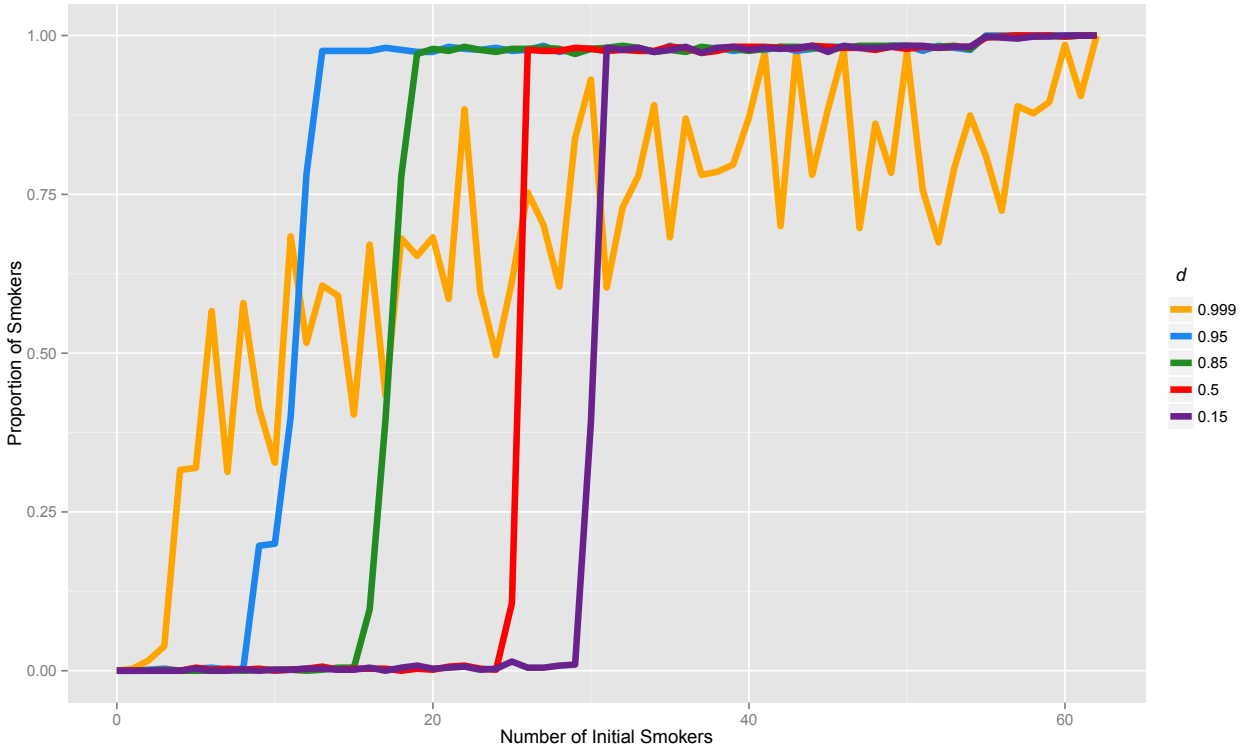


Figure 9.8: The BPRM method smoker proportion predictions with alternative values of d, initial smoker selection being based on largest PR values

The results of the BPRM method with hierarchical smoker selection are displayed in Figure 9.8. While predictions with $d = 0.15$ retain similar behaviour to that observed in Figure 9.7, the variability in results from $d = 0.95$, $d = 0.85$ and $d = 0.5$ has substantially decreased. It is also noted that the point at which smoking becomes a majority behaviour, occurs sooner as d increases. The smaller number of initial smokers required for smoking to become the dominant behaviour, would suggest that agents with a high PR have the ability to spread the smoking message more rapidly. This is because agents are seeking to become similar to the highly PageRanked agents, which in the current scenario are the smokers.

The results of $d = 0.999$ still display a great deal of variability, albeit this decreasing from the results generated with random smoker selection. The variability may be due to the primary focus of the agent PR optimisation process being the network structure (previously

discussed in Section 6.2.2), causing the highest PageRanked agents to continually change. As the highest ranked agents are not remaining consistent, this may result in the previously selected highly ranked smokers no longer possessing high PR values. Thus, the dominant behaviour in the network changes - causing the variability observed.

To provide further evidence of the effect of highly ranked agents upon behavioural dominance in the BPRM method, the process of selecting smoking agents is again altered. A new initial smoker selection process is adopted, with agents possessing the *lowest* PR now being selected to become smokers. The process of introducing an increasing number of initial smokers is repeated until all agents in the network smoke. The results of the lowly ranked smoking agent selection are displayed in Figure 9.9.

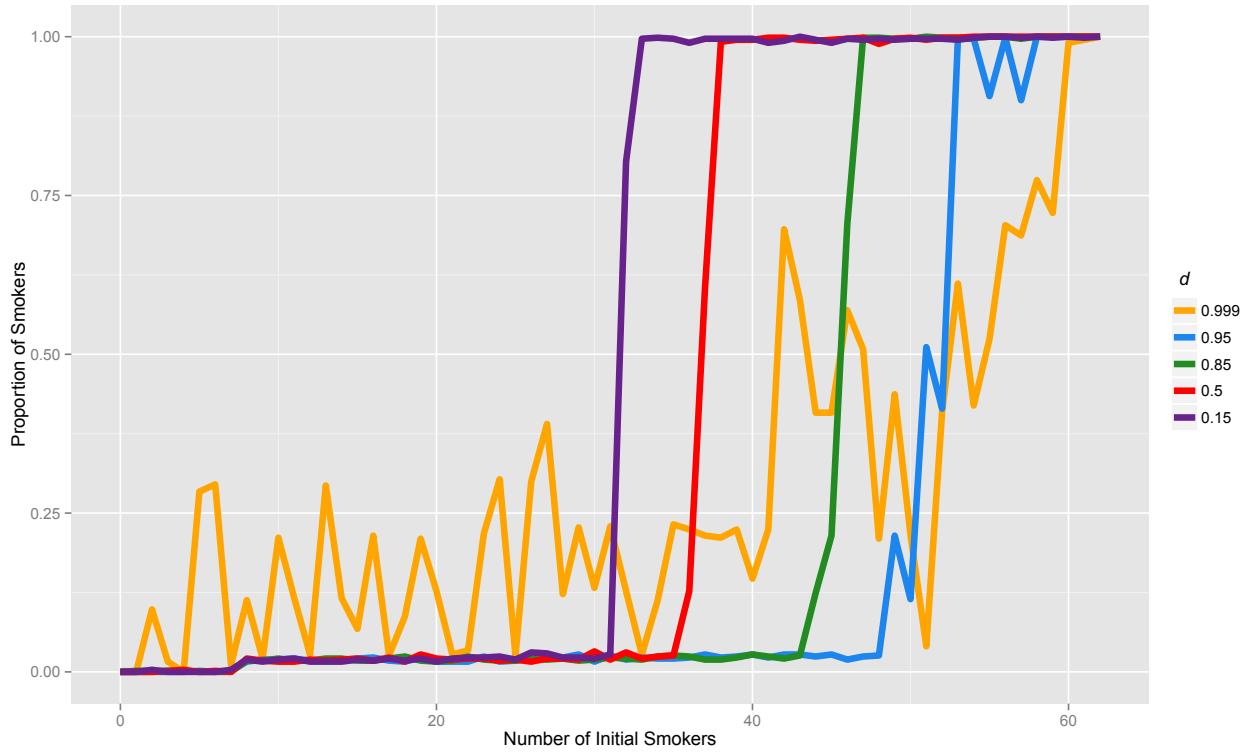


Figure 9.9: The BPRM method smoker proportion predictions with alternative values of d, initial smoker selection being based upon the lowest PR values.

The results demonstrate that the number of agents required for smoking to become a majority behaviour has increased substantially. With $d = 0.95$, $d = 0.85$ and $d = 0.5$, more than 50% of the agents are required to smoke for smoking to become the majority behaviour. This is greater than high-PR smoker selection process (Figure 9.8), which required less

than 50% of agents to smoke for smoking to become a majority behaviour. The results of $d = 0.15$ exhibit the same trend as previous analyses, with the PR of agents being less important - the algorithm focusing upon increasing an agent's similarity with all other agents.

An additional observation on comparison of the results of Figure 9.8 and Figure 9.9, with those of Figure 9.7, is the reduction of variability in the final proportion of smokers as the number of initial smokers increases (for all values of d) - the lines becoming smoother. The previous variability observed (in Figure 9.7), may be due to the random selection of additional smokers. If the randomly selected additional smokers had a high PR, then the resultant proportion of smokers increased; whereas if the additional smokers had low PR values, then the smoker proportion was reduced.

The results have clearly demonstrated the effect of highly PageRanked individuals in the BPRM method, and the resultant effect of their behaviour upon other individuals in the network. It would appear that when a great emphasis is given to the matching of behavioural similarity, agents decide to smoke or not to smoke based upon the majority behaviour. However, when d is increased, the agent's focus becomes their network centrality and emulating the behaviours of highly PageRanked individuals. This emulation having the ability to increase the proportion of smokers more rapidly when highly ranked agents smoke, but also reduce the dominance of smoking when lowly PageRanked individuals smoke. The conclusions of this investigation, and a discussion of the BPRM method as a technique to model behaviour, are discussed in the following section (9.1.3).

9.1.3 BPRM Smoker Model Conclusions

The investigation of the BPRM method smoking predictions, has drawn the following conclusions:

- The BPRM smoker predictions are affected by the dampening constant d;
- The proportion of smokers predicted in the four ASSIST test schools generally decreases as d decreases (from Section 9.1.1);
- The cause of the observed dynamic is a result of specific behaviours being dominant in the system;

- When d is small, the dominant behaviour is simply the behaviour possessed by the majority of agents - the threshold required for a majority being roughly half of the population;

- As d increases, agents seek to match their behaviours to highly ranked individuals. This alters the threshold required for behaviours to become dominant;

- When d is high, variability is introduced as the agents seek to improve their PR more aggressively. As a result, the highly ranked individuals change, causing the dominant behaviour to fluctuate.

In addition to the conclusions drawn above, the BPRM method presents great opportunity for future research. An understanding of the sensitivity of smoker predictions in relation to d has been established, with the interaction between highly ranked agents and behavioural majority clearly defined. Therefore, to make representative smoker predictions in conjunction with social network structures, an appropriate value of d is required.

The selected value of d must encompass the drive to become eigen-central in a social network, but also the need for similarity with other agents. Multiple simulations with real data would be required to find appropriate values of d, potentially offering the ability to find a general value of d that may be used in the explanation of smoking uptake in adolescents. This may provide insight for adolescent smoking interventions and allow the BPRM method to be recognised as a modelling technique for a multitude of behavioural characteristics.

This BPRM analysis has provided a foundation from which future research can be conducted, this new model offering insights regarding the interplay between smoking uptake and social network structure amongst adolescents. However, the research presented offers just one perspective of modelling social influence amongst a population. The following sections (9.2 and 9.3) outline the foundations of alternative approaches to examining the problem at hand. These alternative models offer differing perspectives of model formulation, the conclusions of which will be discussed in Section 9.4.

9.2 Game Theoretical Model

This section introduces an Evolutionary Game Theory (EGT) model of adolescent smoking uptake. The model examines an adolescent's decision to smoke, based upon the smoking decisions of the population as a whole - echoing the concept of a majority behaviour introduced in Secton 9.1.1. This section serves to introduce how an investigation of social influence might be structured in the framework of EGT, presenting directions for future investigation.

Myerson (1991) describes game theory as "the study of mathematical models of conflict and cooperation between intelligent rational decision-makers". Modern game theory as a mathematical disciple is said to have begun with the work of von Neumann & Morgenstern (1944), and has been applied within the context of econometrics (Agarwal & Zeephongsekul, 2013; Shapiro, 1989; Tesfatsion, 2006), political science (Morrow, 1994; Ostrom, 1998; Wood, 2011) and biology (Hammerstein & Selten, 1994; Maynard Smith & Price, 1973).

This section of work describes the basic concepts of game theory (Section 9.2.1), introduces evolutionary stable strategies (Section 9.2.2), outlines the adolescent smoking model (Section 9.2.3), describes the process of finding an Evolutionary Stable Strategy (ESS) in the developed model (Section 9.2.4) and discusses the subsequent conclusions drawn (Section 9.2.5).

9.2.1 Game Theory Introduction

A *normal form game* is an interactive decision problem (the "game") involving a number of individuals (the "players"), which may be represented in tabular form. A *static game* is one in which each player may make only one decision, without prior knowledge of the decisions made by other players - effectively the decisions are simultaneous. Each player is assumed to be *rational*, being aware of their own potential options or *strategies* for playing the game, and the strategies available to other players (Webb, 2007).

A static normal form game consists of:

- A set of N players, indexed by $i \in \{1, 2, ..., N\}$;

- A set of possible strategies for each player, S_i;
- A utility function for each player $u_i : S_1 \times S_2 ... \times S_N \rightarrow \mathbb{R}$

Classic examples of static normal form games include the "battle of the sexes" and "the prisoner's dilemma" (Luce & Raiffa, 1957). For illustrative purposes, an example of the prisoner's dilemma is given below.

Two prisoners are being held in conjunction with a serious crime. They are incarcerated in separate cells and cannot communicate with one another. The police only have enough evidence to charge the prisoners with minor offences. The police concoct the following plan to obtain a confession for the serious offence:

- If one prisoner confesses (known as *defecting*) that both prisoners were perpetrators of the serious crime, the confessor will be set free and the other prisoner will spend 10 years in jail.
- If both prisoners confess to the serious crime (both defect), they each receive 5 years of jail time;
- If neither prisoner confess to the crime (both cooperate), the prisoners can only be charged for the minor offences - receiving only 2 years in jail each.

	Prisoner B: Confess	Prisoner B: Remain Silent
Prisoner A: Confess	(-5,-5)	(0,-10)
Prisoner A: Remain Silent	(-10,0)	(-2,-2)

Table 9.1: Normal form of the prisoners dilemma.

Table 9.1 displays the normal form of the prisoners dilemma. There are two strategies that each prisoner can play: confess or remain silent, with the utilities expressed in number of years "lost". The utilities of prisoner A are given in the first entry of each pair in the table, while those of prisoner B are given in the second. Although it is in the *common* best interest of both players to remain silent (and receive 2 years each), because the prisoners are unable to confer with one another before being interviewed by police - they might act in their *personal* best interests.

Irrespective of the strategy played by the other prisoner, it is always in the prisoner's per-

sonal best interest (or **best response**) to confess (defect). This is a Nash Equilibrium, defined formally as:

Definition 9.2.1. *A Nash Equilibrium in a two player game is a pair of strategies* (σ_1^*, σ_2^*) *such that:*

$$u_1(\sigma_1^*, \sigma_2^*) \geq u_1(\sigma_1, \sigma_2^*) \; \forall \sigma_1 \in S_1 \tag{9.1}$$

and

$$u_2(\sigma_1^*, \sigma_2^*) \geq u_2(\sigma_1^*, \sigma_2) \; \forall \sigma_2 \in S_2 \tag{9.2}$$

where u_i *is the utility for player i,* σ_i *is the strategy for player i and* S_i *is the entire set of strategies for player i (Nash, 1950).*

Thus, given the strategy adopted by the other player, neither player can increase their utility by selecting an alternative strategy (Webb, 2007).

This section has provided an introduction into the ideas of game theory, outlining the basic notation, terminology and concepts that will be discussed in the adolescent game theory model. Further information regarding the history, development and mathematical formulation of game theory may be found in the texts of Myerson (1991) and Webb (2007). This section has defined the concept of a strategy in game theory, with the following section (9.2.2) describing an evolutionary stable strategy in Evolutionary Game Theory (EGT).

9.2.2 Evolutionary Game Theory

This section focuses upon an area of game theory that examines the evolution of strategic behaviour in a population. In the framework of classic game theory, the outcome is dependent upon the rational choices of the players, whereas in Evolutionary Game Theory (EGT), it is the strategies that are of interest. Consider a population of players where each individual is playing the best response the population's strategy (σ^*), then no individual can improve their utility given the current strategies being played - the population is said to be in *equilibrium*. Of interest in EGT is the stability of the equilibrium point. Consider some mutation occurring, causing part of the population to begin playing a different strategy. If the population returns to the equilibrium point, for all such mutations (if they are small enough), then σ^* is said to be an Evolutionary Stable Strategy (ESS).

An evolutionary game can be described in the context of biological reproduction. To il-

lustrate the evolution process, consider a "biological" game with two strategies s_1 and s_2. There are N individuals in the population, with 50% playing strategy s_1 and 50% playing s_2. The population profile (χ) is a vector that gives the probability with which each strategy is being played, therefore, $\chi = (0.5, 0.5)$. The utilities for playing a particular strategy with the current population profile, $u(s, \chi)$, may be interpreted as the number of offspring generated by a player. In the current example, these are set to: $u(s_1, \chi) = 2$ and $u(s_2, \chi) = 8$. When a player generates an offspring, the offspring inherit the strategy played by their parent. Therefore, in the next generation, there will be $\frac{2N}{2}$ individuals playing s_1, while $\frac{8N}{2}$ individuals play s_2; the new probability profile will be $\chi = (0.2, 0.8)$.

The biological game described above, demonstrates how the population evolves based on the utilities (or number of offspring) associated with a given strategy. In the next generation of the game, the utilities for each strategy may change based on the number of individuals playing that strategy. This is known as a *game against the field*, whereby there is no specific opponent for each individual - the utilities being based upon the behaviour of others in the population.

To describe an ESS, consider a population with two strategies $S = \{s_1, s_2\}$, where all individuals adopt a best response σ^* to $\chi = (0.5, 0.5)$. Suppose some genetic mutation occurs and a small proportion of the population (ϵ) decide to use a different strategy σ. The new population profile (including the newly developed mutant population) χ_ϵ is:

$$\chi_\epsilon = (1 - \epsilon)\sigma^* + \epsilon\sigma \tag{9.3}$$

Definition 9.2.2. *The strategy σ^* is an ESS if there exists an $0 < \bar{\epsilon} < 1$ such that for every $0 < \epsilon < \bar{\epsilon}$ and $\sigma \neq \sigma^*$:*

$$u(\sigma^*, \chi_\epsilon) > u(\sigma, \chi_\epsilon) \tag{9.4}$$

This means that no strategy adopted by the new mutant population can produce more offspring than σ^*, therefore, the mutant strategy does not displace the equilibrium of the current population - the mutant strategy becoming extinct. An example to illustrate the process of finding an ESS in a game against the field, may be found in Appendix C.1.

More information about EGT may be found in Weibull (1997) and Webb (2007). Origi-

nally, EGT was proposed by Maynard Smith & Price (1973) in the context of biology, but is now applied to a multitude of problems (Bauch & Bhattacharyya, 2012; Cui et al., 2014; Jalali Naini et al., 2011; Wen et al., 2013). This work aims to apply EGT with respect to smoking in schools, the specific strategies played by the population being either to smoke, or not to smoke. This is discussed further, with the model being outlined, in section 9.2.3.

9.2.3 Adolescent Smoker Model

To develop a theoretical model of adolescent smoking in schools, an evolutionary game is proposed. Smoking has previously been studied in the context of game theory, with Shiell & Chapman (2000) attempting to reduce passive smoking in restaurants. Nyborg & Rege (2003) also developed an EGT model to examine considerate smoking behaviour in public places. However, to the best of this author's knowledge, an EGT model of adolescent smoking in schools has not been addressed in the literature.

In the proposed adolescent smoker EGT model, there are two strategies each student can play: non smoker (ns) or smoker (s), with $S = \{\text{ns}, \text{s}\}$. The proportion of non smokers in the population is α, and the proportion of smokers $1 - \alpha$, meaning that the population profile is $\chi = (\alpha, 1 - \alpha)$. The general strategy $\sigma = (\omega, 1 - \omega)$ induces a population with a proportion of ω non-smokers and $(1 - \omega)$ smokers. Selection of appropriate utilities for non-smokers, $u(\text{ns}, \chi)$, and smokers, $u(\text{s}, \chi)$, is key - as this will ultimately decide the evolution of the population.

To select appropriate utility functions for the proposed model, literature related to adolescent smoker uptake is investigated. The social aspect of adolescent smoking is well documented (Alexander et al., 2001; Biglan et al., 1984; Ennett & Bauman, 1994; Kobus, 2003; Lakon & Valente, 2012; Levinson et al., 2007) but it is of interest to understand the specific gratification received from being a smoker (or non-smoker) relative to a particular population profile. Consideration must also be given to factors that may deter adolescents from smoking. As such, three components are identified as important in the literature: "coolness", "personal cost" and "conformity".

Coolness

Ioannou (2003) identifies coolness and self perception as being key to an adolescent's de-

cision to smoke, with the need to look "cool" amongst peers being of great influence (Norman & Tedeschi, 1989). In a review of the stages of progression in adolescent smoking behaviour, Mayhew et al. (2000) also found an individual's perceptions amongst others (their "coolness") to be important in smoking uptake. Aloise-Young et al. (1996) showed that individuals who were similar to the perceived adolescent smoker stereotype, the stereotype being a cool and sociable individual, were twice as likely to show smoking onset.

To select an appropriate function to represent coolness in a population of smokers and non-smokers, literature relating to definitions of coolness is explored. Cool individuals are said to be those who are risk takers (Bird & Tapp, 2008), with coolness having to encompass originality, attractiveness and subcultural appeal (Sundar et al., 2014). Coolness is said to be ever changing (Nancarrow & Nancarrow, 2007), being a marker for status in a population (Belk et al., 2010), with cool individuals known as "trend setters" (Nancarrow et al., 2002).

The research into the perception of cool, would suggest those individuals who adopt a "risky" behaviour first are cool, with cool individuals wanting to adopt a product or behaviour if it is itself cool (Bird & Tapp, 2008). While adolescents want to be cool, they want to belong to a community (Osterman, 2000). Therefore, there is a need to balance the pursuit of individuality (and coolness) with the need to belong (Hornsey & Jetten, 2004).

The following function is selected to attempt a representation of coolness and belonging in terms of a smoker's utility:

$$\exp(-\alpha)^{\frac{1}{c}} \tag{9.5}$$

where α is the previously defined proportion of non smokers in the population, and *uncoolness*, $0 < c < \infty$, is a parameter used to adjust the perceived coolness of smoking.

The selected coolness function assumes that a smoker will receive the most gratification (or largest utility) if everybody smokes, with the utility gained from smoking increasing as the proportion of smokers increases - representing the need to belong. The c parameter is used to adjust the impact to a smokers utility if the proportion of smokers decreases. If the uncoolness of smoking is low, $c = 0.1$, then smoking is deemed very cool and the gratification from smoking is high even if others are not smoking (Figure 9.10). If the uncoolness of smoking begins to increase, $c = 1$,then smoking becomes less worthwhile if others are not partaking - it being most beneficial to smoke when others are smoking

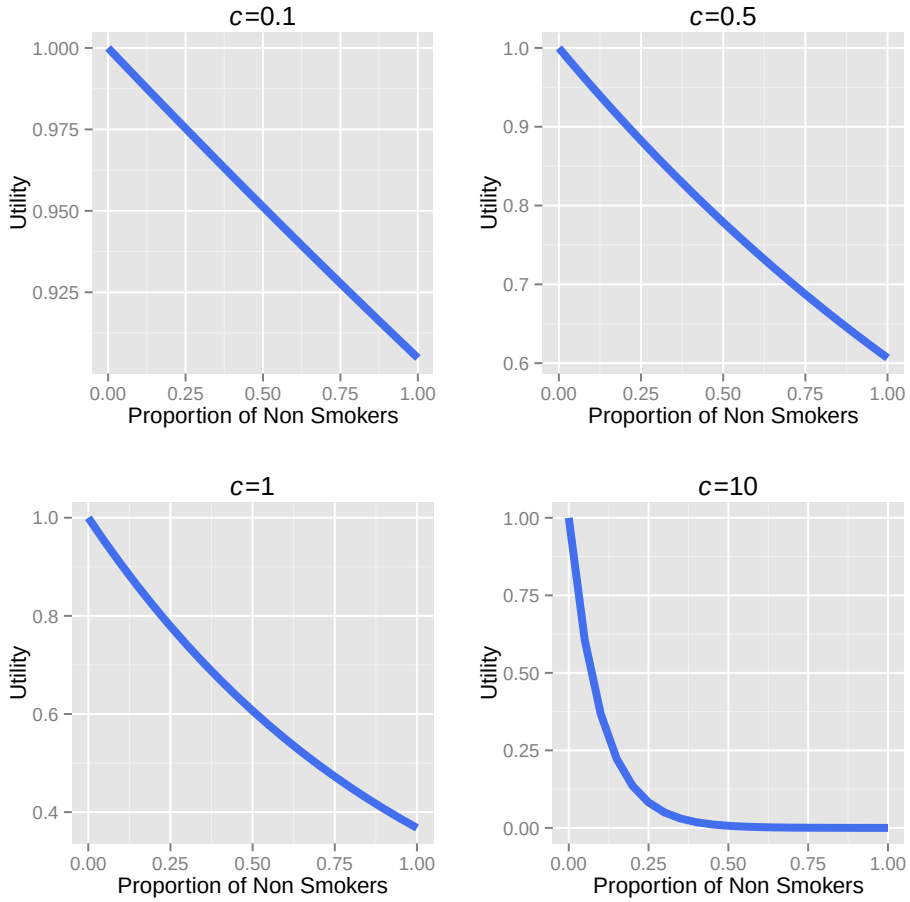


Figure 9.10: The coolness utility function from Equation 9.5 with increasing values of c (uncoolness).

(Figure 9.10). If smoking is particularly uncool, $c = 10$, then it is only worth being a smoker if a majority of students smoke (Figure 9.10). Therefore, as uncoolness increases, the appeal of being a minority smoker reduces - meaning more smokers are required for smoking to appear cool.

The coolness function expresses the positive aspects of smoking with respect to the general population. However, there is a negative impact to smoking which must also be taken into consideration in the smoker utility. This is incorporated into the adolescent smoker model through “personal cost”.

Personal Cost

To represent the negative aspects of smoking, and the subsequent negative impact to a smokers utility, a personal cost function is developed. From an individual perspective, smoking has a number of health risks (Bartecchi et al., 1994; Doll, 2000; US Department of Health and Human Services, 2004), and may be an expensive habit for an adolescent (Montes & Villalbí, 2001; Townsend, 1996); thus, the personal cost of smoking is commonly accepted. However, it is of interest to examine the change in the personal cost of smoking, as the smoker population varies. Three elements of personal cost, from an adolescent population perspective, have been identified: second hand smoking, availability of cigarettes and school intervention.

Second hand smoking, or passive smoking, is the exposure of cigarette smoke to individuals other than the active smoker. The negative cardiovascular effect of second hand smoking is said to be almost as large as smoking (Barnoya & Glantz, 2005), and linked with many other health risks (Correa et al., 1983; Glantz & Parmley, 1991; Trichopoulos et al., 1981), being particularly harmful to adolescents and young children (Asomaning et al., 2008; Carlsen & Carlsen, 2008; Tager, 2008). Evidently, as a smoking population increases, exposure to second hand smoke also increases; therefore, personal health costs increase as the smoker population increases.

The legal age of cigarette purchase in the UK is 18, meaning that the majority of secondary school students cannot legally purchase cigarettes from a licensed tobacconist. Access to cigarettes is said to be from family members and social markets within schools (Emery et al., 1999; Friend et al., 2001; Katzman et al., 2007). Students often acquire cigarettes from intermediaries ("dealers") within the school, who have some cigarette supply system (parents, siblings, friends etc.) (Croghan et al., 2003). Therefore, as demand increases, and a larger proportion of the school population smoke, more pressure may put upon the school dealer to supply cigarettes - potentially reducing their availability and increasing costs.

The work of Fergusson et al. (1995) and Prokhorov et al. (1996) suggests that adolescents who smoke, have trouble quitting in later life. It is of paramount importance to reduce the smoker population in schools, if a reduction to the overall smoker population is sought - as made evident by the breadth of adolescent smoking intervention programmes (Bruvold,

1993; Campbell et al., 2008; Reid et al., 1995; Richardson et al., 2009; Thomas & Perera, 2006). It is assumed that as the smoker population increases, the more difficult it becomes to remain a smoker due to the intervention measures employed by the school; thus, due to the increased effort required to smoke (and avoid the school intervention procedures), personal cost increases.

personal cost increases in resistance to the intervention measures imposed.

From the three population-based perspectives of personal cost discussed, it can be deduced that as the smoker population increases, a smoker's personal cost increases. The following personal cost function is proposed to encompass the negative implications of smoking to a smoker's utility:

$$(1-\alpha)^p \tag{9.6}$$

where $0 \leq p < \infty$ represents the *tolerance* to personal cost.

When tolerance is low, $p = 0$, then the personal cost of smoking is high - irrespective of the proportion of smokers in the population. As tolerance begins to increase, $p = 0.5$, the cost of being a smoker reduces if there is only a small smoking minority in the school. When the tolerance to smoking cost is high, $p = 10$, then the personal costs to a smoker remain low until there is large population of smokers. Figure 9.11 displays the variation in personal cost dependent upon the proportion of smokers in the population, as p increases.

The **personal cost** and **coolness** functions are used to represent the utility received by a smoker, given the strategy profile of the school. Taking into consideration the literature presented, and the coolness and personal cost functions described, the smoker utility is given by:

$$u(\mathrm{s},\chi) = \exp(-\alpha)^{\frac{1}{c}} - (1-\alpha)^p \tag{9.7}$$

Figure 9.12 displays the smoker utility for varying values of c and p. When the tolerance to personal cost is low ($p = 0$), a smoker's utility is negative unless the whole population smokes - indicating that the majority of students are required to smoke for it to be worthwhile enduring the costs. If tolerance to personal cost is high ($p = 10$) and uncoolness is low (0.1), then as the proportion of smokers increases, the utility increases. This continues until the number of smokers in the population becomes large, causing personal costs to increase beyond a sustainable level. However, when tolerance to personal cost is high

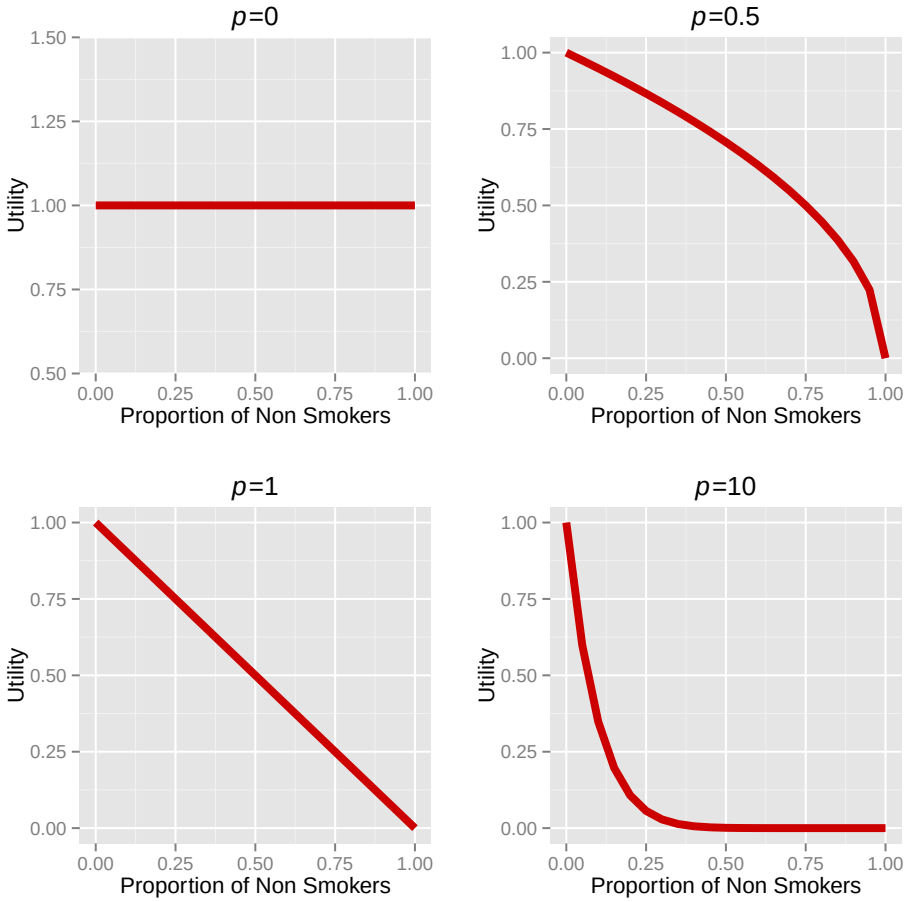


Figure 9.11: The personal cost function from Equation 9.6 with increasing values of p (tolerance to personal cost).

($p = 10$), but smoking is very uncool ($c = 10$), then only a small smoker utility is achieved regardless of the number of smokers.

While the suggested smoker utility function attempts to encapsulate the positive and negative aspects of being a smoker, with respect to the population as a whole, the utility of a non-smoker must also be considered. This is represented through the need for conformity.

Conformity

Aside from the personal health benefits of being a non-smoker, an important aspect in the decision to remain a non-smoker is conformity. Conformity is said to be "the act of changing one's behaviour to match the responses of others" (Cialdini & Goldstein,

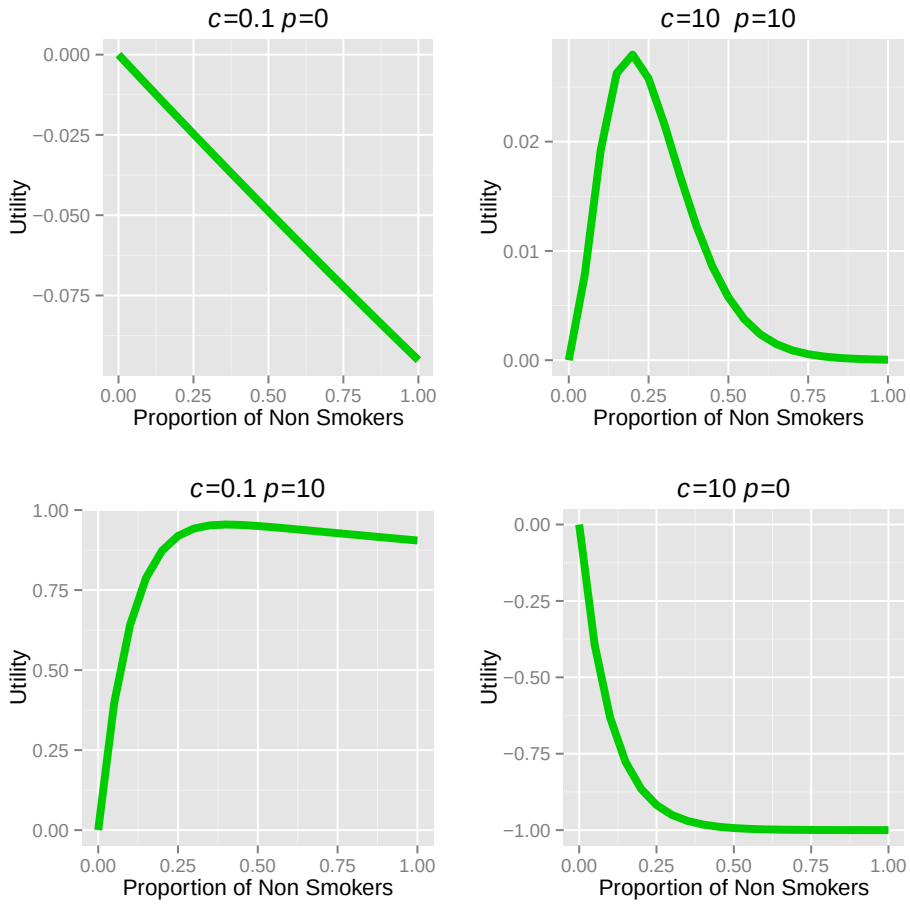


Figure 9.12: The smoker utility function from Equation 9.7 with differing values of c (uncoolness) and p (tolerance to personal cost).

2004), it is the quest to "behave correctly" in a situation and obtain "social approval" from others (Deutsch & Gerard, 1955). Research has demonstrated the power of conformity in encouraging individuals to adopt behaviours they would otherwise not necessarily consider (Deutsch & Gerard, 1955; Milgram, 1963; Sherif, 1937), with Berndt (1979) finding that levels of conformity are particularly high amongst secondary school adolescents.

Smoking behaviours are said to be, in part, a result of adolescents' desire to conform. Peer conformity is identified as an important aspect in the onset and prevention of adolescent smoking (Mcalister et al., 1979), is a risk factor in adolescent smoking uptake (Santor et al., 2000), and has the potential to vary the opinions of a peer group substantially (Hill, 1971). Thus, the utility of being a non-smoker may change based on the levels of conformity present in the population.

To include conformity in the adolescent smoker model, the following function is adopted as the non-smoker utility:

$$u(\text{ns}, \chi) = ((1 - q)(1 - \alpha))^{1-\alpha} \tag{9.8}$$

where $0 \leq q \leq 1$ represents the *need for conformity* with smokers. The non-smoker utility takes into consideration the desire to conform, against the personal benefit received from being a non-smoker.

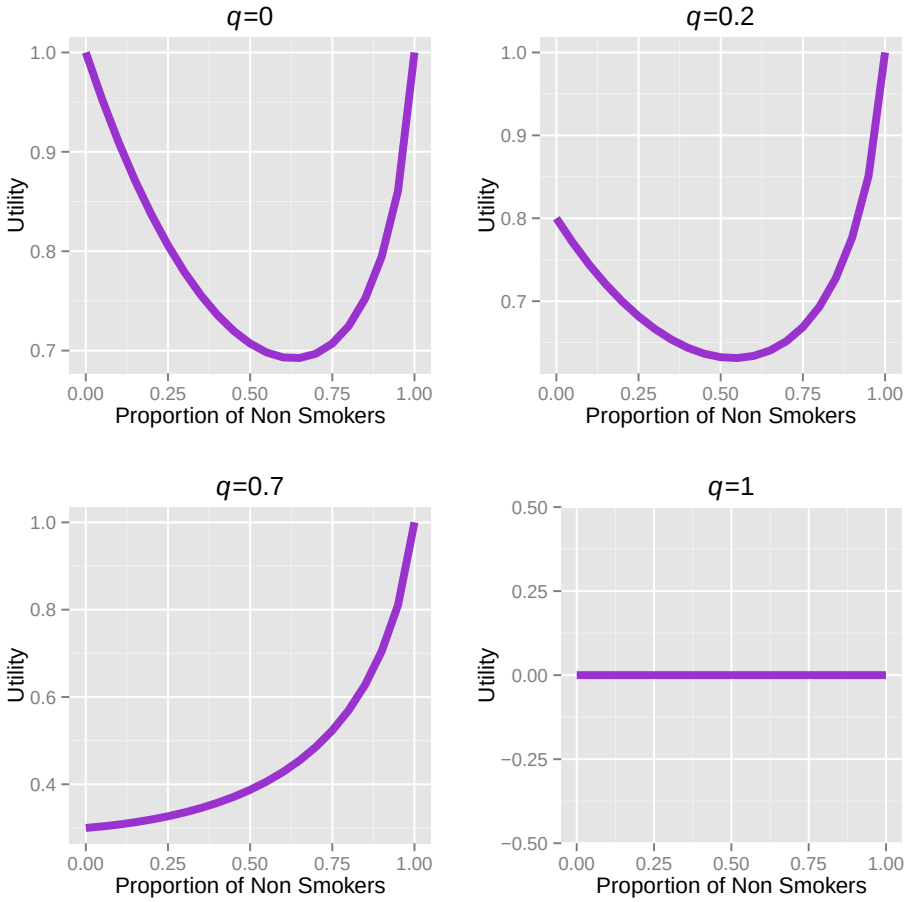


Figure 9.13: The non-smoker utility function from Equation 9.8 with increasing values of q (conformity need).

When the conformity need is low, $q = 0$, and the population are all non-smokers ($\alpha = 1$), the non-smoker utility is high. As the smoker population begins to increase, the non-smokers' personal opinions begin to waver, and the utility begins to decrease. However, at around $\alpha = 0.6$ (40% smoker population), the need to conform is overcome by the

fact that smokers are becoming common place and "unoriginal"; research suggesting that when too many people adopt a new product (or behaviour), it becomes less desirable (Faw, 2013; Jivanda, 2013; Miller, 2013; Zeman, 2011). Then, with a fully smoking population ($\alpha = 0$), the desirability of smoking is negated, and therefore the utility received is again at a maximum (Figure 9.13).

As the need for conformity increases, $q = 0.2$, the "kudos" for being a non-smoker in a smoking majority begins to decrease. When $q = 0.7$, the need for smoker conformity begins to strengthen such that the utility of remaining a non-smoker in a smoker population, does not experience the "originality" surge previously experienced. When the need to conform with smoking is at a maximum, $q = 1$, no utility is received for being a non-smoker irrespective of the population profile. This is to represent that there is no benefit to being a non-smoker in a highly smoker conformist population, even if no current smokers exist. Figure 9.13 illustrates the change in smoker utility, as the desire to conform with smoking behaviours increases.

Model Overview

The literature discussed has outlined a number of important aspects in adolescent smoking, with particular emphasis on the impact to an individual based upon the overall population behaviour. The utilities selected:

$$u(\mathrm{s}, \chi) = \exp(-\alpha)^{\frac{1}{c}} - (1-\alpha)^{p} \tag{9.9}$$

$$u(\mathrm{ns}, \chi) = ((1-q)(1-\alpha))^{1-\alpha} \tag{9.10}$$

have been chosen to best represent the explored literature, while explicitly evaluating the benefits of playing a particular strategy in context of an evolutionary smoker game. Many other factors may also be pertinent in an adolescent's decision to smoke (Simantov et al., 2000; Turner et al., 2006); however, as this is a preliminary investigation using EGT, for simplicity, only the utilities presented shall be explored.

With the development of the utility functions complete, the model can now be explored. If a small enough mutant population is introduced, against which σ^* is stable, then σ^* is an ESS (from Definition 9.2.2). A discussion regarding the process of finding an ESS in the

developed model is presented in the following section (9.2.4).

9.2.4 Finding an ESS

The analysis of the adolescent smoker model, and the identification of its Evolutionary Stable Strategies, is conducted using Sage Mathematical software (Stein, 2014). There are two steps required to find an ESS:

1. finding an equilibrium point;

2. establishing whether the equilibrium point is evolutionary stable.

This section discusses the process by which an ESS is found and provides numerical examples.

The Equilibrium Point

The system is in equilibrium when $u(\text{ns}, \chi) = u(\text{s}, \chi)$, as such, the roots of:

$$((1-q)(1-\alpha))^{1-\alpha} = \exp(-\alpha)^{\frac{1}{c}} - (1-\alpha)^{p}, \tag{9.11}$$

must be established. As this is an intractable equation, a numerical approach is employed. This is conducted using the Brent (2013) method implemented within Sage, which finds roots to certain degree of precision - the default of 1×10^{-10} is selected for this work.

Ideally, the values of c, p and q would be known, or could be estimated from data. Unfortunately, information relating to the specific coolness, personal cost and conformity of smoking is unavailable. As such, a parameter sweep over c, p and q is conducted, finding the values of α that satisfy Equation 9.11.

The selected parameter values relate to a specific smoking scenario, describing the school environment under analysis. For example, when smoking is moderately uncool ($c = 1$), the tolerance to personal cost is moderately low ($p = 1$) and conformity to smoking is high ($q = 1$), then from Equation 9.11:

$$\exp(-\alpha) = (1-\alpha) \tag{9.12}$$

For Equation 9.12, there is an equilibrium point at $\alpha = 0$, when there is a fully smoking population. This is the point at which the population is playing its best response, with $\sigma^* = (1, 0)$. This solution (σ^*) must now be tested to ascertain whether it is an ESS.

Testing Evolutionary Stability

To test if σ^* is evolutionary stable, Definition 9.2.2 is used. Consider a proportion of the population ϵ begins playing some mutant strategy, $\sigma = (m, 1 - m)$. This gives the mutant population:

$$\chi_\epsilon = (1 - \epsilon)\sigma^* + \epsilon\sigma, \tag{9.13}$$

this implies that the proportion of the population playing α in χ_ϵ, is given by :

$$\alpha_\epsilon = (1 - \epsilon)\alpha + m\epsilon. \tag{9.14}$$

Then:

$$u(\sigma^*, \chi_\epsilon) = (1 - \alpha)(\exp(-\alpha_\epsilon)^{\frac{1}{c}} - (1 - \alpha_\epsilon)^p) - \alpha((1 - q)(1 - \alpha_\epsilon))^{1-\alpha_\epsilon} \tag{9.15}$$

and:

$$u(\sigma, \chi_\epsilon) = (1 - m)(\exp(-\alpha_\epsilon)^{\frac{1}{c}} - (1 - \alpha_\epsilon)^p) - m((1 - q)(1 - \alpha_\epsilon))^{1-\alpha_\epsilon} \tag{9.16}$$

If a small enough ϵ can be found such that $u(\sigma^*, \chi_\epsilon) - u(\sigma, \chi_\epsilon) > 0$, then by Definition 9.2.2, σ^* is an ESS.

Returning to the example with $c = 1$, $p = 1$, $q = 1$ and an equilibrium point $\alpha = 0$; if 50% of the population mutate ($\epsilon = 0.5$) and begin playing $\sigma = (m, 1 - m)$, then

$$\chi_\epsilon = 0.5\sigma^* + 0.5\sigma, \tag{9.17}$$

which implies:

$$\alpha_\epsilon = 0.5m. \tag{9.18}$$

Then:

$$u(\sigma^*, \chi_\epsilon) = \exp(-0.5m) + 0.5m - 1 \tag{9.19}$$

and:

$$u(\sigma,\chi_\epsilon) = (1-m)(\exp(-0.5m) + 0.5m - 1) \tag{9.20}$$

As $u(\sigma^*,\chi_\epsilon) - u(\sigma,\chi_\epsilon) > 0$ when $\sigma^* \neq \sigma$, then σ^* is an ESS and the fully smoking population is stable. A graph of the difference between Equation 9.19 and Equation 9.20 is displayed in Figure 9.14.

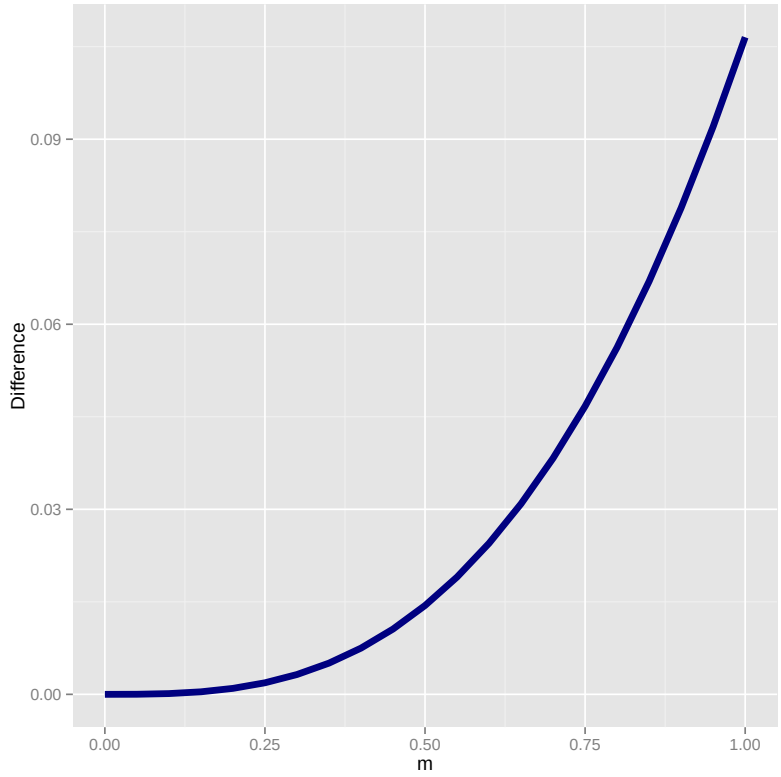


Figure 9.14: Graph of the difference between Equation 9.19 and Equation 9.20 for $m \in [0, 1]$.

Sage is used to evaluate $u(\sigma^*,\chi_\epsilon) - u(\sigma,\chi_\epsilon)$ by conducting a solution sweep of ϵ and m, with the previously defined values of c, p and q. Pseudo code of the ESS finding process in Sage, can be found in Appendix C.2. The Evolutionary Stable Strategies found for the adolescent smoker model are discussed in the following section (9.2.5).

9.2.5 Model Results and Conclusions

A preliminary parameter sweep was conducted to find a region that would produce an ESS. When an ESS was found, the region was explored further with increasing granularity.

Evolutionary stable strategies are discovered when uncoolness is low ($0.05 < c \leq 1$), smoker conformity is at a maximum ($q = 1$), and with varying levels of tolerance to personal cost ($1 \leq p \leq 10$). Figure 9.15 displays the proportion of non-smokers (α) required to achieve an ESS in the selected parameter regions.

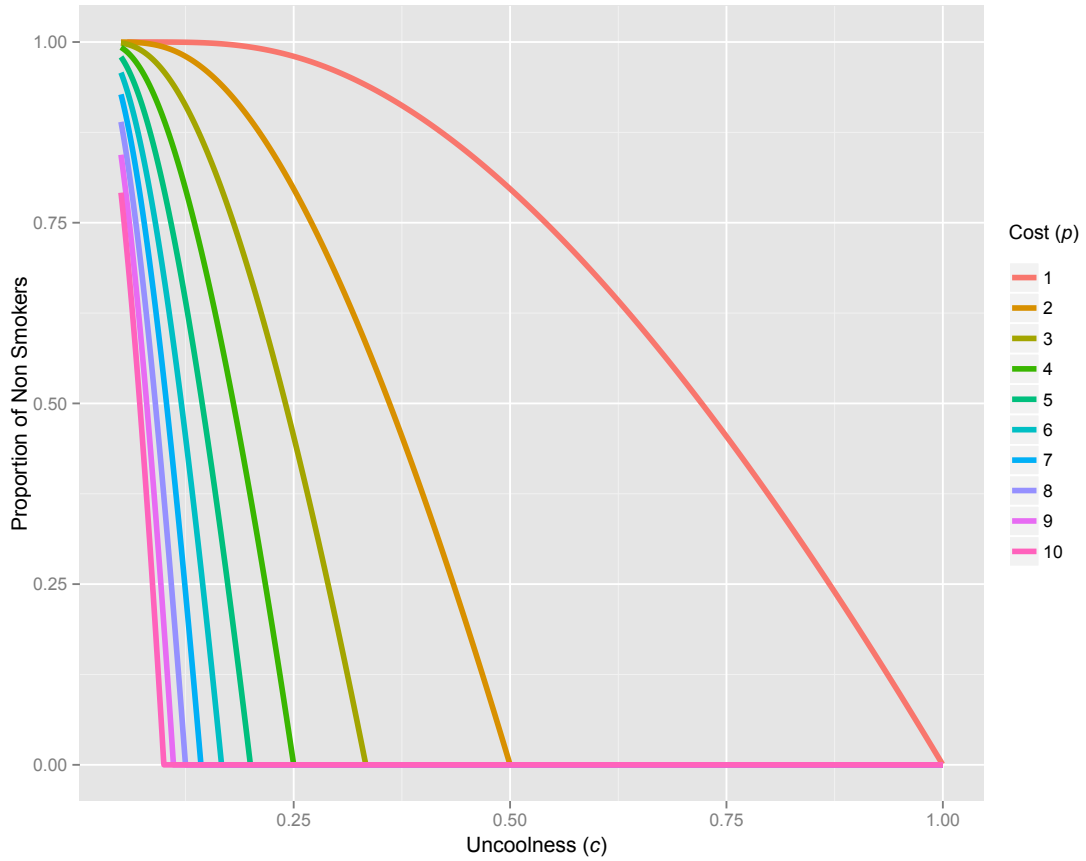


Figure 9.15: The evolutionary stable strategies of the adolescent smoker model, when $q = 1$.

In a population with a high need for conformity to smoking behaviour ($q = 1$), this means that $u(\text{ns}, \chi) = 0$. As a result, when the tolerance to the cost of smoking is low ($p = 1$) and smoking is very cool ($c = 0.05$), an ESS occurs when the proportion of non-smokers is high. As smoking uncoolness begins to increase, the proportion of smokers required to give an ESS increases; this suggests that as smoking is becoming less cool, it is evolutionary stable for more people to smoke. Initially, this may seem counter intuitive, but this is because individuals do not wish to smoke alone - gaining a greater utility for smoking as part of a group.

If smoking is very cool, $c = 0.05$, then the utility from the coolness function is high irrespective of how many individuals actually smoke. However, as smoking becomes uncool, more individuals are required to smoke to obtain a high utility from the coolness function. As previously discussed, because $q = 1$, $u(\text{ns}, \chi) = 0$; thus, to be a stable strategy, $u(\text{s}, \chi) = 0$. This requires $\exp(-\alpha)^{\frac{1}{c}} = (1 - \alpha)^p$, which occurs as $\alpha \to 0$, when $c \to \infty$, for all values of p. Meaning that an ESS is achieved with a higher proportion of smokers, as uncoolness increases.

As the tolerance to smoking costs increase, $p = 10$, evolutionary stable strategies are observed when there is a greater smoking population. This is because the players are trying to achieve their best possible coolness utility, which is when the number of smokers is at a maximum. The increase in tolerance allows more players to smoke, before incurring high personal costs. This results in an ESS with a higher proportion of smokers; Table 9.2 demonstrates the stable proportion of non-smokers (α) decreasing as p increases.

c	p	α
0.05	3	1.00
0.05	4	0.99
0.05	5	0.98
0.05	6	0.96
0.05	7	0.93
0.05	8	0.89
0.05	9	0.84
0.05	10	0.79

Table 9.2: Stable values of α decreasing as p is increasing.

The results of the adolescent game theory model would suggest that, if there are no benefits to being a non-smoker, the idea of smoking is very cool, but there is an extremely high personal cost - then it is evolutionary stable not to smoke. However, if more smokers are required to make smoking cool, then it is evolutionary stable to have a higher proportion of smokers. Furthermore, if the tolerance to smoking behaviours are high, and more smokers are required to make it cool, then it is evolutionary stable for everyone to smoke.

Some example parameters, and whether they induce an ESS, can be observed in Table 9.3. When $u(\text{ns}, \chi) \neq 0$, an ESS could not be found in the parameter sweeps conducted. This demonstrates the volatility of the functions chosen to represent smoking behaviour - mutant strategies being able to unbalance any equilibrium observed. Perhaps with a larger

c	q	p	Stable α	ESS?
0.30	0.76	1.04	N/A	N/A
0.41	0.85	2.64	N/A	N/A
0.48	0.95	3.00	0.61	FALSE
0.50	0.94	2.32	0.30	FALSE
1.00	1.00	1.00	0.00	TRUE

Table 9.3: The stable values of α for various parameters, displaying whether they are an ESS.

parameter search, or alternative utility functions, further adolescent smoking evolutionary stable strategies could be found. However, this work serves to demonstrate the abilities of EGT in modelling complex behavioural dynamics, presenting opportunities for future research.

The results have provided an interesting perspective of adolescent smoking: to achieve a **stable** population of non-smokers, the tolerance to personal costs of smoking must be extremely low, but smoking must be very cool - individuals not requiring the approval of others to adopt the behaviour. This provides initial insights which may be investigated in future research, with a further exploration of adolescent social literature and the development of more representative utility functions required.

The EGT model has presented smoker behaviour as a result of overall population dynamics. This differs from the BPRM model presented in Section 9.1, whereby specific highly eigen-central individuals had the potential to dictate smoking uptake; the specific smoking outcomes being heavily dependent upon the selected consideration of smoker similarity relative to eigen-centrality (parameter d). Although the EGT model considers the individual utilities of the players in the game, it may be considered a more aggregated approach to investigating social influence - the exact position of individuals in a network not being considered. For an additional modelling perspective, compartmental models are considered in Section 9.3.

9.3 SIR Model

The final model of social influence investigated, makes use of a compartmental structure to examine the stages of smoking behaviour in a population. Once again, the model presented

aims to provide an outline of the selected methodology and its application to adolescent smoking - laying the foundations for more representative compartmental models to potentially be developed in future research.

For the investigation of smoking uptake with the proposed model formulation, smoking is interpreted as an *epidemic*. In epidemiology, compartmental models are a common tool used for investigating the spread of infectious diseases (Hethcote, 1994), being originally developed by Kermack & McKendrick (1927) . These epidemiological models have been used to investigate the spread of diseases such as measles (Ferrari et al., 2008; Finkenstädt & Grenfell, 2000; Grenfell et al., 2002), HIV/Aids (Griffiths et al., 2006; Huang et al., 1992; Nowak & May, 1993) and SARS (Ng et al., 2003; Zhou et al., 2004).

The World Health Organisation (WHO) describe tobacco use as an epidemic (World Health Organisation, 2009), with Rowe et al. (1996) suggesting that smoking behaviour is transferred by face-to-face encounters - much like infectious diseases. Various mathematical models related to smoking have been proposed (Darby & Pike, 1988; Ezzati & Lopez, 2003), including generalised models of social and biological contagions (Dodds & Watts, 2005), and compartmental models directly related to adolescent smoking (Rowe et al., 1996, 1992). Thus, investigating smoking in the context of epidemic modelling, appears to be an appropriate alternative approach to investigating social influence.

A basic compartmental model is the SIR, which subdivides the population into three categories: Susceptible, Infected and Recovered. An SIR model is described as having a closed population, meaning that births and deaths do not occur - the total population count being fixed for the duration of the model. For this analysis, the ASSIST population shall be used, with the three SIR compartments described as follows:

- Susceptible (S) - individuals who have never smoked, and thus susceptible to being infected with smoking behaviours;
- Infected (I) - the current smoker population of the system;
- Recovered (R) - individuals who used to smoke, but have since recovered, with recovered individuals assumed to have immunity to further smoking behaviour.

The created smoking SIR ordinary differential equations are constructed such that:

$$\frac{dS}{dt} = -\beta S I \tag{9.21}$$

$$\frac{dI}{dt} = \beta S I - \gamma I \tag{9.22}$$

$$\frac{dR}{dt} = -\gamma I \tag{9.23}$$

where β is the contact and transmission (or uptake) rate of smoking, and γ is the recovery rate (Keeling & Rohani, 2008). A flow diagram of the model is displayed in Figure 9.16.

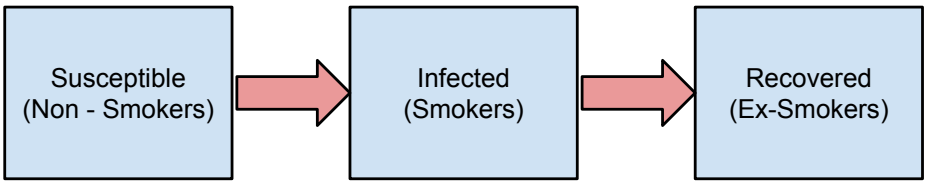


Figure 9.16: Flow diagram of the smoking SIR model.

An important concept of epidemiology is the basic reproductive ratio (R_0), which measures the potential for a disease to spread through a population. Within an SIR model, the reproductive ratio is determined by:

$$R_0 = \frac{\beta}{\gamma} \tag{9.24}$$

If $R_0 > 1$ then the disease outbreak is said to lead to an epidemic, while $R_0 < 1$ indicates that the outbreak will eventually become extinct (Anderson & May, 1992). Some estimated R_0 values for common diseases in human populations include:

- Influenza, $3 < R_0 < 4$ (Murray, 1989);
- Rubella, $6 < R_0 < 7$ (Anderson & May, 1992);
- Measles, $13.7 \leq R_0 \leq 18$ (Anderson & May, 1982).

To create the proposed smoking SIR model, the values of β (uptake rate) and γ (recovery

rate) are obtained from the ASSIST data. All participants from ASSIST are used (irrespective of school type) for this analysis, with the assumption that the population is well mixed - contact patterns and social networks not being considered. It is acknowledged that this is a simplistic assumption, ignoring a great deal of detail regarding social contact. Alternatively, further compartments could also be included; however, the sole purpose of this model is to provide a fundamental understanding of the problem's formulation in a compartmentalised structure - further augmentation being reserved for future research. A full discussion of the limitations of this particular implementation of a compartmentalised structure is presented in the closing remarks of this section.

For simplicity, ASSIST individuals with missing smoking data at any time step (T_0 to T_3) are removed, and ex-smokers who become reinfected are not considered. This gives a total of 7774 individuals with complete smoker data, across the four waves of data collection.

	T_0	T_1	T_2	T_3
Susceptible	7324	6976	6260	5490
Infected	450	715	1304	1808
Recovered	0	83	210	476

Table 9.4: Table of SIR values from ASSIST data.

Rates	$T_0 - T_1$	$T_1 - T_2$	$T_2 - T_3$	Average
Uptake (β)	0.008	0.009	0.010	0.009
Recovery (γ)	0.031	0.015	0.017	0.021

Table 9.5: Table of average monthly smoking uptake and recovery rates from the ASSIST data.

Table 9.4 presents the number of Susceptible, Infected and Recovered individuals at each time period from ASSIST. The monthly uptake rate from t to $t + 1$ is calculated by:

$$\frac{s_t - s_{t+1}}{\bar{n} s_t} \tag{9.25}$$

where s_t is the number of susceptible individuals at time t and $\bar{n}$ is the number of months between t and $t + 1$. The monthly recovery rate is calculated by:

$$\frac{r_{t+1} - r_t}{\bar{n} \bar{I}_t} \tag{9.26}$$

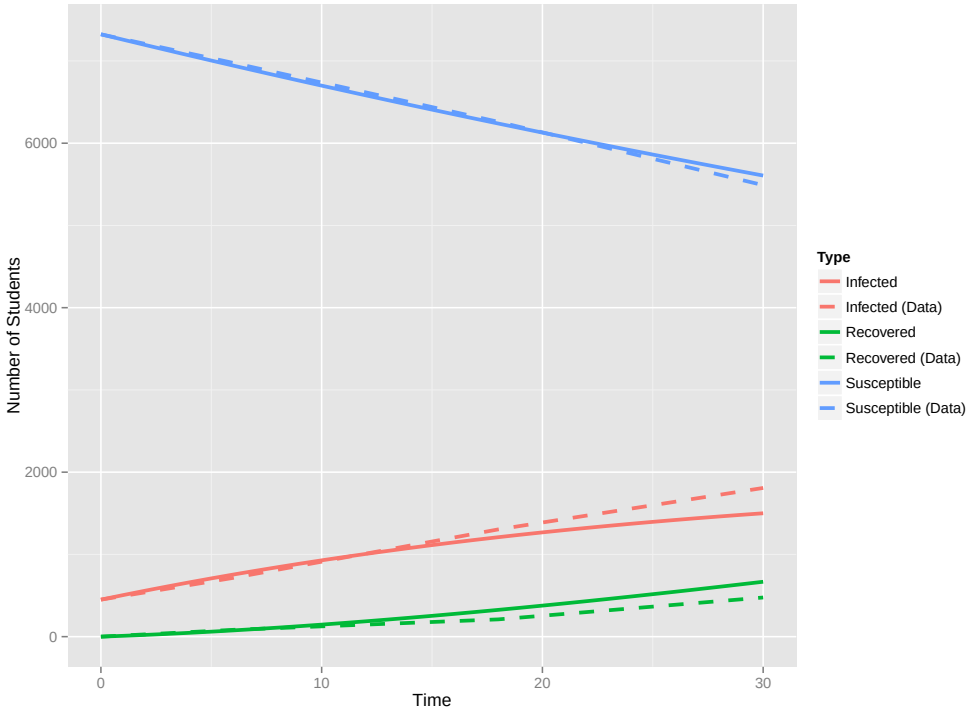


Figure 9.17: SIR model results, with values from the ASSIST data overlaid for 30 months.

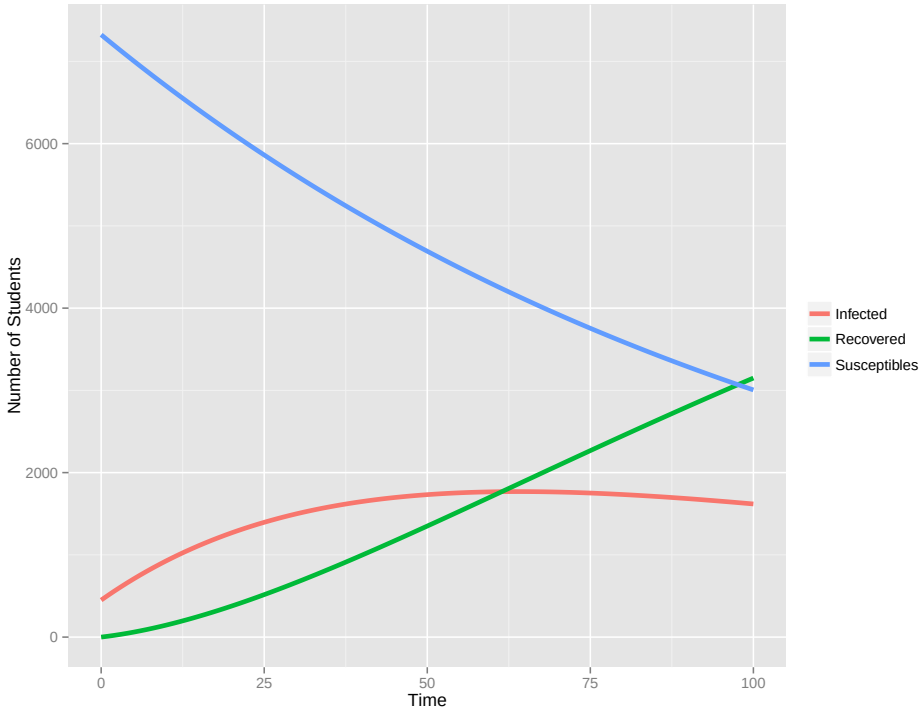


Figure 9.18: SIR model dynamics for 100 months.

where r_t is the number of recovered individuals at time t, $\bar{I}$ is the number of infected individuals at time t and $\bar{n}$ is the number of months between t and $t+1$. Table 9.5 displays the average monthly rates from the ASSIST data. The smoking SIR model is then produced using the initial (T_0) values from Table 9.4, with $\beta = 0.009$ and $\gamma = 0.021$ from Table 9.5.

A graph of the smoker SIR model results, across 30 months, is displayed in Figure 9.17. The graph demonstrates that the SIR model appears to follow the trend of smoking uptake in ASSIST schools well, with predicted uptake and cessation figures being representative of the real data. If the model is run for 100 months (Figure 9.18), the typical SIR curves begin to form, with the number of "infected smokers" recovering at a faster rate than they can infect susceptible individuals. The observed dynamic for 100 months is also a result of the pool of susceptible smokers diminishing greatly over time, a result of the SIR model not allowing feedback into the susceptible state.

The reproductive ratio of smoking is calculated as $R_0 \approx 0.43$. As $R_0 < 1$, this would suggest that the smoking disease outbreak will eventually become extinct in ASSIST schools; hence, smoking behaviour will not become an epidemic. This is contrary to the findings of Section 5.3, which indicated a clear significant increase in school smoking prevalence over time. In terms of the wider UK population, while smoking statistics from ONS (2013b) and Ash.org (2013) indicate a decline in smoking prevalence since 1974, the decline has stalled at around 20% since 2007. Therefore, it would appear that the calculated smoking R_0 is not representative of true smoking dynamics.

The SIR model presented has provided an outline of how social influence may be investigated in a compartmentalised structure. However, the particular model constructed has a number of limitations:

- SIR - the separation of individuals into three compartments is not an appropriate depiction of the cyclic nature of smoking, with ex-smokers having the potential to re-adopt the habit at a later time. The SIR model also fails to capture levels of smoking, which may also have an influence upon the rate of uptake. Alternative compartmental models may be more appropriate, such as an SIRS (Susceptible, Infected, Recovered, Susceptible) structure, or the formulations proposed in Munz et al. (2009) and Miller & Kiss (2014). While further consideration of alternative structures is beyond the scope of this research, the constucted model has laid the

foundation for future research;

- Mixed Population - the assumption that the cohort of ASSIST participants is well mixed is not accurate, as there are individuals from different schools who will not encounter one another. This thesis has extensively demonstrated the importance of social network structure in behavioural influence, as such, this limitation must be overcome if a compartmentalised structure is to be explored further;

- Closed Population - individuals may be influenced (or "infected") by smokers outside of the school population, and new individuals may enter the system and greatly impact smoking uptake. Furthermore, the cohort will eventually leave their respective schools and be introduced into alternative populations.

A further limitation of the model is the assumption that smoking transfers in a manner synonymous with infectious diseases - adolescents needing only to be in close proximity to "infected" smokers to become smokers themselves. This is not the case, with the smoking uptake process said to comprise of many factors (Tyas & Pederson, 1998). However, consideration may be given to such factors when selecting β.

As previously discussed throughout this thesis, social networks are important in the smoking dynamics of an adolescent. While a basic SIR model does not consider network structure, alternative network-based epidemic models have been proposed (Riolo et al., 2001; Rocha et al., 2011); these alternative epidemiological models allow for explicit consideration of graph structures within the "infection" process, accounting for the limitations associated with the assumption of a mixed population. It would appear that, although the basic SIR model presented may not capture the dynamic of adolescent social influence effectively, compartmental models with an underlying network structure may be a viable direction for future research.

This section has provided an introduction to the methods in which social influence (and smoking) may be modelled in an epidemiological context. The SIR model presented, along with the EGT model of Section 9.2, have discussed alternative social influence modelling approaches to that of the BPRM - presenting potential new investigations for future research. Consideration shall be given to all models in the conclusions presented in Section 9.4.

9.4 Social Smoking Outcomes

The methods presented across this chapter have offered alternative approaches to modelling adolescent smoking uptake, with a particular emphasis on techniques which incorporate social factors in a decision making process. Each method offers a unique perspective of an individual's decision to smoke, relative to the population as a whole. Across all the models investigated, there would appear to be a particular parameter which quantifies the smoking behaviour of the population:

- BPRM - the proportion of consideration given to social network structure, d;
- Evolutionary Game Theory - an evolutionary stable strategy (ESS);
- SIR Model - the R_0 of a disease.

Each of the presented model elements, describe a certain point at which some important change occurs in relation to behaviour. The d parameter controls the development of a dominant behaviour, an ESS is a special point of stability in behaviour and the R_0 expresses the point at which a disease becomes an epidemic. In sociological theory, there is a concept known as *the tipping point*, described as the point at which "an idea, trend or social behaviour crosses a threshold, tips and spreads..." (Malcolm, 2000). This concept is also known as a "phase transition" in physical systems.

An example of a tipping point relates to sudden resurgence of "Hush Puppies", the American shoe brand, as market leader in 1996 - taken from (Malcolm, 2000). The brand was on the verge of being phased out by its parent company, selling just 30,000 pairs of shoes a year. Then, in late 1994, young people had started wearing the brand in fashionable bars and clubs in New York. As a result, leading designers requested Hush Puppies for their catwalk shows, causing influential people (such as celebrities) to begin wearing the shoes, resulting in an increase in sales of around 2 million pairs.

The Hush Puppies example, demonstrates how a few individuals of status can have an influence upon behaviour. Literature relating to concepts of the tipping point are observed in marketing theory (Goldenberg et al., 2000; Kotler, 2011; Rogers, 2003), crime (Greene, 1999; Malcolm, 2000) and smoking (Davis, 2000; Wood, 2006). The research surrounding

tipping points, along with the inherent points of change highlighted in the behavioural models, would suggest that modelling adolescent smoking with the proposed models, may provide great insights into smoking uptake behaviours.

Chapter 5 demonstrated how each school responded to the intervention in a unique manner, with the natural progression of smoking in control schools varying greatly. This would suggest that each school has their own tipping point in relation to smoking behaviours, governed by specific factors pertinent to the school environment. The investigation of this important point of behavioural change, may be key in the development of robust cessation methods.

Any of the three presented models could be developed further to investigate school specific smoking tipping points, but the research of this thesis has demonstrated the importance of social networks in behavioural adoption. Furthermore, an individual's smoking behaviour can have higher level influence across the school. This would suggest an ABS perspective, with social network analysis, may be particularly useful.

The BPRM method encompasses the social connection and behavioural aspects required for such analysis, potentially being used as an investigative tool for the further exploration of school specific smoker dynamics. While ABS models of diffusion theory (Remondino, 2008; Schwarz & Ernst, 2009) and opinion dynamics (Hegselmann, 2002; Moore et al., 2011) have been conducted, an investigation specifically related to adolescent smoking behaviours has not been fully explored. This presents great opportunity for future research.

9.5 Chapter Summary

This chapter has presented the BPRM as a method to model the interaction between smoking uptake and friendship selection, demonstrating the impact of highly eigen-central individuals upon population smoking behaviours. Two additional models - Evolutionary Game Theory and SIR - were presented to provide alternative formulations of the investigation, each method offering a differing perspective on the exploration of social influence. Furthermore, this chapter has outlined the viability of the BPRM method to investigate "tipping points" in social systems, proposing novel directions for future research.

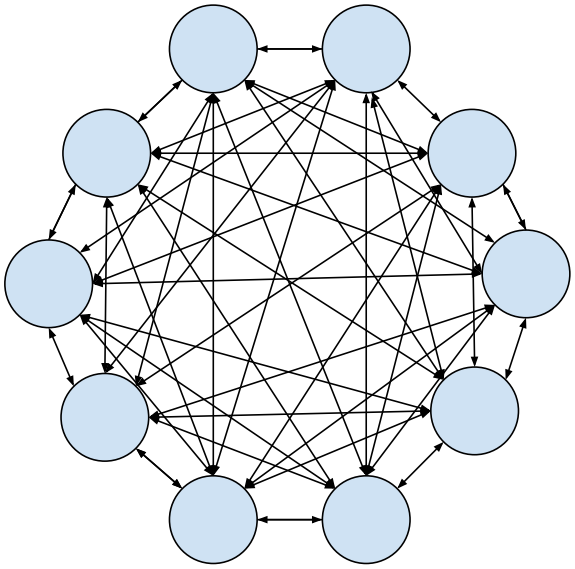
Section 9.1 investigated the underlying smoker behavioural changes of the BPRM method.

The dampening constant, d, was found to have a profound effect upon the subsequent proportion of predicted smokers. To further understand the inner workings of the BPRM method, explanatory tests were conducted - manipulating the initial number of smokers in a test school. The results demonstrated the number of individuals required for smoking behaviour to become dominant, this level varying with alternative values of d. The results also presented the effect of highly ranked individuals, with their behaviours being of particular importance in the identification of a dominant behaviour.

Section 9.2 used an Evolutionary Game Theory model to examine stable smoker or non-smoker strategies in a population. Three theoretical notions, said to be important in adolescent decisions to smoke, were incorporated into the model: coolness, personal cost and conformity. A program to investigate evolutionary stable strategies was created, highlighting specific regions of stability. It was observed that an evolutionary stable non-smoking population could be achieved when smoking alone was very cool, the cost of smoking is high and no utility is gained from not smoking. Again, this research introduced adolescent smoker modelling ideas, which may be further developed in the future.

Section 9.3 provided an introductory account of compartmental models, developing a basic SIR framework of the ASSIST smoker community. The created model represented the data well, however, further augmentation would be required to account for the limitations inherent in the formulated model. In particular, the inclusion of a "Recovered" compartment without the possibility of feedback was highlighted as a key issue. However, the research provided an introduction to modelling smoking as an epidemic, with the consideration of network-based epidemiological models being identified for future research.

Section 9.4 drew together the outcomes of the behavioural modelling research presented. This section highlighted a common theme in the exploratory models discussed, the experiments showing the existence of a tipping point in smoking uptake. The investigation also makes some progress in the identification of factors that determine the position of the tipping point. The BPRM method in particular was identified as a tool to investigate this further, albeit requiring greater development and refinement. This chapter concludes having presented future directions for research not covered in the main body of this thesis, that may be aided by the investigative models presented. With the discussion of the research conducted in this thesis complete, Chapter 10 draws together the overall conclusions and impact of this investigation.



-"A Complete Graph"

10

Conclusions and Recommendations

Chapter 1 introduced the objectives of this research - the investigation of social networks and the influence that their structure may have upon its members. The investigative technique of choice was simulation (Chapter 2), incorporating methods drawn from Link Prediction (LP) literature and Social Network Analysis (SNA). This chapter serves to draw together the conclusions of the research conducted, discussing the main outcomes of the thesis and the directions for future work.

The discussion is structured in the following manner: the specific aims of the research presented in Chapter 1 are revisited, with the key findings in reference to the specified objectives being summarised in Section 10.1; potential for further work is discussed in Section 10.2; recommendations in relation to social network based interventions are proposed in Section 10.3; and Section 10.4 provides the closing remarks of this thesis.

10.1 Research Aims: Revisited

This section compiles the conclusions produced from previous chapters, in reference to the research aims identified in Chapter 1. Each of the research goals outlined shall be addressed in turn, giving particular emphasis to the outcomes gleaned, the specific chapters within the thesis containing its discussion and the novel contributions of the research. The discussion shall also highlight the limitations of the work, with Section 10.2 aiming to use these limitations as a basis for further work.

1. **Apply Agent Based Simulation methods to investigate the effect of social network structures in a theoretical social environment.**

   Chapter 4 used Agent Based Simulation (ABS) methods to investigate the effect of social networks and behavioural factors upon the Peter Principle (PP). The PP states that as individuals ascend in a hierarchical organisation, they become promoted beyond their level of competence. Previous ABS models of the PP suggest promoting at random to avoid this phenomenon, using the assumption that individuals do not retain their competence from lower level positions when adopting managerial roles.

   The findings of a new ABS of the PP (the NBM) - considering social network structure and human behaviour - demonstrated that the social network configuration imposed on the hierarchy, had a substantial effect upon system outcomes. This highlighted that specific network structures, and the placement of particular individuals in a network, is key to the evolution of a connected social system - addressing the first research criterion of this thesis. This sector of research also provided a theoretical underpinning to the empirical analyses conducted in later parts of the thesis, with the NBM demonstrating the applicability of an ABS framework to the investigation of social theory.

   The novel contribution of the NBM is its consideration of social network structure and social theory within the context of an ABS, to the study of managerial incompetence. This advances discussions regarding both the existence of the PP and ways in which to avoid its detrimental effects, providing an introductory analysis of social network effects within this thesis. While the limitations of the NBM primarily relate to the assumptions of the outlined model, these can be addressed with a greater

consideration of social literature and the availability of real world data.

2. **Explore the social network structures of ASSIST to identify important factors in adolescent friendship selection and social influence.**

   Chapter 5 explored the social structures of the ASSIST data, investigating the smoking uptake of adolescents within control and intervention schools. The analysis highlighted specific differences in the responses of individual schools to intervention procedures. Schools with close-knit communities were found to be particularly receptive to intervention diffusion in the early stages of the trial, although this diminished at later time periods. Control schools also contributed to the discussion of social influence, smoking uptake being particularly large in schools with high levels of cohesion.

   The outcomes of the analysis conducted, echoed those of Chapter 4; the overall structure of a social network is important to the evolution of a connected social system, with the position of specific individuals also being key to social influence. In particular, individuals who exhibited high levels of centrality were identified as influential in overall school smoking behaviour - this highlighted the role of centrality as an important factor in adolescent social networks.

   Additionally, the effectiveness of the intervention procedures was also assessed. It would appear that the effect of the intervention diminishes over time, with no quantitative evidence of successful smoking reduction observed at later time periods. The continued evolution of the school social networks was identified as a contributing factor to the attenuation. In particular, the adolescents chosen to diffuse the intervention may be well placed to do so in the initial stages of the trial, however, their network position may become altered over time - reducing their ability to effectively enact their roles. The intervention itself was also suggested as a potential factor in the reduction of a peer supporter's status in the network.

   The ASSIST data analysis highlighted factors pertinent to adolescent friendship selection and social influence, satisfying the second criterion of this research. The role of centrality, and the overall evolution of a social network, are identified as important factors in the diffusion of social influence, contributing directly to the creation of new methods to explore social network evolution and individual behaviour intro-

duced in this thesis. Evidently, a limitation of the social network comparison is that only 18 schools were available for analysis; however, the in-depth analysis of each school provided evidence for the conclusions drawn.

The novel contribution of this sector of research is the application of social analysis techniques to the ASSIST data. While Holliday (2006) analysed the position of nominated peers at T_1 to ascertain their effectiveness in their role, no other study has compared the social network metrics of the ASSIST control and intervention schools. Furthermore, alternative studies have not analysed the full suite of 18 network schools available, across all three time periods; this giving the capacity to provide greater insights into the structure of adolescent social networks.

3. **Develop a new simulation-based approach for the prediction of social network evolution, aiming to incorporate the identified important structural evolution processes of adolescent social networks.**

Chapter 6 outlined the development of a new simulation-based approach for the prediction of social network evolution, PageRank-Max (PR-Max), addressing the third research aim of this thesis. Informed by the analysis of Chapter 5, social network evolution was identified as key to the understanding of social influence; as such, this chapter focused upon the creation of a representative method to model friendship selection.

The created framework provided the ability to predict social network evolution with the newly developed PR-Max algorithm, or one of four existing methods taken from LP literature - Adamic/Adar, Katz, SAB models and PageRank. The existing methods selected were chosen due to their success in a wealth of prior applications, while the PR-Max method was developed to provide an alternative approach based on the optimisation of an agent's eigen-centrality - Chapter 5 concluding centrality important in social network structures.

The inclusion of alternative existing LP methods in the developed simulation, was to provide the ability to assess the accuracy of PR-Max predictions against those of existing methods; thus, a limitation of the study may be the selection of only four existing LP methods to explore in depth. However, given the rigorous selection process of the chosen algorithms (and the time constraints of this research) it was

felt that an appropriate representation of existing LP methods was presented.

The PR-Max algorithm offers a number of novel contributions to both LP and simulation literature. While the Stochastic Actor Based (SAB) method uses simulation as an underlying tool for the generation of statistical models, this thesis is seemingly the first study to structure the LP problem within an ABS framework. The development of the PR-Max method also provides a new approach predicting social network evolution, considering the role of personal eigen-centrality in the friendship selection process. Furthermore, this investigation expands the current literature relating to social applications of simulation, signalling a potential future direction for ABS research.

4. **Evaluate the effectiveness of the developed framework in the prediction of links from the ASSIST dataset, giving particular attention to the differences between schools;**

   Chapter 7 presented the application of the PR-Max algorithm to the ASSIST data. Two types of analysis were conducted: precision - assessing the specific accuracy of the predicted links - and network structure - using the Average Effect Size (AES) to investigate the overall structure of the predicted network. The analysis of Chapter 7 concluded that the PR-Max method was the most successful (of those tested) in predicting the evolution of adolescent friendships, in terms of both precision and network structure.

   The PR-Max method highlighted that centrality (and status) may be an important factor in the evolution of adolescent social networks, especially as the individuals mature - reinforcing the findings of Chapter 5 . This identifies status (an interpretation of eigen-centrality) as a key focus for future investigations of adolescent social networks. Chapter 7 also provided a comparison of control and intervention school Social Network Simulation (SNS) results, demonstrating that intervention schools performed significantly worse in terms of both precision and structural accuracy.

   The results demonstrated the abilities of a simulation-based LP structure in gaining insights unobtainable by conventional SNA. Furthermore, the findings also reinforced conclusions drawn from Chapter 5, as each school had variable responses to the LP methods employed (and their underlying friendship selection processes).

This further identifies the uniqueness of schools, both in terms of friendship selection and intervention response.

The primary limitation in this sector of research, is the availability of school structures for analysis. If a more comparable number of intervention schools had been investigated, a more robust analysis could be produced - providing further weight to the conclusions drawn. An additional limitation is the negative correlation between network size and prediction accuracy for a number of the investigated LP methods; however, this did not appear to be an issue for the PR-Max method, reinforcing its effectiveness in the prediction of links.

The analysis presented appears to have fulfilled the requirements of the fourth criterion. It has demonstrated the success of the PR-Max method in predicting the evolution of adolescent social networks, and contributed a new improved LP method to the literature. The AES network structural analysis also provided a novel approach to examining network predictions, which may be employed across a wealth of studies. Furthermore, this quantitative approach to analysing status, and the search for improved status, in adolescent social networks appears unexplored in previous research.

5. **Create a framework to investigate the interplay between social network structure and smoking behaviours.**

Chapter 8 initialised the investigation of smoking behaviour in a population, through the incorporation of an individual's attributes and behaviours into link predictions. Two alternative LP methods, based upon the PR-Max method, were developed: Behavioural Search and Behavioural PageRank-Max (BPRM). Behavioural Search was deemed inappropriate for modelling the evolution of adolescent social networks, with a limitation being its consideration of dynamic behaviours in a static manner. To rectify this, the BPRM method was developed to give agents the opportunity to alter links or update a specific behaviour.

The BPRM investigation of Chapter 8 focused upon smoking as a changeable behaviour. Small significant improvements to LP precision were observed across each of the test schools explored, when consideration was given to all available aspects of the ASSIST data. This concluded that individual attributes and behaviours are

important in the friendship selection process, but also that friendship selection may impact smoking uptake.

The BPRM improves upon the link predictions of the PR-Max method through the consideration of dynamic behavioural change. Its unique exploitation of PageRank, provides a novel approach to the understanding of network evolution in the context of specific behaviours. A limitation of this research is its application of the BPRM to only four test schools; perhaps with a greater selection of schools, further insights and value could be achieved. Moreover, the BPRM still requires refinement before being adopted as a generic LP method, aspects of which shall be discussed in Section 10.2.

Presented in Chapter 9, is the developed BPRM framework used as a method to investigate the interplay between social network structure and smoking behaviours. When high consideration was given to smoking similarity, agents simply adopted the majority behaviour; however, when network status was important, agents based smoking decisions upon those of highest status in the network. This provided a novel approach to modelling the spread of smoking in conjunction with status, with insight gained into the effect of individuals with high eigen-centrality. A limitation of this model is the inability to quantify the specific balance of similarity and status pertinent to an adolescent; however, this could be a topic of future research.

The created BPRM framework addresses the final criteria of this research, providing a new approach to the investigation of social influence. Additionally, alternative methods of modelling social influence were also presented, an Evolutionary Game Theory (EGT) model and a basic compartmental model (SIR) also being constructed. These additional models demonstrate how the investigation of adolescent smoking and social influence might be structured with an alternative methodology, providing directions for future research. This builds upon the fifth research aim, and demonstrates the breadth of potential that considering a quantitative approach may offer to the investigation of social systems.

10.2 Further Work

This section discusses the opportunities for further work afforded by the conducted research. Each area of interest is introduced by chapter, and a basic outline of the proposed extensions is given.

Chapter 4

As a basic step, the NBM could be extended to include greater aspects of social theory (as discussed in Section 4.6.1), but true insight into the effects of the PP, in relation to promotion methods, cannot be gained unless real world experiments are conducted. Very few studies have attempted to examine the PP in a true workplace environment (Dickinson & Villeval, 2012), this may be due to the subjective nature of job incompetence and the implications of identifying poor performance upon the selected participants. Undoubtedly, a great benefit of creating a simulation model (such as that of the NBM), is the ability to explore alternative scenarios in a safe environment; however, the particular analysis of social structure and influence in the workplace cannot be carried forward effectively unless an appropriate real world context were available.

Chapter 5

While the data analysis of Chapter 5 was extensive, a great deal of investigation may still be conducted into ASSIST. Three extensions in particular are identified:

- *School Environment* - secondary sources of data, such as those relating to regional smoking statistics and school performance, could be cross-referenced against the smoking uptake figures. This provides a greater school context for the investigation;

- *Peer Supporters* - although the nominated peer supporters were discussed in the analysis of Chapter 5, it would be of interest to conduct a thorough investigation into the progression of their network structures and personal attributes;

- *Predicting Smoking Uptake* - A regression model based upon the personal attributes of participants could be created to predict smoking uptake, this giving an alternative view of factors influential in smoking behaviours.

These provide just a few examples of further insights that may be gained from the ASSIST data.

Chapter 7

This study has solely applied the SNS framework to the ASSIST data; therefore, it would be of great interest to investigate the performance of PR-Max upon other networks. The recent work of Sarigöl et al. (2014), demonstrates the importance of centrality in academic literature citation; as such, the PR-Max may particularly excel when exacted upon a network of scientific co-authorships. An analysis of online connections is also of interest, allowing for the examination of PR-Max precision in quantifying internet based relations. Furthermore, analysing a cohort's offline connections alongside their online connections, may provide further insights into social connections and the differences in offline and online friendships.

Chapter 8

Two extensions to the BPRM method research are proposed. The first is the application of the method upon additional data. Further network structures within the ASSIST data could be explored, investigating whether or not improved link predictions are gained across all the schools in the cohort. Additionally, alternative changeable behaviours could be explored upon a wider range of data, developing the framework as a generic tool for investigative research.

The second extension relates to an alteration within the constructed methodology. The improvements in link prediction gained from the BPRM were obtained when all attribute data from ASSIST was used. However, this data contained many variables which may be irrelevant to the LP process - hindering the accuracy achieved; as such, a factor analysis of the data is proposed. This would take the number of observed variables and reduce them into a smaller selection of explanatory factors - the observed variables becoming linear combinations of the newly developed factors. An assumption of factor analysis is that the technique is performed upon continuous data, with the ASSIST data predominantly containing nominal variables; this in itself presents statistical challenges to explore further.

Chapter 9

Each of the models presented in Chapter 9 (BPRM, EGT, SIR) could be taken forward for future research. These include:

- *BPRM Model* - the BPRM model could be used to classify an individual's consideration of similarity and social network structure through extensive investigations of social network data; however, this would require further refinement of the BPRM method as a whole;

- *EGT Model* - additional parameters of the EGT model could be defined, with parametrisation being based on real world data. Literature relating to evolutionary graph theory could also be investigated (D'Onofrio et al., 2013; Javarone & Armano, 2012; Shakarian et al., 2012), whereby a mutant gene takes hold of structured populations. This may be a method to incorporate social networks into the EGT model;

- *SIR Model* - a more representative compartmental structure could be developed, giving better model assumptions. Additionally, examining literature relating to graph theory in epidemic modelling (Miller et al., 2012; Miller & Volz, 2013; Rocha et al., 2011), could present a future direction of this research.

Overall the examples of further work presented, demonstrate the breadth of potential research emanating from this thesis.

10.3 Recommendations

The discussions within this thesis have identified the importance of centrality in adolescent social networks, with said networks continually evolving over time. As such, this section relates the conclusions of the research back to the design of the ASSIST intervention. While this thesis has found no overall quantitative evidence of a reduction in smoking behaviours due to ASSIST, previous research has indicated the success of the trial in specific schools from a qualitative perspective. As such, the Scottish government is piloting ASSIST in a number of secondary schools (The Scottish Government, 2013). This section presents a number of recommendations for the new trial, to potentially gain greater levels of success across a wider variety of schools.

1. **Social Network Analysis**

   In the initial selection of peer supporters, social network analysis was not employed. The message diffusers were simply selected by asking participants to identify: respected fellow students; leaders in sports or group activities; and individuals who are "looked up" to in Year 8. Although these individuals may have the respect of fellow students, they may not necessarily be placed in the best network position to diffuse the intervention. As such, this thesis recommends also selecting individuals based upon their closeness, betweenness and eigen-centrality.

2. **Data**

   A great barrier to the analysis of ASSIST was the availability of data. As questionnaires were completed in paper form, this required a team of data entry clerks to furnish the database, with 34 schools of social network data still remaining unentered. The increased access to mobile phones and tablet computers could allow participants to enter data directly into an online form, avoiding the need for paper based responses - streamlining the data collection process. Evidently some expertise in database structuring techniques would be required to match students and friendship nominations, however, the process will be vastly quicker than manual data entry.

3. **Re-evaluate Peer Supporters**

   This research (and previous investigations) indicate an attenuation of the intervention over time, which may be a result of peer supporters no longer being well placed to exact the intervention. The initial ASSIST study selected the top 17.5% of nominated participants to be peer supporters. To investigate their positioning, the proportion of peer supporters residing in the top 17.5% of eigen-central individuals was calculated for each school; the proportions for T_1 are displayed in Figure 10.1.

   Figure 10.2 depicts the change in the proportion of peer supporters in the eigen-central group between T_3 and T_1. The results demonstrate a reduction in the proportion of highly eigen-central peer supporters over time, for the majority of schools. This suggests that other individuals are replacing the peer supporters in terms of status at T_3. Therefore, this thesis recommends the re-evaluation of peer supporters and the selection of new individuals to exact the intervention at later time periods.

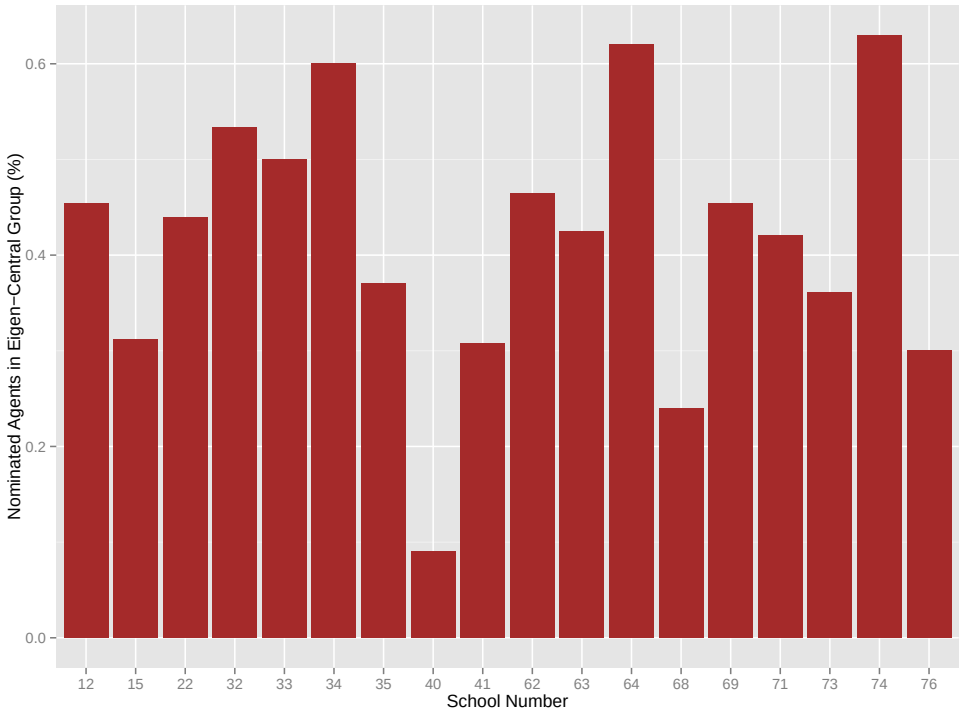


Figure 10.1: The proportion of nominated individuals in the eigen-central group at T_1.

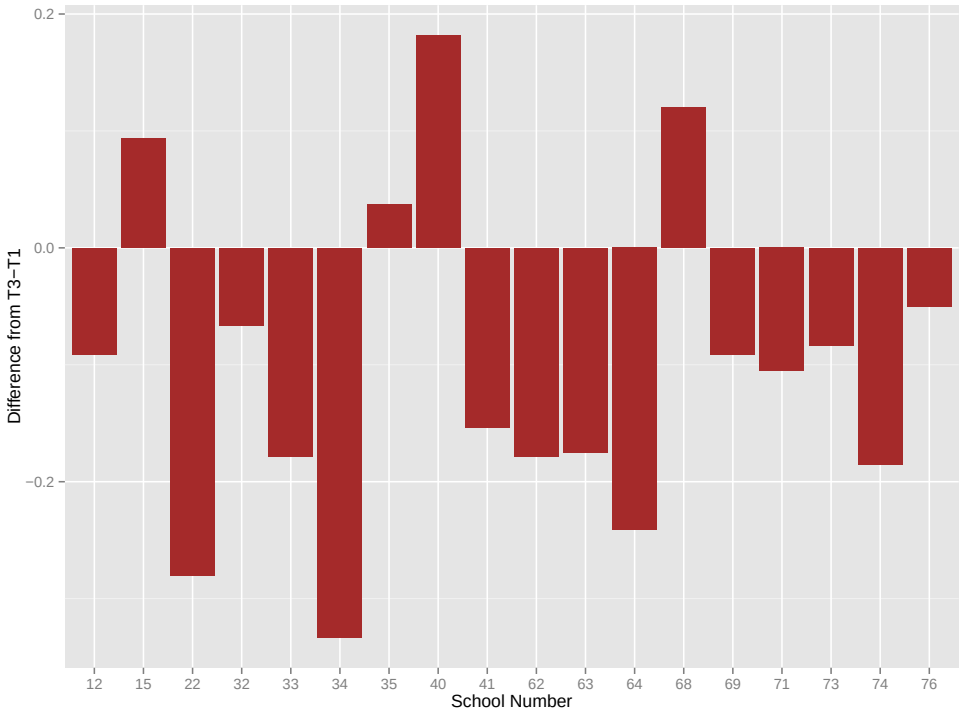


Figure 10.2: The proportional difference in the number of nominated agents in the eigen-central group at T_3 compared to T_1.

4. **Targeted interventions**

   The research has highlighted that schools differ both in their response to intervention and the aspects important in friendship selection. As such, a contextual understanding of the schools may be of benefit. Influential students may not necessarily be individuals who have a great deal of connections, or those who are respected leaders. For example, there may be sports clubs or activities within the schools which breed smoking behaviour, which results in its proliferation throughout the school. Understanding school specific social situations and targeting them, may increase the effectiveness of the intervention - although this would require a greater amount of school participation.

The recommendations have presented four alterations to the ASSIST procedures. In particular, recommendations one (Social Network Analysis) and two (Data) are thought to be of key importance. Recommendation one will allow for the selection of more network-central peer supporters, and recommendation two will allow for quicker access to the data during the trial - meaning that real time analysis can be conducted to assess the effectiveness of the current intervention framework. With greater consideration of SNA and current data, a more successful intervention may potentially be achieved.

10.4 Closing Statements

This thesis has investigated the dynamics of social networks and the influence that their structure may have upon its members. It has presented a novel approach to the prediction of social network evolution, and developed a new method to examine the relationship between social network structure and behavioural influence. Additionally, centrality measures have been highlighted as important in the friendship evolution process and identifying influential individuals in a network. The techniques employed, have played a pivotal role in furthering the understanding of social networks - an area of research which appears to be gathering considerable momentum. Furthermore, this thesis has demonstrated the value of social network analysis, Link Prediction methods and Agent Based Simulation, contributing insights into the individual and collective dynamics of social connection.

Appendices

A The Peter Principle

A.1 NBM Network Statistics

		Common Sense			Peter Principle		
		SF	SW	RAN	SF	SW	RAN
Mean Degree	Best	1.01	1.85	0.86	0.99	1.79	0.84
	Worst	0.98	1.85	0.85	0.95	1.80	0.84
	Random	1.02	1.88	0.85	1.00	1.87	0.84
Network Efficiency	Best	0.29	0.11	0.60	0.30	0.12	0.62
	Worst	0.29	0.12	0.48	0.30	0.11	0.52
	Random	0.30	0.12	0.51	0.30	0.11	0.54

Table A.1: Average steady state network statistics exclusive of warm up period.

Table A.1 refers to the generated network statistics of the NBM. It is evident that a pattern of results may be observed. SW networks produce the highest mean degree, followed by SF networks, with RAN network agents possessing the lowest number of average connections. In terms of Network Efficiency (NE), RAN generates the most connected network, followed by SF and SW respectively.

While mean degree statistics appear to exhibit similar behaviour across promotion methods, NE figures fluctuate - particularly under RAN conditions. The fluctuations may be attributed to the negative envy generated by the worst and random promotion methods. Agents drop competence following an unjust promotion, potentially to the levels of being ejected from the system - an issue regularly encountered in the worst and random promotion methods. Subsequently, agents leave the system and sever previous ties, meaning overall NE reduces.

A.2 γ effect

This section gives a brief description of the effects when varying γ. The γ effect assesses the sensitivity of the simulation in relation to incremental envy, the original selection being $\gamma = 1$. The random methods (PP Random, Figure A.3; CS Random, Figure A.6) show similarity in terms of increase as γ rises, although CS efficiency demonstrates a larger upward trend. This may be due to the CS method insighting more positive envy as $\gamma \rightarrow 5$,

increasing overall efficiency.

CS Best (Figure A.4) and PP worst (Figure A.2) are opposing to one another, potentially due to the maximum efficiency levels discussed in the main body of the thesis; the simulation producing highly competent agents that cannot gain promotion as $\gamma \rightarrow 5$, resulting in a competence peak. Conversely, PP Worst creates poorly performing agents (a product of negative envy) which become eradicated sooner as $\gamma \rightarrow 5$.

Finally CS Worst (Figure A.5) produces variation in terms of topology - Scale Free networks appear markedly different to those of Small World and Random. If the worst agent in being promoted under CS conditions, evidently system efficiency will be minimal; given that the SF network produces a highly connected individual, who bypasses all other promotion rules, they may then precipitate positive envy and increase the competence of those connected. The resulting dynamic is a sizeable increase to efficiency.

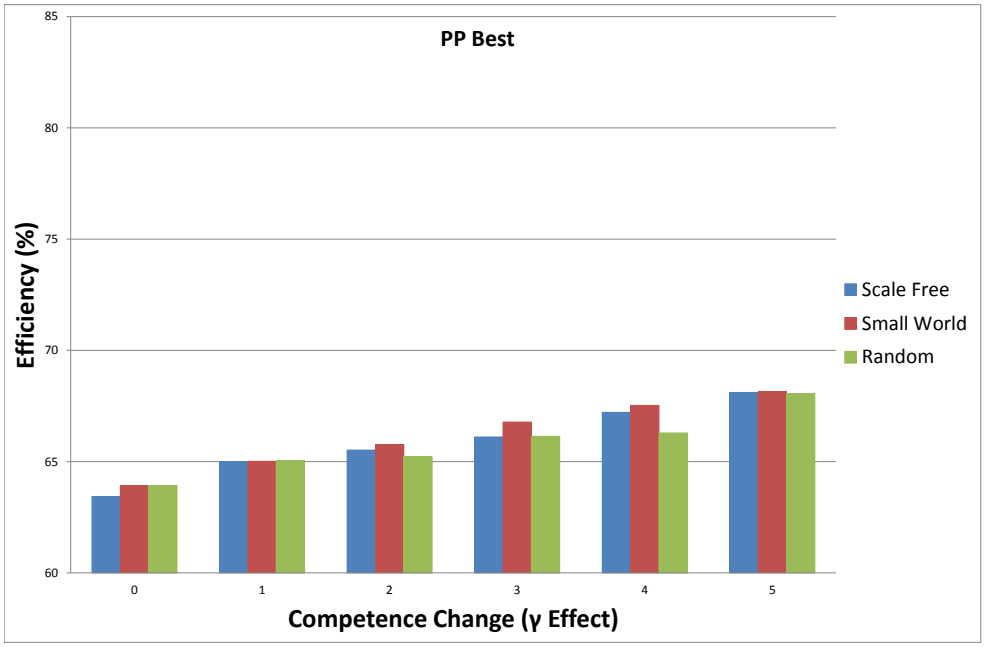


Figure A.1: PP Best γ effect upon averaged steady state efficiencies.

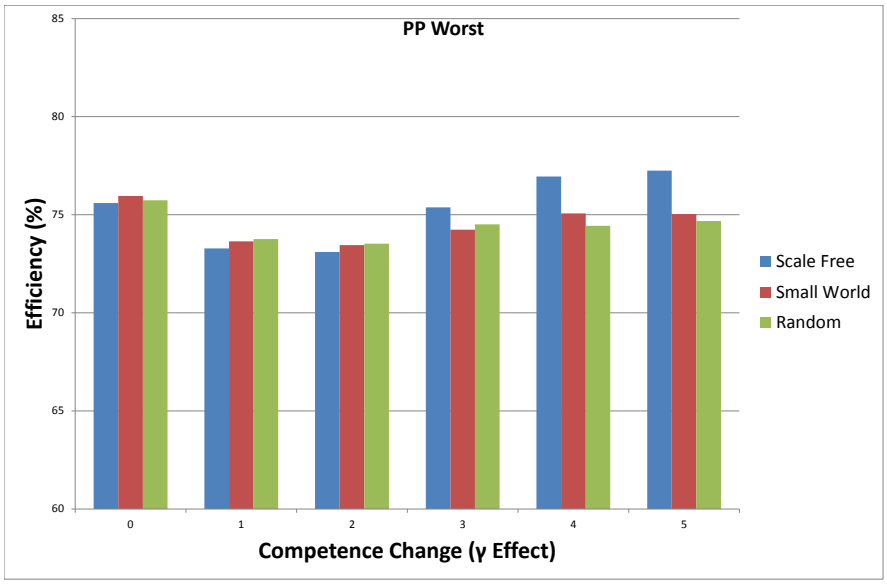


Figure A.2: PP Worst γ effect upon averaged steady state efficiencies.

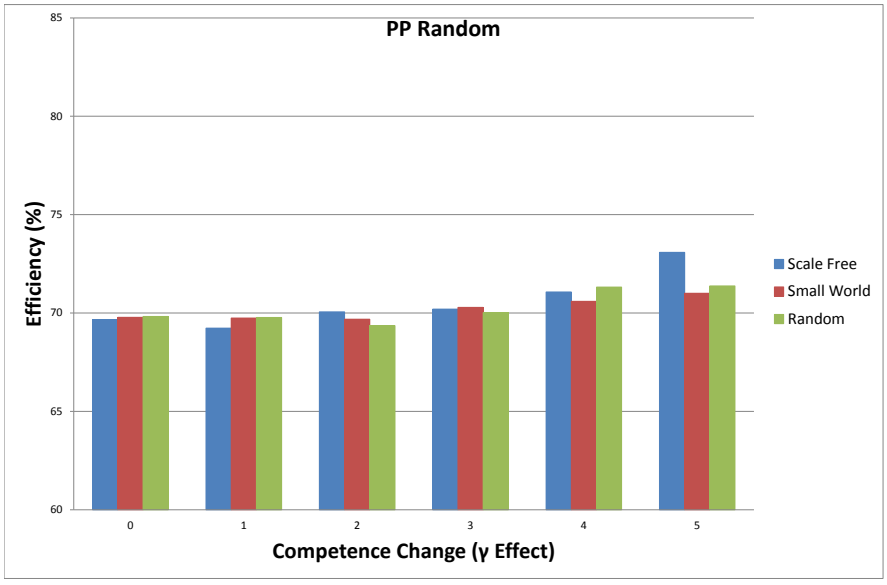


Figure A.3: PP Random γ effect upon averaged steady state efficiencies.

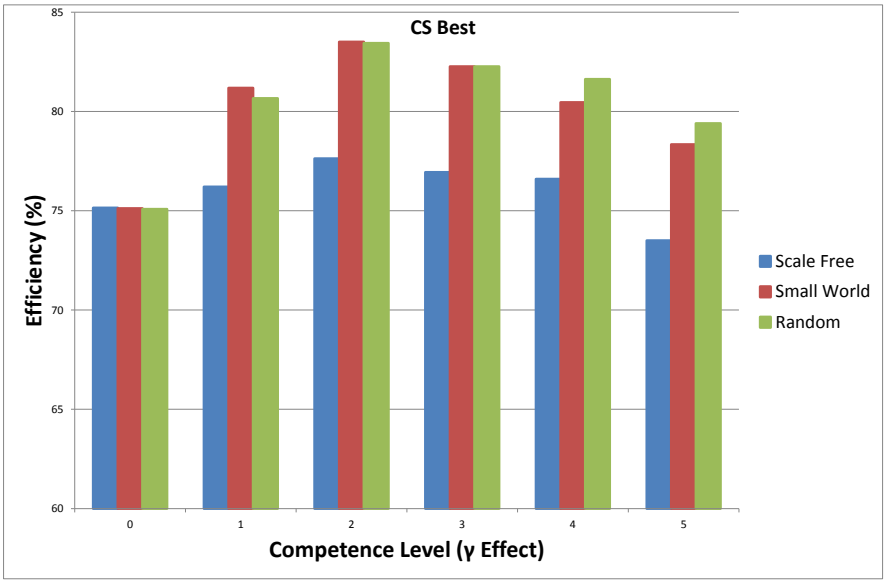


Figure A.4: CS Best γ effect upon averaged steady state efficiencies.

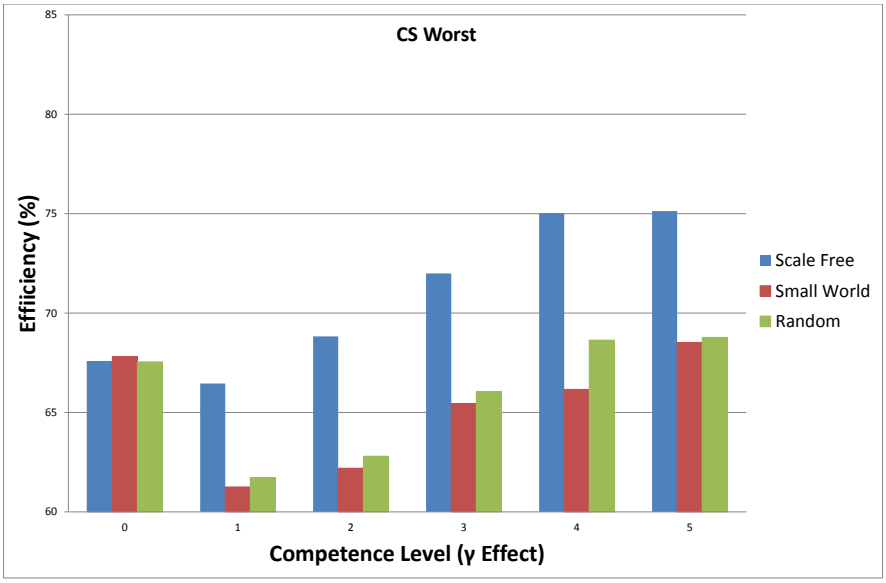


Figure A.5: CS Worst γ effect upon averaged steady state efficiencies.

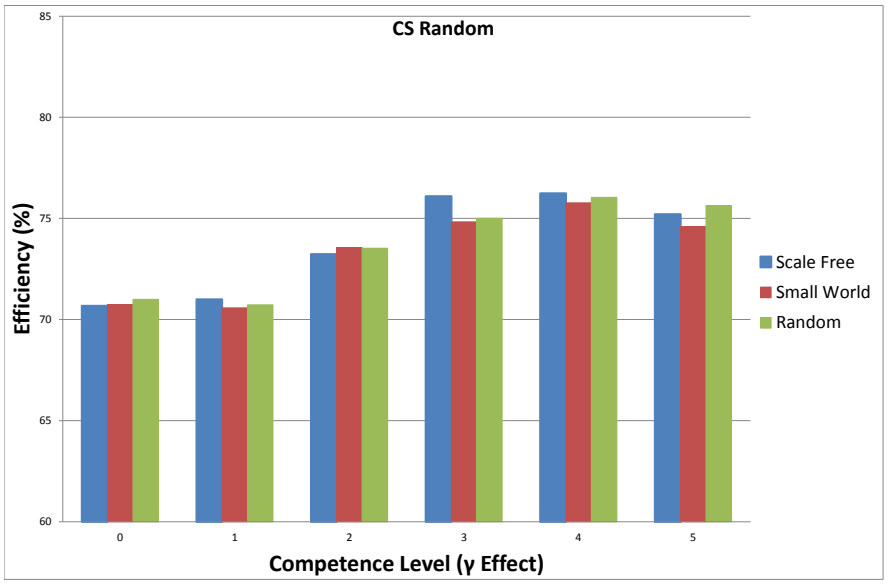


Figure A.6: CS Random γ effect upon averaged steady state efficiencies.

B Data Analysis: ASSIST

This appendix presents a copy of the ASSIST social network questionnaire. The documents overleaf request participants name up to six friend and provide details of their interactions. The particular questionnaire displayed in this appendix was issued to year 10's - the final wave of data collection. A sample of the personal data questionnaires were unavailable for inclusion.

Yr 10

ASSIST RESEARCH PROJECT

FRIENDS QUESTIONNAIRE

Please read these instructions before filling in this questionnaire.

1. ***Do not put your name*** anywhere on the questionnaire.

2. Please fill in the questionnaire on your own and do not talk to anyone.

3. Think about your friends and ***fill in the table on the first page*** with the ***first name and surname*** of up to ***six friends. Please do not use nicknames.*** If you have more than six friends, just write down the names of your six closest friends.

4. You can name ***any of your friends, they don't have to be in your year.*** You ***can include friends who do not go to your school.*** You can name both ***boys*** and ***girls.***

5. Once you have done that, carry on with the rest of the questionnaire, filling in ***one page for each friend you named in the table.*** If you named one friend in the table, please fill out one page. If you named two friends, please fill out two pages, and so on for ***up to six friends.***

6. When you fill in a page, please write the ***first name and surname*** of your friend at the top of the page.

7. If your friend goes to school, please write down the ***name of the school*** they go to.

8. If your friend is at ***your*** school, then write down their ***form/tutor group*** if you know it. If you do not know what form/tutor group your friend is in, but know the year that they are in, please write this down. If you only know their ***form tutor's name*** please write this down instead.

9. If your friend goes to a ***different*** school, please write down which ***year*** they are in or their ***age.***

10. Remember, your answers are ***confidential*** - they will only be seen by the research team.

BEFORE YOU FILL IN THE NEXT PAGES, PLEASE WRITE THE FULL NAMES OF YOUR FRIENDS BELOW.

If you have one friend, please write one name below. If you have two friends, please write two names and so on. If you have more than six friends, write the names of your six closest friends below.

Remember that you can name any of your friends.
The order you name your friends does not matter.
You can include friends who do not go to your school.

	Name of friend (first name and surname)
1	
2	
3	
4	
5	
6	

Remember that the answers you give are confidential.

PLEASE TURN TO THE NEXT PAGE AND ANSWER SOME QUESTIONS ABOUT THE FRIEND/FRIENDS YOU HAVE JUST NAMED.

Name of friend 1 (first name & surname) _____

If at school, which school? _____

If at YOUR school, which form/tutor group? _____ **or form tutor's name?**_____

If at a DIFFERENT school, which year group? _____ **or their age?** _____

Answer the questions on this page for the friend you have named above

1a) Is this friend (Please tick **one box** only)

A best friend ☐ 1 Just a friend ☐ 2

1b) Is this friend (Please tick **one box** only)

A boy ☐ 1 A girl ☐ 2

1c) This friend (Please tick **one box** only)

Is in year 10 at my school ☐ 1

Is in a year below year 10 at my school ☐ 2

Is in a year above year 10 at my school ☐ 3

Is at another school ☐ 4

Has left school ☐ 5

1d) When do you see each other? (Please tick **one box** only)

In school only ☐ 1

In and out of school ☐ 2

Out of school only ☐ 3

1e) How would you describe your friendship? (Please **tick Yes or No for each line**)

	Yes	No
We do activities together (sport, computer games etc.)	☐ 1	☐ 2
We just hang out but don't do activities together	☐ 1	☐ 2
We are close and talk a lot together	☐ 1	☐ 2
We are like each other	☐ 1	☐ 2
We think the same way	☐ 1	☐ 2

C Evolutionary Game Theory

C.1 Further Example

The following example illustrates the process of finding an ESS in a game against the field, adapted from Webb (2007). Consider a population of males and females such that:

- The proportion of males is α, and the proportion of females is $1-\alpha$;
- Each female selects one mate and produces K offspring;
- Males have on average $\frac{(1-\alpha)}{\alpha}$ mates;
- Females decide the sex of the offspring.

There are two strategies available to the females, produce only male offspring or only female offspring. Thus, a general strategy $\sigma = (\omega, 1-\omega)$ produces a population of ω male offspring and $1-\omega$ female offspring. Initially $\chi = (\alpha, 1-\alpha)$, and the utilities of the available strategies are:

$$u(M,\chi) = \frac{1-\alpha}{\alpha}K^2 \tag{C.1}$$

$$u(F,\chi) = K^2 \tag{C.2}$$

Therefore:

$$u(\sigma,\chi) = K^2(\omega\frac{1-\alpha}{\alpha} + (1-\omega)) \tag{C.3}$$

If $\alpha \neq \frac{1}{2}$ then $u(M,\chi) \neq u(F,\chi)$ and the population is not stable, as population profile will vary based on the utilities. To investigate whether $\sigma^* = (0.5, 0.5)$ is an ESS, consider a mutant strategy $\sigma = (p, 1-p)$ and $\chi_\epsilon = (1-\epsilon)\sigma^* + \epsilon\sigma$. This implies:

$$\alpha\epsilon = (1-\epsilon)\frac{1}{2} + p\epsilon = \frac{1}{2} + \epsilon(p - \frac{1}{2}) \tag{C.4}$$

It is known that:

$$u(\sigma^*,\chi_\epsilon) = \frac{1}{2} + \frac{1-\alpha_\epsilon}{2\alpha_\epsilon} \tag{C.5}$$

and:

$$u(\sigma, \chi_\epsilon) = (1 - p) + p\frac{1 - \alpha_\epsilon}{\alpha_\epsilon} \tag{C.6}$$

Then the difference:

$$u(\sigma^*, \chi_\epsilon) - u(\sigma, \chi_\epsilon) = (\frac{1}{2} - p)\frac{1 - 2\alpha_\epsilon}{\alpha_\epsilon} \tag{C.7}$$

If $p < \frac{1}{2}$ then $\alpha_\epsilon < 2$ which means that $u(\sigma^*, \chi_\epsilon) - u(\sigma, \chi_\epsilon) > 0$. Also if $p > \frac{1}{2}$ then $\alpha_\epsilon < 2$ which means that $u(\sigma^*, \chi_\epsilon) - u(\sigma, \chi_\epsilon) > 0$. Therefore, $\sigma^* = (0.5, 0.5)$ is an ESS. This illustrates the process of finding an ESS analytically for a game against the field.

C.2 Pseudo Code For Finding an ESS

Algorithm 1 Finding an ESS

```
for c ← 0.05 to n do
  for p ← 0 to m do
    for q ← 0 to i do
      Find Roots of ((1 − q)(1 − α))^{1−α} − exp(−α)^{1/c} − (1 − α)^p
      for (Set of all Roots) do
        ϵ ← 0.01
        plot ← Plot of u(σ*, χ_ϵ) − u(σ, χ_ϵ) over m ∈ [0, 1]
        if plot> 0 when α ≠ m then
          append to list (c, p, q, α,true)
        else
          append to list (c, p, q, α,false)
        end if
      end for
    end for
  end for
end for
```

D BPRM Smoker Predictions

This appendix presents the tables of BPRM smoking predictions. Table D.1 displays the predicted smoker proportions at T_2 and T_3 from the BPRM method, which may be compared with the true smoking proportions for each of the selected test schools (Table D.2). The accuracy of predictions for each time step is presented in Tables D.3, giving the proportion of agents possessing the correct smoking value.

School	Type	T_2			T_3		
		0.15	0.50	0.85	0.15	0.50	0.85
12	Smoke	0.46	0.26	0.03	0.88	0.82	0.13
	Gender and Ethnicity	0.03	0.03	0.03	0.04	0.04	0.04
	Form	0.11	0.03	0.02	0.51	0.04	0.04
	Nominations	0.74	0.47	0.03	0.74	0.71	0.70
	Levenshtien	0.02	0.02	0.02	0.03	0.02	0.02
33	Smoke	0.60	0.29	0.08	0.59	0.19	0.19
	Gender and Ethnicity	0.07	0.06	0.05	0.19	0.18	0.19
	Form	0.07	0.05	0.05	0.44	0.19	0.19
	Nominations	0.57	0.39	0.06	0.83	0.58	0.19
	Levenshtien	0.02	0.03	0.02	0.22	0.12	0.10
71	Smoke	0.69	0.65	0.28	0.69	0.83	0.30
	Gender and Ethnicity	0.27	0.27	0.27	0.30	0.30	0.30
	Form	0.27	0.27	0.27	0.30	0.30	0.30
	Nominations	0.73	0.76	0.28	0.74	0.72	0.31
	Levenshtien	0.28	0.23	0.23	0.30	0.29	0.27
74	Smoke	0.51	0.24	0.04	0.34	0.15	0.15
	Gender and Ethnicity	0.04	0.05	0.05	0.23	0.15	0.15
	Form	0.04	0.04	0.05	0.38	0.15	0.15
	Nominations	0.60	0.51	0.05	0.61	0.30	0.15
	Levenshtien	0.03	0.03	0.03	0.16	0.14	0.13

Table D.1: BPRM predicted smoker proportions at T_2 and T_3 for each of the similarity matricies with varying values of d.

School	T_2	T_3
12	0.20	0.26
33	0.24	0.36
71	0.37	0.42
74	0.19	0.25

Table D.2: True smoking proportions for the ASSIST school data at T_2 and T_3

School	Type	T_2			T_3		
		0.15	0.50	0.85	0.15	0.50	0.85
12	Smoke	0.49	0.64	0.80	0.21	0.25	0.70
	Gender and Ethnicity	0.80	0.79	0.80	0.77	0.76	0.76
	Form	0.74	0.79	0.80	0.44	0.76	0.77
	Nominations	0.31	0.50	0.79	0.32	0.29	0.29
	Levenshtien	0.80	0.80	0.80	0.75	0.75	0.75
33	Smoke	0.40	0.65	0.79	0.49	0.74	0.74
	Gender and Ethnicity	0.79	0.78	0.78	0.74	0.74	0.74
	Form	0.79	0.78	0.78	0.58	0.73	0.74
	Nominations	0.39	0.52	0.77	0.27	0.49	0.73
	Levenshtien	0.78	0.78	0.78	0.66	0.72	0.69
71	Smoke	0.42	0.44	0.69	0.51	0.40	0.75
	Gender and Ethnicity	0.69	0.68	0.69	0.75	0.75	0.75
	Form	0.69	0.68	0.69	0.76	0.75	0.75
	Nominations	0.35	0.34	0.69	0.36	0.34	0.62
	Levenshtien	0.72	0.69	0.67	0.75	0.75	0.75
74	Smoke	0.41	0.63	0.77	0.49	0.62	0.62
	Gender and Ethnicity	0.77	0.77	0.77	0.55	0.62	0.62
	Form	0.77	0.78	0.77	0.46	0.62	0.62
	Nominations	0.33	0.39	0.77	0.31	0.51	0.62
	Levenshtien	0.78	0.78	0.79	0.61	0.63	0.63

Table D.3: Smoking prediction accuracy for each of the BPRM similarity matrices at T_2 and T_3, for varying values of d.

Bibliography

Abrahams, M. (2010). Random promotion may be best, research suggests. *The Guardian*.

Acosta, P. (2010). Promotion dynamics the Peter Principle: Incumbents vs. external hires. *Labour Economics*, 17(6):975–986.

Adamatzky, A. (2010). *Game of Life Cellular Automata*. Springer, London.

Adamic, L. A. and Adar, E. (2003). Friends and neighbors on the web. *Social networks*, 25(3):211–230.

Agarwal, N. and Zeephongsekul, P. (2013). Psychological Pricing in Mergers & Acquisitions using Game Theory. *Studies in Economics and Finance*, 30(1):22–30.

Alam, S. J. and Geller, A. (2012). Networks in Agent-Based Social Simulation. In Heppenstall, A. J., Crooks, A. T., See, L. M., and Batty, M., editors, *Agent-Based Models of Geographical Systems*, pages 199–216. Springer Netherlands, Dordrecht.

Albert, R. and Barabási, A. L. (2000). Topology of evolving networks: local events and universality. *Physical review letters*, 85(24):5234–7.

Albert, R. and Barabási, A. L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47–97.

Albert, R., Jeong, H., and Barabási, A. L. (2000). Error and attack tolerance of complex networks. *Nature*, 406(6794):378–82.

Alexander, C., Piazza, M., Mekos, D., and Valente, T. (2001). Peers, schools, and adolescent cigarette smoking. *Journal of Adolescent Health*, 29(1):22–30.

Alexander Jr, C. N. (1963). A method for processing sociometric data. *Sociometry*, 26(2):268–269.

Allender, S., Balakrishnan, R., Scarborough, P., Webster, P., and Rayner, M. (2009). The burden of smoking-related ill health in the UK. *Tobacco control*, 18(4):262–7.

Aloise-Young, P. A., Hennigan, K. M., and Graham, J. W. (1996). Role of the self-image and smoker stereotype in smoking onset during early adolescence: a longitudinal study. *Health Psychology*, 15(6):494–7.

AltaVista (2012). www.altavista.com, Last Accessed: Aug 2012.

Anderson, R. M. and May, R. M. (1982). Directly transmitted infectious diseases: control by vaccination. *Science*, 215(4536):1053–1060.

Anderson, R. M. and May, R. M. (1992). *Infectious Diseases of Humans: Dynamics and Control*. OUP, Oxford.

Anderson, R. M., May, R. M., and Boily, M. C. (1991). The spread of HIV-1 in Africa: sexual contact patterns and the predicted demographic impact of AIDS. *Nature*, 352(6336):581–589.

Andrighetto, G., Cranefield, S., Conte, R., Purvis, M., Purvis, M., Savarimuthu, B. T. R., and Villatoro, D. (2013). (Social) Norms and Agent-Based Simulation. In Ossowski, S., editor, *Agreement Technologies*, pages 181–189. Springer Netherlands, Dordrecht.

AnyLogic (2002). http://www.xjtek.com/, Last Accessed: Jan 2014.

Appel, K. and Haken, W. (1977). Every planar map is four colorable. Part I: Discharging. *Illinois Journal of Mathematics*, 21(3):429–490.

Arnett, J. J. (2007). The myth of peer influence in adolescent smoking initiation. *Health education & behavior : the official publication of the Society for Public Health Education*, 34(4):594–607.

Ash.org (2013). Smoking statistics: who smokes and how much, http://ash.org.uk/files/documents/ASH_106.pdf, Last Accessed: Jan 2014.

Ashraf, Q., Gershman, B., and Howitt, P. (2011). Banks, market organization, and macroeconomic performance: an agent-based computational analysis, http://www.nber.org/papers/w17102, Last Accessed: Jan 2014 .

Asomaning, K., Miller, D. P., Liu, G., Wain, J. C., Lynch, T. J., Su, L., and Christiani, D. C. (2008). Second hand smoke, age of exposure and lung cancer risk. *Lung cancer*, 61(1):13–20.

Atkins, B. (2003). *More than a game: The computer game as fictional form*. Manchester University Press, New York.

Au, G. and Paul, R. J. (1996). Visual interactive modelling: A pictorial simulation specification system. *European journal of operational research*, 91(1):14–26.

Audrey, S., Cordall, K., Moore, L., Cohen, D., and Campbell, R. (2004). The development and implementation of a peer-led intervention to prevent smoking among secondary school students using their established social networks. *Health Education Journal*, 63(3):266–284.

Audrey, S., Holliday, J., and Campbell, R. (2006a). It's good to talk: adolescent perspectives of an informal, peer-led intervention to reduce smoking. *Social science & medicine (1982)*, 63(2):320–34.

Audrey, S., Holliday, J., and Campbell, R. (2008). Commitment and compatibility: Teachers' perspectives on the implementation of an effective school-based, peer-led smoking intervention. *Health Education Journal*, 67(2):74–90.

Audrey, S., Holliday, J., Parry-Langdon, N., and Campbell, R. (2006b). Meeting the challenges of implementing process evaluation within randomized controlled trials: the example of ASSIST (A Stop Smoking in Schools Trial). *Health education research*, 21(3):366–77.

Avrachenkov, K. and Litvak, N. (2004). Decomposition of the google pagerank and optimal linking strategy, http://hal.archives-ouvertes.fr/docs/00/07/14/82/PDF/RR-5101.pdf, Last Accessed: Feb 2014.

Axtell, R. (2000). Why agents? On the varied motivations for agent computing in the social sciences, http://www.brookings.edu/es/dynamics/papers/agents/agents.pdf, Last Accessed: May 2014.

Bakker, R. M., Raab, J., and Milward, H. B. (2012). A preliminary theory of dark network resilience. *Journal of policy analysis and management*, 31(1):33–62.

Bala, V. and Goyal, S. (2000). A noncooperative model of network formation. *Econometrica*, 68(5):1181–1229.

Balci, O. (1994). Validation, verification, and testing techniques throughout the life cycle of a simulation study. *Annals of operations research*, 53(1):121–173.

Balci, O. and Ormsby, W. F. (2007). Conceptual modelling for designing large-scale simulations. *Journal of Simulation*, 1(3):175–186.

Bankes, S. C. (2002). Agent-based modeling: a revolution? *Proceedings of the National Academy of Sciences of the United States of America*, 99(Suppl 3):7199–7200.

Banks, J. (1998). *Handbook of Simulation*. John Wiley & Sons, Inc., Hoboken.

Bannister, M., Eppstein, D., Goodrich, M., and Trott, L. (2013). Force-Directed graph drawing using social gravity and scaling. In *Graph Drawing*, pages 414–425. Springer, Berlin.

Barabási, A. L. (2009). Scale-free networks: a decade and beyond. *Science*, 325(5939):412–3.

Barabási, A. L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

Barabási, A. L., Albert, R., and Jeong, H. (1999). Mean-field theory for scale-free random networks. *Physica A: Statistical Mechanics and its Applications*, 272(1-2):173–187.

Barabási, A. L., Albert, R., and Jeong, H. (2000). Scale-free characteristics of random networks: the topology of the world-wide web. *Physica A: Statistical Mechanics and its Applications*, 281(1-4):69–77.

Barabási, A. L. and Bonabeau, E. (2003). Scale-free networks. *Scientific American*, 288(5):50–59.

Barabási, A. L. and Frangos, J. (2003). *Linked: The new science of networks*, volume 71. Perseus Books Group, Cambridge, USA.

Barabâsi, A. L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., and Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3):590–614.

Barnoya, J. and Glantz, S. A. (2005). Cardiovascular effects of secondhand smoke: nearly as large as smoking. *Circulation*, 111(20):2684–98.

Bartecchi, C. E., MacKenzie, T. D., and Schrier, R. W. (1994). The Human Costs of Tobacco Use. *New England Journal of Medicine*, 330(13):907–914.

Bass, F. M. (1969). A New Product Growth for Model Consumer Durables. *Management Science*, 15(5):215–227.

Bassett, D. S. and Bullmore, E. (2006). Small-world brain networks. *The Neuroscientist*, 12(6):512–523.

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks, http://gephi.org/publications/gephi-bastian-feb09.pdf, Last Accessed: May 2014.

Bauch, C. T. and Bhattacharyya, S. (2012). Evolutionary game theory and social learning can determine how vaccine scares unfold. *PLoS computational biology*, 8(4):e1002452.

Bauman, K. E., Foshee, V. A., Linzer, M. A., and Koch, G. G. (1990). Effect of parental smoking classification on the association between parental and adolescent smoking. *Addictive behaviors*, 15(5):413–422.

Baur, M. and Brandes, U. (2005). Crossing reduction in circular layouts. In *Graph-Theoretic Concepts in Computer Science*, pages 332–343. Springer, Berlin.

Bavelas, A. (1950). Communication patterns in task oriented groups. *The Journal of the Acoustical Society of America*, 22(6):725–730.

Bearman, P. S. and Moody, J. (2004). Suicide and friendships among American adolescents. *American journal of public health*, 94(1):89–95.

Becker, M. Y. and Rojas, I. (2001). A graph layout algorithm for drawing metabolic pathways. *Bioinformatics*, 17(5):461–467.

Beeman, D. R. (1981). A public execution of the Peter principle. *Business Horizons*, 24(6):48–50.

Belk, R. W., Tian, K., and Paavola, H. (2010). Consuming cool: Behind the unemotional mask.

Research in Consumer Behavior, 12:183–208.

Bell, P. C. (1991). Visual interactive modelling: The past, the present, and the prospects. *European Journal of Operational Research*, 54(3):274–286.

Berman, E. M., West, J. P., and Richter Jr, M. N. (2002). Workplace relations: Friendship patterns and consequences (according to managers). *Public Administration Review*, 62(2):217–230.

Berndt, T. J. (1979). Developmental changes in conformity to peers and parents. *Developmental Psychology*, 15(6):608–616.

Bernstein, G. and O'Brien, K. (2013). Stochastic agent-based simulations of social networks. In *Proceedings of the 46th Annual Simulation Symposium, Article no: 5*.

Bianchini, M., Gori, M., and Scarselli, F. (2005). Inside pagerank. *ACM Transactions on Internet Technology*, 5(1):92–128.

Biglan, A., McConnell, S., Severson, H. H., Bavry, J., and Ary, D. (1984). A situational analysis of adolescent smoking. *Journal of Behavioral Medicine*, 7(1):109–114.

Bird, S. and Tapp, A. (2008). Social Marketing and the Meaning of Cool. *Social Marketing Quarterly*, 14(1):18–29.

Björk, S. and Juul, J. (2012). Zero-Player Games, http://www.jesperjuul.net/text/zeroplayergames/, Last Accessed: May 2014.

Bloor, M., Frankland, J., Langdon, N. P., Robinson, M., Allerston, S., Catherine, a., Cooper, L., Gibbs, L., Gibbs, N., Hamilton-Kirkwood, L., Jones, E., Smith, R. W., and Spragg, B. (1999). A controlled evaluation of an intensive, peer- led, schools-based, anti-smoking programme. *Health Education Journal*, 58(1):17–25.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D. (2006). Complex networks: Structure and dynamics. *Physics Reports*, 424(4-5):175–308.

Bollobas, B. (2013). *Modern Graph Theory*. Springer-Verlag, London.

Bonabeau, E. (2002). Agent-based modeling: methods and techniques for simulating human systems. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 99, pages 7280–7287.

Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *The Journal of Mathematical Sociology*, 2(1):113–120.

Borgs, C., Chayes, J., Daskalakis, C., and Roch, S. (2007). First to market is not everything: an analysis of preferential attachment with fitness. In *Proceedings of the thirty-ninth annual ACM*

symposium on Theory of computing, pages 135–144. ACM.

Borshchev, A. and Filippov, A. (2004). From system dynamics and discrete event to practical agent based modeling: reasons, techniques, tools. In *Proceedings of the 22nd international conference of the system dynamics society*, page 22.

Boss, L. (1997). Epidemic hysteria: a review of the published literature. *Epidemiologic Reviews*, 19(2):233–43.

Bourdieu, P. (1986). The Forms of Capital. In Richardson, J., editor, *Handbook of Theory and Research for the Sociology of Education*, pages 241–258. Greenwood Press, New York.

Bowler, W. M. (2006). Organizational Goals Versus the Dominant Coalition: A Critical View of the Value of Organizational Citizenship Behavior. *Journal of Behavioral and Applied Management*, 7:258–273.

Box, G. E. P. and Draper, N. R. (1987). *Empirical model-building and response surfaces*. Wiley, Hoboken.

Boyce, W., Torsheim, T., Currie, C., and Zambon, A. (2006). The Family Affluence Scale as a Measure of National Wealth: Validation of an Adolescent Self-Report Measure. *Social Indicators Research*, 78(3):473–487.

Brailsford, S. C. and Hilton, N. A. (2001). A comparison of discrete event simulation and system dynamics for modelling health care systems, http://eprints.soton.ac.uk/35689/, Last Accessed: May 2014.

Brandes, U. (2001). Drawing on physical analogies. In *Drawing Graphs*, pages 71–86. Springer, Berlin.

Brantingham, P. J., Tita, G. E., Short, M. B., and Reid, S. E. (2012). The Ecology of Gang Territorial Boundaries. *Criminology*, 50(3):851–885.

Brass, D. J. (1985). Men's and women's networks: A study of interaction patterns and influence in an organization. *Academy of Management Journal*, 28(2):327–343.

Brent, D. A., Kerr, M., and Goldstein, C. (1989). An outbreak of suicide and suicidal behavior in a high school. *Journal of the American Academy of Child & Adolescent Psychiatry*, 28(6):918–924.

Brent, R. P. (2013). *Algorithms for Minimization Without Derivatives*. Courier Dover Publications, New York.

Breslau, L., Estrin, D., Fall, K., Floyd, S., Heidemann, J., Helmy, a., Huang, P., McCanne, S., and Varadhan, K. (2000). Advances in network simulation. *Computer*, 33(5):59–67.

Bressan, M. and Peserico, E. (2010). Choose the damping, choose the ranking? *Journal of Discrete Algorithms*, 8(2):199–213.

Briggle, A. (2008). Real friends: how the Internet can foster friendship. *Ethics and Information Technology*, 10(1):71–79.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer networks and ISDN systems*, 30(1):107–117.

Brooks, A. C. (2005). Does Social Capital Make You Generous? *Social Science Quarterly*, 86(1):1–15.

Brooks, R., Robinson, S., and Lewis, C. (2001). *Simulation and Inventory Control*. Palgrave Macmillan, Basingstoke.

Brown, B. B., Clasen, D. R., and Eicher, S. a. (1986). Perceptions of peer pressure, peer conformity dispositions, and self-reported behavior among adolescents. *Developmental Psychology*, 22(4):521–530.

Bruggeman, J. (2013). *Social Networks: An Introduction*. Routledge, Abingdon.

Bruvold, W. H. (1993). A meta-analysis of adolescent smoking prevention programs. *American journal of public health*, 83(6):872–80.

Bryan, K. (2006). The $25,000,000,000 eigenvector: The linear algebra behind Google. *SIAM review*, 48(3):569–581.

Bunn, D. W. and Oliveira, F. S. (2001). Agent-based simulation-an application to the new electricity trading arrangements of England and Wales. *IEEE Transactions on Evolutionary Computation*, 5(5):493–503.

Burk, W. J., Steglich, C. E. G., and Snijders, T. A. B. (2007). Beyond dyadic interdependence: Actor-oriented models for co-evolving social networks and individual behaviors. *International Journal of Behavioral Development*, 31(4):397–404.

Burt, R. S. (1992). *Structural Holes: The Social Structure of Competition*. Harvard University Press, Cambridge, MA, USA.

Burt, R. S. (1997). The contingent value of social capital. *Administrative science quarterly*, 42(2):339–365.

Burt, R. S. (2000). The network structure of social capital. *Research in Organizational Behavior*, 22:345–423.

Byrne, J., Heavey, C., and Byrne, P. (2010). A review of Web-based simulation and supporting

tools. *Simulation Modelling Practice and Theory*, 18(3):253–276.

Cacioppo, J. T., Cacioppo, S., Gonzaga, G. C., Ogburn, E. L., and VanderWeele, T. J. (2013). Marital satisfaction and break-ups differ across on-line and off-line meeting venues. In *Proceedings of the National Academy of Sciences of the United States of America*, volume 110, pages 10135–40.

Cahn, R. (2001). Computer Simulation. In *The Coming of Materials Science, Volume 5*. Pergamon, Oxford.

Campbell, R., Starkey, F., Holliday, J., Audrey, S., Bloor, M., Parry-Langdon, N., Hughes, R., and Moore, L. (2008). An informal school-based peer-led intervention for smoking prevention in adolescence (ASSIST): a cluster randomised trial. *Lancet*, 371(9624):1595–602.

Campbell-Meiklejohn, D. and Bach, D. (2010). How the opinion of others affects our valuation of objects. *Current Biology*, 20(13):1165–1170.

Carley, K. (1991). A theory of group stability. *American Sociological Review*, 56(3):331–354.

Carlsen, K. H. and Carlsen, K. C. L. (2008). Respiratory effects of tobacco smoking on infants and young children. *Paediatric respiratory reviews*, 9(1):11–9; quiz 19–20.

Carrington, P. (2005). *Models and methods in social network analysis*. Cambridge University Press, New York.

Casey, M., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., and Slaney, M. (2008). Content-Based Music Information Retrieval: Current Directions and Future Challenges. *Proceedings of the IEEE*, 96(4):668–696.

Cassandras, C. G. and Lafortune, S. (2008). *Introduction to Discrete Event Systems*. Springer, New York.

Castle, C. and Crooks, A. (2006). Principles and concepts of agent-based modelling for developing geospatial simulations. *UCL Working Papers Series*, (110):60.

Center for Disease Control (1994). Suicide contagion and the reporting of suicide: Recommendations from a national workshop. *Morbidity and Mortality Weekly Review*, 43(RR-6):9–18.

Checkland, P. (1981). *Systems Thinking, Systems Practice*. Wiley, Hoboken.

Chen, E. (2012). Edge Prediction in a Social Graph: My Solution to Facebook's User Recommendation Contest on Kaggle, http://blog.echen.me/2012/07/31/edge-prediction-in-a-social-graph-my-solution-to-facebooks-user-recommendation-contest-on-kaggle, Last Accessed: Jan 2014 .

Chen, W., Teng, S., Wang, Y., and Zhou, Y. (2009). On the a-Sensitivity of Nash Equilibria in

PageRank-Based Network Reputation Games. In *Proceedings of the Third International Workshop on Frontiers in Algorithmics*, page 73.

Chen, X. and Zhan, F. B. (2008). Agent-based modelling and simulation of urban evacuation: relative effectiveness of simultaneous and staged evacuation strategies. *Journal of the Operational Research Society*, 59(1):25–33.

Cheng, A. and Friedman, E. (2006). Manipulability of PageRank under sybil strategies. In *First Workshop on the Economics of Networked Systemson*, pages 1–8.

Christakis, N. A. and Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. *The New England journal of medicine*, 357(4):370–9.

Christakis, N. A. and Fowler, J. H. (2008). The collective dynamics of smoking in a large social network. *The New England journal of medicine*, 358(21):2249–58.

Christakis, N. A. and Fowler, J. H. (2010a). Collective Dynamics of Smoking Behavior in a Large Social Network. *New England Journal of Medicine*, 358(21):2249–2258.

Christakis, N. A. and Fowler, J. H. (2010b). *Connected: The Amazing Power of Social Networks and How They Shape Our Lives*. HarperPress, London.

Cialdini, R. B. and Goldstein, N. J. (2004). Social influence: compliance and conformity. *Annual review of psychology*, 55(1974):591–621.

Clark, M. L. (1992). Frienship Similarity During Early Adolescence: Gender and Racial Patterns. *The Journal of Psychology*, 126(4):393–405.

Clayton, S. (1991). Gender Differences in Psychosocial Determinants of Adolescent Smoking. *Journal of School Health*, 61(3):115–120.

Cleverdon, C. W. (1972). On the Inverse Relationship of Recall and Precision. *Journal of Documentation*, 28(3):195–201.

Cocking, D. and Matthews, S. (2000). Unreal friends. *Ethics and Information Technology*, 2(4):223–231.

Cohen, J. M. (1977). Sources of Peer Group Homogeneity. *Sociology of Education*, 50(4):227–241.

Cohen-Charash, Y. and Mueller, J. S. (2007). Does perceived unfairness exacerbate or mitigate interpersonal counterproductive work behaviors related to envy? *Journal of Applied Psychology*, 92(3):666–680.

Correa, P., Fontham, E., Williams Pickle, L., Lin, Y., and Haenszel, W. (1983). Passive Smoking and Lung Cancer. *The Lancet*, 322(8350):595–597.

Costa, L. and Rodrigues, F. (2007). Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1):167–242.

Couclelis, H. (2002). Modeling frameworks, paradigms, and approaches. In *Geographic Information Systems and Environmental Modelling*, pages 1–15. Prentice Hall, London.

Croghan, E., Aveyard, P., Griffin, C., and Cheng, K. K. (2003). The importance of social sources of cigarettes to school students. *Tobacco control*, 12(1):67–73.

Cui, G., Li, M., Wang, Z., and Ren, J. (2014). Analysis and evaluation of incentive mechanisms in P2P networks: a spatial evolutionary game theory perspective. *Concurrency and Computation: Practice and Experience*, Pre-Print.

Cunningham, S. and Vaquera, E. (2012). Is there evidence that friends influence body weight? A systematic review of empirical research. *Social Science & Medicine*, 75(7):1175–1183.

Cuomo, A., Rak, M., and Villano, U. (2012). Process-oriented discrete-event simulation in java with continuations-quantitative performance evaluation. In *SIMULTECH*, pages 87–96.

Cyert, R. M. and March, J. G. (1992). *A Behavioural Theory of the Firm*. Blackwell Publishers, Oxford.

Darby, S. and Pike, M. (1988). Lung cancer and passive smoking: predicted effects from a mathematical model for cigarette smoking and lung cancer. *British journal of cancer*, 58(6):825–831.

Dastkhan, H. and Owlia, M. S. (2014). What are the right policies for electricity supply in Middle East? A regional dynamic integrated electricity model for the province of Yazd in Iran. *Renewable and Sustainable Energy Reviews*, 33:254–267.

Davidsson, P. (2001). Multi agent based simulation: beyond social simulation. In *Multi-Agent-Based Simulation*, pages 97–107.

Davis, R. M. (2000). Moving tobacco control beyond "the tipping point". *BMJ (Clinical research ed.)*, 321(7257):309–10.

de Kerchove, C., Ninove, L., and van Dooren, P. (2008). Maximizing PageRank via outlinks. *Linear Algebra and its Applications*, 429(5):1254–1276.

de Solla Price, D. J. (1965). Networks of Scientific Papers. *Science (New York, NY)*, 149(3683):510.

de Solla Price, D. J. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5):292–306.

Dechesne, F., Di Tosto, G., Dignum, V., and Dignum, F. (2012). No smoking here: values, norms and culture in multi-agent systems. *Artificial Intelligence and Law*, 21(1):79–107.

Deutsch, M. and Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. *The Journal of Abnormal and Social Psychology*, 51(3):629–636.

Dickinson, D. L. and Villeval, M. C. (2012). Job Allocation Rules and Sorting Efficiency: Experimental Outcomes in a Peter Principle Environment. *Southern Economic Journal*, 78(3):842–859.

Dilger, A. (2003). Lazear's Stochastic Interpretation of the Peter Principle: An Empirical Examination based on NBA-Data, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=520922, Last Accessed: May 2014.

Ding, Y. (2011). Applying weighted PageRank to author citation networks. *Journal of the American Society for Information Science and Technology*, 62(2):236–245.

Ding, Y., Yan, E., Frazho, A., and Caverlee, J. (2009). PageRank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Technology*, 60(11):2229–2243.

Dishion, T., Andrews, D., and Crosby, L. (1995). Antisocial boys and their friends in early adolescence: Relationship characteristics, quality, and interactional process. *Child development*, 66(1):139–151.

Djanatliev, A., German, R., Kolominsky-Rabas, P., and Hofmann, B. M. (2012). Hybrid simulation with loosely coupled system dynamics and agent-based models for prospective health technology assessments. In *Proceedings of the 2012 Winter Simulation Conference, Berlin*.

Dodds, P. S., Muhamad, R., and Watts, D. J. (2003). An experimental study of search in global social networks. *Science (New York, N.Y.)*, 301(5634):827–829.

Dodds, P. S. and Watts, D. J. (2005). A generalized model of social and biological contagion. *Journal of theoretical biology*, 232(4):587–604.

Doebelin, E. (1998). *System Dynamics: Modeling, Analysis, Simulation, Design*. Marcell Dekker, New York.

Doll, R. (2000). How It Really Happened: Smoking and Lung Cancer. *American Journal of Respiratory and Critical Care Medicine*, 162(1):4–6.

Domingos, P. and Richardson, M. (2007). Markov Logic: A Unifying Framework for Statistical Relational Learning. In *Proceedings of the ICML-2004 Workshop On Statistical Relational Learning and its connections to other fields.*, volume 339, pages 49–54.

D'Onofrio, A., Cerrai, P., and Gandolfi, A. (2013). Combining Game Theory and Graph Theory to Model Interactions between Cells. In *New Challenges for Cancer Systems Biomedicine*, page 409. Springer, Milan.

Doran, J., Palmer, M., Gilbert, N., and Mellars, P. (1994). The EOS project: modelling Upper Palaeolithic social change. In *Simulating Societies: the computer simulation of social phenomena*.

Drogoul, A., Corbara, B., and Fresneau, D. (1995). MANTA: New experimental results on the emergence of (artificial) ant societies. In *Artificial Societies: the computer simulation of social life*, pages 190–211.

Drogoul, A., Vanbergue, D., and Meurisse, T. (2003). Multi-agent based simulation: Where are the agents? In *Multi-agent-based simulation II*, pages 1–15.

DuBois, D. and Hirsch, B. (1990). School and neighborhood friendship patterns of Blacks and Whites in early adolescence. *Child development*, 61(2):524–536.

Dunbar, R. I. M. (1995). Neocortex size and group size in primates: a test of the hypothesis. *Journal of Human Evolution*, 28(3):287–296.

Dunbar, R. I. M. (1998). The social brain hypothesis. *Evolutionary Anthropology: Issues, News, and Reviews*, 6(5):178–190.

Dunbar, R. I. M. (2012). Social cognition on the Internet: testing constraints on social network size. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 367(1599):2192–201.

Dunlavy, D. M., Kolda, T. G., and Acar, E. (2011). Temporal Link Prediction Using Matrix and Tensor Factorizations. *ACM Transactions on Knowledge Discovery from Data*, 5(2):10.

Eades, P. (1984). A Heuristic for Graph Drawing. *Congressus Numerantium*, 42:149 – 160.

El-Sayed, A. M., Scarborough, P., Seemann, L., and Galea, S. (2012). Social network analysis and agent-based modeling in social epidemiology. *Epidemiologic perspectives & innovations : EP+I*, 9(1):1.

Elwert, F. and Christakis, N. A. (2006). Widowhood and Race. *American Sociological Review*, 71(1):16–41.

EMarketer (2013). Worldwide Social Network Users: 2013 Forecast and Comparative Estimates, http://www.emarketer.com, Last Accessed: May 2014.

Emery, S., Gilpin, E. A., White, M. M., and Pierce, J. P. (1999). How adolescents get their cigarettes: implications for policies on access and price. *Journal of the National Cancer Institute*, 91(2):184–6.

Ennett, S. T. and Bauman, K. E. (1994). The contribution of influence and selection to adolescent peer group homogeneity: The case of adolescent cigarette smoking. *Journal of Personality and*

Social Psychology, 67(4):653–663.

Epstein, J. M. (1996). *Growing Artificial Societies: Social Science from the Bottom Up*. Brookings Institution Press, Washington D.C.

Erdos, P. and Renyi, A. (1959). On random graphs. *Publicationes mathematicae*, 6:290–7.

Eubank, S., Guclu, H., Kumar, V. S. A., Marathe, M. V., Srinivasan, A., Toroczkai, Z., and Wang, N. (2004). Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(May):180–184.

Euler, L. (1736). Solutio problematis ad geometriarn situs pertinentis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae*, 8(128-140).

Evenden, D., Harper, P. R., Brailsford, S. C., and Harindra, V. (2005a). Improving the cost-effectiveness of Chlamydia screening with targeted screening strategies. *Journal of the Operational Research Society*, 57(12):1400–1412.

Evenden, D., Harper, P. R., Brailsford, S. C., and Harindra, V. (2005b). System Dynamics modeling of Chlamydia infection for screening intervention planning and cost-benefit estimation. *IMA Journal of Management Mathematics*, 16(3):265–279.

Ezzati, M. and Lopez, A. D. (2003). Estimates of global mortality attributable to smoking in 2000. *Lancet*, 362(9387):847–52.

Facebook (2013). www.facebook.com, Last Accessed: Feb 2014.

Fairburn, J. A. and Malcomson, J. M. (2001). Performance, Promotion, and the Peter Principle. *Review of Economic Studies*, 68(1):45–66.

Faloutsos, M., Faloutsos, P., and Faloutsos, C. (1999). On power-law relationships of the internet topology. *Applications, technologies, architectures, and protocols for computer communication*, 29(4):251–262.

Farkas, A., Distefan, J., and Choi, W. (1999). Does parental smoking cessation discourage adolescent smoking? *Preventive Medicine*, 28(3):213–218.

Farrell, S., Campbell, C., and Myagmar, S. (2005). Relescope: an experiment in accelerating relationships. In *Conference on Human Factors in Computing Systems*, pages 2–7.

Faw, L. (2013). Is Apple's iPhone No Longer Cool To Teens?, http://www.forbes.com, Last Accessed: May 2014.

Feld, S. (1991). Why your friends have more friends than you do. *American Journal of Sociology*, 96(6):1464–1477.

Feld, S. L. (1981). The focused organization of social ties. *American journal of sociology*, 86(5):1015–1035.

Ferber, J. (1999). *Multi-agent systems: an introduction to distributed artificial intelligence*. Harlow: Addison Wesley Longman.

Fergusson, D. M., Lynskey, M. T., and Horwood, L. J. (1995). The role of peer affiliations, social, family and individual factors in continuities in cigarette smoking between childhood and adolescence. *Addiction*, 90(5):647–59.

Fernández, P. and Madrid, G. (2007). Google's secret and Linear Algebra. *Newsletter of the European Mathematical Society*, 63(March):10–15.

Ferrari, M. J., Grais, R. F., Bharti, N., Conlan, A. J. K., Bjørnstad, O. N., Wolfson, L. J., Guerin, P. J., Djibo, A., and Grenfell, B. T. (2008). The dynamics of measles in sub-Saharan Africa. *Nature*, 451(7179):679–84.

Festinger, L. (1949). The Analysis of Sociograms using Matrix Algebra. *Human Relations*, 2:153–158.

Fetta, A. G., Harper, P. R., Knight, V. A., Vieira, I. T., and Williams, J. E. (2012). On the Peter Principle: An agent based investigation into the consequential effects of social networks and behavioural factors. *Statistical Mechanics and its Applications*, 391(9):2898–2910.

Fetta, A. G., Jones, M., Knight, V. A., and Williams, J. E. (2010). Comparing Discrete Event Simulation and System Dynamics - a detailed investigation, Dissertation, Cardiff University.

Finkenstädt, B. F. and Grenfell, B. T. (2000). Time series modelling of childhood diseases: a dynamical systems approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 49(2):187–205.

Fioretti, G. (2012). Agent-Based Simulation Models in Organization Science. *Organizational Research Methods*, 16(2):227–242.

Fioretti, G. and Lomi, A. (2010). Passing the buck in the garbage can model of organizational choice. *Computational and Mathematical Organization Theory*, 16(2):113–143.

Fisher, J. and Pry, R. (1972). A simple substitution model of technological change. *Technological forecasting and social change*, 3:75–88.

Fishman, G. (1978). *Principles of discrete event simulation.* John Wiley & Sons, Hoboken.

Fishman, G. and Kiviat, P. (1967). Digital computer simulation: statistical considerations. Technical report, RAND Corp, Santa Monica, CA.

Forbes (2013). Google on the Forbes World's Most Valuable Brands List, http://www.forbes.com/companies/google/, Last Accessed: May 2014.

Ford, A. (1997). System dynamics and the electric power industry. *System Dynamics Review*, 13(1):57–85.

Ford, D. K., Truxillo, D. M., and Bauer, T. N. (2009). Rejected But Still There: Shifting the focus in applicant reactions to the promotional context. *International Journal of Selection and Assessment*, 17(4):402–416.

Forrester, J. W. (1958). Industrial Dynamics: A Major Breakthrough for Decision Makers. *Harvard Business Review*, 36(4):37 – 66.

Forrester, J. W. (1961). *Industrial Dynamics*. MIT Press, New York.

Forrester, J. W. (1968). *Principles of Systems*. Wright-Allen Press, Cambridge, MA.

Forrester, J. W. (1994). System dynamics, systems thinking, and soft OR. *System Dynamics Review*, 10(2-3):245–256.

Forrester, J. W. (1995). The beginning of system dynamics. *McKinsey Quarterly*, 4:4–17.

Fowler, J. H. and Christakis, N. A. (2008). Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study. *BMJ: British medical journal*, 337:1–9.

Freeman, L. C. (1979). Centrality in social networks: conceptual clarification. *Social networks*, 1(1968):215–239.

Freeman, L. C., Sociometry, S., and Mar, N. (1977). A Set of Measures of Centrality Based on Betweenness A Set of Measures of Centrality Based on Betweenness. *American Sociological Association*, 40(1):35–41.

Friedkin, N. (1991). Theoretical foundations for centrality measures. *American journal of Sociology*, 96(6):1478–1504.

Friend, K., Carmona, M., Wilbur, P., and Levy, D. (2001). Youths' Social Sources of Cigarettes: The Limits of Youth-Access Policies. In *Contemporary Drug Problems 28*, volume 1, pages 507–526.

Fröding, B. and Peterson, M. (2012). Why virtual friendship is no genuine friendship. *Ethics and information technology*, 14(3):201–207.

Fronczak, A., Fronczak, P., and Holyst, J. (2004). Average path length in random networks. *Physical Review E*, 70(5):056110.

Fruchterman, T. and Reingold, E. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience*, 21(11):1129–1164.

Furnham, A. and Petrides, K. V. (2006). Deciding on promotions and redundancies: Promoting people by ability, experience, gender and motivation. *Journal of Managerial Psychology*, 21(1):6–18.

Gajer, P., Goodrich, M. T., and Kobourov, S. G. (2000). A Multi-dimensional Approach to Force-Directed Layouts of Large Graphs. pages 211–221.

Garcia, R. and Jager, W. (2011). From the Special Issue Editors: Agent Based Modeling of Innovation Diffusion. *Journal of Product Innovation Management*, 28(2):148–151.

Gardner, M. (1970). Mathematical games: The fantastic combinations of John Conway's new solitaire game "life". *Scientific American*, 223(4):120–123.

Garlaschelli, D. and Loffredo, M. (2004). Patterns of link reciprocity in directed networks. *Physical Review Letters*, 93(26):1–4.

Gass, S. (1983). Decision-aiding models: validation, assessment, and related issues for policy analysis. *Operations Research*, 31(4):603–631.

Gertner, J. (2003). 2003: The 3rd Annual Year in Ideas; Social Networks. *New York Times*.

Getoor, L. (2003). Link mining: a new data mining challenge. *ACM SIGKDD Explorations Newsletter*, 5(1):84–89.

Getoor, L. and Diehl, C. (2005). Link mining: a survey. *ACM SIGKDD Explorations Newsletter*, 7(2):3–12.

Gilbert, E. N. (1959). Random Graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144.

Gilbert, N. (2004). Agent-based social simulation: dealing with complexity. *The Complex Systems Network of Excellence*, 9(25):1–14.

Gilbert, N. (2007). *Agent-Based Models (Quantitative Applications in the Social Sciences)*. Sage Publications Ltd, London.

Gilbert, N. (2008). *Researching social life*. Sage Publications Ltd, London.

Gilbert, N. and Troitzsch, K. (2005). *Simulation for the social scientist*. Open University Press, Buckingham; 2nd edition.

Glantz, S. A. and Parmley, W. W. (1991). Passive smoking and heart disease. Epidemiology, physiology, and biochemistry. *Circulation*, 83(1):1–12.

Gleick, J. (1997). *Chaos: Making a new science*. Vintage, London.

Go, M. H., Tucker, J. S., Green, H. D., Pollard, M., and Kennedy, D. (2012). Social distance and homophily in adolescent smoking initiation. *Drug and alcohol dependence*, 124(3):347–54.

Goldenberg, A., Kubica, J., and Komarek, P. (2003). A comparison of statistical and machine learning algorithms on the task of link completion. In *KDD Workshop on Link Analysis for Detecting Complex Behavior*, page 8.

Goldenberg, J., Libai, B., Solomon, S., Jan, N., and Stau, D. (2000). Marketing percolation. *Physica A: Statistical Mechanics and its Applications*, 284(1):335–347.

Gonçalves, B., Perra, N., and Vespignani, A. (2011). Modeling users' activity on twitter networks: validation of Dunbar's number. *PloS one*, 6(8):e22656.

Gong, N. Z., Talwalkar, A., and Mackey, L. (2012). Jointly Predicting Links and Inferring Attributes using a Social-Attribute Network (SAN). In *SNA-KDD*.

Google+ (2013). https://plus.google.com, Last Accessed: Mar 2014.

Google (2013). https://www.google.com, Last Accessed: Apr 2014.

Google Scholar (2013). http://scholar.google.co.uk/, Last Accessed: Apr 2014.

Gorin, A., Wing, R., and Fava, J. (2008). Weight loss treatment influences untreated spouses and the home environment: evidence of a ripple effect. *International Journal of Obesity*, 32(11):1678–1684.

Gotts, N. M., Polhill, J. G., and Law, A. N. R. (2003). Agent-based simulation in the study of social dilemmas. *Artificial Intelligence Review*, 19(1):3–92.

Gould, M. S., Wallenstein, S., and Kleinman, M. (1990). Time-space clustering of teenage suicide. *American journal of epidemiology*, 131(1):71–8.

Granovetter, M. (1973). The strength of weak ties. *American journal of sociology*, 78(6):1360–1380.

Greenberg, J. and Colquitt, J. (2005). *Handbook of Organizational Justice*. Routledge.

Greene, J. A. (1999). Zero Tolerance: A Case Study of Police Policies and Practices in New York City. *Crime & Delinquency*, 45(2):171–187.

Grenfell, B. T., Bjørnstad, O. N., and Finkenstädt, B. F. (2002). Dynamics of measles epidemics: scaling noise, determinism, and predictability with the TSIR model. *Ecological Monographs*, 72(2):185–202.

Gribkovskaia, I., Halskau, O., and Laporte, G. (2007). The bridges of Königsberg - a historical perspective. *Networks*, 49(3):199–203.

Griffiths, J. D., Lawson, Z. F., and Williams, J. E. (2006). Modelling treatment effects in the HIV/AIDS epidemic. *The Journal of the Operational Research Society*, 57(12):1413–1424.

Grimm, V., Berger, U., Bastiansen, F., Eliassen, S., Ginot, V., Giske, J., Goss-Custard, J., Grand, T., Heinz, S. K., Huse, G., Huth, A., Jepsen, J. U., Jørgensen, C., Mooij, W. M., Müller, B., Pe'er, G., Piou, C., Railsback, S. F., Robbins, A. M., Robbins, M. M., Rossmanith, E., Rüger, N., Strand, E., Souissi, S., Stillman, R. A., Vabø, R., Visser, U., and DeAngelis, D. L. (2006). A standard protocol for describing individual-based and agent-based models. *Ecological Modelling*, 198(1-2):115–126.

Guare, J. (1992). *Six Degrees of Separation*. Dramatists Play Service, Inc.

Guimerà, R., Mossa, S., Turtschi, A., and Amaral, L. A. N. (2005). The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(22):7794–9.

Gyöngyi, Z. and Garcia-Molina, H. (2005). Link Spam Alliances. In *Proceedings of the 31st international conference on Very large data bases.*, pages 517–528.

Hammerstein, P. and Selten, R. (1994). Game theory and evolutionary biology. In *Handbook of game theory with economic applications 2*, volume 2, pages 929–993.

Han, X., Zhao, Z., Hadzibeganovic, T., and Wang, B. (2014). Epidemic spreading on hierarchical geographical networks with mobile agents. *Communications in Nonlinear Science and Numerical Simulation*, 19(5):1301–1312.

Harary, F. (1994). *Graph Theory*. Westview Press, Boulder.

Hasan, M. A., Chaoji, V., Salem, S., and Zaki, M. (2006). Link prediction using supervised learning. In *Proc. of SDM 06 workshop on Link Analysis, Counterterrorism and Security*.

Haveliwala, T. (2003). Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE Transactions on Knowledge and Data Engineering*, 15(4):784–796.

Hawe, G. I., Coates, G., Wilson, D. T., and Crouch, R. S. (2012). Agent-based simulation for large-scale emergency response. *ACM Computing Surveys*, 45(1):1–51.

Heath, B. (2010). *The History, Philosophy, and Practice of Agent-Based Modeling and the Development of the Conceptual Model for Simulation Diagram*. PhD thesis, Wright State University.

Hedges, L. V. and Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. Academic Press, San Diego.

Hegselmann, R. (2002). Opinion dynamics and bounded confidence: models, analysis and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3).

Heidemann, J., Klier, M., and Probst, F. (2010). Identifying Key Users in Online Social Networks: A PageRank Based Approach. In *Proceedings of the 31st International Conference on Information Systems, Saint Louis*, volume 4801, pages 1–22.

Helbing, D. and Balietti, S. (2012). Agent-Based Modeling. In Helbing, D., editor, *Social Self-Organization*, Understanding Complex Systems, pages 25–70. Springer, Berlin.

Helbing, D., Farkas, I., and Vicsek, T. (2000). Simulating dynamical features of escape panic. *Nature*, 407(6803):487–90.

Helleringer, S. and Kohler, H.-P. (2007). Sexual network structure and the spread of HIV in Africa: evidence from Likoma Island, Malawi. *AIDS (London, England)*, 21(17):2323–32.

Herlocker, J. L., Konstan, J. a., Terveen, L. G., and Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22(1):5–53.

Hernández Encinas, A., Hernández Encinas, L., Hoya White, S., Martín del Rey, A., and Rodríguez Sánchez, G. (2007). Simulation of forest fire fronts using cellular automata. *Advances in Engineering Software*, 38(6):372–378.

Hethcote, H. W. (1994). A thousand and one epidemic models. In *Frontiers in mathematical biology*, pages 504–515.

Hill, D. (1971). Peer group conformity in adolescent smoking and its relationship to affiliation and autonomy needs. *Australian Journal of Psychology*, 23(2):189–199.

Hill, R., McIntyre, G. A., Tighe, T. R., and Bullock, R. K. (2003). Some Experiments with Agent-Based Combat Models. *Military Operations Research*, 8(3):17–28.

Hill, R., Miller, J., and McIntyre, G. (2001). Applications of discrete event simulation modeling to military problems. In *Proceedings of the 2001 Winter Simulation Conference*, pages 780–788.

Hiscock, R., Bauld, L., Amos, A., Fidler, J. a., and Munafò, M. (2012). Socioeconomic status and smoking: a review. *Annals of the New York Academy of Sciences*, 1248:107–23.

Hodas, N., Kooti, F., and Lerman, K. (2013). Friendship Paradox Redux: Your Friends Are More Interesting Than You. *arXiv:1304.3480*.

Hofmann, R. T. and Carole (2004). Evaluation of free Java-libraries for social-scientific agent based simulation. *Journal of Artificial Societies and Social Simulation*, 7(1).

Holliday, J. (2006). *Identifying and Using Influential Young People for Informal Peer-Led Health*

Promotion. PhD thesis.

Holliday, J. C., Audrey, S., Moore, L., Parry-Langdon, N., and Campbell, R. (2009). High fidelity? How should we consider variations in the delivery of school-based health promotion interventions? *Health Education Journal*, 68(1):44–62.

Holliday, J. C., Rothwell, H. A., and Moore, L. A. R. (2010). The relative importance of different measures of peer smoking on adolescent smoking behavior: cross-sectional and longitudinal analyses of a large British cohort. *The Journal of adolescent health : official publication of the Society for Adolescent Medicine*, 47(1):58–66.

Hollingworth, W., Cohen, D., Hawkins, J., Hughes, R. A., Moore, L. A. R., Holliday, J. C., Audrey, S., Starkey, F., and Campbell, R. (2012). Reducing smoking in adolescents: cost-effectiveness results from the cluster randomized ASSIST (A Stop Smoking In Schools Trial). *Nicotine & tobacco research : official journal of the Society for Research on Nicotine and Tobacco*, 14(2):161–8.

Hollocks, B. W. (2006). Forty years of discrete-event simulation - a personal reflection. *Journal of the Operational Research Society*, 57(12):1383–1399.

Holme, P. (2005). Network reachability of real-world contact sequences. *Physical Review E*, 71(4):046119.

Homer, J., Milstein, B., Wile, K., Trogdon, J., Huang, P., Labarthe, D., and Orenstein, D. (2010). Simulating and evaluating local interventions to improve cardiovascular health. *Preventing chronic disease*, 7(1):A18.

Hopcroft, J. and Sheldon, D. (2008). Network reputation games.

Hornsey, M. J. and Jetten, J. (2004). The Individual Within the Group: Balancing the Need to Belong With the Need to Be Different. *Personality and Social Psychology Review*, 8(3):248–264.

Hortulanus, R., Machielse, A., and Meeuwesen, L. (2005). *Social Isolation in Modern Society*. Routledge.

Huang, C. Y., Tsai, Y. S., and Wen, T. H. (2010). Simulations for epidemiology and public health education. *Journal of Simulation*, 4(1):68–80.

Huang, W., Cooke, K. L., and Castillo-Chavez, C. (1992). Stability and bifurcation for a multiple-group model for the dynamics of HIV/AIDS transmission. *SIAM Journal on Applied Mathematics*, 52(3):835–854.

Huang, W., Hong, S., and Eades, P. (2007). Effects of Sociogram Drawing Conventions and Edge Crossings in Social Network Visualization. *Journal of Graph Algorithms and Applications*,

11(2):397–429.

Huang, Z., Li, X., and Chen, H. (2005). Link prediction approach to collaborative filtering. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, pages 141–142, New York.

Huerre, S. (2010). Agent-based crowd simulation tool for theme park environments. In *23rd International Conference on Computer Animation & Social Agents*.

Huhns, M. and Singh, M. (1998). *Readings in agents*. Morgan Kaufmann.

Huisman, M. and Snijders, T. A. B. (2003). Statistical analysis of longitudinal network data with changing composition. *Sociological Methods & Research*, 32(2):253–287.

Hunter, J. E. and Shotland, R. L. (1974). Treating data collected by the "Small World" method as a Markov process. *Social Forces*, 52(3):321–332.

IBM (2011). IBM SPSS Statistics for Windows.

Igor, M., Filiposka, S., Gramatikov, S., Trajanov, D., and Kocarev, L. (2010). Game Theoretic Approach for Discovering Vulnerable Links in Complex Networks. *Novel Algorithms and Techniques in Telecommunications and Networking*, pages 211–216.

Improbable-Research (2010). www.improbable.com, Last Accessed: Apr 2014.

Instagram (2013). www.instagram.com, Last Accessed: Mar 2014.

Ioannou, S. (2003). Young people's accounts of smoking, exercising, eating and drinking alcohol: being cool or being unhealthy? *Critical Public Health*, 13(4):357–371.

Ipsen, I. C. F. and Selee, T. M. (2008). PageRank Computation, with Special Attention to Dangling Nodes. *SIAM Journal on Matrix Analysis and Applications*, 29(4):1281–1296.

Iragne, F., Nikolski, M., Mathieu, B., Auber, D., and Sherman, D. (2005). ProViz: protein interaction visualization and exploration. *Bioinformatics (Oxford, England)*, 21(2):272–4.

Jalali Naini, S. G., Aliahmadi, A. R., and Jafari-Eskandari, M. (2011). Designing a mixed performance measurement system for environmental supply chain management using evolutionary game theory and balanced scorecard: A case study of an auto industry supply chain. *Resources, Conservation and Recycling*, 55(6):593–603.

James, L., Choi, C. C., Ko, C. H. E., McNeil, P. K., Minton, M. K., Wright, M. A., and Kim, K. (2008). Organizational and psychological climate: A review of theory and research. *European Journal of Work and Organizational Psychology*, 17(1):5–32.

Javarone, M. and Armano, G. (2012). A Fitness Model for Epidemic Dynamics in Complex Networks. In *Signal Image Technology and Internet Based Systems*, pages 793–797.

Jiang, B. and Jia, T. (2011). Agent-based simulation of human movement shaped by the underlying street structure. *International Journal of Geographical Information Science*, 25(1):51–64.

Jivanda, T. (2013). Facebook 'dead and buried' as teens switch to Snapchat and Whatsapp. *The Independent*.

Jones, T. F., Craig, A. S., Hoy, D., Gunter, E. W., Ashley, D. L., Barr, D. B., Brock, J. W., and Schaffner, W. (2000). Mass psychogenic illness attributed to toxic exposure at a high school. *New England Journal of Medicine*, 342(2):96–100.

Juan, A. A., Faulin, J., Pérez-Bernabeu, E., and Domínguez, O. (2013). Simulation-Optimization Methods in Vehicle Routing Problems: A Literature Review and an Example. In *Modeling and Simulation in Engineering, Economics, and Management*, pages 115–124. Springer, Berlin.

Junchao, Z., Junjie, C., Song, J., and Zhao, R.-X. (2013). Monte Carlo Based Personalized PageRank on Dynamic Networks. *International Journal of Distributed Sensor Networks*, 2013(2):1–8.

Kadushin, C. (2012). *Understanding Social Networks: Theories, Concepts, and Findings*. Oxford University Press, New York.

Kamada, T. and Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31(1):7–15.

Kandel, D. (1978). Homophily, selection, and socialization in adolescent friendships. *American journal of Sociology*, 84(2):427–436.

Kandel, E. and Lazear, E. (1992). Peer pressure and partnerships. *Journal of political Economy*, 100(4):801–817.

Kane, J. (1970). Dynamics of the Peter principle. *Management Science*, 16(12):800–811.

Kashima, Y., Wilson, S., Lusher, D., Pearson, L. J., and Pearson, C. (2013). The acquisition of perceived descriptive norms as social category learning in social networks. *Social Networks*, 35(4):711–719.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43.

Katzman, B., Markowitz, S., and McGeary, K. (2007). An empirical investigation of the social market for cigarettes. *Health Economics*, 1039(February):1025–1039.

Keeling, M. J. and Rohani, P. (2008). *Modeling Infectious Diseases in Humans and Animals*.

Princeton University Press, NJ, USA.

Kelly, J. A., Murphy, D. A., Sikkema, K. J., Mcauliffe, T. L., Roffman, R. A., and Solomon, L. J. (1997). Randomised, controlled, community-level HIV-prevention intervention for sexual-risk behaviour among homosexual men in US cities. Community HIV Prevention Research Collaborative. *The Lancet*, 350(9090):1500–1505.

Kelly, J. A., St Lawrence, J. S., Stevenson, L. Y., Hauth, a. C., Kalichman, S. C., Diaz, Y. E., Brasfield, T. L., Koob, J. J., and Morgan, M. G. (1992). Community AIDS/HIV risk reduction: the effects of endorsements by popular people in three cities. *American journal of public health*, 82(11):1483–9.

Kempe, A. (1879). On the geographical problem of the four colours. *American journal of mathematics*, 2(3):193–200.

Kermack, W. O. and McKendrick, A. G. (1927). A Contribution to the Mathematical Theory of Epidemics. *Proceedings of the Royal Society London*, 115(772):700–721.

Kermack, W. O. and McKendrick, A. G. (1932). Contributions to the Mathematical Theory of Epidemics. II. The Problem of Endemicity. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 138(834):55–83.

Kiesling, E., Günther, M., Stummer, C., and Wakolbinger, L. M. (2011). Agent-based simulation of innovation diffusion: a review. *Central European Journal of Operations Research*, 20(2):183–230.

Kiesner, J., Kerr, M., and Stattin, H. (2004). "Very important persons" in adolescence: going beyond in-school, single friendships in the study of peer homophily. *Journal of adolescence*, 27(5):545–560.

Killen, M. (2007). Children's social and moral reasoning about exclusion. *Current Directions in Psychological Science*, 16(1):32–36.

Killen, M. and Stangor, C. (2001). Children's social reasoning about inclusion and exclusion in gender and race peer group contexts. *Child development*, 72(1):174–186.

Kite, S., Wood, C., Taylor, S. J. E., and Mustafee, N. (2011). SakerGrid: simulation experimentation using grid enabled simulation software. In *Proceedings of the 2011 Winter Simulation Conference*, pages 2283–2293.

Klausner, J. D., Wolf, W., Fischer-Ponce, L., Zolt, I., and Katz, M. H. (2000). Tracing a syphilis outbreak through cyberspace. *JAMA The Journal of the American Medical Association*, 284(4):447–449.

Kleijnen, J. P. (1995). Verification and validation of simulation models. *European Journal of*

Operational Research, 82(1):145–162.

Kleinberg, J. (1998). Authoritative sources in a hyperlinked environment. In *ACM-SIAM Synposium on Discrete Algorithms*, volume 46, pages 668–677.

Kleinberg, J. and Easley, D. (2010). *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, Cambridge, USA.

Kleinfeld, J. S. (2002). The small world problem. *Society*, 39(2):61–66.

Kluger, B. D. and McBride, M. E. (2011). Intraday trading patterns in an intelligent autonomous agent-based stock market. *Journal of Economic Behavior & Organization*, 79(3):226–245.

Kobus, K. (2003). Peers and adolescent smoking. *Addiction*, 98(Suppl 1):37–55.

Kornbluh, M. and Little, D. (1976). The nature of a computer simulation model. *Technological forecasting and social change*, 9(1-2):3–26.

Kosinski, M., Stillwell, D., and Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America*, 110(15):5802–5.

Kotiadis, K. and Robinson, S. (2008). Conceptual modelling: knowledge acquisition and model abstraction. In *Proceedings of the 40th Winter Simulation Conference*, pages 951–958.

Kotler, P. (2011). Reinventing Marketing to Manage the Environmental Imperative. *Journal of Marketing*, 75(4):132–135.

Krebs, V. E. (2002). Mapping Networks of Terrorist Terrorist Cells. *Networks*, 24(3):43–52.

Krebs, V. E. (2013). A Brief Introduction to Social Network Analysis by Orgnet, http://www.orgnet.com/sna.html, Last Accessed: May 2014.

Kretschmer, H. and Kretschmer, T. (2007). A new centrality measure for social network analysis applicable to bibliometric and webometric data. *Collnet Journal of Scientometrics and Information Management*, 1(1):1–7.

Kudo, H. and Dunbar, R. (2001). Neocortex size and social network size in primates. *Animal Behaviour*, 62(4):711–722.

Kunegis, J. (2007). *On the Spectral Evolution of Large Networks*. PhD thesis, University of Koblenz-Landau.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600.

Lakon, C. M. and Valente, T. W. (2012). Social integration in friendship networks: the synergy of network structure and peer influence in relation to cigarette smoking among high risk adolescents. *Social science & medicine*, 74(9):1407–17.

Lambiotte, R. and Ausloos, M. (2005). Uncovering collective listening habits and music genres in bipartite networks. *Physical Review E*, 72(6):066107.

Lambiotte, R., Blondel, V. D., de Kerchove, C., Huens, E., Prieur, C., Smoreda, Z., and Van Dooren, P. (2008). Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications*, 387(21):5317–5325.

Lane, D., Monefeldt, C., and Rosenhead, J. (2000). Looking in the wrong place for healthcare improvements: A system dynamics study of an accident and emergency department. *Journal of the operational Research Society*, 51(5):518–531.

Langton, C. (1997). *Artificial life: An overview*. MIT Press, Cambridge, USA.

Langville, A. and Meyer, C. (2004). Deeper Inside PageRank. *Internet Mathematics*, 1(3):335–380.

Lareau, A. (1987). Social class differences in family-school relationships: The importance of cultural capital. *Sociology of education*, 60(2):73–85.

Law, A. (2009). How to build valid and credible simulation models. In *Proceedings of the 2009 Winter Simulation Conference*, pages 24–33.

Law, A. and Kelton, W. D. (1999). *Simulation Modeling and Analysis (Industrial Engineering and Management Science Series)*. McGraw-Hill Science/Engineering/Math.

Lazear, E. R. (2004). The Peter Principle: a theory of decline. *Journal of Political Economy*, 112(February):141–163.

Leavitt, H. (1951). Some effects of certain communication patterns on group performance. *The Journal of Abnormal and Social Psychology*, 46(1):38–50.

Leskovec, J. and Horvitz, E. (2008). Planetary-scale views on a large instant-messaging network. *Proceedings of the 17th international conference on World Wide Web*, pages 915–924.

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet physics doklady*, 10(8):707–710.

Levinson, A. H., Campo, S., Gascoigne, J., Jolly, O., Zakharyan, A., and Tran, Z. V. (2007). Smoking, but not smokers: identity among college students who smoke cigarettes. *Nicotine & tobacco research : official journal of the Society for Research on Nicotine and Tobacco*, 9(8):845–52.

Levy, D. T. and Friend, K. (2002). A Simulation Model of Policies Directed at Treating Tobacco Use and Dependence. *Medical Decision Making*, 22(1):6–17.

Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., and Christakis, N. (2008). Tastes, ties, and time: A new social network dataset using Facebook.com. *Social Networks*, 30(4):330–342.

Lewis, T. G. (2011). *Network Science: Theory and Applications*. John Wiley & Sons, Hoboken.

Liben-Nowell, D. and Kleinberg, J. (2007). The link prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031.

Lichtenwalter, R., Lussier, J., and Chawla, N. (2010). New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 243–252.

Light, J. M., Rusby, J. C., Nies, K. M., and Snijders, T. A. B. (2013). Antisocial Behavior Trajectories and Social Victimization Within and Between School Years in Early Adolescence. *Journal of Research on Adolescence*.

Liljeros, F., Edling, C. R., Amaral, L. A. N., Stanley, H. E., and Å berg, Y. (2001). The web of human sexual contacts. *Nature*, 411(6840):907–908.

Lin, N. (1976). *Foundations of social research*. McGraw-Hill, New York.

Liu, Y., Liang, M., Zhou, Y., He, Y., Hao, Y., Song, M., Yu, C., Liu, H., Liu, Z., and Jiang, T. (2008). Disrupted small-world networks in schizophrenia. *Brain*, 131(4):945–961.

Lockwood, P. and Kunda, Z. (1997). Superstars and me: Predicting the impact of role models on the self. *Journal of Personality and Social Psychology*, 73(1):91–103.

Lockwood, P. and Kunda, Z. (1999). Increasing the salience of one's best selves can undermine inspiration by outstanding role models. *Journal of Personality and Social Psychology*, 76(2):214–228.

Lofgren, E. T. and Fefferman, N. H. (2007). The untapped potential of virtual game worlds to shed light on real world epidemics. *The Lancet infectious diseases*, 7(9):625–9.

Lopez-Rojas, E., Axelsson, S., and Gorton, D. (2013). RETSIM: A shoe store agent-based simulation for fraud detection. In *25th European Modeling and Simulation Symposium*, number c, pages 25–34.

Lospinoso, J. and Schweinberger, M. (2011). Assessing and accounting for time heterogeneity in stochastic actor oriented models. *Advances in data analysis*, 5(2):147–176.

Loyo, H. K., Batcher, C., Wile, K., Huang, P., Orenstein, D., and Milstein, B. (2013). From model

to action: using a system dynamics model of chronic disease risks to align community action. *Health promotion practice*, 14(1):53–61.

Lü, L., Medo, M., Yeung, C. H., Zhang, Y. C., Zhang, Z. K., and Zhou, T. (2012). Recommender systems. *Physics Reports*, 519(1):1–49.

Lü, L. and Zhou, T. (2011). Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150–1170.

Luce, R. D. and Perry, A. D. (1949). A method of matrix analysis of group structure. *Psychometrika*, 14(1):95–116.

Luce, R. D. and Raiffa, H. (1957). *Games and Decisions: Introduction and Critical Survey*. Courier Dover Publications, New York.

Lycos Search (2013). www.lycos.com, Last Accessed: Aug 2013.

Lynch, P. (2008). The origins of computer weather prediction and climate modeling. *Journal of Computational Physics*, 227(7):3431–3444.

Ma, N., Guan, J., and Zhao, Y. (2008). Bringing PageRank to the citation analysis. *Information Processing & Management*, 44(2):800–810.

Macal, C. M. and North, M. J. (2005). Tutorial on agent-based modeling and simulation. In *Proceedings of the 37th conference on Winter simulation*, pages 2–15.

Macy, M. W. and Willer, R. (2002). FROM FACTORS TO ACTORS : Computational Sociology and Agent-Based Modeling. *Annual Review of Sociology*, 28(1):143–166.

Maddala, G. S. (1983). *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge University Press, New York.

Maier, F. H. (1998). New product diffusion models in innovation management - a system dynamics perspective. *System Dynamics Review*, 14(4):285–308.

Malaga, R. (2008). Worst practices in search engine optimization. *Communications of the ACM*, 51(12):147–150.

Malcolm, G. (2000). *The tipping point: how little things can make a big difference*. Abacus; New Ed edition.

Malleson, N., Heppenstall, A., See, L., and Evans, A. (2013). Using an agent-based crime simulation to predict the effects of urban regeneration on individual household burglary risk. *Environment and Planning B: Planning and Design*, 40(3):405–426.

Mao, H. Y. (2006). The relationship between organizational level and workplace friendship. *The International Journal of Human Resource Management*, 17(10):1819–1833.

Mao, L. (2014). Modeling triple-diffusions of infectious diseases, information, and preventive behaviors through a metropolitan social network - An agent-based simulation. *Applied Geography*, 50:31–39.

Marchiori, M. (1997). The quest for correct information on the web: Hyper search engines. *Computer Networks and ISDN Systems*, 29(8-13):1225–1235.

Marsden, P. (1987). Core discussion networks of Americans. *American sociological review*, 52(1):122–131.

Marsili, M. (2004). The Rise and Fall of a Networked Society. *Proceedings of the National Academy of Sciences of the United States of America*, 101(6):1439–1442.

Martin, R. H. and Raffo, D. (2000). A model of the software development process using both continuous and discrete models. *Software Process: Improvement and Practice*, 5(2-3):147–157.

Maslow, A. (1943). A theory of human motivation. *Psychological review*, 50(4):370–396.

MASON (2012). cs.gmu.edu/˜eclab/projects/mason, Last Accessed: Mar 2014.

Matsuo, Y., Hasida, K., Tomobe, H., and Ishizuka, M. (2003). Mining social network of conference participants from the Web. *Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003)*, pages 190–193.

Mayhew, K. P., Flay, B. R., and Mott, J. A. (2000). Stages in the development of adolescent smoking. *Drug and alcohol dependence*, 59(Suppl 1):S61–81.

Maynard Smith, J. and Price, G. R. (1973). The logic of animal conflict. *Nature*, 246:15–18.

Mcalister, A., Perry, C., and Maccoby, N. (1979). Adolescent smoking: Onset and prevention. *Pediatrics*, 63(4):650–658.

McKenna, K. (2002). Relationship formation on the Internet: What's the big attraction? *Journal of social issues*, 58(1):9–31.

Mei, S., Sloot, P., Quax, R., Zhu, Y., and Wang, W. (2010a). Complex agent networks explaining the HIV epidemic among homosexual men in Amsterdam. *Mathematics and Computers in Simulation*, 80(5):1018–1030.

Mei, S., van de Vijver, D., Xuan, L., Zhu, Y., and Sloot, P. (2010b). Quantitatively evaluating interventions in the influenza A (H1N1) epidemic on China campus grounded on individual-based simulations. *Procedia Computer Science*, 1(1):1675–1682.

Mercken, L., Moore, L., Crone, M. R., De Vries, H., De Bourdeaudhuij, I., Lien, N., Fagiano, F., Vitória, P. D., and Van Lenthe, F. J. (2012a). The effectiveness of school-based smoking prevention interventions among low- and high-SES European teenagers. *Health education research*, 27(3):459–69.

Mercken, L., Sleddens, E. F. C., de Vries, H., and Steglich, C. E. G. (2013). Choosing adolescent smokers as friends: the role of parenting and parental smoking. *Journal of adolescence*, 36(2):383–92.

Mercken, L., Snijders, T. A., Steglich, C., and de Vries, H. (2009). Dynamics of adolescent friendship networks and smoking behavior: Social network analyses in six European countries. *Social Science & Medicine*, 69(10):1506–1514.

Mercken, L., Snijders, T. A. B., Steglich, C., Vertiainen, E., and de Vries, H. (2010). Smoking-based selection and influence in gender-segregated friendship networks: a social network analysis of adolescent smoking. *Addiction (Abingdon, England)*, 105(7):1280–9.

Mercken, L., Steglich, C., and Sinclair, P. (2012b). A longitudinal social network analysis of peer influence, peer selection, and smoking behavior among adolescents in British schools. *Health Psychology*, 31(4):450–459.

Mesch, G. and Talmud, I. (2006). The Quality of Online and Offline Relationships: The Role of Multiplexity and Duration of Social Relationships. *The Information Society*, 22(3):137–148.

Metro Reporters (2013). New version of London Underground map shows circles are the way forward. *Metro*.

Meyer, C. (2000). *Matrix Analysis and Applied Linear Algebra*. SIAM, ISBN: 0-89871-454-0.

Milgram, S. (1963). Behavioral Study of Obedience. *Journal of abnormal psychology*, 67(4):371–378.

Milgram, S. (1967). The small world problem. *Psychology Today*, 2(1):60–67.

Milgram, S., Mann, L., and Harter, S. (1965). The lost-letter technique: A tool of social research. *Public Opinion Quarterly*, 29(3):437–438.

Miller, D. (2013). Facebook's so uncool, but it's morphing into a different beast, http://theconversation.com/facebooks-so-uncool-but-its-morphing-into-a-different-beast-21548, Last Accessed: May 2014.

Miller, J. and Page, S. (2007). *Complex Adaptive Systems: An Introduction to Computational Models of Social Life*. Princeton University Press, NJ, USA.

Miller, J., Slim, A., and Volz, E. (2012). Edge-based compartmental modelling for infectious

disease spread. *Journal of The Royal Society Interface*, 9(70):890–906.

Miller, J. C. and Kiss, I. Z. (2014). Epidemic spread in networks: Existing methods and current challenges. *arXiv:1403.1957*.

Miller, J. C. and Volz, E. M. (2013). Incorporating Disease and Population Structure into Models of SIR Disease in Contact Networks. *PloS one*, 8(8):e69162.

Miller, J. H. and Page, S. E. (2004). The standing ovation problem. *Complexity*, 9(5):8–16.

Minar, N., Burkhart, R., Langton, C., and Askenazi, M. (1996). The Swarm Simulation System : A Toolkit for Building Multi-agent Simulations.

Mingers, J. and Rosenhead, J. (2001). *Rational analysis for a problematic world revisited*. John Wiley & Sons, Hoboken; 2nd Edition.

Mislove, A., Koppula, H. S., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. (2008). Growth of the flickr social network. In *In Proceedings of the first workshop on Online social networks*, pages 25–30, New York, New York, USA. ACM Press, New York.

Mitchell, M. (2009). *Complexity: A guided tour*. OUP, USA.

Mizuta, H. and Steiglitz, K. (2000). Agent-based simulation of dynamic online auctions. In *Winter Simulation Conference*, pages 1772–1777.

Mo, Y., Ren, B., Yang, W., and Shuai, J. (2014). The 3-dimensional cellular automata for HIV infection. *Physica A: Statistical Mechanics and its Applications*, 399:31–39.

Mohnen, S. M., Völker, B., Flap, H., Subramanian, S. V., and Groenewegen, P. P. (2013). You have to be there to enjoy it? Neighbourhood social capital and health. *European journal of public health*, 23(1):33–9.

Montes, A. and Villalbí, J. R. (2001). The price of cigarettes in the European Union. *Tobacco control*, 10(2):135–6.

Moody, J. and White, D. (2003). Structural cohesion and embeddedness: A hierarchical concept of social groups. *American Sociological Review*, 68(1):103–127.

Moore, T., Finley, P., and Linebarger, J. (2011). Extending Opinion Dynamics to Model Public Health Problems and Analyze Public Policy Interventions. *on Complex Systems*, 19(5):391–4.

Morecroft, J. and Robinson, S. (2005). Explaining puzzling dynamics: comparing the use of system dynamics and discrete-event simulation. In *Proceedings of the 23rd International Conference of the System Dynamics Society*, pages 1–32.

Morgeson, F. P. and Ryan, A. M. (2009). Reacting to Applicant Perspectives Research: What's next? *International Journal of Selection and Assessment*, 17(4):431–437.

Morrow, J. (1994). *Game theory for political scientists*. Princeton University Press, NJ, USA.

Munz, P., Hudea, I., Imad, J., and Smith, R. (2009). When zombies attack!: mathematical modelling of an outbreak of zombie infection. *Infectious Disease Modelling Research Progress*, 4.

Murray, J. D. (1989). *Mathematical Biology*. Springer-Verlag, New York.

Mustafee, N., Katsaliaki, K., and Taylor, S. J. E. (2010). Profiling Literature in Healthcare Simulation. *Simulation*, 86(8-9):543–558.

Mustafee, N. and Taylor, S. J. E. (2009). Speeding up simulation applications using WinGrid. *Concurrency and Computation: Practice and Experience*, 21(11):1504–1523.

Myerson, R. B. (1991). *Game theory: analysis of conflict*. Harvard University Press, Cambridge, USA.

Nagel, K. and Raschke, E. (1992). Self-organizing criticality in cloud formation? *Physica A: Statistical Mechanics and its Applications*, 182(4):519–531.

Najlis, R., Janssen, M., and Parker, D. (2001). Software tools and communication issues. In *Agent-Based Models of Land-Use and Land-Cover Change Workshop*, pages 17 – 30.

Nancarrow, C. and Nancarrow, P. (2007). Hunting for cool tribes. In *Consumer Tribes*, page 339. Routledge.

Nancarrow, C., Nancarrow, P., and Page, J. (2002). An analysis of the concept of cool and its marketing implications. *Journal of Consumer Behaviour*, 1(4):311–322.

Nance, R. and Sargent, R. (2002). Perspectives on the evolution of simulation. *Operations Research*, 50(1):161–172.

Nash, J. (1950). Equilibrium points in n-person games. *Proceedings of the national academy of sciences of the United States of America*, 36(1):48–49.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, 33(1):31–88.

Near, A. M., Blackman, K., Currie, L. M., and Levy, D. T. (2013). Sweden SimSmoke: the effect of tobacco control policies on smoking and snus prevalence and attributable deaths. *European journal of public health*, pages 1–8.

Negahban, A. and Smith, J. S. (2014). Simulation for manufacturing system design and operation: Literature review and analysis. *Journal of Manufacturing Systems*, pages 1–21.

Newman, I. and Ward, J. (1989). The influence of parental attitude and behavior on early adolescent cigarette smoking. *Journal of School Health*, 59(4):150–152.

Newman, M. E. J. (2001). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E*, 64(1):016132.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM review*, 45(2):167–256.

Newman, M. E. J., Forrest, S., and Balthrop, J. (2002a). Email networks and the spread of computer viruses. *Physical Review E*, 66(3):035101.

Newman, M. E. J. and Park, J. (2003). Why social networks are different from other types of networks. *Physical Review E*, 68(3):036122.

Newman, M. E. J., Watts, D. J., and Strogatz, S. H. (2002b). Random graph models of social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(Suppl 1):2566–2572.

Ng, T. W., Turinici, G., and Danchin, A. (2003). A double epidemic model for the SARS propagation. *BMC infectious diseases*, 3(19):1–16.

Niazi, M. and Hussain, A. (2011). Agent-based computing from multi-agent systems to agent-based models: a visual survey. *Scientometrics*, 89(2):479–499.

Nicholson, C. F. and Kaiser, H. M. (2008). Dynamic market impacts of generic dairy advertising. *Journal of Business Research*, 61(11):1125–1135.

Nichter, M., Vuckovic, N., Quintero, G., and Ritenbaugh, C. (1997). Smoking experimentation and initiation among adolescent girls: qualitative and quantitative findings. *Tobacco control*, 6(4):285–295.

Niemeijer, R. (1973). Some applications of the notion of density. In Boissevain, J. and Mitchell, J., editors, *Network Analysis Studies in Human Interaction*. Mouton, Paris.

Nikolai, C. and Madey, G. (2009). Tools of the Trade: A Survey of Various Agent Based Modeling Platforms. *Journal of Artificial Societies and Social Simulation*, 12(2).

Norman, N. M. and Tedeschi, J. T. (1989). Self-Presentation, Reasoned Action, and Adolescents' Decisions to Smoke Cigarettes. *Journal of Applied Social Psychology*, 19(7):543–558.

North, M., Howe, T., Collier, N., and Vos, R. (2005). The repast simphony runtime system. In

Proceedings of the Agent 2005 Conference on Generative Social Processes, Models, and Mechanisms, pages 151–158.

Nouman, A., Anagnostou, A., and Taylor, S. J. E. (2013). Developing a Distributed Agent-Based and DES Simulation Using poRTIco and Repast. In *2013 IEEE/ACM 17th International Symposium on Distributed Simulation and Real Time Applications*, pages 97–104.

Nowak, M. and May, R. M. (1993). AIDS pathogenesis: mathematical models of HIV and SIV infections. *Aids*, 7:S3–S18.

Nyborg, K. and Rege, M. (2003). On social norms: the evolution of considerate smoking behavior. *Journal of Economic Behavior & Organization*, 52(3):323–340.

Odda, T. (1979). On properties of a well-known graph or what is your ramsey number? *Annals of the New York Academy of Sciences*, 328:166–172.

O'Madadhain, J., Hutchins, J., and Smyth, P. (2005). Prediction and ranking algorithms for event-based network data. *ACM SIGKDD Explorations Newsletter*, 7(2):23–30.

ONS (2013a). Opinions and Lifestyle Survey, 2012. *Report*, (September).

ONS (2013b). Opinions and Lifestyle Survey, Smoking Habits Amongst Adults, 2012. *Report*.

ONS (2013c). Statistical Bulletin Labour Market Statistics, September 2013. *Report*, (April).

Opsahl, T. (2013). Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. *Social Networks*, 35(2):159–167.

Opsahl, T. and Panzarasa, P. (2009). Clustering in weighted networks. *Social Networks*, 31(2):155–163.

Ormerod, P. (2008). Emergent scale-free social networks in history: burning and the rise of English Protestantism. *Cultural Science*, 1(1):1–29.

Ormerod, P. and Wiltshire, G. (2009). 'Binge' drinking in the UK: a social network phenomenon. *Mind & Society*, 8(June):1–13.

Osgood, D. W., Feinberg, M. E., Wallace, L. N., and Moody, J. (2013). Friendship group position and substance use. *Addictive behaviors*, 39(5):923–933.

Osterman, K. F. (2000). Students' Need for Belonging in the School Community. *Review of Educational Research*, 70(3):323–367.

Ostrom, E. (1998). A Behavioral Approach to the Rational Choice Theory of Collective Action: Presidential Address, American Political Science Association, 1997. *The American Political*

Science Review, 92(1):1–22.

O'Sullivan, D. (2004). Complexity science and human geography. *Transactions of the Institute of British Geographers*, 29(3):282–295.

Otto, P. (2008). A system dynamics model as a decision aid in evaluating and communicating complex market entry strategies. *Journal of Business Research*, 61(11):1173–1181.

Outkin, A. V. (2012). An Agent-based Model of the Nasdaq Stock Market: Historic Validation and Future Directions. In *CSSSA Annual Conference*, pages 18–21.

Oxford Dictionary (2010). *Oxford English Dictionary*. OUP, Oxford.

Page, L. and Brin, S. (1999). The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Page, R. (2000). Brief history of flight simulation. In *SimTecT 2000 Proceedings*, pages 1–11.

Palmer, R. G., Arthur, W. B., Holland, J. H., LeBaron, B., and Tayler, P. (1994). Artificial economic life: a simple model of a stockmarket. *Physica D: Nonlinear Phenomena*, 75(1):264–274.

Pan, X., Han, C. S., Dauber, K., and Law, K. H. (2007). A multi-agent based framework for the simulation of human and social behaviors during emergency evacuations. *Ai & Society*, 22(2):113–132.

Parker, J. G. and Asher, S. R. (1987). Peer relations and later personal adjustment: are low-accepted children at risk? *Psychological bulletin*, 102(3):357–89.

Parker, J. G. and Asher, S. R. (1993). Friendship and Frendship Quality in Middle Childhood: Links with Peer Group Acceptance and Feelings of Lonliness and Social Dissatisfaction. *Developmental Psychology*, 29(4):611–621.

Parks-Yancy, R. (2006). The Effects of Social Group Membership and Social Capital Resources on Careers. *Journal of Black Studies*, 36(4):515–545.

Parunak, H., Savit, R., and Riolo, R. (1998). Agent-based modeling vs. equation-based modeling: A case study and users' guide. In *Multi-Agent Systems and Agent-Based Simulation*, pages 10–25.

Paxton, S. and Schutz, H. (1999). Friendship clique and peer influences on body image concerns, dietary restraint, extreme weight-loss behaviors, and binge eating in adolescent girls. *Journal of abnormal Psychology*, 108(2):255–266.

Pearson, M. and Michell, L. (2000). Smoke rings: social network analysis of friendship groups, smoking and drug-taking. *Drugs-education Prevention and Policy*, 7(1):21–37.

Pearson, M., Sieglich, C., and Snijders, T. (2006). Homophily and assimilation among sport-active adolescent substance users. *Connections*, 27(1):47–63.

Peter, L. J. (1972). *The Peter Prescription*. William Morrow & Company, Inc., New York.

Peter, L. J. and Hull, R. (1969). *The Peter Principle: Why Things Always Go Wrong*. W. Morrow.

Petersen, J. (1898). Sur le théorème de Tait. *L'Intermédiaire des Mathématiciens*, 5:225–227.

Phuoc, N., Kim, S., Lee, H., and Kim, H. (2009). PageRank vs. Katz Status Index, a Theoretical Approach. In *Proceedings of the 2009 Fourth International Conference on Computer Sciences and Convergence Information Technology*, pages 1276–1279.

Pidd, M. (2003). *Tools for Thinking: Modelling in Management Science.* John Wiley & Sons, Chichester; 2nd edition.

Pidd, M. (2004). *Computer Simulation in Management Science*. John Wiley & Sons, Chichester.

Pillai, S., Suel, T., and Cha, S. (2005). The Perron-Frobenius theorem: some of its applications. *Signal Processing Magazine, IEEE*, 22(2):62–75.

Pinski, G. and Narin, F. (1976). Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Information Processing & Management*, 12(5):297–312.

Pluchino, A., Garofalo, C., and Inturri, G. (2013). Agent-based simulation of pedestrian behaviour in closed spaces: a museum case study. *arXiv preprint arXiv:1302.7153*, page 14.

Pluchino, A., Rapisarda, A., and Garofalo, C. (2010). The Peter principle revisited: A computational study. *Physica A: Statistical Mechanics and its Applications*, 389(3):467–472.

Pluchino, A., Rapisarda, A., and Garofalo, C. (2011). Efficient promotion strategies in hierarchical organizations. *Physica A: Statistical Mechanics and its Applications*, 390(20):3496–3511.

Pollet, T. V., Roberts, S. G. B., and Dunbar, R. I. M. (2011a). Extraverts Have Larger Social Network Layers. *Journal of Individual Differences*, 32(3):161–169.

Pollet, T. V., Roberts, S. G. B., and Dunbar, R. I. M. (2011b). Use of social network sites and instant messaging does not lead to increased offline social network size, or to emotionally closer relationships with offline network members. *Cyberpsychology, behavior and social networking*, 14(4):253–8.

Pooyandeh, M. and Marceau, D. J. (2013). A spatial web/agent-based model to support stakeholders' negotiation regarding land development. *Journal of environmental management*, 129:309–23.

Popescul, A. and Ungar, L. (2003). Statistical relational learning for link prediction. In *Proc. of the Workshop on Learning Statistical Models from Relational Data*.

Potterat, J. J., Muth, S. Q., Rothenberg, R. B., Zimmerman-Rogers, H., Green, D. L., Taylor, J. E., Bonney, M. S., and White, H. A. (2002). Sexual network structure as an indicator of epidemic phase. *Sexually Transmitted Infections*, 78(1):i152–i158.

Prenkert, F. and Føgesvold, A. (2014). Relationship strength and network form: An agent-based simulation of interaction in a business network. *Australasian Marketing Journal (AMJ)*, 22(1):15–27.

Proctor, C. H. and Loomis, C. (1951). Analysis of sociometric data. In Jahoda, M., Deutsch, M., and Cook, S., editors, *Research methods in Social Relations*, pages 561–586. Dryden Press, New York.

Prokhorov, A., Pallonen, U., and Fava, J. (1996). Measuring nicotine dependence among high-risk adolescent smokers. *Addictive behaviors*, 21(I):117–127.

Pruyt, E. (2004). System Dynamics Models of Electrical Wind Power. In *The 22th International Conference of the System Dynamics Society, Oxford, England.*

Pujol, J., Sanguesa, R., and Delgado, J. (2002). Extracting Reputation in Multi Agent Systems by Means of Social Network Topology. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 1*, pages 467–474.

Putnam, R. D. (2001). *Bowling Alone*. Simon and Schuster, New York.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raab, J. and Milward, H. B. (2003). Dark Networks as Problems. *Journal of Public Administration Research and Theory*, 13(4):413–439.

Raczynski, S. (2006). *Modeling and Simulation: The Computer Science of Illusion*. Wiley, Chichester.

Railsback, S. F., Lytinen, S. L., and Jackson, S. K. (2006). Agent-based Simulation Platforms: Review and Development Recommendations. *Simulation*, 82(9):609–623.

Rankin, A. and Philip, P. (1963). An epidemic of laughing in the Bukoba district of Tanganyika. *Central African Journal of Medicine*, 2(2):128–128.

Raspe, H., Hueppe, A., and Neuhauser, H. (2008). Back pain, a communicable disease? *International journal of epidemiology*, 37(1):69–74.

Rattigan, M. and Jensen, D. (2005). The case for anomalous link discovery. *ACM SIGKDD Explorations Newsletter*, 7(2):41–47.

Rayner, K. E., Schniering, C. A., Rapee, R. M., Taylor, A., and Hutchinson, D. M. (2013). Adolescent girls' friendship networks, body dissatisfaction, and disordered eating: examining selection and socialization processes. *Journal of abnormal psychology*, 122(1):93–104.

Rees, M. A., Kopke, J. E., Pelletier, R. P., Segev, D. L., Rutter, M. E., Fabrega, A. J., Rogers, J., Pankewycz, O. G., Hiller, J., Roth, A. E., Sandholm, T., Ünver, M. U., and Montgomery, R. A. (2009). A Nonsimultaneous, Extended, Altruistic-Donor Chain. *New England Journal of Medicine*, 360(11):1096–1101.

Regenmortel, M. V. (2004). Emergence in biology. In *Modelling and simulation of biological processes in the context of genomics*.

Regnerus, M. D. (2006). The Parent-Child Relationship and Opportunities for Adolescents' First Sex. *Journal of Family Issues*, 27(2):159–183.

Reid, D. J., McNeill, A., and Glynn, T. J. (1995). Reducing the prevalence of smoking in youth in Western countries: an international review. *Tobacco control*, 4(3):266–277.

Reingold, E. and Tilford, J. (1981). Tidier drawings of trees. *IEEE Transactions on Software Engineering*, SE-7(2):223–228.

Remondino, M. (2008). Diffusion of Innovation in a Social Environment: A Multi Agent Based Model. *Tenth International Conference on Computer Modeling and Simulation (uksim 2008)*, pages 573–578.

Repast (2013). repast.sourceforge.net, Last Accessed: Mar 2014.

Reynolds, C. (1987). Flocks, herds and schools: A distributed behavioral model. *ACM SIGGRAPH Computer Graphics*, 21(4):25–34.

Rhee, M. (2007). The time relevance of social capital. *Rationality and Society*, 19(3):367–389.

Riach, K. and Wilson, F. (2007). Don't screw the crew: Exploring the rules of engagement in organizational romance. *British Journal of Management*, 18(1):79–92.

Richard, E., Baskiotis, N., Evgeniou, T., and Vayatis, N. (2010). Link discovery using graph feature tracking. In *Advances in Neural Information Processing Systems 23*, pages 1966–1974.

Richardson, L., Hemsing, N., Greaves, L., Assanand, S., Allen, P., McCullough, L., Bauld, L., Humphries, K., and Amos, A. (2009). Preventing smoking in young people: a systematic review of the impact of access interventions. *International journal of environmental research and public health*, 6(4):1485–514.

Riolo, C. S., Koopman, J. S., and Chick, S. E. (2001). Methods and measures for the description of epidemiologic contact networks. *Journal of urban health : bulletin of the New York Academy of Medicine*, 78(3):446–57.

Ripley, R., Snijders, T., and Preciado, P. (2012). Manual for RSiena, http://www.stats.ox.ac.uk/˜snijders/siena/RSiena_Manual.pdf.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407.

Roberts, M. (2013). Tube Map in Circles, http://london-underground.blogspot.co.uk/2013/01/tube-map-in-circles.html.

Robinson, S. (1994). *Successful Simulation: A Practical Approach to Simulation Projects*. McGraw-Hill Publishing Co., Berkshire.

Robinson, S. (1999). Three sources of simulation inaccuracy (and how to overcome them). In *Winter Simulation Conference Proceedings*, pages 1701–1708.

Robinson, S. (2001). Soft with a hard centre: discrete-event simulation in facilitation. *Journal of the Operational Research Society*, 52(8):905–915.

Robinson, S. (2004). *Simulation: The Practice of Model Development and Use*. John Wiley & Sons, Chichester.

Robinson, S. (2005). Discrete-event simulation: from the pioneers to the present, what next? *Journal of the Operational Research Society*, 56(6):619–629.

Robinson, S. (2007). The future's bright the future's... Conceptual modelling for simulation! *Journal of Simulation*, 1(3):149–151.

Robinson, S. (2008). Conceptual modelling for simulation Part I: definition and requirements. *Journal of the Operational Research Society*, 59(3):278–290.

Robinson, S., Nance, R. E., Paul, R. J., Pidd, M., and Taylor, S. J. E. (2004). Simulation model reuse: definitions, benefits and obstacles. *Simulation Modelling Practice and Theory*, 12(7-8):479–494.

Rocha, L. E. C., Liljeros, F., and Holme, P. (2011). Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts. *PLoS computational biology*, 7(3):1–9.

Rogers, E. M. (2003). *Diffusion of Innovations, 5th Edition*. Free Press, New York.

Ronald, N., Dignum, V., Jonker, C., Arentze, T., and Timmermans, H. (2012). On the engineering of agent-based simulations of social activities with social networks. *Information and Software*

Technology, 54(6):625–638.

Rothenberg, R. B., Sterk, C., Toomey, K. E., Potterat, J. J., Johnson, D., Schrader, M., and Hatch, S. (1998). Using social network and ethnographic tools to evaluate syphilis transmission. *Sexually Transmitted Diseases*, 25(3):154–160.

Rowe, D. C., Chassin, L., Presson, C., and Sherman, S. J. (1996). Parental Smoking and the "Epidemic" Spread of Cigarette Smoking. *Journal of Applied Social Psychology*, 26(5):437–454.

Rowe, D. C., Chassin, L., Presson, C. C., Edwards, D., and Sherman, S. J. (1992). An "Epidemic" Model of Adolescent Cigarette Smoking. *Journal of Applied Social Psychology*, 22(4):261–285.

Royston, G., Dost, A., Townshend, J., and Turner, H. (1999). Using system dynamics to help develop and implement policies and programmes in health care in England. *System Dynamics Review*, 15(3):293–313.

Rudolph, K. D., Lansford, J. E., Agoston, A. M., Sugimura, N., Schwartz, D., Dodge, K. A., Pettit, G. S., and Bates, J. E. (2013). Peer Victimization and Social Alienation: Predicting Deviant Peer Affiliation in Middle School. *Child development*, pages 1–16.

Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, 31(4):581–603.

Salovey, P. and Rothman, A. J. (1991). *Envy and Jealousy: Self and society*. Guilford Press, New York.

Salter-Townshend, M. (2012). Analysing my Facebook friends. *Significance*, 9(4):40–42.

Salvy, S. and Haye, K. D. L. (2012). Influence of peers and friends on children's and adolescents' eating and activity behaviors. *Physiology & behavior*, 106(3):369–378.

Sankoff, D. and Kruskal, J. B. (1983). *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison*. Addison-Wesley Publication, Reading.

Santor, D. A., Messervey, D., and Kusumakar, V. (2000). Measuring Peer Pressure, Popularity, and Conformity in Adolescent Boys and Girls: Predicting School Performance, Sexual Attitudes, and Substance Abuse. *Journal of Youth and Adolescence*, 29(2):163–182.

Sargent, R. G. (2005). Verification and validation of simulation models. In *Proceedings of the 37th conference on Winter simulation.*

Sarigöl, E., Pfitzner, R., and Scholtes, I. (2014). Predicting Scientific Success Based on Coauthorship Networks. *arXiv:1402.7268*.

Sawaguchi, T. and Kudo, H. (1990). Neocortical development and social structure in primates.

Primates, 31(April):283–289.

Sawyer, S. M., Afifi, R. A., Bearinger, L. H., Blakemore, S.-J., Dick, B., Ezeh, A. C., and Patton, G. C. (2012). Adolescence: a foundation for future health. *Lancet*, 379(9826):1630–40.

Schaefer, D., Haas, S., and Bishop, N. (2012). A Dynamic Model of US Adolescents' Smoking and Friendship Networks. *American journal of public health*, 102(6):1–15.

Schank, T. and Wagner, D. (2005). Approximating Clustering Coefficient and Transitivity Basic Definitions. *Journal of Graph Algorithms and Applications*, 9(2):265–275.

Schaubroeck, J. and Lam, S. S. K. (2004). Comparing lots before and after: Promotion rejectees' invidious reactions to promotees. *Organizational Behavior and Human Decision Processes*, 94(1):33–47.

Schelling, T. C. (1971). Dynamic models of segregation. *Journal of mathematical sociology*, 1(2):143–186.

Schuhbauer, M. (2012). http://www.small-world-network.com, Last Accessed: Feb 2012.

Schutte, J. G. and Light, J. M. (1978). The relative importance of proximity and status for friendship choices in social hierarchies. *Social psychology*, 41(3):260–264.

Schwarz, N. and Ernst, A. (2009). Agent-based modeling of the diffusion of environmental innovations - An empirical approach. *Technological Forecasting and Social Change*, 76(4):497–511.

Scott, J. (2005). *Models and Methods in Social Network Analysis (Structural Analysis in the Social Sciences)*. Cambridge University Press, New York.

Sentse, M., Dijkstra, J. K., Salmivalli, C., and Cillessen, A. H. N. (2013). The dynamics of friendships and victimization in adolescence: a longitudinal social network perspective. *Aggressive behavior*, 39(3):229–38.

Shakarian, P., Roos, P., and Johnson, A. (2012). A review of evolutionary graph theory with applications to game theory. *Bio Systems*, 107(2):66–80.

Shapiro, C. (1989). The theory of business strategy. *The Rand journal of economics*, 20(1):125–137.

Shazam (2013). www.shazam.com, Last Accessed: Feb 2014.

Sherif, M. (1937). An experimental approach to the study of attitudes. *Sociometry*, 135(3503):554–5.

Shiell, A. and Chapman, S. (2000). The inertia of self-regulation: a game-theoretic approach to

reducing passive smoking in restaurants. *Social science & medicine*, 51(7):1111–9.

Shimbel, A. (1953). Structural parameters of communication networks. *The Bulletin of Mathematical Biophysics*, 15(4):501–507.

Sias, P., Heath, R., and Perry, T. (2004). Narratives of workplace friendship deterioration. *Journal of Social and Personal Relationships*, 21(3):321–340.

Siebers, P. O., Macal, C. M., Garnett, J., Buxton, D., and Pidd, M. (2010). Discrete-event simulation is dead, long live agent-based simulation! *Journal of Simulation*, 4(3):204–210.

Sieverding, M. (2009). 'Be Cool!': Emotional costs of hiding feelings in a job interview. *International Journal of Selection and Assessment*, 17(4):391–401.

Simantov, E., Schoen, C., and Klein, J. D. (2000). Health-Compromising Behaviors: Why Do Adolescents Smoke or Drink? *Archives of Pediatric Adolescent Medicine*, 154(10):1025–1033.

Simon, H. (1955). On a class of skew distribution functions. *Biometrika*, 42(3/4):425–440.

Singh, S. (2002). Erdos-Bacon Numbers. *The Telegraph*.

Skovholt, K. and Svennevig, J. (2006). Email Copies in Workplace Interaction. *Journal of Computer-Mediated Communication*, 12(1):42–65.

Skyrms, B. and Pemantle, R. (2000). A dynamic model of social network formation. *Proceedings of the National Academy of Sciences of the United States of America*, 97(16):9340–6.

Smith, J. R. and Louis, W. R. (2008). Do as we say and as we do: the interplay of descriptive and injunctive group norms in the attitude-behaviour relationship. *The British journal of social psychology / the British Psychological Society*, 47(Pt 4):647–66.

Snijders, T. A. B. (1996). Stochastic actor-oriented models for network change. *Journal of mathematical sociology*, 21(1-2):149–172.

Snijders, T. A. B. (2001). The statistical evaluation of social network dynamics. *Sociological methodology*, 31(1):361–395.

Snijders, T. A. B. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure*, 3(2):1–40.

Snijders, T. A. B., Steglich, C. E. G., and Schweinberger, M. (2007a). Modeling the co-evolution of networks and behavior. In *Longitudinal models in the behavioral and related sciences*, pages 1–32.

Snijders, T. A. B., Steglich, C. E. G., and Schweinberger, M. (2007b). Modeling the co-evolution

of networks and behavior. pages 1–32.

Snijders, T. A. B., Van de Bunt, G. G., and Steglich, C. E. G. (2010). Introduction to stochastic actor-based models for network dynamics. *Social Networks*, 32(1):44–60.

Soffer, S. and Vázquez, A. (2005). Network clustering coefficient without degree-correlation biases. *Physical Review E*, 71(5):057101.

Solomonoff, R. (1952). An exact method for the computation of the connectivity of random nets. *The bulletin of mathematical biophysics*, 14(2):153–157.

Solomonoff, R. and Rapoport, A. (1951). Connectivity of random nets. *The Bulletin of Mathematical Biophysics*, 13(2):107–117.

Song, Z. (2006). Addiction and Cessation Functions in the Agent Based Smoking Model, http://www.brookings.edu/, Last Accessed: May 2014.

SoundHound (2013). www.soundhound.com, Last Accessed: Feb 2014.

Stam, C., Jones, B., Nolte, G., Breakspear, M., and Scheltens, P. (2007). Small-World Networks and Functional Connectivity in Alzheimer's Disease. *Cerebral Cortex*, 17(1):92–99.

Starkey, F., Audrey, S., Holliday, J., Moore, L., and Campbell, R. (2009). Identifying influential young people to undertake effective peer-led health promotion: the example of A Stop Smoking In Schools Trial (ASSIST). *Health education research*, 24(6):977–88.

Starkey, F., Moore, L., Campbell, R., Sidaway, M., and Bloor, M. (2005). Rationale, design and conduct of a comprehensive evaluation of a school-based peer-led anti-smoking intervention in the UK: the ASSIST cluster randomised trial [ISRCTN55572965]. *BMC public health*, 5:43.

Steglich, C. (2013). Analysis of Longitudinal Social Network Data using SIENA. In *Proceedings of the Sunbelt XXXIII conference, Berlin, Germany*.

Steglich, C., Sinclair, P., Holliday, J., and Moore, L. (2012). Actor-based analysis of peer influence in A Stop Smoking In Schools Trial (ASSIST). *Social Networks*, 34(3):359–369.

Steglich, C., Snijders, T., and West, P. (2006). Applying SIENA: An Illustrative Analysis of the Coevolution of Adolescents' Friendship Networks, Taste in Music, and Alcohol Consumption. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 2(1):48–56.

Stein, W. (2014). http://www.sagemath.org/, Last Accessed: March 2014.

Stewart, W. J. (2009). *Probability, Markov Chains, Queues, and Simulation: The Mathematical Basis of Performance Modeling*. Princeton University Press, NJ, USA.

Sumter, S. R., Bokhorst, C. L., Steinberg, L., and Westenberg, P. M. (2009). The developmental pattern of resistance to peer influence in adolescence: will the teenager ever be able to resist? *Journal of adolescence*, 32(4):1009–21.

Sundar, S. S., Tamul, D. J., and Wu, M. (2014). Capturing "cool": Measures for assessing coolness of technological products. *International Journal of Human-Computer Studies*, 72(2):169–180.

Suweis, S., Simini, F., Banavar, J. R., and Maritan, A. (2013). Emergence of structural and dynamical properties of ecological mutualistic networks. *Nature*, 500(7463):449–52.

SWARM (2012). http://www.swarm.org, Last Accessed: Mar 2014.

Sweetser, A. (1999). A comparison of system dynamics (SD) and discrete event simulation (DES). In *17th International Conference of the System Dynamics Society*, pages 20–23.

Sylvester, J. J. (1878). Chemistry and Algebra. *Nature*, 17(432):277–296.

Tadelis, S. (2012). *Game Theory: An Introduction*. Princeton University Press, NJ, USA.

Tager, I. B. (2008). The effects of second-hand and direct exposure to tobacco smoke on asthma and lung function in adolescence. *Paediatric respiratory reviews*, 9(1):29–37.

Tako, A. and Robinson, S. (2009). Comparing discrete-event simulation and system dynamics: users' perceptions. *Journal of the Operational Research Society*, 60(3):296–312.

Tampubolon, G., Subramanian, S. V., and Kawach, I. (2013). Neighbourhood Social Capital And Individual Self-Rated Health In Wales. *Health Economics*, 22(November):14–21.

Taskar, B., Wong, M. F., Abbeel, P., and Koller, D. (2003). Link prediction in relational data. In *Advances in neural information processing systems*.

Taylor, S. J. E., Balci, O., Cai, W., Loper, M. L., Nicol, D. M., and Riley, G. (2013). Grand challenges in modeling and simulation: expanding our horizons. In *Proceedings of the 2013 ACM SIGSIM conference on Principles of advanced discrete simulation.*, pages 409–414.

Taylor, S. J. E., Eldabi, T., Riley, G., Paul, R. J., and Pidd, M. (2009). Simulation modelling is 50! Do we need a reality check? *Journal of the Operational Research Society*, 60:S69–S82.

Taylor, S. J. E. and Robinson, S. (2006). So where to next? A survey of the future for discrete-event simulation. *Journal of Simulation*, 1(1):1–6.

Taylor, S. J. E., Turner, S. J., Strassburger, S., and Mustafee, N. (2012). Bridging the gap: A standards-based approach to OR/MS distributed simulation. *ACM Transactions on Modeling and Computer Simulation*, 22(4):18.

Tesfatsion, L. (2003). Agent-based computational economics: modeling economies as complex adaptive systems. *Information Sciences*, 149(4):262–268.

Tesfatsion, L. (2006). Agent-based computational economics: A constructive approach to economic theory. In *Handbook of Computational Economics, Vol. 2: Agent-Based Computational Economics*, number October 2003, pages 831–880.

The Associated Press (2013). Number of active users at Facebook over the years. *Yahoo! News*.

The Scottish Government (2013). Creating a Tobacco-Free Generation: A Tobacco Control Strategy for Scotland, http://www.scotland.gov.uk/Resource/0041/00417331.pdf.

Thomas, R. E. and Perera, R. (2006). School-based programmes for preventing smoking. *Cochrane Database of Systematic Reviews 2006*, 3:CD001293.

Thompson, C. (2009). Random Promotions. *The New York Times*.

Todd, P. (1997). Searching for the next best mate. In *Simulating social phenomena*, pages 419–436. Springer, Berlin.

Tollis, I. G., Battista, G. D., Eades, P., and Tamassia, R. (1998). *Graph Drawing: Algorithms for the Visualization of Graphs*. Prentice Hall.

Townsend, J. (1996). Price and consumption of tobacco. *British Medical Bulletin*, 52(1):132–142.

Transport For London (2014). London Undergound Map.

Trichopoulos, D., Kalandidi, A., Sparros, L., and Macmahon, B. (1981). Lung cancer and passive smoking. *International Journal of Cancer*, 27(1):1–4.

Turner, K., West, P., Gordon, J., Young, R., and Sweeting, H. (2006). Could the peer group explain school differences in pupil smoking rates? An exploratory study. *Social science & medicine (1982)*, 62(10):2513–25.

Twitter (2013). www.twitter.com, Last Accessed: Mar 2014.

Tyas, S. L. and Pederson, L. L. (1998). Psychosocial factors related to adolescent smoking: a critical review of the literature. *Tobacco control*, 7(4):409–20.

Ugander, J., Karrer, B., Backstrom, L., and Marlow, C. (2011). The anatomy of the facebook social graph. *arXiv preprint*, pages 1–17.

US Department of Health and Human Services (2004). The health consequences of smoking: a report of the Surgeon General, http://www.ncbi.nlm.nih.gov/pubmed/20669512.

Valente, T. W. and Davis, R. L. (1999). Accelerating the Diffusion of Innovations Using Opinion Leaders. *Annals of the American Academy of Political and Social Science*, 566:55–67.

van Eck, P. S., Jager, W., and Leeflang, P. S. H. (2011). Opinion Leaders' Role in Innovation Diffusion: A Simulation Study. *Journal of Product Innovation Management*, 28(2):187–203.

van Roosmalen, E. H. and McDaniel, S. A. Adolescent smoking intentions: Gender differences in peer context.

Verzi, S. J., Apelberg, B., Rostron, B., Brodsky, N. S., and Brown, T. J. (2012). An Agent-Based Approach for Modeling Population Behavior and Health with Application to Tobacco Use, http://www.sandia.gov/CasosEngineering/docs/PSM_9_18_2012_PLOS.pdf.

Viana, J., Brailsford, S., Harindra, V., and Harper, P. R. (2014). Combining discrete-event simulation and system dynamics in a healthcare setting: A composite model for Chlamydia infection. *European Journal of Operational Research*, In Press.

Viana, J., Rossiter, S., Channon, A. A., Brailsford, S. C., and Lotery, A. (2012). A multi-paradigm, whole system view of health and social care for age-related macular degeneration. In *Proceedings of the 2012 Winter Simulation Conference*, pages 1–12.

von Neumann, J. (1966). *Theory of Self-Reproducing Automata*. University of Illinois Press, Champaign, Illinois.

von Neumann, J. and Morgenstern, O. (1944). *Theory of Games and Economic Behavior.* Princeton University Press, NJ, USA.

Walter, F. E., Battiston, S., and Schweitzer, F. (2009). Personalised and dynamic trust in social networks. In *Proceedings of the third ACM conference on Recommender systems*, page 197. ACM Press, New York.

Wang, A. (2003). An Industrial Strength Audio Search Algorithm. In *International Conference on Music Information Retrieval*, pages 7–13.

Wang, A. (2006). The Shazam music recognition service. *Communications of the ACM*, 49(8):44–48.

Wasserman, S. and Faust, K. Cambridge University Press, Cambridge, USA.

Wasserman, S. and Galaskiewicz, J. (1994). *Advances in Social Network Analysis: Research in the Social and Behavioral Sciences*. SAGE Publications, Inc.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442.

Waxman, B. (1988). Routing of multipoint connections. *IEEE Journal on Selected Areas in Communications*, 6(9):1617–1622.

Webb, J. N. (2007). *Game Theory: Decisions, Interaction and Evolution*. Springer-Verlag, London.

Weibull, J. W. (1997). *Evolutionary Game Theory*. MIT Press, Cambridge, USA.

Wen, Y., Peng, J., and Tong, P. (2013). Research on the Strategic Alliance Between Hospitals and Suppliers Based on Evolutionary Game Theory. In Qi, E., Shen, J., and Dou, R., editors, *International Asia Conference on Industrial Engineering and Management Innovation*, pages 1215–1225. Springer, Berlin.

White, H. (1970). Search parameters for the small world problem. *Social forces*, 49(2):259–264.

Wilensky, U. (1999). NetLogo. http://ccl.northwestern.edu/netlogo/, Last Accessed: May 2014.

Wilkie, C., Macdonald, S., and Hildahl, K. (1998). Community case study: suicide cluster in a small Manitoba community. *Canadian journal of psychiatry. Revue canadienne de psychiatrie*, 43(8):823–8.

Wilson, C., Boe, B., Sala, A., Puttaswamy, K. P., and Zhao, B. Y. (2009). User interactions in social networks and their implications. In *Proceedings of the 4th ACM European conference on Computer systems*, pages 205–218. ACM, New York.

Wilson, R. A. (2002). *Graphs, Colourings and the Four-Colour Theorem*. Oxford University Press, Cambridge, USA.

Wing, R. and Jeffery, R. (1999). Benefits of recruiting participants with friends and increasing social support for weight loss and maintenance. *Journal of consulting and clinical psychology*, 61(1):132–138.

Wolfram, S. (1983). Statistical mechanics of cellular automata. *Reviews of Modern Physics*, 55(3):601–644.

Wood, P. (2011). Climate change and game theory. *Annals of the New York Academy of Sciences*, 1219:153–70.

Wood, R. S. (2006). Tobacco's Tipping Point: The Master Settlement Agreement as a Focusing Event. *Policy Studies Journal*, 34(3):419–436.

World Health Organisation (2009). Gender, Women, and the Tobacco Epidemic, http://www.who.int/tobacco/publications/gender/women_tob_epidemic/en/. pages 29–50.

World Health Organisation (2013). Tobacco Facts, http://www.who.int/tobacco/mpower/tobacco_facts/en/.

Xing, W. and Ghorbani, A. (2004). Weighted PageRank algorithm. In *Proceedings. Second Annual Conference on Communication Networks and Services Research*, pages 305–314. Ieee.

Xu, J. and Chen, H. (2008). The topology of dark networks. *Communications of the ACM*, 51(10):58.

Yahoo! (2013). www.yahoo.com, Last Accessed: Aug 2013.

Yan, E. and Ding, Y. (2011). Discovering author impact: A PageRank perspective. *Information Processing & Management*, 47(1):125–134.

Yang, S. H., Long, B., and Smola, A. (2011). Like like alike: joint friendship and interest propagation in social networks. In *Proceedings of the 20th international conference on World wide web*, pages 537–546.

Yule, G. U. (1925). A Mathematical Theory of Evolution, Based on the Conclusions of Dr. J. C. Willis, F.R.S. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 213(402-410):21–87.

Yung, D. (2012). Personalized Pagerank for link prediction. http://shom83.blogspot.co.uk/, Last Accessed: May 2014.

Zeman, E. (2011). HTC: iPhone Not 'Cool' Anymore. *Information Week*.

Zhang, T., Gensler, S., and Garcia, R. (2011). A Study of the Diffusion of Alternative Fuel Vehicles: An Agent-Based Modeling Approach. *Journal of Product Innovation Management*, 28(2):152–168.

Zhou, Y., Ma, Z., and Brauer, F. (2004). A discrete epidemic model for SARS transmission and control in China. *Mathematical and Computer Modelling*, 40(13):1491–1506.

Zhou, Z., Chan, W. K. V., and Chow, J. H. (2009). Agent-based simulation of electricity markets: a survey of tools. *Artificial Intelligence Review*, 28(4):305–342.

Zhu, J., Hong, J., and Hughes, J. (2004). PageCluster: Mining conceptual link hierarchies from Web log files for adaptive Web site navigation. *ACM Transactions on Internet Technology (TOIT) special issue on "Machine Learning for the Internet"*, 4(2):185–208.

Zimmermann, P. (2004). Attachment representations and characteristics of friendship relations during adolescence. *Journal of experimental child psychology*, 88(1):83–101.

Zuckerberg, M. (2013). http://facebook.com/zuck, Last Accessed: Dec 2013.