Privacy Loss and Exploitation in e-Commerce Preference Searching

Rhys Smith

Cardiff University School of Computer Science

August 2009

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Cardiff University.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Abstract

An area of e-commerce that is very much an active area of research is that of using an individual's preferences to enhance search. The development of this research area, and the model used to produce all existing methods, has an implicit assumption that the vendor to whom the consumer is releasing their preference information is trustworthy. This assumption results in two major issues: the certainty of privacy loss, and the potential for exploitation.

Motivated by a wide ranging investigation into the concept and history of privacy and the methods used to protect it, along with the conclusion drawn from this investigation that the previously used methods of privacy protection via legal means can no longer keep pace with technological evolution, this thesis presents an alternative approach to searching with a consumer's preferences that enables the main goal of preference searching whilst also minimising privacy loss and the potential for exploitation.

A proof of concept implementation of this approach, called "Gradual Partial Release", is presented. Essentially, its aim is to minimise privacy loss and exploitation by splitting a consumer's preferences up into multiple subsets of these preferences – *partial* release – to be released one at a time to the vendor – *gradual* release – until sufficient results are returned.

Three different Gradual Partial Release algorithms, that split up preferences into subsets in different ways, are presented, along with measures enabling quantitative measurement of privacy loss and exploitation to allow evaluation of their effectiveness.

An evaluation was performed of the effectiveness and efficiency of the Gradual Partial Release algorithms, comparing the effectiveness (in terms of minimising of privacy loss and exploitation) of each algorithm and to the current approach to preference searching. Experiments show that the proposed Gradual Partial Release approach enables the basic idea of preferences searching whilst simultaneously offering the possibility of reduced privacy loss and reduced exploitation.

ACKNOWLEDGEMENTS

I would like to gratefully acknowledge the support and supervision of Dr Jianhua Shao: without his help, support, and dedication, this thesis would not exist.

I would also like to thank my colleagues in Information Services at Cardiff University; all who have been extremely supportive in my quest to complete this work whilst also performing the duties of my job. I would also like to thank the EPSRC who provided the initial grant funding that enabled this work.

Finally, is difficult to overstate my gratitude to my partner, Sarah, along with my friends and family; without all of their support it is doubtful that my sanity would have survived the PhD experience. Some may argue whether or not that statement is true; but anyway, to all, a heartfelt thank you.

CONTENTS

Acknowledgements iv
List of Figures ix
List of Tables xi
List of Algorithms xii

1 Introduction **1**
1.1 Problem Statement 3
1.2 Major Contributions 4
1.3 Thesis Structure 5

2 Privacy: Concept and Development **7**
2.1 The Nature of Privacy 7
2.2 The Development of Privacy 11
2.2.1 The Infancy of Privacy 11
2.2.2 The Legal Age of Privacy 13
2.2.3 Privacy in the Technological Age 15
2.2.4 Privacy in the Information Age 19
2.2.5 The Future of Privacy 21
2.3 Summary 21

3 Protecting Privacy in e-Commerce **23**
3.1 Privacy Enhancing Technologies 23
3.1.1 Anonymous/Pseudonymous Techniques 23
3.1.1.1 Anonymising the Transport Medium 24
3.1.1.2 Credential Systems 27
3.1.1.3 Privacy in databases 29
3.1.1.4 Summary of Anonymous/Pseudonymous Techniques 31
3.1.2 Onymous Techniques 32

3.1.2.1 Decision Helping Techniques 34
3.1.2.2 Enforcement Techniques 35
3.1.2.3 Summary of Onymous Techniques 36
3.1.3 Summary of Privacy Enhancing Technologies 36
3.2 Enhanced Search Techniques 39
3.2.1 Obtaining Preferences 39
3.2.1.1 Explicit Expression 39
3.2.1.2 Implicit Collection 41
3.2.2 Computing with Preferences 42
3.2.2.1 Explicit Usage 42
3.2.2.2 Implicit Usage 43
3.2.3 Summary of Enhanced Search Techniques 45
3.3 Summary 45

4 Privacy Loss and Exploitation when Preference Searching 47
4.1 e-Commerce 47
4.2 e-Commerce Search 49
4.2.1 Limitations and Issues 51
4.3 Enhanced Searching using Preferences 52
4.3.1 Limitations and Issues 54
4.3.1.1 Privacy 55
4.3.1.2 Exploitation 56
4.4 A New Model 58
4.4.1 Gradual Partial Release 59
4.5 Summary 62

5 Gradual Partial Release of Preferences 64
5.1 Preferences 64
5.1.1 Relative Preference Operators 66
5.1.2 Discrete Value Preferences 67
5.1.3 Range Value Preferences 68
5.1.4 Attribute Preferences 69
5.2 Preference Set 70

6.2.3.1 Privacy Loss 101
6.2.3.2 Exploitation 104
6.2.4 Summary 106
6.3 Further Analysis of GPR Approach 106
6.3.1 Overview of Experiment 107
6.3.2 Experiment Details 107
6.3.3 Results 109
6.3.3.1 Varying Size of Vendor's Catalogue 109
6.3.3.2 Varying Number of Attributes in Vendor's Catalogue 112
6.3.3.3 Varying Number of Values per Attribute in Vendor's Catalogue 114
6.3.3.4 Varying Number of Attributes in Consumer's Preferences 117
6.3.3.5 Varying Number of Values per Attribute in Consumer's Preferences 120
6.4 Summary 124

7 Conclusions **125**
7.1 Privacy in the Information Age 125
7.2 Preference Searching 126
7.3 Quantifying privacy loss and exploitation 127
7.4 Gradual Partial Release 127
7.5 Concluding Remarks 128

A List of Publications **129**

References **130**

List of Figures

4.1 Current e-Commerce Search Scenario 49
4.2 Current e-Commerce Preference Search Scenario 52
4.3 Proposed e-Commerce Preference Search Scenario 60

5.1 Example of a DVP represented as a graph 68
5.2 Example of an Attribute Preference represented as a graph 69
5.3 Attribute Preference with Values 70
5.4 Preference Subset with values removed 74
5.5 Potential preference relations between values 75
5.6 Preference Subset with attributes removed 76
5.7 HFS SCA - Iterative Process 88
5.8 HFS SCA - Preference Subsets Created 89
5.9 SQ SCA - Iterative Process 89
5.10 SQ SCA - Preference Subsets Created 90
5.11 RD SCA - Iterative Process 92
5.12 RD SCA - Preference Subsets Created 92

6.1 Example of Generated Preference Set 99
6.2 Average Privacy Loss, uniformly distributed stock 102
6.3 Distribution of Privacy Loss, uniformly distributed stock 103
6.4 Average Exploitation, uniformly distributed stock 105
6.5 Distribution of Exploitation, uniformly distributed stock 105
6.6 Privacy Loss vs Size of Vendor's Catalogue 109
6.7 Exploitation vs Size of Vendor's Catalogue 110
6.8 Runtime vs Size of Vendor's Catalogue 110
6.9 Queries vs Size of Vendor's Catalogue 111
6.10 Network Traffic vs Size of Vendor's Catalogue 111
6.11 Privacy Loss vs Attributes in Vendor's Catalogue 112
6.12 Exploitation vs Attributes in Vendor's Catalogue 112
6.13 Runtime vs Attributes in Vendor's Catalogue 113
6.14 Queries vs Attributes in Vendor's Catalogue 113

6.15 Network Traffic vs Attributes in Vendor's Catalogue 114
6.16 Privacy Loss vs Values per Attribute in Vendor's Catalogue 114
6.17 Exploitation vs Values per Attribute in Vendor's Catalogue 115
6.18 Runtime vs Values per Attribute in Vendor's Catalogue 116
6.19 Queries vs Values per Attribute in Vendor's Catalogue 116
6.20 Network Traffic vs Values per Attribute in Vendor's Catalogue . . . 117
6.21 Privacy Loss vs Attributes in Consumer's Preferences 118
6.22 Exploitation vs Attributes in Consumer's Preferences 119
6.23 Runtime vs Attributes in Consumer's Preferences 119
6.24 Queries vs Attributes in Consumer's Preferences 120
6.25 Network Traffic vs Attributes in Consumer's Preferences 120
6.26 Privacy Loss vs Values per Attribute in Consumer's Preferences . . 121
6.27 Exploitation vs Values per Attribute in Consumer's Preferences . . . 122
6.28 Runtime vs Values per Attribute in Consumer's Preferences 122
6.29 Queries vs Values per Attribute in Consumer's Preferences 123
6.30 Network Traffic vs Values per Attribute in Consumer's Preferenecs . 123

List of Tables

3.1 Anonymous/Pseudonymous Privacy Enhancing Technologies 33
3.2 Onymous Privacy Enhancing Technologies 37

6.1 Privacy Loss, uniformly distributed stock 102
6.2 Exploitation, uniformly distributed stock 104

List of Algorithms

5.1 Completely Ordering a Preference Set 83
5.2 GPR SCA - Highly Focused Subsets 87
5.3 GPR SCA - Single Query 89
5.4 GPR SCA - Relax Down 91
5.5 Gradually Releasing Preference Subsets 93

CHAPTER 1

INTRODUCTION

The dawning of the information age has brought about a world where information itself is valuable: possession of intellectual, personal, social, and economic information about oneself can create opportunities for an individual and provide them with social and economic advantages. On the other hand, it has enabled a world where an individual can unwittingly lose control of this valuable information while trying to take advantage of new technologies born of the age. This loss of control and possession of personal information represents both an abstract loss of privacy and a concrete loss of valuable information.

One specific aspect of the information age is e-commerce; this is an aspect containing many prime examples of new technologies that manifest privacy-related issues to participating individuals. These issues have had a major effect on e-commerce: it is a virtually indisputable fact that concerns about privacy in the current climate of e-commerce are limiting its acceptance and usage by the public by a notable amount [48, 66, 85, 92, 97, 136, 180, 181]. Worryingly, the future seems destined to produce even greater possibilities for such invasions of individual privacy, and in a much less obvious and more insidious way than ever before. Given that current privacy concerns are already hampering the current growth of e-commerce, future privacy violations can only further intensify public disquiet and further impede the growth of e-commerce. Thus, to help maximise the future growth of e-commerce, considerable diligence is needed from the research community in identifying such privacy issues with new technologies – both existing and forthcoming – and to work to minimise and alleviate these concerns. Indeed, influential companies such as IBM and Microsoft have recently supported this view [91, 118].

One area of e-commerce technology is the area of information filtering and

retrieval. This area is one that has remained somewhat stagnant since its inception: most e-commerce websites still only provide a simple binary keyword search option to help individuals to locate items that they desire to purchase. The consumer would typically interact with such a website by declaring their search terms in a carefully structured, and very limited, way. This information would be sent to a vendor and turned into a query to search their stock database; the list of results matching the query is then returned to the consumer.

Two particular shortcomings associated with such binary keyword based information filtering and retrieval techniques are the related and equally undesirable problems of *information overload* and *empty result sets*. Information overload is the problem of large amounts of results being returned due to too broad a search being submitted, many of which are likely to be substandard, while empty result sets are the opposite problem of no results being returned due to too specific a search being submitted. Both are equally undesirable from a usability perspective as they are likely to frustrate consumers, potentially resulting in a consumer giving up after a few failed search attempts.

To overcome these shortcomings, improved search techniques are being developed with an emphasis that is shifting away from the simple paradigm described above and edging towards advanced interfacing between business and customer (for example, using personalisation [23]). Such advanced techniques work by making use of a consumer's personal preferences for searching: the consumer's preferences are used by the vendor to assist with the evaluation of database queries by helping target search results towards these preferences. For example, instead of a consumer issuing a search for a used car with strict binary search terms equating to "Make is BMW; Colour is Black" – which may return no results due to its exact-match specificity – a consumer could instead issue a query that equates to "I'd like a BMW or a Mercedes; I like Black, but Silver would be okay".

All work developed thus far in this area is based upon what we have called the "complete release" paradigm. This assumes that the consumer's preferences are to be sent in full to the vendor, who will subsequently analyse the preferences, perform the necessary calculations, and return the best matching results from their stock database. An assumption implicit to this paradigm is that the vendor is entirely trustworthy in the handling of these preferences. In practice, however, the main goal of most business is to turn as large a profit as possible, so if an opportunity to increase profit presents itself then this assumption can no longer definitely hold to be valid. Thus, the paradigm can be shown to present two main

problems, both related to the issue of trust:

- Firstly, there are some obvious inherent privacy implications. In order to take advantage of these newly designed search technologies, a consumer has to release the entirety of their preference information (personal information about their likes and dislikes) to an entity that they generally would – and should – have no good reason to trust. The issue here is fairly simple - consumers have a right to keep the information about their preferences private if they so wish, but in order to use the enhanced search technologies they have to give up this right.

- Secondly, there is a potential problem of exploitability. Given a hypothetical world where we assume a consumer has correctly expressed all of their preferences and passed them across to a vendor, and that the vendor is purely interested in maximising the possibility of making a sale to that individual, then the new technology will indeed achieve what it is designed to achieve – blindly maximising the quality of the search results for its customers. If we assume that the vendor is not trustworthy, however, we can easily foresee the possibility that they could instead use the information gifted to them by the consumer to maximise many things other than simply the quality of results returned to the participating individual. The most obvious of these possibilities is simply maximising profit.

Due to these issues of privacy loss and the potential for exploitation inherent in the currently proposed approach to preference-enhanced search techniques, a new consumer-centric approach is needed to address these issues and add to the arsenal of tools available to the privacy-conscious consumer.

1.1 PROBLEM STATEMENT

The problem inherent with the current preference-enhanced search techniques is that they release all of a consumer's preference information to a vendor. Therefore, to avoid a loss of privacy and a potential for exploitation, it is clear that a new consumer-centric approach should consider how the release of preference information to the vendor may be "disguised" and "controlled" in such a way that the preference is largely kept private yet the benefits of preference-enhanced search are still seen by the consumer.

In this thesis, we investigate the issue of how to control the release of preference information to a vendor. Our hypothesis is this: by suitably partitioning preference information and gradually releasing them to a vendor, we can develop a preference search method that is more effective than existing techniques in terms of privacy loss, potential exploitation, and search utility. That is, we will be able to enable the retrieval of a set of items that must closely match the consumer's preferences while measurably minimising the amount of preferences released and the exploitation that could have occurred using existing techniques.

1.2 MAJOR CONTRIBUTIONS

The overriding contribution of this thesis is a new approach to enabling preference-enhanced searching that does not require participating consumers to release all of their private preference information at once; simultaneously reducing their privacy loss and the possibility for exploitation. The new approach resulted in a basic framework that allows the *gradual partial release* of a consumer's preferences – aiming to gradually send a certain fraction of the consumer's full preferences to the vendor. This framework, implemented as a proof of concept, is shown to successfully enable the goal of preference-enhanced searching while also minimising individual privacy loss and the possibility of exploitation of these preferences.

More specifically, this thesis makes the following contributions:

- A comprehensive analysis of the concept of privacy and historical development of its protection; and how this has affected the world of e-commerce;
- Methods designed to measure privacy loss and exploitation in the context of user preferences and search;
- A new method of preference-enhanced searching called "Gradual Partial Release" (GPR) that aims to gradually send a subset of the consumer's preferences to the vendor, enabling preference searching while simultaneously minimising privacy loss and exploitation of these preferences;

Looking from a broader perspective, these contributions have two intended consequences. The first, and most obvious, is that the beginnings of a new tool for the privacy-conscious individual has been created that could be useful in their everyday dealings with the world of e-commerce. Secondly, it is demonstrated that,

with some additional effort, it can be possible to design new e-commerce technologies and techniques with a more consumer-centric viewpoint. It is hoped that this effort may lead to e-commerce technologies that may be developed in the future to follow a more privacy-friendly path.

1.3 THESIS STRUCTURE

The remainder of this thesis consists of the following:

Chapter 2 explores the history and development of privacy throughout the ages, concentrating on how developing technologies have affected this history. This information provides a solid understanding of the current world of privacy, giving a solid context within which to place the work of this thesis.

Chapter 3 examines existing technologies and techniques that have been created and designed for privacy protection in the world of e-commerce, discussing their core concepts and commenting on their effectiveness. It also examines existing work in the area of the use of preferences in order to enhance search techniques. Together, this gives us an understanding of where the work of this thesis fits in amongst other relevant work.

Chapter 4 discusses in detail the problem this thesis deals with. It examines the development of search techniques into preference-enhanced searching, looks at the current "complete release" approach used to enable this, and demonstrates the problems that this approach exhibits. A detailed model and scope of the problem is developed, allowing the problem to be precisely understood, thus enabling potential solutions. A new model is then introduced that aims to help mitigate against these issues.

Chapter 5 details the new model, called "Gradual Partial Release" (GPR). A set of measures designed to allow evaluation of the effectiveness of GPR is next introduced. The full detail of the GPR approach is then presented, including some algorithms that implement GPR in different ways.

Chapter 6 presents an evaluation of the GPR approach and algorithms as implemented in the form of a proof of concept system. The experimental setup is discussed, along with the methodology used to evaluate the system. The evaluation itself considers the effectiveness of the proposed GPR algorithms and compares them to the existing approach.

Chapter 7 summarises the findings from previous chapters, drawing some conclusions, and providing some suggestions as to what the next steps of work should be.

Chapter 2

Privacy: Concept and Development

Privacy is a core theme of this thesis; and an idea fundamental to the work of this thesis is the idea that individual privacy is something that is worth protecting wherever possible. It would, however, be remiss to simply assume that this is indeed true; therefore, an examination of the very idea of privacy is necessary. Several obvious relevant questions that need answering quickly present themselves: What exactly is privacy? Why is it important? Where did the idea start and how has it changed throughout its lifetime? How has technology influenced these changes? This chapter aims to answer these questions, and more; giving a solid foundation on which the remainder of this thesis is built.

2.1 The Nature of Privacy

Privacy, as a concept, is highly interesting. Perhaps its most striking feature is the fact that nobody seems able to agree upon what it actually *is*. The "right to privacy" has inspired considerable debate in many fields of thinking: including the areas of law, philosophy, sociology, politics, and more recently, computer science. This debate is fascinating, complex, and at times rather surprising. Furthermore, how this right to privacy fares in the context of the world of e-commerce is an even more contentious issue. Thus, our first stop in our journey of understanding privacy will be in this neighbourhood of the nature of privacy.

At first glance, the idea of privacy seems fairly intuitive. When pressed upon to elucidate this idea of privacy in a clear-cut all-encompassing definition and defence of privacy, however, people consistently flounder. Thus, debate about this most

impalpable of human values has raged throughout recorded history. These debates became prominent in the philosophy/sociology literature in the latter half of the 20th century: Benn and Gaus' anthology [18], Schoeman's anthology [153], and Weintraub and Kumar's anthology [176] have collected many of the important arguments presented throughout these decades.

The most rudimentary area of this debate is one which is concerned with the defence of privacy: why is privacy important, and what does retaining privacy bring to an individual? Many theorists have produced defences of privacy that state that privacy is a highly important human value necessary for many aspects of an individual's moral and social being, such as: privacy being a requirement of the ability to develop diverse and meaningful relationships [68, 139]; privacy being a basic aspect of individual personality and integrity [68]; privacy being a requirement for human dignity and retaining one's uniqueness and autonomy [19]; and privacy being a necessary prerequisite for intimacy [71].

Another area of debate – probably the most fundamental and essential of areas – is the *definition* of privacy. The fact is that nobody has yet produced a single agreed-upon definition for the right to privacy – and this is not at all surprising since an individual's conception of privacy is based partly upon their society's general conception of privacy [59, 178, 179] and partly upon their own life's experiences and general social attitudes. Despite this difficulty, many people have, of course, attempted to produce all-encompassing definitions of "privacy". Schoeman noted that they all fall into three main categories [153]:

(a) The *right* an individual has in being able to control access to personal information about themselves;

(b) The *measure of control* an individual has over information about themselves, or who has sensory access to them;

(c) The *state* of limited access to an individual and their personal information.

Each of the three proposed categories of definitions has properties that are pleasing in the attempt to produce a single unified definition of privacy, however, they each contain some major problems. The first category is simply a statement about privacy that assumes that privacy is a morally significant human value and therefore something sacred that should be protected, but does not say *why* this should be the case, and indeed does not define what privacy actually *is*. The second category's critics argue that counter-examples can easily be created that disprove

the definition (although its proponents refute this claiming that such examples are not realistic and to produce them "would be to engage in irony" [68]). The third category inherently poses the question as to whether privacy is desirable, and to what extent. It also raises further questions about the difference between privacy itself and the right to privacy – examples where one can be said to have lost privacy but not had one's right to privacy violated (and the converse) are easily constructed.

More recently, theorists have argued that the reason that no-one has yet produced a single unified concept of privacy is that privacy is far too complex to actually capture in a single (relatively) simple definition – therefore it should instead be treated as a collection of related concepts. Benn and Gaus, for example, suggest that privacy is a wide-ranging social concept that shapes how an individual sees and interacts with society [18]. Many studies have been carried out investigating the cultural relativity of privacy: most social theorists claim that privacy is (in some way) recognised and institutionalised in all societies (e.g. [123, 179]), with only a few notable exceptions (e.g. [10]).

The attempt to define privacy has further been complicated with the transition from the industrial society to the technological society through to the information society. In the information society, an individual's concept of self is expanded into provinces additional to the traditional: it gets expressed through, and is affected by, technology and the projection of one's identity to additionally include one's online identity. This addition of a component that exists in a totally different kind of space shifts and blurs the public/private boundary, making an attempt to define privacy even more difficult than ever before. Thus, increasing numbers of theorists have started to subscribe to the view that privacy is in fact a collection of concepts rather than one specific concept.

Not all theorists, on the other hand, have views that are complementary to the idea of privacy. Some of the most common arguments against privacy are those such as the view put forward by Prosser and Thomson who hold that privacy is non-distinctive - there is no "right to privacy" [138, 160]. They argue that while privacy is important, thinking of privacy as something special is unproductive, as any interest that could be categorised as a privacy interest could be equally well explained and better protected by considering other interests or rights, such as property rights and the right to bodily security. This view is not actually 'anti-privacy' *per se*, but can be considered critical of it as it questions the whole foundations of privacy as previously discussed. When investigating privacy from

a purely legal standpoint, Volokh came to a similar conclusion – that privacy is possibly best protected in law by relying on contractual protections [172].

Delving deeper into the criticism of privacy, a highly sceptical view of privacy was produced by Wasserstrom who argued that withholding information about oneself might be morally equivalent to deception, and therefore socially undesirable [174]. He suggests that views on privacy encourage individuals to feel more vulnerable than they should simply by accepting the notion that there are thoughts and actions which should make one feel ashamed or embarrassed, and that privacy encourages hypocrisy and deceit.

Another critical view – one possibly more interesting in the context of e-commerce – sceptical of the reasons for privacy were first introduced by Posner and Stigler [134, 135, 156]. Posner argued that privacy interests are non-distinctive, and are better thought of in economic terms [135]. He posits that information can have value – people will incur costs to discover it – and that there therefore are two economic goods: "privacy" and "prying". He however regards them as instrumental rather than "final" goods, allowing them to be analysed economically. According to this idea, people do not desire privacy for privacy's sake, but for the economic or social advantage that it gives them. His view is that privacy should only be protected when allowing access to information would reduce its value. He classifies personal information such as "my ill health, evil temper, even my income" [135] as facts that should not generally be protected since the main motive for concealment is often to mislead others, or for private economic gain. Since corporate gains enhance the economy more than individual gains, he concludes that defence of individual privacy is hard to justify as it can negatively impact these more "important" corporate gains. Whether one agrees with this stance of course depends entirely on whether one is a corporation or an individual. Etzioni also espoused a similar view that individual rights with regards to privacy can be trumped if doing so benefits society at large [55]. Critics of this viewpoint question whether the idea that such conditions whereby economic benefits can be reaped exist at all [124]. Thinking specifically about e-commerce, these views of privacy do not take into account the facts presented in the Chapter 1 – that favouring corporate over individual privacy leads to a reduction in the amount of e-commerce an individual will engage in. Nevertheless, these are highly interesting viewpoints from the perspective of this thesis, as they suggest the idea that privacy loss can be measured from an economical point of view, rather than a purely social point of view.

One thing that all of these arguments about the nature of privacy – both those

for and those against privacy – have in common is that they all agree on the core existence of the concept of privacy: to debate whether something is materially good or bad must necessarily mandate the existence of the subject of the debate. Thus, virtually all theorists have acknowledged that the idea of privacy is indeed a real concept. Additionally, although a single, all-encompassing view of privacy has not yet been presented (and seems unlikely to ever be presented), individual privacy, however defined, is seen by the vast majority of theorists as a highly important human value and therefore an area that deserves to be fiercely protected in all areas of life. Thus, the assumption made in this thesis that an individual's privacy is something worth saving is indeed a valid assumption to make, and provides the necessity for the work presented in this thesis.

2.2 THE DEVELOPMENT OF PRIVACY

Given an increased understanding about the nature of privacy, the second stop on the journey of understanding privacy is to look at the development of the idea of privacy throughout the ages. This will give a historical perspective that can be built upon to theorise about the future of this development.

The idea of privacy, as with any other human sociological creation, is not absolute and static – developments in society itself have grown and shaped this human value for most of recorded human history. The development of privacy has gone through several main stages in the lead up from its initial articulation in ancient times to its current incarnation in today's world. Each of these stages can be differentiated by their unique view of privacy and how it has been protected.

2.2.1 THE INFANCY OF PRIVACY

The concept of privacy has roots dating back many millennia. In approximately 350 BC in his treatise *Politics* [11], Aristotle distinguished between the public sphere of the city and its political activities (polis) and the private sphere of the household and its domestic life (oikos). Aristotle claimed that the private sphere has an inherent hierarchical structure whose physical assets provide the material ability for the citizen to act in the public political sphere. Thus, the presence of the "private" is a necessity for the smooth running of the "public".

This idea of public and private spheres were embodied in Greco-Roman society, where the public sphere was not just a metaphorical place but an actual physical

space — the Roman *forum* and the Greek *agora* — where public and legal affairs were discussed and where any free man could directly participate in the running of public life [89].

Greek society originally conceived of no separation between the two spheres. The concept of an individual being separate from the polis was brought to the Greek peninsula in the fifth century B.C. by radical Sophists, teaching that the human being was the measure of all things - not the city or gods as was the prevailing philosophy [147]. This view was alien and highly radical but slowly permeated through Greek society, being discussed by the great Greek philosophers and ultimately inspiring Aristotle's *Politics* as an attempt to solve this opposition between the opposing views.

The Romans further developed the idea of the private sphere as opposed to the public sphere. In Roman society, the notion of public equated to the good of the state and its sovereignty, while the notion of private equated to the interests of the individuals in the empire [81, 175]. These notions were actually eventually accepted as fundamental enough that they were incorporated into late Roman Law - in the first chapter of the two sections of the *Corpus Juris Civilis*, the compilation of Roman Law issued by Emperor Justinian in 533-34 AD [175]. Indeed, the words 'public' and 'private' have a Roman origin[1].

As the Roman Empire fell into decline and the middle ages advanced upon Europe, the idea of the separation of the state and the individual, the public and the private, fell out of public consciousness. Society at the time simply did not have any significant distinction between public and private, mainly due to the feudal system of rule which was based upon kinship and bonds of loyalty - basically a network of personal dependent links in which there was no distinction between state and individual [175].

The revival of the separation of public and private began in more modern history with firstly the occurrence of the enlightenment and later with the growth of capitalism, as the separation of the sovereignty and the citizen once more occurred and public political society came back into prevalence. The development of bourgeois society also had heavy influences on this development of the modern public sphere [81].

At this point in the history of privacy, the basic tenet of privacy was slowly being recognised once more as a concrete human value, and thus support for treating it

[1]Public originally comes from the Latin "populus" (*public*), while private originally comes from the Latin "privus" (single, alone)

as such began to grow. With these occurrences the concept of privacy entered a new age as it began to be afforded legal protections in an effort by the legislators to protect this most basic of human rights.

2.2.2 THE LEGAL AGE OF PRIVACY

One of the first areas in the legal arena in which privacy has been deliberated was in English common law, where the idea of the right to privacy has been discussed for at least the last few centuries in numerous legal cases and judgements over the years. One example of this was articulated by Mr. Justice Yates in 1769 in a case actually centred around common law copyright (Millar v. Taylor [185]):

> "It is certain every man has a right to keep his own sentiments, if he pleases: he has certainly a right to judge whether he will make them public, or commit them only to the sight of his friends"

These early English legal cases paved the way for many further privacy developments. Since many American legal practices have their roots in English common law, it has been argued by many that one of the most basic and well-known of these developments is the U.S. constitution – many arguments have been put forth supporting the idea that the distinction of the private sphere is enshrined within it; for example, the Fourth Amendment protects the physical private sphere of an individual against unreasonable violation, and many have read this to include electronic property and communications. Other of the amendments also arguably have privacy interests (the search and seizure limits of the third and fourth; and the self-incrimination limits of the fifth).

This basic idea of privacy being incorporated into the very foundation of the United States further paved the way for one of the most famous, influential and oft-cited articles about privacy, published in 1890 by Warren and Brandeis [173]. Their article mainly focused upon the privacy violation that can occur due to the public dissemination of information about an individual that the individual would rather stay private (the article was inspired by the press intruding in Mr. Warren's private life – the reporting of the wedding of his daughter). During the course of the article the authors discuss many aspects of privacy, including control over one's private thoughts, and they connect it to many other values, such as an individual's "right to be left alone" and the respect due an individual's "inviolate personality". Somewhat sensibly, they do not, however, attempt to define what

privacy actually *is*. Their main argument is that in order for law to properly protect privacy, privacy needs explicit legal recognition, as simply applying other legal arguments to protect privacy — such as using copyright law or contract law — is inadequate. A significant claim that they make is that privacy is a specific human interest, connected to the moral character, and that this interest is more important in the present than it has been in the past – the importance of privacy is *increasing*.

In the wake of Warren and Brandeis' article, privacy related cases slowly began to surface over the next several decades that supported the article's views. This continued until 1965 when what is known as the "constitutional right to privacy" was explicitly recognised by the US Supreme Court (*Griswold vs. Connecticut*, 381 U.S. 479 - 1965) [86]. Until this point, protection of privacy in U.S. law was simply viewed as being necessary to the protection of other more well-established rights, and were therefore dealt with as such. Justice Douglas wrote in the US Supreme Court Decision that the case in question concerned "a relationship lying within the zone of privacy created by several fundamental constitutional guarantees" – the amendments previously mentioned, each of which creates different zones or "penumbras" of privacy [86]. Some people are now arguing that this ruling should be taken further and another amendment to the constitution should be added that *explicitly* recognises the right to privacy in a specific and fundamental manner.

Outside of the USA during these times, many other countries were developing laws that protected people's privacy. For example, in 1789 in France *La Déclaration des droits de l'Homme et du citoyen* (The Declaration of the Rights of Man and of the Citizen) [125] was adopted by the ruling government, which included privacy guarantees to its post-revolutionary citizens. Meanwhile, on a wider scale, two highly important developments occurred.

Firstly, in 1948 (December 10th), the General Assembly of the United Nations adopted the Universal Declaration of Human Rights [164], wherein privacy was enumerated in Article 12:

> "No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks."

However, since the Universal Declaration of Human rights was so wide-ranging in scope, it never garnered the international consensus necessary to become a

binding treaty. To solve this problem, the declaration was spilt into two binding covenants - with the privacy guarantees becoming part of the International Covenant on Civil and Political Rights (Article 17) [165] created in 1966, which has been ratified by 149 parties worldwide.

Secondly, in 1950, the European Convention on Human Rights [43] (officially the Convention for the Protection of Human Rights and Fundamental Freedoms) was adopted by most Council of Europe member states, wherein privacy was enumerated in Article 8 (right to respect for private life):

> (a) Everyone has the right to respect for his private and family life, his home and his correspondence.
>
> (b) There shall be no interference by a public authority with the exercise of this right except such as is in accordance with the law and is necessary in a democratic society in the interests of national security, public safety or the economic well-being of the country, for the prevention of disorder or crime, for the protection of health or morals, or for the protection of the rights and freedoms of others.

The history of privacy in its legal age shows that as societies have risen and fallen in their haphazard progression and western society's current incarnation slowly emerged, the importance placed upon the ideal of privacy as a fundamental human right has increased enormously, and the subsequent necessity for its protection has thus also grown and gained in importance. However, the nature of protection of privacy has changed as this legal age of privacy gradually gave way to the current modern ages of privacy. This transition was initiated by the ever-forward march of technology, which has affected both the concept of privacy and the mechanisms of its protection in some major ways.

2.2.3 PRIVACY IN THE TECHNOLOGICAL AGE

The privacy implications of ever-evolving technology are not exactly new: Warren and Brandeis' "right to be left alone" came from a time when privacy was threatened by a new technology that allowed photographs to be included in mass-circulation newspapers. During the 118 years since their article, the problems posed by technology have increased wildly, from such "simple" examples as telephones being introduced (leading to objections being raised that they "permitted intrusion [...] by solicitors, purveyors of inferior music, eavesdropping operators,

and even wire-transmitted germs" [63]) through to far more complex recent technologies; while privacy protection has struggled to keep apace. Tuerkheimer classified these problems into two broad categories: surveillance and personal data protection [163].

Surveillance is the monitoring of behaviour[2], and many new surveillance technologies have posed new and complex privacy issues. Some examples of privacy issues brought forth by these technologies include wireless telematics networks embedded in lamp posts tracking the location of one's automobile at all times [101], automatic toll payment systems based on RFID and automobile number plate recognition [64], facial recognition systems employed widely tracking the location of individuals, national identity cards [137], monitoring of employees in a workplace [93], SMS [82, 117], Instant Messaging [80, 120], location awareness services on mobile telephones [90], blanket coverage of CCTV [22, 127], monitoring systems deployed in the homes of the elderly [17], implanted sensors monitoring the health of the sick, monitoring of computer traffic, email and telephone calls, and governmental surveillance systems such as Carnivore and Echelon [61]. All of these technologies are able to violate an individual's privacy in some major ways, as in extreme cases a person can be monitored 24 hours a day without any awareness as to this fact.

So far, there has been little legislation guarding against privacy violations caused by such surveillance technologies, except in a few specific cases: such as when the basic protections that the Fourth Amendment brings Americans against unreasonable violation was extended in the latter half of the last century to cover developments in electronic surveillance (*Katz vs. United States*, 389 U.S. 479 - 1967)), requiring U.S. government agencies to obtain a court order giving permission to use a wiretap. In fact, however, even this little legislation has been overridden by the "Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism Act", more commonly known as the Patriot Act [170]. This Act expanded the authority of the United States government (more specifically the F.B.I.) to, amongst other things, search telephone and e-mail communications and medical, financial, and other records. They can, through the use of "National Security Letters", do this without a court order. Since its introduction in 2001, several legal challenges have been brought against the act, and US courts have ruled at least one provision unconstitutional. Outside of the U.S. many countries have enacted similar, but weaker, legislation. Thus, legislation is

[2] "Surveillance" comes from the French, meaning literally "Watching Over"

not solving the problem of privacy violation caused by surveillance technologies, and has not even begun to attempt to legislate against private company carrying out such activities.

Personal data protection is the well-known privacy issue created by the proliferation of databases containing information about individuals' lives, habits, preferences, and personal histories. From virtually every place one goes to shop, from the supermarket to the video store to the bookstore to the pharmacy, a steady stream of information about customers pours into vast warehouses of data. These treasure-troves of knowledge can help companies in their drive to ultimate efficiency and therefore the good of the customers; at least, that is the argument used to support this. Undoubtedly this is true. However, it can also give a business a huge unfair advantage in their dealings with the customers as it leaves them holding all the cards in the game of commerce – while most of the individuals playing do not know for sure how to play the game, or even that they are playing at all. The time may come when everything it is possible to know about an individual is stored somewhere, and there are no technological guarantees that this information cannot be accessed by anyone, at any time.

In an attempt to counter this problem, governments in many countries have enacted legislation that specifically attempts to protect the privacy of this data: for example, in the UK and Sweden there is a legal restriction on any entity possessing any kind of personal information without the explicit consent of the data subject, and every entity that does store such data has to register this fact with the government. These types of legislation originated from a set of guidelines called the Fair Information Practices (FIPS), developed in the early 1970s as part of an investigation by the then US Department of Health, Education and Welfare concerned with citizens rights with regards medical records [169].

The FIPS were a set of desirable practises with regards to the processing and storage of data within computer systems, advocating such things as limiting the collection of data to that which is necessary for the application involved and allowing individuals to view the information help about them. This in turn led to the Organisation of Economic Development issuing a set of guidelines (the OECD privacy guidelines) based on the FIPS ideas which set out the minimum standards for data collection, storage, processing and dissemination that both the public and the private sector should adhere to [132]. These guidelines are commonly consulted by nations and private organisations when drafting privacy laws and policies.

One notable collection of privacy laws created based upon the FIPS was created

in Europe: in 1981 personal data became specifically highly protected when the Council of Europe created the *Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data* [44]. This convention obliged the member states of the Council of Europe to enact legislation concerning the automatic processing of personal data, resulting in different types of regulatory models and approaches being adopted [119]. These ranged from the "self-help" approach (where there is no government interference and it is up to the countries' citizens to challenge inappropriate practises and bring them to the attention of the courts) to the "registration" / "licensing" approach (where a government takes full control of ensuring that personal data about its citizens is not misused).

Over the following decade or so, a range of diverging legislation was enacted in EU countries. The European Commission realised that the different, and oft contradictory, approaches and laws enacted by its member states would impede the free flow of data within the EU. Therefore the European Commission decided to build upon their previous work and harmonise data protection regulation across member states, proposing the Directive on the Protection of Personal Data in 1995 (officially *Directive 95/46/EC on the protection of individuals with regard to the processing of personal data and on the free movement of such data*) [60]. All members of the EU had to transpose this legislation into their internal law by 1998, and all did. This directive required that many things become legally mandatory, including the creation of government data protection agencies and the registration of any databases of personal information with these agencies. However, a 2003 EU report concluded that European citizens do not understand what rights they actually have under the data protection legislation [42], partly leading the EU to publish a plan calling for common rules on the form in which companies publish privacy policies in order that they are simpler to understand [57].

One of the stipulations of the 1995 Directive was that personal data would only be allowed to be transferred to non-EU countries if the country provided an adequate level of protection of the data protection. However, while the EU and the United States both ostensibly share the goal of enhancing privacy protection for their citizens, the EU and the US take fundamentally different approaches to achieving this. As has been discussed, the EU approach relies on comprehensive legislation to mandate its corporate citizens to be respectful of an individual's privacy; the US, however, uses an approach that mixes small amounts of legislation with large amounts of self-regulation, relying on its corporate citizens to behave responsibly. This, however, means that it is impossible to show that the US can

provide an adequate enough level of protection, and means that the directive could have significantly hampered many US companies from engaging in transatlantic transactions with European customers. In an attempt to solve this problem, the European Commission and the US Department of Commerce jointly developed what is known as the "Safe Harbor" framework[3], which was adopted in 2000. Under the safe harbor agreement, US companies can choose to register and enter the safe harbor (self certifying annually): agreeing to comply with the agreement's rules and regulations (which includes elements such as notice, choice, access, and enforcement). EU organisations can then ensure they are sending personal information to US companies in accordance with EU rules by checking that the US company they are dealing with is on the list of participating companies.

Other notable examples of things influenced by the FIPS include HIPAA (the US Health Insurance Portability and Accountability Act of 1999) and privacy work around various policies and technologies such as P3P [46] (see Section 3.1.2.1), data management [99], RFID [69], ubiquitous computing [110], video surveillance [56] and biometrics [58].

In this age of privacy, technology has advanced so far and so fast that the approach of protecting privacy through legal means is not as effective as it once was: technological development is far outpacing the ability of the legal system to react and adapt to new developments – and in many cases has overstepped the line where the legal system is able to protect privacy at all. In an effort to keep up, instead of continuing the past legal developments calling for an all-encompassing legal and moral protection of privacy, the legislators have instead had to mitigate against specific privacy violations as they appear – resulting in today's legal landscape containing a mishmash of various legal protections and requirements that help guard against only the occasional isolated pocket of privacy violation.

2.2.4 PRIVACY IN THE INFORMATION AGE

In the information age that we live in the nature of privacy changes in some interesting ways. Moor has discussed how in the information age information is "greased" – quick moving and with uses impossible to imagine when it was initially entered onto a computer [121]. The privacy implications of this one development alone are staggering. It is now possible to give personal information to one entity

[3]http://export.gov/safeharbor

(whether purposefully or without realising it) for one specific purpose or reason, only for it to be transferred to another entity and used for an entirely different (and possibly objectionable) purpose, *ad infinitum*. Moor therefore argues that we need to create "zones of privacy" which allow individuals to control levels of access to private information differently in different situations. He argues that it is important to think of privacy as an amalgamation of the competing ideas that privacy is either about controlling one's information or about the state of limited access to one's information, saying that although the control theory is highly desirable it is impossible to totally control one's information in a computerised society.

Obviously, Moor's conclusion is based upon the assumption that individuals really do care about the privacy of their personal information in this greased world of the information age. However, not everyone believes this to be the case. One point often raised supporting their opinion is that while most individuals state they are concerned about privacy, many of them then go online and give away personal information with apparently no thought or hesitation. This, they argue, seems to suggest that retaining privacy of their personal information online is not as important to people as they like to think, and therefore it can safely be ignored. To counter this argument, Syverson briefly discussed this view and produced some preliminary evidence debunking it in [158], while Acquisti discussed how it is "unrealistic to expect individual rationality in this context", since consumers who wish to protect their privacy "might not do so because of psychological distortions well documented in the behavioral literature", or that they may incorrectly "perceive the risks from not protecting their privacy as significant" [3].

Looking more deeply at the statistics of public views on privacy, a number of studies have been carried out looking at the people's privacy preferences when dealing with commercial entities. Some of the first of these were carried out by Alan Westin, followed by Privacy & American Business (a research company founded by Alan Westin) [136, 180], while more recent work backing up the conclusions of these first studies being provided by the Eurobarometer survey in the EU [59], by Ackerman et al [2], and by Jenson et al [96]. The conclusions of all of the studies are that while consumers in general maintain that they have a high level of concern about privacy, when faced with real life situations these consumers can be split into three main categories termed the *privacy fundamentalists*, the *pragmatic majority*, and the *marginally concerned*. Privacy fundamentalists are extremely concerned about privacy, unwilling to provide private information under almost any circumstances. Pragmatists have privacy concerns, but are willing to give certain private

information for pertinent reasons when assured by privacy protection measures. The marginally concerned are willing to give private information under almost any circumstances. While the distributions of consumers in these three categories has varied over time [182], generally, the percentages seen are around the 15-25% mark of privacy fundamentalists, 40-60% pragmatists, and 15-25% marginally concerned. Thus, while some individuals may reveal private information at any time, the majority of people (some 55-85%) are at least partially concerned about their privacy online. To further this investigation, Tsai et al performed a study whose findings showed that when a consumer had privacy information made easily available, they tended "to purchase from merchants that offer more privacy protection and even pay a premium to purchase from such merchants" [162].

These studies prove that, despite the point of view of some people such as ex-Sun Microsystems CEO Scott McNealy – who once famously said "Privacy is dead, deal with it" – the public themselves are in fact concerned about privacy. However, the greased nature of information makes the privacy of information something very difficult to protect.

2.2.5 THE FUTURE OF PRIVACY

These conclusions suggest that a new model of privacy protection is necessary in the information age if the laudable goal of protecting a consumer's privacy is to be achieved. Instead of relying upon legal protections alone to guarantee that no entity can violate a consumer's privacy, the future of privacy protection seems likely to need to be a mixture of legal protections and technologies that put consumers back in control of their information: thus allowing the consumer to flexibly protect their own privacy to the extent that they wish, releasing certain amounts of personal information for whatever reasons they wish, tailored to the specific circumstance they are faced with.

2.3 SUMMARY

Privacy is a complex and contentious concept that has been debated vociferously ever since its first inception. It is generally recognised as a core human value that is important for many various reasons. Over time, the methods used to protect this core value have changed as the nature of potential privacy violations have changed; and the speed of this change has increased dramatically in the last few centuries

as humanity progressed through the industrial age to the technological age. In this age, most privacy violations against an individual were due to calculated acts by specific entities, without the individual's consent, and the best defence was specific legislation protecting individuals against such acts. However, as we move into the information age, information has become "greased", resulting in privacy violations that can likely be due to entities misusing data that was collected from individuals for apparently legitimate reasons. Thus, protection of privacy becomes a new challenge as individuals have freely given this information away for a specific purpose – and once this information is out in the wild, it is impossible to keep track of, or recapture. One way to answer this challenge suggested by this thesis is to create technologies that put consumers back in control of their personal private information. The next chapter explores this conclusion in more detail by examining methods of protecting privacy in the context of e-commerce in the information age.

Chapter 3

Protecting Privacy in e-Commerce

The previous chapter examined the idea of privacy and its development in a general sense. Since the focus of this thesis is on addressing the problem of using a consumer's preferences to enhance e-commerce search techniques while preserving privacy and reducing exploitation, this chapter will explore the idea of privacy and privacy protection in the more specific area of e-commerce. This will help to understand where the contributions of this thesis fit within the range of current research in two relevant areas – that of Privacy Enhancing Technologies, and that of Preference Searching.

3.1 Privacy Enhancing Technologies

Desire for consumer privacy had led some in the research community to design technologies that aim to uphold the ideal of protecting this privacy in the e-commerce environment. These technologies can generally be placed into one of two categories: anonymous technologies and non-anonymous technologies; or to use the correct but oft-ignored antonym of anonymous – "onymous" technologies. The methods by which the technologies of these two categories attempts to preserve individual privacy differ both in fundamental philosophy and in application.

3.1.1 Anonymous/Pseudonymous Techniques

Anonymising/pseudonymising technologies attempt to achieve unlinkability between a consumer and any of their personal information. That is, they aim to

secure the privacy of a consumer's personal information by simply trying to ensure that any personal information released to an organisation cannot be linked to a consumer's real identity. Thus, a consumer could make use of preference-enhanced search techniques by releasing their preferences with impunity, safe in the knowledge that the vendor could not link this preference information back to them as a person. These kinds of technologies will not, however, help with the issue of a vendor exploiting the knowledge contained in these preferences to the detriment of the consumer.

There are a range of levels of anonymity available: from the truly anonymous – no one can find out who you really are; through the pseudo-anonymous – your identity is generally not known but can be obtained if deemed necessary and with enough hard work; through to the pseudonymous – where a consumer can create a range of virtual identities for use in various circumstances. Throughout this spread individual privacy is maintained – as although an attacker trying to harvest personal information is able to gather large quantities of it, the material gathered cannot (normally) be linked back to a specific individual.

User anonymity, at whatever level, can be achieved through one of three main methods: anonymising the transport medium; allowing anonymous but accountable transactions (credential systems); and 'scrubbing' of data stored by an organisation.

3.1.1.1 Anonymising the Transport Medium

One method of enforcing anonymity between consumers and their prospective e-commerce vendors is to ensure that any communications between the two occur in such a way that the original identity of the consumer cannot be gleaned through examination of any of the communications, or by eavesdropping on communication patterns. So for example, when a consumer browses a vendor's website and buys items from them, this type of technology will aim to prevent the e-business from ever knowing exactly with whom they are dealing. Technologies have been created that attempt to achieve this goal, with varying degrees of success.

One of the simplest ways for a consumer to achieve anonymity in this fashion is to set up an account with a free email service such as Hotmail[4] (now Windows Live Hotmail), Yahoo Mail[5] or Google Mail[6]. This allows a consumer to have

[4]http://www.hotmail.com/
[5]http://mail.yahoo.com/
[6]http://www.gmail.com/

a point of contact for communications – enabling them to be members of online communities, have accounts with vendors, etc – whilst retaining anonymity as there is no easy way for anyone to link such an email account to the individual who controls it. Hushmail[7] takes this idea further offering end-to-end encryption for its users ensuring eavesdroppers cannot breach the privacy of this setup. This simple approach to anonymity through free email systems however requires that the consumer trust the email service provider completely – that they will not log communication details such as IP addresses during email sessions. Additionally, this approach is fast becoming impossible in recent times as the majority of these services increasingly require personal details to sign up. That being said, there is generally nothing to stop consumers signing up with false details.

A step up in technological complexity leads us to a well known tool to achieve anonymous web browsing – Anonymizer[8]. When a consumer uses this service to view items on the internet or submit information to remote sites, it is done with all communications being routed through Anonymizer's servers – thus the remote site has no way of detecting the IP address or identity of the consumer. However, this kind of technique also requires a trusted third-party – in this case Anonymizer itself. This is simply because Anonymizer's servers (or the user's ISP) can certainly identify the consumer if they so desired.

To achieve complete anonymity on the internet, tools are needed that do not rely on a trusted third-party. In this vein, Reiter and Rubin created a system called Crowds [142, 143] that operates by grouping consumers into large groups (crowds). Consumers connect to this crowd, and instead of directly issuing requests to internet servers, they give it to the crowd. The request gets passed around the members of the crowd randomly until it eventually gets submitted to the intended recipient. To use Crowds is essentially to play 'pass the parcel' with users' requests. The recipient of the request thus cannot identify who in the crowd issued the initial request – as it is equally likely to have been any of the members of the crowd. However, malicious behaviour by any rogue crowd members can affect the usefulness and reliability of the system – although they cannot compromise the anonymity of any of the other members through such behaviour.

Another step up in complexity of technology leads to technologies that use encryption to assist with solving the problem. A well known technology of this type that spurned numerous off-shoots was created by Chaum in 1981 [32]. A

[7] http://www.hushmail.com/
[8] http://www.anonymizer.com/

Chaum Mix is a system based upon public key cryptography that allows people to communicate via email while remaining anonymous to each other (and any global eavesdropper), all without needing any guarantees as to the security of the underlying communications system. It does this by ensuring that messages passing through the system are of equal size, cryptographically changing them and then sending the messages to their recipients in a different order. This makes it very difficult for even a global eavesdropper to link an incoming message and its sender to an outgoing message and its recipient. Chaum Mixes can be improved by linking mixes together to create a 'cascade' of Mixes in order to further provide security guarantees and not to have to require a person to trust one single mix server.

One of the off-shoots of Chaum Mixes was created by Goldschlag, Reed and Syverson [75, 76]. They designed an anonymous communication architecture called *Onion Routing* which can be used by any protocol capable of being adapted to use a proxy service. It is built upon the idea of using a network of dynamic real-time Chaum Mixes. Onion routing allows bi-directional, near real-time connections that are highly resistant to eavesdropping and traffic analysis. The user submits an encrypted request in the form of an onion – a layered data structure specifying the properties of the connection at each point along the route (including cryptographic information and keys). Each point can only decrypt its layer, finding out only where the next point in the route is, except for the final point which decrypts its onion to find the request to send and whom to send it to, which it then does. Thus, a recipient only knows the identity of the person at the end of the chain. This first generation of Onion Routing, however, never developed much beyond a proof-of-concept system that ran on a single machine. The Freedom Network [14, 21] designed and operated by Zero-Knowledge Systems Inc. was a commercial implementation of a variant of Onion Routing, routing IP connections through intermediate nodes to provide users with anonymity. This was, however, a commercial failure [74].

In 2003, Acquisti, Dingledine, and Syverson examined the current state of decentralised anonymity infrastructures, identifying the then current issues that were limiting the deployment of such systems and drawing conclusions about what the next generation of these systems should do in outline form [4]. This partly led to a new second generation Onion Routing system (supported by the Electronic Frontier Foundation[9]) called Tor[10] (The Onion Router) being presented [52]. This new generation of communication service added many useful features designed to

[9]http://www.eff.net
[10]http://www.torproject.org/

make Onion Routing secure, efficient, and usable enough for real world use. It still, however, lacked an user interface good enough to enable large scale adoption, leading to a "grand challenge" being issued to develop such a user interface[11].

All of these anonymising technologies however have one general major caveat - they only work properly (provide anonymity) if certain conditions are met. Some of these conditions have been noted by Clayton, Danezis and Kuhn [41]: for example, attacks can be made that will reveal a supposedly anonymous individual's IP address by doing such things as making use of client-side scripting, sending images which can be tracked when loaded by the user (web bugs), cookie stealing, and many other methods. Clayton *et al*'s conclusion is that it is important to not only think about the anonymity properties of communication channels, but to also consider ways of protecting anonymity throughout the entire system. To use the old adage, 'A chain is only as secure as its weakest link'.

3.1.1.2 Credential Systems

Besides anonymising the transport medium, another method of enabling anonymity between a customer and a vendor is through the use of a credential system (sometimes referred to as a pseudonym system). In a credential system, consumers are known to the vendor they are doing business with only by a pseudonym (or *nym*). A single consumer can use different pseudonyms with different vendors, and these cannot be linked together by any member of the system. However, a vendor can issue a credential to a pseudonym, who can then prove possession of this to another vendor revealing *only* that the consumer owns a credential. A certification authority sometimes plays an important role in guaranteeing that the system is used properly and that users can be trusted. What this means in e-commerce transactions is that these technologies enable consumers to buy items from an e-business by proving certain facts: that they are eligible to buy the item, that they have given a payment, that they are old enough to buy an age-restricted item, etc., all without the vendor knowing exactly with whom they are dealing.

The idea and basic framework of credential systems were first introduced by Chaum in 1985 [30], who soon after published a full model with security proofs using RSA as a one-way function [31]. This model, however, requires a trusted third party (the certification authority) which manages the transfers of consumers' credentials between organisations. To relax this constraint, Damguård developed

[11] http://tor.eff.org/gui

a model that only needs a third party to be involved - it does not necessarily have to be trusted [49]. It does this by using the idea of multi-party computations [77], resulting in a scheme that is provably immune to forgery by malicious entities in the system. This model however is not meant for implementation as some of the methods used are too inefficient for heavy use. A practical version of this work was produced by Chen [33]; however, it requires that the certification authority behave honestly, thus negating the advantage of Damguård's approach.

One weakness of all of these models of credential systems is that there is nothing to stop a consumer from sharing his pseudonym and/or credentials with other consumers. For example if a consumer was issued with a certificate proving that they are over 18, they could then share this with an under-age consumer who could use it to purchase age-limited products. While this is in fact acceptable from the perspective of protecting privacy, it may result in business transactions that are not valid or intended (and possibly illegal). In an attempt to solve this problem, Lysyanskaya, Rivest and Sahai produced a model [114] that includes the presumption that the consumer's private key (linked to the corresponding public key in the credential system) is something that they are motivated to keep secret – for example, it could be their digital signature key. If a consumer then shared a credential with another consumer, the other consumer would have access to this secret. If the secret was the digital signature key, the other consumer could then forge signatures on any documents in the original consumer name. This idea was termed *non-transferability*. As with Damguård's model however, this model is not directly usable in practice due to the reliance on methods that are too inefficient to use in practice.

Camenisch and Lysyanskaya further developed this idea of non-transferability [25] producing another model for a credential system that additionally has the optional property of allowing a consumer's identity to be revealed if the consumer misuses their credential or uses it in an illegal transaction. This model, however, requires that the certification authority is trusted to do their job properly – this risk can be minimised, however, through distribution of the tasks of the certification authority, weakening the trust assumptions. Camenisch and Van Herreweghen described a prototype of a system based upon this model [24].

The idea of credential systems, along with the lack of success of PKI (Public Key Infrastructure) systems, has led to the development of PMI (Privilege Management Infrastructures), AAI (Authentication and Authorisation Infrastructures) and FAM (Federated Access Management) systems. As such, several proposals for

working systems and implementations of these proposals have been put forward: such as Kerberos [105, 126], Microsoft Passport (now Windows Live ID)[12], the Liberty Framework[13] [112], and SAML [26]. Work in these areas has been performed by many different communities, notably including the Grid / e-Science community (e.g. PERMIS [28, 50]). However, given the difference in requirements of these communities to those of e-commerce, many do not fit in the world of e-commerce – a view shared by Schlaeger and Pernul who surveyed many of these proposals, concluding that none of them "is perfectly suitable for b2c e-commerce" [149].

3.1.1.3 Privacy in databases

Another area examined by the research community is the area of enabling anonymity of existing sets of consumer information held by a vendor (both personal and transactional). There are two main reasons this may be desirable: the vendor may have a desire to keep such valuable information where individual identification of a customer is not necessary, and removing this possibility saves them from having to deal with privacy, security and data protection issues (e.g. to understand overall purchasing trends over time); and the vendor may wish to pass parts of this information on to third parties without any information being personally identifiable.

Techniques to accomplish this were originally pioneered to help solve privacy issues in statistical databases. In these databases it is the statistical information about the data — rather than the data itself — that is important, thus methods have been considered that can keep the statistics of the data-set valid whilst keeping the individual data itself private. Two excellent surveys of this area were produced by Adam and Worthmann [5] and Shoshani [155]. Broadly, the methods that attempt to accomplish this ideal of privacy of database information can be split into three main categories: *query restriction*, *data perturbation* and *output transformation* [5].

The goal of technologies that fall into the *query restriction* category is to retain privacy of individual data items by restricting the information that can be released. In this approach only queries that obey specific criteria are allowed, in an effort to prevent information about specific data items becoming known. The problem of this approach is that only a small subset of possible queries are allowed, reducing the usefulness of the database. Similar to this idea is query auditing [34], where an audit trail of all queries that have been performed is kept and every new query is

[12] http://www.passport.net
[13] http://www.projectliberty.org

checked for possible compromise. If a possible compromise is detected, the query is disallowed. The main problem with this approach, however, is that an entity can use the results of queries along with the knowledge of what queries have been allowed and denied to infer data and thus compromise the privacy of the data contained in the database.

The goal of *data perturbation* is to modify the original database in such a way that the overall statistics remain valid while individual items are changed, thus preserving privacy of individual records. A very basic method of achieving a limited degree of confidentiality is simple rounding of numerical data. This naive approach can be improved slightly by randomly rounding or by adding random noise with a mean of zero [171]. This idea has been further developed, for example in [6, 7]. Other methods have been proposed that fall into this category, including data swapping, which involves swapping each item in the database with another one from the same distribution, thus creating a new database with supposedly the same statistics [141]. However, by its very nature the process of modifying the original data will always alter the overall statistics at least slightly, as well as making the original data meaningless, making the information gained less useful overall.

The final category of *output perturbation* allows a database to permanently store the original data and perform queries on it, however the *results* of any query performed are altered such that the original data cannot be inferred before being returned to the user. Methods that achieve this include adding a random perturbation to query results (with increasing variance as queries are repeated) [16].

One prevalent example of such a technique in this area is Sweeney's idea of k-anonymisation [157]. The basic idea of this approach is to make each record of a table identical to at least $k-1$ other records over a chosen set of attributes, thus enabling anonymity by ensuring that when a record is released, its linkage to a single individual cannot be identified. This idea has since been refined in many ways (e.g. [111, 113, 115]) and a good survey of this work has been presented by Ciriani et al [40].

All of these techniques have one major technical drawback however – they do not adequately satisfy the conflict of being able to provide high quality statistics while simultaneously preventing disclosure of individual information [5]. Also, attempting to achieve the privacy of individual data itself is a very hard problem to overcome. Denning and Denning discussed the details of this problem – whereby statistical information "contain[s] vestiges of the original information; a snooper

might be able to reconstruct this information by processing enough summaries" [51] – drawing on work both by themselves and Schlorer (e.g. [150, 151, 152]).

From an e-commerce and consumer's point of view there is a big drawback with these methods: any consumers wishing to enter into a business relationship with a vendor have to trust that vendor's promise that once their personal information is received and stored it will be anonymised. Given the public view on privacy in e-commerce as discussed in Chapter 1 this trust may not be something easy to come by for many vendors, and any breach of trust by any vendor has the potential to destroy any trust built up between the public and all vendors.

3.1.1.4 Summary of Anonymous/Pseudonymous Techniques

A summary of anonymous/pseudonymous techniques surveyed is presented in Table 3.1, where we compare each of the techniques by their basic architecture, indicate whether they are usable in practical circumstances, and indicate in which application areas they can be used successfully. As can be seen, the majority of the anonymous/pseudonymous privacy enhancing techniques are practically usable in areas involving web applications (which includes e-commerce), but require a trusted third party in their architecture. This is potentially a serious limiting factor for the usefulness of this type of technology: it requires that entities exist that consumers trust entirely (and all the problems this entails); also as e-commerce develops further and e-businesses become ever more distributed, it may not be realistic to require the existence of a trusted third party in order to protect privacy for consumers.

Of course, there are other privacy-enhancing techniques available. One such area of techniques is the area of steganography [98, 100] – the science of hiding communication between two parties in such a manner that an eavesdropper would not know that a message exists. However, while these techniques are highly developed, none are realistically applicable to the field of e-commerce as it stands today (e.g. in the case of steganography, hiding the fact that an individual e-commerce transaction was occurring from an eavesdropper would only guard against privacy invasions from that eavesdropper – which is not where the types of privacy invasion that this thesis is interested in is occurring).

So, where anonymous/pseudonymous techniques are both desirable and technically plausible, they are an excellent method of preserving the privacy of a consumer's personal information, including information about their preferences.

However, such techniques cannot help at all with the problem of exploitation of a consumer's preferences.

3.1.2 ONYMOUS TECHNIQUES

Preserving privacy by staying anonymous is not always possible. Sometimes a consumer is *required* to be identified by a vendor in order to receive their services. Examples of cases where this may be true are the use of digital libraries where payment is required to access the data (and therefore credit card details, and thus identity can be revealed) or standard Business-to-Consumer (B2C) e-commerce – if a consumer orders a book from Amazon, payment details as well as delivery details are required. While anonymous e-cash systems have been developed that have the potential to solve the first problem, any material goods one orders still need to be delivered: therefore an address needs to be supplied, therefore identity can be discovered. Solutions to this problem have also been proposed – a simple example being that a consumer set up a PO Box. However, none of the solutions to this problem developed so far would realistically work in the real world of mass e-commerce, where the average individual would simply not bother engaging in e-commerce if it meant having to go well out of one's way to retain privacy by enacting one of the solutions.

Onymous technologies have started to be developed to counter this problem. The main philosophy behind onymous technologies is not to attempt to withhold a consumer's identity from a vendor, instead attempting to help consumers preserve the privacy of some of their information – or at the very least to help them to make informed decisions about which entities can be trusted. This area of technology is becoming increasingly important in recent times as more resources on the internet slowly move away from the free-service model and as security concerns are pressing for fully accountable and identifiable transactions.

Additionally, the enabling of a consumer to either minimise the amount of personal information released to a vendor, or to only release such information to those vendors it has made an informed judgement about, allows for that consumer to not only minimise loss of privacy but also has the potential for helping with the second problem addressed in this thesis; that of minimising exploitation of personal information.

Onymous technologies fall largely into two distinct groups: those that help consumers to make informed decisions when transacting in e-commerce, and those

Table 3.1: Anonymous/Pseudonymous Privacy Enhancing Technologies

		Architectural Requirements			Practically Usable	Application Areas		
Technology Type	References	Trusted 3rd Party	3rd Party	Decentralised		Email	Web	Other
Anonymous Techniques:								
Anonymous Email		X			X	X		
"Anonymizer"		X			X		X	
"Crowds"	[142, 143]			X	X		X	
Simple Chaum Mix	[32]	X			X	X		
Network of Chaum Mixes	[32]			X	X	X		
"Onion Routing"	[75, 76]			X	X		X	X
Credential Systems (CS):								
"Mixnet"	[30, 31]	X			X		X	
Damguård's CS	[49]		X				X	
Chen's CS	[33]	X			X		X	
Lysyanskaya *et al*'s CS	[114]	X					X	
Camenisch *et al*'s CS	[25]	X			X		X	
"Idemix"	[24]	X			X		X	
Database Privacy:								
Query Restriction	[34]	X			X	N/A	N/A	N/A
Data Perturbation	[6, 7, 141, 171]	X			X	N/A	N/A	N/A
Output Perturbation	[17]	X			X	N/A	N/A	N/A

that actively attempt to actually enforce the preservation of privacy.

3.1.2.1 Decision Helping Techniques

One of the simplest methods of helping maintain consumer privacy onymously are those which aim to guide a consumer in the decision about which vendors can be trusted to be (relatively) respectful of their privacy and which to avoid. Various certification programmes exist that implement this idea; Anton and Earp summarised these programmes [9]. One of the more famous examples is TRUSTe, the 'online privacy seal'. TRUSTe is simply a programme designed to help an individual make a choice over which websites they can trust and enter into business with. The TRUSTe organisation issues a 'trustmark' to e-businesses that adhere to TRUSTe's privacy principles – practices approved by the U.S. Department of Commerce and the Federal Trade Commission – and which allow oversight to ensure that they follow through on their promises. Moores and Dhillon [122] conducted a study into the effectiveness of these methods and concluded that while they can be effective for the organisations that participate in these programmes and abide by their stated privacy principles, the overall perception of trust in e-commerce is still heavily damaged by the majority of organisations that do not participate. Thus, this solution to the trust problem of e-commerce is not effective enough to make a significant difference.

A more technological solution of the decision helping techniques proposed is the Platform for Privacy Preferences (P3P) [45, 140], a World-Wide Web Consortium (W3C) specification from Cranor et al. An overview of the history of P3P was presented by Hochheiser [87]. Its main philosophy is that if individuals *have* to give up some privacy in order to transact with an e-business, they should be able to at least make an *informed* choice as to which e-businesses they wish to interact with. To achieve this, P3P enabled e-businesses make available their P3P policy - a set of privacy practices that their website and company adhere to. P3P enabled individuals create their own policy, deciding what privacy practices they find acceptable. These two items come together when a user visits an e-business' website, where a P3P agent under the individual's control compares the two policies, informing the individual about their similarities and whether they match. This process allows individuals to tailor their relationships with different e-businesses, releasing different amounts of personal information accordingly. However, P3P has not seen wide adoption – in 2006 only 21% of e-commerce sites from a sample

taken from Froogle[14] had a P3P policy [53]. Also, P3P has one main drawback – an e-business' P3P policy only states what their policies are, it does not ensure these policies are actually enforced. This lack of enforcement, alongside a lack of motivation amongst e-businesses [87] and a lack of a good user interface to involve the user in the decision making process [1] are the suggested main reasons for this lack of adoption.

3.1.2.2 Enforcement Techniques

In an attempt to counter the main drawback with P3P, Ashley, Powers and Schunter created an extended variant of it which works towards enterprise-wide enforcement of P3P policies [12, 13]. Organisations create an "Enterprise Privacy Policy" which is then enforced by a protected system holding the consumer's data. The system grants or denies attempts to access information and creates an audit trail that can be requested by the consumer. While this is a good step forward for P3P enforcement – ensuring that employees of the company can only access a consumer's data for agreed upon reasons – the consumer must still assume that the company has indeed protected its system, and therefore the company as a whole is trustworthy.

A more consumer-centric privacy-enforcement solution was presented by Elovici, Shapira and Maschiach [54]. They presented a model for hiding information about group interests of a group of individuals who share a common point of access to the internet. The model works by generating faked transactions in various fields of interest in an attempt to prevent the real group profile being inferred. The raison d'être for the model is to allow individuals within the group to identify themselves to, and thence make use of, various services – such as digital libraries, specialised databases, etc. – without allowing eavesdroppers to infer a common group interest. This could be used for example to prevent someone inferring the direction of research within research groups in rival companies. The measure of the model's success is based upon measuring the "degree of confusion" the system can inflict upon eavesdroppers.

Another technological solution to preserving privacy by enforcement was presented by Rezgui, Ouzzani, Bouguettaya and Medjahed [144]. Their system concentrates on preserving privacy of a citizen's personal information in web-services in general, and in e-government applications in particular. Its main thrust is in

[14] http://froogle.google.com/

enforcing privacy by requiring entities which wish to access data to provide credentials proving that they are allowed to access it, filtering out data that they are not allowed to access, and finally by delivering the data through *mobile privacy preserving agents* which enforces the privacy of the data on the remote site. However, the security of mobile agents is a major problem that has not yet been addressed adequately (see [144] for an overview of this problem), consequently any solutions that include the use of mobile agents cannot currently contain any security guarantees.

3.1.2.3 Summary of Onymous Techniques

Onymous techniques can be very useful in e-commerce as they support the ideal of consumer privacy — attempting to maximise the amount preserved — whilst still allowing fully onymous verifiable transactions to occur between consumer and e-business. Compared to anonymising technologies there have been relatively few technologies of this type proposed – possibly because realistically usable methods of maintaining consumer privacy onymously spring to mind less readily than their anonymous counterparts. A summary of the techniques surveyed is presented in Table 3.2, where we compare each of the techniques by their basic philosophy (whether they are decision helping or enforcement), whether they require a trusted third party, and whether they can be applicable to the general area of e-commerce. As we can see, with the relatively few onymous privacy enhancing technologies so far proposed, only one actually fulfils the ultimate goal of enforcement of consumer privacy in e-commerce without requiring a trusted third party to operate (outside of the work by this author): the technology proposed by Elovici, Shapira and Maschiach [54]. More work in this area is clearly desirable.

3.1.3 SUMMARY OF PRIVACY ENHANCING TECHNOLOGIES

Current privacy enhancing technologies are making significant achievements in the arena of preserving the privacy of a consumer and their personal information through achieving user anonymity. In the cases where anonymity is possible (and indeed desirable), and the caveats mentioned are managed satisfactorily, this is a very viable and effective approach and definitely merits the attention it currently receives. However, more work is required into the area of designing systems that can manage these caveats automatically, and also into relaxing the requirement of trusted third parties in their architectures. This will then provide total

Table 3.2: Onymous Privacy Enhancing Technologies

		Privacy Philosophy			
Technology	References	Helper	Enforcement	Requires Trusted 3rd Party	Applicable to E-Commerce?
"TRUSTe"	[122]	X		X	X
"P3P"	[47, 140]	X		X	X
"E-P3P"	[12, 13]		X	X	X
ESM's	[54]		X		X
ROBM's	[144]		X	X	

anonymity/pseudonymity, rather than just concentrating on the basic methods and protocols that provide anonymity at the base level.

It is not always possible, however, for a consumer to remain anonymous. This is especially true in the area of e-commerce – where names, payment details, delivery details, etc. are required when one buys material goods over the internet and enters into a contract with that vendor. This, coupled with the fact that such technologies cannot help guard against exploitation of a consumer's personal information, leads to the conclusion that onymous technologies are needed to allow an individual to preserve their privacy whilst simultaneously fulfilling the practical necessity for fully accountable and identifiable transactions.

So far, in this area of preserving privacy in onymous circumstances only a few technologies have been produced: the main contender (at least as far as take-up is concerned) being the Platform for Privacy Preferences (P3P). However, this technology does not actually enforce individual privacy, it is a technology to help guide decision making about whom to trust. More technologies, and specifically technologies that attempt to enforce the preservation of privacy, are required.

If it is accepted that there is a need to produce new and more advanced onymous technologies then a need develops to understand fundamentally how consumer privacy in the e-commerce context may be *measured.* That is, for example, whether some types of information are worth more than others, whether there is any correlation between what businesses and individuals value the most, and if certain kinds of private information hold more economic power (and thus bargaining power). Establishing these measures will then give a sound basis for developing onymous technologies, and allow full evaluation and comparison of any technologies that may be created in future.

Finally, there is a need to study privacy enhancing technologies in the light of security requirements. Allowing any individual to transact anonymously on the internet is sometimes considered a potential security problem in today's world. However, in the specific context of e-commerce, it is easy to argue that any individual who truly desires to be anonymous can simply choose to not transact electronically and instead go into a physical shop in person and pay with cash. Thus, stopping anonymous e-transactions may not get rid of this potential security problem; it only removes it from e-commerce shifting it to a different area. More studies are needed in order to understand how the two sets of requirements can be meaningfully balanced in e-commerce; onymous technologies seem possible to play an important role in viable solutions.

3.2 ENHANCED SEARCH TECHNIQUES

The idea of using an individual's preferences to help with searching information is becoming increasingly common in information systems in general; the use of this idea in the world of e-commerce is a good example of this trend. In this particular circumstance preferences are mainly used to filter and personalise information before it is presented to an individual.

Conceptually, the work required to enable preference-enhanced searching can be split into two areas: the initial act of obtaining preference information, and the subsequent act of using this information to help computational problems. While the first area naturally leads to the second, both can be viewed as essentially disparate problems.

3.2.1 OBTAINING PREFERENCES

The initial obtaining of consumer preference information by a vendor can happen in one of two ways: via explicit expression or implicit collection. Each of these are explored in more detail next; however, irrespective of which method is used, the very fact that personal information has been obtained assuredly leads to questions around privacy issues caused by the storage and usage of such information. To mitigate against this, such personal information needs to be treated carefully by all parties.

3.2.1.1 Explicit Expression

The most obvious way for a vendor to obtain a consumer's preferences is for the vendor to explicitly collect relevant preference information directly from a consumer in what can be considered a just-in-time manner – by essentially simply asking the consumer to express their preferences manually to the vendor.

Examples of this method are diverse. One approach is to simply ask a consumer to enter preference information relevant to a specific domain on a web page that essentially represents a preference-enhanced search form. An example of this is the interface employed by Kießling and Köstler, used to demonstrate their preference search technology called Preference SQL [104] where the consumer was simply asked to describe various aspects of their perfect used car in a manner fairly similar to existing standard search webpages. Another example of this type of approach

that is commonly seen is the style of interface deployed on dating websites, where an individual first describes themselves, then expresses various preferences of the type of person they are looking for [62]. This style of explicit expression of preferences has the advantage that consumers should be able to get to grips with the interface quickly due to the similarity to existing search interfaces, and that it is fairly intuitive for consumers to understand how their preferences will be used. A major issue with this approach, however, is the fact that an upfront investment in both time and effort is required of the consumer to perform the process of expressing their preferences – and the more preference information is required, the longer it will take to express it. Another, more esoteric, issue that is sometimes encountered is that in certain contexts (such as expressing information and preferences on dating websites) people have the tendency to lie (or at least exaggerate) [84]. So, depending on what use is going to be made of these preferences, the degree of "trust" that can be placed upon this information may vary considerably.

Another, less direct, approach to explicitly gathering preferences is for a vendor to ask a consumer to rate each item they purchase. Given enough ratings, a prediction of that consumer's preferences can be made by examining common properties of different items with similar ratings. The more ratings that are made, the more accurate the prediction may be. This is a form of what is known as "relevance feedback" [146]. It could be argued that this is in fact a form of implicit collection of preferences, since the consumer is not directly expressing their preferences, it is included here as an explicit collection method since the consumer is explicitly driving the preference gathering mechanism and will be aware of the consequences of their actions. While this does not require the upfront investment in time that the more direct approach requires, it does require a continuing dialogue between consumer and vendor – and due to this slow process of elucidating their preferences, the consumer may not see the results of their effort for some time.

In general, explicit expression of preferences is a simple but effective approach to collecting a users preferences in a manner that enables the consumer to be aware of the collection – and therefore make an informed choice to provide the information in exchange for an enhanced e-commerce experience. However, if a large amount of accurate preference information is required, the consumer needs to invest substantial time and effort in order to provide this information – and this requirement may require more time and effort than many consumers are prepared to give.

3.2.1.2 Implicit Collection

An indirect way for a vendor to obtain a consumer's preferences is to implicitly infer a particular consumer's preferences given their previous recorded behaviour. This will include things such as past purchase history, past search history, etc. This style of preference collection is a form of data mining called usage mining. Some good overviews of the whole area of data mining most relevant to e-commerce were presented by Madria et al, and by Van Wel and Royakkers [116, 177].

A typical example of the kind of information that can be made of through usage mining would be the vast data warehouses of purchase history operated by supermarkets for those customers with loyalty cards. Whenever such a customer buys a set of items and has their loyalty card scanned, the supermarket can associate those purchases with all past purchases of that person. To give an idea of the potential scale of this information, a report in 2006 indicated that the data warehouse for the Sainsbury's supermarket chain in the UK, holding the previous two year's sales information, included 15 billion individual item sales across 1 billion transactions from 22 million customers[15]. This represents a staggering amount of information about the preferences of many millions of UK consumers.

A more complex implicit method of gathering preferences was introduced in the area of "collaborative filtering", whose basic idea is that consumers who have expressed similar preferences in the past are potentially likely to have similar preferences in the future. Thus, a vendor can predict a consumer's preferences based both on that person's own explicitly expressed preferences and that of other consumers with similar preferences. How collaborative filtering derives preferences is discussed further in Section 3.2.2.2.

In general, implicit collection of preferences is an effective way of inferring a consumer's preferences without requiring that the consumer make any extra effort. There are two main downsides of this approach, however. Firstly, accurately inferring a consumer's preferences can be a very complex task requiring substantial effort on behalf of the vendor. Secondly comes a major ethical issue – since a consumer is not directly involved in the process they may not be aware that such information is being collected, or of how it could be potentially used. Thus, they may have no opportunity to either consent to, or withhold consent from, the process.

[15] http://www.retailtech.nl/rubrieken/?id=274&page=detail

3.2.2 COMPUTING WITH PREFERENCES

Once a consumer's preferences have been obtained by a vendor, the next step is for that vendor to make use of this information for various computational purposes that attempt to enhance the e-commerce experience of the consumer. Again, this can be split into two main categories: explicit and implicit use of preferences. Each of these are explored in more detail next; however, irrespective of which method is used, the usage of consumer preference information potentially leads to issues centred around the exploitation of such information – and possibly in a manner deemed unacceptable by the average consumer. This kind of exploitation can include that discussed in this thesis.

3.2.2.1 Explicit Usage

The first main way for a vendor to use a consumer's preferences, however gathered, is to explicitly use the preference information to help enhance specific consumer-initiated operations – the main such operation being a search operation. Work in this area has largely centred around defining frameworks for embedding preferences directly into query languages to enable efficient preference based search. The work can be split into the two main approaches - *quantitative* and *qualitative* based areas.

The *quantitative* approach specifies preferences between tuples of a database in an instance of a relation using *scoring functions*, whereby a numeric score is associated with each tuple of a query result set. Examples of this approach include Agrawal and Wimmers' work [8], Chomicki's work [36], and Hristidis et al's work [88]. The scoring function is defined according to the user's stated preferences. If the score of a tuple t_1 is greater than the score of a tuple t_2, then t_1 is the preferred tuple. If two tuples had the same score then no preference between the two would be defined. This approach lays the necessary groundwork for allowing a preferred ordering of tuples within a database to be understood; it does not however provide a mechanism of associating a consumer's preferences with the tuples such that a preferred ordering can be achieved.

The *qualitative* approach fills the gap left by the quantitative approach. Here, preferences between tuples are specified more indirectly, usually using *binary preference relations* which act upon attributes of tuples, where the preferences used are those gathered by the vendor. Chomicki discussed how preference relations can be defined in one of two main methods: using logical formulas, or using spe-

cial preference constructors [38]. Examples of the former include Chomicki's own work [36, 37]; examples of the latter include that of Kießling [102].

Whichever method is used for defining the preference relation itself, the mechanism by which a preference relation is embedded into a relational query language (such as SQL) is through the use of relational operators which extend the base language specification. The preference relation is expressed in these relational operators, which are used by the DBMS to select the most preferred tuples from its dataset. Several different implementations of this approach have been proposed, such as Chomicki's *winnow* operator [36, 37], Kießling's *BMO* (Best Match Only) operator [102, 104], Torlone Ciaccia's *Best* operator [161], and Borzsonyi et al's *skyline* operator [20].

Since preference relations are incorporated into a relational query language through the use of new operators, optimisation and evaluation of queries can occur both by usual such methods and by new additional techniques that specifically target this new type of information. Examples of such new techniques include algorithms for evaluating general preference queries [37, 72] and algorithms for optimising preference queries [15, 20, 35, 38, 39, 72, 107, 133, 184]. Rizzi recently outlined the main research issues faced by the community in handling user preferences on OLAP cubes [145].

The usage of preferences for operations specifically requested by the consumer, such as performing a preference-enhanced search of the vendor's catalogue, represents a usage of preferences that is easy for the consumer to understand the implications of, and exactly how their preferences are being used. The majority of relevant work in this area has concentrated on the direct embedding of preference information into relational query languages. While this means that preference-enhanced search can now be done in a relatively efficient and effective manner, to enable this the consumer must give all of their preference information to the vendor.

3.2.2.2 Implicit Usage

A less direct way that a consumer's preferences can be used by a vendor can be viewed as an implicit use of preferences, meaning that the preferences are used for vendor-initated e-commerce operations (rather than consumer-initiated), i.e. any kind of personalisation that the vendor performs without an explicit request from the consumer.

A typical example of such an implicit use of preferences is a "recommender system". The idea behind these types of systems is to use information, including preference information of a particular person, to help recommend any items that a person may be interested in from a group of items so large that a person is unlikely to be able to manually find the items themselves. Examples of this include messages of interest in Usenet newsgroups, web sites on the internet and objects for sale on an e-commerce site. In the latter example the vendor may – unprompted by the consumer – display items that they think the consumer may have an interest in. Recommender systems are a technology widely adopted in e-commerce, the most famous of which is probably that used by Amazon; indeed, Jeff Bezos, CEO of Amazon.com, is widely quoted as saying "If I have 3 million customers on the Web, I should have 3 million stores on the Web" – meaning that every single web page displayed to the consumer should be generated via input from their recommender system.

The first collaborative recommender system, called Tapestry, was created by Goldberg et al [73], who coined the term "collaborative filtering" (collaborative meaning that the information used to identify items that may be of interest comes from a group of people, rather than just the person involved). Tapestry was designed to enable a person to find documents in a large document store; previous comments by other people were used as the source of information. While a successful demonstrator, it had a few major issues: it only worked with small groups of people, and to use it people had to enter very specific queries – somewhat defeating the purpose of collaborative filtering. Many systems based upon the idea of collaborative filtering followed; notable initial examples include GroupLens [106] and PHOAKS [159], both recommender systems designed to help users navigate Usenet newsgroup postings. Looking more specifically from the e-commerce perspective, Schafer, Konstan, and Riedl present a good review of the various commerical recommender systems [148] that were available at the height of the dot com era, while more recently Lin discussed recommender systems as one of several primary technologies important to e-commerce [95], concluding that "although recommender systems are clearly useful to buyers, they are especially valuable to sellers".

The implicit usage of preferences represents a manner of usage that could change the way e-commerce vendors interact with consumers and in fact – given examples such as Amazon.com – it could be easily argued that this has already happened. Its main use, however, seems to be in helping vendors sell items to consumers that the consumer was not particularly looking for when they started their interaction

with the vendor; rather than helping the consumer when they are looking for a specific type of item.

3.2.3 SUMMARY OF ENHANCED SEARCH TECHNIQUES

There are many approaches to both collecting, and subsequently using, a consumer's preferences. However, the collection of preferences always potentially leads to privacy issues centred around the fact that the vendor has possession of such personal information; while the usage of preferences always potentially leads to exploitation issues centred around the fact that the vendor can make use of this preference information to the point that the consumer may consider it an abuse of their trust – if they ever found out. The first of these problems can be alleviated by making sure that consumers know that their preference information is being gathered and for what purpose it is to be used. The second of the problems can be managed either by the vendor stating upfront to what purposes the consumer's preferences will be used, and sticking to their promise, in some verifiable manner (e.g. taking part in a certification process like TRUSTe, as described in Section 3.1.2.1), or by using technology that aims to limit the amount of preference information sent to the vendor, thus minimising the amount of privacy loss and exploitation that could take place.

3.3 SUMMARY

This chapter has examined two main areas of work related to this thesis: existing Privacy Enhancing Technologies (PETs), and existing ways and means of obtaining a consumer's preferences and using them to enhance the e-commerce experience.

The main method used by PETs in their aim to achieve protection of a consumer's privacy is to enable a consumer to conduct anonymous or pseudonymous transactions with a vendor; thus, it does not matter what preference information is released since the vendor cannot link this information back to a specific consumer. However, it was shown that this basic idea does not work in certain circumstances, such as when a consumer is interacting with a vendor and needs a physical item to be delivered to a real address that is linkable back to the consumer. In those cases, a different type of PET is needed: a non-anonymous ("onymous") technology. These are designed to enhance a consumer's privacy not by remaining anonymous but by minimising the amount of information that is given to a vendor.

Obtaining and using a consumer's preference can happen in a variety of ways. No matter which method is used, however, the collection of a consumer's preferences assuredly leads to questions around privacy, since the vendor has knowledge of a key part of an individual's psyche which many individuals wish to protect, while the actual usage of a consumer's preferences to help with computational problems such as search potentially leads to questions around the possibility for exploitation – whether this information is being used above and beyond what the consumer would deem reasonable. Thus there is an obvious link between a consumer's preferences and privacy: since a consumer's preferences are a conceptual representation of a key part of their psyche, they represent information that the consumer may or may not wish to keep private.

Drawing these two conclusions and the conclusions of the previous chapter together, it can be seen there is at present a gap in knowledge – the lack of a technology that is designed to control the release of a consumer's preferences to allow preference-enhanced search in e-commerce in a privacy aware manner. The next chapter takes this conclusions and examines the area of preference-enhanced searching in detail, outlining what characteristics this onymous PET should have.

Chapter 4

Privacy Loss and Exploitation when Preference Searching

This chapter examines issues concerning the use of using a consumer's preferences to enhance the search process in e-commerce in detail. A model underpinning current e-commerce search is first discussed; this is then used to provide a precise description of the process of preference searching. Given this detail, the problems inherent in the design of current implementations of preference searching and the consequent implications for consumer privacy and the possibility for exploitation are analysed. Finally, a new model that aims to help mitigate against these problems is introduced.

4.1 e-Commerce

Pick up any newspaper, trade magazine, or academic journal, and chances are that some of the content focuses on e-Business or e-Commerce. But what exactly *are* e-Business and e-Commerce? While the classic terms "business" and "commerce" can usually be used interchangeably, their modern derivations "e-Business" and "e-Commerce" cannot. One of the most cited definitions of e-Business is that of IBM, one of the key players in the evolution of e-Business, who described it in 1997 as "the transformation of key business processes through the use of Internet technologies" [29]. e-Commerce is commonly defined as a subset of e-Business – the part dealing with the buying and selling of products and services by businesses and consumers over the Internet.

e-Commerce plays an important part in a modern economy. Looking at the latest figures available at the time of writing, those of the year 2007, e-Commerce

sales specifically were estimated to be worth UK£163 billion to the U.K. economy and US$124 billion to the U.S. economy (note that these figures cannot be directly compared due to different statistical analysis methodologies), growing at a rate of 30% in the U.K. and 20% in the U.S., as compared to the previous year [129, 130, 131, 167, 168]. While e-Commerce retail only represented a fraction of total retail sales (7.7% in the U.K. and 3.2% in the U.S.), the fact that e-Commerce growth rates have been outpacing total retail growth rates, coupled with the fact that 15% of UK businesses had sold goods online in 2006 – which was over double that of the 6.9% figure seen just 4 years previously in 2002 [129] – leads to the conclusion that e-Commerce is already fairly important to businesses worldwide. Furthermore, if the growth rates seen thus far continue along the same trends, then this importance is only going to increase in magnitude.

However, privacy concerns from the public have hampered the growth of e-commerce to a notable degree; one report by Forrester Research estimated that in 2001 US$15 billion of sales were lost due to privacy concerns of consumers [66], while a report by Jupiter Research estimates that online retail sales would have been approximately 24% higher in 2002 if such privacy concerns had been dealt with effectively [97]. In 1999, Forrester Research predicted US$184 billion of US online retail sales in 2004 [65]; the actual figures seen were approximately US$71 billion [166]. Many other reports with a similar theme have been published over the years (e.g. [48, 85, 92, 136, 180, 181], and while different reports focus on different aspects of the problem, all agree that there is a deficit between potential e-Commerce levels and those actually seen, owing to privacy concerns of the public. One study expounding this conclusion was Gartner's survey in June 2005 which showed that the problem has not abated over time, and predicted that these privacy concerns will only continue to further inhibit e-Commerce growth rates in the coming years [70].

The background to the reports of privacy concerns is that of a world where consumers only directly give a small amount of private information to e-Commerce vendors, such as name and address; the rest of the private information has to be gleaned by the vendor from purchasing habits of its consumers and other such statistical information. However, the world is changing: new technologies are being developed in which the consumer is required (at least, if they wish to use such technologies) to directly give the e-Commerce vendor larger amounts of personal information. One such emerging technology is addressed in this thesis, which is an area of active research, is in enhancing the e-commerce experience through making

use of a consumer's personal preferences.

4.2 E-COMMERCE SEARCH

If e-Commerce is the process of buying and selling products and services over the internet, then a commonly seen part of this process is one whereby a consumer visits the website of a vendor and spends some time browsing or searching the vendor's catalogue of available items until they find an item they wish to purchase; they then inform the vendor that they wish to purchase this item and send payment and delivery information to the vendor. Looking specifically at the case where the consumer is searching for an item, and not just browsing, this is described in more detail in Figure 4.1.

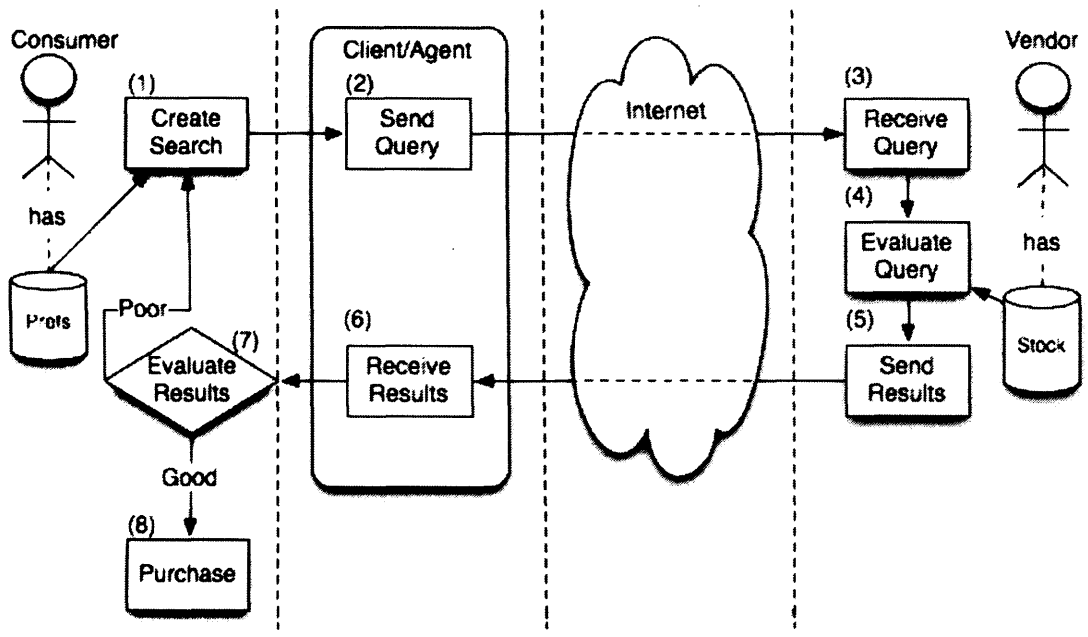

Figure 4.1: Current e-Commerce Search Scenario

The two main players involved in the process of searching are the consumer and the vendor. The consumer is a person who wishes to buy an item while the vendor is a real world business with an electronic presence on the internet and items to sell. Sitting between the consumer and the vendor is an agent usually a web browser residing on the consumer's computer. The consumer has a set of preference criteria over the type of item that they wish to buy; this criteria exists only informally within their mind. Their main aim is to find an item that fits their preference criteria. The vendor has a set of stock available for sale and represented in a datastore. Without loss of generality, we assume that stock

data is stored in a single relation $S(A_1, A_2, \ldots, A_k)$, where each A_i is an attribute having an associated domain, D_i $1 \leq i \leq k$. Thus, a set of of available items are represented as a collection of tuples forming an instance of S where each tuple $t \in D_1 \times D_2 \times \ldots \times D_k$.

EXAMPLE 1

Assume that we have a consumer who is wishing to purchase a used car. They have a set of preferences which states that the car is to be either a Mercedes or BMW (Mercedes is preferred), in Silver or Black (Silver is preferred); the make is more important to them than the colour; and they are willing to pay up to £35,000.

Also, assume that we have a vendor with a small collection of used cars for sale, and a website with a search interface. Car stocks are stored in a Cars(Make, Model, Colour, Engine Size, Electric Windows, Price) *table, as shown below:*

id	*Make*	*Model*	*Colour*	*Engine Size*	*Electric Windows*	*Price*
1	*BMW*	*325i*	*Black*	*5000*	*no*	*£30,000*
2	*BMW*	*M5*	*Silver*	*4500*	*yes*	*£50,000*
3	*Ford*	*Focus*	*Red*	*1300*	*no*	*£4,000*
4	*Mercedes*	*SLK300*	*Silver*	*3500*	*yes*	*£28,500*
5	*Mercedes*	*SL315*	*Blue*	*3200*	*no*	*£15,000*
6	*Mercedes*	*SL615*	*Yellow*	*3000*	*yes*	*£12,000*
7	*Toyota*	*MR2*	*Sonic Shadow*	*1998*	*yes*	*£6,000*
8	*Toyota*	*Prius*	*Black*	*1200*	*yes*	*£8,000*

The process of searching in e-commerce consists of the following steps:

(1) Create Search:

The consumer, using their internally held preferences, interacts with the vendor's search function on a web page. They create an initial search query, Q_1, which will contain a single search term for one or more of the attributes in the Cars table. Thus Q_1 represents a manually constructed subset of the consumer's complete preferences. The initial search will likely be looking for the consumer's "perfect" item. Using the example above, Q_1 will equate to a search for a Silver Mercedes costing less than £35,000.

(2) Send Query:

Q_1 is submitted to the vendor over the internet (either in a plaintext or encrypted format).

(3) Receive Query:

The vendor receives Q_1.

(4) Evaluate Query:
The vendor will take the Q_1 and evaluate it with respect to the Cars table, returning a set of results, R_1 that satisfy the search criteria of Q_1. R_1 will obviously be a subset of Cars ($R_1 \subseteq Cars$). Using the example above, R_1 would contain a single item – item 4 (Silver Mercedes costing £28,500) since this falls within the scope of the query.

(5) Send Results:
The vendor will send R_1 to the consumer, again over the internet.

(6) Receive Results:
The consumer receives R_1 in the form of a human readable HTML page which displays each item in R_1.

(7) Evaluate Results:
The consumer, using their internally held preferences, evaluates the results that were returned. If they decide that the results are not acceptable (for example, the result set is empty or too large, or they do not desire the items presented), then they may return to step (1) and formulate a new search query (Q_2) by changing the search terms. This process may repeat one or more times, until acceptable results are returned or the consumer ends the process. When, however, the results returned were acceptable to the consumer, they may decide to purchase an items in the results. Using the example above, the consumer may decide to purchase the item presented to them.

(8) Payment:
If the consumer has decided to purchase an item, they would then go through a process of sending payment, and if necessary, delivery details to the vendor.

4.2.1 LIMITATIONS AND ISSUES

While this basic e-commerce search methodology is currently widely used and relatively easy to implement it suffers from a number of limitations, most notably *Information overload* and *empty result sets.*

Information overload is the problem of large amounts of results being returned due to too broad a search being submitted (many of which are likely to be sub-standard with respect to the user's preferences); while empty result sets are the opposite problem of no results being returned due to too specific a search being

submitted. Both are equally undesirable as they waste the time of the consumer, they waste the bandwidth of both consumer and vendor, and they are likely to frustrate consumers which may lead them to stop searching and even try a competitor's business after a few failed attempts.

The root cause of both information overload and empty result sets is the fact that the standard form e-commerce search takes is closely linked to the technology used to perform the search – database queries in the form of *hard constraints* – meaning that searching works in an exact-match manner using specific keywords as specified by the consumer. This is in sharp contrast to the real world, where the consumer's preferences are likely to be much more complex than what can be specified in this simple form, and where the consumer may often be prepared to compromise on less ideal results if they were all that were available.

Other limitations exist that are mostly focused around security, for example, the possibility that privacy can be lost to third parties intercepting communications, however, these are not dealt with in this thesis.

4.3 ENHANCED SEARCHING USING PREFERENCES

New search techniques have been developed aimed at overcoming the related problems of information overload and empty result sets. These new techniques make use of a consumer's personal preferences for searching, as shown in Figure 4.2.

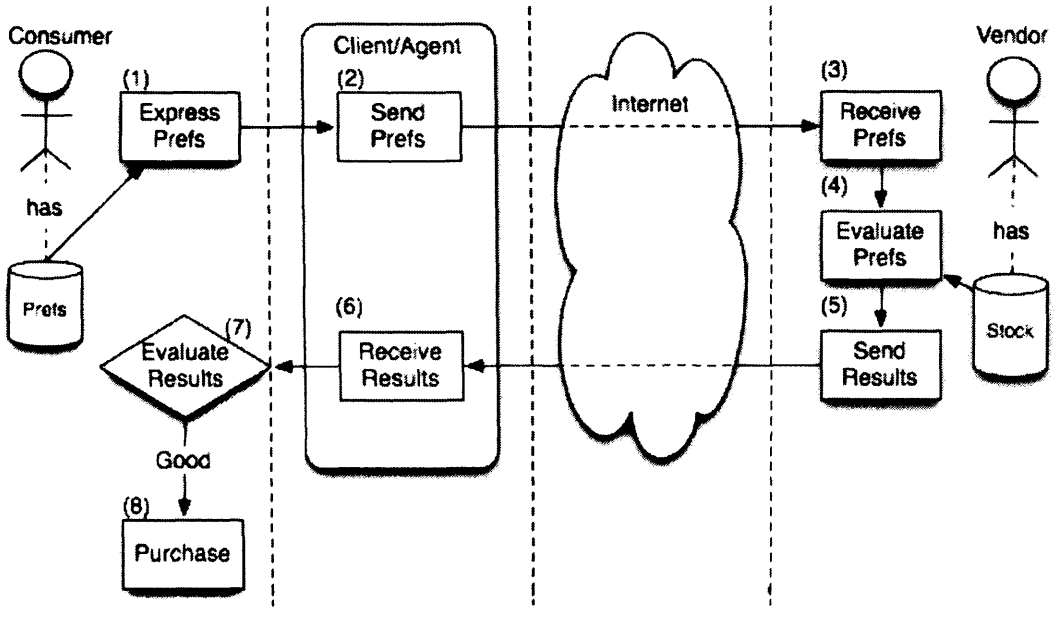

Figure 4.2: Current e-Commerce Preference Search Scenario

The two main players involved in the process of searching are the same as those seen in standard e-commerce search, but there is a difference in how the search is performed.

We illustrate the differences by running through the same example as before.

(1) Express Preferences:
The consumer, using their internally held preferences, interacts with the vendor's new preference search function on a web page. They express all of their preferences that they consider relevant for this search. Using the example above, this will equate to a search for a Mercedes or BMW (Mercedes is preferred), in Silver or Black (Silver is preferred), the make is more important to them than the colour, and they are willing to pay up to £35,000. In addition, the consumer also indicates that they would like to see the top two items.

(2) Send Preferences:
All of the preferences are submitted to the vendor over the internet (either in a plaintext or encrypted format).

(3) Receive Preferences:
The vendor receives the preferences.

(4) Evaluate Preferences:
The vendor will evaluate the preferences with respect to the Cars table, returning a set of results, R that satisfy the preferences. R will obviously be a subset of Cars ($R \subseteq Cars$). R would contain the two items from Cars that most closely match the preferences – item 4 (Silver Mercedes costing £28,500) and item 1 (Black BMW costing £30,000) – since these fall within the scope of the preference query.

(5) Send Results:
The vendor will send R to the consumer, again over the internet.

(6) Receive Results:
The consumer receives R in the form of a human readable HTML page which displays each item in R.

(7) Evaluate Results:
The consumer, using their internally held preferences, evaluates the results that were returned. If they decide that the results are not acceptable (i.e.

if the results shown were too far down their preference ordering), they may abandon the process. If, however, any of the results returned were acceptable to the consumer, they may decide to purchase one of them. Using the example above, the consumer may decide to purchase the first item presented to them.

(8) Payment:
If the consumer has decided to purchase an item, they would then go through a process of sending payment, and if necessary, delivery details to the vendor.

Highlighting the differences between the models of preferences search and standard search, it can be seen:

- The consumer, instead of manually mentally constructing a single search from their preferences, now detail the whole of their preferences. Given that these preferences have now left a conceptual representation within the consumer's mind, there now needs to exist a formal method of representing such preferences and a method in which the consumer can express them to the vendor. The consumer also expresses the amount of results that they wish returned.

- The vendor, instead of accepting a simple binary keyword search terms, now accepts a preference query, PQ. Conceptually, PQ is used by the vendor to create a function, $\delta(PQ, s_i)$, which computes a preference score for each of the items in their table, where a higher value of δ indicates an item that more closely matches PQ. Those items with the highest δ values are the top most preferred items, and a set of these, containing the amount of items the consumer requested, are returned to the consumer as a result set R.

Thus, given this formulation of this new model of preference search, the work that has been accomplished thus far concentrates on the mechanics of how the vendor can take a set of preferences and efficiently evaluate it with respect to stock database.

4.3.1 LIMITATIONS AND ISSUES

While this new model clearly addresses the issues it set out to solve – namely, those of information overload and empty result sets – an assumption implicit to the model, and therefore to all of the technologies so far created, is that the vendor

is entirely trustworthy. This assumption means that all existing implementations thus far have unwittingly adopted an approach that we have deemed the "complete release" paradigm.

The essence of "complete release" is that since the vendor is assumed to be entirely trustworthy, the consumer happily gives the entirety of their preference information to them. In practice, however, the main goal of most vendors is to turn as large a profit as possible for its shareholders: if an opportunity to increase profit presents itself, then the assumption that the vendor is entirely trustworthy is one that cannot realistically hold to be true in all cases. Even if we assume that the majority of vendors were to be trustworthy, the fact that some may not be – and that the consumer has no way of identifying which are and which are not – means that the consumer has to either accept the risk or assume that none are trustworthy.

Thus, the model of preference searching as currently used presents two main problems.

4.3.1.1 Privacy

The relationship between privacy and a consumer's preferences is fairly simple in premise – preferences over different items, and the relative preferences between the items, represent a key part of the psyche of a consumer, thus representing a significant and valuable set of private information. Given the conclusions drawn in Chapter 2 that information in the age in which we live is greased (once released, a consumer has lost control of it), one way to attempt to minimise this loss of control is to attempt to minimise the release of the information in the first place. Thus, consumers who wish to retain as much privacy as possible need to release as little preference information as possible.

If a consumer wishes to make use of preference search techniques, they must release some amount of preference information. While they may only release it in small amounts about specific categories of items to an individual vendor (i.e. either intentionally or unintentionally minimising preference release) thus mitigating the issue somewhat, the minimisation of even this small amount of privacy loss is still vitally important because vendors can potentially collude, linking together small amounts of information they each glean about a consumer, building up an almost complete picture of a consumer's preferences. This, coupled with personally identifiable information (PII) about the consumer, such as name and address, could

in the extreme case result in an almost total and comprehensive loss of privacy.

Even without the use of preference search techniques, if a consumer is to buy an item from a vendor, then a small amount of preference information generally has to be given away – the item will have characteristics that are obviously liked by the consumer – notwithstanding some of the techniques in Chapter 3, which were evaluated as not being generally applicable to e-Commerce. So, if a consumer wishes to take part in e-Commerce then they cannot eradicate *all* preference information from being released to a vendor; however, a target that is possible, and should be aimed for, is to minimise the *needless* release of preferences wherever possible. This should make the possibility of vendors piecing together information about a person's preferences harder to do with less accurate results.

The current model of preference searching is based upon the idea that the entirety of a consumer's stated preferences are communicated to the vendor. All current preference searching implementations are centred around this model. While this does mean that they are all relatively efficient since various search optimisation techniques can be used, they each fail on this privacy issue. Preference searching technologies that are little more careful about what preference information they release to the vendor are needed.

4.3.1.2 Exploitation

The Oxford English Dictionary[16] defines exploitation as "The action of turning to account for selfish purposes, using for one's own profit". Given the specific scenario of exploitation in e-Commerce discussed in this thesis, we notionally extend and focus this definition of exploitation to introduce the idea of "preference exploitation", giving it a meaning of "a vendor making use of the information contained within a consumer's preferences to their advantage, and potentially the consumer's disadvantage".

The information contained within a consumer's preferences are potentially subject to many different types of preference exploitation, all of which share a common theme: the vendor, given a preference query, would return a set of results that is collectively not the most preferred set of results that could be returned, because the vendor would rather sell the items in the set they return. Using the δ function defined earlier this equates to the vendor returning an "exploited" result set, R_e, rather than the "unexploited" result set, R_u, even though $\delta(P, R_e) < \delta(P, R_u)$.

[16] http://dictionary.oed.com/

The simplest way to demonstrate this idea is through the use of a few examples. Looking from the point of view of individual items (with no loss in generalisation for sets of items), for example, a vendor could attempt to:

- Maximise Profit: the vendor could artificially inflate the preferred ranking of less preferred items whose price is greater. i.e. $if\ \exists s_y, s_x \in S\ s.t.\ (\delta(P, s_y) - \delta(P, s_x) \leq \epsilon)$ AND $(Price(s_y) \geq Price(s_x))$ (where S is the stock data and ϵ is a value tuned by the vendor that indicates how closely preferred s_y must be to s_x such that it is still likely the consumer will want to buy s_y) *then* the vendor returns s_y instead of s_x.

- Maximise the range of stock available: the vendor could artificially inflate the preferred ranking of items with large quantity levels in an attempt to only sell items with low quantity levels to consumers who are looking for that specific item. i.e. $if\ \exists s_y, s_x \in S\ s.t.\ (\delta(P, s_y) - \delta(P, s_x) \leq \epsilon)$ AND $(Quantity(s_y) \geq \alpha)$ AND $(Quantity(s_x) \leq \beta)$ (where S and ϵ are as defined previously, and α and β are values set by the vendor indicating their desired overstocking level of s_y and minimum stock level of s_x respectively), then the vendor returns s_y instead of s_x.

- Artificially promote specific items: a supplier of a particular item could provide inducements to a vendor to artificially inflate the preferred ranking of their particular product. i.e. $if\ \exists s_y, s_x \in S\ s.t.\ (\delta(P, s_y) - \delta(P, s_x) \leq \epsilon)$ AND $(s_y \in T)$, (where S and ϵ are as defined previously, and T is the set of items the vendor wishes to inflate the rankings of.

Thus, given adequate preference information from a consumer, a vendor is able to exploit the information contained within to their own advantage, and the consumer's disadvantage; thus preference exploitation is a real potential problem in the world of preference searching.

Looking closer at the notion of preference exploitation, it can be seen that exploitation of this form can be thought of as being akin to price discrimination. Price discrimination is commonly defined as being the sale of identical goods or services from the same provider at different prices to different customers, depending on their willingness to pay. In general, economists are proponents of this economic practise since definite economic advantages have been demonstrated, while the general public are usually opponents; vociferously disproving of it as they perceive it as being unfair to them and downright exploitative [128]. Exploitation of a

consumer's preferences, however, goes one step further than price discrimination: it allows the sale of *different, less preferred*, goods or services from the same vendor for an *unchanged or increased price.* In the worst case, a consumer could end up paying more for an item they desire less.

From a vendor's point of view, preference exploitation is eminently sensible: they may argue that all they are enabling is a simple process such as maximisation of profit; a core goal any competent business strives for. From a consumer's point of view, however, making use of enhanced search techniques which were supposed to make their e-commerce interactions easier and better, but which instead leads to their exploitation, is obviously not desirable. Given the public dislike of price discrimination, this comparable but demonstrably more unfair and exploitative process could potentially seriously harm the public view of e-commerce, and therefore there is a real need to create technologies that mitigate against the possibility.

4.4 A NEW MODEL

To address the issues regarding privacy and exploitation inherent in the current model of the problem represented by existing implementations of preference searching, an adapted model is necessary. Such a model would have the following constraints and characteristics:

- A Vendor, V, has a dataset of available stock items S.
- A Consumer, C, acting through an agent, A, has a set of preferences P.
- $\delta(P, S_i)$ is a function which computes a preference score for an item $s_i \in S$ – where a higher δ indicates s_i more closely matches P and a lower δ indicates s_i less closely matches P.
- The vendor cannot be assumed to be "honest", i.e. V is not just interested in returning C an item from S, S_x, s.t. s_x most closely matches P (such that a sale is likely) but also both in gathering as much preference information about C as possible, and in potentially exploiting the information contained within P to instead return an item $s_y \in S$ $s.t$ $\delta(P, s_y) \leq \delta(P, s_x)$ but V wants to sell more rather than s_x for some reason.
- The consumer is no longer just interested in finding the item s_x inS that most closely matches their preference information in P – i.e. the item in S with

the largest δ. This time, they are also requiring that the size of P handed to the vendor is minimised, as are the chances that the vendor can exploit the information in P to return s_y instead of s_x, where $\delta(P, s_y) \leq \delta(P, s_x)$ but the vendor would rather sell S_y.

- To achieve the goal of minimising the size of the released preferences, instead of the consumer simply giving P to the vendor, they instead should give PS (where $PS \subseteq P$). The vendor will then evaluate it against S and against their own requirements, returning s_z.

- Note that now s_x is the item in S which most closely matches P, s_y would be the item that V would return to C given P, and s_z is the item that V returns to C given PS. Thus, C is attempting to find the s_z with the lowest $\delta(PS, s_x) - \delta(PS, s_z)$.

Thus, given this new formulation of the problem, the work that needs to be accomplished centres on how the consumer can find the correct PS from P such that the vendor returns an item $s_z \in S$ with a minimal $\delta(PS, s_x) - \delta(PS, s_z)$ while not giving away too many of their preferences (PS is a minimal subset of P), i.e. the consumer receives the benefits of using their preferences to search without losing too much preference information.

4.4.1 Gradual Partial Release

An alternative approach to preference searching has been created based on the new model – "Gradual Partial Release" (GPR) – based upon the idea of sending a single subset of the consumer's full set of preferences (partial release), and of repeating this process with different subsets of the full set of preferences until satisfactory results are returned (gradual partial release). The core idea of this approach is conceptually fairly simple and is described in Figure 4.3.

Again, we illustrate the differences between the new and the previously described approaches through the same example as before.

(1) Express Preferences:
The consumer, using their internally held preferences, interacts with an agent, expressing their preferences relevant to the preference search they wish to perform, i.e. P is given to A. They also express a number, n, which represents the number of results they wish to retrieve. Using Example 1,

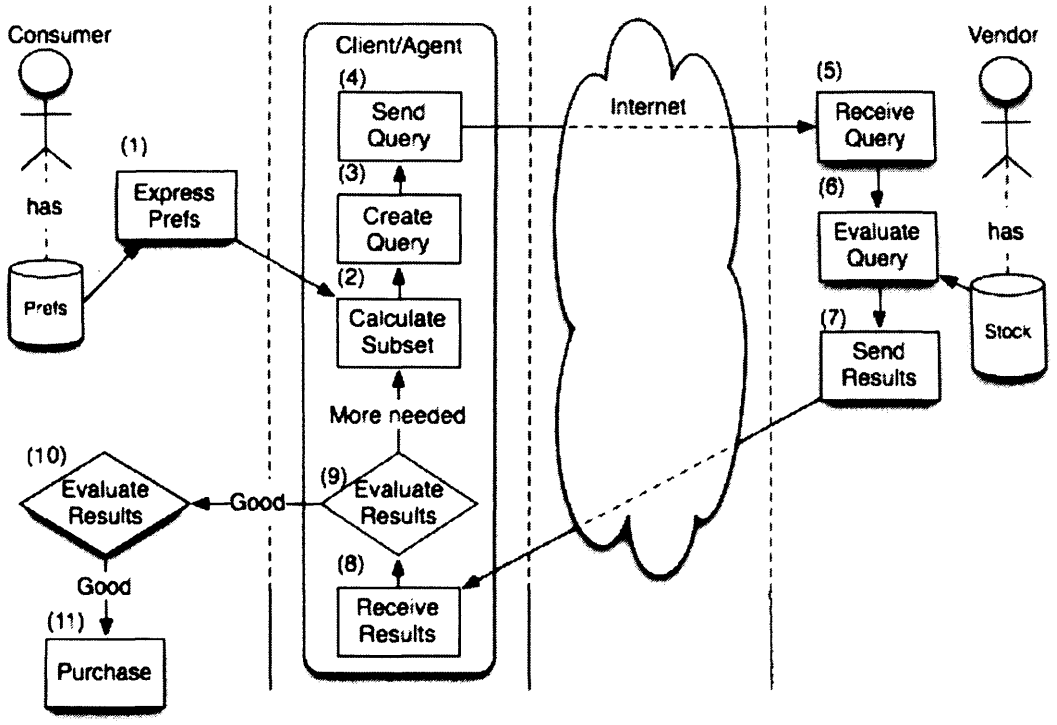

Figure 4.3: Proposed e-Commerce Preference Search Scenario

P will equate to a search for a Mercedes or BMW (Mercedes is preferred), in Silver or Black (Silver is preferred), the make is more important to them than the colour, and they are willing to pay up to £35,000. Assume that the consumer also indicated that they would like to see the top two items.

(2) Calculate Subset:

Using a preference subset creation algorithm, (to be discussed in Chapter 5 the agent creates a subset of P, PS_1. This subset can contain preference information for one or more of the attributes in the Cars table. If the preference subset creation algorithm aimed to first create a subset representing the consumer's "ideal" item, the preference subset would equate to a query for a Silver Mercedes, costing up to £35,000.

(3) Create Query:

The agent takes the preference subset, PS_1 and converts it into a query, Q_1. This could either be a simple binary keyword search query, or the existing work on preference searching could be utilised and a preference query created.

(4) Send Query:

Q_1 is submitted to the vendor over the internet.

(5) Receive Query:
The vendor receives Q_1.

(6) Evaluate Query:
The vendor will take Q_1 and evaluate it with respect to the Cars table, returning a set of results R_1 (where $R_1 \subseteq Cars$) that satisfy the search criteria in Q_1. Using Example 1, R_1 would contain a single item – item 4 (Silver Mercedes costing £28,500) since this falls within the scope of the query.

(7) Send Results:
The vendor will send R_1 to the consumer, again over the internet.

(8) Receive Results:
The agent receives R_1.

(9) Evaluate Results (by Agent):
The agent takes R_1 and evaluates its contents, using criteria such as n results being required (but possibly other criteria as well). If more results are required, the agent may formulate a new search query by going back to step 2 and calculating a different subset of P to release and the process will repeat. The items in each successive R_m received will be combined into one overall result set, R. When its evaluation criteria are fulfilled, the agent will then present R to the consumer it is acting for. If R becomes larger than n, the agent will evaluate the results and presents the n items within R that best match P.

(10) Evaluate Results (by Consumer):
The consumer using their internally held preferences, P, evaluates the results that were returned. If the results were acceptable, they may decide to purchase one of the items in the results.

(11) Payment:
If the consumer has decided to purchase an item, they would then go through a process of sending payment, and if necessary, delivery details to the vendor.

Highlighting the differences between the models of the new approach to preference search and the existing approach to preference search, as previously described, it can be seen that:

- The consumer, instead of expressing the whole of their preferences, P, directly to a vendor, now expresses it more locally to an agent acting on their behalf.
- The agent takes these preferences and splits them up into a series of subsets of the whole preferences, in order that these can be released gradually to the vendor.

Thus, an implementation that enables this approach needs the following:

(a) A way of allowing a consumer to express preferences to an agent;

(b) A method of storing these preferences within an agent in a form that allows operations necessary to partition preferences;

(c) A preference subset creation algorithm to take a complete set of preferences and split it up into one or more preference subsets;

(d) A way of sending these preference subsets to a vendor and returning retrieved results.

This approach of splitting up a consumer's preference into subsets and gradually releasing them could be both a technology that directly meets the aims of minimising privacy loss and exploitation, and an enabler of further methods of fulfilling these aims in a more effective manner. Directly, this approach will allow the possibility that only a certain amount of the full set of preference information of the consumer will have to be released, meaning they do not lose control of this information, and also meaning there is less information available to the vendor to practise exploitation. As an enabling technology, this approach allows for methods such as sending preference queries with some "fake" preference information as a constituent part in order to further enhance the privacy of the preference information that is given out (in a process similar to [154]); however, this idea is not discussed further in this thesis.

4.5 SUMMARY

This chapter has analysed the issues around search in e-Commerce. The model of standard e-commerce search was examined and some of its issues were discussed – notably information overload or empty result sets. A model of using a consumer's

preferences to help combat these problems has been introduced by the community; this was in turn examined and issues inherent to this model were highlighted: issues around the related areas of privacy and exploitation, caused by the fact that the model is based upon giving all of a consumer's preference information to a vendor and trusting them with this personal information. The conclusion was drawn that to combat these problems an alternative model of enable preference searching was needed; such an alternative model was then proposed, based on the idea of an agent acting on behalf of the consumer gradually releasing only parts of the consumer's preferences to a vendor. The next chapter will consider the new model in detail.

Chapter 5

Gradual Partial Release of Preferences

Privacy loss and the potential for exploitation are potential problems when using a consumer's preferences to assist with searching in e-commerce. The previous chapter proposed an alternative approach to the idea of preference searching that mitigates against these problems. This chapter details an implementation of this approach. It first introduces definitions and detail necessary to describe the implementation, followed by details of the implementation itself.

5.1 Preferences

To allow preference searching in e-commerce using the proposed approach, the consumer must express their preference information to their agent, and the agent needs to store the information in a form that it is able to work with and perform computation upon. Various frameworks and preference languages have been explored by the community over the years that would enable this.

Lacroix and Lavency presented one of the first complex studies of preference queries [109], proposing an extension to domain relation calculus whereby preferences can be expressed in the form of logical conditions. However, their work did not contain any formal language for expressing this logic, and was only able to capture fairly simple preferences [37].

The first work that addressed the expression of preferences in any detail was accomplished by both by Kießling and Güntzer [103], and Köstler et al [108], and by Govindarajan et al [78, 79]; these two independent sets of work centred on the specific area of deductive databases and proposed extensions to Datalog to

accomplish their goals. While both sets of work were comprehensive, they both require a specialised query evaluation engine to work.

Following this, three groups of researchers all presented separate (but mostly similar) frameworks for formulating preferences. Agrawal and Wimmers first introduced a framework for expressing and combining preferences [8] that works by specifying preferences between tuples of a relation using *scoring functions*, whereby a numeric score is associated with each tuple of a query result set. However, it does not provide a mechanism of associating a consumer's preferences with the tuples such that a preferred ordering can be achieved. The other two pieces of work, however, do. These were Kießling and Köstler's Best Match Only (BMO) query model [102, 104], and Chomicki's Winnow operator [37]. Both were similar in scope, but Chomicki's work defined preferences in the form of logical formulas, while Kießling and Köstler's work defined preferences using special preference constructors.

Alongside these general frameworks for formulating preferences, more specific implementations were introduced. An example of this is the *skyline* operator, introduced in 2001 by Borzsonyi et al [20], which when given a set of preferences over numeric attributes will search for a set of "interesting" tuples, where interesting means that it is not dominated by any other tuple. For example, hotel *h1* dominates hotel *h2* if *h1* is cheaper than *h2* but both have the same rating. The problem being solved by the skyline operator is also known as the maximal vector problem. This has been implemented by extending the SQL syntax to allow a "SKYLINE OF ..." clause to be added to a standard SELECT statement. While a very powerful idea, it is limited to working with numeric attributes only, and thus is not flexible enough for our purposes.

The work presented in this thesis is an enhancement to the general concept of preference searching, and not to any specific existing preference searching technology. Thus, for the purposes of this thesis, the requirements of a preference framework are simply to allow preferences to be expressed and evaluated in a technology-independent manner. This should mean that the results will be relevant to any preference framework. However, rather than starting completely from scratch, the work of Kießing (i.e. the work presented in [102]) will be used as a base – since it gives a set of base preference constructors that allow fairly complex preferences to be expressed in a manner that has a sound logical underpinning, while remaining simple to use, without loss of generality.

A consumer's preferences over a specific category of items will naturally exist

in the form of preferred values across a range of attributes used to describe items within that category. For example, when thinking about a car they wish to buy, a consumer might have preferences about the make, engine size, and price: they may prefer the make to be "BMW", followed by "Mercedes" or "Ford"; and the engine size to be between 1500cc and 2000cc. These may be considered as "value preferences", i.e. preferences expressed over values of a specific attribute of an item. The first is an example of a "discrete value preference" (preferences expressed over specific values); the second an example of a "range value preference" (preferences expressed over a range of values). The consumer might also have a preference about which of these attributes are more preferred: they may consider their preferences over make to be more more important than those over engine size. This may be considered an "attribute preference", i.e. a preference expressed over a set of value preferences.

Within both such preferences a preferred ordering is specified, or can be inductively assembled, allowing a consumer to specify a preference ordering (or lack thereof) between two discrete values within a value preference, between sets of continuous values within a range preference, or between value preferences within an attribute preference. The preference operators that enable this, along with a more in-depth discussion of each of the types of preferences, are now detailed.

5.1.1 RELATIVE PREFERENCE OPERATORS

To allow complex preference relations to be directly expressed, or inductively built, two preference operators are required. These are the *Pareto* and *Prioritised* accumulation operators [102].

DEFINITION 1 - **Pareto Accumulation Operator**

The pareto accumulation operator ($\otimes$) defines a relation between two values (v_1 and v_2), where $v_1 \otimes v_2$ indicates that v_1 is considered equally as preferred as v_2.

DEFINITION 2 - **Prioritised Accumulation Operator**

The prioritised accumulation operator ($\&$) defines a relation between two values (v_1 and v_2), where $v_1 \,\&\, v_2$ indicates that v_1 is more preferred than v_2.

Based on these two operators, we introduce two types of value preference and an attribute preference in the following sections.

5.1.2 DISCRETE VALUE PREFERENCES

A Discrete Value Preference (DVP) is a Value Preference defined on a single attribute. Let $v_1, v_2, \ldots, v_k$ be a set of values of a discrete attribute A. A DVP on A is expressed in the form of $A = exp$, where exp involves $v_1, v_2, \ldots, v_k$ and the two preference operators, and is formed according to the following grammar:

```
<dvp> ::= <equal-values>
        | <equal-values> <prioritised-op> <equal-values>
<equal-values> ::= <value>
                 | <equal-values> <pareto-op> <value>
<pareto-op> ::= ⊗
<prioritised-op> ::= &
<value> ::= <character> | <value> <character>
<character> ::= A | B | C | D | E | F | G | H | I | J | K | L | M | N
| O | P | Q | R | S | T | U | V | W | X | Y | Z | 0 | 1 | 2 | 3 | 4 |
5 | 6 | 7 | 8 | 9 | 0
```

For example, $Colour = Silver \otimes Black \,\&\, Red \otimes Green \otimes White$ is a DVP, and it expresses that Silver or Black are equally preferred, followed by the less preferred set of values of Red, Green or White (equally preferred). Note that when operators appear in a single DVP the Pareto operator ($\otimes$) takes precedence over the Prioritised operator (&) in computation. That is, $Colour = Silver \otimes Black \,\&\, Red \otimes Green \otimes White$ is equivalent to $Colour = (Silver \otimes Black) \,\&\, (Red \otimes Green \otimes White)$. Also note that all values expressed in this form are required to be distinct, hence cyclic preference relations (e.g. $v_1 \,\&\, v_2 \,\&\, v_1$) are not valid[17].

A DVP may be represented as a partially directed graph: each vertex represents a value in the expression; vertices connected by a directed edge represent a Prioritised relationship (&); and vertices connected by an undirected edge represent a Pareto relationship ($\otimes$) preference relationship. For simplicity of presentation, a set of vertices connected by undirected edges can also be condensed as a supernode. Figure 5.1 gives an example of a DVP.

It is easy to see from the properties of a DVP that there should be no isolated vertices in a graph representing a DVP as this graph is acyclic.

[17]More expressive preference frameworks exist, and the reader is referred to [38, 102] for further details. For the purposes of this thesis, we only consider a simple preference framework.

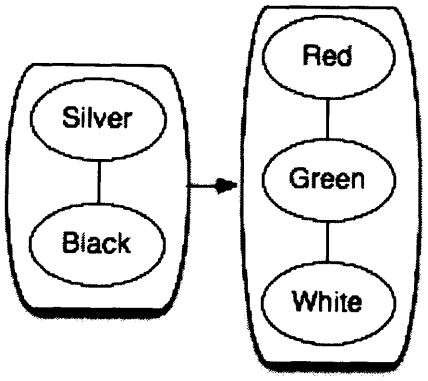

Figure 5.1: Example of a DVP represented as a graph

5.1.3 RANGE VALUE PREFERENCES

A Range Value Preference (RVP) is similarly defined to that of a DVP, except that it is specified over a range, rather than a set of discrete values. To express an RVP, a standard set of Range Preference Operators, as used by Kießling's et al. [102, 104], may be used. These are "GREATER THAN (X)" (a preference for values greater than X), "LESS THAN (X)" (a preference for values less than X), "BETWEEN (X,Y)" (a preference for values between X and Y), "HIGHER" (a preference for higher values), and "LOWER" (a preference for lower values).

An RVP on an attribute A is expressed in the form of A = *exp*, where *exp* is constructed according to the following grammar:

```
<rvp> ::= <rvp-gt> | <rvp-lt> | <rvp-bet> | <rvp-higher> | <rvp-lower>
<rvp-gt> ::= "GREATER THAN" <value>
<rvp-lt> ::= "LESS THAN" <value>
<rvp-bet> ::= "BETWEEN" <value> <value>
<rvp-higher> ::= "HIGHER"
<rvp-lower> ::= "LOWER"
<value> ::  = <numbers> | <numbers> <decimal-point> <numbers>
<decimal-point> ::  = "."
<numbers> ::  = <number> | <number><numbers>
<number> ::  = 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 0
```

For example, a consumer may wish to define a preference over an attribute of Engine Size for those items with values less than 3000. This would be expressed as $EngineSize = LESS\,THAN\ 3000$.

Range Value Preferences do not explicitly contain the Prioritised and Pareto preference accumulation operators, however, they can be implicitly inferred. The

Greater Than, *Less Than*, and *Between* operators will split a range of values up into two groups of equally preferred values, with a prioritised relation between them. For example, *Greater Than* x splits a range up into a group of values greater than the value x, all equally preferred, which are preferred to the group of values less than the value x, all equally preferred.

Note that if further preferences over a range are desired, then a DVP may be used: for example, suppose a consumer wished to express a preference of Engine Size less than 5 litres, but prefers 2 or 3 litres to 4 litres, they could express a DVP of $EngineSize = 2L \otimes 3L \,\&\, 4L$.

5.1.4 ATTRIBUTE PREFERENCES

Discrete and Range Value Preferences, as described above, allow preferences to be stated over values of a single attribute. To allow preferences to be defined across the multiple attributes that usually describe available items the notion of an "Attribute Preference" is now introduced.

An Attribute Preference, AP, is a set of Value Preferences, $VP_1, VP_2, \ldots, VP_n$, with a preference accumulation operator defined sequentially between them. The grammar for constructing an AP is the same as that for a DVP, except that attributes are substituted for values. For example, supposing Colour and Make were equally preferred, with Gears less preferred: this would be expressed as $AP = Make \otimes Colour \,\&\, Gears$, or as a graph shown in Figure 5.2.

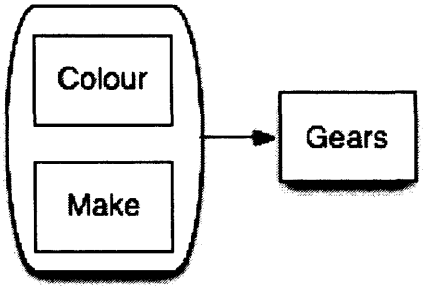

Figure 5.2: Example of an Attribute Preference represented as a graph

The interpretation of one attribute being preferred to another in an AP is essentially "importance" – in the sense of which attribute should assume precedence when being evaluated. For example, suppose we have two DVPs, $Make = Mercedes \& BMW$ and $Colour = Silver \& Black$, along with an AP, $AP = Make \& Colour$ (Figure 5.3 shows this preference information represented graphically). This indicates that the Make attribute is more preferred than the Colour attribute.

Thus, if given two items, a Black Mercedes and a Silver BMW, the Make attribute takes precedence, meaning that the preferred item would be that containing the more preferred value of the Make attribute: the Black Mercedes, since Mercedes is more preferred than BMW. If however, the AP had been $AP = Make \otimes Colour$ (i.e. Make and Colour are equally preferred), then no precedence between the two attributes would be indicated and both items would be considered equally preferred.

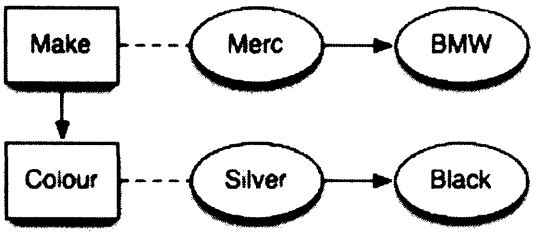

Figure 5.3: Attribute Preference with Values

It should be noted, however, that while an AP defines computational precedence between attributes thus allowing values from different attributes to be assessed as more or less "important", as in the previous example, this does not mean that a preference relationship between values from different attributes can be *directly* established. In other words, using the previous example, it could not be said that $Make = Mercedes$ is more preferred than $Colour = Silver$, since attempting to assess the relative preference between values of different attributes makes no sense in the given preference framework.

5.2 PREFERENCE SET

When a consumer has specified a set of Value Preferences (over values in a single attribute), and a set of Attribute Preferences (over attributes of a table), we have a Preference Set $\mathcal{P}$. In this section we observe the properties associated with, and establish some measures for, $\mathcal{P}$, before considering how $\mathcal{P}$ may be split up and sent to the vendor to allow preference search while minimising privacy loss and the potential for exploitation in the next section.

5.2.1 FULLY DEFINED AND COMPLETELY ORDERED

Two properties that $\mathcal{P}$ may have are *Fully Defined* and *Completely Ordered*; these properties are introduced as they are useful during the process of splitting $\mathcal{P}$ up into subsets, as discussed in Section 5.4

If $\mathcal{P}$ meets the following criteria then it is known as "Fully Defined": all attributes referred to within $\mathcal{P}$ are connected, i.e. there are no pairs of attributes without a defined preference relation (either explicitly stated or that can be deduced); and for each DVP within $\mathcal{P}$, all values within it are connected, i.e. there are no pairs of values without a defined preference relation (either explicitly stated or that can be deduced). Fully Defined Preference Sets have the following property: given any two attributes, or two values within any one of its attributes, a defined preference relation can be shown. Hence, there can be no ambiguity when assessing the relative preference of one value to another within a DVP, or the relative preference of one attribute to another within an AP.

If $\mathcal{P}$ meets the following criteria then it is known as "Completely Ordered": all Attribute Preferences within $\mathcal{P}$ have a specified prioritised ordering (i.e. there are no pairs of attributes defined as equally preferred within the Attribute Preference); and all DVPs have a specified prioritised ordering (i.e. there are no pairs of values defined as equally preferred within the DVP). Completely Ordered Preference Sets have the following property: given any two attributes, or two values within any one of its DVPs, a definite preferred ordering can be directly established.

These properties of $\mathcal{P}$ are important to the approach to preference searching outlined in this thesis. If parts of $\mathcal{P}$ are to be released in a gradual process according to some algorithm, that algorithm will need to be able to calculate the more preferred value of a set of values in a Value Preference, and the more preferred attribute in an Attribute Preference, in order to decide which to release. A completely ordered $\mathcal{P}$ can clearly help with that task.

Given that the grammar of expressing a single DVP requires all values within the DVP to be connected, and the grammar of expressing a single AP requires all values within the AP to be connected, it is easy to see that $\mathcal{P}$ is guaranteed to be fully defined. To obtain a completely ordered $\mathcal{P}$, we may impose some default assumptions on a fully defined $\mathcal{P}$. We will describe this process in Section 5.4.1.3.

We assume that $\mathcal{P}$ contains a single AP. This is a reasonable assumption because it is relatively straightforward to combine multiple APs into a single AP, based on the grammatical rules defined for an AP given in the previous section, and using

a process similar to that employed to obtain a completely ordered $\mathcal{P}$ as described in Section 5.4.1.3. Thus, for the remainder of this thesis, we will consider that $\mathcal{P}$ contains a single AP only.

5.2.2 SIZE OF INFORMATION IN PREFERENCES

Attempting to assess a quantitative "amount" of preference information within a set of preferences stated by a consumer is an area that previous work in the area of expressing and defining preferences has not needed to explore; this is because such work has typically concentrated on the expressing of preferences and the mechanics of the evaluation of this information over a set of items. However, if a technology is to take these preferences, split them into subsets, and evaluate which subsets it should send (and in which order) then this area of work becomes necessary as it will inform this evaluation process.

When considering a DVP, an obvious and relatively meaningful way to measure the amount of preference information within it is to simply examine how many values have been specified by the consumer. Thus a DVP with four values specified contains twice as much preference information as a DVP with only two values specified. Thus, we define the preference information contained within a DVP, denoted by $||DVP||$, as the number of values present in the DVP.

Measuring the amount of preference information contained with a RVP, however, is distinctly less obvious. For example, a RVP that indicates a consumer wishes the image quality of a digital camera to be between 4 and 8 megapixels could suggest a single value (4-8 megapixels), 5 values (4,5,6,7, and 8 megapixels), or 50 values (4.0, 4.1, 4.2, ... ,7.8, 7.9, and 8.0 megapixels). If the first – specifying a single value – then there is a question of how specifying between 4 and 8 megapixels compares to that between 6 and 8 megapixels. If the second or third – turning the preference into a series of discrete values and counting – then how does one specify which is the correct interval? This problem is very similar in nature to that seen in the worlds of machine learning and data mining, where much work has been done on "discretising" (or "quantising") continuous data into discrete intervals. These methods are developed primarily to help reduce time complexity when dealing with datasets with continuous data and to derive useful information from the data. Catlett presented a seminal overview of work in this area almost 20 years ago [27], mainly focused on the machine learning area, while more recently several texts on data mining include sections on discretising data, e.g. [67, 83, 183].

A simplistic approach – and one adopted for the proof of concept implementation of GPR in this thesis – is that a single RVP represents a single piece of preference information. The reasoning behind this is twofold. Firstly, it could be argued that an individual piece of preference information is one that focuses a search query in a single way. A RVP will either focus the query in one way in exactly the same way as a single value in a DVP, i.e. splitting a set of items into two – those items that match this preference and those that do not – or by instructing the query to order the items in a certain manner (i.e. the *HIGHER* or *LOWER* RVPs). Secondly, and more pragmatically, this will allow the basic idea of gradual release of subsets of a consumer's preference information to be investigated without engaging in a large amount of work of discovering the best method of discretising preference information in these circumstances. Using a less simple approach, for example "discretising" the data and measuring the amount of items in the resulting set, could potentially become a very complex problem given that a different approach to discretising may be needed for different domains that preferences were expressed over. This is left for future work. Thus, we define $||RVP|| = 1$.

Given a way of measuring preference information contained within both types of Value Preferences, measuring the preference information with a Preference Set $\mathcal{P}$ as a whole, denoted $||\mathcal{P}||$, is a matter of calculating the sum of the size of its constituent Value Preferences. That is, let $\mathcal{P}$ be a Preference Set containing a single AP over $VP_1, VP_2, \ldots, VP_n$. Then, $||\mathcal{P}|| = \sum_{i=1}^{n} ||VP_i||$.

5.2.3 PREFERENCE SUBSETS

To enable the idea of partial release of a consumer's preferences, the concept of a *subset* of a consumer's preference is necessary. A Preference Subset ($\mathcal{PS}$) is essentially the Preference Set $\mathcal{P}$ with zero or more preference values removed. An example of this is shown in Figure 5.4, where each value that will be kept is shaded and each value that will be removed is not.

A $\mathcal{PS}$ with zero values removed is, of course, equal to $\mathcal{P}$, and equates to the existing approach to preference search – that of releasing all of the consumer's preferences in one go. A $\mathcal{PS}$ with all values removed would represent a search with no preference information and would thus equate to a non preference-enhanced search. So, for the purpose of this thesis, we will consider those $\mathcal{PS}$ where $0 < ||\mathcal{PS}|| < ||\mathcal{P}||$ only.

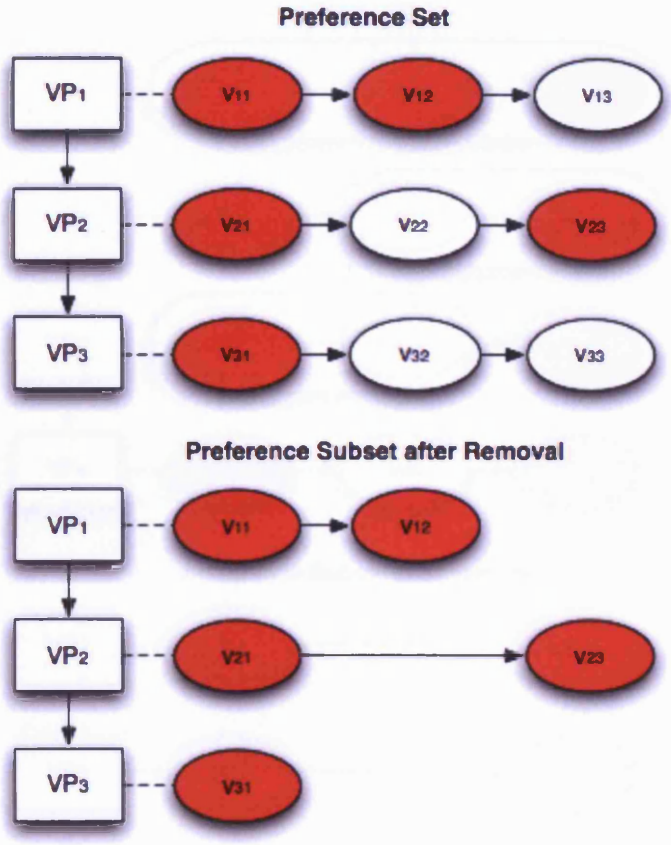

Figure 5.4: Preference Subset with values removed

When a value within a DVP is removed from $\mathcal{P}$, ordering will need to be maintained within the DVP. If the value were a maximal (minimal) value, then no work is to be done as the value is simply removed from the DVP, along with the preference operator following (preceding) it. If the value were to lie between the maximal or minimal values, however, then ordering is maintained by inferring the preference relation between its neighbours. In graph terms, if the value was part of a supernode of equally preferred values, that supernode's preference relations remain unaltered. If the value was a single value with a prioritised ordering on either side of it, then the prioritised ordering would be inherited. For example, suppose we have $DVP_1 = v_1 \,\&\, v_2 \,\&\, v_3$ and v_2 were to be removed, then the subset would be $DVP_1 = v_1 \& v_3$. Figure 5.5 shows each combination of preference relation around a value, and how the ordering is inherited in each case.

If *all* values within a single DVP were to be removed, then that attribute itself will be removed from the AP too. Ordering between Value Preferences will be maintained in the same way as for values within a DVP. An example of this is

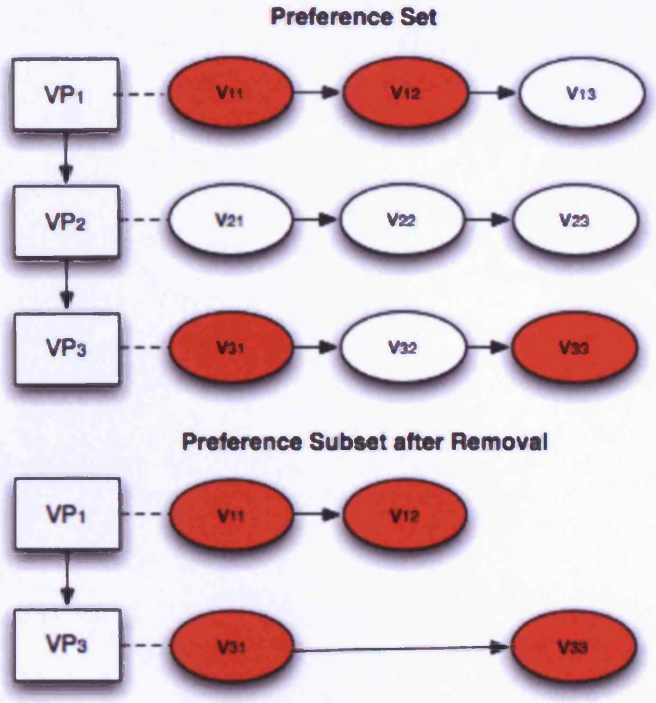

Figure 5.6: Preference Subset with attributes removed

5.2.4 Measuring Preference Loss

Given the definitions of Preference Sets and Preference Subsets, and a set of methods that measure the preference information contained within each, then a way of measuring preference loss has been enabled – comparing the preference information that is contained within $\mathcal{P}$ to that released in the form of subsets.

If only one such subset were to be released to a vendor, the calculation of the amount of preference information released would be a simple matter of comparing the size of preference information in the Preference Subset released to that in the full Preference Set. That is,

$$\text{Preference Loss}(\mathcal{P}, \mathcal{PS}) = \frac{||\mathcal{PS}||}{||\mathcal{P}||} = \frac{\sum_{j=1}^{|\mathcal{PS}|} ||VP'_j||}{\sum_{i=1}^{|\mathcal{P}|} ||VP_i||} \tag{5.1}$$

where $|\mathcal{P}|$ and $|\mathcal{PS}|$ are the numbers of Value Preferences in $\mathcal{P}$ and $\mathcal{PS}$ respectively, VP'_j is a $VP \in \mathcal{P}$ with zero or more values removed, and $||\ .\ ||$ is a function as defined in Section 5.2.2.

So, the privacy loss measure will be a numerical value between 0 and 1, where:

- 1 would indicate complete preference loss – i.e. all preference information in $\mathcal{P}$ was contained in $\mathcal{PS}$;
- 0 would indicate complete preference retention – i.e. $\mathcal{PS}$ was empty;
- values in between 1 and 0 would indicate that some amount of preference loss has occurred; the closer the value to zero, the less the amount of preference loss.

In situations where multiple subsets are released, however, then measuring the preference information lost is a little more complicated, and can be split into two scenarios – that in which the Vendor is unable to link together separate queries as having originated from the same source, and that where the vendor is able to do this. We term this *linkability.*

In the first of these scenarios, where the vendor is unable to link together multiple queries as being from the same source (for example, if a technology such as Crowds was used, as discussed in Section 3.1.1.1), then each subset released would have its own separate measurement of preference loss as described above, while the preference loss as a whole could be effectively measured as being that of the largest measure of preference loss seen, usng the measure defined above. That is:

$$\text{Preference Loss}(\mathcal{P}, \mathcal{PS}_1, \ldots, \mathcal{PS}_k) = Max\left(\frac{||(\mathcal{PS}_1||}{||\mathcal{P}||}, \ldots, \frac{||(\mathcal{PS}_k||}{||\mathcal{P}||}\right) \quad (5.2)$$

This is because the preference loss would not be cumulative, i.e. a vendor would only "know" the preference information given to it as being from that consumer on a per query basis. Not being able to link together multiple such sets of information renders them useful only separately.

In the second of these scenarios, where the vendor is able to link together queries as coming from the same source, preference loss as a whole could be effectively measured as that of the union of all subsets released. This is because the vendor could theoretically take each of the subsets released to it, connect them together, and form an overall picture of the preferences given all preference information included in each subset. Thus:

$$\text{Preference Loss}(\mathcal{P}, \mathcal{PS}_1, \ldots, \mathcal{PS}_k) = \frac{\|(\mathcal{PS}_1 \cup \ldots \cup \mathcal{PS}_k)\|}{\|\mathcal{P}\|}$$

$$= \frac{\sum_{j=1}^{|(\mathcal{PS}_1 \cup \ldots \cup \mathcal{PS}_k)|} \|VP'_j\|}{\sum_{i=1}^{|\mathcal{P}|} \|VP_i\|} \tag{5.3}$$

Note that techniques such as constructing "fake" queries to send to the vendor could be used in order to confuse the overall picture of the preferences that could be formed, however, that idea is not dealt with in this work.

5.2.5 PREFERENCE LOSS AND PRIVACY LOSS

We have argued and identified in previous chapters that a technology designed to control the release of a consumer's preferences to allow preference-enhanced search in e-commerce in a privacy aware manner is needed; since to protect privacy we need to protect (or control) the release of personal information. Information about one's preferences – i.e. one's likes and dislikes – are intuitively and obviously personal private information, and therefore equating preference loss to privacy loss is one seemingly meaningful approach of measuring privacy loss in the specific circumstances of preference searching.

As measures have been created that attempt to measure such preference loss – by assessing the amount of preference information within a Preference Set, and within a collection of Preference Subsets derived from this Preference Set and subsequently released to a Vendor – then these measures can be used as a way of measuring privacy loss. Thus, for the purposes of this thesis:

$$\text{Privacy Loss}(\mathcal{P}) = \text{Preference Loss}(\mathcal{P}) \tag{5.4}$$

Given the difficulties of coming to an agreement on what "privacy" means (as discussed in Chapter 2), the formulation of a general measure of privacy loss would be an incredibly complicated – if not impossible – task, and any solutions proposed would undoubtedly cause numerous disagreements amongst the community and would thus be less than useful. Thus we make no claims that the measure of

privacy loss employed here is more generally useful – or even particularly correct – in these circumstances. However, it does seem a reasonable way to measure an aspect of privacy loss and therefore serves a useful purpose for our work.

5.3 EXPLOITATION

The theoretical idea of *preference exploitation* of a consumer's preferences was introduced in Chapter 4; there it was described as "a vendor making use of the information contained within a consumer's preferences to their advantage, and potentially the consumer's disadvantage". In practice, this vendor would achieve this by first evaluating the consumer's preference query against their set of stock creating a set of *unexploited* results, RS_u. A process of exploitation would then occur, where RS_u will be distorted (e.g. by the vendor removing items from it), before returning a set of *exploited* results, RS_e, to the consumer.

5.3.1 MEASURING EXPLOITATION

In order to accurately measure the "amount" of exploitation that has taken place, access to both the unexploited and exploited result sets is necessary, so that a comparison can be drawn. Such a measure is necessary for the purposes of this thesis if it is to demonstrate that the GPR approach can help minimise preference exploitation. In the experiments we will assume that both are available.

Access to both RS_u and RS_e gives the necessary input for measuring the difference between the two. Many possible metrics for the difference could be constructed. However, given that RS_u and RS_e can be viewed as simple mathematical sets, the notion of "relative complement" (or "set theoretic difference") from mathematical set theory will be used.

The relative complement of RS_e in RS_u, or $RS_u \setminus RS_e$, is the set of items that exist in RS_u but not RS_e, i.e. $RS_u \setminus RS_e = \{x : x \in RS_u \text{ and } x \notin RS_e\}$. Clearly $RS_u \setminus RS_e$ shows the set of items that have been removed from RS_u due to the process of vendor exploitation. Examining the cardinality of that set, i.e. $|RS_u \setminus RS_e|$, will give a simple measure of the amount exploitation that has occurred.

$$\text{Exploitation}(RS_u, RS_e) = |RS_u \setminus RS_e| \tag{5.5}$$

For example, assume $RS_u = \{1, 2, 3, 4, 5\}$, and the vendor removes items 1 and

3, adding 6 and 7, to create $RS_e = \{2,4,5,6,7\}$. In this case, $RS_u \setminus RS_e = \{1,3\}$. and $|RS_u \setminus RS_e| = 2$

This measure will thus return a positive integer value, k, where

- $k = 0$ indicates that there are no items in the unexploited result set that are absent from the exploited result set, and therefore the exploitation process did not remove anything;
- $k > 0$ indicates that k items are in the unexploited result set that are absent from the exploited result set, and therefore have been removed by the exploitation process. The higher the value of k, the more items have been removed, and the more exploitation has occurred.

The measure given above gives only a basic indication of the amount of exploitation that has occurred, as it does not, for example, give any indication as to the relative "quality" of items removed (e.g. taking into account whether it was the top most preferred item that was removed, or the least preferred item in the set). But this measure suffices for the purpose of testing our hypothesis in this thesis.

5.4 GRADUAL PARTIAL RELEASE

Now we introduce the Gradual Partial Release (GPR) method. The basic idea of GPR is that a consumer's preferences are to be split into subsets and gradually released to a vendor. This process will involve four main areas:

- Preparing for GPR;
- Creation of a collection of Preference Subsets;
- Gradual release of this collection (including stopping criteria); and
- Post-processing of the results.

Each of these areas are discussed in turn in the following.

5.4.1 PREPARING FOR GRADUAL PARTIAL RELEASE

Before the GPR process can be occur, the consumer's preferences need eliciting, representing internally, and possibly modifying to ensure that GPR works correctly.

5.4.1.1 Eliciting Preferences

In the proof of concept implementation, preferences will be stated directly within the system without a user interface. Preference Sets are built by expressing preferred values of attributes, stating the preference relationship between values, and finally stating the preference relation between the attributes themselves. This allows complex preferences to be built up in an easy to use manner. This could potentially be extended into a visual metaphor able to be used by real world consumers, given the development of a good UI.

5.4.1.2 Representing Preferences

Preferences are represented internally in the form of a graph, as previously discussed. This allows for easy traversal of a consumer's preferences using well establish graph traversal techniques.

5.4.1.3 Establishing Completely Ordered Preferences

Given an arbitrary Preference Set, it can be seen that if constructed using the preference grammar previously introduced, $\mathcal{P}$ must necessarily be fully defined. However, given a fully defined $\mathcal{P}$, it can be seen that while it is possible for it to initially be completely ordered – if the consumer used the prioritised accumulation operator (&) between *every* value specified in each DVP and between *every* attribute pair – it does not *necessarily* contain such an ordering. This lack of assured ordering may cause problems for GPR algorithms that require a definite preference ordering between any pair of values. To avoid encountering these ambiguities, $\mathcal{P}$ will be made completely ordered.

To completely order $\mathcal{P}$, essentially, the partially directed graph of $\mathcal{P}$ needs to be made directed – i.e. all undirected edges within $\mathcal{P}$, and its subgraphs, need to be made directed edges. A decision is needed on which direction an undirected edge should take. There are different ways this could be achieved; for example:

- A simple random process (the rationale being that if the consumer expressed no preference between a pair of values then it does not really matter how the order is imposed);
- A process whereby the value stated first by the consumer becomes the preferred value (the rationale being that since this value was stated by the con-

sumer first, it is possible that it is slightly more preferred than the second. If not, then it does not really matter anyhow);

- A manual process of asking the consumer to impose an order; or
- If the agent is to attempt to build up an estimation of the vendor's catalogue for the purpose of assessing exploitability, then using statistics from the estimated catalogue to decide which of the values is more likely to retrieve results. For example, if the amount of stock matching a preferred value of an attribute of "Make" matched five items while the amount of stock matching a preferred value of an attribute of "Memory" contained ten items, the optimal attribute to select as "most important" would be "Memory", as releasing the values in this attribute is more likely to return more results.

The proof of concept implementation uses the second of these possibilities – that of introducing a directed edge in the direction that the values were stated. Thus the first value will have an induced preference over the second. The algorithm used to do this is shown in Algorithm 5.1. Lines 2 to 10 induce an ordering in each DVP, first iterating through every DVP (line 2), then each value in each of these (line 3); for every vertex (value) its edges are checked for directionality (line 5). If they are undirected, then that edge is converted into a directed edge, with the start vertex set as the current vertex and the end vertex the vertex at the other end of the edge (line 6). Lines 12 to 17 induce an ordering between all attributes that preferences were expressed over, and it works in a similar way, checking the edges between each Value Preference vertex, setting undirected edges to be directed.

5.4.2 SUBSET CREATION

A Preference Subset, $\mathcal{PS}$, of a consumer's full Preference Set, $\mathcal{P}$, represents a portion of a consumer's preferences. Which "parts" of $\mathcal{P}$ get placed into a particular $\mathcal{P}$ is the responsibility of what we have termed the *Subset Creation Algorithm* (SCA).

The SCA is required to create one or more preference subsets. This will enable single Preference Subsets to be released in turn by the agent until a sufficient amount of results have been received (as specified by the consumer) or until the collection of preference subsets is exhausted.

The creation and population of a Preference Subset is a non-trivial task. Firstly, it has the potential of making subsets of a wide variety of sizes, ranging from a single item from $\mathcal{P}$ through to the whole of $\mathcal{P}$. The size of the subset is influenced

Algorithm 5.1 Completely Ordering a Preference Set

Require: The consumer has expressed their preferences in the form of $\mathcal{P}$

```
// First, induce order in each Discrete Value Preference's subgraph
for all DVP ∈ P do
  for all value ∈ DVP do
    for all edge connected to value do
      if edge is undirected then
        Set edge is Directed
      end if
    end for
  end for
end for
// Next, induce order between each Attribute
for all VP ∈ P do
  for all edge connected to VP do
    if edge is undirected then
      Set edge is Directed
    end if
  end for
end for
```

by two separate things: how many VPs are included in the subset, and how many values per DVP are included in the subset.

Firstly, the fewer attributes that preferences are expressed over in the preference subset, the less preference information being released; thus it will minimise the privacy loss and chances of exploitation. However, it is relatively more likely to retrieve results from a search as its focus is lost (i.e. more items will potentially match the query), thus a collection of subsets with fewer attributes will likely require fewer subsets to be released. However, if the amount of results retrieved is greater than the amount requested by the consumer the agent will have to evaluate those results returned to calculate which of those n it should present to the consumer in order to avoid the problem of information overload that preference searching is meant to solve. Conversely, the more attributes that preferences are expressed over in the preference subset the more focused the query will become mitigating this issue – but to the detriment of privacy loss and the chances for exploitation.

Secondly, the fewer items per DVP in the Preference Subset, the less preference information is being released. Thus, it will minimise the privacy loss and chances of exploitation. However, it is relatively less likely to retrieve results from a search as its range is very narrow; thus a collection of small subsets will likely require the release of many of them. This will increase the time taken for the process

to occur. Conversely, a "larger" Preference Subset (i.e. containing many of the items from P), the greater the amount of preference information is being released – thus probably needlessly compromising the privacy of the consumer whilst also increasing the chances of exploitation. On the other hand, it is relatively more likely to retrieve results for a search; thus a collection of large subsets will likely require the release of only a small amount of them. However, if amount of results retrieved is greater than the amount requested by the consumer the agent will have to evaluate those results returned to calculate which of those n it should present to the consumer in order to avoid the problem of information overload that preference searching is meant to solve.

Thus, the goal of a "good" SCA is to balance these issues and create a collection of Preference Subsets that will, when gradually released, try to achieve a balance between retrieving a set of results close to the optimal results (i.e. those that would be returned by an honest Vendor) and releasing the bare minimum amount of preference information necessary to achieve this.

Note that given the assumption stated in Section 5.2.2 that RVPs are considered to contain only a single piece of preference information, then they should be treated by an SCA in the same way that a DVP with only a single value stated would be. Thus a preference subset can either contain the RVP or not.

There are many possible approaches a SCA could take to create a collection of Preference Subsets. Three such approaches have been created for the purposes of this thesis that concretely illustrate some of the possibilities that a SCA can take and that demonstrate the basic GPR idea. The first approach (deemed the *Highly Focused Subsets* SCA) represents the "smallest" end of the preference subset scale: an SCA that creates many preference subsets of minimum size. The second approach (deemed the *Single Query* SCA) represents the "largest" end of the preference subset scale: an SCA that creates a single preference subset containing all of the consumer's preferences. Finally, the third approach (deemed the *Relax Down* SCA) represents an example of an SCA that sits somewhere in the middle of the scale. Each of these are examined in detail next.

Note that no claims are made that any of these SCAs are particularly efficient or particularly elegant; they simply demonstrate the range of possibilities that an SCA could take. More complex and clever SCAs can be imagined. For example, an SCA that could firstly estimate statistical properties about the stock database of the vendor (similar to methods employed to categorise the content of "hidden web" databases [94]) and then use this estimation to inform the way it creates

subsets can be easily imagined – although not easily implemented. Such work can clearly improve the model proposed in this thesis, and we suggest that these more advanced forms of subset creation and release should be explored in the future.

5.4.2.1 SCA - Highly Focused Subsets

The Highly Focused Subsets SCA is based on the idea that each subset should be as tightly focused as possible. Thus, for each individual preference subset sent to the vendor, both privacy loss and the chances of exploitation are minimised. However, it is likely that many subsets will have to be released to a vendor in order to retrieve enough results, since each search query will be relatively unlikely to return results. If the vendor is able to link together queries then this means this SCA is unlikely to reduce overall privacy loss by a great deal, but exploitation should be minimised if each separate query was evaluated separately, since minimising the amount of preference information released per subset reduces the potential for exploitation of that information. When the assumption is made that vendors cannot link together queries, both privacy loss and exploitation are minimised.

The approach taken by this SCA is that the Preference Subset Collection created should contain Preference Subsets that sequentially attempt to find items with a gradually decreasing preference match; i.e. the first Preference Subset should search for an item that is the consumer's "perfect" potential item (i.e. the item with characteristics that match the combination of the most preferred value of each attributes as specified by the consumer); the second Preference Subset should search for the second most preferred potential item (i.e. an item with the combination of the most preferred value for each attribute apart from the least important – which should now have the second most preferred value), etc. When the least preferred value of that attribute is reached, the next least preferred attribute is iterated through in the same way, and the process starts again. This pattern will carry on until a Preference Subset is created containing the combination of the least preferred value from each attribute; this means that all possible combinations have been exhausted.

An algorithm to achieve this is presented as Algorithm 5.2. Lines 2 and 3 set up two arrays: currentPosition and maxPosition. The first of these represents a pointer to the position of the value of each attribute that should be released, the second represents how many values there are in each attribute. Lines 6 to 9 then populate these arrays: the currentPosition array is populated entirely with

the value "1" – representing the most preferred value of each attribute – while the maxPosition array is populated with the number of values present in each attribute of $\mathcal{P}$. Next, lines 12 to 28 create all of the preference subsets, until the last possible combination of values is reached – the least preferred value of each attribute. In this loop, line 13 creates a new $\mathcal{PS}$, and lines 16 to 18 take the value at the current position pointer for each attribute and inserts them into the $\mathcal{PS}$. The remainder of the work to be done is in setting up the position pointers ready for the next time around the loop. Lines 22 to 27 then iterate over each of the attributes, checking to see whether the position pointer has overflowed past the last value. If it has, lines 23 to 26 set the position pointer for that attribute back to "1" – the most preferred value – and shifts the position pointer to the next attribute. Finally, lines 31 to 35 insert the least preferred values of each attribute to create the final preference subset.

Generally, this SCA can create a collection up to $\prod_{i=1}^{|\mathcal{P}|} ||VP_i||$ Preference Subsets – potentially a very large amount for any non-trivially sized $\mathcal{P}$. Each Preference Subset created will contain $|\mathcal{P}|$ items of preference information. For example, suppose a consumer expresses the following preferences: $Make$ = $Toyota$ & $Ford$; $Colour$ = $Silver$ & $Black$ & Red; $Gears$ = $Manual$ & $Automatic$; and AP = $Make$ & $Colour$ & $Gears$. The *Highly Focused Subsets* SCA iterates through the Preference Set as shown in Figure 5.7, and creates the corresponding queries shown in Figure 5.8.

5.4.2.2 SCA - Single Query

The Single Query SCA is based on the extreme possibility of minimising the amount of Preference Subsets created and sent – to only send one containing essentially the consumer's full preference set, $\mathcal{P}$, but with all preference ordering information converted to a Pareto preference, so that when this is sent to the vendor, the chances for exploitability are lessened as the vendor lacks vital preference information – it will know which preferred values the consumer has specified, but not know which of these values are more preferred. However, the amount of preferred values released to the vendor is maximal in this approach. Thus, this SCA is more effective in combating exploitation than privacy loss.

The approach taken by this SCA is that the Preference Subset Colletion will only create one Preference Subset which will contain all values held in $\mathcal{P}$, but with all values holding an equally preferred status. An algorithm that will create the

Algorithm 5.2 GPR SCA - Highly Focused Subsets

1: // Create Arrays
2: Create array $currentPosition[]$ with size $|\mathcal{P}|$
3: Create array $maxPosition[]$ with size $|\mathcal{P}|$
4: .
5: // Populate Arrays with pointers to most preferred values
6: **for** $i = 1$ to $i = |\mathcal{P}|$ **do** {For each attribute in $\mathcal{P}$}
7: $currentPosition[i] \Leftarrow 1$ {Most preferred value of this attribute}
8: $maxPosition[i] \Leftarrow ||VP_i||$ {Amount of values in this attribute}
9: **end for**
10: .
11: // Create Preference Subsets
12: **while** ($currentPosition \neq maxPosition$) **do** {While not at the least preferred element of each attribute}
13: Create new empty $\mathcal{PS}$
14: .
15: // Insert the values at current pointer positions into $\mathcal{PS}$
16: **for** ($i = 1$ to $i = |\mathcal{P}|$) **do** {For each attribute in P}
17: $VP'_{i1} \Leftarrow VP_{\{i\}\{currentPosition\}}$
18: **end for**
19: .
20: // Set current position pointer to next least preferred configuration, ready for next $\mathcal{PS}$
21: $currentPosition[|\mathcal{P}|]++$ {Increment current pointer of least important attribute}
22: **for** $i = |\mathcal{P}|$ to $i = 1$ **do** {For each attribute in $\mathcal{P}$}
23: **if** $currentPosition[i] > maxPosition[i]$ **then** {If pointer for this attribute has overflowed past least preferred value}
24: $currentPosition[i] \Leftarrow 1$ {Go back to the start for this attribute}
25: $currentPosition[i-1]++$ {Go to next value in the next attribute}
26: **end if**
27: **end for**
28: **end while**
29: .
30: // Insert the values at final pointer positions into the final $\mathcal{PS}$
31: Create new empty $\mathcal{PS}$
32: .
33: **for** ($i = 1$ to $i = |\mathcal{P}|$) **do** {For each attribute in $\mathcal{P}$}
34: $VP'_{i1} \Leftarrow VP_{\{i\}\{currentPosition\}}$
35: **end for**

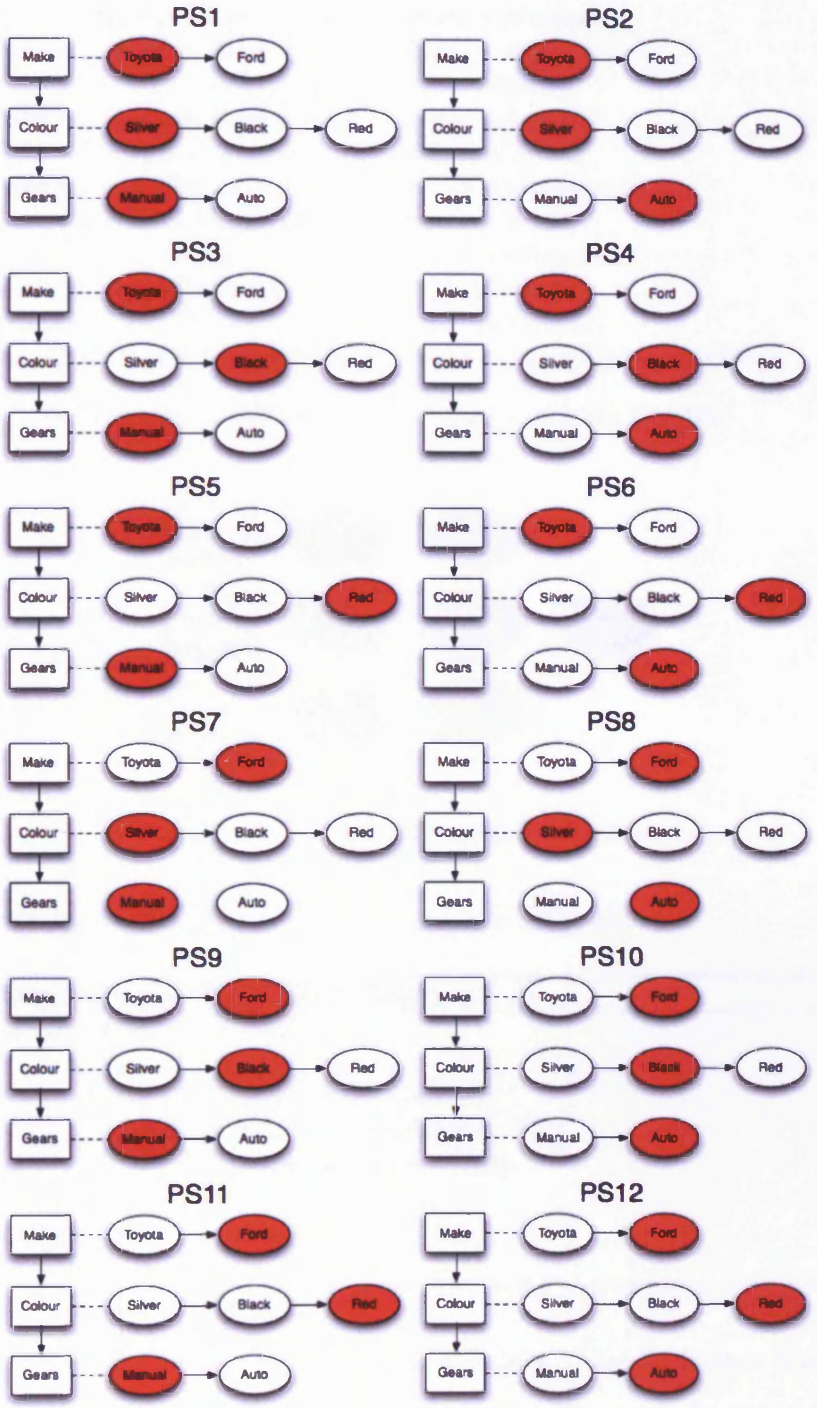

Figure 5.7: HFS SCA - Iterative Process

PS1: *Make* = *Toyota*, *Colour* = *Silver*, *Gears* = *Manual*
PS2: *Make* = *Toyota*, *Colour* = *Silver*, *Gears* = *Automatic*
PS3: *Make* = *Toyota*, *Colour* = *Black*, *Gears* = *Manual*
PS4: *Make* = *Toyota*, *Colour* = *Black*, *Gears* = *Automatic*
PS5: *Make* = *Toyota*, *Colour* = *Red*, *Gears* = *Manual*
PS6: *Make* = *Toyota*, *Colour* = *Red*, *Gears* = *Automatic*
PS7: *Make* = *Ford*, *Colour* = *Silver*, *Gears* = *Manual*
PS8: *Make* = *Ford*, *Colour* = *Silver*, *Gears* = *Automatic*
PS9: *Make* = *Ford*, *Colour* = *Black*, *Gears* = *Manual*
PS10:*Make* = *Ford*, *Colour* = *Black*, *Gears* = *Automatic*
PS11: *Make* = *Ford*, *Colour* = *Red*, *Gears* = *Manual*
PS12: *Make* = *Ford*, *Colour* = *Red*, *Gears* = *Automatic*
Note: For all of the Preference Subsets, AP = *Make* & *Colour* & *Gears*

Figure 5.8: HFS SCA - Preference Subsets Created

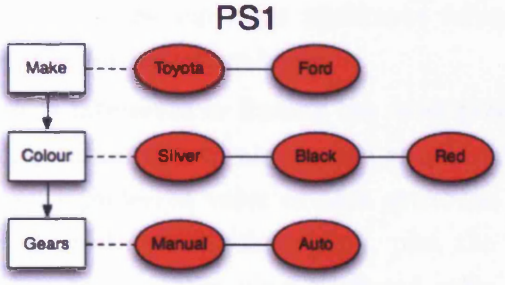

Figure 5.9: SQ SCA - Iterative Process

subset collection using this approach is shown in Algorithm 5.3.

Algorithm 5.3 GPR SCA - Single Query

```
1: Create new empty PS
2: .
3: // Insert all values of all attributes into PS
4: for i = 1 to i = |P| do {For each attribute in P}
5:    for j = 1 to j = ||VP_i|| do {For each value}
6:       VP'_ij ⇐ VP_ij {Add value to VP_i}
7:    end for
8: end for
```

This SCA will create a single Preference Subset. This Preference Subset will contain $\sum_{i=1}^{|\mathcal{P}|} ||DVP_i||$ items of preference information. For example, suppose a consumer expresses a preference set as in the previous example. The *Single Query* SCA iterates through $\mathcal{P}$ as shown in Figure 5.9, and creates the corresponding queries shown in Figure 5.10.

PS1: $Make = Toyota \otimes Ford$, $Colour = Silver \otimes Black \otimes Red$, $Gears = Manual \otimes Automatic$
Note: $AP = Make \,\&\, Colour \,\&\, Gears$

Figure 5.10: SQ SCA - Preference Subsets Created

5.4.2.3 SCA - Relax Down

The Relax Down SCA sits somewhere in between the two extremes of the previous two SCAs. This starts with the absolute minimum amount of preference information in a single Preference Subset (as per the first $\mathcal{PS}$ created by the Highly Focused Subsets SCA), but then gradually relaxes the search criteria by adding a value at a time into the Preference Subset - i.e. it adds more and more preference information, in the order of less important preference values to most important preference values, into the subsequent subsets.

Since this approach is interested in finding the most preferred results for the consumer, then the first subset should aim to find an item that is the consumers ideal item, i.e. the most preferred value of each attribute. The second subset should aim to find the results of the first query, plus the next most desirable possibility for the consumer, i.e. the most preferred value from each attribute apart from the least preferred attribute, which should contain the first and second most preferred values. This will continue, until all values of the least preferred attribute are added, at which point the next least preferred attribute is relaxed by choosing its next value, and the process starts again. This pattern will carry on until a subset is created containing the all preferred values from all attributes.

An algorithm that will create subsets using this approach is presented as Algorithm 5.4. Lines 2 and 3 create two pointers, *currentAttr* and *currentValue*, which combine to point to the location in $\mathcal{P}$ that the algorithm has currently relaxed to. Since the algorithm starts out fully unrelaxed, at the consumer's ideal item, then the initial values populated point to the first value of the last attribute. Next, a while loop between lines 6 and 39 creates the Preference Subsets, one at a time, until there are no more values to relax in $\mathcal{P}$. In this loop, lines 10 to 30 iterate through each attribute and inserts the correct values for that attribute into $\mathcal{PS}$ – only the first value if the attribute has yet to be relaxed (lines 13 to 15), the values up until the *currentValue* pointer if the attribute is in the process of being relaxed (lines 18 to 22), and all values in the attribute if that attribute has already been fully relaxed (lines 25 to 29). Finally, lines 35 to 38 check to see whether the pointer is now pointing past the last value, and if it is it moves the pointer to the

first value of the next attribute to relax.

Algorithm 5.4 GPR SCA - Relax Down

```
1: // Create pointers and populate inital values
2: currentAttr ⇐ |P| {Start at the last Attribute}
3: currentValue ⇐ 1 {Start at its first value}
4: .
5: // Create Preference Subsets
6: while currentAttr ≠ 0 do {While there are still attributes and values to relax}
7:    Create new empty PS
8:    .
9:    // Iterate through each attribute
10:   for i = 1 to i = |P| do {For each Attribute in P}
11:      .
12:      // If the attribute hasn't yet been relaxed, only insert the first value
13:      if currentAttr < i then
14:         VP'_i1 ⇐ VP_i1
15:      end if
16:      .
17:      // If the attribute is the one currently being relaxed, insert its values up
         until the current position pointer
18:      if currentAttr = i then
19:         for j = 1 to j = currentValue do {For each relaxed value in VP_i}
20:            VP'_ij ⇐ VP_ij
21:         end for
22:      end if
23:      .
24:      // If the attribute has already been relaxed, insert all of its values
25:      if currentAttr > i then
26:         for j = 1 to j = ||VP_i|| do {For each value in VP_i}
27:            VP'_ij ⇐ VP_ij
28:         end for
29:      end if
30:   end for
31:   .
32:   // Relax the pointer ready by moving the pointer to the next value
33:   currentValue + +
34:   // If the pointer overflowed (is pointing past the last value) then move the
      pointer to the first item of the next attribute
35:   if currentValue > ||VP_i|| then
36:      currentValue ⇐ 1
37:      currentAttr − −
38:   end if
39: end while
```

This SCA will create up to $\sum_{i=1}^{|\mathcal{P}|} \|VP_i\|$ Preference Subsets, and the size of the Preference Subsets created will range from $|\mathcal{P}|$ items of preference information in

the first subset, up to $\sum_{i=1}^{|\mathcal{P}|} \|VP_i\|$ items of preference information. For example, suppose a consumer expresses a preference set as in the previous example. The *Relax Down* SCA iterates through $\mathcal{P}$ as shown in Figure 5.11, and creates the corresponding queries shown in Figure 5.12.

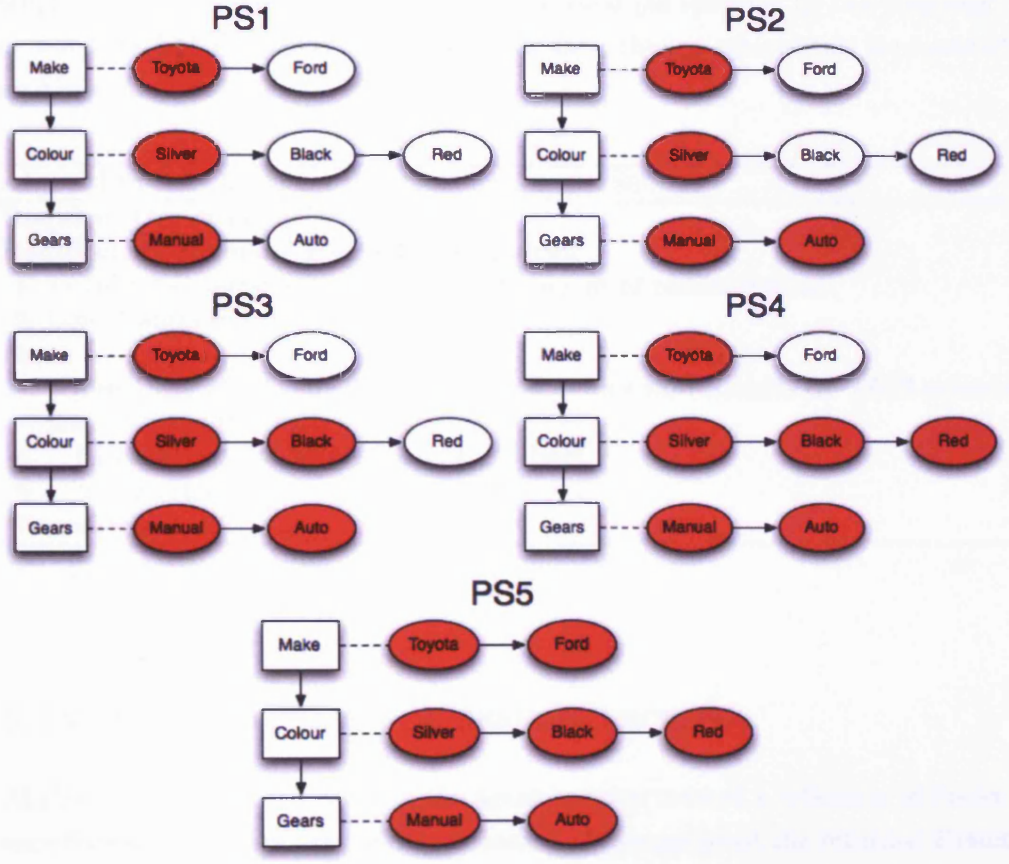

Figure 5.11: RD SCA - Iterative Process

PS1: *Make* = *Toyota*, *Colour* = *Silver*, *Gears* = *Manual*
PS2: *Make* = *Toyota*, *Colour* = *Silver*, *Gears* = *Manual* & *Automatic*
PS3: *Make* = *Toyota*, *Colour* = *Silver* & *Black*, *Gears* = *Manual* & *Automatic*
PS4: *Make* = *Toyota*, *Colour* = *Silver* & *Black* & *Red*, *Gears* = *Manual* & *Automatic*
PS5: *Make* = *Toyota* & *Ford*, *Colour* = *Silver* & *Black* & *Red*, *Gears* = *Manual* & *Automatic*
Note: *AP* = *Make* & *Colour* & *Gears*

Figure 5.12: RD SCA - Preference Subsets Created

5.4.3 GRADUAL RELEASE OF SUBSETS

Given that the agent has created a collection of Preference Subsets using a particular SCA, the next step is for these Preference Subsets to be gradually released to the vendor. The agent will do this, working its way from the first $\mathcal{PS}$ to the last, stopping either when enough results are returned (as specified by the consumer) or when the last $\mathcal{PS}$ has been released. An algorithm to achieve this is presented as Algorithm 5.5.

Algorithm 5.5 Gradually Releasing Preference Subsets

Require: One or more PS
Require: $maxResults$ specified by Consumer
1: // Initialise variable to keep track of amount of results fetched
2: $numResults \Leftarrow 0$
3: .
4: **while** More PS Available AND $numResults < maxResults$ **do** {Still subsets and not enough results}
5: Release next PS to vendor, get RS back
6: $numResults \Leftarrow numResults + |RS|$
7: **end while**

5.4.4 POST-PROCESSING OF RESULTS RECEIVED

At this point in the GPR process, the agent has now created a collection of Preference Subsets to release, has released them, and has gathered the returned Result Set. The SCA will have attempted to ensure that the amount of results in that result set is equal to the number of results requested by the consumer, however, if the last query sent retrieved multiple items then it is possible that it could contain more. In this case, the agent will need to assess the Result Set w.r.t $\mathcal{P}$, so that it can show the consumer only the amount of items they requested (so as to stop information overload).

Given that the Highly Focused Subsets SCA is essentially a brute-force method of calculating a preferred ordering of items, the agent can use the same algorithm to calculate the most preferred n items out of the p items in the Result Set. The use of this method does not represent a particularly efficient way of performing the necessary calculations, given the relative inefficiency of the HFS SCA; it should, however, produce the required results for the proof of concept implementation.

5.5 SUMMARY

The idea of Gradual Partial Release of a consumer's preferences to a vendor has been developed and detailed, and the specific aspects of how to achieve this idea have been discussed. In order to be able to describe the approach in full, a series of definitions necessary for describing various aspects of consumer, vendor, and preference searching was detailed. A series of measures was discussed that will enable the evaluation of the effectiveness of different GPR algorithms. Finally, three examples of GPR algorithms were detailed, including one which will emulate the results (but not the process) of the current paradigm of preference searching, in order to evaluate the relative effectiveness of the approach. These GPR algorithms are significant as they are the mechanism by which the release of a consumer's preferences to enhance search can be controlled in a privacy aware manner.

Chapter 6

Results and Evaluation

The Gradual Partial Release (GPR) approach to preference searching was detailed in Chapter 5. This approach was implemented in the form of a proof of concept system; this was then used as the basis for a process of evaluation. This chapter presents the details of this evaluation firstly of the GPR approach as a whole, and secondly of the specific GPR algorithms in particular.

The evaluation process was split into two main areas. The first area looked at the overall effectiveness of the proposed GPR approach for preference searching (using all three GPR algorithms), examining how well it worked with respect to the main desired outcomes of reducing privacy loss and exploitation; including how this was affected by varying input. The second area looked at its efficiency, examining how the GPR approach (using all three GPR algorithms) worked in terms of speed, network traffic, and other quantitative measures.

This chapter first discusses the environment in which the experiments were performed, including how problem instances and test data were generated; this is followed by the results of the evaluation itself and conclusions consequently drawn.

6.1 Experimental Setup

This section describes the various components of the setup used to perform experiments aimed at evaluating the effectiveness of the GPR approach. This includes what is actually going to be tested, the data that is going to be used within the experiments, and the environment in which the experiments were performed.

6.1.1 Overview of Experiments

Analysis of the GPR approach was split into two main phases of experiments: firstly analysing the overall effectiveness of the GPR approach with regards to the main aims of the GPR approach (minimising privacy loss and exploitation); and secondly analysing how different types of input (i.e. differently sized and configured vendor catalogues and consumer preference sets) affect the effectiveness and efficiency of each of the GPR algorithms.

6.1.2 Generation of Problem Instances

When considering experiments designed to evaluate the GPR approach, there are two sets of data that must first be present: the consumer's preferences, and the vendor's catalogue. In all experiments both sets of data will be generated in order to exert complete control over statistical aspects of the data sets used and thus allow more meaningful evaluation to take place. Some discussion of how this data was generated is necessary before results are discussed, in order that the context of the results can be fully understood.

6.1.2.1 Generating the Vendor's Catalogue

The first set of data necessary for each experiment is the vendor's catalogue. During experiments, the structure of the vendor's catalogue is as follows:

- An integer "ID" field. This will hold a unique identifier for each item in the catalogue, and will be the primary key of the table;
- A series of n integer fields, labelled "Attr_n", where n is the number of attributes specified in each experiment; these fields will hold the values of each attribute describing each item in the catalogue; and
- A decimal "price" field. This will hold the price of each item in the catalogue.

The catalogue was configured in this way so as to be a representation of a completely generic vendor's catalogue, with no bias from any specific type of attribute. A price field is included separately as an item's price is represented in a decimal format in order to more closely match the real world.

EXAMPLE 2

An example of what the structure of a vendor's catalogue with 4 attribute fields specified is shown in the following table:

id	*Attr_1*	*Attr_2*	*Attr_3*	*Attr_4*	*Price*

When populating the table for each experiment, the following rules are used to create the values:

- For each of the main attribute fields, an integer value will be randomly created between 1 and k, where k is the number of values per attribute as specified for each experiment. The method used to create this random value uses a uniform distribution or a normal distribution as required by the experiment. The reason for using a uniform distribution is to create items in the catalogue representing a wide range of unbiased attribute information, while a normal distribution is used to evaluate whether the change in distribution changes the outcome of the experiment. If a normal distribution is used, the mean value is taken as $k/2$, with the standard deviation as $k/6$, ensuring a good spread of data between 1 and k.

- A decimal value will be randomly created for the price field between j and k, where j and k are the minimum and maximum price as specified for each experiment. The method used to create this random value uses either a uniform distribution or a normal distribution, as specified for each experiment. If a normal distribution is used, a mean value and standard deviation is also required as an input to the data generation method. The reason for using a uniform distribution is, as with the main attribute fields, to create items in the catalogue with a wide range of unbiased pricing information. Separately, using a normal distribution in a separate set of experiments will allow the evaluation to see if having items clustered around certain prices makes a difference in the effectiveness or efficiency of the GPR approach.

This generated data attempts to represent a highly generic catalogue of available items with no particular bias as to type of attribute that could possibly describe such items.

EXAMPLE 3

An example of what an excerpt of a vendor's catalogue with 4 attribute fields and 3 values per attribute might look like is shown in the following table:

id	*Attr_1*	*Attr_2*	*Attr_3*	*Attr_4*	*Price*
35	*1*	*2*	*1*	*3*	*£9,458*
36	*3*	*3*	*2*	*1*	*£7,209*
37	*3*	*2*	*3*	*3*	*£2,374*
38	*2*	*1*	*2*	*2*	*£8,093*

6.1.2.2 Generating the Consumer's Preferences

The second set of data necessary for each experiment is the consumer preference set that would be used to search the vendor's catalogue. Preferences are specified over m of the n attributes in the vendor's catalogue, where m is specified in each experiment. The values in each attribute chosen will be integers from 1 to k, where k is the number of values per attribute specified in each experiment. Relative preferences between the items can all be expressed as prioritised relations (&). The values simply represent a subset of the values available in the vendor's catalogue, and do not need a more complex generation method since the values these would correlate with in the vendor's catalogue are already randomly generated; thus this will be sufficient for testing the effectiveness of the GPR approach. As for the preferred ordering within the preferences, recall that prior to any GPR process the consumer's preferences will go through a process of having a complete order imposed upon them, as described in Section 5.4.1.3; thus whatever relative preferences are specified the preference set will end up being completely ordered anyway.

An example of what a preference set with 3 attribute fields and 4 values per attribute might look like is shown in Figure 6.1.

6.1.3 VENDOR EXPLOITATION

In order to assess whether the GPR approach has minimised the amount of exploitation that could potentially occur, then the vendor in our experiments will need to attempt such exploitation where possible. A simple exploitation model has been used in all experiments – the vendor is attempting to maximise profit by

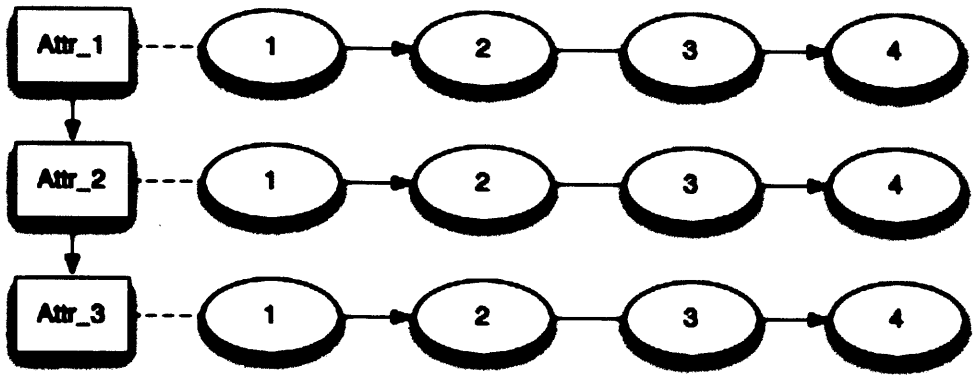

Figure 6.1: Example of Generated Preference Set

taking the original, unexploited result set, RS_u that it calculates, and dropping 20% of the items within that have the lowest prices.

6.1.4 ENVIRONMENT

The proof of concept software implementation was christened "PEEPS" – Privacy Enabled Enhanced Preference Searching – and was implemented in the Java programming language. All java code implemented was written using Sun's JRE 6.0 specification.

All experiments with the software took place on the 64 bit Sun 1.6.0 (r4) JVM for Linux; this JVM in turn ran on a 64 bit RHEL 5.2 OS. The database that the Vendor Agent connected to to store its catalogue was MySQL v5.0.45, connecting to Java using the MySQL Connector/j 3.1, distributed by MySQL AB. The MySQL instance ran on the same machine as the JVM in order to lessen the effects of network latency when evaluating the software performance.

The hardware used to run these experiments on was one quad-core Intel Core2 Extreme QX6700 (4x 2.66GHz cores with 8MB L2 cache), 4GB 800MHz DDR2 memory, and a pair of 500GB E-SATA300 hard disks formatted with the XFS file system.

6.2 EFFECTIVENESS OF GPR APPROACH

The first phase of analysis aimed to understanding how well the GPR approach worked in general – and the three GPR algorithms presented in Section 5.4.2 in

particular – perform with respect to the two main aims of minimising consumer privacy loss and exploitation. Comparisons were drawn between the GPR approach and the existing approach to preference searching as one of the GPR algorithms (the Single Query algorithm) was designed to mimic the outcomes of the existing approach.

6.2.1 Overview of Experiment

A series of individual experiments were set up and performed, where for each experiment all variables (vendor catalogue size, number of attributes, etc) were created afresh using a randomising function, and a new set of test data (i.e. a vendor catalogue and a corresponding consumer preference set) was generated. These experiments were performed in two groups – firstly with the random data generation method used to create the stock catalogue using a uniform distribution, and secondly using a normal distribution.

A benchmarking process was first run upon this set of test data which calculated the actual top 10 preferred results in the vendor's catalogue w.r.t. the consumer's preferences. Next, each GPR algorithm was then run which attempted to obtain the top 10 preferred results, where the vendor was using the simple exploitation algorithm previously discussed. For each algorithm the privacy loss and exploitation that occurred was measured and recorded, along with the time taken and the amount of network traffic produced. 3000 of these separate experiments were performed for each of the two groups of experiments.

The reasoning behind this experimental design was to produce a set of results that could quantifiably demonstrate whether the GPR approach in general, and each GPR algorithm in particular, was able to reduce the levels of privacy loss and exploitation that would be seen if they were used as compared to the existing approach. The series of random scenarios created aimed to lessen any bias that may occur if the values of the variables were accidentally chosen such that they described a scenario particularly well suited for a particular algorithm. Using both a uniform then a normal distribution allows for evaluation as to whether the properties of the vendor's catalogue change the effectivess of the GPR approach.

6.2.2 EXPERIMENT DETAILS

In this first set of experiments, the random function used (the Apache Commons[18] RandomData[19] java classes) used a uniform or a normal distribution to create the values that constrained the variables describing the experiment. The limits imposed upon each of the variables were chosen manually simply to enact restrictions around the process and were designed to enable a very wide range of possibilities. The limits were as follows:

- Vendor Catalogue parameters:
  - Database size: Random, between 1 and 250,000
  - Number of Attributes, n: Random, between 5 and 30
  - Number of Values per Attribute, k: Random, between 5 and 30
  - Price: Random, between £999.99 and £49,999.99
- Consumer Preference parameters:
  - Number of Attributes: Random, between 1 and n
  - Number of Values per Attributes: Random, between 1 and k
  - Number of items to fetch: 10

The test data generated to populate the vendor's catalogue and consumer's preferences to fit each experiment was created as discussed in Section 6.1.2.

6.2.3 RESULTS

Discussion of the results gathered when using a uniform distribution to generate the vendor's catalogue while performing the series of experiments is split into the two main evaluative areas – privacy loss and exploitation.

6.2.3.1 Privacy Loss

Table 6.1 shows the basic overall statistics of the privacy loss during the experiments, Figure 6.2 graphs the mean privacy loss, and Figure 6.3 graphs the distribution of privacy loss.

Note that Privacy Loss is measured as defined in Equation 5.2 in Section 5.2.4.

[18] http://commons.apache.org/
[19] Package org.apache.commons.math.random

	Mean	(Standard Deviation)	Min	Max	Total Privacy Loss seen
GPR Approach (Highly Focused Subsets)	0.16	0.04	0.13	1	0.1% of experiments
GPR Approach (Relax Down)	0.26	0.16	0.13	1	0.2% of experiments
Existing Approach (Single Query)	1	0	1	1	100% of experiments

Table 6.1: Privacy Loss, uniformly distributed stock

Some degree of privacy loss occurred in 100% of experiments for all GPR algorithms. This was as expected, since in order to perform any search enhanced with preferences, some degree of preference information has to be released.

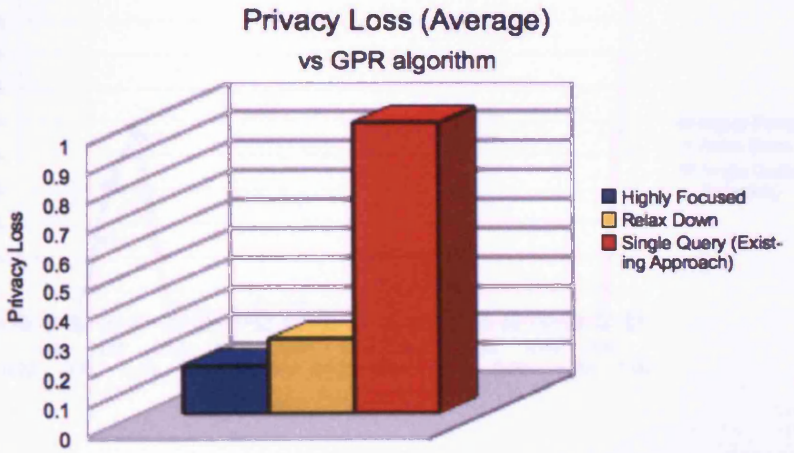

Figure 6.2: Average Privacy Loss, uniformly distributed stock

The GPR algorithm that imitates the existing approach (the Single Query algorithm) shows a consistent level of privacy loss of 1.0 in all experiments – i.e. total privacy loss at all times. This was as expected, as this algorithm releases all preference information in the first (and only) query it sends.

The Highly Focused Subsets and Relax Down GPR algorithms, however, show a reduction in the level of privacy loss in almost all cases: the Highly Focused Subsets algorithm in around 99.9% of cases and the Relax Down algorithm showing a reduction in over 99.8% of cases. The Highly Focused Subsets algorithm was seen to generally reduce the amount of privacy loss to about one fifth of that of the existing approach; the Relax Down algorithm to between one quarter of that of the existing approach. This represents a significant reduction in privacy loss of the consumer's preferences.

The distribution seen for the Highly Focused Subsets algorithm is a result of the fact that it releases a preference subset containing an amount of preference information that is equal to the amount of attributes that preferences were expressed over (it releases a single value per attribute). Thus the privacy loss seen is a direct proportion of this amount of information and the amount of values expressed in total in $\mathcal{P}$. Given the bounds established in these experiments for the random creation of a consumer's preference set, this averages out to the figures shown in the graph of distribution. The distribution seen for the Relax Down algorithm is also affected similarly due to the way it creates the preference subsets.

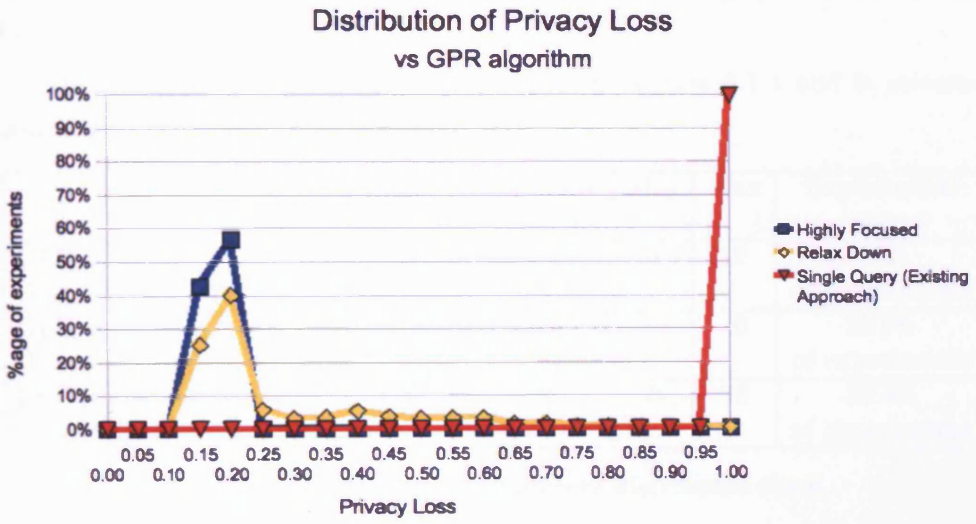

Figure 6.3: Distribution of Privacy Loss, uniformly distributed stock

Note that for normally distributed catalogue data similar patterns in the values and distribution of privacy loss were observed during experimentation, and are thus not detailed further here.

When compared with the existing approach, the two GPR algorithms that implement the idea of the GPR approach demonstrably reduced the loss of preference information, and therefore privacy loss, of a consumer's preferences as compared to the existing approach in the vast majority of experiments. The Highly Focused Subsets algorithm performed better by reducing the amount of privacy loss to a greater extent, managing to give away less than a quarter of the consumer's preferences at any one time; the Relax Down algorithm fared less well but still managed a reduction of privacy loss by generally giving away less than a half of the consumer's preferences. This is as expected due to the manner in which the two algorithms work: the former sending very highly focused subsets all contain-

ing a minimum of preference information, while the latter gradually increases the amount of preference information released as it iterates through subsets. Thus, it has been demonstrated that the GPR approach has the potential to significantly reduce privacy loss as compared to the current approach, and that if a consumer wishes to maximise privacy loss the Highly Focused algorithm is the most effective.

6.2.3.2 Exploitation

Table 6.2 shows the basic overall statistics of exploitation during the experiments, Figure 6.4 graphs the mean exploitation, while Figure 6.5 graphs the distribution of exploitation.

Note that exploitation occurs as described in Section 6.1.3 and is measured according to Equation 5.5 in Section 5.3.1.

	Mean	(Standard Deviation)	Min	Max	Exploitation seen
GPR Approach (Highly Focused Subsets)	0.26	1.39	0	10	8.7% of experiments
GPR Approach (Relax Down)	0.92	2.54	0	10	26.9% of experiments
Existing Approach (Single Query)	1.38	2.88	0	10	33.3% of experiments

Table 6.2: Exploitation, uniformly distributed stock

Exploitation was seen in roughly one third of experiments where the stock catalogue data was uniformly distributed, and roughly two thirds where normally distributed. This difference will be due to the fact that a stock catalogue with normally distributed data will have groups of items with similar attribute values, thus for queries which find those groups of items exploitation will be easier to perform since the vendor has a wider choice of items it can remove from the result set.

The GPR algorithm that imitates the existing approach (the Single Query algorithm) shows exploitation in all experiments where exploitation was seen by any algorithm. The average exploitation was measured at 1.38 (standard deviation of 2.88) for uniformly distributed stock catalogue data, and 2.41 (standard deviation of 3.43) for normally distributed.

The algorithms that implement the new GPR approach, however, reduce exploitation in roughly 90% of experiments where the Highly Focused Subsets algorithm is used, and roughly 30% of experiments where the Relax Down algorithm

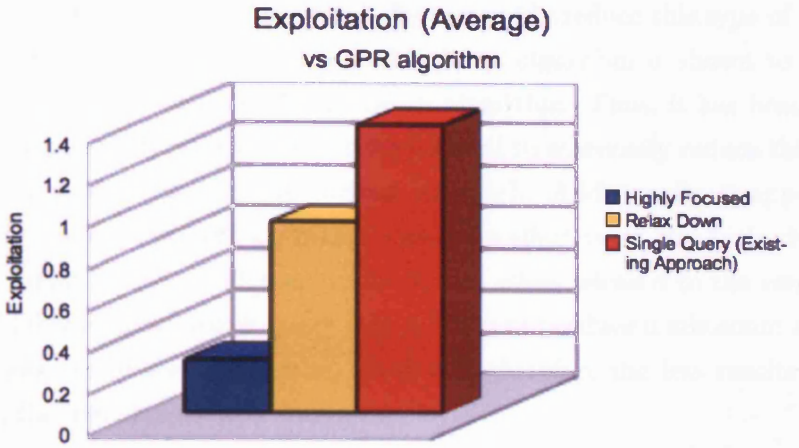

Figure 6.4: Average Exploitation, uniformly distributed stock

is used against uniformly distributed stock data and 50% against normally distributed stock data. In fact, the Highly Focused Subsets algorithm completely eradicates exploitation in more than two thirds of experiments where it occurred with the existing approach; the Relax Down algorithm completely eradicated it in around 20%. In those cases where exploitation was not eradicated completely, it was reduced – the Highly Focused Subsets algorithm significantly reducing the amount seen; and the Relax Down algorithm reducing it to a lesser, but demonstrable, amount.

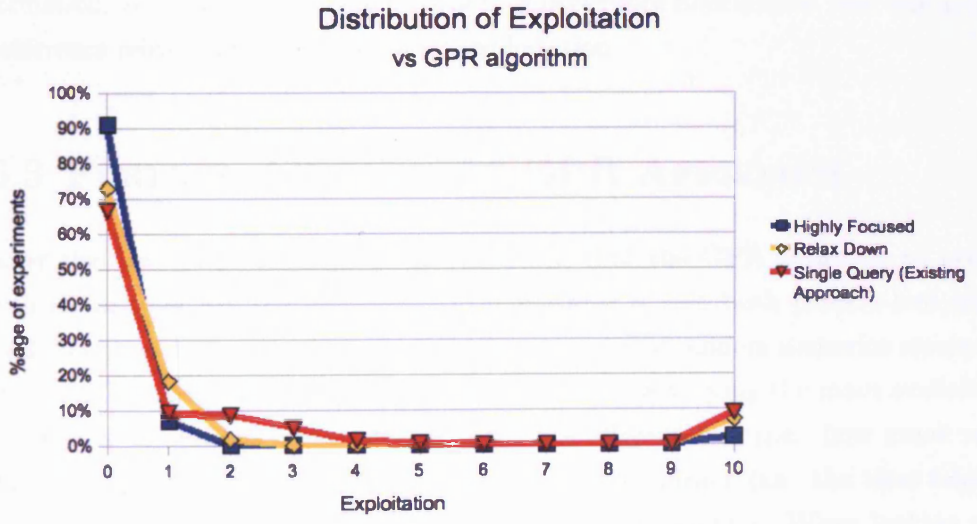

Figure 6.5: Distribution of Exploitation, uniformly distributed stock

When compared with the existing approach, the two GPR algorithms that im-

plement the idea of the GPR approach demonstrably reduce this type of exploitation. As expected, the Highly Focused Subsets algorithm is shown to be more effective at this task than the Relax Down algorithm. Thus, it has been demonstrated that the GPR approach has the potential to materially reduce this type of exploitation as compared to the current approach. Additionally, it appears that the Highly Focused Subsets algorithm was more effective at this task, due to the fact that highly focused preference subsets will, when released to the vendor, represent a highly focused search query that is likely to produce a minimum of results; and the less preference information given and therefore the less results that are returned, the less exploitation can occur.

6.2.4 SUMMARY

The results presented in this section have shown that while it is by no means guaranteed that the GPR approach to preference searching will always reduce privacy loss and exploitation as compared to the current approach to preference searching, it has been demonstrated that it is likely to do so in the majority of cases – and the amount of reduction seen can be significant.

Of the algorithms that implement the GPR approach, the Highly Focused Subsets algorithm was shown to be most effective at reducing both privacy loss and exploitation, due to the method that algorithm employs in releasing preference information: sending highly focused subsets of preference information that minimise preference release and the chances for exploitation.

6.3 FURTHER ANALYSIS OF GPR APPROACH

After the first phase of analysis demonstrated that the GPR approach to preference searching has the potential to significantly reduce both privacy loss and exploitation of a consumer's preferences in the series of random scenarios created, the second phase of analysis aimed to understand how varying the main available parameters of such scenarios affected both the effectiveness (i.e. how much reduction in privacy and exploitation occurred) and efficiency (i.e. the time taken and communications incurred) of each of the GPR algorithms. When looking at the effectiveness of the GPR approach, comparisons were drawn between the GPR approach and the existing approach to preference searching as one of the GPR algorithms (the Single Query algorithm) was designed to mimic the output of

the existing approach. Such comparisons cannot, however, be drawn with regards efficiency since this algorithm does not mimic the actual computational implementation of the existing approach – and is undoubtedly far less efficient than these relatively mature implementations. Given that this work is a proof of concept and makes no claims as the relative efficiency of the approach, however, this is not seen as an issue.

6.3.1 Overview of Experiment

A series of individual experiments were set up and performed. Fixed values for each of the five parameters (vendor catalogue size, number of attributes, etc) were chosen initially. Then, for each of these five parameters, the fixed values chosen for the other four parameters were used while the remaining one was varied from a low to a high number over a series of intervals. For each of the chosen intervals 1000 experiments were performed, where for each a new set of test data (i.e. a vendor catalogue and a corresponding consumer preference set) was generated given the same values for the five parameters.

For each of these experiments, a benchmarking process was first run upon the set of test data which calculated the top 10 preferred results in the vendor's catalogue w.r.t. the consumer's preferences. Next, each GPR algorithm was then run which attempted to obtain the top 10 preferred results, where the vendor was using the simple exploitation algorithm previously discussed. For each algorithm the privacy loss and exploitation that occurred were measured and recorded, along with the time taken and the amount of network traffic produced.

The reasoning behind these experiments was to produce a set of results that quantifiably demonstrate how the effectiveness and efficiency of each of the GPR algorithms is effected given varying input. Varying a single variable at a time, while the others remained fixed makes the results seen to show this information, while running many experiments for each of the varied values chosen and taking average results across each series enables the results to be as unbiased as possible by any particular random sets of test data generated.

6.3.2 Experiment Details

The fixed values chosen for the experiment parameters were as follows: a vendor catalogue size of 30,000, with a number of attributes of 15 and number of values per attribute of 15; a consumer preference set with a number of attributes of 5

and a number of values per attribute of 5. These particular numbers were chosen as representing a scenario that is not too simple, somewhat realistic, and that the benchmarking process and all GPR algorithms can work through relatively quickly. The focus of this set of experiments is to understand trends that may emerge as parameter values vary, rather than looking at the actual numbers reported during the experiments. This means that any bias that may be present in the scenario that these parameter values describe should not be an issue.

The random function used a uniform distribution to create the values that constrained the values of the parameters describing the experiment. The variance of the parameters (minimum, maximum, and interval) were chosen manually to allow a range of values to be seen but that enable efficient evaluation. The parameter variance chosen were as follows:

- Vendor Catalogue parameters:
  - Database size: 10,000 to 100,000; intervals of 10,000
  - Number of Attributes, n: 5 to 24; intervals of 1
  - Number of Values per Attribute, k: 7 to 16, intervals of 1
- Consumer Preference parameters:
  - Number of Attributes: 1 to 10, increments of 1
  - Number of Values per Attributes: 5 to 14; increments of 1

For each of the parameter values, 1000 experiments were performed, where each experiment used a newly created set of test data using the same parameter values. Thus, 10,000 experiments were performed when varying the vendor's catalogue size; 20,000 experiments when varying the number of attributes in the vendor's catalogue; 10,000 experiments when varying the number of values per attribute in the vendor's catalogue; 10,000 experiments when varying the number of attributes in the consumer's preferences; and 10,000 experiments when varying the number of values per attribute in the consumer's preferences.

The test data generated to populate the vendor's catalogue and consumer's preferences to fit each experiment was created as previously discussed, using the uniform method when generating the vendor's calogue. Note that only a single stock catalogue distribution was considered in order to keep the series of experiments unbiased.

6.3.3 Results

The results of this set of experiments are split into discussing the results of each of the parameters varied.

6.3.3.1 Varying Size of Vendor's Catalogue

The first of the areas examined look at how the effectiveness and efficiency of the GPR algorithms is affected by the size of the vendor's catalogue. Figures 6.6 - 6.7 show how privacy loss and exploitation were affected as the size of the vendor's catalogue changes, while Figures 6.8 - 6.10 show how runtime, number of queries generated, and communications traffic generated were consequently affected.

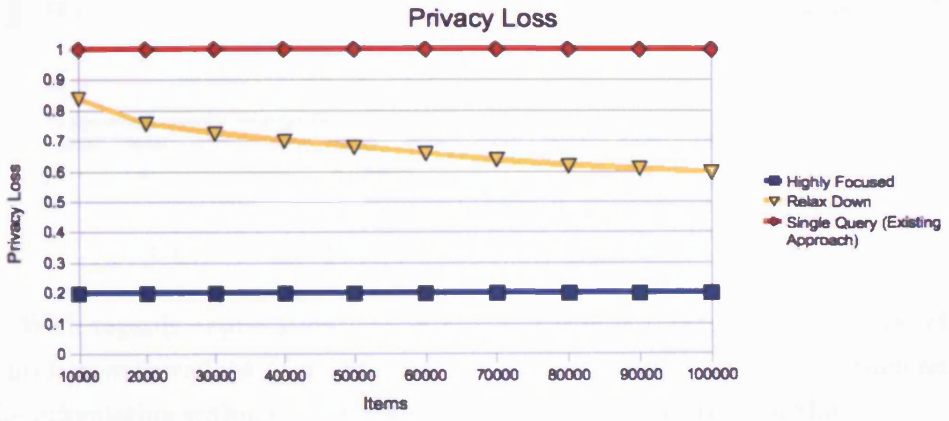

Figure 6.6: Privacy Loss vs Size of Vendor's Catalogue

With regards the privacy loss seen, the size of the vendor's catalogue has no effect on the Highly Focused Subsets or Single Query GPR algorithms. This is as expected, since the former algorithm always creates queries of a fixed size relative only to the amount of attributes in the consumer's preferences, while the latter algorithm always releases all preferences.

However, the privacy loss when using the Relax Down algorithm decreases as the size of the catalogue increases, since as the catalogue size increases the query selectivity increases, meaning the Relax Down algorithm will need to send less queries to retrieve the same amount of results: given that the algorithm works by making each subsequent query "bigger", less queries means the last query will be smaller. When the vendor's catalogue is sufficiently small the privacy loss should become total, like that of the Single Query algorithm (and therefore the existing

approach), since it will ultimately need to create a query containing all preferences; whereas when the vendor's catalogue is sufficiently large the privacy loss should become minimal, like that of the Highly Focused Subsets algorithm, since the first query containing just a single value per attribute within the consumer's preferences should retrieve the desired amount of results. The implication of this is that the larger the amount of items a particular vendor has available, the smaller the privacy loss that should occur when using this algorithm.

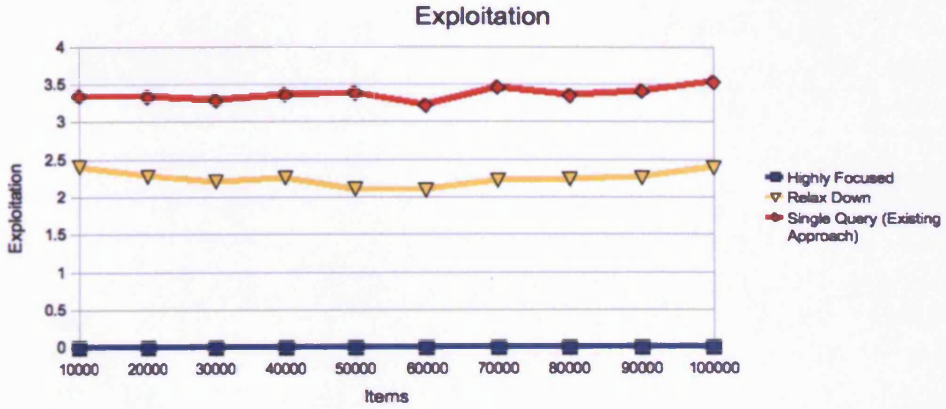

Figure 6.7: Exploitation vs Size of Vendor's Catalogue

With regards exploitation, the size of the vendor's catalogue has little effect. This is as expected since exploitations act purely upon the consumer's preferences: the information within the vendor's catalogue has no bearing on this.

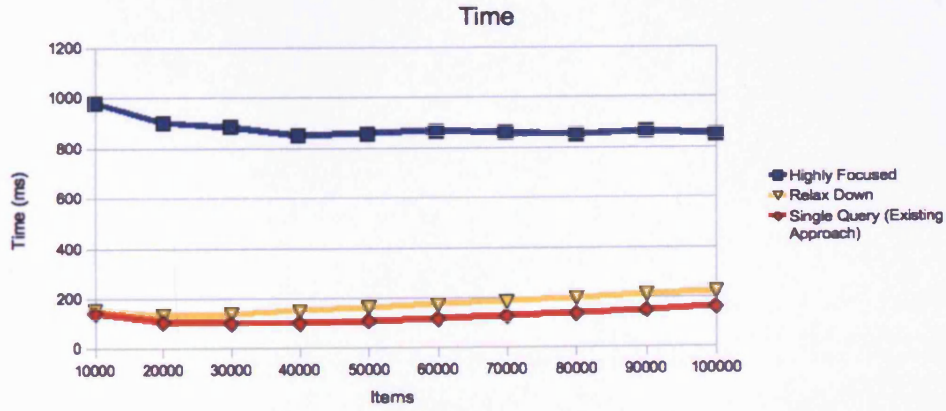

Figure 6.8: Runtime vs Size of Vendor's Catalogue

Looking at the efficiency measures, the results show that varying the amount of items in the vendor's catalogue only appreciably affects the runtime of the Highly

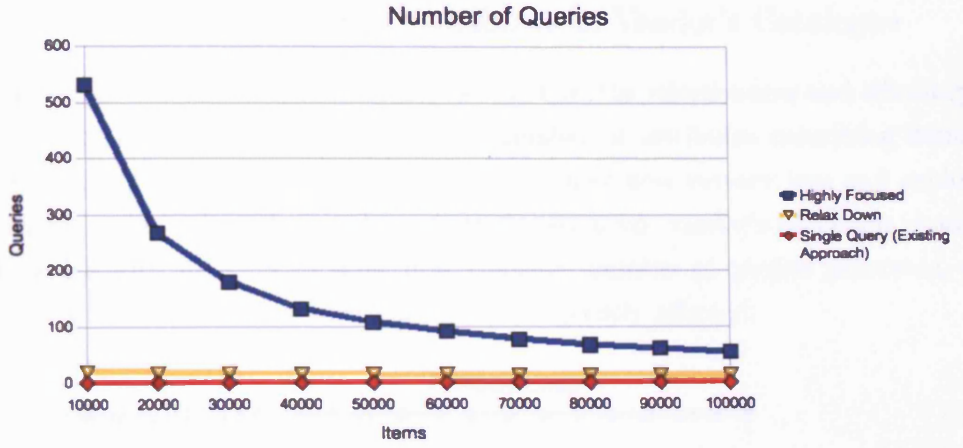

Figure 6.9: Queries vs Size of Vendor's Catalogue

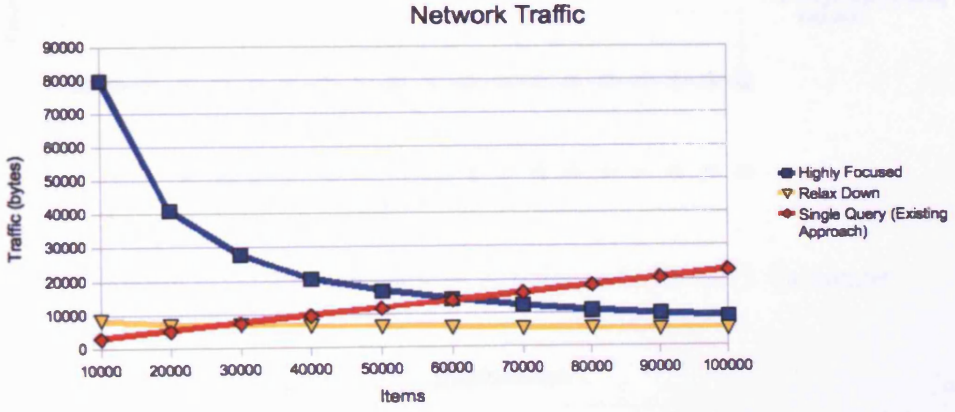

Figure 6.10: Network Traffic vs Size of Vendor's Catalogue

Focused Subsets algorithm, which shows a gradual decrease with an exponential trend. This is due to the fact that the number of queries sent by the Highly Focused Subsets GPR algorithm is seen to increase exponentially as the number of items (and therefore query selectivity) decreases, since the likelihood that a single very focused query will find matching items in the vendor's catalogue decreases as the amount of items decreases. The network traffic for this algorithm follows this trend, as each query and response generates network traffic. The implications here are that when the query selectivity is low, the Highly Focused Subsets GPR algorithm may generate an undesirably large amount of queries. The other GPR algorithms show no such issues.

6.3.3.2 Varying Number of Attributes in Vendor's Catalogue

The second of the areas examined looks at how the effectiveness and efficiency of the GPR algorithms is affected by the number of attributes describing items in the vendor's catalogue. Figures 6.11 - 6.12 show how privacy loss and exploitation were affected as the number of attributes in the vendor's catalogue changes, while Figures 6.13 - 6.15 show how runtime, number of queries generated, and communications traffic generated were consequently affected.

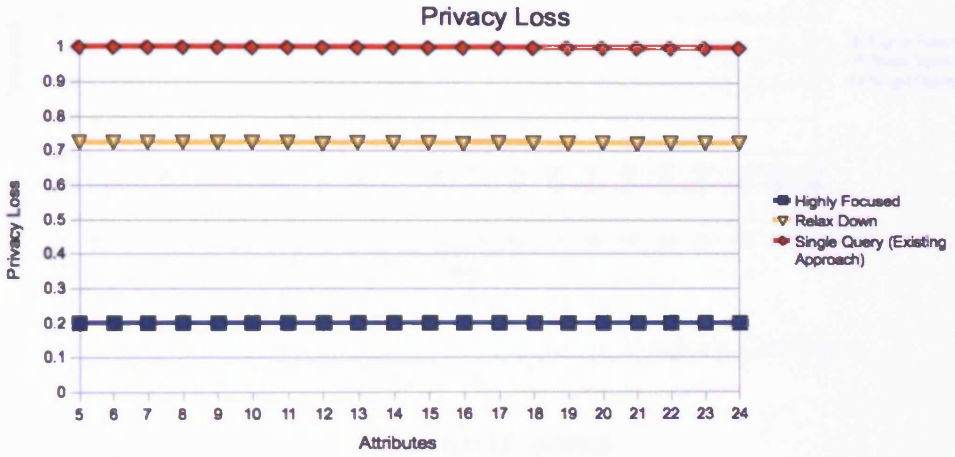

Figure 6.11: Privacy Loss vs Attributes in Vendor's Catalogue

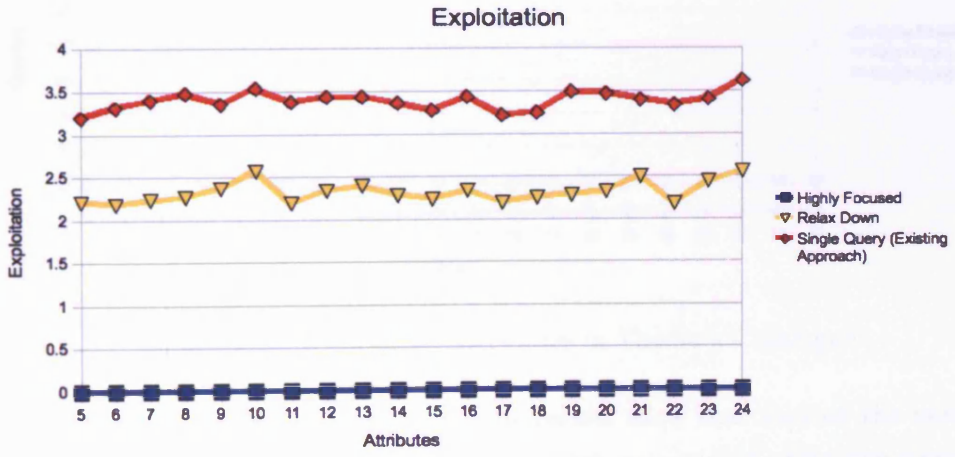

Figure 6.12: Exploitation vs Attributes in Vendor's Catalogue

Neither privacy loss nor exploitation are affected by the size of the vendor's catalogue; all values seen remain fairly constant with no trend emerging. This is

as expected as only the attributes relevant to the consumer's preferences should have any effect on any part of the preference searching process, so adding more attributes while keeping the consumer preferences static will not have any effect on the GPR process, whichever algorithm is used.

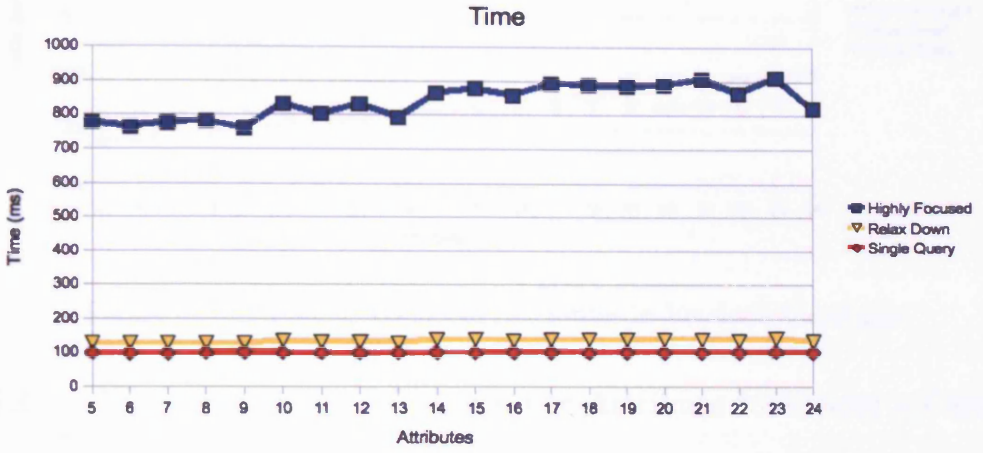

Figure 6.13: Runtime vs Attributes in Vendor's Catalogue

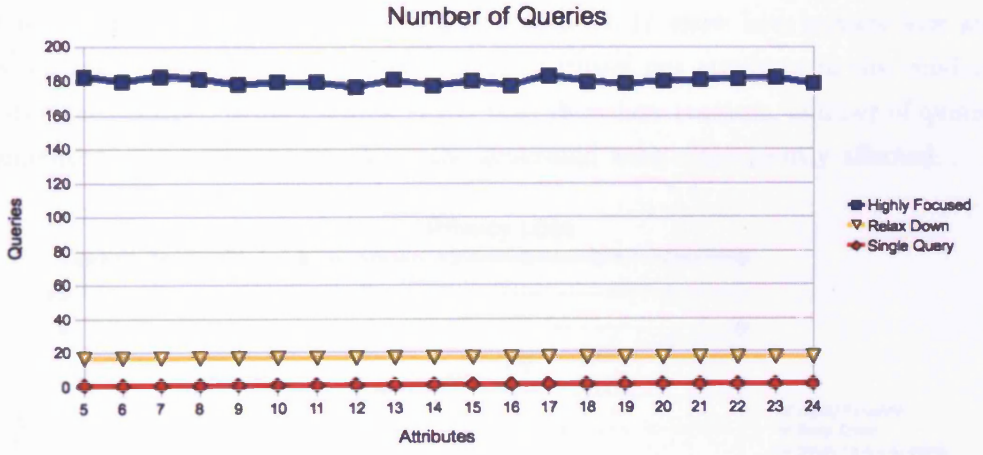

Figure 6.14: Queries vs Attributes in Vendor's Catalogue

Looking at the efficiency measures, the results show that varying the number of attributes in the vendor's catalogue does not appreciably affect the runtime, network traffic, or number of queries generated by any of the GPR algorithms. The slight gradual increase seen in network traffic of the Single Query GPR algorithm is simply due to the fact that the size of the results returned will increase as there is more information to return to the consumer.

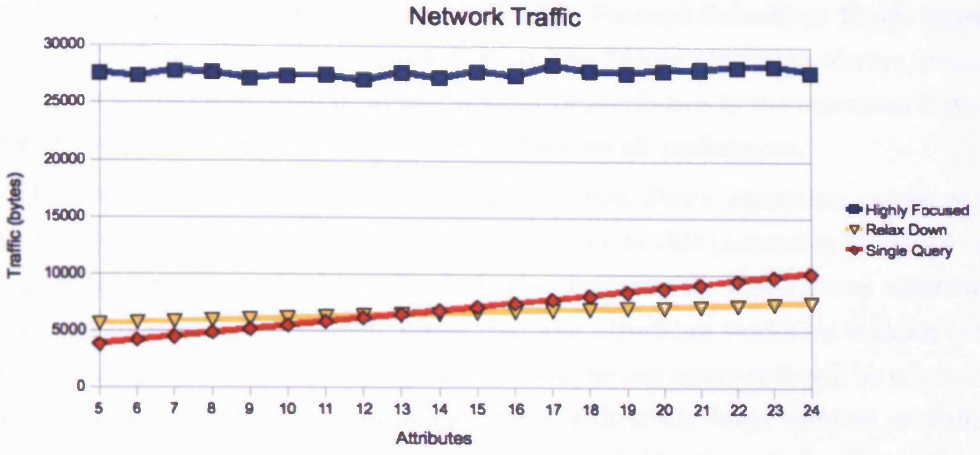

Figure 6.15: Network Traffic vs Attributes in Vendor's Catalogue

6.3.3.3 Varying Number of Values per Attribute in Vendor's Catalogue

The third of the areas examined looks at how the effectiveness and efficiency of the GPR algorithms is affected by the number of values in each attribute describing items in the vendor's catalogue. Figures 6.16 - 6.17 show how privacy loss and exploitation were affected as the number of values per attribute in the vendor's catalogue changes, while Figures 6.18 - 6.20 show how runtime, number of queries generated, and communications traffic generated were consequently affected.

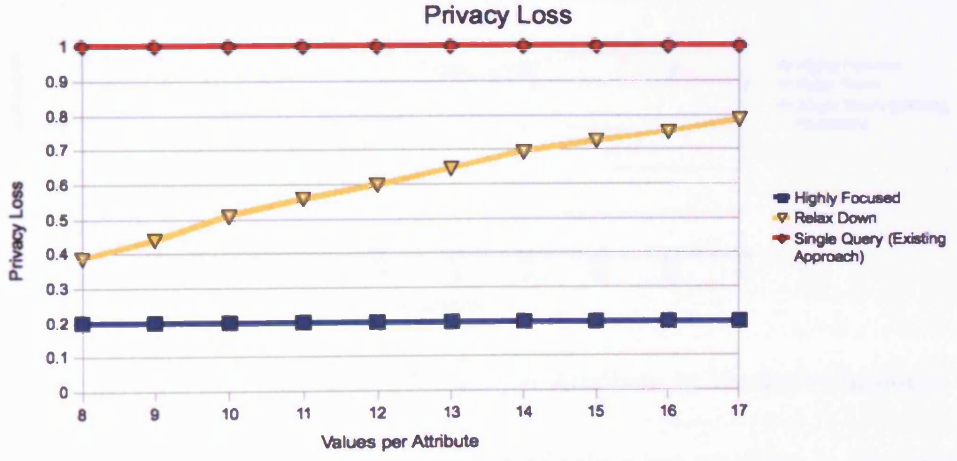

Figure 6.16: Privacy Loss vs Values per Attribute in Vendor's Catalogue

With regards the privacy loss seen, the number of values per attribute in the

vendor's catalogue has no affect for the Highly Focused Subsets or Single Query GPR algorithms. This is as expected, since the former algorithm always creates queries of a fixed size relative to the amount of attributes in the consumer's preferences, while the latter algorithm always releases all preferences.

However, the privacy loss when using the Relax Down algorithm increases as the number of values per attribute increases, since as this parameter increases the chances that any query finds results decreases, meaning the Relax Down algorithm will need to send more queries: given that the algorithm works by making each subsequent query "bigger", more queries means the last query sent will be relatively larger. When the vendor's catalogue has a sufficiently large number of values per attribute the privacy loss should become total, like that of the Single Query algorithm (and therefore the existing approach), since it will need to create a query containing all preferences; whereas when the vendor's catalogue has a sufficiently small number of values per attribute the privacy loss should become minimal, like that of the Highly Focused Subsets algorithm, since the first query containing just a single value per attribute within the consumer's preferences should retrieve the desired amount of results. The implication of this is that if using the Relax Down algorithm, the wider the range of items available from the vendor, the more the privacy loss seen is likely to be.

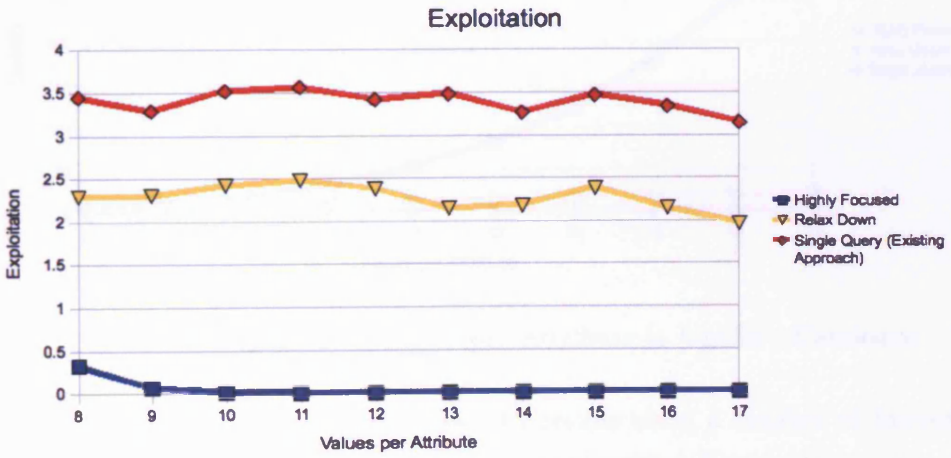

Figure 6.17: Exploitation vs Values per Attribute in Vendor's Catalogue

With regards exploitation, the number of values per attribute in the vendor's catalogue has no effect. This is as expected as only the attributes relevant to the consumer's preferences should have any effect on any part of the preference searching process, so adding more values per attributes while keeping the consumer

preferences static will not have any effect on the GPR process, whichever algorithm is used.

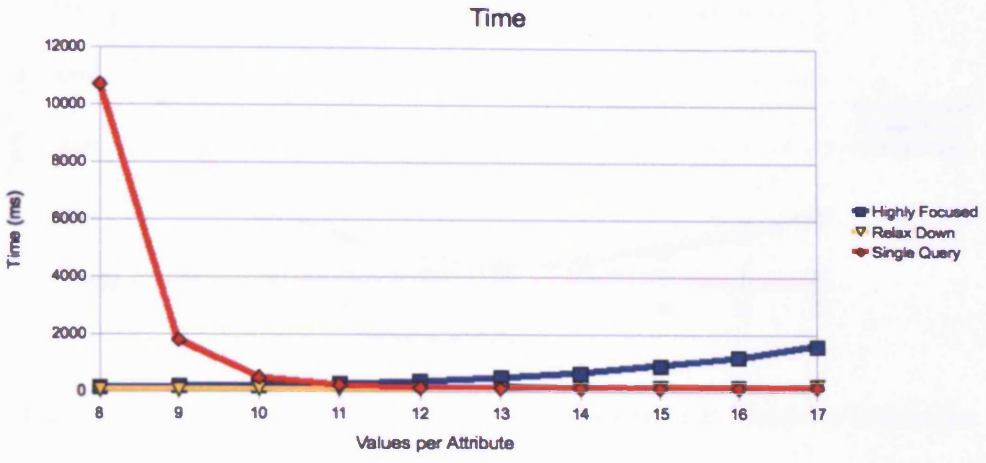

Figure 6.18: Runtime vs Values per Attribute in Vendor's Catalogue

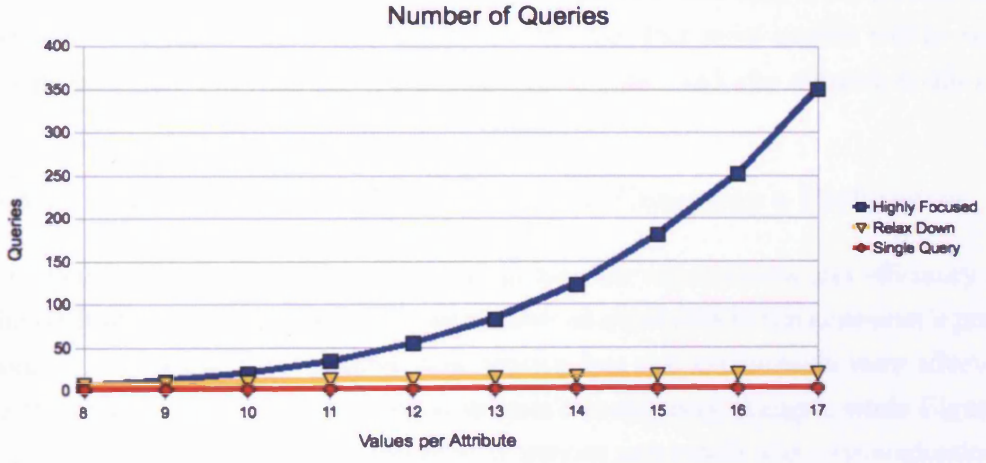

Figure 6.19: Queries vs Values per Attribute in Vendor's Catalogue

Looking at the efficiency measures, the results show a number of interesting things. The time taken for the Single Query algorithm becomes very large as it approaches the number of values per attribute in the consumer's preferences; this is because the single query sent will match more and more items in the vendor's catalogue, and therefore the time taken for the query evaluation in the database and the amount of post-processing of the results by the agent to select the most preferred n items increases. This is shown in the amount of network traffic seen for this algorithm - the larger the result set returned, the more traffic is seen. The

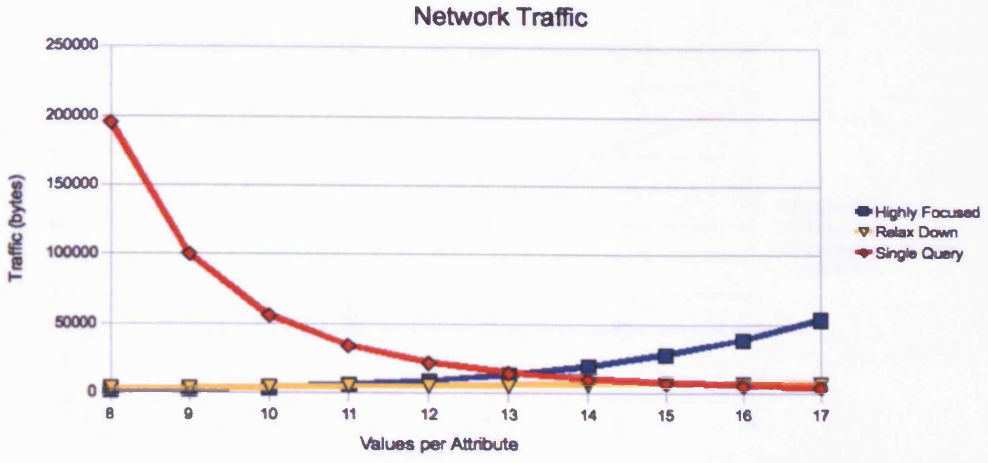

Figure 6.20: Network Traffic vs Values per Attribute in Vendor's Catalogue

time taken for the Relax Down algorithm is fairly static, on the other hand, whilst the time taken for the Highly Focused Subsets algorithm increases gradually as this particular parameter increases; this is due to the fact that each of its queries will become less likely to find matching items, and therefore more queries will be sent – a process that will take increasing amounts of time (and also network traffic).

6.3.3.4 Varying Number of Attributes in Consumer's Preferences

The fourth of the areas examined looks at how the effectiveness and efficiency of the GPR algorithms is affected by the number of attributes in the consumer's preferences. Figures 6.21 - 6.22 show how privacy loss and exploitation were affected as the number of attributes in the consumer's preferences changes, while Figures 6.23 - 6.25 show how runtime, number of queries generated, and communications traffic generated were consequently affected.

With regards the privacy loss seen, the number of attributes in the consumer's preference has no affect for the Highly Focused Subsets or Single Query GPR algorithms. This is as expected, since the former algorithm will always send preference subsets of a fixed size relative to the number of values per attribute – not the number of attributes – while the latter algorithm always releases all preferences. Thus, the number of attributes in the consumer's preferences has no bearing on this.

However, the privacy loss when using the Relax Down algorithm increases as the number of attributes increases, since as this increases the chances that any

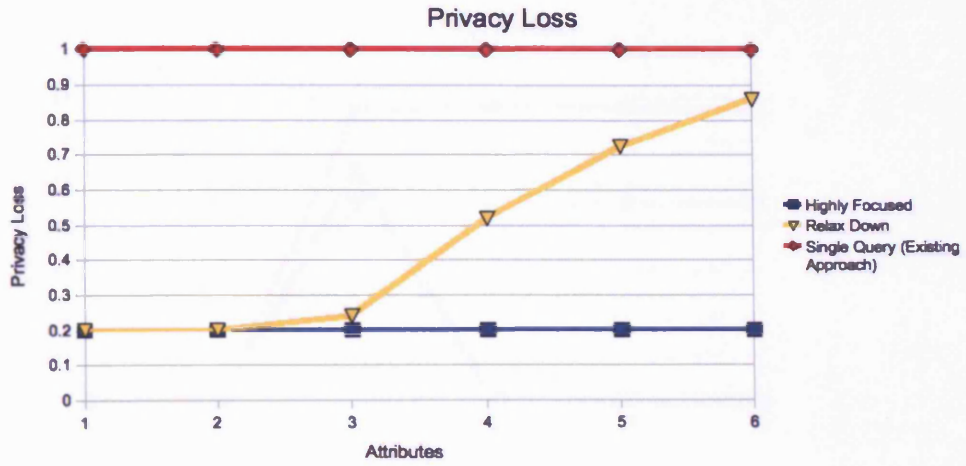

Figure 6.21: Privacy Loss vs Attributes in Consumer's Preferences

query finds results decreases, meaning the Relax Down algorithm will need to send more queries: given that the algorithm works by making each subsequent query "bigger" (adding more and more of the preferred values), more queries means the last query will be relatively larger. When the vendor's catalogue has a sufficiently large number of values per attribute the privacy loss should become total, like that of the Single Query algorithm (and therefore the existing approach), since it will need to create a query containing all preferences; whereas when the vendor's catalogue has a sufficiently small number of values per attribute the privacy loss should become minimal, like that of the Highly Focused Subsets algorithm, since the first query containing just a single value per attribute within the consumer's preferences should retrieve the desired amount of results. The implication of this is that if the consumer is using the Relax Down algorithm, the more attributes they express preferences over, the more potential for privacy loss there will be.

With regards exploitation, the amount seen for the Single Query and Relax Down algorithms increases as the amount of attributes in the consumer's preferences increases; due to the fact that there is more information for the vendor to exploit. The trend seen for the Highly Focused Subsets algorithm is very interesting. There is no exploitation seen for small amounts of attributes, since there is very little preference information sent and therefore very little information to exploit; conversely, there is no exploitation seen for large amounts of attributes, since the queries created are so targeted that they are also hard to exploit. However, in between these values lies a very specific area where exploitation is possible and is seen; this level is just where the first query sent becomes sufficiently targeted that

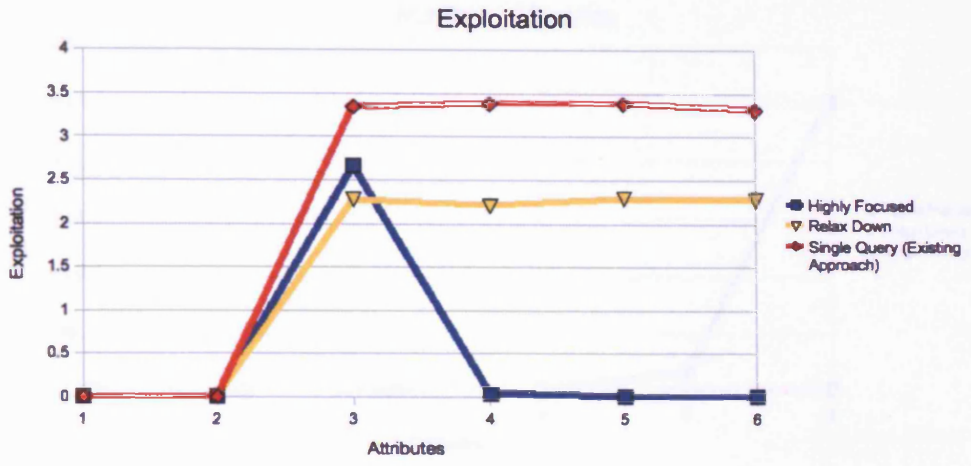

Figure 6.22: Exploitation vs Attributes in Consumer's Preferences

it stops retrieving more than the n results requested and multiple queries have to be sent.

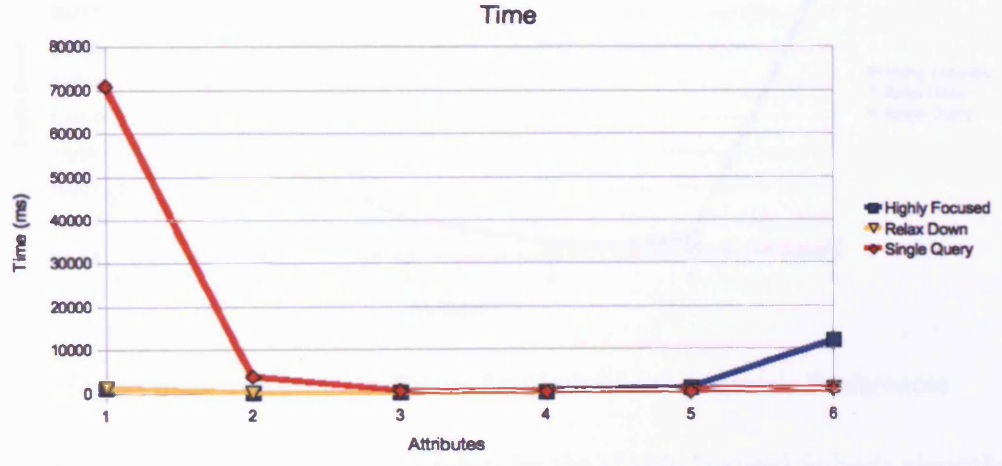

Figure 6.23: Runtime vs Attributes in Consumer's Preferences

Looking at the efficiency measures, it can be seen that the Single Query algorithm takes a large time to run when a consumer expresses preferences over only a small amount of attributes; this is because the single query sent will match many items in the vendor's catalogue, and therefore the time taken for the query evaluation in the database and the amount of post-processing of the results by the agent to select the most preferred n items increases. This is shown in the amount of network traffic seen for this algorithm - the larger the result set returned, the more traffic is seen. The time taken for the Relax Down algorithm is fairly static,

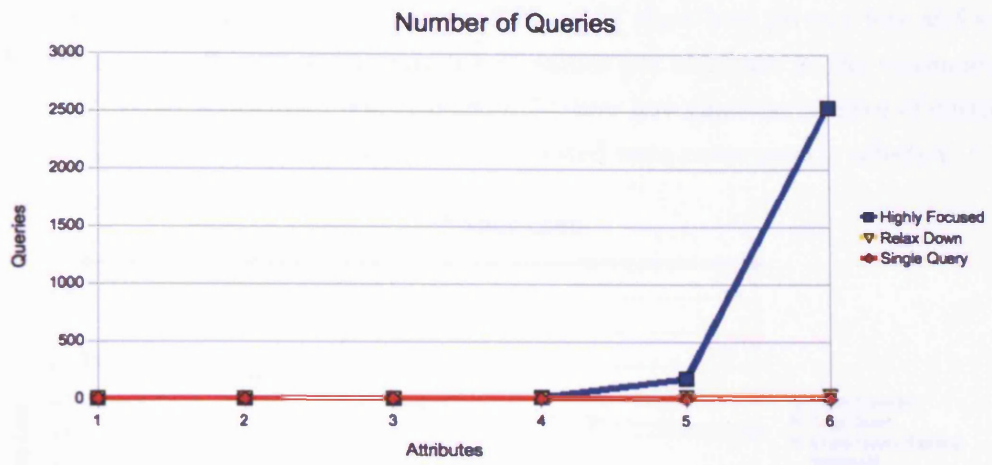

Figure 6.24: Queries vs Attributes in Consumer's Preferences

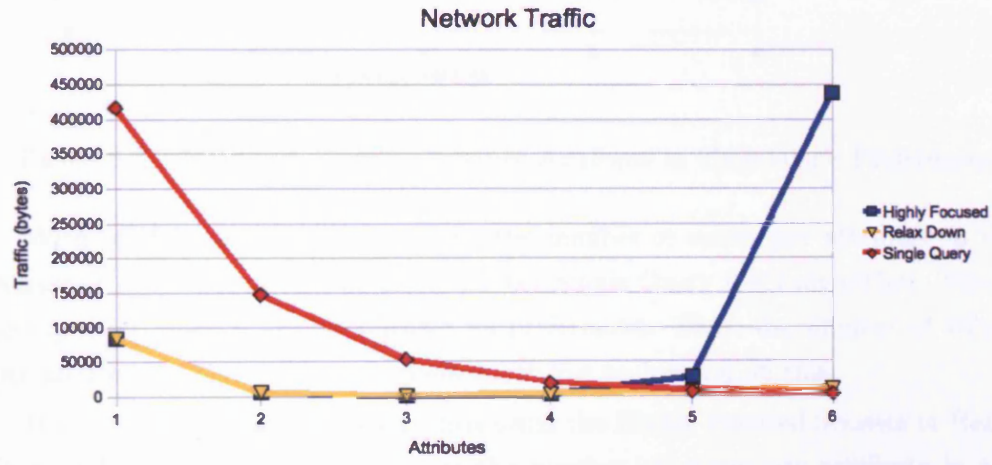

Figure 6.25: Network Traffic vs Attributes in Consumer's Preferences

on the other hand, whilst the time taken for the Highly Focused Subsets algorithm increases gradually as this particular parameter increases; this is due to the fact that each of its queries will become less likely to find matching items, and therefore more queries will be sent – a process that will take increasing amounts of time (and also network traffic).

6.3.3.5 Varying Number of Values per Attribute in Consumer's Preferences

The fifth and final of the areas examined looks at how the effectiveness and efficiency of the GPR algorithms is affected by the number of values in each attribute

of the consumer's preferences. Figures 6.26 - 6.27 show how privacy loss and exploitation were affected as the number of values per attribute in the consumer's preferences changes, while Figures 6.28 - 6.30 show how runtime, number of queries generated, and communications traffic generated were consequently affected.

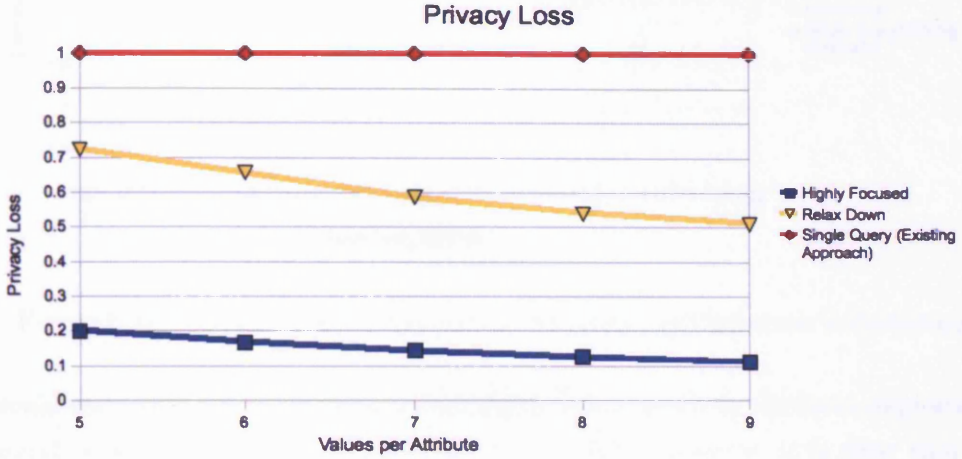

Figure 6.26: Privacy Loss vs Values per Attribute in Consumer's Preferences

With regards the privacy loss seen, the number of values per attribute in the consumer's preferences has no affect for the Single Query GPR algorithm. This is as expected, since it always releases all preferences. Thus, the number of values per attribute in the consumer's preferences has no bearing on this.

However, the privacy loss seen when using the Highly Focused Subsets or Relax Down GPR algorithms decreases as the number of values per attribute in the consumer's preferences increases, since for both algorithms a larger amount of these values means that a query containing a certain amount of preferences represents a smaller proportion of the overall amount of preferences. The implication of this is that if the consumer expresses preferences over larger amounts of values for a given number of preferences, the relative privacy loss seen is likely to be less. This does not mean that less preference information will be released – just that the amount of preference information released as compared to the amount specified will decrease.

With regards exploitation, the number of values per attribute in the consumer's preferences has no real effect. This, for the Highly Focused Subsets and Relax Down algorithms, is as expected – since these algorithms split up a consumer's preferences and release the values gradually, minimising the chances of exploitation. For the Single Query algorithm, however, this result is counter intuitive – it

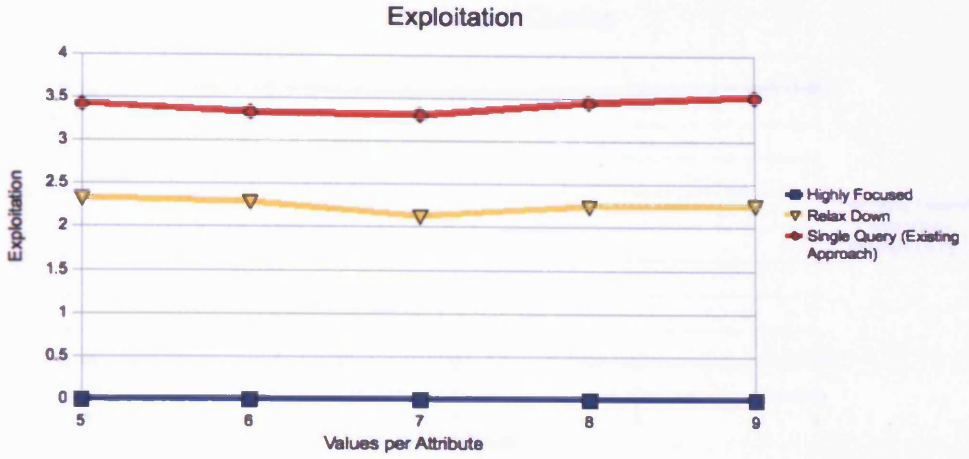

Figure 6.27: Exploitation vs Values per Attribute in Consumer's Preferences

would seem that the more information there was to exploit, the more exploitation would occur. Looking in more detail at the results, however, it is clear that this trend is seen because the single query released finds a minimum of many hundreds of results, and given the type of exploitation configured at the vendor, this exploitation is unlikely to remove items that happen to be in the list of top 10 most preferred. Given a set of preferences and a vendor's catalogue where the single query released find a number of results close to the amount requested by the consumer, however, a trend of exploitation increasing in line with increasing values per attribute would be expected.

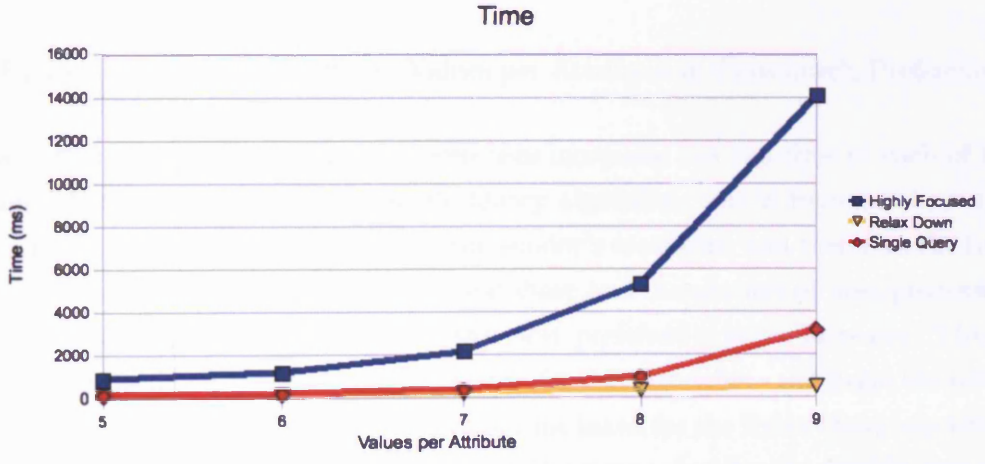

Figure 6.28: Runtime vs Values per Attribute in Consumer's Preferences

Looking at the efficiency measures, the results show that as the amount of values

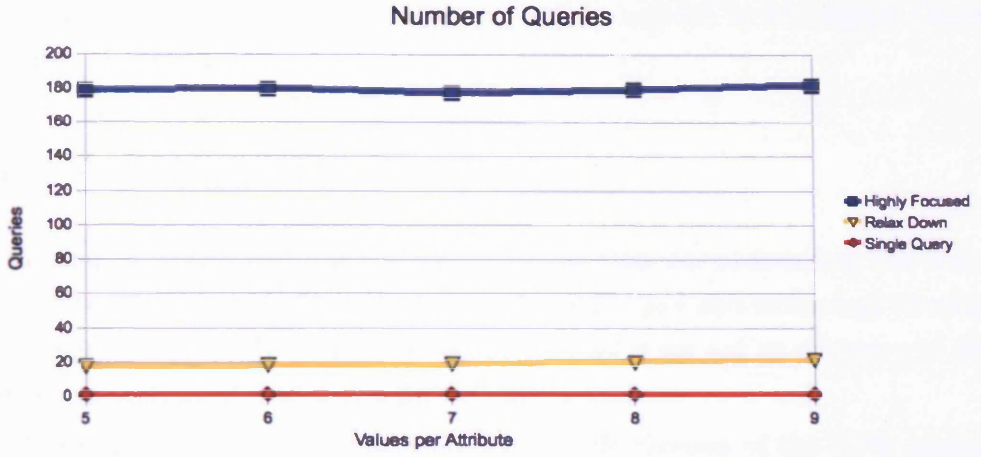

Figure 6.29: Queries vs Values per Attribute in Consumer's Preferences

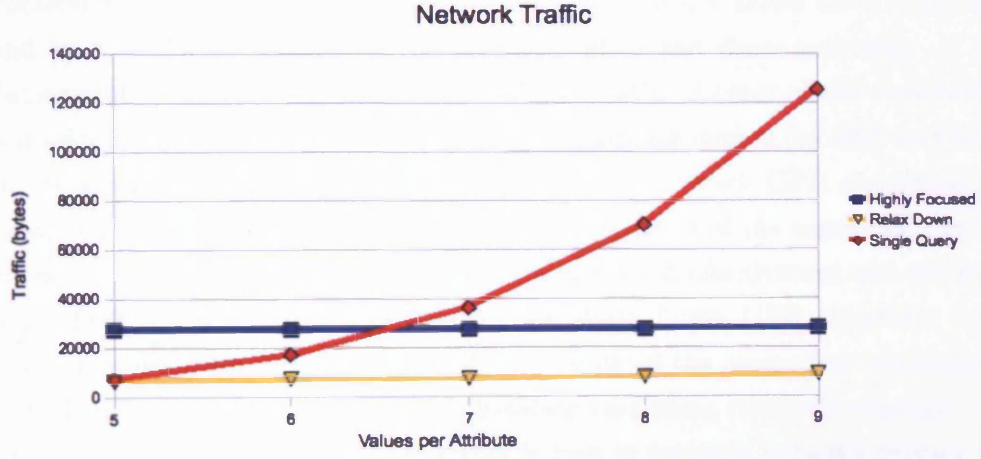

Figure 6.30: Network Traffic vs Values per Attribute in Consumer's Preferenecs

per attribute in the consumer's preferences increases, the run time of each of the algorithms increases. For the Single Query algorithm; this is because the single query sent will match many items in the vendor's catalogue, and therefore the time taken for the query evaluation in the database and the amount of post-processing of the results by the agent to select the most preferred n items increases. This is shown in the amount of network traffic seen for this algorithm - the larger the result set returned, the more traffic is seen. The time taken for the Relax Down algorithm is fairly static, on the other hand, whilst the time taken for the Highly Focused Subsets algorithm increases gradually as this particular parameter increases; this is due to the fact that the implementation of GPR pre-creates all subsets before it starts to gradually release them – and the more values per attribute in the

consumer's preferences, the more subsets will be created by the Highly Focused Subsets algorithm.

6.4 SUMMARY

This chapter has detailed the evaluation process that was undertaken; this was designed to allow an evaluation of whether the techniques and technology developed in this work (the GPR process) achieved the goal it set out to achieve – to allow preference-enhanced search in a privacy aware manner.

In the first set of experiments, the overall effectiveness of the GPR approach was examined; the experiments showed that, as compared to the existing approach, the GPR approach reduced privacy loss in almost all cases (99.9% for the Highly Focused Subsets GPR algorithm and 99.8% for the Relax Down GPR algorithm) and by a significant degree (by roughly four fifths and three quarters). It also demonstrably reduced exploitation (in 90% and 30% of cases where exploitation was seen in the existing approach), and to a significant degree (by 80% and 33%). In the second set of experiments, the performance of each GPR algorithm was examined in more detail by varying different properties of the consumer's preferences and vendor's catalogue and analysing how both effectiveness and efficiency were affected. Of the three algorithms, the Relax Down GPR algorithm is the most preferred method to use, since it best balanced the competing requirements of both reducing privacy loss and exploitation and being relatively efficient. The Highly Focussed Subsets GPR algorithm is best in terms of reducing privacy loss and exploitation where efficiency is not a factor.

Thus, the evaluation demonstrated that the GPR approach to preference searching can reduce both privacy loss and exploitation with no penalty in terms of achieving the goal of retrieving the most preferred results with respect to the consumer's preferences, and has thus demonstrated that the proposed alternative approach to preference searching in e-commerce has merit.

Chapter 7

Conclusions

In this final chapter, we summarise the research carried out in this thesis, highlighting the major contributions from this work and areas where further work is deemed necessary.

7.1 Privacy in the Information Age

Privacy is a concept which has intrigued society and inspired heated debate throughout the whole course of civilisation. Methods used for the protection of privacy have changed through the ages as new thinking and new technologies have altered the perception of privacy and the ways in which it can be lost. The age in which we currently find ourselves is often called the "information age", and this is an age characterised by extreme amounts of information being available about people – much of it likely to be considered as "private" by the individuals concerned – at the click of a mouse button. Information itself has become valuable, and information about people even more so. Unfortunately, the information age is an age where an individual can unwittingly lose control of this valuable information about themselves, and the legal methods typically employed to protect privacy in recent decades are struggling to keep up. Given the "greased" nature of such information, once control of it has been lost by the individual it is pretty much impossible to get back. Thus, technologies that strive to place the control of the release of this information in the hands of the consumer, including those for supporting e-commerce applications and activities, are needed to protect privacy in the information age.

7.2 PREFERENCE SEARCHING

The concept of using a consumer's preferences to help enhance their e-commerce experience through personalisation is one that has attracted increasing attention from the research community in recent years. One specific area of work has been in the use of a consumer's preferences to help enhance search techniques.

"Standard" search techniques simply allow a consumer to express keywords that they wish to search for; this often results in either empty result sets (if the keywords expressed defined too narrow a query for the stock catalogue being searched resulting in no items being found) or information overload (if the keywords expressed defined too wide a query resulting in an unwieldily large amount of items being found). *Preference Searching* has been proposed as a way of solving this problem. The basic idea is that instead of issuing queries containing hard constraints, a query is constructed to contain soft preferences. The model used in all current implementations of this idea is for a consumer to release all of their preference information to the vendor to enable this.

However, the release of this preference information to a vendor means a loss of privacy and the potential for exploitation. A loss of privacy is obvious in this case and is caused by the fact that the consumer is required to release private personal information to a vendor. The release of this information also presents chances for exploitation – the possibility that the vendor can take this preference information and inject their own preferences, distorting the result set to their advantage – and the consumer's disadvantage – in a manner transparent to the consumer. All current approaches to preference searching are based on this model, so all suffer this issue.

This research proposed an alternative, consumer-centric, approach to the problem of preference searching. It shifts the computation involved in using preferences to enhance search from the vendor side to the consumer side. The consumer has an agent working on their behalf that enables preference searching through the release of portions of the consumer's preferences in a carefully controlled manner that aims to retrieve preferred results. This approach allows a set of most preferred items to be calculated whilst also minimising firstly privacy loss by only releasing those parts of the consumer's preferences necessary to find their most preferred items; and secondly the chances of exploitation by minimising the amount of preference information available for the vendor to exploit.

7.3 QUANTIFYING PRIVACY LOSS AND EXPLOITATION

To support the proposed alternative approach to preference searching, measures are required that allow the quantification of private information contained within preferences and to assess the loss of it during the release process. To the best of our knowledge, no such measures currently exist, specifically quantifying preference privacy. In this work we attempted to take a step forward in this area by proposing two measures.

One method of viewing privacy loss of a consumer's preferences is to think of it as a function of how much preference information is released to the vendor relative to how much exists. Our measure of privacy loss is based on this observation, and produces a numerical indication of how much privacy has been lost that takes into account how much preference information has been released to a vendor. The proposed measure is relatively simple, but is adequate for the purposes of proving our concept. Constructing more generally applicable measures of privacy loss in this context can be an interesting area of future work.

Exploitation, on the other hand, is a conceptually simpler problem. The amount of exploitation of a set of results is essentially a function of how much that set of results has been distorted by the vendor for their own purposes. Thus, constructing a measure of exploitation is a matter of measuring the difference between the result set that *should have* been sent to the consumer, and the result set that *was* sent to the consumer. We propose a measure of exploitation that produces a numerical indication of how much exploitation has occurred that is based on to the cardinality of the difference between the two sets. Again, this is a relatively simple way of measuring exploitation, but sufficient for the purposes of verifying the hypothesis of this work. Measures considering more complex exploitation models (e.g. relaxing the unlinkability assumption, or assuming that the vendor is able to use data mining to establish consumer preferences) are worth investigating in the future.

7.4 GRADUAL PARTIAL RELEASE

A major contribution of this work is a proof of concept implementation of the proposed consumer-centric model, called Gradual Partial Release. This implementation aimed to achieve the goals of the new model by taking a consumer's preferences, splitting this up into subsets of these preferences, and releasing the subsets until the amount of results specified by the consumer are returned.

Implementing GPR involved several major areas of work. Firstly, a basic preference framework was required, in order that preferences could be expressed into the system. A preference framework, built on some existing work, was introduced. This allowed preferred values, or a range of values, to be expressed across multiple attributes. This framework has a limited expressive power in terms of the variety of preferences that it can represent, but it allowed us to verify our hypothesis. Integrating the GPR approach with a more complex and expressive preference framework is viewed as the next stage of work for this approach.

Once preferences have been expressed, they need splitting into subsets of preferences. This thesis presented three such algorithms. One essentially duplicated the manner in which the existing approach works by releasing a single subset containing all preferences; this was to allow comparison between the existing approach and the new approach. A second algorithm aimed to retrieve preferred results in a privacy-preserving and exploitation-reducing manner by sending subsets that were as highly focussed as possible. However, while each subset released contains only a small amount of information, many subsets may need to be released in a highly inefficient manner. A third algorithm took this idea and attempted to make it more efficient by starting out with a subset that was as highly focused as possible; preferences were then added one at a time to subsequent subsets in order to balance privacy loss and exploitation while finding preferred results in an efficient manner. More advanced algorithms are possible further work, and could include ideas such as first using query probing techniques to estimate a vendor's catalogue in order to target subsets such that they are more likely to retrieve preferred results.

7.5 CONCLUDING REMARKS

Privacy is an important human value that should be protected wherever possible. An area of active research where this is not currently being achieved is that of using a consumer's private preference information to enhance the search process in e-commerce. In this work we have we proposed an alternate approach to preference searching, called Gradual Partial Release, along with methods that implement this approach. Our experiments confirmed that the approach is effective – in terms of finding a set of preferred results while reducing privacy loss and exploitation. We therefore believe that GPR represents a promising way forward in reducing privacy loss and exploitation while achieving the goal of preference-enhanced searching in e-commerce.

Appendix A

List of Publications

- Smith, R and Shao, J. Preserving privacy when preference searching in e-commerce. In *Proceedings of the 2003 ACM workshop on Privacy in the electronic society (WPES)*, ACM Press, pp. 101–110.
- Smith, R and Shao, J. Privacy and e-commerce: a consumer-centric perspective. *Electronic Commerce Research 7*, 2 (2007), 89–116.

REFERENCES

[1] ACKERMAN, M. S., AND CRANOR, L. Privacy critics: UI components to safeguard users' privacy. In *Extended abstracts on Human factors in computing systems (CHI)* (1999), ACM Press, pp. 258–259.

[2] ACKERMAN, M. S., CRANOR, L. F., AND REAGLE, J. Privacy in e-commerce: examining user scenarios and privacy preferences. In *Proceedings of the 1st ACM conference on Electronic commerce* (1999), ACM Press, pp. 1–8.

[3] ACQUISTI, A. Privacy in electronic commerce and the economics of immediate gratification. In *Proceedings of the 5th ACM conference on Electronic commerce (EC)* (2004), ACM Press, pp. 21–29.

[4] ACQUISTI, A., DINGLEDINE, R., AND SYVERSON, P. On the economics of anonymity. In *Financial Cryptography: 7th International Workshop, FC 2003* (2003), R. N. Wright, Ed., vol. 2742 of *Lecture Notes in Computer Science*, Springer-Verlag, pp. 84–102.

[5] ADAM, N. R., AND WORTHMANN, J. C. Security-control methods for statistical databases: a comparative study. *ACM Computing Surveys (CSUR) 21*, 4 (1989), 515–556.

[6] AGRAWAL, D., AND AGGARWAL, C. C. On the design and quantification of privacy preserving data mining algorithms. In *Proceedings of the Twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of Database Systems (PODS)* (2001), ACM Press, pp. 247–255.

[7] AGRAWAL, R., AND SRIKANT, R. Privacy-preserving data mining. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data* (2000), ACM Press, pp. 439–450.

[8] AGRAWAL, R., AND WIMMERS, E. L. A framework for expressing and combining preferences. *ACM SIGMOD Record 29*, 2 (2000), 297–306.

[9] ANTÓN, A. I., AND EARP, J. B. A requirements taxonomy for reducing web site privacy vulnerabilities. *Requirements Engineering 9*, 3 (2004), 169–185.

[10] ARENDT, H. *The Human Condition.* University of Chicago Press, 1958.

[11] ARISTOTLE. Politics, 350 B.C. http://classics.mit.edu/Aristotle/politics.html.

[12] ASHLEY, P., HADA, S., KARJOTH, G., AND SCHUNTER, M. E-P3P privacy policies and privacy authorization. In *Proceedings of the ACM workshop on Privacy in the Electronic Society* (2002), ACM Press, pp. 103–109.

[13] ASHLEY, P., POWERS, C., AND SCHUNTER, M. From privacy promises to privacy management: a new approach for enforcing privacy throughout an enterprise. In *Proceedings of the 2002 workshop on New security paradigms* (2002), ACM Press, pp. 43–50.

[14] BACK, A., GOLDBERG, I., AND SHOSTACK, A. Freedom systems 2.1 security issues and analysis. White paper, Zero Knowledge Systems, Inc., May 2001.

[15] BARTOLINI, I., CIACCIA, P., AND PATELLA, M. Efficient sort-based skyline evaluation. *ACM Transactions on Database Systems 33*, 4 (2008), 1–49.

[16] BECK, L. L. A security mechanism for statistical database. *ACM Transactions on Database Systems (TODS) 5*, 3 (1980), 316–3338.

[17] BECKWITH, R. Designing for ubiquity: The perception of privacy. *IEEE Pervasive 2*, 2 (2002), 40–46.

[18] BENN, S. I., AND GAUS, G. F. *Public and Private in Social Life.* St. Martins Press, 1983.

[19] BLOUSTEIN, E. J. Privacy as an aspect of human dignity: an answer to Dean Prosser. *New York University Law Review 39* (1964), 962–1007.

[20] BORZSONYI, S., KOSSMANN, D., AND STOCKER, K. The skyline operator. In *Proceedings of the 17th International Conference on Data Engineering* (2001), IEEE Computer Society, p. 421.

[21] BOUCHER, P., SHOSTACK, A., AND GOLDBERG, I. Freedom systems 2.0 architecture. White paper, Zero Knowledge Systems, Inc., December 2000.

[22] BRIN, D. *The Transparent Society.* Perseus Books, 1998.

[23] BRUSILOVSKY, P., KOBSA, A., AND NEJDL, W., Eds. *The Adaptive Web, Methods and Strategies of Web Personalization* (2007), vol. 4321 of *Lecture Notes in Computer Science*, Springer-Verlag.

[24] CAMENISCH, J., AND HERREWEGHEN, E. V. Design and implementation of the idemix anonymous credential system. In *Proceedings of the 9th ACM conference on computer and communications security (CCS)* (2002), ACM Press, pp. 21–30.

[25] CAMENISCH, J., AND LYSYANSKAYA, A. An efficient system for non-transferable anonymous credentials with optional anonymity revocation. In *Proceedings of the International Conference on the Theory and Application of Cryptographic Techniques (EUROCRYPT)* (2001), vol. 2045 of *Lecture Notes in Computer Science*, Springer-Verlag, pp. 93–118.

[26] CANTOR, S., KEMP, J., PHILPOTT, R., AND MALER, E. Assertions and protocols for the oasis security assertion markup language (SAML) v2.0. Technical report, OASIS, 2005. http://saml.xml.org/saml-specifications.

[27] CATLETT, J. On changing continuous attributes into ordered discrete attributes. In *Proceedings of European Working Session on Learning (EWSL)*, vol. 482 of *Lecture Notes in Computer Science.* Springer-Verlag, 1991, pp. 164–178.

[28] CHADWICK, D. W., ZHAO, G., OTENKO, S., LABORDE, R., SU, L., AND NGUYEN, T. A. PERMIS: a modular authorization infrastructure. *Concurrency and Computation: Practice and Experience 20*, 11 (August 2008), 1341–1357.

[29] CHAFFEY, D., MAYER, R., ELLIS-CHADWICK, F., AND JOHNSTON, K. *Internet Marketing: Strategy, Implementation and Practice.* Pearson Education, 2006.

[30] CHAUM, D. Security without identification: transaction systems to make big brother obsolete. *Communications of the ACM 28*, 10 (1985), 1030–1044.

[31] CHAUM, D., AND EVERTSE, J.-H. A secure and privacy-protecting protocol for transmitting personal information between organizations. In *Proceedings on Advances in cryptology—CRYPTO '86* (1987), Springer-Verlag, pp. 118–167.

[32] CHAUM, D. L. Untraceable electronic mail, return addresses, and digital pseudonyms. *Communications of the ACM 24*, 2 (1981), 84–90.

[33] CHEN, L. Access with pseudonyms. In *Proceedings of the International Conference on Cryptography: Policy and Algorithms* (1995), vol. 1029 of *Lecture Notes in Computer Science*, Springer-Verlag, pp. 232–243.

[34] CHIN, F. Y., AND OZSOYOGLU, G. Auditing and inference control in statistical databases. *IEEE Transactions on Software Engineering 8*, 6 (1982), 574–582.

[35] CHO, S., AND BALKE, W.-T. Order-preserving optimization of twig queries with structural preferences. In *Proceedings of the 2008 international symposium on Database engineering & applications (IDEAS)* (2008), ACM Press, pp. 219–229.

[36] CHOMICKI, J. Querying with intrinsic preferences. In *Proceedings of the 8th International Conference on Extending Database Technology* (2002), Springer-Verlag, pp. 34–51.

[37] CHOMICKI, J. Preference formulas in relational queries. *ACM Transactions on Database Systems 28*, 4 (2003), 427–466.

[38] CHOMICKI, J. Semantic optimization techniques for preference queries. *Information Systems 32*, 5 (2007), 670–684.

[39] CHOMICKI, J., GODFREY, P., GRYZ, J., AND LIANG, D. Skyline with presorting, 2003. Poster, In IEEE International Conference on Data Engineering (ICDE).

[40] CIRIANI, V., DI VIMERCATI, S. D. C., FORESTI, S., AND SAMARATI, P. k-anonymity. In *Secure Data Management in Decentralized Systems*, vol. 33 of *Advances in Information Security*. Springer-Verlag, 2007, pp. 323–353.

[41] CLAYTON, R., DANEZIS, G., AND KUHN, M. G. Real world patterns of failure in anonymity systems. In *Proceedings of the 4th International Workshop on Information Hiding (IHW)* (January 2001), vol. 2137 of *Lecture Notes in Computer Science*, Springer-Verlag, pp. 230–244.

[42] COMMISSION OF THE EUROPEAN COMMUNITIES. First report on the implementation of the data protection directive (95/46/ec). Technical report

com(2003) 265 final, European Commission, Brussels, Belgium, December 2003.

[43] COUNCIL OF EUROPE. The european convention on human rights, 1950. http://conventions.coe.int/treaty/en/Treaties/Html/005.htm.

[44] COUNCIL OF EUROPE. Convention for the protection of individuals with regard to automatic processing of personal dataconvention for the protection of individuals with regard to automatic processing of personal data, 1981. http://conventions.coe.int/treaty/en/Treaties/Html/108.htm.

[45] CRANOR, L., LANGHEINRICH, M., AND MARCHIORI, M. A P3P preference exchange language 1.0 (APPEL 1.0). Working draft, W3C, April 2002. http://www.w3.org/TR/P3P-preferences/.

[46] CRANOR, L., LANGHEINRICH, M., MARCHIORI, M., PRESLER-MARSHALL, M., AND REAGLE, J. The platform for privacy preferences 1.0 (P3P1.0) specification, 2000. http://www.w3.org/TR/P3P/.

[47] CRANOR, L. F. The role of privacy advocates and data protection authorities in the design and deployment of the platform for privacy preferences. In *Proceedings of the 12th annual conference on Computers, freedom and privacy* (2002), ACM Press, pp. 1–8.

[48] CYBER DIALOGUE. Cyber dialogue survey reveals lost revenue for retailers due to widespread consumer privacy concerns, 2001. http://www.cyberdialogue.com/.

[49] DAMGÅRD, I. B. Payment systems and credential mechanisms with provable security against abuse by individuals. In *Proceedings on Advances in cryptology (CRYPTO 88)* (1990), vol. 403 of *Lecture Notes in Computer Science*, Springer-Verlag New York, Inc., pp. 328–335.

[50] DAVID W CHADWICK, A. O. The PERMIS X.509 role based privilege management infrastructure. In *Proceedings of the 7th ACM Symposium On Access Control Models And Technologies (SACMAT)* (June 2002), pp. 135–140.

[51] DENNING, D. E., AND DENNING, P. J. Data security. *ACM Computing Surveys (CSUR) 11*, 3 (1979), 227–249.

[52] DINGLEDINE, R., MATHEWSON, N., AND SYVERSON, P. Tor: The second-generation onion router. In *Proceedings of the 13th USENIX Security Symposium* (August 2004), The USENIX association, pp. 303–320.

[53] EGELMAN, S., CRANOR, L. F., AND CHOWDHURY, A. An analysis of P3P-enabled web sites among top-20 search results. In *Proceedings of the 8th international conference on Electronic commerce (ICEC)* (2006), ACM Press, pp. 197–207.

[54] ELOVICI, Y., SHAPIRA, B., AND MASCHIACH, A. A new privacy model for hiding group interests while accessing the web. In *Proceeding of the ACM workshop on Privacy in the Electronic Society* (2002), ACM Press, pp. 63–70.

[55] ETZIONI, A. *The limits of privacy.* Basic Books, 1999.

[56] EUROPEAN COMMISSION ARTICLE 29 WORKING PARTY. Opinion 4/2004 on the processing of personal data by means of video surveillance. Technical report 11750/02/en wp 89, European Commission, Brussels, Belgium, 2004.

[57] EUROPEAN COMMISSION ARTICLE 29 WORKING PARTY. Opinion on more harmonised information provisions. Technical report 11987/04/en wp 100, European Commission, Brussels, Belgium, November 2004.

[58] EUROPEAN COMMISSION ARTICLE 29 WORKING PARTY. Working document on biometrics. Technical report 12168/02/en wp80, European Commission, Brussels, Belgium, 2004.

[59] EUROPEAN OPINION RESEARCH GROUP EEIG. Special eurobarometer data protection executive summary. Technical report special eurobarometer 196, European Commission, Brussels, Belgium, December 2003.

[60] EUROPEAN PARLIAMENT. Directive 95/46/ec of the european parliament and of the council of 24 october 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. *Official Journal of the European Communities L 281* (1995), 31–50.

[61] EUROPEAN PARLIAMENT, TEMPORARY COMMITTEE ON THE ECHELON INTERCEPTION SYSTEM. Report on the existence of a global system for the interception of private and commercial communications (echelon interception system) (2001/2098(ini)), July 2001.

[62] FIORE, A. T., AND DONATH, J. S. Online personals: an overview. In *CHI '04 extended abstracts on Human factors in computing systems (CHI)* (2004), ACM Press, pp. 1395–1398.

[63] FISCHER, C. S. *America Calling: A Social History of the Telephone to 1940.* University of California Press, 1995.

[64] FORESTI, G. L., MAHONEN, P., AND REGAZZONI, C. S., Eds. *Multimedia Video-Based Surveillance Systems: Requirements, Issues and Solutions.* Kluwer Academic Publishers, 2000.

[65] FORRESTER RESEARCH. Post-web retail, September 1999. http://www.forrester.com/.

[66] FORRESTER RESEARCH. Privacy concerns cost e-commerce $15 billion, September 2001. http://www.forrester.com/.

[67] FREITAS, A. A. *Data Mining and Knowledge Discovery with Evolutionary Algorithms.* Springer Verlag, 2002.

[68] FRIED, C. Privacy [a moral analysis]. *Yale Law Journal 77* (1968), 475–493.

[69] GARFINKEL, S. Adopting fair information practices to low cost rfid systems. In *Proceedings of Ubiquitous Computing (UBICOMP) Privacy Workshop* (2002).

[70] GARTNER RESEARCH. Increased phishing and online attacks cause dip in consumer confidence, June 2005. http://www.gartner.com/.

[71] GERSTEIN, R. S. Intimacy and privacy. *Ethics 89* (1978), 76–81.

[72] GODFREY, P., SHIPLEY, R., AND GRYZ, J. Maximal vector computation in large data sets. In *Proceedings of the 31st international conference on Very large data bases (VLDB)* (2005), VLDB Endowment, pp. 229–240.

[73] GOLDBERG, D., NICHOLS, D., OKI, B. M., AND TERRY, D. Using collaborative filtering to weave an information tapestry. *Communications of the ACM 35*, 12 (1992), 61–70.

[74] GOLDBERG, I. Privacy-enhancing technologies for the internet, ii, five years later. In *Privacy Enhancing Technologies (PET)* (April 2002), R. Dingledine and P. F. Syverson, Eds., vol. 2482 of *Lecture Notes in Computer Science*, Springer-Verlag, pp. 209–213.

[75] GOLDSCHLAG, D., REED, M., AND SYVERSON, P. Onion routing. *Communications of the ACM 42*, 2 (1999), 39–41.

[76] GOLDSCHLAG, D. M., REED, M. G., AND SYVERSON, P. F. Hiding routing information. In *Information Hiding* (1996), pp. 137–150.

[77] GOLDWASSER, S. Multi party computations: past and present. In *Proceedings of the sixteenth annual ACM symposium on Principles of distributed computing (PODC)* (1997), ACM Press, pp. 1–6.

[78] GOVINDARAJAN, K., JAYARAMAN, B., AND MANTHA, S. Preference logic programming. In *Proceedings of the 12th International Conference on Logic Programming* (1995), MIT Press, pp. 731–745.

[79] GOVINDARAJAN, K., JAYARAMAN, B., AND MANTHA, S. Preference queries in deductive databases. *New Generation Computing 19*, 1 (2001), 57–86.

[80] GRINTER, R. E., AND PALEN, L. Instant messaging in teenage life. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW)* (2004), ACM Press, pp. 21–30.

[81] HABERMAS, J. *The structural transformation of the public sphere: An inquiry into a category of bourgeois society. Translated by Thomas Burger.* Polity, 1962 trans 1989.

[82] HÄKKILÄ, J., AND CHATFIELD, C. Toward social mobility: "it's like if you opened someone else's letter": User perceived privacy and social practices with sms communication. In *Proceedings of Human Computer Interaction with Mobile Devices and Services (MobileHCI)* (September 2005), ACM Press, pp. 219–222.

[83] HAN, J., AND KAMBER, M. *Data Mining: Concepts and Techniques.* Morgan Kaufmann, 2001.

[84] HANCOCK, J. T., TOMA, C., AND ELLISON, N. The truth about lying in online dating profiles. In *Proceedings of the SIGCHI conference on Human factors in computing systems (CHI)* (2007), ACM Press, pp. 449–452.

[85] HARRIS INTERACTIVE. First major post-9/11 privacy survey finds consumers demanding companies do more to protect privacy, 2002. http://www.harrisinteractive.com/.

[86] HELSCHER, D. Griswold v. connecticut and the unenumerated right of privacy. *Northern Illinois university Law Review 1*, 7 (1994).

[87] HOCHHEISER, H. The platform for privacy preference as a social protocol: An examination within the U.S. policy context. *ACM Transactions on Internet Technology (TOIT) 2*, 4 (2002), 276–306.

[88] HRISTIDIS, V., KOUDAS, N., AND PAPAKONSTANTINOU, Y. Prefer: a system for the efficient execution of multi-parametric ranked queries. *ACM SIGMOD Record 30*, 2 (2001), 259–270.

[89] HUMPHREYS, S. C. Public and private interests in classical athens. *The Classical Journal 73*, 2 (December 2007), 97–104.

[90] IACHELLO, G., SMITH, I., CONSOLVO, S., CHEN, M., AND ABOWD, G. D. Developing privacy guidelines for social location disclosure applications and services. In *Proceedings of Symposium on Usable Privacy and Security (SOUPS)* (July 2005), ACM Press, pp. 65–76.

[91] IBM CORPORATION. Privacy is good for business - interview with IBM Chief Privacy Officer, Harriet Pearson, 2007. http://www.ibm.com/innovation/us/customerloyalty/harriet_pearson_interview.shtml.

[92] IBM GLOBAL SERVICES. IBM multi-national consumer privacy survey, 1999. Conducted by Louis Harris & Associates, Inc.

[93] INTERNATIONAL LABOR ORGANIZATION. Workers privacy part ii: monitoring and surveillance in the workplace conditions of work. *Special series on workers privacy, Digest 12*, 1 (1993).

[94] IPEIROTIS, P. G., GRAVANO, L., AND SAHAMI, M. Probe, count, and classify: categorizing hidden web databases. In *Proceedings of the 2001 ACM SIGMOD international conference on Management of data* (2001), ACM Press, pp. 67–78.

[95] JAY LIN, K. E-commerce technology: Back to a prominent future. *IEEE Internet Computing 12*, 1 (Jan-Feb 2008), 60–65.

[96] JENSEN, C., POTTS, C., AND JENSEN, C. Privacy practices of internet users: Selfreports versus observed behavior. *International Journal of Human-Computer Studies 63* (2005), 203–227.

[97] JUPITER RESEARCH. Seventy percent of us consumers worry about online privacy, but few take protective action, 2002. http://www.jupiterresearch.com/.

[98] KAHN, D. The history of steganography. In *Proceedings of the First International Workshop on Information Hiding* (1996), vol. 1174 of *Lecture Notes in Computer Science*, Springer-Verlag, pp. 1–5.

[99] KARAT, C.-M., KARAT, J., BRODIE, C., AND FENG, J. Evaluating interfaces for privacy policy rule authoring. In *Proceedings of SIGCHI Conference on Human Factors in Computing Systems* (April 2006), ACM Press, pp. 83–92.

[100] KATZENBEISSER, S., AND PETITCOLAS, F. A., Eds. *Information Hiding Techniques for Steganography and Digital Watermarking.* Artech House, Inc., 2000.

[101] KEWNEY, G. Wireless lamp posts take over the world!, 2004. http://www.theregister.co.uk/content/69/34894.html.

[102] KIESSLING, W. Foundations of preferences in database systems. In *Proceedings of the Twenty-Eighth International Conference on Very Large Data Bases (VLDB), 2002* (2002), VLDB Endowment, pp. 311–322.

[103] KIESSLING, W., AND GÜNTZER, U. Database reasoning - a deductive framework for solving large and complex problems by means of subsumption. In *Proceedings of the Third Workshop on Information Systems and Artificial Intelligence: Management and Processing of Complex Data Structures* (London, UK, 1994), Springer-Verlag, pp. 118–138.

[104] KIESSLING, W., AND KÖSTLER, G. Preference sql - design, implementation, experiences. In *Proceedings of the Twenty-Eighth International Conference on Very Large Data Bases (VLDB), 2002* (2002), VLDB Endowment, pp. 990–1001.

[105] KOHL, J. T., NEUMAN, B. C., AND T'SO, T. Y. The evolution of the kerberos authentication system. In *Distributed Open Systems.* IEEE Computer Society Press, 1994, pp. 78–94.

[106] KONSTAN, J. A., MILLER, B. N., MALTZ, D., HERLOCKER, J. L., GORDON, L. R., AND RIEDL, J. Grouplens: applying collaborative filtering to usenet news. *Communications of the ACM 40*, 3 (1997), 77–87.

[107] KOSSMANN, D., RAMSAK, F., AND ROST, S. Shooting stars in the sky: an online algorithm for skyline queries. In *Proceedings of the 28th international conference on Very Large Data Bases (VLDB* (2002), VLDB Endowment, pp. 275–286.

[108] KÖSTLER, G., KIESSLING, W., THÖNE, H., AND GÜNTZER, U. Fixpoint iteration with subsumption in deductive databases. *Journal of Intelligent Information Systems 4*, 2 (1995), 123–148.

[109] LACROIX, M., AND LAVENCY, P. Preferences; putting more knowledge into queries. In *Proceedings of the 13th International Conference on Very Large Data Bases* (San Francisco, CA, USA, 1987), VLDB '87, Morgan Kaufmann Publishers Inc., pp. 217–225.

[110] LANGHEINRICH, M. Privacy by design principles of privacy-aware ubiquitous systems. In *Proceedings of Ubiquitous Computing (UBICOMP)* (2001), Springer-Verlag, pp. 273–291.

[111] LI, N., LI, T., AND VENKATASUBRAMANIAN, S. t-closeness: Privacy beyond k-anonymity and l-diversity. In *IEEE 23rd International Conference on Data Engineering, 2007 (ICDE)* (April 2007), IEEE Press, pp. 106–115.

[112] LIBERTY ALLIANCE PROJECT. Liberty Identity Assurance Framework 1.1. Tech. rep., Liberty Alliance, June 2008. http://www.projectliberty.org/liberty/resource_center/papers.

[113] LOUKIDES, G., AND SHAO, J. Capturing data usefulness and privacy protection in k-anonymisation. In *Proceedings of the 2007 ACM symposium on applied computing (SAC)* (2007), ACM Press, pp. 370–374.

[114] LYSYANSKAYA, A., RIVEST, R. L., SAHAI, A., AND WOLF, S. Pseudonym systems. In *Proceedings of the 6th Annual International Workshop on Selected Areas in Cryptography (SAC)* (2000), vol. 1758 of *Lecture Notes in Computer Science*, Springer-Verlag, pp. 184–199.

[115] MACHANAVAJJHALA, A., KIFER, D., GEHRKE, J., AND VENKITASUBRAMANIAM, M. L-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD) 1*, 1 (2007), 3.

[116] MADRIA, S. K., BHOWMICK, S. S., NG, W. K., AND LIM, E. P. Research issues in web data mining. In *DataWarehousing and Knowledge Dis-*

covery, vol. 1676 of *Lecture Notes in Computer Science*. Springer-Verlag, 1999, p. 805.

[117] MARCH, W., AND FLEURIOT, C. Girls, technology and privacy: "is my mother listening?". In *Proceedings of the SIGCHI conference on Human Factors in computing systems (CHI)* (2006), ACM Press, pp. 107–110.

[118] MICROSOFT CORPORATION. Protecting americans privacy, 2007. http://www.microsoft.com/issues/essays/2007/03-20ProtectingPrivacy.mspx.

[119] MILBERG, S. J., BURKE, S. J., SMITH, H. J., AND KALLMAN, E. A. Values, personal information privacy, and regulatory approaches. *Communications of the ACM 38*, 12 (1995), 65–74.

[120] MITNICK, K., AND SIMON, W. *The Art of Deception: Controlling the Human Element of Security*. Wiley, 2002.

[121] MOOR, J. H. Towards a theory of privacy in the information age. *SIGCAS Computers and Society 27*, 3 (1997), 27–32.

[122] MOORES, T. T., AND DHILLON, G. Do privacy seals in e-commerce really work? *Communications of the ACM 46*, 12 (2003), 265–271.

[123] MURPHY, R. F. Social distance and the veil. In *Philosophical dimensions of privacy: An anthology*. Cambridge University Press, 1984, ch. 2, pp. 34–54.

[124] MURPHY, R. S. Property rights in personal information: An economic defense of privacy. *Georgetown Law Journal 84* (1996), 2381.

[125] NATIONAL ASSEMBLY OF FRANCE. Declaration of the rights of man and of the citizen, 1789. http://www.hrcr.org/docs/frenchdec.html.

[126] NEUMAN, B. C., AND TS'O, T. Kerberos: An authentication service for computer networks. *IEEE Communications 32*, 9 (September 1994), 33–38.

[127] NORRIS, C., AND ARMSTRONG, G. *The maximium surveillance society: The rise of CCTV*. Berg, 1999.

[128] ODLYZKO, A. Privacy, economics, and price discrimination on the internet. In *Proceedings of the 5th international conference on Electronic commerce* (2003), ACM Press, pp. 355–366.

[129] OFFICE FOR NATIONAL STATISTICS (ONS). E-commerce survey of business, 2006. http://www.statistics.gov.uk/statbase/Product.asp?vlnk=6645.

[130] OFFICE FOR NATIONAL STATISTICS (ONS). E-commerce survey of business, 2007. http://www.statistics.gov.uk/pdfdir/ecom1108.pdf.

[131] OFFICE FOR NATIONAL STATISTICS (ONS). News release - e-commerce survey of business, 2007. http://www.statistics.gov.uk/pdfdir/ecomnr1108.pdf.

[132] ORGANIZATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT. Guidelines on the protection of privacy and transborder flows of personal data. Tech. rep., Organization for Economic Co-operation and Development, 1980.

[133] PAPADIAS, D., TAO, Y., FU, G., AND SEEGER, B. An optimal and progressive algorithm for skyline queries. In *Proceedings of the 2003 ACM SIGMOD international conference on Management of data* (2003), ACM Press, pp. 467–478.

[134] POSNER, R. An economic theory of privacy. *Regulation* (1978), 19–26. (May/June).

[135] POSNER, R. A. An economic theory of privacy. In *Philosophical Dimensions of Privacy: An Anthology*. Cambridge University Press, 1984, ch. 15, pp. 333–345.

[136] PRIVACY AND AMERICAN BUSINESS. Consumer privacy attitudes: a major shift since 2000 and why. *Privacy and American Business Newsletter 10*, 6 (2003).

[137] PRIVACY INTERNATIONAL. National id cards. http://www.privacyinternational.org/issues/idcard/_index.html.

[138] PROSSER, W. L. Privacy [a legal analysis]. *Harvard Law Review 48* (1960), 338–423.

[139] RACHELS, J. Why privacy is important. *Philosophy & Public Affairs 4*, 4 (1975), 323–333.

[140] REAGLE, J., AND CRANOR, L. F. The platform for privacy preferences. *Communications of the ACM 42*, 2 (1999), 48–55.

[141] REISS, S. P. Practical data-swapping: the first steps. *ACM Transactions on Database Systems (TODS) 9*, 1 (1984), 20–37.

[142] REITER, M. K., AND RUBIN, A. D. Crowds: anonymity for web transactions. *ACM Transactions on Information Systems Security 1*, 1 (1998), 66–92.

[143] REITER, M. K., AND RUBIN, A. D. Anonymous web transactions with crowds. *Communications of the ACM 42*, 2 (1999), 32–48.

[144] REZGUI, A., OUZZANI, M., BOUGUETTAYA, A., AND MEDJAHED, B. Preserving privacy in web services. In *Proceedings of the fourth international workshop on Web information and data management* (2002), ACM Press, pp. 56–62.

[145] RIZZI, S. Olap preferences: a research agenda. In *Proceedings of the ACM tenth international workshop on Data warehousing and OLAP (DOLAP)* (2007), ACM Press, pp. 99–100.

[146] RUTHVEN, I., AND LALMAS, M. A survey on the use of relevance feedback for information access systems. *The Knowledge Engineering Review 18*, 2 (2003), 95–145.

[147] SAXONHOUSE, A. W. Classical greek conceptions of public and private. In *Public and Private in Social Life*. St. Martins Press, 1983, ch. 15, pp. 363–384.

[148] SCHAFER, J. B., KONSTAN, J. A., AND RIEDL, J. E-commerce recommendation applications. *Journal of Data Mining and Knowledge Discovery 5*, 1-2 (2001), 115–152.

[149] SCHLAEGER, C., AND PERNUL, G. Authentication and authorisation infrastructures in b2c e-commerce. In *Proceedings of the Sixth International Conference on Electronic Commerce and Web Technologies (EC-Web'05)* (2005), vol. 3590 of *Lecture Notes in Computer Science*, Springer-Verlag, pp. 306–315.

[150] SCHLÖRER, J. Identification and retrieval of personal records from a statistical data bank. *Methods of Information in Medicine 14*, 1 (1975), 7–13.

[151] SCHLÖRER, J. Confidentiality and security in statistical data banks. In *Proceedings of Workshop on Data Documentation: Some Principles and Applications in Science and Industry* (1977), W. Gaus and R. Henzler, Eds., Verlag Dokumentation, pp. 101–123.

[152] SCHLÖRER, J. Security of statistical databases: multidimensional transformation. *ACM Transactions on Database Systems 6*, 1 (1981), 95–112.

[153] SCHOEMEN, F. D. *Philosophical dimensions of privacy: An anthology.* Cambridge University Press, 1984.

[154] SHAPIRA, B., ELOVICI, Y., MESHIACH, A., AND KUFLIK, T. PRAW—a privacy model for the web: Research articles. *Journal of the American Society for Information Science and Technology 56*, 2 (2005), 159–172.

[155] SHOSHANI, A. Statistical databases: Characteristics, problems, and some solutions. In *VLDB '82: Proceedings of the 8th International Conference on Very Large Data Bases* (1982), Morgan Kaufmann Publishers Inc., pp. 208–222.

[156] STIGLER, G. J. An introduction to privacy in economics and politics. *Journal of Legal Studies, 9* (1975).

[157] SWEENEY, L. k-anonymity: a model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10*, 5 (2002), 557–570.

[158] SYVERSON, P. The paradoxical value of privacy. In *Proceedings of the 2nd Annual Workshop on Economics and Information Security (WEIS)* (2003).

[159] TERVEEN, L., HILL, W., AMENTO, B., MCDONALD, D., AND CRETER, J. Phoaks: a system for sharing recommendations. *Communications of the ACM 40*, 3 (1997), 59–62.

[160] THOMSON, J. J. The right to privacy. *Philosophy & Public Affairs 4*, 4 (1975), 295–314.

[161] TORLONE, R., AND CIACCIA, P. Finding the best when it's a matter of preference. In *Proceedings 10th Italian National Conference on Advanced Database Systems (SEBD), 2002* (2002), pp. 347–360.

[162] TSAI, J., EGELMAN, S., CRANOR, L., AND ACQUISTI, A. The effect of online privacy information on purchasing behavior: An experimental study. In *Proceedings of the 6th Workshop on the Economics of Information Security (WEIS)* (2007).

[163] TUERKHEIMER, F. M. The underpinnings of privacy protection. *Communications of the ACM 36*, 8 (1993), 69–73.

[164] UNITED NATIONS. The universal declaration of human rights, 1948. http://www.un.org/Overview/rights.html.

[165] UNITED NATIONS GENERAL ASSEMBLY RESOLUTION 2200A (XXI). International covenant on civil and political rights, 1966. http://www2.ohchr.org/english/law/ccpr.htm.

[166] UNITED STATES CENSUS BUREAU. 2004 e-commerce multi-sector report, May 2006. http://www.census.gov/estats.

[167] UNITED STATES CENSUS BUREAU. 2006 e-commerce multi-sector report, May 2008. http://www.census.gov/estats.

[168] UNITED STATES CENSUS BUREAU. 2007 e-commerce multi-sector report, May 2009. http://www.census.gov/estats.

[169] UNITED STATES DEPARTMENT OF HEALTH EDUCATION AND WELFARE. Records, computers and the rights of citizens, report of the secretarys advisory committee on automated personal data systems. Tech. rep., United States Department of Health Education and Welfare, 1973.

[170] UNITED STATES HOUSE OF REPRESENTATIVES. Usa patriot act, 2001. (Public Law 107-56 [H.R. 3162]).

[171] US DEPTARTMENT OF COMMERCE - OFFICE OF FEDERAL STATISTICAL POLICY AND STANDARDS. Statistical policy working paper 2: report on statistical disclosure and disclosure avoidance techniques, 1978. http://www.fcsm.gov/working-papers/sw2.html.

[172] VOLOKH, E. Personalization and privacy. *Communications of the ACM 43*, 8 (2000), 84–88.

[173] WARREN, S. D., AND BRANDEIS, L. D. The right to privacy [the implicit made explicit]. *Harvard Law Review 4*, 5 (1890), 193–220.

[174] WASSERSTROM, R. A. Privacy: Some arguments and assumptions. In *Philosophical Dimensions of Privacy: An Anthology.* Cambridge University Press, 1984, ch. 14, pp. 317–332.

[175] WEINTRAUB, J. The theory and politics of the public/private distinction. In *Public and private in thought and practise.* University of Chicago Press, 1997, ch. 1, pp. 1–42.

[176] WEINTRAUB, J., AND KUMAR, K. *Public and private in thought and practise.* University of Chicago Press, 1997.

[177] WEL, L. V., AND ROYAKKERS, L. Ethical issues in web data mining. *Ethics and Information Technology 6*, 2 (2004), 129–140.

[178] WESTIN, A. F. *Privacy and freedom.* Atheneum Publishers, 1967.

[179] WESTIN, A. F. The origins of modern claims to privacy. In *Philosophical dimensions of privacy: An anthology.* Cambridge University Press, 1984, ch. 3, pp. 56–74.

[180] WESTIN, A. F. Equifax-harris consumer privacy survey, 1991. New York: Louis Harris & Associates.

[181] WESTIN, A. F. Equifax-harris consumer privacy survey, 1994. New York: Louis Harris & Associates.

[182] WESTIN, A. F. Opinion surveys: What consumers have to say about information privacy, S.o.C: The house committee on energy and commerce, trade, and consumer protection, 2001.

[183] WITTEN, I. H., AND FRANK, E. *Data Mining: practical machine learning tools and techniques.* Morgan Kaufmann, 2005.

[184] WOLF-TILO BALKE, U. G., AND ZHENG, J. X. Efficient distributed skylining for web information systems. In *Advances in Database Technology (EDBT)*, vol. 2992 of *Lecture Notes in Computer Science.* Springer-Verlag, 2004, pp. 256–273.

[185] YATES, J. Millar vs. Taylor. In *4 Burr.* 1769, pp. 2303–2379.