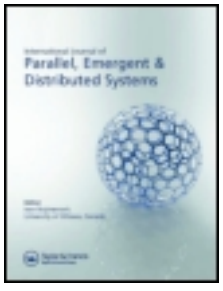

International Journal of Parallel, Emergent and Distributed Systems
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gpaa20

COSMOS: Towards an integrated and scalable service for analysing social media on demand

Peter Burnap[a], Omer Rana[a], Matthew Williams[b], William Housley[b], Adam Edwards[b], Jeffrey Morgan[b], Luke Sloan[b] & Javier Conejero[c]
[a] School of Computer Science & Informatics, Cardiff University, Cardiff, UK
[b] School of Social Sciences, Cardiff University, Cardiff, UK
[c] University of Castilla-La Mancha (UCLM), Albacete, Spain
Published online: 06 May 2014.

To cite this article: Peter Burnap, Omer Rana, Matthew Williams, William Housley, Adam Edwards, Jeffrey Morgan, Luke Sloan & Javier Conejero (2014): COSMOS: Towards an integrated and scalable service for analysing social media on demand, International Journal of Parallel, Emergent and Distributed Systems, DOI: 10.1080/17445760.2014.902057

To link to this article: http://dx.doi.org/10.1080/17445760.2014.902057

International Journal of Parallel, Emergent and Distributed Systems, 2014
http://dx.doi.org/10.1080/17445760.2014.902057

COSMOS: Towards an integrated and scalable service for analysing social media on demand

Peter Burnap[a], Omer Rana[a]*, Matthew Williams[b], William Housley[b], Adam Edwards[b], Jeffrey Morgan[b], Luke Sloan[b] and Javier Conejero[c]

[a]*School of Computer Science & Informatics, Cardiff University, Cardiff, UK;* [b]*School of Social Sciences, Cardiff University, Cardiff, UK;* [c]*University of Castilla-La Mancha (UCLM), Albacete, Spain*

(*Received 4 March 2014; accepted 4 March 2014*)

The growing number of people using social media to publish their opinions, share expertise, make social connections and promote their ideas to an international audience is creating data on an epic scale. This enables social scientists to conduct research into ethnography, discourse analysis and analysis of social interactions, providing insight into today's society, which is largely augmented by social computing. The tools available for such analysis are often proprietary and expensive, and often non-interoperable, meaning the rapid marshalling of large data-sets through a range of analyses is arduous and difficult to scale. The collaborative online social media observatory (COSMOS), an integrated social media analysis tool is presented, developed for open access within academia. COSMOS is underpinned by a scalable Hadoop infrastructure and can support the rapid analysis of large data-sets and the orchestration of workflows between tools with limited human effort. We describe an architecture and scalability results for the computational analysis of social media data, and comment on the storage, search and retrieval issues associated with massive social media data-sets. We also provide an insight into the impact of such an integrated on-demand service in the social science academic community.

Keywords: computational sociology; information search and retrieval; performance evaluation; text analysis

1. Introduction

In recent years, social networking applications such as Twitter, Facebook and Google + have created an open digital platform that has empowered global citizens to publish their opinions, share expertise, make social connections and promote their ideas, work and themselves to an international audience. The engagement of certain proportions of society with social networks offers the exciting potential for researchers interested in observing and interpreting society to apply established theory and methods to the emerging online society, and innovate to produce new methods. With over a billion active users, social networks are producing massive amounts of data (volume) on a real-time basis (velocity) with implicit sociological attributes such as belief, opinion, behaviour, societal structure and influence (variety). These data exhibit the key traits of what is often referred to as Big Data – volume, velocity and variety. Behaviour and interaction has been empirically studied for centuries by social scientists, with rigorously applied methods and analyses being developed to yield meaningful results. In the age of Big Data and the increasingly interconnected online society, there is a new challenge – to support the development of

*Corresponding author. Email: p.burnap@cs.cardiff.ac.uk

interoperable and scalable methods and tools that can be applied to digitised social behaviour produced via social networks – or Big Social Data, and that are usable by those with the expertise to interpret the analysed data, i.e. social scientists.

The uptake of social computing coincided with a seminal article published by Savage and Burrows in which they anticipate a 'coming crises of empirical sociology' due to the ability of large corporate organisations to marshall and analyse such Big Data to their benefit [24]. They argue that traditional social science methodologies such as focus groups and surveys are now outdated in the face of these new data, and fail to capture the nuance of online society. However, there is emerging interest in applying digital methodologies to support social media analysis for the understanding of community networks and cohesion [9,34], communication patterns [7] and topic-specific sentiment and opinion [29], for the benefit of society, as well as for financial gain. Social media is also being used for forecasting purposes, such as to predict the outcome of political elections [32] or revenue of new movies [2]. This poses the question whether such digital methodologies could mitigate the 'coming crises' to some degree if the tools to support their application were available, and thus the motivation of this work is to provide such tools in a manner accessible to social science.

We present the collaborative online social media observatory (COSMOS), a distributed digital social research platform, providing on-demand analytics for the purposes of observing and inductively interpreting socially significant evidence gathered via the emerging uptake of social computing (e.g. Twitter, Facebook, Google+, Blogs, News reporting agencies). The volume of data produced on a daily basis requires significant computational resources to analyse. For example, COSMOS collects ~3.5–4 million tweets a day via the Twitter Streaming API. The collection has been operating for 20 months and has collected ~2.5 billion tweets. To perform a longitudinal analysis of, say public opinion and sentiment, around a socially significant event (e.g. a political campaign, change of legislation, world sporting event, etc.) could require analysis of several weeks' worth of data. However, social media analysis may require several 'tweaks' to the study parameters, and therefore requires a more interactive way of analysing data. For example, age, gender, location and topic of study within the event may change, as sociological hypotheses are formed and tested. Therefore, the computational analysis must be able to complete much faster to give a more acceptable wait-time for the researcher. The researcher should be able to invoke the computational resources to support large-scale data analysis on-demand and resources need to be dynamically allocated depending on the size of the job.

Throughout the article we will refer to social scientists as the end users of COSMOS. Though this is not the only group of potential users, COSMOS is designed to support researchers interested in analysing large data-sets of socially significant data, but who lack the technical skills to collect, archive and engage with data and open source software tools that perform this task. A variety of tools already exist to perform sentiment, content and social network analysis, and while COSMOS has developed new tools for social media analysis, the novelties of COSMOS (as a data analytics platform) are (i) that it provides a hosted and scalable computational infrastructure to support the automated collection and an on-demand analysis of massive amounts of data, from social media sources as well as other digital sources such as government-curated and administrative data. The on-demand analysis phase supports the processing of millions of data-points that would take much longer on a desktop computer and facilitates data linking between largely non-interoperable sources and (ii) it comprises an integrated user interface to a number of analytical services, designed for users with relatively limited computing knowledge, and

to visualise Big Social Data in such a way that points of interest clearly stand out and support inductive reasoning and further hypothesis postulation, and can be easily marshalled between previously non-interoperable analytic services as a dynamic workflow. The translation of data structures between services is transparent to the user, as is the fact that COSMOS is underpinned by a Hadoop infrastructure to support large-scale analytics. Though, importantly, the algorithms and processes behind the scenes are all open for methodological inspection and critique [8]. In Section 1.1, we describe related work and compare COSMOS with other related approaches. In Section 2, the overall architecture of the COSMOS system, along with the various services that it currently implements and supports are presented. Performance results of the Hadoop-based implementation are presented in Section 3, using a data-set consisting of 15 million tweets representing a collection from a global event. In Section 5, we reflect on the ethical implications of a system such as COSMOS that programmatically collects, archives and analyses publicly available data published by global citizens, followed by conclusions in Section 6.

1.1 Related work

Various academic approaches to analyse Big Social Data have some shared objectives with COSMOS. For example, Prometheus is a peer-to-peer service that collects social data from a number of sources and applies social inferencing techniques, but it is more focused on privacy-aware social data management than supporting social scientific insight [12]. Truthy tracks political memes in Twitter and helps to detect smear campaigns, and other misinformation in the context of US political elections [11]. However, the target of its analysis is pre-defined and not controlled by the researcher. These are encouraging related works as they demonstrate a requirement for Big Social Data analysis at a large scale driven by sociological insight.

Over recent years, there has also been an abundance of commercial social media analysis tools created primarily to facilitate online advertising and marketing, from companies such as Radian 6 (now part of SalesForce), Palantir and IBM. Indeed, Twitter who produce large volumes of social media data make use of data manipulation tools such as PigLatin and underpin analytics with a Hadoop cluster. Many of these tools, however, are difficult to use in the context of academic research due to their cost and/or proprietary closed (black box) technology that does not lend itself to methodological inspection and attributing confidence to the results.

COSMOS proposes the use of a suite of openly available text and network analysis tools, which enable questions to be formulated that are relevant for academic researchers and for analysing self-reported data. There are also free tools that can support some types of analysis, such as sentiment analysis tools, SentiStrength [30] and LIWC [23], and social network analysis tools, Gephi [3] and NodeXL [26]. LIWC also provides many more categories for text than sentiment and tension and is widely used in the social sciences for text analysis (primarily based on the use of particular word groupings), through a licence fee. However, these tools require a certain degree of technical sophistication to collect and structure data into a tool-specific file format, such as CSV or GraphML. It is also often difficult to use them directly in a scalable manner – as access to source code for these tools is not available. Running a series of analyses requires the execution of separate processes with data being transferred manually between tools. COSMOS provides a single interface to a number of tools, with no data collection overhead and automated translation of input files from one format to another. For example, a user could extract data from the COSMOS

archive of Twitter data, containing the term *olympics* posted between May and August 2012, and pose a number of sociological questions such as: What are the key words and hashtags being used to describe events? How does sentiment change over time and in relation to key events? Who is talking to whom about this event and are there any leading actors in the social network? Are there gender-based differences in opinion at certain points in time? And can we identify clusters of geo-located data? Furthermore, each question may lead to another question, which requires a different tool to answer. Therefore, the results of one analysis could prescribe the refinement of the data-set (e.g. by time or keyword) and the appropriate tool to answer the next question. This is supported by COSMOS without needing to handle any data or configure any tools. Perhaps even more important is the paradigm of inductive reasoning via the visual filtering of large data-sets which has been applied in fields such as medical imaging, virtual prototyping and geotechnical data-sets, but is yet to be applied to social media in the way afforded by COSMOS. The human eye is extremely adept at anomaly detection. Using the 'olympics' example, if visualised longitudinally, a spike in sentiment at a particular point in time can be identified, which may lead to a refinement of the data-set to focus on the data related to that point in time and marshall it through a social network analysis to reveal the key actors at that time, or through a geospatial analysis to identify geographic distribution or clusters of tweets at this time. Then, the time frame could be moved back a week, and the same analysis run, to see how the situation has changed.

2. COSMOS: architecture and functionality

In this section, we describe the data collection, sampling and archiving process used in COSMOS. We also outline how COSMOS can be used in real-time or using historic data to undertake observation and visualisation of Big Social Data for longitudinal studies, mining millions of tweets archived using the NoSQL database MongoDB, and scaled to support data analysis on Twitter data using the OpenNebula system, with Map/Reduce-based analysis using Hadoop and through the use of the KVM hypervisor.

2.1 Data collection

COSMOS programmatically collects data from a number of sources using publicly accessible APIs – with a particular focus on Twitter in this article. Twitter has become established as a data source for public opinion and behaviour mining. It is possible to understand public mood [5], opinion [21,29], tension and cohesion [6,9] and communication patterns [7] from Twitter data. COSMOS also has a persistent connection to the UK Police API, which provides crime data on a local district level for the previous month, and being linked to the UK Office for National Statistics, who hold census data (as well as many other curated data-sets), which includes for example, district-level unemployment figures, ethnic composition of the community and population size. This allows further social scientific insight through the linkage of data sources and the analysis and verification of opinion and behaviour identified through Twitter, with reference to local crime rates, types of crime, unemployment levels, population density, and ethnic composition in a community.

COSMOS collects a random 1% sample from the Twitter API (commonly referred to as the 'spritzer'). It is also possible to collect 10% from the API ('gardenhose') or the full 100% of all tweets ('firehose'). However, the data storage requirements for 100% are impractical for many academic establishments. Previous investigations of a sample of 113

million tweets from the spritzer have shown that, in the UK, the gender demographic of Twitter users mirrors that of the UK census within 0.1% and their geographic distribution is also in proportion to the population density of the UK [25]. Only 0.85% of this sample included geo-coded tweets where the user has agreed to submit the longitudinal and latitudinal coordinates of their position when they sent the tweet. This is the only way to guarantee an accurate location for the tweet. This amounts to 30,000 tweets per day that can be accurately geo-located. Furthermore, related research has demonstrated that the spritzer provides a promising representation of the topics, actors and behaviour exhibited in the firehose [15].

Table 1 provides statistics for a 14-day sample data-set from the COSMOS archive, collected from the spritzer stream. Although each tweet (including metadata) is only 0.57 KB on average, with more than 54 million tweets, the archive requires nearly 30 GB disk space. Projecting the statistics for 100% of all Twitter data (the firehose), we would expect the collection size to approach nearly 3 TB for a 14-day period. Because of its size, the Twitter stream collected daily by COSMOS is stored in a NoSQL database. We choose MongoDB because of its Hadoop connector to support parallel data analysis (see Section 2.3). MongoDB stores data in collections, which are effectively databases of individual data objects. Twitter data are obtained through the Twitter API in JSON format, which is suitable for direct storage in a MongoDB collection. The Twitter API provides various metadata for each tweet collected, including a status element (including the text of the tweet, the time it was posted, geo-location, etc.), and a user element (including profile name, profile location, etc.). Full details of the structure of Twitter metadata can be found in [6].

For speed of search, MongoDB collections can be indexed. For instance, to run a keyword search across the 14 days of Twitter data (54 million tweets) without indexing takes approximately 10 seconds. When indexed using a unigram index, the search is almost instantaneous. However, MongoDB requires the index for a collection to be held in RAM and, as Table 1 shows, the size of an index of unigrams in the collection is 11.56 GB. Therefore, RAM is likely to be exhausted with a collection larger than a few months. To overcome this, we archive collections on a monthly basis, although not ideal this situation is improved when distributing the data using the Hadoop environment to multiple machines. Each collection includes ~100 million tweets and is unigram indexed for optimum information retrieval speed. The choice to index by unigram, as opposed to n-gram is to reduce the size of the index. The choice to index words as opposed to other demographic data that can be derived from Twitter data, such as geography, gender of the tweeter or language was taken for two reasons. First, because not all tweets include demographic data and we wanted to include all data items in the search; and second because essentially we are providing a service to analyse opinion, communication and behaviour in an online environment where studies are most likely to be conducted around

Table 1. Twitter MongoDB collection statistics.

	Spritzer (1%) (actual)	Firehose (100%) (projected)
Collection size	29.56 GB	2.88 TB
Count	54,727,358	5.472 Billion
Avg. object size	0.57 KB	–
Unigram index size	11.56 GB	1.13 TB

a particular topic or event. This is supported by a great deal of literature focusing on analysing topics and events such as politics [32], sports [9], social events [16] and health and well-being [4], to name a few. The primary search terms in many cases are identified from the text itself with demographic data being a secondary focus. Each monthly collection requires ~24 GB RAM to operate. However, when conducting longitudinal studies across a number of months, collections need to be loaded and unloaded dynamically. COSMOS handles this transparently to the user by identifying the search range (e.g. May–September 2013), and swapping collections iteratively. This provides a valuable service to researchers who are not traditionally trained in handling large volumes of data, but are adept at inductively inferring sociological trends and anomalies using computer-assisted qualitative data analysis (CAQDAS) tools to process data, which allow the iterative and automated analysis of data-sets into a structured and annotated form. Such tools are generally used in the social sciences to support activities such as transcription analysis, text interpretation and content analysis. Furthermore, given the archival format, parallel searches across a number of months are possible. While CAQDAS tools such as Atlas.ti and NVIVO normally operate off a single data-set on the desktop, COSMOS has access to a cluster of 32 high-performance machines, each with 48 GB RAM; therefore we can search up to 32 months in parallel using a distributed implementation of Hadoop, providing hitherto inaccessible computational power to social scientific research.

2.2 *Data analysis*

COSMOS currently provides nine modes of analysis, at an individual tweet level or at a corpus level, which include tweeter gender and sentiment analysis at an individual level, and social network analysis at the corpus level. The following types of analysis are supported: *Individual Level*: gender detection, language detection, sentiment analysis, tension detection, qualitative overview and geospatial location; and at a *corpus level*: keyword/tweet frequency analysis, social network analysis, crime and census data and hashtag usage analysis. We present these in more detail in this section.

Gender detection is used to derive the portrayed gender of the person who posted the tweet (the tweeter). The gender detection algorithm works by extracting the profile name from each tweet's metadata, then extracting the first name of the tweeter, and maps this onto a database of 44,000 names [14] that have been manually classified as male, female and unisex. There are limitations to this approach in that the tweeter could use a fake name that could falsely represent their gender, or indeed a non-person name (e.g. 'God', 'The Bomb'). Furthermore, the name classification is mostly based on names originating in the European region, which requires extensions to include a wider range of names. COSMOS provides a breakdown of the filtered data-set as a percentage tweets in the corpus posted by male, female, unisex and unknown gender users. It also uses gender detection in other analysis, such as sentiment analysis and geospatial location – which will be discussed later in this section.

Language detection is used to determine the language used in the text of the tweet. Language detection is important for a number of reasons, including the analysis of language use (e.g. proportion of languages mentioning keywords, and geospatial distribution), and the pre-processing of data to remove 'noise' and reduce the computational payload for further analytics. Language is detected using the Language detection library for Java, which accepts a text string as input and outputs a decision as to whether it can be classified into one of 52 possible languages or unknown.

Sentiment analysis is a form of opinion mining that generally requires the identification of an entity on which the opinion is focused (e.g. a person); attributes of the entity (e.g. the person's political perspective); views, attitudes or feelings towards the entity and its attributes (commonly defined as sentiment); an opinion holder and a time at which the sentiment was expressed [13]. The approach has been used to determine emotional differences between genders on MySpace [31], and study levels of positive and negative sentiment in Facebook [1] and Twitter comments [21]. The outcome of sentiment analysis is also often subjective and based on the existence of a list of keywords in a message. We have therefore validated the 'ground truth' of the sentiment analysis tool in COSMOS with human users and compared the human annotated sentiment score with results returned from COSMOS [9]. This is a semi-automated approach where human input is required to tailor the machine's interpretation of what is positive or negative, and can dramatically increase the speed at which a general opinion on a topic can be obtained. COSMOS has integrated the SentiStrength sentiment analysis tool, which provides a positive and negative score for each text it examines. The average + ve and − ve sentiment scores are plotted on a line chart with time as the horizontal axis and a range of −5 to +5 on the vertical axis, representing variation in sentiment over time, and making peaks and troughs obvious to the user (see Figure 1). COSMOS presents three charts. Individual charts for male and female users, and an aggregated chart. A noted limitation to this is that the individual charts may be skewed by inaccuracies in classifying gender. However, the separation of the gender-specific sentiment charts enables the study of sentiment differences based on gender in relation to different events. For instance, Figure 1 illustrates a noticeable difference between males and females in relation to this event.

Tension detection is a tool that was developed specifically for COSMOS. It implements a conversation analysis methods – membership categorisation analysis – combined with lexicons of expletive terms, tension-specific degradation terms (e.g. racist, homophobic or disabilist terms), and attribution terms, to classify short text (e.g. tweets) into a range of classes on a three-point ordinal scale from low to high tension. This has been robustly tested for precision, recall and accuracy, and performed very well in a study of how online racial tension manifests in online society in relation to a real-world event [9]. Like sentiment analysis, the tension detection tool assigns a numeric value from 0 to 3 for tension, which is plotted in COSMOS on the same chart as sentiment (see Figure 1). This offers an interesting insight into how tension relates to sentiment and, of course, how tension changes over time in relation to key real-world events.

Qualitative overview provides a list of the text in all tweets. This can comprise all tweets within a specified time range, tweets that match the parameters identified using the filters, or a combination of both. The qualitative text of each tweet is displayed (though not the identity of the tweeter), along with two attributional annotations: the gender of the tweeter and the sentiment scores (positive and negative) calculated based on the content of the tweet. This gives the end user an *at a glance* view of the filtered data-set, which can support users in identifying key topics, events, opinions and perspectives from the text, as well as outliers with extreme + ve or − ve sentiment and clusters of male and female tweets.

Keyword frequency analysis visualises the occurrence of specified keywords (or tweet frequency if no keywords are specified) as a bar chart over time. This allows the user to visually identify points of high and low activity in relation to an event or topic. COSMOS visualises frequency using three units of time, by day, hour and minute, each visualised on its own time line (see Figure 2). Each bar in the chart represents a period of time (i.e. day, hour and minute) and users can mouse over each bar and COSMOS will display the text of

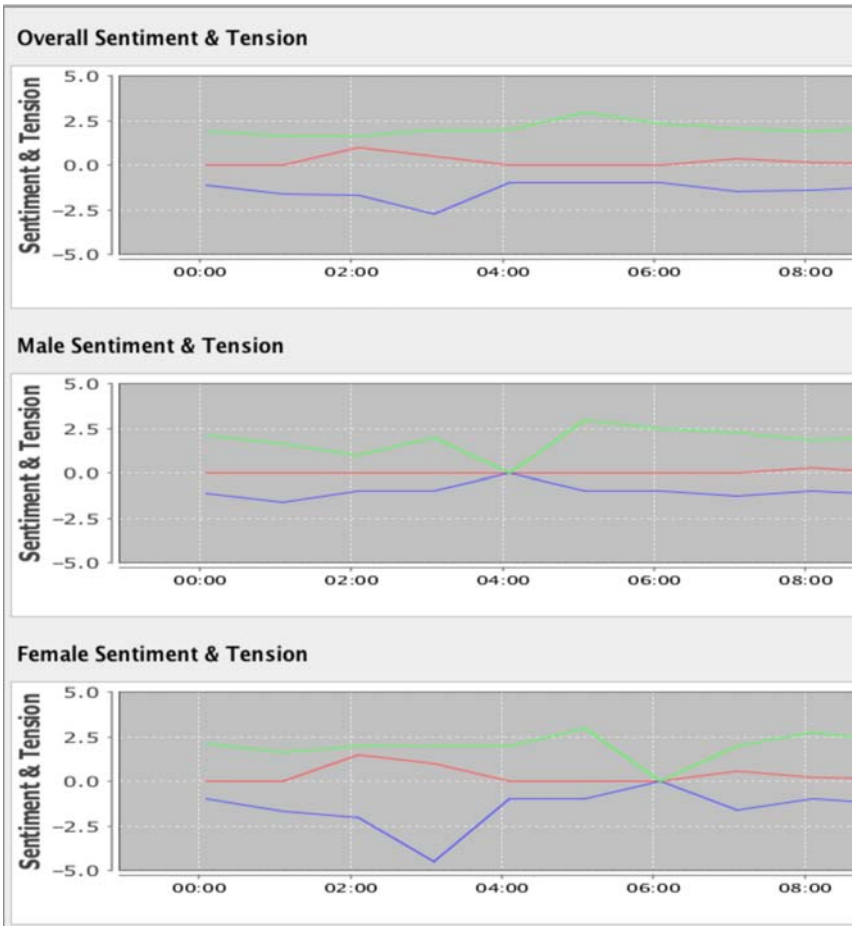

Figure 1. COSMOS sentiment analysis. The green line represents positive sentiment, the red line represents tension and the blue line represents negative sentiment.

the tweets posted during that period. This provides the *at a glance* view to enable the visual identification of key attributes of the tweets (e.g. topic). Sliders can then be used to select a range of bars in the chart, thereby refining the data-set to extend and contract the time frame as a scopic tool to analyse only those tweets posted within the selected range. For instance, if the frequency of 14 days of tweets using the keyword *olympics* were visualised, and there was a spike in frequency at day 4, the sliders could be used to refine the data-set to focus on only tweets posted during day. The hourly frequency during the selected time frame is then plotted by hour below the daily frequency. Hovering the mouse over each hour displays the text of the tweets posted during that period. Suppose there were two peaks, between 09:00 and 11:00 and 13:00 and 15:00 on day 4 – the sliders can be moved to each peak, at which point the minute-by-minute keyword frequency is visualised below the minute frequency. At the finest level of granularity, a list of tweets posted during a single minute of interest can be viewed from a starting point of many days.

Geospatial distribution plots data points on a map. For example, tweets containing a geo-location can be plotted on a map. Our research suggests that, assuming our 1% sample

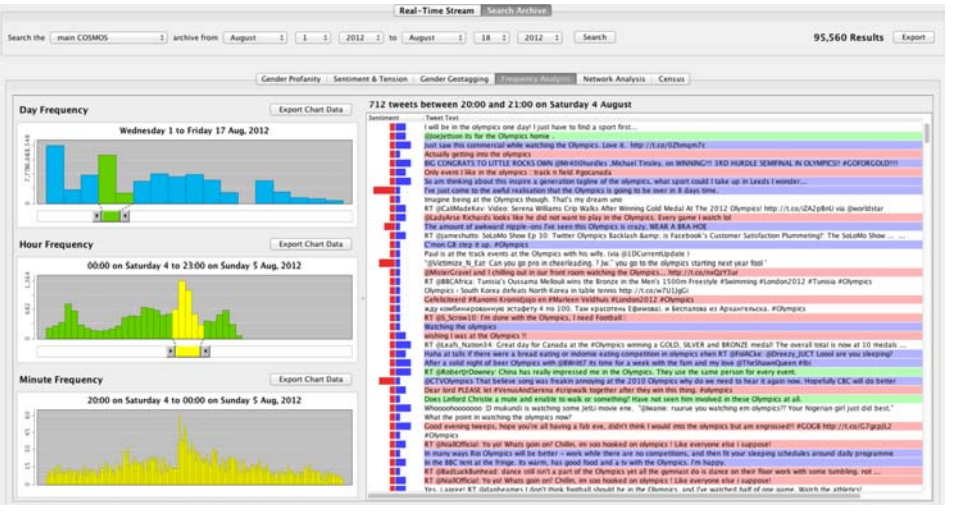

Figure 2. COSMOS frequency analysis.

is indeed random, only 0.85% of tweets are geo-tagged [25]. Further methods of deriving location are from the profile metadata and using keyword matching to identify text referring to a place. Crime and census data are an example of how social media can be used to augment existing sources of administrative and curated data to enhance our understanding of social phenomena. COSMOS presents the user with an outline map of Greater London split by local authority (LA). London has been used as an area around which to test the data augmentation process and eventually COSMOS will be able to display any UK geographical area. Drawing on Census data that is currently stored locally but will be sourced from the ONS API, each LA is coloured according to an aggregated demographic variable – the percentage of the population that is unemployed. Other information about an LA is displayed in a side panel as the cursor moves over the geographical area of interest including ethnic composition and total population. When the user clicks on the map the position of the cursor is converted into a geospatial coordinate compatible with the Police API, which receives a query requesting the crime statistics for the area in question. When COSMOS receives the crime information it is presented in another panel on the platform as a pie chart, where each slice represents a particular type of crime. The Police API provides data at a smaller geographical level than LA; therefore, a user may click in a different area of the same polygon and retrieve a different set of crime information for a different area (generally referred to as 'neighbourhoods'). For a social scientist, this ability to link geospatial, census and crime data via an interactive map is powerful and allows the testing of hypotheses linking area characteristics and deprivation to criminal activity. COSMOS also enables layering of geolocated tweets over a map and allows users to read the tweet content, thus linking digitally generated content with crime and socio-economic characteristics and allowing us to test for links between what people are saying and the context in which they tweet. As an example, we may wish to investigate whether there is a relationship between hate speech on Twitter, hate crime, deprivation (via unemployment) and the ethnic composition of the area in which the tweet was made. Alternatively, we could search for tweets about illness, populate the map with census information on general health and look for a correlation between the two. Through the future integration of other curated and administrative data-sets we could look in detail at

health-related Twitter content and links to traffic patterns, pollution from heavy industries, weather and so on.

Social network analysis with respect to online interactions and behaviour is of significant importance for social science. One means of augmenting the analysis of the networked society is through the comparison and contrast of online and offline social networks to question whether online communications 'signal', even anticipate, offline social relations. An example of this is the investigation of urban governance, which conventionally has focused on key arenas of elite decision-making, be they elected legislatures and committees or governing 'regimes' of state and corporate elites and the pressures to which they are subject by organised campaigning groups [28]. Work also exists on the use of social network analysis in emergency response management, such as work by Bruns and Liang [6] and how communication breakdown in such events could lead to an impact on team functioning [27]. Capturing user-generated social media communications about issues of urban governance can augment this conventional study of elites and pressure groups by registering popular sentiment about governing regimes and their policy agendas throughout the course of their administrations and not just at the particular moments that elections are called or conventional opinion-polls are conducted. Figure 3 is visualising a network of over 17,000 tweets collected over a period of 3 weeks from the Cardiff area. The purpose of the network is to highlight prominent user accounts in the city. Each tweet is analysed to determine whether it contains any interaction with another Twitter user. This can include a direct mention or a retweet. Both can be visualised in this way to quantify the number of interactions between users – indicating level of influence in the network. It also helps to identify important connecting users – those with high Eigenvector centrality who provide a link between users of high levels of influence.

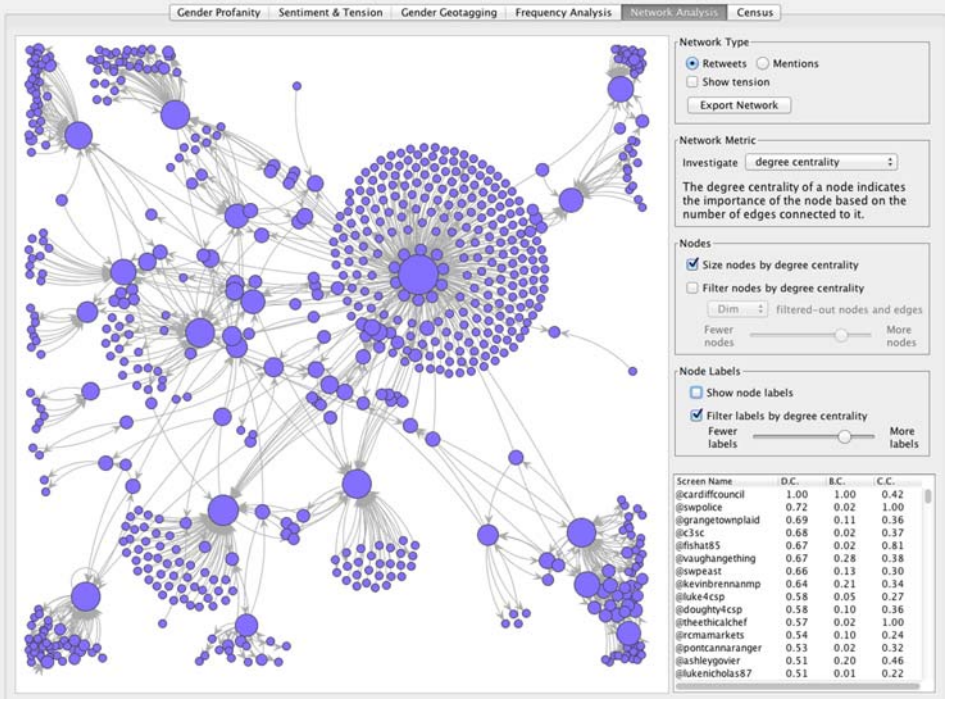

Figure 3. Social network analysis – retweet network visualisation.

This can be used, for example, to identify users who are highly active in a political discussion so that we can analyse their political position, or to identify users who provide a link between influential politicians from different parties.

2.3 Architecture

COSMOS enables the analysis of both real-time and archived data. To support real-time analysis, as illustrated in Figure 4, COSMOS has a single connection to the Twitter spritzer API stream and implements a multi-threaded connection handler to allow multiple analytical tools to receive and analyse the stream in parallel. The individual analysis tools enables processing of each incoming tweet, typically at a rate between 25 and 75 per second, depending on current Twitter activity. The reason for analysing tweets as they arrive from the stream is twofold: (i) it reduces the computational overhead and researcher wait-time when needing to conduct experiments on large numbers of tweets, as the individual tweets have already been analysed; (ii) it allows us to index the data-sets using derived tweet attributes. For example, we could index by tweeter gender or positive/negative sentiment. This may be useful if a study requires gender analytics or a further analysis of tweets expressing extremely negative sentiment. The results of this real-time

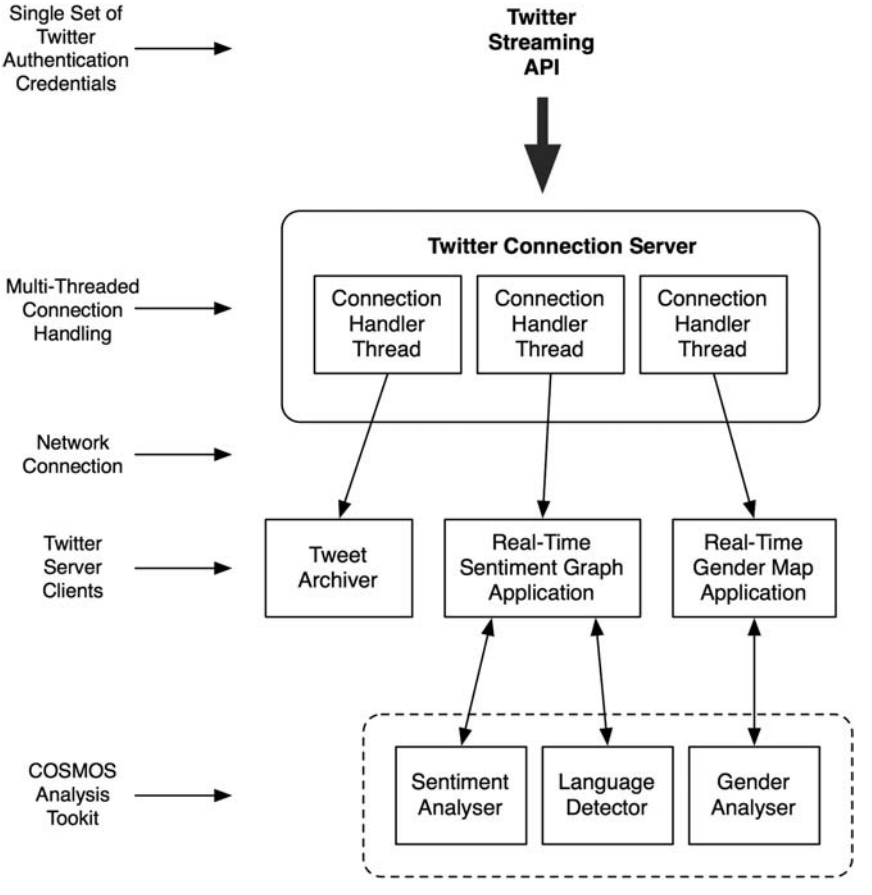

Figure 4. COSMOS streams and real-time analysis.

analysis are then visualised for the user, with the user interface being updated every second. Users can refine the stream using these results as filters. For example, a user could show only tweets posted in English, posted from within the UK, by male users, etc. During an evolving event, such as a political campaign, natural disaster or sporting event, COSMOS thus provides real-time monitoring as a service, with flexible refinement of the 1% Twitter stream. It has been reported that the largest event-based Twitter throughput was 15,358 tweets per second when Spain scored the winning goal against Italy in the European football championships in 2012. This throughput is many orders of magnitude higher than the expected norm and would require further analysis to stress test the real-time analytics.

In addition to real-time processing, each tweet received through the Twitter stream is archived. Along with the metadata provided by the Twitter API, the COSMOS-derived individual-level tweet attributes are stored as an attribute of the tweet to support rapid data filtering. Returning to the real-time example, a user could request only tweets posted in English, posted from within the UK, by male users, from the archive. This could also be refined by time. To retrieve these data without having completed prospective analysis would require the selection of all tweets within the selected time period, several types analysis (as outlined in Section 2.2) and subsequent selection of the relevant tweets. This not only slows down the information retrieval exercise but also is effectively a waste of processing time and power because the next user may make the exact same request. Storing the attributes as they are derived reduces this overhead. Furthermore, it enables the data collections to be indexed by any of these fields so that they can be optimised for particular types of searches. For example, indexing on Geospatial Location could reduce the search time for a geographically focused query. Beyond analysing the COSMOS archive of the spritzer, bespoke archives can also be analysed using COSMOS. For example, in February 2012, a record 13.7 million tweets were posted during Superbowl 46. This requires an underpinning scalable infrastructure to facilitate a large-scale on-demand analysis of tweets. We were, therefore, particularly interested in how analytical experiments in COSMOS can be conducted at scale. COSMOS aims to support researchers in performing empirical longitudinal studies (e.g. public response and reaction to political agenda, changes in legislation, views on national identity, etc.) and for this we need to process significantly larger volumes of archived data (for longer periods of time – ranging from 1 day to months or even years), for which terabytes of archived tweets need to be processed. This is generally not viable on a single machine; therefore, in order to handle this scenario we have developed a Hadoop-based implementation, which exploits the Map/Reduce paradigm. Map/Reduce supports parallel processing of large data-sets by splitting the data into '*n*' subsets and 'mapping' the data onto '*n*' Hadoop Worker Nodes, where '*n*' can be dynamically determined depending on the number of nodes in a cluster. Each node processes the data it is passed and the 'reduce' phase combines the resulting outputs. Figure 5 illustrates our deployment, where archived tweets are loaded into COSMOS as MongoDB collections where we assume they conform to the Twitter data schema as returned by the spritzer API. The user can define the parameters for their query, which leads to the selection of data from the MongoDB archive.

Hashtags are a user-defined convention that is associated with Twitter but is increasingly being used within other social networks and media. In terms of social media, the hashtag convention 'tags' tweets or similar content in terms of user defined 'topics' that in turn provide for enhanced searchability and discovery through the Twitter platform by other users. Consequently, the identification of hashtags via COSMOS, in relation to archived and real-time Twitter data streams, enables the identification of user-generated

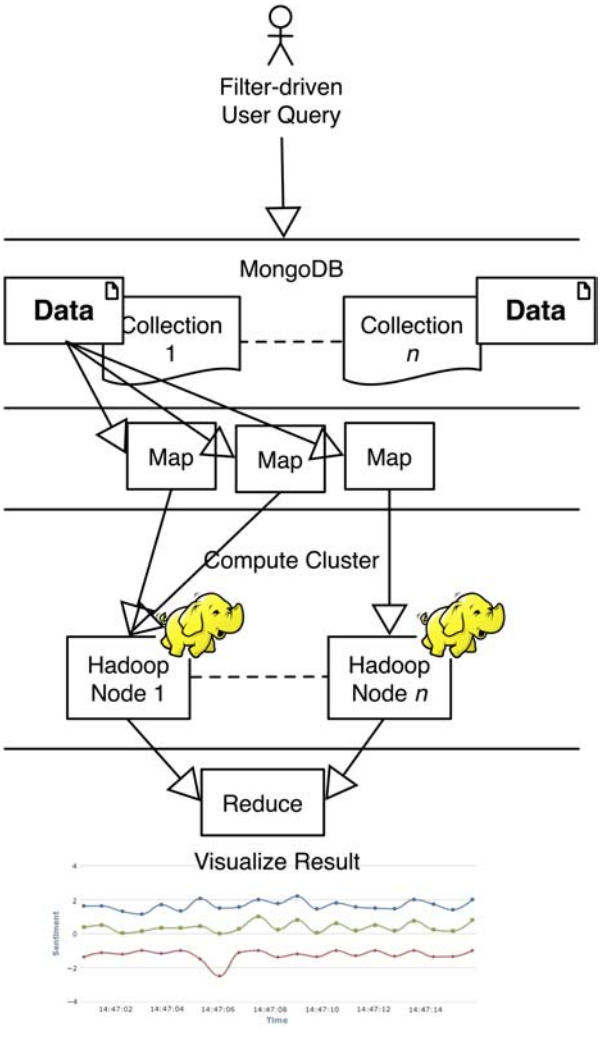

Figure 5. COSMOS: Hadoop and MongoDB architecture.

themes and topics. This, in turn, offers further semantic refinement and exploration, given that pre-identified or trending hashtags can be incorporated into search parameters that can be used to filter and collate Twitter data prior to further interrogation through the use of other tools and methods identified in this article. The fact that hashtags represent an identifiable signature of user-generated 'topic identification' means that the organisation and collation of data around hashtags provide for a robust and defensible strategy for assembling archives of interest especially when they are connected to clearly definable time lines and events, e.g. the olympics or civil incidents. As a form of data identification, they provide an opportunity to discover and explore semantic linkage that reflects the self-organisational work of users and key social institutions and agencies via social media. Hashtags can also inform strategies for topic detection.

The Hadoop deployment has been tested with the SentiStrength [30], the performance results of which are discussed in Section 3. The deployment is designed such that the executable is swappable when required; therefore, sentiment analysis could be replaced by

another type of analysis. The resulting improvement gained by deploying the analysis to Hadoop would be influenced by the execution time and performance of different tools. Hadoop generally requires the data to be stored in the Hadoop Distributed File Systems (HDFS) format, but also provides a MongoDB streaming interface that supports the splitting of MongoDB data-sets into a number of subsets so that they can be mapped across a number of Hadoop nodes. The streaming interface provides an end-to-end workflow that reduces the processing time required to analyse large data-sets on demand. To support further analytics beyond COSMOS, the gender, language, sentiment and tension scores, key-word/tweet frequency and geospatial distribution metrics are exportable (to a CSV format), so that further statistical analyses such as regression and time series models can be developed. Social network analysis metrics are exported in GEXF and GraphML format, for use with other tools, such as Gephi and NodeXLfor further investigation.

3. Performance results

COSMOS makes use of Hadoop to undertake analysis on the archived Twitter data. The Hadoop infrastructure is deployed over a virtualised platform, with one virtual machine (VM) hosting a single Hadoop worker process. Determining how such a virtualised Hadoop deployment should be supported across a Cloud environment is also challenging, as Cloud infrastructure parameters and virtual cluster configurations (i.e. number of VMs to support, how many VMs should be mapped to a physical machine taking account of computation and I/O constraints) can influence Hadoop performance and have an impact on resource usage. The lack of any behaviour models for achieving such resource management provides an opportunity to consider various optimisation techniques. Consequently, we focus on determining how data-intensive computation could be carried out over such an environment.

To provide performance metrics for the COSMOS Hadoop infrastructure, we used a Cloud infrastructure at the Universities of Cardiff and Castilla-La Mancha. The Cardiff University local infrastructure is composed of a cluster computer (Viglen ix4600) with one compute node and 2 Xeon e5620 CPUs (4 Cores + Hyperthread), 24 GB of main memory, and 4 TB of storage). It has CentOS 6.2 Linux. The University of Castilla-La Mancha (UCLM) local infrastructure (known as Vesuvius) is composed of 10 compute nodes, with 2 Xeon e5462 CPU (4 Cores), 32 GB of main memory and 60 GB of storage each. It also has a headnode, with the same configuration but with 1 TB of storage shared between all the compute nodes using NFS through a Gigabit Ethernet network. All of them have CentOS 6.2 Linux. For the experimentation, a test tweet archive of 15 million tweets was used to represent a longitudinal study of tweets posted on a particular topic. This is representative of the data produced in relation to a large event, such as the aforementioned Superbowl 46 corpus of 13.7 million tweets. To conduct a post-event analysis, we aim to produce results as fast as possible for social commentators and researchers to analyse. We used Hadoop to execute the sentiment analysis tool using the 15 million tweets as input. The results obtained from this experiment using the Cardiff Cloud (Figure 6, top) show an improvement on the performance when a number of Hadoop workers are compared with sequential execution. It is shown that although the four Hadoop configurations (1, 2, 4 and 8 workers) have the same amount of resources in common (but split between the total number of workers) within the same compute node, increasing the number of workers generally leads to an improvement in performance, bringing the analysis down from a $\sim$10 min to a $\sim$3 min. For social commentators, this provides a much faster response time to social events, and for social researchers, this provides a very significantly reduced wait

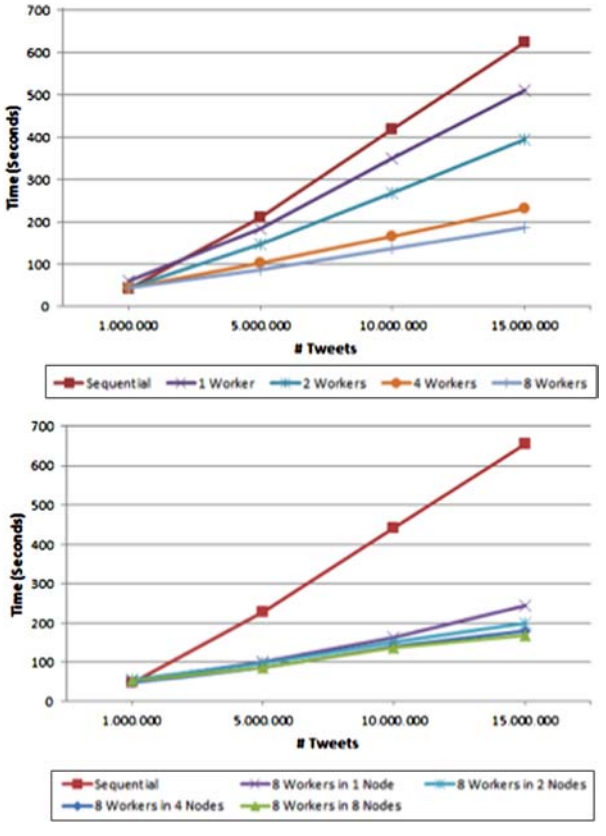

Figure 6. Hadoop Scalability (Cardiff top, UCLM bottom).

time between analyses. After observing the behaviour of the Cardiff Cloud with different worker configurations, we focused on the 8-worker configuration (as it gave better performance results at Cardiff) in order to experiment with the UCLM Cloud and evaluate whether this deployment configuration, over a distributed cluster, affects performance. In this scenario, the workers are distributed across the cluster and the results obtained are illustrated in Figure 6 (bottom). Workers are divided as 4 per node, 2 per node and then 1 per node – a node in this instance being a physical machine as opposed to a VM. Spreading the workers in this way does have an impact on the performance but not as much as with different worker configurations. The best results are obtained with one worker per compute node.

Comparing the results from both Cloud infrastructures it can also be observed that for the best scenario (8 workers deployed across 8 nodes), the behaviours are quite similar, independent of where they are deployed. However, comparing the performance with the 8-worker configuration deployed in the same compute node at UCLM Cloud with the result obtained from the same configuration at Cardiff Cloud, it can be stated that Cardiff Cloud achieves slightly better performance. This is due to its compute node being more powerful than a UCLM Cloud compute node. This performance penalty can be reduced by spreading all the 8 workers across 4 compute nodes (2 workers per compute node). It is, therefore, useful to note that in a realistic heterogeneous Hadoop cluster, one is likely to see variable performance based on the hardware of the node used. Performance, as in our

scenario, is also impacted by the type of Java VM being used to execute the sentiment analysis algorithm. Focusing specifically on the deployment strategy, it can be observed that results from the UCLM Cloud have shown a performance increase as the level of distribution of the VMs increases. This difference shows that when they are deployed within the same physical machine (even without CPU and main memory over-commitment), workers interfere with the operation of other workers, resulting in reduced performance. This fact means that running this application on a virtual cluster in a production Cloud environment may suffer perturbations from other VMs deployed within the same physical machine. This suggests that for best performance COSMOS should be deployed on a high-performance computing cluster of physical nodes, rather than virtualised across a number of cores.

4. User evaluation

The COSMOS platform is aimed at Social Scientist end users and as such the ease of use and intuitive understanding of the system are important to its success [19,33]. We have conducted initial system usability studies with the aim of informing a large-scale usability and user experience study in the near future. Three participants performed four tasks using the COSMOS software. The one male and two female participants were postgraduate students in the School of Social Science at Cardiff University. Further additional evaluation has been carried out through beta-testing of COSMOS. All participants represent the target user group of the COSMOS software. None of the participants had used or seen the COSMOS system before this evaluation session. The evaluation tested 'first-use by new users' of the archive search mode of the COSMOS software. The real-time mode or the longer-term productivity of COSMOS was not tested in these initial trials. The evaluation was divided into three parts: a group introduction, individual evaluations and a short group discussion. In the first part of the evaluation session, the three participants were given an introduction to the COSMOS software as a group. The presentation outlined the purpose of the software, the data-sets available for interrogation with the software and the methods of visualising and interrogating those data-sets with the software. In the second part of the evaluation session, each participant was asked to perform a set of four tasks individually. Each participant was asked to read an information and instruction sheet to ensure that each participant received the same information from the evaluator. Each participant was asked to use the talk-aloud protocol to explain their thought process as they completed the tasks. All three participants were particularly adept at talk-aloud and required little or no prompting to talk aloud throughout their session. The four tasks required the participants to interact with the COSMOS software in a variety of ways:

(1) Identify anomalies (by day) in the time line of tweets between 1st and 15th August 2012 using frequency analysis and the *TeamGB* keyword;
(2) Determine gender differences in sentiment around the time of any anomalies;
(3) Identify the geographic distribution of tweets during any anomalies;
(4) For the entire period, and without using keywords, identify users who are (i) mentioned frequently and (ii) retweeted frequently.

In the first task, two of the three participants identified the peaks in the Frequency Analysis tab as gold-medal wins for the British Olympic team. Both participants cited Mo Farah's win as an example. In the second task, two of the three participants identified as an anomaly a large dip in the sentiment lines in the charts of the Sentiment & Tension tab.

Although this dip was due to missing data rather than an Olympics-related event, this does demonstrate the potential for line charts to help users identify anomalies. All three participants used other parts of the COSMOS software – particularly the frequency analysis tab – to explore their own definitions of anomalous data. In the third task, all participants were able to use the map in the geotagging tab to view the geographic distribution of the search results as a whole. However, none of the participants investigated the geographic distribution of tweets around the time of an anomaly identified in the first task. One reason may be participants' unfamiliarity with the need to reformulate queries to search around the time of an event. To make this type of exploration easier, a direct link between selections in one visualisation and data in other visualisations is needed.

In the fourth task, participants used a different strategy than expected to investigate frequently retweeted and mentioned Twitter users. Although retweets and mentions are the basis of the networks provided by the network analysis tab, the lack of links between nodes in networks produced from the one per cent Twitter stream produced a network that participants found unhelpful for completing this task. Instead, all the participants explored the qualitative overview and frequency analysis tabs that provide the text of the tweets. This alternate strategy demonstrates the need for supporting multiple representations of the same data.

As a general comment, it is encouraging that all the participants were able to explore the software while performing the tasks and were able to follow alternative strategies to use the flexibility of the software to achieve data filtering, processing and visualisation tasks. All three participants had never used the COSMOS software before. As such, this initial evaluation tested the usability of COSMOS for first-time users, rather than the usability and productivity of the software for longer-term intermediate users. In the third part of the evaluation session, the participants discussed their likes and dislikes of the COSMOS software with the evaluator. The participants made many suggestions for improvements on both the functionality and aspects of the interfaces presented. These suggestions have been taken on board and changes made ahead of a second, larger scale usability study currently under way.

In the post-evaluation discussion, the participants highlighted two different approaches to the new software. One participant was comfortable enough with new software to learn first through exploration and then by reading a manual when required. The other participant preferred to read a few pages of instructional material before use. To support this latter style of new-user interaction, the COSMOS development team should provide a 'cheat sheet' that provides a brief introduction to each of the data visualisation styles, as well as the controls for selecting and refining data selections. The cheat sheet could be web-based as part of a wider set of COSMOS documentation.

5. Ethics

While Twitter Terms and Conditions allow for the non-commercial harvesting and archiving of data, analysts must also contend with legal and moral obligations. Principal concerns include anonymity, confidentiality, informed consent and the potential for harm. The analysis of social media necessitates an engagement with 'moral architecture' – the practice where moral and ethical procedures become inscribed into the workflow of digital data observatories [22]. The Association of Internet Researchers (AoIR) ethical guidelines emphasise three areas of concern: human subjects online, data and personhood and the public/private divide [10]. Online environments problematise the notion of the human subject. Does the automation of some online actions call into question the definition of

human subjects in social media research? Furthermore, social media may redefine the construction of 'person-hood' and the 'self', questioning the presence of the human subject in online interactions. Can we say an avatar is a person with a self? Is digital information an extension of a person? More fundamentally, are 'avatars' human subjects? In some cases, this may be clear-cut: emails, instant message chat and newsgroup posts are easily attributable to the persons that produced them. However, when dealing with aggregate information in 'Big Social Data' observatories, such as collective sentiment scores for sub-groups of Twitter users, the connection between the object of research and the person who produced it is more indistinct. Attribute data on very large groups of anonymised twitter users could be said to constitute non-personalised information, more removed from the human subjects that produced the interactions as compared to, say, an online interview. In these cases, the AoIR guidelines state 'it is possible to forget that there was ever a person somewhere in the process that could be directly or indirectly impacted by the research'. Anonymisation procedures for Big Social Data are in their infancy and researchers are yet fully cognizant of the factors that may result in 'deductive disclosure' of identity and subsequent potential harms [17,18]. A further concern is that people who use social media can perceive their interaction as private [20]. This can question the use of data aggregators that make accessible to the public, data on interactions that were intended for private consumption. The AoIR guidelines state that social, academic and regulatory delineations of the public–private divide may not hold in online contexts and as such 'privacy is a concept that must include a consideration of expectations and consensus' within context. At this stage 'ethical engineering' is in its infancy as far as social scientific analysis of social media is concerned. However, the use of proxies, data augmentation, archiving and harvesting need to be informed and developed within an emerging ethical context that is able to balance the digital public sphere with commercial interests, the privacy and protection of individual citizens and the requirements of critical and public social science.

6. Conclusion

In this article, we have presented an integrated set of social media analysis tools that can support the rapid marshalling of social media data-sets, where the outcome of one type of analysis can inform the researcher's choice of which tool to use next. This inductive methodology has been used in other domains such as healthcare, but is yet to be applied in the social sciences for social media analysis. User evaluation has proven encouraging and suggests that the tool is inherently usable in support of answering questions that could provide insight into online society.

Furthermore, an example of large-scale Twitter analysis using a sentiment analysis tool has provided results that indicate a scalable on-demand service when deployed in a virtualised cluster environment, though results also suggest that using a collection of physical nodes would reduce interference between virtual instances and improve performance. The results suggest that the time taken to process massive social media data-sets can be significantly reduced, which could enable the rapid and iterative analysis of such data to inductively interpret the data in near-real time as events occur and produce large amount of self-reported data that can be programmatically collected from APIs that provide publicly accessible interfaces. Finally, we have argued that such tools, when deployed in an ethical manner, could have a significant impact on the types of analysis available to academics with world leading expertise in social theory.

As COSMOS is routinely harvesting social media communications, it can depict changes in the structure and content of these networked communications in near (if not

real) time and over longer periods of months, the duration of electoral campaigns or terms of office for particular administrations. In turn, this analysis of online communications enables comparison with signals of offline behaviour such as conventional opinion-polling, actual election results, the progress of legislative agendas, responses to particular events, crises, scandals and so forth during a term of office. COSMOS can facilitate this comparison and contrast of the online and offline signals of urban governance where it can access administrative (e.g. election results, minutes of committee meetings) and curated (e.g. opinion polling, broadcast and print journalism) data-sets that are also digitised and streamed through an API. Finally, this analytical functionality provides COSMOS with the potential to re-orientate social scientific analysis of urban governance and other social phenomena around the effect of online communications in driving, not simply signalling, offline behaviour. In this regard, a major impact of COSMOS for social science is the revelation of how constitutive, not just representative, social media communications are of offline social organisation, change and identification including, in the context of urban governance, the advent of various 'information wars' as campaigning groups join state and corporate elites in more strategic use of social media to shape policy agendas and advance their interests.

Acknowledgements

This work was supported by the UK Joint Information Systems Committee through the project 'COSMOS: Supporting Empirical Social Scientific Research with a Virtual Research Environment'; the UK Economic and Social Research Council (ESRC) through grant ES/J009903/1; and the Spanish Government through grant TIN2012-38341-C04-04 and an FPI scholarship associated with grant TIN2009-14475-C04-03.

References

[1] J. Ahktar and S. Soria, *Sentiment Analysis: Facebook Status Messages*, Stanford University, 2009.

[2] S. Asur and B.A. Huberman, *Predicting the future with social media,* in *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01 (WI-IAT '10)*, Vol. 1. IEEE Computer Society, Washington, DC, USA, 2010, pp. 492–499.

[3] M. Bastian, S. Heymann, and M. Jacomy, *Gephi: an open source software for exploring and manipulating networks*, in *Proceedings of 3rd International AAAI Conference on Weblogs and Social Media (ICWSM), San Jose, CA*, 2009.

[4] J. Bollen, B. Goncalves, G. Ruan, and H. Mao, *Happiness is assortative in online social networks*, Artif. Life 17 (2011), pp. 237–251.

[5] J. Bollen, A. Pepe, and H. Mao, *Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena*, in *Fifth International AAAI Conference on Weblogs and Social Media (ICWSM), Barcelona*, 2009.

[6] A. Bruns and Y. Liang, *Tools and methods for capturing Twitter data during natural disasters*, First Monday 17 (2012).

[7] A. Bruns and S. Stieglitz, *Quantitative Approaches to Comparing Communication Patterns on Twitter*, J. Technol. Hum. Serv. 30 (2012), pp. 160–185.

[8] P. Burnap, N.J. Avis, and O.F. Rana, *Making sense of self-reported socially significant data using computational methods*, International J. Social. Res. Methodl. 16 (2013), pp. 215–230, 2013/05/01.

[9] P. Burnap, O. Rana, N. Avis, M. Williams, W. Housley, A. Edwards J. Morgan, and L. Sloan, *Detecting tension in online communities with computational twitter analysis,* Technol. Forecasting Social Change, 2013, Available at http://dx.doi.org/10.1016/j.techfore.2013.04.013, ISSN 0040-1625.

[10] C. Ess, *Ethical decision-making and Internet research: Recommendations from the AoIR Ethics Working Committee. Association of Internet Researchers (AoIR)*, 2002. Available at http://www.aoir.org/reports/ethics.pdf
[11] R. Jacob, M. Conover, M. Meiss, B. Gonçalves, A. Flammini, and F. Menczer, *Detecting and tracking political abuse in social media*, Presented at the International Conference on Weblogs and Social Media (ICWSM) 2011.
[12] N. Kourtellis, J. Finnis, P. Anderson, J. Blackburn, C. Borcea, and A. Iamnitchi, *Prometheus: User-controlled P2P social data management for socially-aware applications*, in *Middleware 2010*, Springer, Berlin, 2010, pp. 212–231.
[13] B. Liu, *Sentiment Analysis and Opinion Mining*, Morgan & Claypool, San Rafael, CA, 2012.
[14] J. Michael, *40000 Namen, Anredebestimmung anhand des Vornamens*. Available at: http://www.heise.de/ct/ftp/07/17/182/ 2007.
[15] F. Morstatter, J. Pfeffer, H. Liu, and K. Carley, *Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose*, CoRR abs/13065204 (2013).
[16] D. Murthy, *Digital ethnography: An examination of the use of new technologies for social research*, Sociology 42 (2008), pp. 837–855.
[17] A. Narayanan and V. Shmatikov, *Robust de-anonymization of large sparse datasets (How to break anonymity of the Netflix prize dataset.)*, in *IEEE Symposium on Security & Privacy*, IEEE, Oakland, CA, 2008, pp. 111–125.
[18] A. Narayanan and V. Shmatikov, *De-anonymizing social networks*, in *IEEE Symposium on Security & Privacy*, IEEE, Oakland, CA, 2009, pp. 177–183.
[19] J. Nielsen, *Usability Engineering*, Morgan Kaufmann, San Francisco, 1993.
[20] H. Nissenbaum, *Privacy in Context: Technology, Policy, and the Integrity of Social Life*, Stanford University Press, Stanford, 2010.
[21] A. Pak and P. Paroubek, *Twitter as a corpus for sentiment analysis and opinion mining*, in *Proceedings of the Seventh conference on International Language Resources and Evaluation LREC'10, Valletta, Malta, European Language Resources Association ELRA, May 2010*, Morgan Kaufmann, San Francisco, 2010.
[22] M. Parker, *Ethical and moral dimensions of e-research*, in *World Wide Research: Reshaping the Sciences and Humanities*, W.H. Dutton, I. Goldin and P.W. Jeffreys, eds., The MIT Press, Cambridge, MA, 2010, pp. 241–244.
[23] J.W. Pennebaker, M.E. Francis, and R.J. Booth, *Linguistic Inquiry and Word Count: LIWC*, Lawrence Erlbaum Associates, Mahwah, NJ, 2001, p. 2001.
[24] M. Savage and R. Burrows, *The coming crisis of empirical sociology*, Sociology 41 (2007), pp. 885–899.
[25] L. Sloan, J. Morgan, W. Housley, M.L. Williams, A. Edwards, P. Burnap, and O. Rana, *Knowing the tweeters: Deriving sociologically relevant demographics from twitter*, Sociological Res. Online, 18 (2013).
[26] M. Smith, N. Milic-Frayling, B. Shneiderman, E. Mendes Rodrigues, J. Leskovec, and C. Dunne, 2010. *NodeXL: A free and open network overview, discovery and exploration add-in for Excel 2007/2010*. Available from the Social Media Research Foundation at http://nodexl.codeplex.com/
[27] M. Stojmenovic and G. Lindgaard, *Social network analysis and communication in emergency response simulations*, J. Organ. Comput. Electron. Commerce. 2014. doi:10.1080/10919392.2014.896729
[28] G. Stoker, *Public-private partnerships and urban governance*, in *Partnerships in Urban Governance: European and American Experience*, J. Pierre, eds., Macmillan, London, pp. 34–51.
[29] M. Thelwall, K. Buckley, and G. Paltogou, *Sentiment in Twitter Events*, J. Am. Soc. Inf. Sci. Technol. 62 (2011), pp. 406–418.
[30] M. Thelwall, K. Buckley, G. Paltogou, D. Cai, and A. Kappas, *Sentiment strength detection in short informal text*, J. Am. Soc. Inf. Sci. Tech. 61 (2010), pp. 2544–2558.
[31] M. Thelwall, D. Wilkinson, and S. Uppal, *Data mining emotion in social network communication: Gender differences in MySpace*, J. Am. Soc. Inf. Sci. Tech. 61 (2010), pp. 190–199.
[32] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe, *Predicting elections with Twitter: What 140 characters reveal about political sentiment*, in *International AAAI Conference on Weblogs and Social Media*, Washington, DC, 2010, AAAI Press, Menlo Park, CA, pp. 178–185.

[33] C. Wharton, *The cognitive walkthrough method: A practitioner's guide*, in *Usability Inspection Methods*, John Wiley & Sons, NY, pp. 105–140.
[34] M Williams, A. Edwards, W. Housley, P. Burnap, O. Rana, N. Avis, J. Morgan, and L. Sloan., *Policing cyber-neighbourhoods: Tension monitoring and social media networks*, Policing Soc. 23 (2013), pp. 1–21.