Social Information Landscapes:
Automated Mapping of Large Multimodal, Longitudinal Social Networks

Eugene Ch’ng
School of Computer Science, Big Data and Visual Analytics Lab
University of Nottingham Ningbo China
Science and Engineering Building
199 Taikang East Road, 315100 Zhejiang Ningbo, China
eugene.chng@nottingham.edu.cn

Abstract

Purpose – This article presents a Big Data solution as a methodological approach to the automated collection, cleaning, collation and mapping of multimodal, longitudinal datasets from social media. The article constructs Social Information Landscapes.

Design/methodology/approach – The research presented here adopts a Big Data methodological approach for mapping user-generated contents in social media. The methodology and algorithms presented are generic, and can be applied to diverse types of social media or user-generated contents involving user interactions, such as within blogs, comments in product pages and other forms of media, so long as a formal data structure proposed here can be constructed.

Findings – The limited presentation of the sequential nature of content listings within social media and Web 2.0 pages, as viewed on Web browsers or on mobile devices, do not necessarily reveal nor make obvious an unknown nature of the medium; that every participant, from content producers, to consumers, to followers and subscribers, including the contents they produce or subscribed to, are intrinsically connected in a hidden but massive network. Such networks when mapped, could be quantitatively analysed using social network analysis (e.g., centralities), and the semantics and sentiments could equally reveal valuable information with appropriate analytics. Yet that which is difficult is the traditional approach of collecting, cleaning, collating and mapping such datasets into a sufficiently large sample of data that could yield important insights into the community structure and the directional, and polarity of interaction on diverse topics. This research solves this particular strand of problem.

Research limitations/implications – The automated mapping of extremely large networks involving hundreds of thousands to millions of nodes, over a long period of time could possibly assist in the proving or even disproving of theories. The goal of this article is to demonstrate the feasibility of using automated approaches for acquiring massive, connected datasets for academic inquiry in the social sciences.

Practical implications – The methods presented in this article, and the Big Data architecture presented here have great practical values to individuals and institutions which have low budgets. The software-hardward integrated architecture uses open source software, and the social information landscapes mapping algorithms are not difficult to implement.

Originality/value – The majority of research in the literatures uses traditional approach for collecting social networks data. The traditional approach is slow, tedious and does not yield a large enough sample for the data to be significant for analysis. Whilst traditional approach collects only a small percentage of data, the original methods presented could possibility collect entire datasets in social media due to its scalability and automated mapping techniques.

Keywords – Social network analysis, big data, mapping, online communities, social media, multimodal network, longitudinal network

Paper type Research Paper

1. Introduction

What are Social Information Landscapes (SIL) and how will it be useful for studying social networks? Here, Social Information Landscapes can be defined as "*the automated mapping of large topological networks from instantaneous contents, sentiments and users reconstructed from social media channels, events and user generated contents within blogs and websites, presented virtually as a graph that encompasses, within a timescale, all connections between followers, active users, comments and conversations within a social rather than a physical space*". The key phrase 'automated mapping' is essential here as the mapping of very large networks is essential as network behaviour may be different in massive networks in relation to their emergent behaviours and the way they self-organise. Larger networks are also ideal subjects for studying geodesic distance, centrality and density. SIL was first mentioned in an article dealing with a Big Data funnelling approach [1], and subsequently in corresponding articles on the study of online community formation and decline when research results were published [2,3].

The onset of the Internet age has made our world smaller, as offline communities connect to and as virtual communities, space and time are, in a sense compressed to within the social medium that facilitates community interactions. Insights into the behaviour of virtual communities in the age of social media requires a Big Data approach in mapping the interactions as social networks, for the interactions of online communities for a single topic can occur over large geographical distances and may span many months involving thousands to millions of participants from highly diverse demographics. Such interactions may also be multimodal, involving not only the users, but also the content of the interactions linked between the two parties. As such, the traditional approach of manually mapping such networks can be tedious and would not yield a large enough sample of the subject for social network analysis, due to the fact that the collection, collation and pre-processing of datasets are frequently too large to manage with human resource alone. This article presents a Big Data solution as a methodological approach to the automated collection, cleaning, collation and mapping of multimodal, longitudinal datasets from social media. The methods and algorithms presented here are generic, and can be applied to diverse types of social media or user-generated contents involving user interactions, such as within blogs, comments in product pages and other forms of media, so long as a formal data structure can be constructed.

The mapping of extremely large networks involving hundreds of thousands to millions of nodes could possibly prove or disprove theories in the social sciences. The goal of this article is to demonstrate the feasibility of using automated approaches for acquiring massive, connected datasets for academic inquiry in the social sciences.

The article begins with a background of the of the research work, it then continues with the method for mapping social networks, covering the need for a Big Data architecture, identifying suitable asynchronous and distributable open source technologies that are scalable in terms of volume of data and velocity of incoming data streams. The latter section of the methodology covers the focus of this article – the mapping of Social Information Landscapes. The section describes the data structures and algorithms for mapping, reconfiguration and storage of dynamic and static networks. Section 4 demonstrates the feasibility of the methods in Section 3. Finally, the article concludes with a discussion and future work.

2. Background

The limited presentation of the sequential nature of content listings within social media, as viewed on Web browsers or on mobile devices, do not necessarily reveal nor make obvious an unknown nature of the medium; that every participant, from content producers, to consumers, to followers and subscribers, including the contents they produce or subscribed to, are intrinsically connected in a hidden but massive network. Such networks when mapped, could be quantitatively analysed using methods in social network analysis (e.g., centralities), and the semantics and sentiments could equally reveal valuable information with appropriate analytics. Yet that which is difficult is the tediousness of collecting, cleaning, collating and mapping such datasets into a data structure that could yield important insights into the community structure and the directional, and polarity of interaction on diverse topics. It seems that the majority of research that investigate social networks have not looked beyond the traditional approach of manually collecting and mapping such networks, to a Big Data approach where mapping processes are automated. The research community that invariably focused their attention on social networks should genuinely consider the fact that any online activities is a sample of a population, and that manual data collection, due to its tediousness, will perhaps harvest a tiny sample and, as a consequence, yield a minor percentage of the full online content – a sample of a sample of a population. Unfortunately, such datasets will invariably be biased. A Big Data approach using scalable technologies is a viable approach in the era of Big Data, using machines to collect machine-collated contents.

Massive quantities of information generated by people are being tapped by diverse groups within the domains of computer science, physics, mathematics, social sciences, business and economics with the hope that it will answer questions in their disciplines. In the physics domain for example, the Large Hadron Collider at CERN churns out vast quantities of information. Within computer science and the engineering disciplines, the development of connected devices defined as the Internet of Things (IoT) generates massive information about human activities, systems and environments from sensors and video surveillance. Computational social science on the other hand offers 'the capacity to collect and analyse data with an unprecedented breadth and depth and scale' [4]. Regardless of which disciplines, the fact is that 'these massive amount of information can be tracked and measured with unprecedented fidelity' [5]. Anderson continues, "This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves." Whilst Anderson's statement appeared to be controversial to many, there is truth in the fact that taking a Big Data approach could have significant contribution to a new instrument for the vehicle of knowledge – 'change the instrument, and you will change the entire social theory that goes with them' (Latour, 2010. p.9).

Data can be 'Big' in different ways, and is observed by Manovich [7] as data sets that are sufficiently large to require supercomputers. It is loosely defined by Jacobs as "data whose size forces us to look beyond the tried-and-true methods that are prevalent at that time" [8]. However, Big Data is not only characterised by its size, but by its relationality with other data [9] – "Big Data is fundamentally networked". Big Data can be characterised by three Vs – *Volume*, *Variety*, and *Velocity*. These Vs, together with four others are briefly cover here, and how they are related to the context of this paper – relationality and the networked nature of data.

Volume refers to the enormity of data that have been created from the past to the future in ever increasing rate. *Velocity* refers to the speed of data that is being created, stored and analysed. *Variety* refers to the number of types of raw data that is scattered, unstructured and complex. The 'Big' in the term 'Big Data' describes the three Vs, for all can potentially be 'Big' to an extent that the datasets will render a standard computer insufficient for storing, processing and visualising incoming data in real-time (or near real-time), without which data becomes meaningless as the big picture will not be known. Four recent Vs have been proposed as part of the goal to define the properties of Big Data. The first is *Veracity*, which refers to the credibility, or accuracy of the data that a research may acquire or process, such as false information that are spread widely online via social media costing huge loss of resources [10,11]. *Variability* refers to the meaning in which a piece of information may be associated with, and how that meaning may differ in different context or at different times. The third is *Visualisation*, which extracts patterns in data in a consumable format to extract meanings and inform data consumers. The previous Vs may eventually add *Value* to institutions and organisations within the digital economy when raw data is transformed via the Big Data process into information that facilitates decision-making. The permutations of a set of research problems associated with Big Data's *Volume*, *Variety*, *Velocity*, *Veracity*, *Variability*, *Visualisation* and *Value* are manifold, and the mapping of Social Information Landscapes is one such area. Mapping SIL potentially faces challenging issues related to all these terms during the process of social network analysis. The present article concentrates on volume and velocity, which pave the way for resolving the other terms.

Social Information Landscapes of a similar nature may be obtained, but they are perhaps a limited subset of the definition. For example, Java *et al.* [12] investigated a very small community network based on key terms within a gaming circle who also share daily experiences with each other. Jansen *et al.* [13] examine Starbucks' customer networks with Starbucks as a central node based on the frequency of tweets in "word of mouth" communication. Huberman *et al.* [14] reveals a hidden network of actual friends within a subset network via the removal of followers who were not part of an actual friendship. Sakaki *et al.* [15] compared information diffusion in three Twitter network types: earthquake, typhoon and new Nintendo game news. A larger study by Kwak *et al.* [16] constructs retweets trees of 'Air France Flights' and examined their temporal and spatial characteristics for retweets pattern, depth, and speed. The mapping and analysis of large scale SIL may indeed provide deeper information about a social network than conventional statistical analysis or previous studies of partial networks will allow.

The hidden social information within social media channels is important for revealing insights on the spatial organisation of the social network. This requires the mapping of the "social space". Bourdieu [17] states that the social world, a 'social topology', can be represented as a space with several dimensions constructed on the basis of principles of differentiation or distribution constituted by the set of properties active within the social universe capable of conferring strength and power within the sphere of that universe. Agents and groups, according to Bourdieu, are defined by their relative positions within space as a set of objective power relations, i.e., field of forces as economic, cultural, social capital, or symbolic capital (prestige, reputation, etc.) that affect other agents who enter that space. Social networks are similar in nature. Whilst one may suggest that physical interactions are irreplaceable, the benefits that computer-based social networking has provided society cannot be ignored. Social networks increase our range of human connectedness beyond the boundary of a participant's geographical location. With regards to the time of interaction, communications sent now may be retrieved and responded to, much later by other participants. This invariably opens up a broad range of opportunities as space and time, in the eye of a user are 'compressed' to within a digital display, allowing diverse communications from large demographics groups. In this context, Social Information Landscapes are an important aspect for social science inquiries using social network analysis.

The next section provides a technical foundation for the construction of SILs.

3. Methods for Mapping Social Information Landscapes

Two basic types of social network analysis exist. The first of which is ego network analysis, the second is complete network analysis. Ego network analysis is conducted via surveys where the 'ego' responds to questions asked about the ego's immediate network. Each ego is sampled from a large population in order to assess the person's network size, diversity, and etc. The second type of analysis views the complete networks, such as all existing relationships between a set of nodes. The research here is posited in the latter type, although data from large-scale surveys used for collecting ego networks can also be automatically mapped.

There are two types of social networks which can be mapped:

1. Follower – Followee or "Content Provider" – "Content Consumer" network. This network consists mainly of inactive nodes with a more or less permanent connection between them. Followers such as within Twitter, LinkedIn, or subscribers, and in YouTube and Instagram for example, are within this type of network. The growth of these types of network is slower. Such networks have a larger temporal distance between follower and followee.
2. Activity networks consist of active participant interactions, which have a shorter temporal distance between each link. The growth of the network size can be rapid, particularly with viral contents. Examples of activity networks are active conversations within Twitter, Facebook, under the comments section of YouTube, Google+, and product pages or blogs.

Both network types are possible with the method described here. Section 3.1 begins the methodology by suggesting appropriate technology for such work that makes it possible to deal with the volume and velocity of data in the Big Data context.

3.1 Accessible Big Data Software-Hardware Integrated Architecture

This section introduces and suggests accessible technologies for Big Data scaling issues but do not delve into the details of their implementation. Implementation details are available within the technical manuals of these technologies and a summary guide in the same context is available [1]. Two requirements are set prior to mapping social networks. The first of which is that such a system must be accessible to a wide range of individuals and institutions with low budget constraints. The second requirement must ensure that all technologies used are scalable in terms of volume and velocity. When volume and velocity can be managed, data integrity and completeness can be assured.

Both requirements are easily achievable for individuals and institutions, fortunately. Open source scalable technologies, packaged as Application Programming Interfaces (API) and software libraries are available and are accessible for the development and deployment to inexpensive hardware such as commodity machines. Alternately they may be deployed to servers or Cloud services. Figure 1 is a summary diagram of the software architecture deployed on the author's institutional Linux platform (and works with Windows and Mac OSes). In (1), unstructured data enters through the server "edge nodes" via Node.js, various algorithms work together to clean and structure data into schemas of key-value pairs. Data either enters through Redis (if real-time in-memory processing is needed), or directly into MongoDB, distributed across MongoDB Shards. The asynchronous nature of the "edge nodes" using Node.js ensures that data streams that scale in volume or velocity (viral contents) are all received. The software integration is similar across developmental servers, Cloud services, or within in-house servers or commodity machines. Commodity machines (low-budget clusters of PCs) are the most inexpensive way for deploying the system – Each Node.js application uses only a single CPU core, thus the suitability of using cheap hardware. In (2), datasets are queried through either Redis, or directly from MongoDB, via conventional algorithms that do not need high-efficiency algorithms as the velocity of data streams can now be controlled. Mapping algorithms can be programmed in either Node.js, or using any other languages that support connections with MongoDB or Redis.

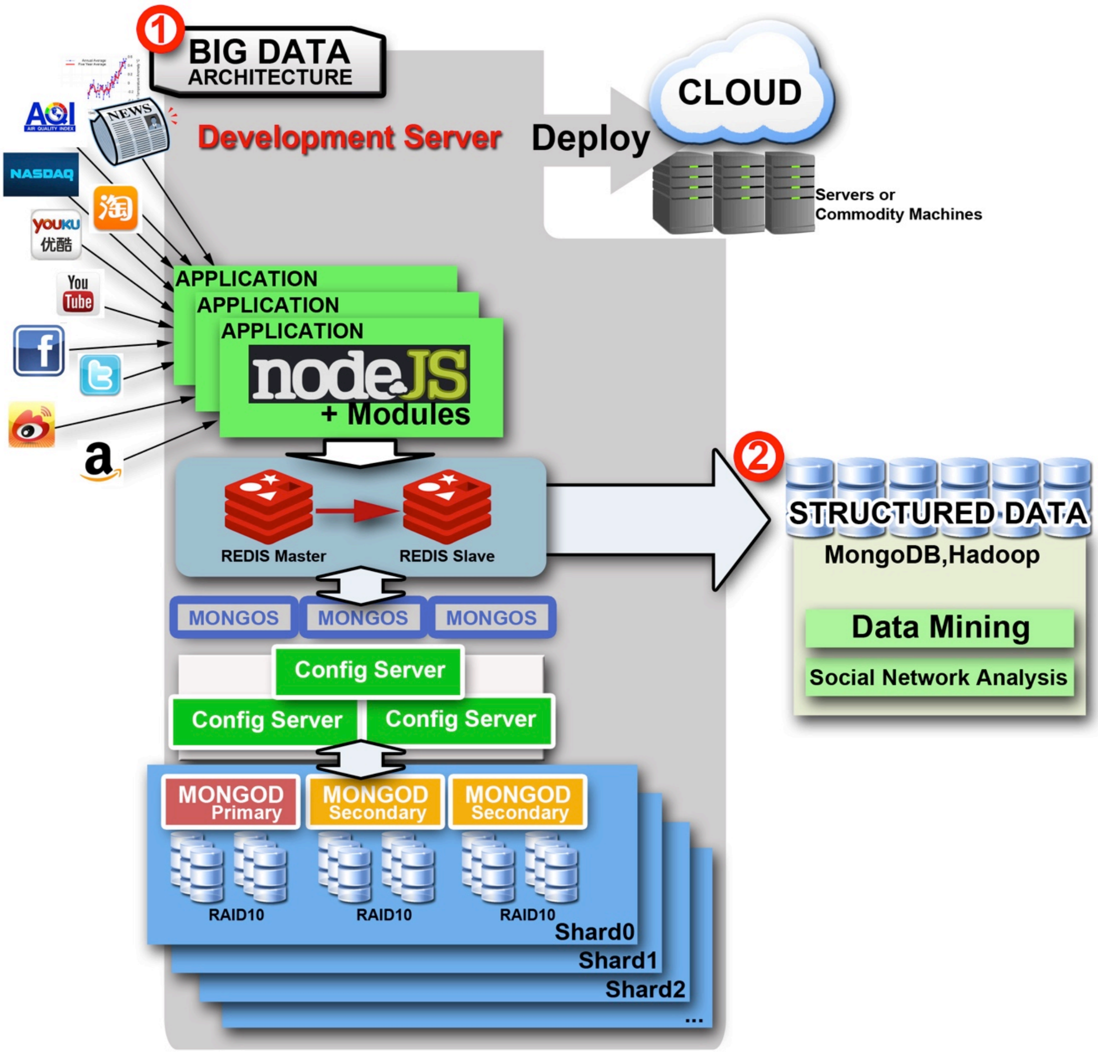


Figure 1. A summary diagram illustrating the Big Data software-hardware integrated architecture. 1) A Bid Data architecture integrating various open source scalable technologies. 2) Structured data used for research. In the context of this paper, social network analysis, and data (graph) mining

Secondary data types are important to SIL. The availability of secondary data types is dependent on the social media API and the target Web page. These statistical data are usually available (friends count, # of subscribers, # of followers, # of posts, etc.) and could be used to compare with the growth and evolution of a network over time. Geo-location, time zones, language, gender and textual contents provide other means of informing the status of each node in relation to the SIL. The research questions asked and the hypotheses tested within each research inquiry are different, and therefore, the necessity for collection of secondary data types depends on the project at hand.

3.1.1 Accessing Social Media Data and Web Crawling

The volume and velocity of incoming data associated with mapping SIL may involve accessing social media Application Programming Interfaces (API) such as Twitter, Facebook, Weibo, QQ, YouTube or Youku, and online stores such as Amazon, Taobao, for example. It is expected that certain API may limit the availability of their repositories due to internal financial motive of the company at stake. Data after all is the new oil and has great value to business profitability. Where API limit data access, Web-crawling and scraping algorithms can be used to extract data from JavaScript's Document Object Model (DOM) from websites, which any graduate programmer from a reputable institution would have learned. This article will not cover the details of accessing social media data via API, nor will it cover Web-scraping technologies, which individuals or institutional programmers will be able to manage. These topics are beyond the scope of this paper, which focuses on methods for mapping SIL.

It is imperative that we use social media data for testing the methods presented here. The choice of social media should generate sufficiently large volumes of data with potentially high velocity. One of these types of social media is Twitter. Twitter generates around 500 million tweets per day. In any topical viral hash tags, the tweets in Twitter could easily reach tens of thousands of posts in a short span of time. This meets the first two categorical Vs of Big Data – *Volume* and *Velocity*. Twitter data will be used for testing the methods covered in this paper, using the *streaming API* which capture more complete datasets than the limited REST API for polling endpoints.

3.1.2 Asynchronous I/O

The volume of data is expected to be in the orders of magnitude, the velocity will have a higher rate, considering that hundreds of millions of contents are generated every second. Such volume of data multiplied by the velocity of incoming streams from the Web will definitely crash conventional server-side applications. A scalable network layer that is event-driven, with a non-blocking I/O model that is capable of data-intensive application with the ability to push and pull data from multiple platforms is needed. Node.js built on Google's V8 JavaScript runtime, an open source JavaScript engine developed for Google Chrome's Web browser is used. V8 compiles to native code prior to execution, optimised and re-optimised dynamically at runtime and is built for browser-based and stand-alone high-performance applications. Node.js supports clustering but a PC with a single core is generally sufficient for supporting large input streams. Node.js is purely a server-side application and therefore, integration algorithms will need to be written for various API and Web Services. A suite of basic libraries and modules accompany Node.js (HTTP, Sockets.io, etc., amongst which are modules that connects to social media). The asynchronous nature of Node.js algorithms makes it capable of pulling large volumes of data in increasingly high velocity on a server side application [1].

3.1.3 Scalable Data Storage

The big truth about Big Data is that it is easier to get the data in than out (Jacobs, 2009, p.4) from Relational Database Management Systems (RDBMS). The storage of Big Data in the Terabytes require a format that can be stored, but accessed quickly and processed on the fly in real-time or near real-time. Raw data from websites and social media will be inconsistent and unstructured. As such, indexing may be difficult. As noted by Jacobs "To achieve acceptable performance for highly order-dependent queries on truly large data, one must be willing to consider abandoning the purely relational database model".

A NoSQL (key-value pairs) with non-relational features that scales massively and horizontally with a number of solutions that makes it easy to shard across distributed systems is the key to storing structured social media data. There are benefits in using NoSQL databases of scientific data storage. One of the most important aspects of statistical correlation is the storage of as much data as a project can possible obtain. Storing all data for filtering later is a better approach than not having the full amount of data, for data lost means opportunity lost forever. This is important as websites conduct routine archive and removal of old data. Social media and social network API limit access to past data (for example, Twitter restricts access to stored data and only releases 10% of their API search). Streaming real-time data from API however is unrestricted and therefore is mandatory, which means that all incoming data streams have to be stored.

MongoDB, a cross-platform open source freeform document database that stores document in dynamic schemas as JavaScript Object Notation abbreviated JSON (MongoDB terms it BSON). JSON is a lightweight format that can be used for structuring transmitting data between a server and Web application and is convenient as a data format for Node.js' JavaScript. In this project, the replication of MongoDB is used for increasing the distributed storage and throughput of data. All incoming data are parsed, categorised and stored in MongoDB shards via Node.js.

Redis is another open source, networked, in-memory, advanced key-value store data structure server for fast access of structured data. This is possible as Redis holds the whole dataset in memory with configuration to allow semi-persistence of data via snapshotting. Redis can be positioned as an intermediary between MongoDB and other necessary services (e.g., Hadoop) within the Big Data architecture, such as the processing of datasets, data mining, data and social networks, mapping, and visualisation. Both MongoDB and Redis are scalable and supports clustering.

3.2 Mapping Social Networks

Social media contains hidden networks and can be constructed as SIL. The components of an SIL consist of users, contents and links that made up the landscape. The subsequent sub-section covers the methods and algorithms for mapping, reconfiguring and storing SIL.

3.2.1 Mapping Concepts

Figure 2 presents three models of extremely small multimodal networks, which can be used to compare the differences between various node communication scenarios. The Twitter social network is used as an example here for reasons stated earlier, and for a clearer understanding of the data to be presented in Section 4, but any multimodal network can be mapped in a similar fashion. Visualisation uses the Gephi graph visualisation package, but any network visualisation application can be used. Gephi's layout feature (ForceAtlas and ForceAtlas 2) and ranking for setting up the colour scheme. The average degree and average path length is used for calculating the statistics of the network, these are degree and betweenness centralities.

In Figure 2, on the left is a larger node with many retweets in-degrees (directional edges) pointing towards node 'TwitterC'. On the left of TwitterC are four protruding nodes, which were deliberately 'pulled out' of the cluster for visual clarity. They are Twitter users who retweeted TwitterC's tweet. There is a thick arrow coming out of each of the Twitter user node, with another arrow pointing towards the retweet node (the post) from it, which points back towards TwitterC. All retweets have a similar pattern. This records both the Twitter user and the retweets as a multimodal connection.

At the bottom right of the figure is a small cluster consisting of 5 interactions, with TwitterE tweeting to TwitterF, G and H. Twitter G tweets a post to TwitterF, TwitterH and TwitterG.

The cluster of nodes with two larger nodes TwitterA and TwitterB at the top right of the figure is a conversational network, typical of an online activity network. Such a configuration may reveal a community of some sort [2]. At the bottom of the community cluster is a group of nodes with directed edges from TwitterB showing conversations "talks to @TwitterA with text" directed towards TwitterA (directed edges from the nodes), these group of nodes are responsible for the thick arrow pointing towards TwitterA. At the top right of the cluster, TwitterA tweeted a node with a double directional edge with the dummy text "A text not targeting anybody". This is a tweet (or a comment) without a target. The amount of the interaction defines the weight of the directed edge. The edge between TwitterB and TwitterD is not as thick but is thicker than the other edges, showing more interactions between the two nodes. Other edges and nodes show the direction of conversation, which gives it a distinct visual pattern as compared to the retweets. The unique arrangements of the nodes give rise to distinct features within a large network, as we shall see in Section 4. Activity nodes such as the clusters show here may allow us to identify and isolate communities from small and large-scale networks.

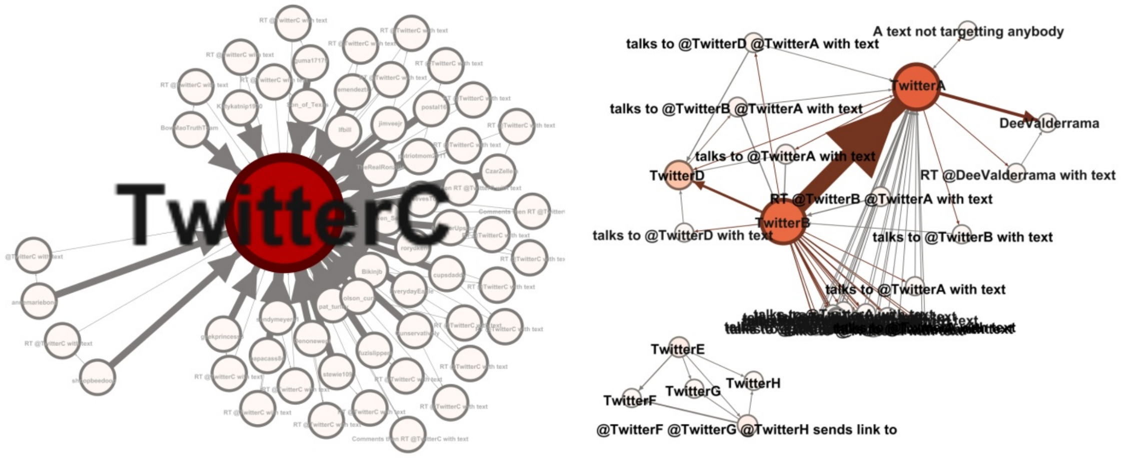

TwitterC
RT @TwitterC with text
Comments on RT @TwitterC with text
talks to @TwitterD @TwitterA with text
A text not targetting anybody
TwitterA
talks to @TwitterB @TwitterA with text
DeeValderrama
talks to @TwitterA with text
TwitterD
RT @DeeValderrama with text
RT @TwitterB @TwitterA with text
TwitterB
talks to @TwitterD with text
talks to @TwitterB with text
talks to @TwitterA with text
TwitterE
TwitterG
TwitterH
TwitterF
@TwitterF @TwitterG @TwitterH sends link to

Figure 2. A representation of a small section of three Social Information Landscapes. The unique configuration of nodes and edges make it possible to isolate activities in social media. The visualisation uses the Gephi graph visualisation software.

3.2.2 Data Structures for Dynamic and Static Graphs

A standard data format GEXF (Graph Exchange XML Format) can be used for representing social networks. GEXF is a language for describing complex network structure and their associated data and dynamics. Whilst data can be output to any of the network-based formats (GraphML, XGMML, SVG, GDF), GEXF preceded the other formats as it is relatively mature, robust, flexible and includes time as one of the dynamics feature. The GEXF format provides a way to visualise time-based networks that other formats do not. The algorithms within the mapping system should consist of *nodes* representing users and posts (tweets, comments, posts, etc.), *edges* representing links between users, and between users and their posts. In multimodal networks, posts are included as nodes so that the expression of the landscape has distinctive features when visualised using a combination of pre-set algorithms in some visualisation software (Gephi for example). All informational entities should be recorded with a timestamp so that the evolution of occurrence and growth of the network can be reconstructed in a dynamical way.

The data structure of a typical GEXF file is in the listings below, with a graphical representation in Figure 5.6. In the figure, *N* refers to a node (a person) with its corresponding number whilst *P* refers to a post (a tweet, a comment, etc.).

Listing 1 shows the top and tail of the data structure of a typical GEXF file, with a graphical representation of the network. Between the `<nodes></nodes>` and `<edges></edges>` tags are all the nodes and edges of the graph, shown in Figures 3 and 4.

```
<?xml version='1.0' encoding='UTF-8'?>
<gexf xmlns='http://www.gexf.net/1.2draft' xmlns:viz='http://www.gexf.net/1.2draft/viz'
version='1.2'>
<meta lastmodifieddate='113-6-4'>
        <creator>Eugene Chng</creator>
        <description>Twitter Map</description>
</meta>
<graph defaultedgetype='directed' idtype='string' mode='dynamic' timeformat='dateTime'
start='2013-05-18T05:00:58.975+0100' end='2013-05-18T06:58:25.584+0100'>
        <attributes class='node' mode='dynamic'>
                <attribute id='description' title='Description' type='string'/>
        </attributes>
        <attributes class='edge' mode='dynamic'>
                <attribute id='weight' title='Weight' type='float'/>
        </attributes>
<nodes count='12'>
        ... // all nodes here
</nodes>
<edges count='16'>
</edges>
        ... // all edges here
</graph>
</gexf>
```

Listing 1. The top and tail of the data structure of a typical GEXF file, with a graphical representation of the network. Between the `<nodes></nodes>` and `<edges></edges>` tags are all the nodes and edges of the graph, shown in Figure 3 and 4.

Particular attention should be paid to a few lines of code and the attributes of the tags in Listing 1 and Figures 3 and 4. Specifically, due to the nature of the dynamic graph, the element of time (start and end attributes) is important. In the `<graph>` tag (bold), the start and end attributes shows the entire timespan of the network. The other tags are the `<nodes>` and `<edges>` tags, which shows the size of the nodes (Figure 3) and edges (Figure 4) in the network.

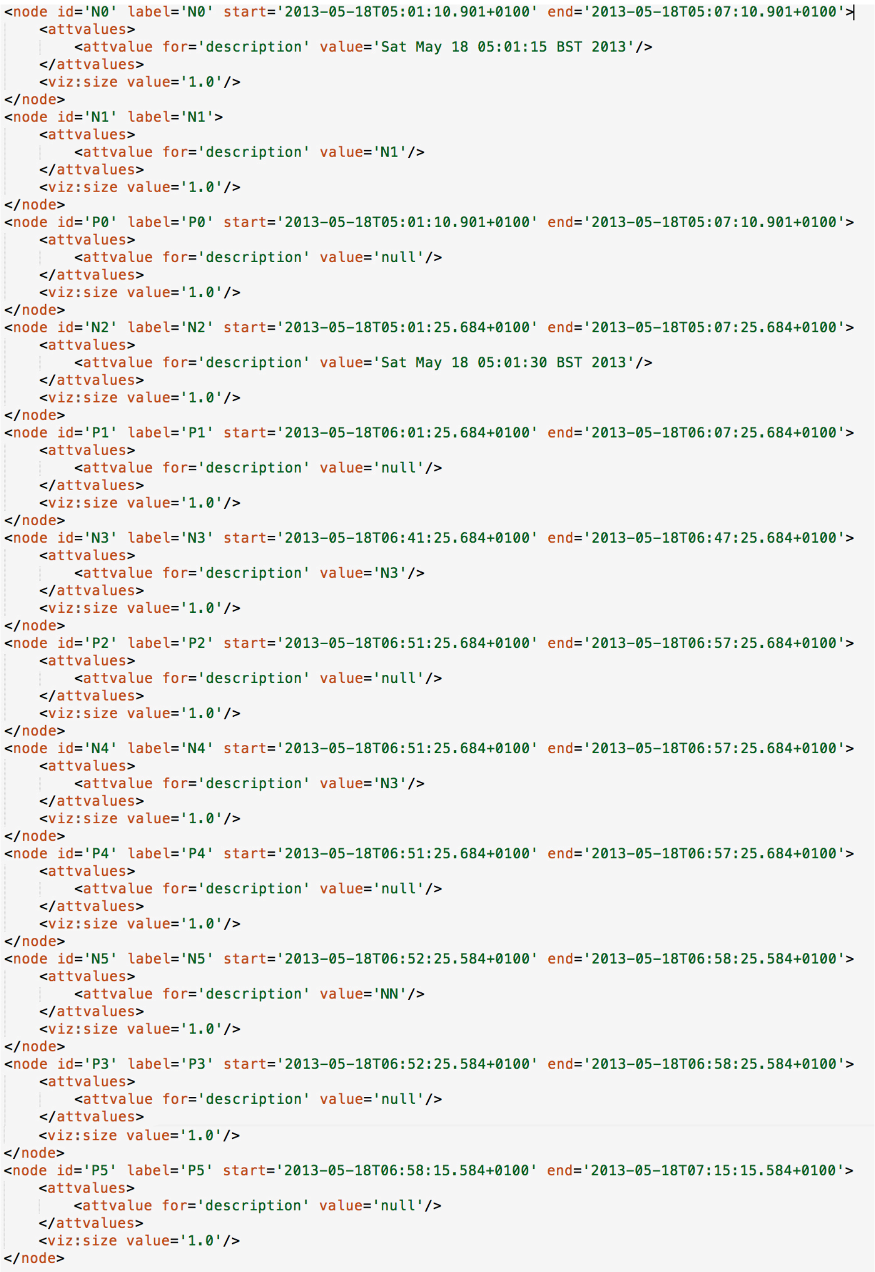
Figure 3 shows a listing of all the nodes of the network. The most important attributes in the `<node>` tag are the id and the label. The id sets the uniqueness of the id for graph processing software and the label is displayed visually as a label on the nodes (see Figure 5).

Figure 4 shows a listing of all the edges of the network. Three most important attributes are the id, the source and the target, which describes the connection between the source and the target node in Figure 3. The id "N4-N1" of an edge has two `<attvalue>` tags showing a value of 2:

```
<attvalue for='weight' value='2' />
```

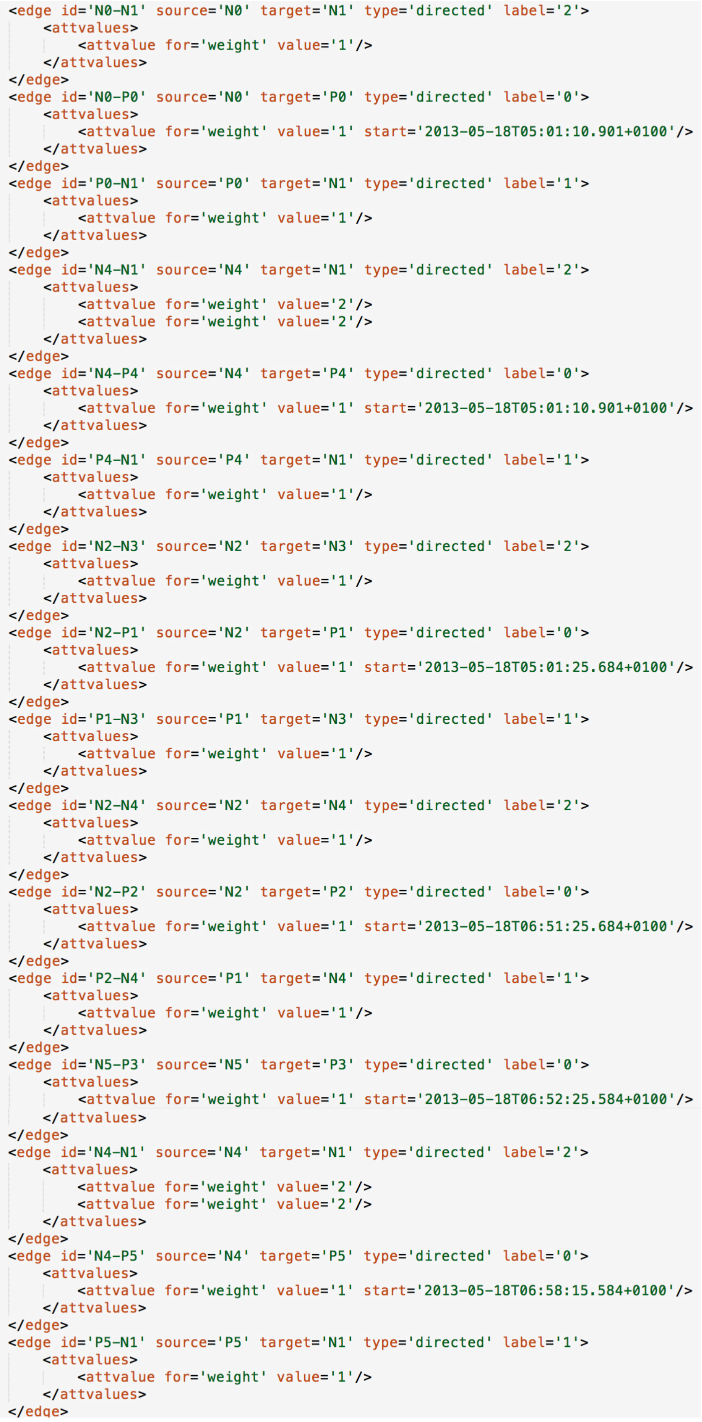
This is read by graph visualisation software (i.e., Gephi) as the weight of the connection. *N4*'s connection to *N1* in Figure 5 has a thicker line because there were two interactions:

```
N4 -> P4 -> N1
N4 -> P5 -> N1
```



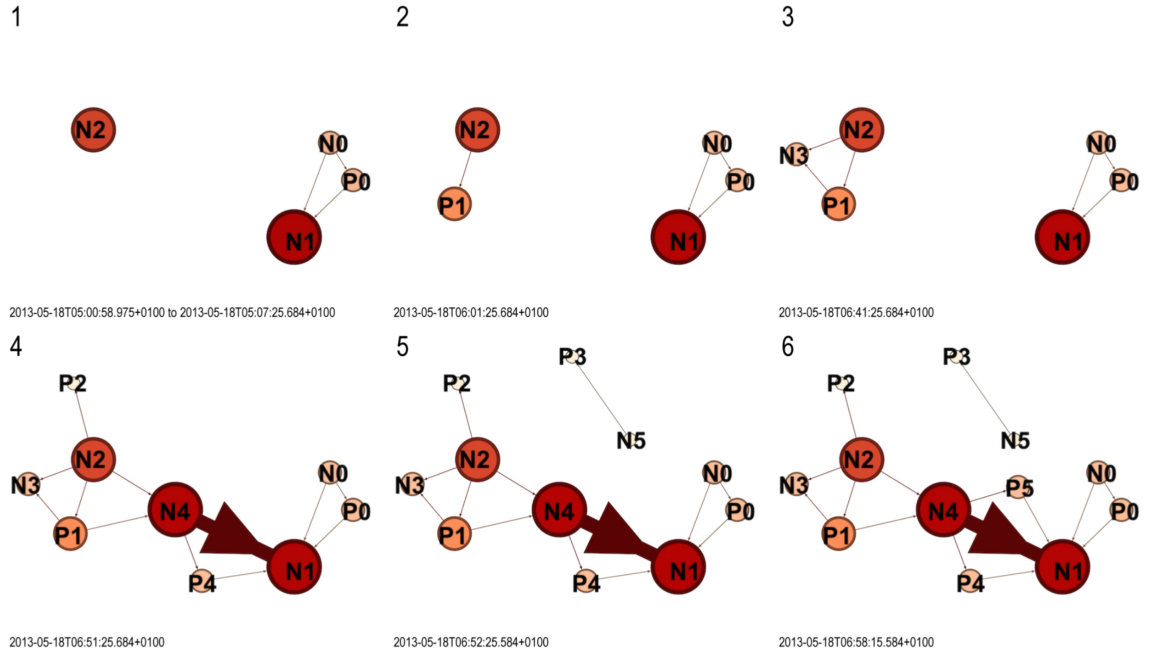
```
<node id='N0' label='N0' start='2013-05-18T05:01:10.901+0100' end='2013-05-18T05:07:10.901+0100'>
    <attvalues>
        <attvalue for='description' value='Sat May 18 05:01:15 BST 2013'/>
    </attvalues>
    <viz:size value='1.0'/>
</node>
<node id='N1' label='N1'>
    <attvalues>
        <attvalue for='description' value='N1'/>
    </attvalues>
    <viz:size value='1.0'/>
</node>
<node id='P0' label='P0' start='2013-05-18T05:01:10.901+0100' end='2013-05-18T05:07:10.901+0100'>
    <attvalues>
        <attvalue for='description' value='null'/>
    </attvalues>
    <viz:size value='1.0'/>
</node>
<node id='N2' label='N2' start='2013-05-18T05:01:25.684+0100' end='2013-05-18T05:07:25.684+0100'>
    <attvalues>
        <attvalue for='description' value='Sat May 18 05:01:30 BST 2013'/>
    </attvalues>
    <viz:size value='1.0'/>
</node>
<node id='P1' label='P1' start='2013-05-18T06:01:25.684+0100' end='2013-05-18T06:07:25.684+0100'>
    <attvalues>
        <attvalue for='description' value='null'/>
    </attvalues>
    <viz:size value='1.0'/>
</node>
<node id='N3' label='N3' start='2013-05-18T06:41:25.684+0100' end='2013-05-18T06:47:25.684+0100'>
    <attvalues>
        <attvalue for='description' value='N3'/>
    </attvalues>
    <viz:size value='1.0'/>
</node>
<node id='P2' label='P2' start='2013-05-18T06:51:25.684+0100' end='2013-05-18T06:57:25.684+0100'>
    <attvalues>
        <attvalue for='description' value='null'/>
    </attvalues>
    <viz:size value='1.0'/>
</node>
<node id='N4' label='N4' start='2013-05-18T06:51:25.684+0100' end='2013-05-18T06:57:25.684+0100'>
    <attvalues>
        <attvalue for='description' value='N3'/>
    </attvalues>
    <viz:size value='1.0'/>
</node>
<node id='P4' label='P4' start='2013-05-18T06:51:25.684+0100' end='2013-05-18T06:57:25.684+0100'>
    <attvalues>
        <attvalue for='description' value='null'/>
    </attvalues>
    <viz:size value='1.0'/>
</node>
<node id='N5' label='N5' start='2013-05-18T06:52:25.584+0100' end='2013-05-18T06:58:25.584+0100'>
    <attvalues>
        <attvalue for='description' value='NN'/>
    </attvalues>
    <viz:size value='1.0'/>
</node>
<node id='P3' label='P3' start='2013-05-18T06:52:25.584+0100' end='2013-05-18T06:58:25.584+0100'>
    <attvalues>
        <attvalue for='description' value='null'/>
    </attvalues>
    <viz:size value='1.0'/>
</node>
<node id='P5' label='P5' start='2013-05-18T06:58:15.584+0100' end='2013-05-18T07:15:15.584+0100'>
    <attvalues>
        <attvalue for='description' value='null'/>
    </attvalues>
    <viz:size value='1.0'/>
</node>
```

Figure 3. Nodes of the graph in Figure 5, inserted within the `<nodes></nodes>` *tags in Listing 1.*



```
<edge id='N0-N1' source='N0' target='N1' type='directed' label='2'>
    <attvalues>
        <attvalue for='weight' value='1'/>
    </attvalues>
</edge>
<edge id='N0-P0' source='N0' target='P0' type='directed' label='0'>
    <attvalues>
        <attvalue for='weight' value='1' start='2013-05-18T05:01:10.901+0100'/>
    </attvalues>
</edge>
<edge id='P0-N1' source='P0' target='N1' type='directed' label='1'>
    <attvalues>
        <attvalue for='weight' value='1'/>
    </attvalues>
</edge>
<edge id='N4-N1' source='N4' target='N1' type='directed' label='2'>
    <attvalues>
        <attvalue for='weight' value='2'/>
        <attvalue for='weight' value='2'/>
    </attvalues>
</edge>
<edge id='N4-P4' source='N4' target='P4' type='directed' label='0'>
    <attvalues>
        <attvalue for='weight' value='1' start='2013-05-18T05:01:10.901+0100'/>
    </attvalues>
</edge>
<edge id='P4-N1' source='P4' target='N1' type='directed' label='1'>
    <attvalues>
        <attvalue for='weight' value='1'/>
    </attvalues>
</edge>
<edge id='N2-N3' source='N2' target='N3' type='directed' label='2'>
    <attvalues>
        <attvalue for='weight' value='1'/>
    </attvalues>
</edge>
<edge id='N2-P1' source='N2' target='P1' type='directed' label='0'>
    <attvalues>
        <attvalue for='weight' value='1' start='2013-05-18T05:01:25.684+0100'/>
    </attvalues>
</edge>
<edge id='P1-N3' source='P1' target='N3' type='directed' label='1'>
    <attvalues>
        <attvalue for='weight' value='1'/>
    </attvalues>
</edge>
<edge id='N2-N4' source='N2' target='N4' type='directed' label='2'>
    <attvalues>
        <attvalue for='weight' value='1'/>
    </attvalues>
</edge>
<edge id='N2-P2' source='N2' target='P2' type='directed' label='0'>
    <attvalues>
        <attvalue for='weight' value='1' start='2013-05-18T06:51:25.684+0100'/>
    </attvalues>
</edge>
<edge id='P2-N4' source='P1' target='N4' type='directed' label='1'>
    <attvalues>
        <attvalue for='weight' value='1'/>
    </attvalues>
</edge>
<edge id='N5-P3' source='N5' target='P3' type='directed' label='0'>
    <attvalues>
        <attvalue for='weight' value='1' start='2013-05-18T06:52:25.584+0100'/>
    </attvalues>
</edge>
<edge id='N4-N1' source='N4' target='N1' type='directed' label='2'>
    <attvalues>
        <attvalue for='weight' value='2'/>
        <attvalue for='weight' value='2'/>
    </attvalues>
</edge>
<edge id='N4-P5' source='N4' target='P5' type='directed' label='0'>
    <attvalues>
        <attvalue for='weight' value='1' start='2013-05-18T06:58:15.584+0100'/>
    </attvalues>
</edge>
<edge id='P5-N1' source='P5' target='N1' type='directed' label='1'>
    <attvalues>
        <attvalue for='weight' value='1'/>
    </attvalues>
</edge>
```

Figure 4. Edge nodes of the graph in Figure 5, inserted within the `<edges></edges>` tags in Listing 1.



2013-05-18T05:00:58.975+0100 to 2013-05-18T05:07:25.684+0100

2013-05-18T06:01:25.684+0100

2013-05-18T06:41:25.684+0100

2013-05-18T06:51:25.684+0100

2013-05-18T06:52:25.584+0100

2013-05-18T06:58:15.584+0100

Figure 5. A sequence diagram showing time-based dynamic graph, with nodes and edges appearing according to the timestamps. The graph is described in the GEXF data structure in Figure 3. The numbering shows the sequence. The dates are in the order of when new nodes and edges appear and connect to the network.

The only disadvantage of the GEXF file format is perhaps the size attributed to the XML tags used for enclosing and describing the contents. In extremely large dataset, an alternative can be used, by removing the time element if they are not important to the research inquiry. For static maps which are much smaller in file size, a GDF file can be generated (see Section 3.2.6 for algorithm) as an output. Listing 2 is a GDF version of the GEXF graph file. The file is self-explanatory. The first section beginning with `nodedef>` followed by a list of attributes describing each line of node properties are the nodes. The second section beginning with `edgedef>` followed by the connection and the weight are edges.

```
nodedef>name VARCHAR,timeStamp VARCHAR,label VARCHAR,
N0,2013-05-18T05:01:10.901+0100,N0,
N1,2013-05-18T05:01:10.901+0100,N1,
P0,2013-05-18T05:01:10.901+0100,P0,
N2,2013-05-18T05:01:25.684+0100,N2,
P1,2013-05-18T06:01:25.684+0100,P1,
N3,2013-05-18T06:41:25.684+0100,N3,
P2,2013-05-18T06:51:25.684+0100,P2,
N4,2013-05-18T06:51:25.684+0100,N4,
P4,2013-05-18T06:51:25.684+0100,P4,
N5,2013-05-18T06:52:25.584+0100,N5,
P3,2013-05-18T06:52:25.584+0100,P3,
P5,2013-05-18T06:58:15.584+0100,P5,
edgedef>node1 VARCHAR,node2 VARCHAR,weight INT,
N0,N1,2,
N0,P0,0,
P0,N1,1,
N4,N1,2,
N4,P4,0,
P4,N1,1,
N2,N3,2,
N2,P1,0,
P1,N3,1,
N2,N4,2,
N2,P2,0,
P1,N4,1,
N5,P3,0,
N4,N1,2,
N4,P5,0,
P5,N1,1,
```

Listing 2. The listing shows the static graph version of the GEXF file format. The file format is GDF.

3.2.3 Mapping Algorithm

The mapping algorithm maps social media activities into an SIL. There are three components in such a network – the user, the post and a collection of users mentioned in the post. A user *v* participates in social media on a topic (collected with a hash tag, or keywords, within a product page or a media channel). The total collection of users within the topic is *V*. Each user may comment, tweet, or post a statement *p*, which may contain mentions of a user *m*. Within a post *p*, there may be a collection of mentions of other users *M*. *e* is an edge between two nodes and *E* is the collection of all the edges. The summary of the sets is:

v: a user object, also containing all the attributes associated with the user (timestamp, post, etc.)
V: a collection of all the users and their attributes
n: a node representing a user, containing all the information in v
N: all nodes in the network
p: a post by a user
m: a user mentioned within the post
M: all users mentioned within the post
e: a single edge between two nodes
E: a collection of edges

At the start of the algorithm, each user with the associated information is queried from the database and stored $V \leftarrow \{\{v\} : v \in V\}$. For each user, first assign the post $p \leftarrow$ the post of v. Get all the mentions from the post $M \leftarrow \{\{m\} : m \in M\}$. Check if the user exists in the node collection *N*, if not, add the user to the node collection Add(v) to N. If no users *m* are mentioned in the post *p*, add the post as a node AddNode(p). If users are mentioned in the post, and if any user is not in the collection *N*, add the user to the collect AddNode(m). The procedure CheckMentions(M) is as follows: For all mentions, add an edge between the user and the post AddEdge(n,p), add an edge from the post to the user mentioned AddEdge(p,m), finally, increase the weight of the edge by 1 between *n* and *m* Increment(w, 1) of n and m. If the user exists in the collection of nodes call the procedure CheckMentions(M). Finally, the data is output as a GEXF file.

```
$V \leftarrow \{\{v\} : v \in V\}$
FOR all $v \in V$ DO
    $p \leftarrow$ the post of v
    $M \leftarrow \{\{m\} : m \in M\}$ in mentions m of the post p of v
    IF $v \notin N$ THEN Add(v) to N, DO
        AddNode(p)
        IF count(M) < 1 DO
          AddEdge(n,p)
        ELSE
            proc CheckMentions(n);
                FOR all $m \in M$
                        IF $m \notin N$ THEN AddNode(m) to N
                                AddEdge(n,p), AddEdge(p,m), AddEdge(n,m)
                                Increment(w, 1) of n and m
                        END-IF
                END-FOR
            end-proc
        END-IF
    ELSE
        CheckMentions(n)
    END-IF
END-FOR
Output(GEXF)
```

Listing 3. The mapping algorithm

3.2.4 Reconfiguration Algorithm

The SIL mapped with the algorithm in Section 3.2.3 has this configuration:

$n \to p \to m,\ n \to m$

The network can be reconfigured to $n \to p \to m$ with the algorithm below, this simply 'unwraps' the complete connection by severing *n* and *m*'s edge:

```
N ← {{n} : n ∈ N } from GEXF file
FOR all n ∈ N DO
    RemoveEdge(n,m)
END-FOR
Output(GEXF)
```

Listing 4. The reconfiguration algorithm: ‘unwrapping’ the network

The network can also be reconfigured to remove all post, containing users only $n \to m$, with the algorithm below:

```
N ← {{n} : n ∈ N } from GEXF file
FOR all n ∈ N DO
    RemoveEdge(n,p), RemoveEdge(p,m), RemoveNodes(p)
END-FOR
Output(GEXF)
```

Listing 5. The reconfiguration algorithm: retaining users and removing posts

3.2.5 Dynamic Graph Output

The GEXF dynamic graph output (Section 3.2.2) has a simple procedure Output(GEXF). The header is written first by streaming a text output and inserting information where needed into the header, such as the start and end timestamps of the first appearance of, and the last appearance of the final node, and subsequently the opening and closing tags with the node and edge counts that encloses the individual node and edge tags. For each node, write the node tag WriteNode(n) with associated information such as the start and end timestamps. For each edge, write the edge tag WriteEdge(e) with associated source and target attributes, and add the weight attributes for visualisation purposes.

```
proc Output(GEXF);
    WriteHeader(N)
    WriteNodesTag(Count(N))
        FOR all n ∈ N DO
            WriteNode(n)
        END-FOR
    CloseNodesTag()
    WriteEdgesTag(Count(E))
        FOR all e ∈ E in N, DO
            WriteEdge(e)
        END-FOR
    CloseNodesTag()
    WriteFooter()
end-proc
```

Listing 6. GEXF file output

3.2.6 Static Graph Output

The GDF format (Section 3.2.2, Listing 2) is a static format that has the same expression as the dynamic version, apart from the removal of the time element. The output and algorithm is straightforward. For the nodes and edges respectively, the labels are being written first before the corresponding node and edge lists are appended.

```
proc Output(GDF);
    writeNodeLabels(N);
    FOR all n ∈ N DO
        WriteNode(n)
    END-FOR
    writeEdgeLabels(N);
    FOR all e ∈ E in N, DO
        WriteEdge(e)
    END-FOR
end-proc
```

Listing 7. GDF file output algorithm

4. Visualising Mapped Social Information Landscapes

This section demonstrates the feasibility of the Big Data architecture and the algorithms used for mapping Social Information Landscapes. The first subsection looks at the capability of the methods in managing the volume and velocity of topical datasets, followed by samples expressions of SIL retrieved from the topical datasets. The final subsections inspect the details of some of the graphs.

4.1 Volume and Velocity

Figure 6 are graphs of viral outbursts of trending data captured from Twitter with the Big Data architecture. Each point within the graphs is a five hourly recording of Twitter activities as a SIL with the associated hash tags (caption at the top), mapped as an SIL using the GEXF file format. The #MH370 dataset has over 21 data points for example, whilst the #RoyalBaby dataset has 91 data points. The volume and velocity of each trending topic can be observed, particularly at the spikes of the data point at particular events released by news channels. #MH370 has one of the largest data point at around 500 Megabytes on the 24 March 2014, all the data points in total amounted to over 14 Gigabytes. The graphs show different signatures. These signatures are by-products of the mapping algorithms and are potentially a useful source of information. For example, the spikes in the #FreeJahar dataset are news releases in associated Twitter accounts of news channels, out of which the majority of the data are retweets. However, in the case of the #FreeJahar, communities were identified within those datasets (see [2]), and community activities were heightened in the spiked data points. The #PRU13 is a Malaysian political election dataset, it is easy to see that 5th May of 2013 has the highest activity as it was the day of the election. There is another spiked data point a little later, which is related to news of swapped ballot boxes during a supposedly deliberate electrical black out. The hills and valleys of data points for other datasets may just be the heightened activities during the day and the easing off of tweets during the night (e.g., the #McCann dataset).

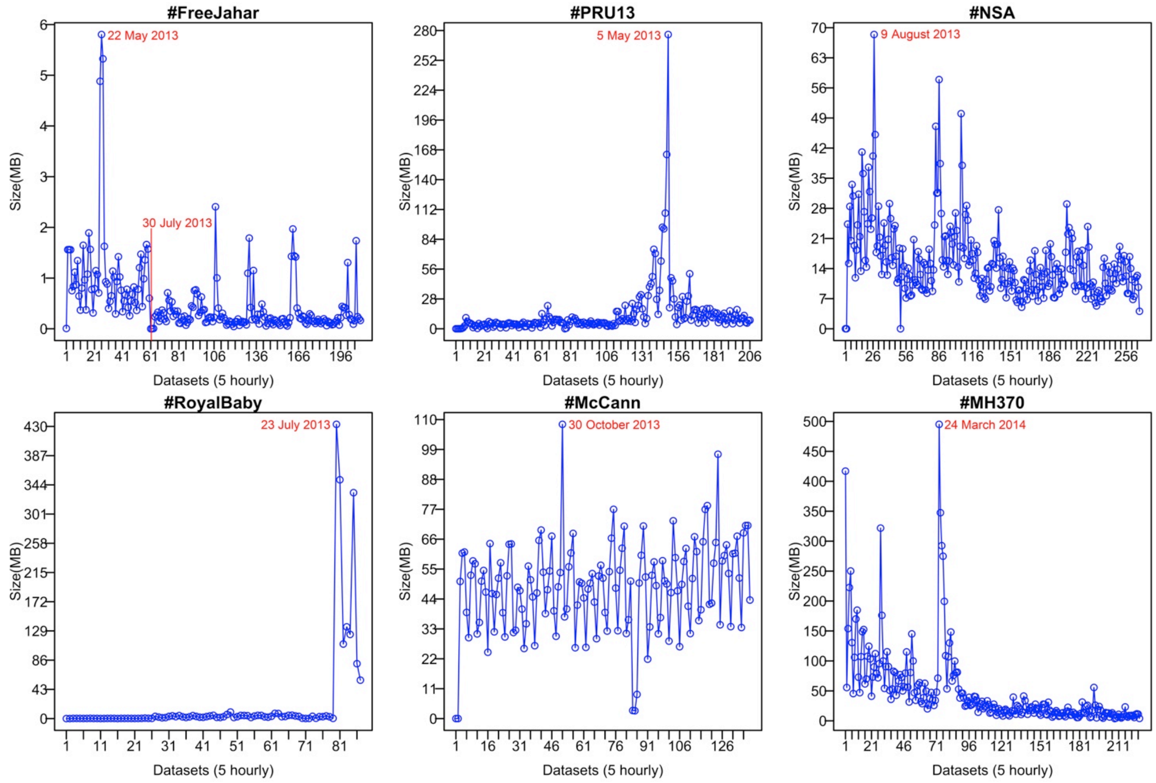


Figure 6. Graphs of viral outbursts of trending data captured from Twitter with the Big Data architecture. Each point in the graphs is a five hourly recording of a Twitter SIL with the associated hash tags (caption at the top). The #MH370 datasets have over 21 points for example, whilst the #RoyalBaby dataset has 91 data points.

4.2 Expressions of Social Information Landscapes

In Figure 7 are six SILs sampled from the topical datasets in Figure 6. Here, the value of the methods for mapping SIL is demonstrated. Each SIL (a data point) is a recording of five hourly activities in Twitter. Each SIL shows a different landscape expression. The degree centrality (number of in-degrees and out-degrees) is used for measuring each node, resulting in the more important nodes (larger in size) with a higher degree centrality within the multimodal activity networks. The number of nodes in each SIL is show at the bottom of each graph. The graphs described with the GEXF format is visualised with the Gephi graph visualisation package.

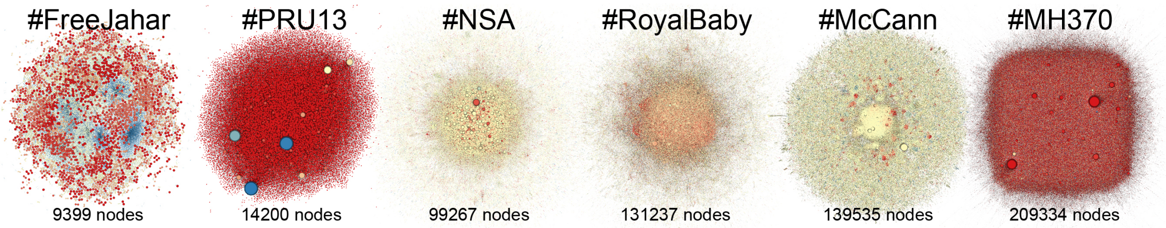


Figure 7. Each SIL here is a data point sampled from the topical graphs in Figure 6. Each SIL (a data point) is a recording of five hours of activities in Twitter. They show different expressions. The important nodes (larger in size) are nodes with a higher degree centrality within the multimodal activity networks.

The largest graph in Figure 7 has over 200,000 nodes (#MH370), the number of nodes and edges in the graph is a challenging problem as it may be difficult for certain graph visualisation packages to process. However, many of the nodes may not be important to social network analysis. The unimportant nodes are usually unconnected, or the nodes may be individuals who posted only a single tweet. Unless semantic or sentiment analysis is conducted on the full set, these nodes can be safely removed from the graph. Figure 8 is a set of graphs with nodes which have a limited number of edges removed from the graphs in Figure 7, the caption on the top right corner of each inset shows the reduced node and edge size, the parenthesis (e.g., <3) is the filter used for reducing nodes with edges less than a number, in the case of #PRU is nodes with edges <3. The smaller graphs are useful for efficient graph processing using graph visualisation software whilst maintaining the quality and meaning of the SIL.

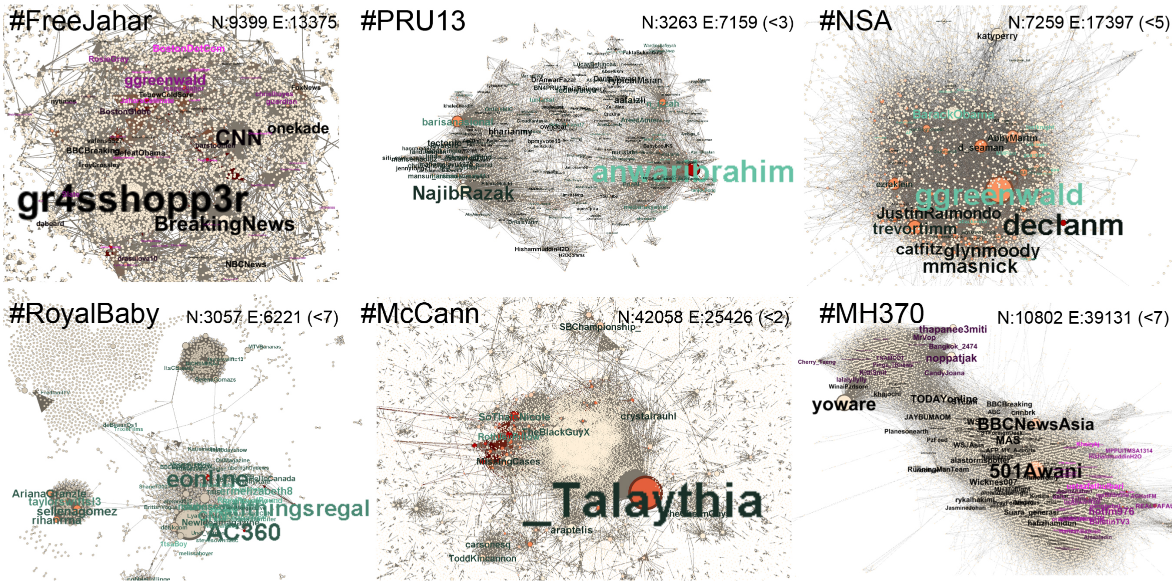


Figure 8. The SIL datasets here are decimated graphs from Figure 7. Most singular nodes have been removed for more efficient graph processing using graph visualisation software whilst maintaining the quality and meaning of the SIL.

The different signatures in each of the SIL are obvious here. The signature of each SIL is due to the nature of the activities, the background of the topic and the intentions of the social media users. It can be observed that users who interacted more are closer together in that they form natural clusters. The #FreeJahar dataset which has very interesting phenomenon has been analysed in detail [2].

The other datasets are also of considerable interests to academics in the social sciences. Here, a brief description of the social network expression is sufficient until further results from the analysis are published.

In the #PRU graph, two large political parties contended during the Malaysian general election bridged by activists, the opposition leader @AnwarIbrahim is flanked by media channels and have large degree and betweenness centralities, an indication of heightened activities which greatly strengthened the opposition party's stand. In the controversial #NSA news, within the graph, a Guardian correspondent is highly active, with clusters of activities directly below in the SIL. Note that #KatyPerry is above in a separate but linked community. The glamorous birth of the #RoyalBaby were mentioned by celebrities (e.g., @selenagomez, etc) with pockets of discussions. The reawakening of the Madeleine #McCann kidnapping case due to new evidence led to separate clusters of varying degree of intense conversations. A single MH370 SIL datapoint contains two separate clusters of important posts by activists.

The brief analysis and visualisation of the above datasets raise some interesting questions. For example, why are there many more distinct clusters of activities within the #McCann SIL as compared to the other datasets? Is this an indication that there were many kidnapping related experiences in the separate clusters of social media users? Whilst the answers to these questions are interesting, they are beyond the scope of this paper.

4.3 Network Structures and Centralities

In this section, three datasets captured from social media are shown. The datasets demonstrate the details of the algorithms in mapping SIL.

Figure 9 is a capture and mapping of a SIL 15 July 2013 for 9 days on the press release related to the publication of the discovery of "World's Oldest Calendar" [18]. There were 44 datasets (5 hourly each data point), which have been combined into a single map in Figure 9. There were 5973 nodes (users and their tweets) and 6100 edges (connecting users and tweets). The relative importance of the nodes within the SIL in diffusing information uses the in- and out-degree centrality (larger, redder nodes).

A of Figure 9 shows the complete dataset composed of 9 days of activities since 15 July 2013 using the mapping algorithm (7187 nodes, 10322 edges). B shows an unwrapped version of the SIL using the reconfiguration algorithm in 3.2.4. The isolated nodes have been removed, leaving only connected nodes in the network (2060 nodes, 2203 edges). C shows a closer view of an isolated cluster within the SIL, illustrating the detailed connections between different types of nodes (users and posts).

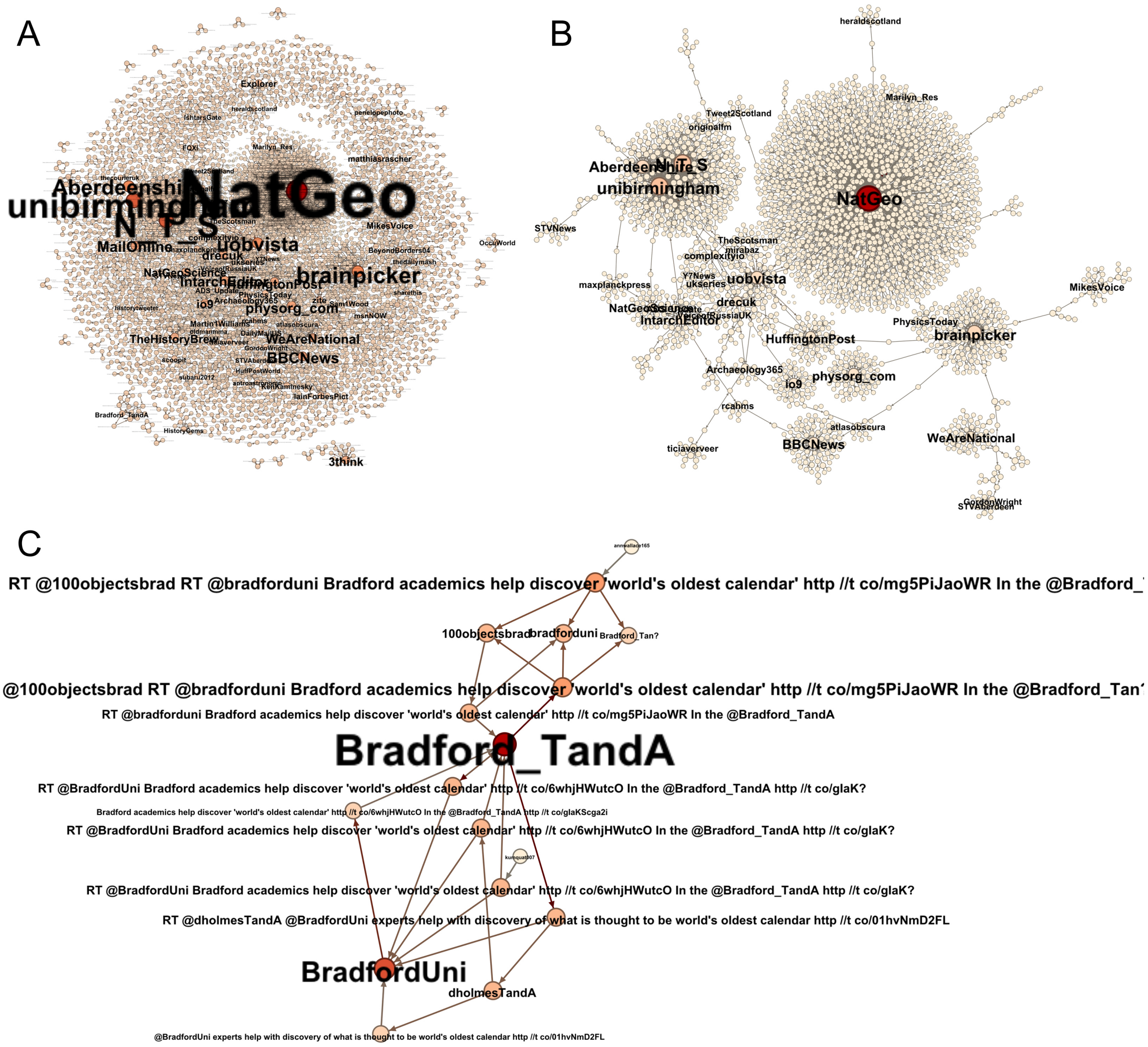

A
B
C
NatGeo
unibirmingham
Aberdeenshire
N T S
uobvista
brainpicker
physorg_com
MailOnline
drecuk
IntarchEditor
HuffingtonPost
io9
TheHistoryBrew
WeAreNational
BBCNews
Explorer
heraldscotland
Marilyn_Res
matthiasrascher
MikesVoice
Bradford_TandA
3think
Tweet2Scotland
originalfm
STVNews
TheScotsman
mirabaz
complexityio
maxplanckpress
Y7News
ukseries
NatGeoScience
PhysicsToday
Archaeology365
rcahms
atlasobscura
ticiaverveer
GordonWright
STVAberdeen
RT @100objectsbrad RT @bradforduni Bradford academics help discover 'world's oldest calendar' http //t co/mg5PiJaoWR In the @Bradford_
100objectsbrad
bradforduni
Bradford_Tan?
@100objectsbrad RT @bradforduni Bradford academics help discover 'world's oldest calendar' http //t co/mg5PiJaoWR In the @Bradford_Tan
RT @bradforduni Bradford academics help discover 'world's oldest calendar' http //t co/mg5PiJaoWR In the @Bradford_TandA
Bradford_TandA
RT @BradfordUni Bradford academics help discover 'world's oldest calendar' http //t co/6whjHWutcO In the @Bradford_TandA http //t co/glaK?
Bradford academics help discover 'world's oldest calendar' http //t co/6whjHWutcO In the @Bradford_TandA http //t co/glaKScga2i
RT @BradfordUni Bradford academics help discover 'world's oldest calendar' http //t co/6whjHWutcO In the @Bradford_TandA http //t co/glaK?
RT @BradfordUni Bradford academics help discover 'world's oldest calendar' http //t co/6whjHWutcO In the @Bradford_TandA http //t co/glaK?
RT @dholmesTandA @BradfordUni experts help with discovery of what is thought to be world's oldest calendar http //t co/01hvNmD2FL
BradfordUni
dholmesTandA
@BradfordUni experts help with discovery of what is thought to be world's oldest calendar http //t co/01hvNmD2FL

Figure 9. Trending tweets of press releases related to the publication of discovery of the "World's Oldest Calendar". A) The complete dataset composed of 9 days of activities since 15 July 2013 using the mapping algorithm (7187 nodes, 10322 edges). B) An unwrapped version of the SIL using the reconfiguration algorithm in 3.2.4. Isolated nodes have been removed, leaving only connected nodes in the network (2060 nodes, 2203 edges). C) A closer view of an isolated cluster within the SIL, showing the detailed connections between different types of nodes (users and posts).

Figure 10 is a visualisation of an SIL sampled from one of the MH370 data points. Posts have been removed, leaving only user nodes within the network of activities, which forms a small world network [3]. The important nodes have higher degree centralities (larger and redder nodes). The nodes are news correspondence of major news channels. The main activities here are retweets from users of news events. Note that unlike the #FreeJahar SIL in Figure 11, there are no conversational clusters.

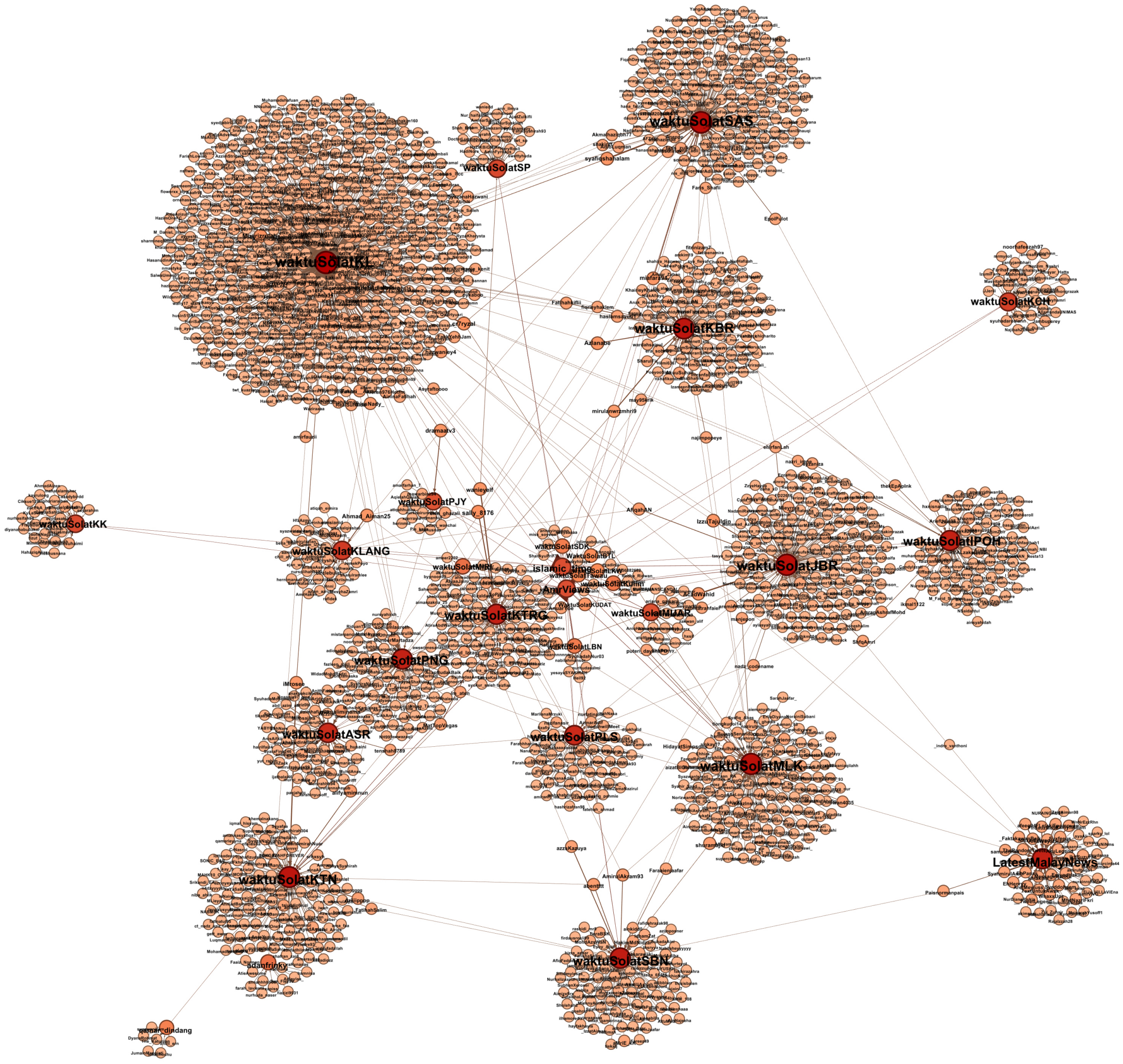


Figure 10. An SIL sampled from one of the MH370 data points. Posts have been removed, leaving only user nodes within the network. The main activities here are retweets of news events.

Figure 11 is a visualisation of the largest five-hourly #FreeJahar SIL. The expression of the SIL is typical of a network with heightened conversational activities, the persistence of the cluster indicates that a community has formed. The SIL is mapped with the algorithm in 3.2.3 and has not been reconfigured. The nodes are colour coded – black nodes are retweets, blue nodes are conversational posts, the shades of red are Twitter users. The reddest nodes are ones with higher measure of betweenness centrality, a sign that they are the 'gate keepers' of information. An analysis of the #FreeJahar dataset collected and mapped with the methods described here is available [2]. The paper investigates the formation, development and decline of the #FreeJahar community and elaborates the phenomenon discovered here.

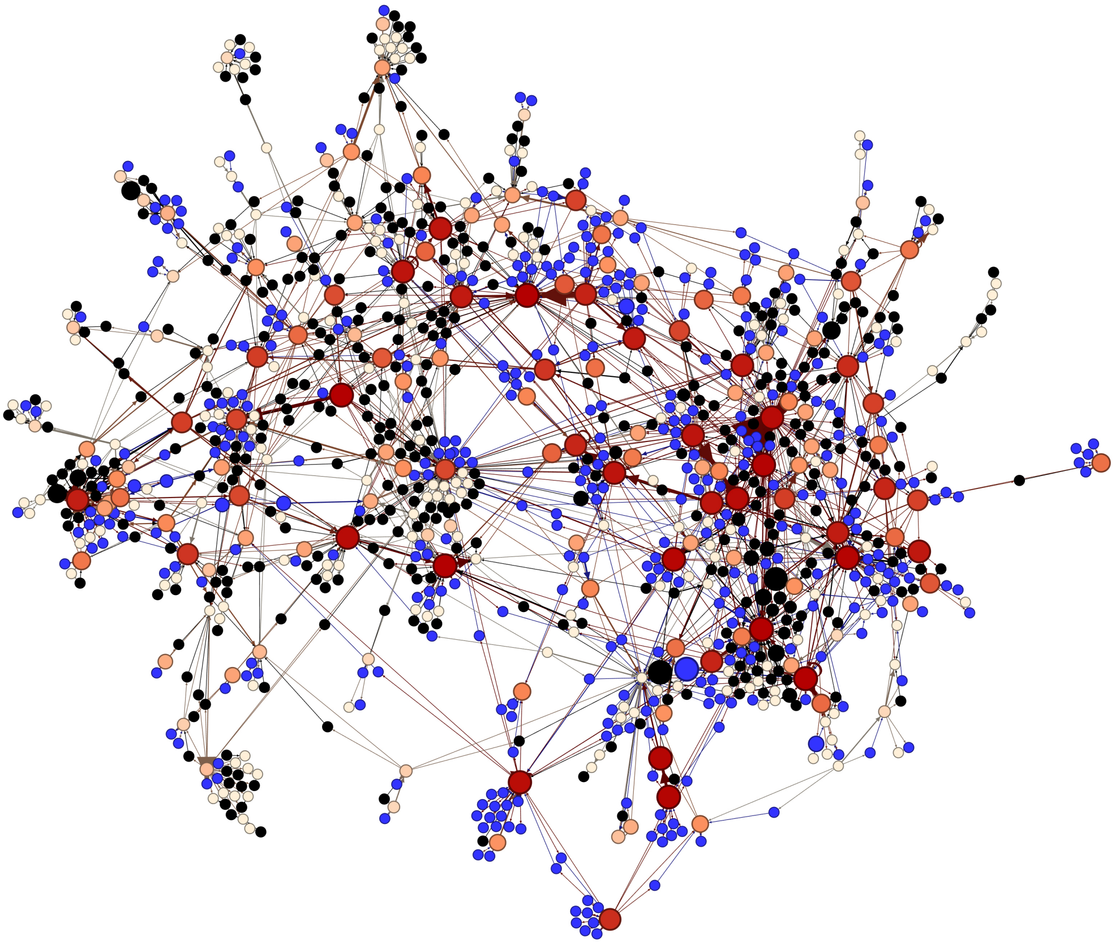


Figure 11. Visualisation of the #FreeJahar SIL. The expression of the SIL is typical of a network with heightened conversational activities. Black nodes are retweets, blue nodes are conversational posts, the shades of red are Twitter users. The reddest nodes are ones with higher measure of betweenness centrality.

5. Discussion

Social media certainly increases our range of activities, and collectively in the global sense, the human connectedness beyond the boundary of a participant's geographical location. The time of interaction between participants and communications sent now may be retrieved and responded to, much later by other participants due to Information Technology infrastructures. Thus, the distance between time and space has been greatly shortened. This invariably opens up a broad range of opportunities as space and time, in the eye of a user are 'compressed' to within a digital display, allowing diverse communications from extremely large demographics groups. Participants interact without thinking much about the information networks that they are building, and users reading social media are oblivious to the fact that a hidden network exists. This is due to the fact that the limited presentation of the sequential nature of content listings within social media, as viewed on Web browsers or on mobile devices do not necessarily reveal nor make obvious an unknown nature of the medium; that every participant, from content producers, to consumers, to followers and subscribers, including the contents they produce or subscribed to, are intrinsically connected in a hidden but potentially massive network. These hidden networks are a valuable source of information for scientific inquiry. Such networks when mapped and analysed using appropriate methods could reveal valuable information. These multimodal networks, together with information about the nodes, direction of edges, the polarity of interactions, the structure of the graph and secondary data captured longitudinally form a Social Information Landscape. Social Information Landscapes have become an important aspect of contemporary scientific inquiry, and the importance will only increase in the future when all things both men and machines are digitally connected.

In this article, the definition of Social Information Landscapes is set, together with a detail of the methods for mapping such landscapes. The mapping of SIL undoubtedly requires a Big Data approach, for the temporal range may be great if meaningful longitudinal datasets are to be acquired. Volume and velocity have been a major challenge in acquiring social media datasets, this may mean that manual approaches in collecting and mapping networks will no longer be feasible. To address the issue of volume and velocity, a Big Data architecture integrated with scalable open source libraries and API is provided as a solution. The solution uses server-side asynchronous I/O software developed for connecting to and streaming data from social media and the Web. The system is designed so that incoming data is immediately stored in scalable and distributed NoSQL databases, and when the data collection is complete, separate algorithms are used for streaming structured data in stable, evenly spaced chunks for mapping SIL. A set of algorithms were presented in section 3 for mapping, reconfiguring and storing SIL, together with patterns for dynamic and static data formats that could be used with graph visualisation packages. Section 4 demonstrates the feasibility of the Big Data architecture and the usefulness of mapping large multimodal and longitudinal data by presenting real-world social media datasets as examples.

The methodology presented here has laid a foundation for mapping multimodal and longitudinal networks. However, much work is required for the future, especially with regards to issues of Big Data. One of the issues is the (1) *processing of extremely large graphs* with hundreds of thousands of nodes and edges. In the datasets given within this article, the largest graph reaches 200k nodes in a single data point, in actual fact however, a SIL is a combination of all the data points as the activities are continuous through time. This will bring the size of nodes to over one million with edges doubling the size of the nodes. On a Dell PowerEdge C6100 with 2x Intel Xeon X5660 Processor (2.80GHz, 12M Cache, 6.40 GT/s QPI, Turbo, HT), 1333MHz Max Memory, 48GB Memory (6x8GB Dual Rank RDIMMs), 200k nodes is manageable. In larger social media involving months to years of data collection, the size could easily reach tens of millions. Extremely large graph processing is therefore an issue and needed parallel and distributed processing. The second issue is (2) *real-time mapping of networks*. Unlike the data management covered in this article, which stores all datasets prior to processing them in a stable stream, real-time processing may need to discard unwanted or redundant raw data as they may be too big for storage, keeping only data which has been structured. This invariably requires real-time mapping and configuration of graphs, implying that parallel and distributed processing is required. The final issue is (3) *large SIL visualisation*, which involves (1) and (2) will require distributed and parallel processing, particularly with GPGPUs. A single GPGPU with thousands of processing cores within a single computer is a suitable platform for developing algorithms that manage all three issues discussed above. The Big Data and Visual Analytics Lab [19] is at present working with NVIDIA in various research projects in developing algorithms in all three issues as part of our future work.

Acknowledgements
The author acknowledges the financial support from the International Doctoral Innovation Centre, Ningbo Education Bureau, Ningbo Science and Technology Bureau, China's MoST and The University of Nottingham. The project is partially supported by NBSTB Project 2012B10055.

References

1. Ch'ng E (2014) The Value of Using Big Data Technology in Computational Social Science. The 3rd ASE Big Data Science 2014, Tsinghua University 4-7 August. Beijing, China.: Association for Science and Engineering. doi:http://dx.doi.org/10.1145/2640087.2644162.

2. Ch'ng E (2015) The Bottom-Up Formation and Maintenance of a Twitter Community: Analysis of the #FreeJahar Twitter Community. Ind Manag Data Syst In press.

3. Ch'ng E (2015) Local Interactions and the Emergence and Maintenance of a Twitter Small-World Network. Soc Netw In press.

4. Lazer D, Pentland AS, Adamic L, Aral S, Barabasi AL, et al. (2009) Life in the network: the coming age of computational social science. Sci (New York, NY) 323: 721.

5. Anderson C (2008) The end of theory. Wired Mag 16.

6. Latour B (2010) 10 Tarde's idea of quantification. Soc after Gabriel Tarde debates assessments: 145.

7. Manovich L (2011) Trending: the promises and the challenges of big social data.

8. Jacobs A (2009) The pathologies of big data. Commun ACM 52: 36–44.

9. Boyd D, Crawford K (2011) Six provocations for big data. A Decad Internet Time Symp Dyn Internet Soc Soc Sci Res Netw.

10. Bontcheva K (2014) Lie Detector on the Way to Test Social Media Rumours. BBC Technol. Available: http://www.bbc.co.uk/news/technology-26263510.

11. Bontcheva K (2014) EU Project to Build Lie Detector for Social Media.

12. Java A, Song X, Finin T, Tseng B (2007) Why we twitter: understanding microblogging usage and communities. Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis. ACM. pp. 56–65.

13. Jansen BJ, Zhang M, Sobel K, Chowdury A (2009) Twitter power: Tweets as electronic word of mouth. J Am Soc Inf Sci Technol 60: 2169–2188.

14. Huberman B, Romero D, Wu F (2008) Social networks that matter: Twitter under the microscope. Available SSRN 1313405.

15. Sakaki T, Okazaki M, Matsuo Y (2010) Earthquake shakes Twitter users: real-time event detection by social sensors. Proceedings of the 19th international conference on World wide web. ACM. pp. 851–860.

16. Kwak H, Lee C, Park H, Moon S (2010) What is Twitter, a social network or a news media? Proceedings of the 19th international conference on World wide web. ACM. pp. 591–600.

17. Bourdieu P (1985) The social space and the genesis of groups. Theory Soc 14: 723–744.

18. Gaffney VL, Fitch S, Ramsey E, Yorston R, Ch'ng E, et al. (2013) Time and a Place: A luni-solar "time-reckoner" from 8th millennium BC Scotland. Internet Archaeol 34.

19. Ch'ng E (2015) Big Data and Visual Analytics Lab. Available: http://www.nottingham.edu.cn/en/science-engineering/departments/computer-science/research/groups/bdva/big-data-and-visual-analytics.aspx. Accessed 19 February 2015.