Future Generation Computer Systems 110 (2020) 976–983

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Towards a large-scale twitter observatory for political events

Senaka Fernando [a], Julio Amador Díaz López [a], Ovidiu Şerban [a], Juan Gómez-Romero [b], Miguel Molina-Solana [a,*], Yike Guo [a]

[a] Imperial College London, UK
[b] Universidad de Granada, Spain

ARTICLE INFO

Article history:
Received 16 May 2019
Received in revised form 20 September 2019
Accepted 27 October 2019
Available online 31 October 2019

Keywords:
Twitter analytics
Large-scale visualisation
Big data
Social media
Scalable resolution display environments

ABSTRACT

Explosion in usage of social media has made its analysis a relevant topic of interest, and particularly so in the political science area. Within Data Science, no other techniques are more widely accepted and appealing than visualisation. However, with datasets growing in size, visualisation tools also require a paradigm shift to remain useful in big data contexts. This work presents our proposal for a Large-Scale Twitter Observatory that enables researchers to efficiently retrieve, analyse and visualise data from this social network to gain actionable insights and knowledge related with political events. In addition to describing the supporting technologies, we put forward a working pipeline and validate the setup with different examples.

1. Introduction

Social media provide a wealth of information that reveals insights into current affairs and ongoing events [1,2], with recent years seeing an increasing interest in online social networks exploration [3]. Availability of unprecedented amounts of data about human interactions from different social networks opens the possibility of using this information to leverage knowledge about the diversity of social behaviour and the activity of individuals [4,5].

The focus of social data analysis is essentially on user-generated content and its dynamics. These data are rich, diverse and abundant, which makes them a relevant source not only for data science [6] but also in computational social science, social media analytics, complex systems and knowledge discovery with social Big Data.

As the use of online social networks has become mainstream, their role as agents for social change has also increased [7]. Their use in events ranging from the spread of political news, election campaigns and protests has demonstrated their importance in the context of fomenting social change. While this was a well-known fact for researchers in the political sciences, it became widely evident during the 2016 US presidential elections, in which misinformation (the so-called *fake news*) had a visible presence. Therefore the use and analysis of social media data in the context of political sciences has become relevant.

Data from social media can be analysed in traditional ways as it has been done for several decades in political science. However, with the growing amount of data that social media makes available, the tools have become more and more complex, posing and added difficulty to these researchers.

Visualisation has proven to be an appropriate means to uncover hidden patterns in data and to effectively communicate those findings to both expert audiences and layman. It is particularly useful when little is known about the source data and the analysis goals are vague.

On the other hand, scalable resolution display environments (SRDEs) are becoming increasingly popular as a mechanism to provide viewers with an immersive visualisation experience when working with large and complex datasets. This has been proven to be an effective means of gaining insight by collaborative group work. The size and arrangement of SRDEs often encourage active participation of researchers of multiple disciplines, which is crucial for exploring and understanding data coming from social networks.

One of the lines of research within the Data Science Institute at Imperial College London is precisely on ways to better understand and communicate the information hidden within large datasets, and particularly so in those coming from social media. Visualisation offers an intuitive mechanism to drive such an analysis specially if complemented with a suitable set of tools for interaction.

Recently, our research group proved that graph visualisations, backed by a large-scale visualisation environment, can be

* Corresponding author.
E-mail addresses: senaka.fernando15@imperial.ac.uk (S. Fernando), j.amador@imperial.ac.uk (J. Amador Díaz López), o.serban@imperial.ac.uk (O. Şerban), jgomez@decsai.ugr.es (J. Gómez-Romero), mmolinas@ic.ac.uk (M. Molina-Solana), y.guo@imperial.ac.uk (Y. Guo).

https://doi.org/10.1016/j.future.2019.10.013
0167-739X/

very helpful to understand the dynamics of different phenomena; e.g. Bitcoin transactions [8], online discussions [9] and biosurveillance [10]. These projects showed how filtering, grouping and highlighting specific patterns help to increase the usability of large graph visualisations. Nevertheless, we also experienced that the scalability of the implemented platforms is limited [11], which calls for new initiatives.

However, visualisation alone is not enough if not powered by a flexible and scalable underlying platform that enables the adequate collection and processing of the data. This work puts forward such a working pipeline, and thus the main contribution of our manuscript is to describe this end-to-end product to analyse and visualise social media data for the purpose of understanding political events.

Twitter can be considered the most studied online social network, with plenty of studies relying on it [1,10,12]. This social media platform provides an efficient and effective communication medium for one-on-one interactions and broadcast calls. In Twitter users post messages (originally limited to 140 characters and extended to 280 in September 2017) known as tweets. Every day, around 500 million tweets are produced, and Twitter has 330 million regular users [13]. The tweets are easily accessible via Twitter's API, either as a real-time stream or bulk search using a RESTful API.

Particularly, scholars have used Twitter data to analyse the way in which people discuss candidates and party leaders during elections in Germany [14], the US [15], and the UK [16,17]. Because Twitter is used to share information, opinions and online petitions, this social network provides us with an important source of data useful to analyse misinformation and its spread.

The manuscript is organised as follows: the next section provides a review of the context of our work. Section 3 describes our proposed pipeline for a Twitter Observatory, and Section 4 showcases some of the use-cases we have used it so far at Imperial College London. The manuscript concludes pointing out some existing limitations and future lines of action.

2. Related work

This section presents the context of our work by reviewing previous efforts on large-scale visualisation, and social media for political events.

2.1. Visualisation

Visualisation is about communicating and perceiving data, both abstract and scientific, by means of the human visual system. Although it was not until the 1980s [18,19] that data visualisation found a place in research, it has since then proved to be an effective way of gaining insights into data [20] both at the initial and final stages of the analysis.

Visual data exploration, in particular, is especially useful when little is known about the data and the exploration goals are vague. In this context, visual data exploration can be viewed as a hypothesis-generation process [20]. In this process, the user can interactively shift and adjust the analysis goals. Hypotheses might get validated or rejected on a visual basis, and new ones can be introduced. According to Shneiderman [21], visual data exploration follows a three-step process comprising overview, zoom and filter, and details-on-demand.

As datasets grow in complexity and scale, so have done their visualisations, with many research teams and engineers developing hardware and software tools to enable large-scale visualisations on big screens and video-walls. The hypothesis is that more visual space will foster greater speed, accuracy, comprehension and confidence in the data analysis and interpretation. Pioneers on this topic were the works by the Electronic Visualization Laboratory at the University of Illinois at Chicago, and their CAVE [22] and CAVE2 [23] environments. Other examples of large visualisation environments reported in literature include [24–27].

Over the past 20 years, SRDEs have evolved from environments that supported resolutions of several megapixels to those that support resolutions of several gigapixels. The sheer scale of these systems makes it impossible for them to be built using a few screens or being powered by one or few compute nodes. Systems such as these are built to have 100s of screens controlled by 100s of compute nodes. These, therefore, require specialised software that is capable of rendering and seamlessly interacting with content that is displayed across multiple screens.

One of the advantages of high-resolution visualisations is the ability to combine in the same display the aforementioned three steps (overview, zoom and filter and details-on-demand) described by Shneiderman. Besides the obvious technical difficulties that large visualisation poses (e.g. screen synchronisation, high-resolution displays, bandwidth), it also prompts a shift from a single-user workflow towards a more collaborative one. This has critical advantages for the comprehension of data and identification of insights, and it is one of the key benefits of this type of systems.

Roberts et al. [28] recently reflected on the new challenges and opportunities that lie ahead for visualisation. Some of these challenges are addressed in our present work, namely, large-scale immersive visualisation and human-like querying. They both aim at achieving a better comprehension of the growing datasets currently available, in intuitive ways for the users.

To perform this study, we leveraged the large-scale visualisation observatory built at Imperial College London premises, which features a distributed rendering cluster with 64 46″ HD screens. These screens have been arranged in a 6 m diameter cylinder with 16 columns of 4 monitors. The 313-degree immersive space has a total resolution of 132 M pixels across 37.31 m^2.

2.2. Visualising Twitter data

In fact, the idea of a Twitter Observatory is not new, and was prominently hinted at by Basaille et al. [29]. However, their work is largely focused on providing a reliable and scalable means for harvesting and storing tweets which can then be analysed by different pluggable components. They deliberately leave aside the (big data) visualisation part and only provide examples of displaying aggregated measures over the dataset.

When analysing Twitter data there are two main perspectives to it: the networked approach (i.e. how tweets are linked to each other and to their authors), and textual approach (in which the text is analysed). In fact, both approaches are complementary and can offer different insights. Any Twitter Observatory would need to take both into account.

Regarding the networked approach, it is of particular interest the relationships between tweets, hashtags, user mentions, and retweets. A network analysis of these entities has often unravelled interesting insights [17].

The increase in the relative size of datasets to account for real-world problems has forced researchers and engineers to move to distributed and parallel proposals [30] in order to explore and process large graphs, and to calculate different measures over them. Machine learning research has made great advancements in developing scalable and distributed algorithms, but they often lack support for interactivity. Similarly, human–computer principles for interaction and visualisation are difficult to scale to the degree needed with the large size of current datasets.

This networked approach is after all similar to those of graph and linked-data (which we have studied elsewhere [9,11]). Some

challenges of networked data visualisation are context adaptation, users and data heterogeneity, supporting different tasks (query, combination, filtering, etc.), and more recently, performance [31]. Gephi (a general-purpose graph visualisation tool [32]) supports data retrieval from several sources, calculation of graph measurements and automatic graph layout. However, it suffers from performance problems with medium-size graphs.

To the best of our knowledge, there are no specialised tools to visualise social media in the context of politics at a very large resolution.

2.3. Political analysis with social media

Within the realm of political participation, social networks play a prominent role. Specifically, social networks are an important conduit of political information, central to participation and, most important, relevant as it is through social networks that political trust is built. Political networks determine how partisan information – or misinformation – spread, how we vote, or whether or not we join a rally. Additionally, by analysing network structures it is possible to uncover how political networks affect our behaviour, the way in which we organise in, and, ultimately, how our institutions are built.

With the rise of online social media, a plethora of network data has become available in the last ten years. Political scientists have profited from this trend and adopted online social media data as an important instrument within their toolbox. In particular, early political science research centred in representing and understanding network structures. This gives way to questions such as *why do we observe certain patterns?*, *how do these patterns affect power structures?* and finally led to the study of how social networks formed: *are such patterns occurring at random?*, or *are individuals following strategic patterns to form networks?*

Formation of networks leads to different implications such as different diffusion patterns for political information; which ultimately affect the way in which individuals learn (*do opinion leaders have an out-weighted influence in social structure? how does the wisdom of the crowds affect political decisions?*) and to the study of how such structures affect individual behaviour. In the end, political institutions and systems are built through networks of individuals. This is the key reason why studying such structures has become particularly important for social (and particularly political) scientists.

2.4. Ethical considerations

While social media data and its public availability have enabled and fostered a large number of research projects, the ethical considerations of using such data have often been diminished and sidelined. Even though this paper is focused on the technical tool, we strongly believe that researchers using such tools should be knowledgeable on the ethical issues that their use might pose.

It is a fact that researchers across the social sciences are routinely harvesting twitter data and publishing identifiable content without any protection on sensitive data or users' consent. While this is possible according to Twitter's terms of service, which clearly states that public tweets will be made available to third parties, several researchers have questioned whether or not this is ethically acceptable [33].

It is commonly accepted that the terms of service of social media networks provide adequate provision to cover its harvesting and usage for social science research. However, if the data is enriched with additional metadata on the users (which might include sensitive personal features) derived from algorithms or other means, the legal issue of privacy may be compounded [33].

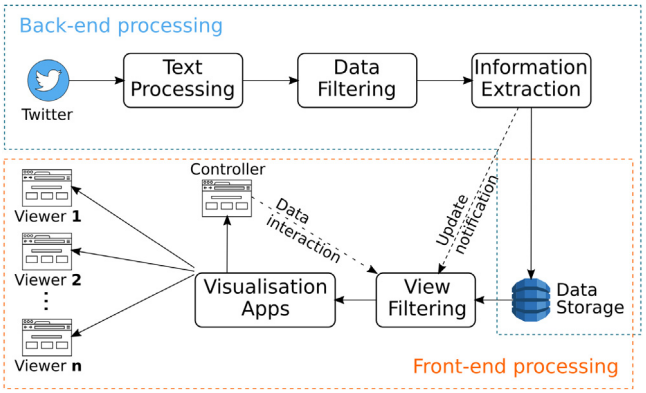

Fig. 1. The system architecture of the Twitter Observatory.

Our proposed Twitter Observatory is agnostic with respect to the data shown and the inferences made on them. It is the role of the researcher to implement the necessary provisions so that no ethical breach exists in their interpretation of the data, even when completely covered by the terms of service.

3. Design and architecture

Due to the volume of data in today's social media, any analysis and visualisation platform requires a dedicated architecture to achieve acceptable performance in this context. Those architectures rely on distributed system, and so we propose a distributed processing pipeline for the back-end, a storage solution and a distributed rendering front-end (i.e. SRDE). This design is flexible enough to allow a data flow even when one of the front-end or back-end components are missing. Moreover, the SRDE is distributed by design, with multiple renderers using the same dataset to show graphical information on different shards of the data. In our design the sharding happens both on the storage and viewer side. Fig. 1 shows our system architecture that separates the back-end from the front-end processing pipelines.

We have chosen to focus the rest of the paper on long-term occurring political patterns in Twitter data, which requires several months worth of data. This can be acquired either in real-time or using a snapshot obtained from the Twitter Search API. In our approach we assume the data is pre-processed as a stream for better parallel processing, therefore the source would be treated in the same way. Twitter provides data in JSON format containing the text of the tweet, metadata and user information. The metadata usually contains the location of the tweet (if the user acknowledged Twitter to make this information public), hashtags and mentions. Both APIs allow pre-filtering by location, usernames, keywords and hashtags to reduce the amount of data to be pre-processed.

3.1. Process & analysis

The processing pipeline can be described in two stages: (1) a common one with all steps applied to all datasets; (2) a particular one consisting on specific classification and filtering steps for different applications. The parameters and methods for the second stage are described in the experiment section.

Among the common steps, text cleaning is initially perform to tokenize, parse and normalise it. Secondly, while some of the hashtags are automatically extracted by Twitter and provided as metadata, this process is sometimes incomplete. Therefore, we employ additional steps to extract all hashtags from the text, if required.

All the data required for the pre-processing steps is stored into a Hadoop Distributed File System [34], a fault-resilient file system for storage, and the posterior processing is done with a Spark cluster deployed on our premises. This approach allows us to transparently allocate more hardware resources into the pre-processing steps when needed.

The processed data is saved back into our storage solution, described in the next section. Once the information extraction has been completed, the visualisation front-end is notified that a new batch of data is available. This allows us to build a feedback link between the front-end and back-end services and assure near real-time updates if needed.

3.2. Storage

In the described scenario, temporal storage of the (big) data is necessary and therefore it needs to be taken into account. While regular text files could be used, random queries over them are complicated.

In fact, databases have been the solution of choice to store data in structured and scalable ways since many decades ago. And particularly, relational databases were highly optimised to deal with this task. The last decade has seen the development of alternative models focusing on different aspects of the data or relaxing some of the constraints of relational. In particular, no-SQL databases (e.g. MongoDB) and graph databases (Neo4j) have gained great traction.

As explained in [29], the way the data is stored and made available is relevant to support different types of queries and applications, and to support efficiency and scalability of operation.

Particularly, having previous knowledge on likely queries can greatly improve the performance and speed of the database at querying time, as indexes can be pre-computed. What is more, the data can be separated into different computing instances. Precisely, one particular clustering strategy we carry out is related with events: collection of tweets within a specific data range and matching a specific search criterion. We thereafter group one or more events together and store them in separate databases based on their scale.

For events that involve a much larger footprint of data, we have explored the option of dissecting the dataset even further by storing retweet and mention networks separately. This is only ideal for some use-cases where search queries do not involve filtering based on comparisons spanning across both retweet and mention networks but produces significantly better results by reducing computation and rendering overheads. Fortunately, this approach works well for a majority of the graphs we visualise, because it is very rare that we have queries so specific that requires both retweet and mention networks to be stored in a single database.

Our implementation of the suggested pipeline relies on a distributed Neo4j database.

3.3. Display

SAGE2 [35], DisplayCluster [27], Chromium [36], CGLX [26] and OVE [37] are examples of software frameworks designed exclusively for SRDEs. Through learning and years of evolution, modern SRDE software frameworks are almost always browser-based. This therefore makes it much easier to develop software for a twitter observatory using web application technologies.

Modern SRDE frameworks such as SAGE2 and OVE support rendering of web content at a resolution that is independent and much greater than that of the actual screens themselves making it possible to display very large visualisations. They also provide built-in applications that natively support the rendering of certain types of data formats such as (a) tiled maps, (b) deep zoomable images, (c) graphs and (d) playback of audio and video. The frameworks themselves deal with the distribution of the content and synchronised rendering, animation and playback. Thanks to this kind of portability, the rich diversity of Twitter data can be harnessed.

For example, OVE supports rendering large graphs using Sigma.js [38] and Neo4J [39]. However, the larger the dataset the longer it takes to run queries, the greater the network bandwidth and system resource consumption. The greatest impact is at the client-side where browsers can accommodate only a certain portion of the dataset before crashing due to insufficient system resources. OVE overcomes this limitation in two ways:

Firstly, the dataset can be sharded and distributed among multiple Neo4J instances by partitioning the graph both horizontally and vertically. SRDEs are capable of tiling content across multiple browsers and support arrangements such as a browser per screen. OVE supports more than one browser per screen making it even more scalable. Therefore, based on the granularity and the density of the data, OVE can support one or more data-sources per screen. By sharding and distributing the dataset based on the volume of the data, each data-source (i.e. Neo4J database) will therefore have a finite and manageable demand.

And secondly, OVE also supports transparent overlays. This makes it possible to render graphs or any other content as a series of layers. It is often the case that running a database query with multiple joins tend to be highly resource-intensive and very slow. Thanks to the transparent overlay support of OVE, we no longer need to run join queries but can source data from a cluster of databases using a set of queries that can now be executed in parallel, which increases the responsiveness and also the rendering speeds. Transparent overlays also make filtering and annotating datasets much easier to implement and highly performant.

3.4. Interaction

Finally, a crucial aspect of visualisation is the ability to interact with the visual display. In a previous paper [9], we presented a fuzzy query mechanism to highlight ‘relevant’ tweets, demonstrating its potential when paired with SRDEs.

In graph data, the key interactions are filtering, pan and zooming, and they enable switching between overall and specific viewpoints, and focusing the attention of the audience towards a portion of the dataset at a time. Though not most useful in simpler graphs, this becomes very important when trying to understand large quantities of data within an immersive SRDE. Frameworks such as OVE [37] have built-in support for all operations related to interaction (in particular through OVE’s Networks app).

However, to make the application even more interactive, we also extended the default controller with a personalised one. With it, the presenter needs not key in queries in the OData format (which is the one that the OVE API expects). Instead, he/she is presented with a simplified user interface accepting various search criteria which will subsequently be converted into the OData format and then passed onto OVE.

Finally, annotations and group discussions around a pre-rendered twitter dataset are made possible using the Whiteboard App of the OVE framework [37]. Annotations can be saved and retrieved at a later time.

4. Experiments

In this section we demonstrate several of the capabilities that our existing infrastructure enables with two relevant examples: data from the Brexit referendum and data from the 2016 US presidential elections.

4.1. Brexit referendum

Firstly, we look into exploring the polarisation of users who tweeted for and against the UK leaving the EU during the period of the Brexit referendum poll. Data collected from the 6th of January 2016 until the 30th of June 2016 is used. The dataset contains a number of randomly selected tweets from around the world, out of which a clear majority originated from within the UK and Western Europe.

We assign a polarisation score *ps* on the range $[-1, 1]$ to a user based on an aggregate measure based on the presence of several hashtags (they were described in an earlier work [17]) across all of their tweets. We saturate the values to the given scale. A score of -1 means that a user is the strongest supporter for leaving the EU whilst a score of 1 means the user is the strongest supporter for remaining in the EU. All users depart from a scoring of 0 (i.e. a neutral tweet) and their preference for hashtags for and against remaining in the EU would move their score towards 1 and -1, accordingly. Users who merely retweeted an existing tweet of another user were given the score of the original tweet.

We chose to represent this data as graph, with nodes representing individual tweets, and edges capturing retweet and mention links. All tweets were located in the horizontal axis according to their creation date, and in the vertical axis according to its polarisation score at that moment. Fig. 2 presents a snapshot of this visualisation.

Retweets were assigned the *ps* of the tweet they were retweeting. Finally, to avoid visual clutter of tweets with a polarisation score of 0, we add a small random noise to them so they are placed in a small interval of 5% around 0.

The colours of red and green were chosen to create a strong contrast between each camp of users, with green associated with those in support of remaining in the EU ($ps \in (0, 1]$) and red associated with those in support of leaving ($ps \in [-1, 0)$). A neutral grey colour was given to both retweets (regardless of their *ps*) and to tweets that were not polarised.

Visually, we can clearly spot that the polarisation towards leaving the EU: leavers were marginally more strongly opinionated and supported compared to the remain camp. This is confirmed in an accompanying visual (Fig. 3 shows the daily average of the polarisation score) and is also consistent with the outcome of the Brexit referendum on the 23rd of June 2016.

We were also able to confirm our previous observations [17]; namely that a few users in twitter are found to be making a majority of the number of tweets, (some of which gain significant popularity), while a majority of the users would simply take their opinion forward by simply retweeting these popular tweets. This is a contributing factor with regards to changing and amplifying a user's opinion. It was also relevant to confirm that some of those opinion-drivers were actually bots.

4.2. 2016 US presidential elections

For a second illustrative example we used a small dataset collected around the 2016 US election, containing around 150,000 tweets. The scientific goal in this occasion is to understand the types of interactions between polarised communities. We investigate the interactions between the communities considering retweets and mentions (Fig. 4 together, and Fig. 5 separated by colours).

The users within the dataset were classified into four classes according to the support they expressed in their tweets towards either Donald Trump, Hillary Clinton, Bernie Saunders or others. We assigned a colour to each one of those classes: green for Trump, red for Hillary, purple for Saunders, and grey for the others.

Table 1
Summary of distribution of tweets among candidates. *% mentions* and *% retweets* refer to the percentage out of the total number of tweets overall.

	% tweets	% mentions	% retweets
Clinton	28.56	2.35	4.06
Saunders	19.79	3.33	3.59
Trump	36.13	1.53	3.29
Unclassified	15.52	11.89	0.07

Edges (with the colour of their origin) were drawn among users whenever there was a retweet or a mention in any of their tweets. This result is depicted in Fig. 4).

In addition to that, we also colour-coded the edges in white if they were a retweet, and yellow for mentions. This result is depicted in Fig. 5 and summarised in Table 1.

The main visual finding from both figures is that the polarised communities tend to retweet others from their own communities and contribute towards further amplifying their opinion. They cherry-picked tweets (preferably originating from one of the three candidates' twitter handles) and retweeted them in an attempt to amplify the message.

The mentions network however were not strictly limited to the polarised communities and it was observed that there was a lot of cross-talk between those tweeting (the yellow edges in Fig. 5 are more prominent between clusters). We believe that these are examples of users making an attempt to point out weaknesses of an opposing individual or to express hatred or negative feedback.

Observations such as above make it easier to understand patterns of information propagation among polarised micro-communities and as well as the ability to study their opinions or even identify key individuals within each polarised community based on the popularity of their tweets.

5. Conclusions

This work has presented our proposal for a real-time large-scale twitter observatory as an advance platform for social analytics on social media. Our proposal leverages on existing technologies (and particularly OVE, our distributed visualisation framework) and put forward a connected pipeline.

Our tests have confirmed that our design (including visualisation) scales to suit graphs involving a few billion data points from Twitter. The OVE framework is designed to be scalable to support even larger graphs. This is by clustering the OVE core server and also the corresponding applications such as the Networks App. Though we have not attempted to so far, we believe that this makes it possible to render graphs that are ten-fold or even 100-fold larger, which will be equivalent to visualising a twitter dataset involving 500 billion tweets over a 3 year period.

Although the work we have described here is focused on political events, the proposed observatory could be easily extended to other areas of interest. As literature has shown often, one-size-for-all solutions are rare and hence we strongly believe that a problem analysis and work with experts is needed to finely adapt the system to specific users' requirements. This is particularly true in research environments (as the one our observatory aims to support).

An additional advantage of a facility such as Imperial's Data Observatory (which we have not specially remarked in this paper) is its social facet which naturally fosters the interactions between teams, and facilitates the collaborative evaluation and exploration of graphs and patterns. Joint data exploration by members of a team is another way of gaining deeper insights on the data, compared with doing it alone.

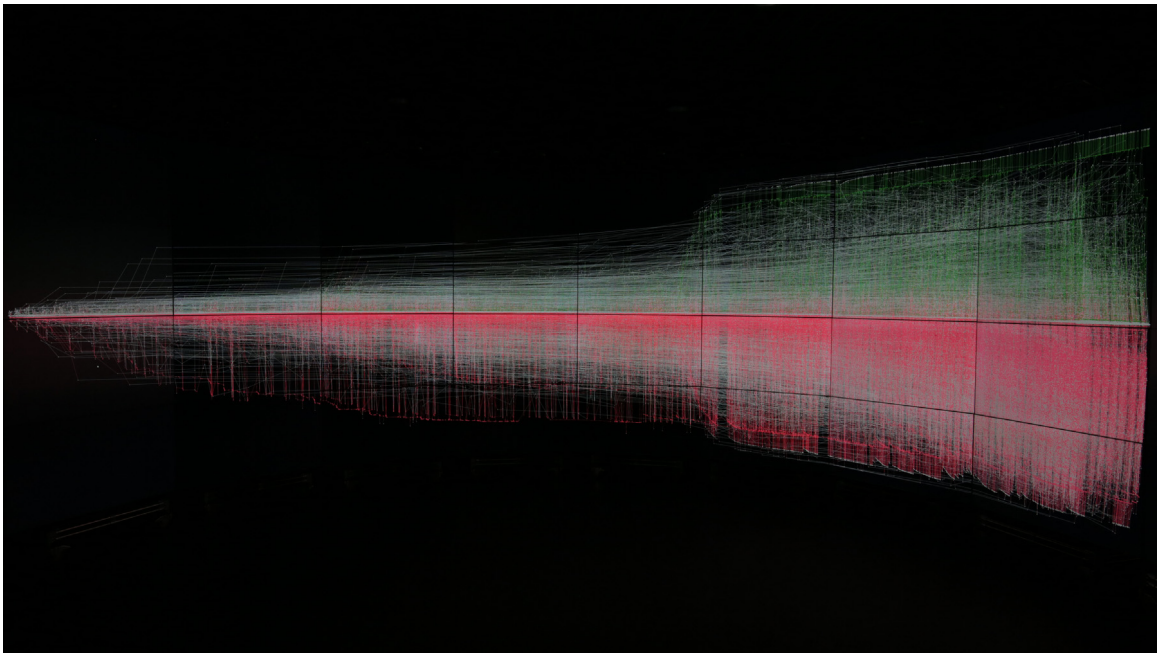

Fig. 2. Photograph of the result of displaying a timeline with polarisation of tweets for the Brexit referendum dataset rendered in a 15 360 × 4320 resolution. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

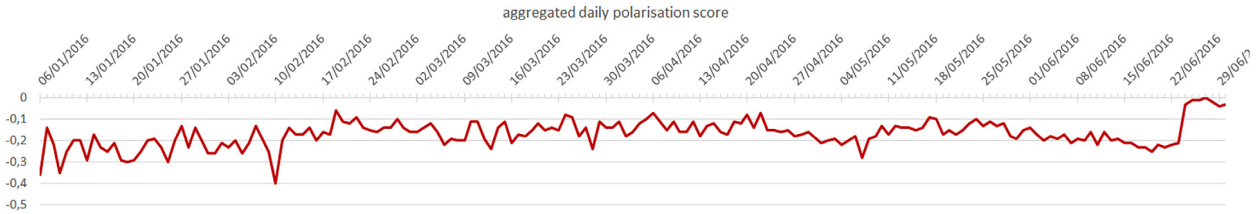

Fig. 3. Aggregated daily polarisation score as the average of all tweets of one day. During the whole period, the polarisation is stronger towards the leave side.

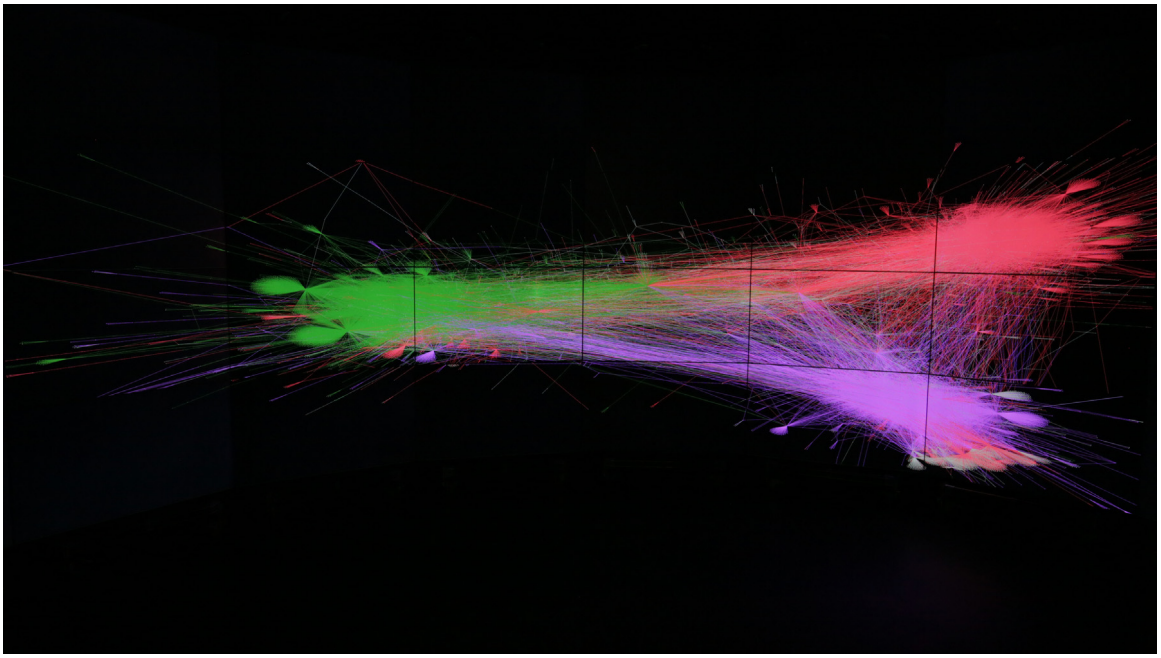

Fig. 4. Visual result of interactions (retweets and mentions) between communities during the 2016 US presidential election. Colours correspond to the supported candidate. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

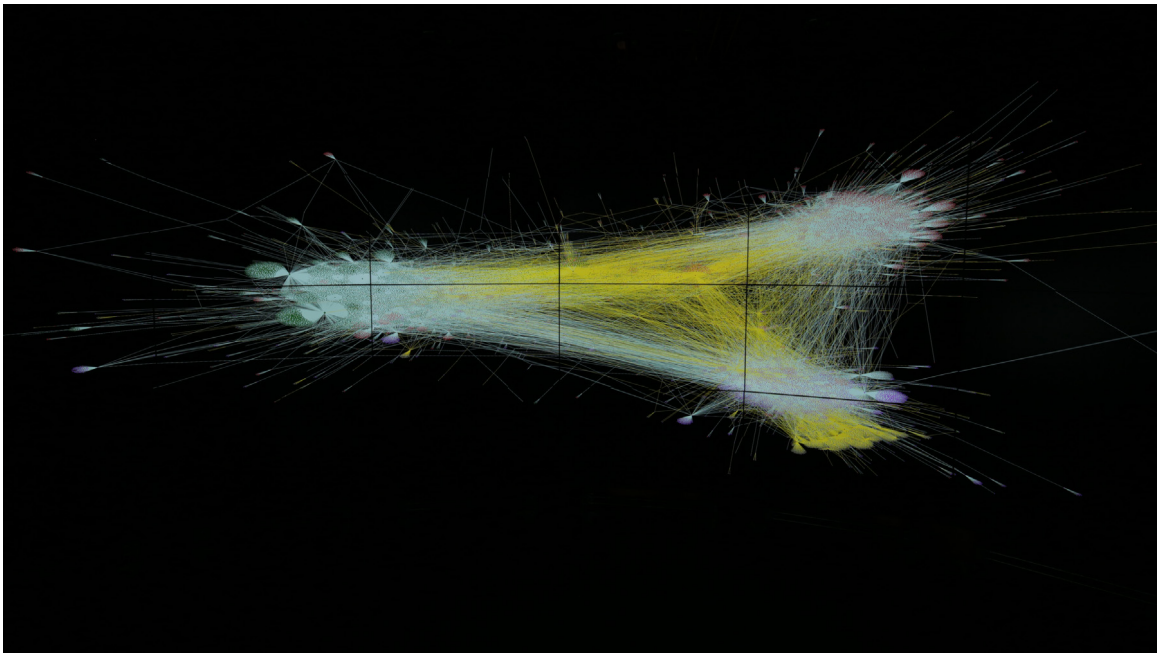

Fig. 5. Visual result of interactions (retweets in blue, mentions in yellow) between communities during the 2016 US presidential election. Users (nodes) remain in the same position as in Fig. 4. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

While the proposed framework is flexible enough to account for both historical and real-time data, our current implementation only covers the former. Supporting real-time analytics and visualisation is a capability that will benefit monitoring tasks, such as those of spotting abnormal activity, political topic/sentiment changes, etc.

Regarding the tool, we have only presented here the visualisation of historical data in a networked way. The observatory will not be complete until other summarising analysis and visualisation dashboards around the dataset can be displayed along the graph data. While this capability exist, we are still missing the automatic connection between both pieces.

We are also exploring the opportunities for more ecological and immersive interaction interfaces (after all, a large visualisation facility call for different ways of interacting with the data), such as gesture-based feedback, voice-controlled querying/filtering, and a touch user interface allowing users to select, for instance, specific data points in a visualisation and learn more about its interconnections or extract a portion of a graph into a separate partition of the space.

It is possible to implement the full pipeline described here with open-source solutions, using the description provided. All code implementing the visualisation part of the Twitter observatory (i.e. the OVE framework) is open-source and available at [37]. We invite interested readers to check it out, experiment with it and contribute to its development.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors acknowledges Chris Snowden's MSc project for his initial ideas on visualising the Brexit dataset. M. Molina-Solana was funded by the European Union's H2020 R&I programme under the Marie Sklodowska—Curie grant agreement No. 743623.

References

[1] D. Spina, A. Zubiaga, A. Sheth, M. Strohmaier, Processing social media in real-time, Inf. Process. Manage. 56 (3) (2019) 1081–1083, http://dx.doi.org/10.1016/j.ipm.2018.06.006.
[2] S.D. Roy, W. Zeng, Social Multimedia Signals, Springer, 2015, http://dx.doi.org/10.1007/978-3-319-09117-4.
[3] J. Ausserhofer, A. Maireder, National politics on twitter, Inf. Commun. Soc. 16 (3) (2013) 291–314, http://dx.doi.org/10.1080/1369118X.2012.756050.
[4] C. Piña-García, J.M. Siqueiros-García, E. Robles-Belmont, G. Carreón, C. Gershenson, J.A.D. López, From neuroscience to computer science: a topical approach on twitter, J. Comput. Soc. Sci. 1 (1) (2018) 187–208, http://dx.doi.org/10.1007/s42001-017-0002-9.
[5] X. Lu, C. Brelsford, Network structure and community evolution on twitter: Human behavior change in response to the 2011 Japanese earthquake and tsunami, Sci. Rep. 4 (2014) http://dx.doi.org/10.1038/srep06773.
[6] E. Ferrara, P.D. Meo, G. Fiumara, R. Baumgartner, Web data extraction, applications and techniques: A survey, Knowl.-Based Syst. 70 (2014) 301–323, http://dx.doi.org/10.1016/j.knosys.2014.07.007.
[7] S. González-Bailón, J. Borge-Holthoefer, A. Rivero, Y. Moreno, The dynamics of protest recruitment through an online network, Sci. Rep. 1 (2011) http://dx.doi.org/10.1038/srep00197.
[8] D. McGinn, D. Birch, D. Akroyd, M. Molina-Solana, Y.-k. Guo, W. Knottenbelt, Visualizing dynamic Bitcoin transaction patterns, Big Data 4 (2) (2016) 109–119, http://dx.doi.org/10.1089/big.2015.0056.
[9] M. Molina-Solana, D. Birch, Y.-k. Guo, Improving data exploration in graphs with fuzzy logic and large-scale visualisation, Appl. Soft Comput. 53 (2017) 227–235, http://dx.doi.org/10.1016/j.asoc.2016.12.044.
[10] O. Şerban, N. Thapen, B. Maginnis, C. Hankin, V. Foot, Real-time processing of social media with SENTINEL: A syndromic surveillance system incorporating deep learning for health classification, Inf. Process. Manage. 56 (3) (2019) 1166–1184, http://dx.doi.org/10.1016/j.ipm.2018.04.011.
[11] J. Gómez-Romero, M. Molina-Solana, A. Oehmichen, Y. Guo, Visualizing knowledge graphs: A performance analysis, Future Gener. Comput. Syst. 89 (2018) 224–238, http://dx.doi.org/10.1016/j.future.2018.06.015.
[12] J. Amador Díaz López, C. Piña-García, Political participation in Mexico through twitter, in: Procs. International Workshop on Complex Networks and their Applications, Springer, 2016, pp. 607–618.
[13] Statista, Number of monthly active twitter users worldwide, https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/.
[14] A. Tumasjan, T.O. Sprenger, P.G. Sandner, I.M. Welpe, Election forecasts with twitter: How 140 characters reflect the political landscape, Soc. Sci. Comput. Rev. 29 (4) (2011) 402–418, http://dx.doi.org/10.1177/0894439310386557.
[15] K. McKelvey, J. DiGrazia, F. Rojas, Twitter publics: how online political communities signaled electoral outcomes in the 2010 US house election, Inf. Commun. Soc. 17 (4) (2014) 436–450, http://dx.doi.org/10.1080/1369118X.2014.892149.

[16] F. Franch, (Wisdom of the crowds)^2: 2010 UK election prediction with social media, J. Inf. Technol. Polit. 10 (1) (2013) 57–71, http://dx.doi.org/10.1080/19331681.2012.705080.
[17] J. Amador Diaz Lopez, S. Collignon-Delmar, K. Benoit, A. Matsuo, Predicting the brexit vote by tracking and classifying public opinion using twitter data, Stat. Polit. Policy 8 (1) (2017) 85–104, http://dx.doi.org/10.1515/spp-2017-0006.
[18] E.R. Tufte, The Visual Display of Quantitative Information, second ed., Graphics Press, 2001.
[19] B.H. McCormick, Visualization in scientific computing, ACM SIGBIO Newslett. 10 (1) (1988) 15–21, http://dx.doi.org/10.1145/43965.43966.
[20] D.A. Keim, Visual exploration of large data sets, Commun. ACM 44 (8) (2001) 38–44, http://dx.doi.org/10.1145/381641.381656.
[21] B. Shneiderman, The eyes have it: a task by data type taxonomy for information visualizations, in: Procs. 1996 IEEE Symposium on Visual Languages, 1996, pp. 336–343, http://dx.doi.org/10.1109/VL.1996.545307.
[22] C. Cruz-Neira, D.J. Sandin, T.A. DeFanti, R.V. Kenyon, J.C. Hart, The CAVE: audio visual experience automatic virtual environment, Commun. ACM 35 (6) (1992) 64–73.
[23] A. Febretti, A. Nishimoto, T. Thigpen, J. Talandis, L. Long, J. Pirtle, T. Peterka, A. Verlo, M. Brown, D. Plepys, et al., Cave2: a hybrid reality environment for immersive simulation and information analysis, in: Is&T/Spie Electronic Imaging, International Society for Optics and Photonics, 2013, p. 864903.
[24] C. Papadopoulos, K. Petkov, A. Kaufman, K. Mueller, The reality deck - immersive gigapixel display, IEEE Comput. Graph. Appl. 35 (1) (2015) 33–45, http://dx.doi.org/10.1109/MCG.2014.80.
[25] K. Li, M. Hibbs, G. Wallace, O. Troyanskaya, Dynamic scalable visualization for collaborative scientific applications, in: Proc. 19th IEEE Int. Parallel and Distributed Processing Symposium, IIPDPS'05, IEEE Computer Society, 2005, http://dx.doi.org/10.1109/IPDPS.2005.183.
[26] K.-U. Doerr, F. Kuester, CGLX: a scalable, high-performance visualization framework for networked display environments, IEEE Trans. Vis. Comput. Graphics 17 (3) (2011) 320–332, http://dx.doi.org/10.1109/TVCG.2010.59.
[27] G.P. Johnson, G.D. Abram, B. Westing, P. Navrátil, K. Gaither, Displaycluster: An interactive visualization environment for tiled displays, in: Cluster Computing (CLUSTER), 2012 IEEE International Conference on, IEEE, 2012, pp. 239–247, http://dx.doi.org/10.1109/CLUSTER.2012.78.
[28] J. Roberts, P. Ritsos, S. Badam, D. Brodbeck, J. Kennedy, N. Elmqvist, Visualization beyond the desktop–the next big thing, IEEE Comput. Graph. Appl. 34 (6) (2014) 26–34, http://dx.doi.org/10.1109/MCG.2014.82.
[29] I. Basaille, S. Kirgizov, E. Leclercq, M. Savonnet, N. Cullot, Towards a twitter observatory: A multi-paradigm framework for collecting, storing and analysing tweets, in: 2016 IEEE Tenth International Conference on Research Challenges in Information Science, RCIS, 2016, pp. 1–10, http://dx.doi.org/10.1109/RCIS.2016.7549324.
[30] D.S. Banerjee, A. Kumar, M. Chaitanya, S. Sharma, K. Kothapalli, Work efficient parallel algorithms for large graph exploration on emerging heterogeneous architectures, J. Parallel Distrib. Comput. 76 (2015) 81–93, http://dx.doi.org/10.1016/j.jpdc.2014.11.006.
[31] A.-S. Dadzoe, M. Rowe, Approaches to visualising linked data: A survey, Semant. Web 2 (2) (2011) 84–124, http://dx.doi.org/10.3233/SW-2011-0037.
[32] M. Bastian, S. Heymann, M. Jacomy, Gephi: An open source software for exploring and manipulating networks, in: Procs. 3rd International Conference on Weblogs and Social Media, ICWSM, 2009, pp. 361–362.
[33] M.L. Williams, P. Burnap, L. Sloan, Towards and ethical framework for publishing twitter data in social research: taking into account users' views, online context and algorithmic estimation, Sociology 51 (6) (2017) 1149–1168, http://dx.doi.org/10.1177/0038038517708140.
[34] K. Shvachko, H. Kuang, S. Radia, R. Chansler, The hadoop distributed file system, in: Proc. 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies, MSST, 2010, pp. 1–10, http://dx.doi.org/10.1109/MSST.2010.5496972.
[35] L. Renambot, T. Marrinan, J. Aurisano, A. Nishimoto, V. Mateevitsi, K. Bharadwaj, L. Long, A. Johnson, M. Brown, J. Leigh, SAGE2: A collaboration portal for scalable resolution displays, Future Gener. Comput. Syst. 54 (2016) 296–305, http://dx.doi.org/10.1016/j.future.2015.05.014.
[36] G. Humphreys, M. Houston, R. Ng, R. Frank, S. Ahern, P.D. Kirchner, J.T. Klosowski, Chromium: a stream-processing framework for interactive rendering on clusters, ACM Trans. Graph. 21 (3) (2002) 693–702, http://dx.doi.org/10.1145/566654.566639.
[37] Data Science Institute, Imperial College London, OVE - Open visualisation environment, https://github.com/ove.
[38] A. Jacomy, G. Plique, Sigma.js, http://sigmajs.org/.
[39] J. Webber, A programmatic introduction to neo4j, in: Procs. 3rd Annual Conference on Systems, Programming, and Applications: Software for Humanity, ACM, 2012, pp. 217–218.

Senaka Fernando is a Ph.D. candidate and a Research Assistant at the Data Science Institute of the Imperial College London. Senaka is a Computer Science and Engineering graduate with over 10 years working experience in the industry in the distributed enterprise middleware domain. At the Data Science Institute Senaka is also the Systems Architect of the Data Observatory and the main contributor its Open Visualisation Environment (OVE) project. Senaka's research interests involve data visualisation in large high resolution display environments.

Julio Amador is a Research Fellow at Imperial College London and holds a Ph.D. in Economics from the University of Essex. His area of expertise is applied machine learning (ML). Julio has held different research positions, both in the UK and abroad. His research includes big-data studies of online political participation and applying ML to categorise public opinion and automatically identifying fake news. Julio is currently dedicated to the study of misinformation.

Ovidiu Şerban (ʃerban) is a Research Associate at the Data Science Institute (DSI), Imperial College London. His current work includes real-time Natural Language Processing and Large Scale Visualisation Systems. Ovidiu's research topics are Natural Language Processing, Machine Learning, Affective Computing and Interactive System Design. He holds a joint Ph.D. from Normandy University (France) and "Babeş-Bolyai" University (Romania), while working at LITIS Laboratory in France.

Juan Gómez-Romero is a Research Fellow at the Computer Science and Artificial Intelligence department of Universidad de Granada since 2013. He received his degree in Computer Science (2004) and his Ph.D. (2008) in Intelligent Systems from the same university. He worked as a visiting professor in the Applied Artificial Intelligence Group of Universidad Carlos III de Madrid from 2008 to 2013. His research interests focus on the use of semantic representation models and machine learning techniques to perform automatic reasoning towards higher-level information fusion. He has participated in more than 20 projects in the areas of security, ambient intelligence and energy efficiency.

Miguel Molina-Solana is a Marie Curie Research Fellow at the Data Science Institute (DSI) at Imperial College London, working on the DATASOUND project. Before, he was a Research Associate at the DSI, and a postdoc researcher at the Department of Computer Science and Artificial Intelligence of University of Granada. He holds a Ph.D. and a M.Sc. in Computer Science from University of Granada, and a M.Sc. in Soft Computing and Intelligent Systems. His research experience comprises work in the areas of Data Mining, Machine Learning and Knowledge representation applied in different areas such as Music, Energy management and Healthcare.

Yike Guo is a Professor of Computing Science in the Department of Computing at Imperial College London. He is the founding Director of the Data Science Institute at Imperial College, as well as leading the Discovery Science Group in the department. Professor Guo also holds the position of CTO of the tranSMART Foundation, a global open source community using and developing data sharing and analytics technology for translational medicine.