OSoMe: The IUNI Observatory on Social Media

Clayton A. Davis[*1,2], Giovanni Luca Ciampaglia[1,3], Luca Maria Aiello[†4], Keychul Chung[2], Michael Conover[†5], Emilio Ferrara[†6], Alessandro Flammini[1,2,3], Geoffrey Fox[2], Xiaoming Gao[†7], Bruno Gonçalves[†8], Przemyslaw Grabowicz[†9], Alex Hong[2], Pik-Mai Hui[2], Scott McCaulay[3], Karissa McKelvey[†10], Mark Meiss[†11], Snehal Patil[†4], Chathuri Peli Kankanamalage[3], Valentin Pentchev[3], Judy Qiu[2], Jacob Ratkiewicz[†11], Alex Rudnick[†11], Benjamin Serrette[3], Prashant Shiralkar[1,2], Onur Varol[1,2], Lilian Weng[†12], Tak-Lon Wu[†13], Andrew Younge[2], and Filippo Menczer[1,2,3]

[1]*Center for Complex Networks and Systems Research, Indiana University, USA*
[2]*School of Informatics and Computing, Indiana University, USA*
[3]*Network Science Institute, Indiana University, USA*
[4]*Yahoo! Inc, USA*
[5]*LinkedIn Inc, USA*
[6]*Information Sciences Institute, University of Southern California, USA*
[7]*Facebook Inc, USA*
[8]*Center for Data Science, New York University, USA*
[9]*Max Planck Institute for Software Systems, Germany*
[10]*US Open Data, USA*
[11]*Google Inc, USA*
[12]*Affirm Inc, USA*
[13]*Amazon Inc, USA*

[*]Corresponding author: claydavi@umail.iu.edu
[†]Work done at Indiana University.

Abstract

The study of social phenomena is becoming increasingly reliant on big data from online social networks. Broad access to social media data, however, requires software development skills that not all researchers possess. Here we present the IUNI Observatory on Social Media, an open analytics platform designed to facilitate computational social science. The system leverages a historical, ongoing collection of over 70 billion public messages from Twitter. We illustrate a number of interactive open-source tools to retrieve, visualize, and analyze derived data from this collection. The Observatory, now available at `osome.iuni.iu.edu`, is the result of a large, six-year collaborative effort coordinated by the Indiana University Network Science Institute.

Introduction

The collective processes of production, consumption, and diffusion of information on social media are starting to reveal a significant portion of human social life, yet scientists struggle to get access to data about it. Recent research has shown that social media can perform as 'sensors' for collective activity at multiple scales (Lazer et al., 2009). As a consequence, data extracted from social media platforms are increasingly used side-by-side with — and sometimes even replacing — traditional methods to investigate hard-pressing questions in the social, behavioral, and economic (SBE) sciences (King, 2011; Moran et al., 2014; Einav and Levin, 2014). For example, interpersonal connections from Facebook have been used to replicate the famous experiment by Travers and Milgram (1969) on a global scale (Backstrom et al., 2012); the emotional content of social media streams has been used to estimate macroeconomic quantities in country-wide economies (Bollen et al., 2011; Choi and Varian, 2012; Antenucci et al., 2014); and imagery from Instagram has been mined (De Choudhury et al., 2013; Andalibi et al., 2015) to understand the spread of depression among teenagers (Link et al., 1999).

A significant amount of work about information production, consumption, and diffusion has been thus aimed at modeling these processes and empirically discriminating among models of mechanisms driving the spread of memes on social media networks such as Twitter (Guille et al., 2013). A set of research questions relate to how social network structure, user interests, competition for finite attention, and other factors affect the manner in which information is disseminated and why some ideas cause viral explosions while others are quickly forgotten. Such questions have been address both in an empirical and in more theoretical terms.

Examples of empirical works concerned with these questions include geographic and temporal patterns in social movements (Conover et al., 2013b,a; Varol et al., 2014), the polarization of online political discourse (Conover et al., 2011b,a, 2012), the use of social media data to predict election outcomes (DiGrazia et al., 2013) and stock market movements (Bollen et al., 2011), the geographic diffusion of trending topics (Ferrara et al., 2013), and the lifecycle of information in the attention economy (Ciampaglia et al., 2015).

On the more theoretical side, agent-based models have been proposed to explain how limited individual attention affects what information we propagate (Weng et al., 2012), what social connections we make (Weng et al., 2013b), and how the structure of social and topical networks can help predict which memes are likely to become viral (Weng et al., 2013a, 2014; Nematzadeh et al., 2014; Weng and Menczer, 2015).

Broad access by the research community to social media platforms is, however, limited by a host of factors. One obvious limitation is due to the commercial nature of these services. On these platforms, data are collected as part of normal operations, but this is seldom done keeping in mind the needs of researchers. In some cases researchers have been allowed to harvest data through programmatic interfaces, or APIs. However, the information that a single researcher can gather through an API typically offers only a limited view of the phenomena under study; access to historical data is often restricted or unavailable (Zimmer, 2015). Moreover, these samples are often collected using ad-hoc procedures, and the statistical biases introduced by these practices are only starting to be understood (Morstatter et al., 2013; Ruths and Pfeffer, 2014; Hargittai, 2015).

A second limitation is related to the ease of use of APIs, which are usually meant for software developers, not researchers. While researchers in the SBE sciences are increasingly acquiring software development skills (Terna et al., 1998; Raento et al., 2009; Healy and Moody, 2014), and intuitive user interfaces are becoming more ubiquitous, many tasks remain challenging enough to hinder research advances. This is especially true for those tasks related to the application of fast visualization techniques.

A third, important limitation is related to user privacy. Unfettered access to sensitive, private data about the choices, behaviors, and preferences of individuals is happening at an increasing rate (Tene and Polonetsky, 2012). Coupled with the possibility to manipulate the environment presented to users (Kramer et al., 2014), this has raised in more than one occasion deep ethical concerns in both the public and the scientific community (Kahn et al., 2014; Fiske and Hauser, 2014; Harriman and Patel, 2014; Vayena et al., 2015).

These limitations point to a critical need for opening social media platforms to researchers in ways that are both respectful of user privacy requirements and aware of the needs of SBE researchers. In the absence of such systems, SBE researchers will have to increasingly rely on closed or opaque data sources, making it more difficult to reproduce and replicate findings — a practice of increasing concern given recent findings about replicability in the SBE sciences (Open Science Collaboration, 2015).

Our long-term goal is to enable SBE researchers and the general public to openly access relevant social media data. As a concrete milestone of our project, here we present an *Observatory on Social Media* — an open infrastructure for sharing public data about information that is spread and collected through online social networks. Our initial focus has been on Twitter as a source of public microblogging posts. The infrastructure takes care of storing, indexing, and analyzing public collections and historical archives of big social data; it does so in an easy-to-use way, enabling broad access from scientists and other stakeholders, like journalists and the general public. We envision that data and analytics from social media will be integrated within a nation-wide network of social observatories. These data centers would allow access to a broad range of data about social, behavioral, and economic phenomena nationwide (King, 2011; Moran et al., 2014; Difranzo et al., 2014).

Our team has been working toward this vision since 2010, when we started collecting public tweets to visualize, analyze, and model meme diffusion networks.[1] The IUNI Observatory on Social Media (OSoMe) presented here is developed through a collaboration between the Indiana University Network Science Institute (IUNI, `iuni.iu.edu`), the IU School of Informatics and Computing (SoIC, `soic.indiana.edu`), and the Center for Complex Networks and Systems Research (CNetS, `cnets.indiana.edu`). It is available at `osome.iuni.iu.edu`.

Data Source

Social media data possess unique characteristics. Besides rich textual content, explicit information about the originating social context is generally available. Information often includes timestamps, geolocations, and interpersonal ties. The Twitter dataset is a prototypical example (McKelvey and Menczer, 2013b,a). The Observatory on Social Media

[1]The website `truthy.indiana.edu` was created to host our first demo, motivated by the application of social media analytics to the study of "astroturf," or artificial grassroots social media campaigns orchestrated through fake accounts and social bots (Ratkiewicz et al., 2011b). The *Truthy* nickname was later adopted in the media to refer to the entire project.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2008v1 | CC-BY 4.0 Open Access | rec: 29 Apr 2016, publ: 29 Apr 2016

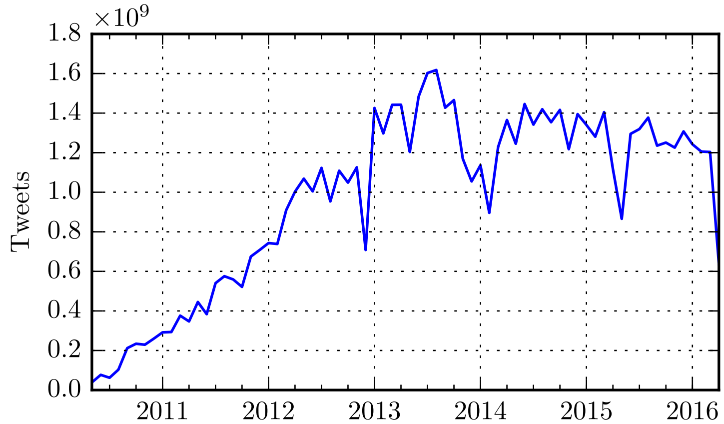

Figure 1: Number of monthly messages collected and indexed by OSoMe. System failures have caused occasional interruptions of the collection system.

is built around a Terabyte-scale historical (and ongoing) collection of approximately **70 billion public tweets** to date. The data has been collected from a random 10% stream sample of public Twitter posts and dates back to mid 2010.[2] The high-speed stream from which the data originates has a rate that ranges in the order of $10^6 - 10^8$ tweets/day. Figure 1 illustrates the growth of the Twitter collection over time.

System Architecture

Performing analytics at this scale presents specific challenges. The most obvious has to do with the design of a suitable architecture for processing such a large volume of data. This requires a scalable storage substrate and efficient query mechanisms.

The architecture the Observatory builds upon the Apache Big Data Stack (ABDS) framework (Jha et al., 2014; Qiu et al., 2014; Fox et al., 2014). Development has been driven over the years by the need for increasingly demanding social media analytics applications (Gao et al., 2011; Gao and Qiu, 2013, 2014; Gao et al., 2014, 2015; Wu et al., 2016). A key idea behind our enhancement of the ABDS architecture is the shift from standalone systems to modules; multiple modules can be used within existing software ecosystems. In particular, we have focused our efforts on enhancing two well-known Apache modules, Hadoop (The Apache Software Foundation, 2016b) and HBase (The Apache Software Foundation, 2016a).

The architecture is illustrated in Figure 2. The *data collection system* receives data from the Twitter Streaming API. Data are first stored on a temporary location and then loaded into a distributed storage layer on a daily basis. At the same time, *long-term backups* are stored on tape to allow recovery in case of data loss or catastrophic events.

The design of the *NoSQL distributed DB* module was guided by the observation that queries of social media data often involve unique constraints on the textual and social context such as temporal or network information. To address this issue, we leveraged

[2]Research based on this data was deemed exempt from review by the Indiana University IRB under Protocol #1102004860.

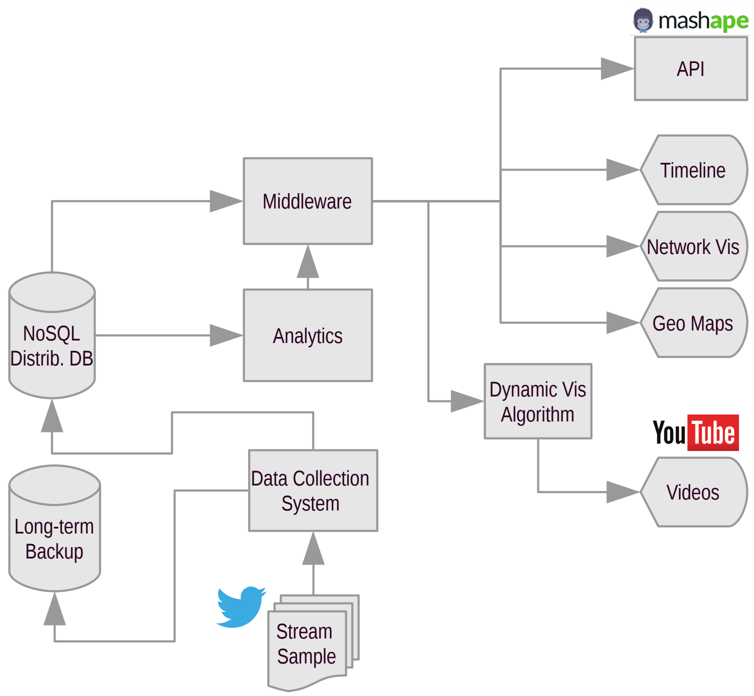

Figure 2: Flowchart diagram of the OSoMe architecture. Arrows indicate flow of data.

the HBase system as the storage substrate and extended it with a flexible indexing framework. The resulting *IndexedHBase* module (Wiggins et al., 2016) allows one to define fully customizable text index structures that are not supported by current state-of-the-art text indexing systems, such as Solr (The Apache Software Foundation, 2016c). The custom index structures can embed contextual information necessary for efficient query evaluation.

The pipelines commonly used for social media data analysis consist of multiple algorithms with varying computation and communication patterns. For example, building the network of retweets of a given hashtag will take more time and computational resources than just counting the number of posts containing the hashtag. Moreover, the temporal resolution and aggregation windows of the data could vary dramatically, from seconds to years. A number of different processing frameworks could be needed to perform such a wide range of tasks. To design the *analytics* module of the Observatory we choose Hadoop, a standard framework for Big Data analytics. We use YARN (The Apache Software Foundation, 2016d) to achieve efficient execution of the whole pipeline, and integrate it with IndexedHBase. An advantage deriving from this choice is that the overall software stack can dynamically adopt different processing frameworks to complete heterogeneous tasks of variable size.

A distributed message-passing task queue, and an in-memory key/value store implement the *middleware* layer needed to connect the backend of the Observatory with the frontend apps. We use Celery (Solem and Contributors, 2016) and RabbitMQ (Pivotal Software, Inc, 2016) to implement such layer.

The Observatory user interface follows a modular architecture too, and is based on a number of apps, which we describe in greater detail in the following section. Three of the apps (*Timeline*, *Network visualization*, and *Geographic maps*) are directly accessible within OSoMe through Web interfaces. We rely on the popular video-sharing service YouTube for the fourth app, which generates meme diffusion movies (*Videos*) using a fast *dynamic visualization algorithm* (Grabowicz et al., 2014) specifically designed for temporal networks. Finally, the Observatory provides access to raw data via a programmatic interface (*API*).

Applications

Storing and indexing tens of billions of tweets is of course pointless without a way to make sense of such a huge trove of information. The Observatory lowers the barrier of entry to social media analysis by providing users with several ready-to-use, Web-based data visualization tools. Visualization techniques allow users to make sense of complex data and patterns (Card, 2009), and let them explore the data and try different visualization parameters (Rafaeli, 1988). In the following, we give a brief overview of the available tools.

It is important to note that, in compliance with the Twitter terms of service (Twitter, Inc., 2016), OSoMe does not provide access to the content of tweets. However, researchers can obtain numeric object identifiers in response to their queries. This information can then be used to retrieve tweet content via the official Twitter API.

Temporal Trends

The *Trends* tool produces time series plots of the number of tweets including one or more given hashtags; it can be compared to the service provided by Google Trends, which allows users to examine the interest toward a topic reflected by the volume of search queries submitted to Google over time.

Users may specify multiple terms in one query, in which case all tweets containing any of the terms will be computed; and they can perform multiple queries, to allow comparisons between different topics. For example, let us compare the relative tweet volumes about the World Series and the Superbowl. We want our Super Bowl timeline to count tweets containing any of #SuperBowl, #SuperBowl50, or #SB50. Since hashtags are case-insensitive and we allow trailing wildcards, this query would be "`#superbowl*, #sb50`." Adding a timeline for the "`#worldseries`" query results in the plot seen in Figure 3. Each query on the Trends tool takes on the order of five seconds; this makes the tool especially suitable for interactive exploration of Twitter conversation topics.

Diffusion and Co-occurrence Networks

In a diffusion network, nodes represent users and an edge drawn between any two nodes indicates an exchange of information between those two users. For example, a user could rebroadcast (*retweet*) the status of another user to her followers, or she could address another user in one of her statuses by including a mention to their user handle (*mention*). Edges have a weight to represent the number of messages connecting two nodes. They may also have an intrinsic direction to represent the flow of information. For example, in the retweet network for the hashtag #IceBucketChallenge, an edge from user i to user j indicates that j retweeted tweets by i containing the hashtag #IceBucketChallenge. Similarly, in a mention network, an edge from i to j indicates that

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2008v1 | CC-BY 4.0 Open Access | rec: 29 Apr 2016, publ: 29 Apr 2016

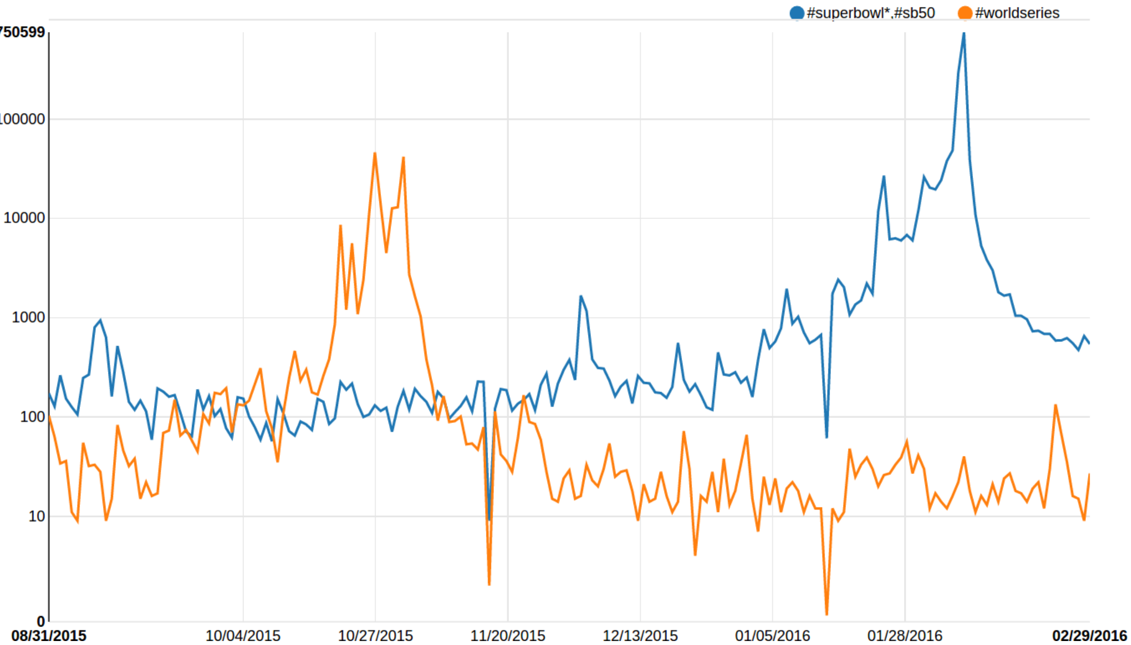

Figure 3: Number of tweets per day about the Super Bowl (in blue) and the World Series (in orange), from September 2015 through February 2016. The Y-axis is in logarithmic scale, shifted by one to account for null counts. The plot shows two outages in the data collection that occurred around mid-November 2015 and mid-January 2016.

i mentioned *j* in tweets containing the hashtag. Information diffusion network, sometimes also called information cascades, have been the subject of intense study in recent years (Gruhl et al., 2004; Weng et al., 2012; Bakshy et al., 2012; Weng et al., 2013b,a; Romero et al., 2011).

Another type of network visualizes how hashtags co-occur with each other. Co-occurrence networks are also weighted, but undirected: nodes represent hashtags, and the weight of an edge between two nodes is the number of tweets containing both of those hashtags.

OSoMe provides two tools that allow users to explore diffusion and and co-occurrence networks.

Interactive Network Visualization

The *Networks* tool enables the visualization of how a given hashtag spreads through the social network via retweets and mentions (Figure 4) or what hashtags co-occur with a given hashtag. The resulting network diagrams, created using a force-directed layout (Kamada and Kawai, 1989), can reveal topological patterns such as influential or highly-connected users and tightly-knit communities. Users can click on the nodes and edges to find out more information about the entities displayed — users, tweets, retweets, and mentions — directly from Twitter. Network are cached to enable fast access to previously-created visualizations.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2008v1 | CC-BY 4.0 Open Access | rec: 29 Apr 2016, publ: 29 Apr 2016

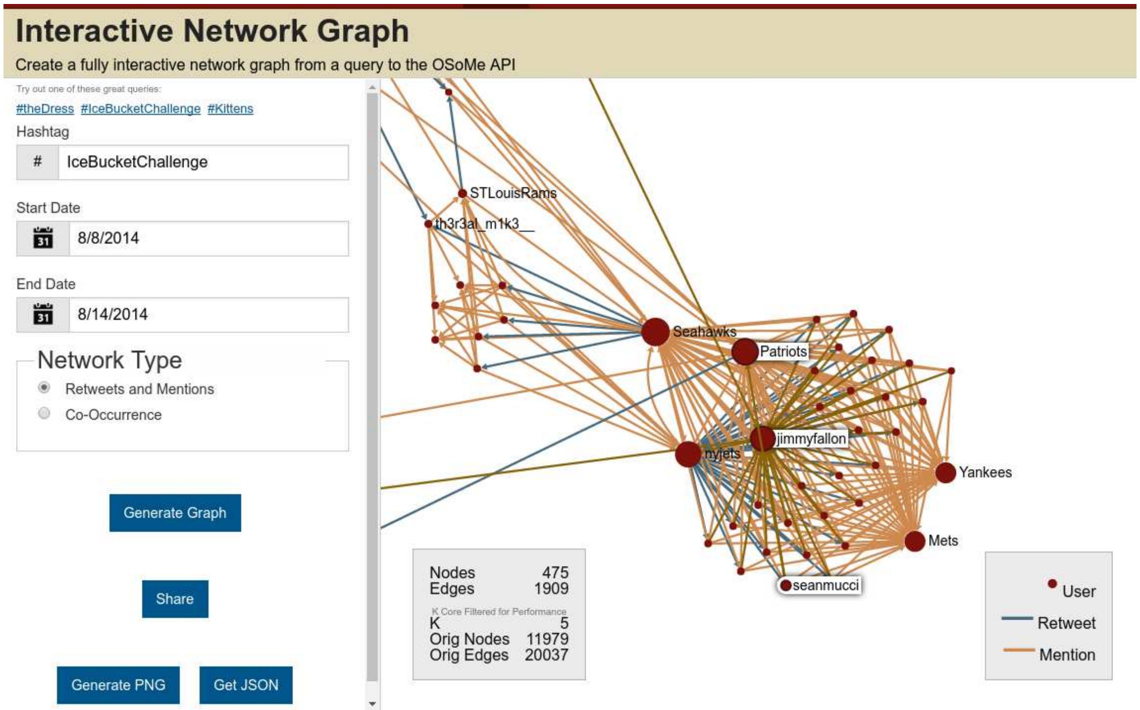

Figure 4: Interactive Network Visualization Tool. A detail of the network of retweets and mention for a hashtag commonly linked to "Ice Bucket Challenge," a popular Internet phenomenon from 2014. The size of a node is proportional to its strength (weighted degree). For visualization purposes, the size of large networks is reduced by extracting their k-core (Alvarez-Hamelin et al., 2005) with k sufficiently large to display 1,000 nodes or less ($k = 5$ in this example). The detail shows the patterns of mention and information broadcasting occurring between celebrities, as the viral challenge was taking off.

Animations

Because tweet data are time resolved, the evolution of a diffusion or co-occurrence network can be also visualized over time. Currently the *Networks* tool visualizes only static networks aggregated over the entire search period specified by the user; we aim to add the ability to observe the network evolution over time, but in the meantime we also provide the *Movies* tools, an alternative service that lets users generate animations of such processes (Figure 5). We have successfully experimented with fast visualization techniques in the past, and have found that edge filtering is the best approach for efficiently visualizing networks that undergo a rapid churn of both edges and nodes. We have therefore deployed a fast filtering algorithm developed by our team (Grabowicz et al., 2014). The user-generated videos are uploaded to YouTube, and we cache the videos in case multiple users try to visualize the same network.

Geographic maps

Online social networks are implictly embedded in space, and the spatial patterns of information spread have started to be investigated in recent years (Ferrara et al., 2013; Conover et al., 2013a). The *Maps* tool enables the exploration of information diffusion

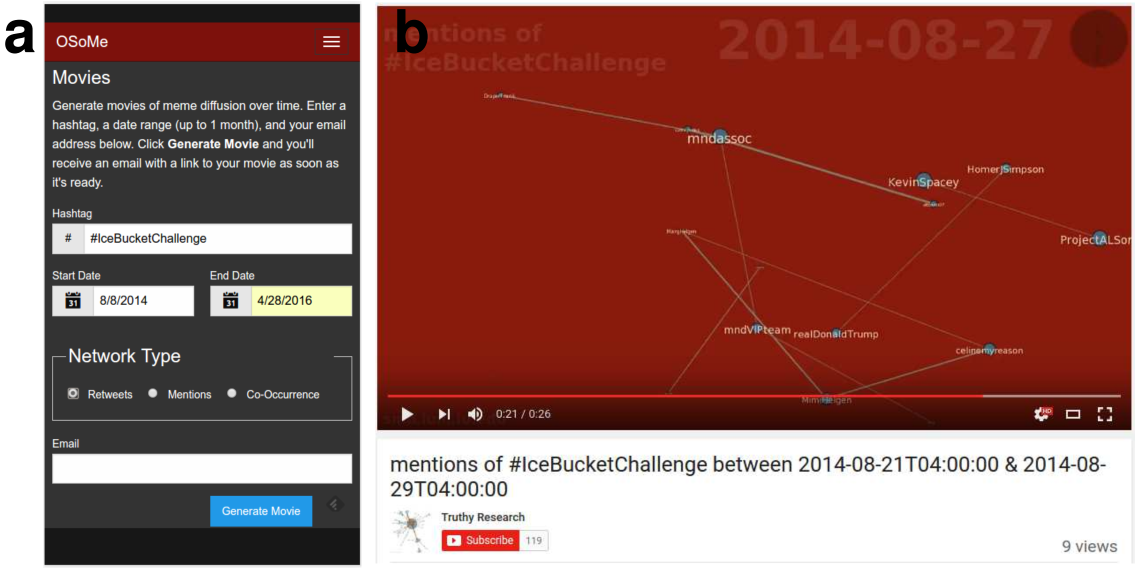

Figure 5: Temporal information diffusion movies. (**a**) The interface of the *Movies* tool let users specify a hashtag, a temporal interval, and the type of diffusion ties to visualize (retweets, mentions, or hashtag co-occurrence). (**b**) Example of a generated movie frame, showing a retweet network for the #IceBucketChallenge hashtag.

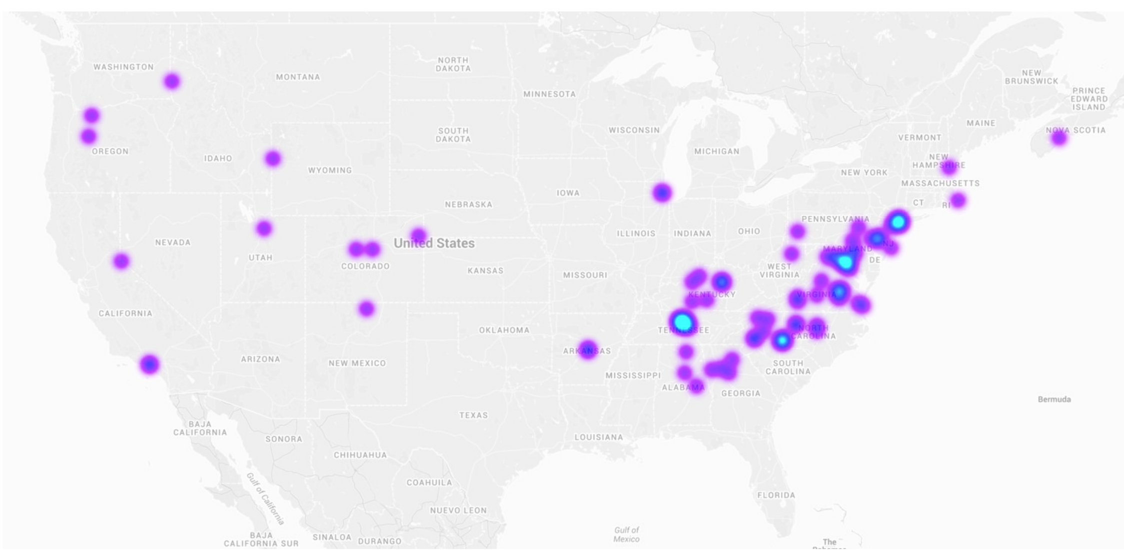

Figure 6: Heatmap of tweets containing the hashtag #snow on January 22, 2016, the day of a large snowstorm over the Eastern United States.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2008v1 | CC-BY 4.0 Open Access | rec: 29 Apr 2016, publ: 29 Apr 2016

through geographic space and time. A subset of tweets (ranging between ≈ 3% in the historical data and ≈ 0.3% in recent years) contain exact latitude/longitude coordinates in their metadata. By aggregating these coordinates into a heatmap layer superimposed on a world map, one can observe the geographic signature of the attention being paid to a given meme. Figure 6 shows an example. Our online tool goes one step further, allowing the user to explore how this geographic signature evolves over a specified time period, via a slider widget.

It takes between 30 and 90 seconds to prepare one of these visualizations *ex novo*. We hope to reduce this lead time with some backend indexing improvements. To enable exploration, we cache all created heatmaps for a period of one week. While cached, the heatmaps can be retrieved instantly, enabling other users to browse and interact with these previously-created visualizations. In the future we hope to experiment with overlaying diffusion networks on top of geographical maps, for example using multi-scale backbone extraction (Serrano et al., 2009) and edge bundling techniques (Selassie et al., 2011).

API

We expect that the majority of users of the Observatory will interact with its data primarily through the tools described above. However, since more advanced data needs are to be expected, we also provide a way to export the data for those who wish to create their own visualizations and develop custom analyses. This is possible either within the tools, via export buttons, and through a read-only HTTP API.

The OSoMe API is deployed via the Mashape management service. Four public methods are currently available. Each takes as input a time interval and a list of tokens (hashtags and/or usernames):

- `tweet-id`: returns a list of tweet IDs mentioning at least one of the inputs in the given interval;
- `counts`: returns a count of the number of tweets mentioning each input token in the given interval;
- `time-series`: for each day in the given time interval, returns a count of tweets matching any of the input tokens;
- `user-post-count`: returns a list of user IDs mentioning any of the tokens in the given time frame, along with a count of matching tweets produced by each user.

Conclusion

The IUNI Observatory on Social Media is the culmination of a large collaborative effort at Indiana University that took place over the course of six years. We hope that it will facilitate computational social science and make big social data easier to analyze by a broad community of researchers, reporters, and the general public. The lessons learned during the development of the infrastructure may be helpful for future endeavors to foster data-intensive research in the social, behavioral, and economic sciences.

We encourage the research community to create new social media analytic tools by building upon our system. For example, one could mashup the OSoMe API with the

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2008v1 | CC-BY 4.0 Open Access | rec: 29 Apr 2016, publ: 29 Apr 2016

BotOrNot API (Davis et al., 2016), also developed by our team, to evaluate the extent to which Twitter campaigns are sustained by social bots.

The opportunities that arise from the Observatory, and from computational social science in general, could have broad societal impact. Systematic attempts to mislead the public on a large scale through "astroturf" campaigns and social bots have been uncovered using big social data analytics, inspiring the development of machine learning methods to detect these abuses (Ratkiewicz et al., 2011a; Ferrara et al., in press; Subrahmanian et al., 2016). Allowing citizens to observe how memes spread online may help raise public awareness of the potential dangers of social media manipulation.

Acknowledgements

The authors would like to acknowledge Alessandro Vespignani and Johan Bollen for discussions leading to the early vision of an Observatory on Social Media; and Gary Miksik, Allan Streib, and Koji Tanaka for their kind assistance with system administration. This work was supported in part by NSF (grants CCF-1101743 and OCI-1149432), the J.S. McDonnell Foundation (grant 220020274), the Swiss National Science Foundation (fellowship PBTIP2_142353), the Lilly Endowment, the Center for Complex Networks and Systems Research (CNetS), the Digital Science Center (DSC), and the Indiana University Network Science Institute (IUNI). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agencies. Finally, we are deeply grateful to Twitter for supporting computational social science research, including the efforts described in this paper, by granting our lab elevated access to the public stream of tweets.

References

J. I. Alvarez-Hamelin, L. Dall'Asta, A. Barrat, and A. Vespignani. Large scale networks fingerprinting and visualization using the k-core decomposition. In *Advances in Neural Information Processing Systems 18 (NIPS)*, pages 41–50, 2005.

N. Andalibi, P. Ozturk, and A. Forte. Depression-related imagery on instagram. In *Proc. 18th ACM Conf. Companion on Computer Supported Cooperative Work & Social Computing (CSCW)*, pages 231–234, 2015. doi: 10.1145/2685553.2699014.

D. Antenucci, M. Cafarella, M. Levenstein, C. Ré, and M. D. Shapiro. Using social media to measure labor market flows. Working Paper 20010, National Bureau of Economic Research, March 2014.

L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna. Four degrees of separation. In *Proceedings of the 4th Annual ACM Web Science Conference*, WebSci '12, pages 33–42, 2012. doi: 10.1145/2380718.2380723.

E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The role of social networks in information diffusion. In *Proceedings of the 21st ACM International Conference on World Wide Web*, pages 519–528, 2012.

J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

S. Card. Information visualization. In *Human-computer interaction: design issues, solutions, and applications*, pages 181–216. CRC Press, 2009.

H. Choi and H. Varian. Predicting the present with google trends. *Economic Record*, 88 (s1):2–9, 2012.

G. L. Ciampaglia, A. Flammini, and F. Menczer. The production of information in the attention economy. *Scientific Reports*, 5:9452, 2015. doi: 10.1038/srep09452.

M. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer. Predicting the political alignment of Twitter users. In *Proc. 3rd IEEE Conference on Social Computing (SocialCom)*, 2011a.

M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, A. Flammini, and F. Menczer. Political polarization on Twitter. In *Proc. 5th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011b.

M. Conover, C. Davis, E. Ferrara, K. McKelvey, F. Menczer, and A. Flammini. The geospatial characteristics of social movement communication networks. *PLoS ONE*, 8 (3):e55957, 2013a.

M. Conover, E. Ferrara, F. Menczer, and A. Flammini. The digital evolution of occupy wall street. *PLoS ONE*, 8(3):e64679, 2013b.

M. D. Conover, B. Gonçalves, A. Flammini, and F. Menczer. Partisan asymmetries in online political activity. *EPJ Data Science*, 1:6, 2012.

C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer. Botornot: A system to evaluate social bots. In *Proc. WWW Developers Day Workshop*, 2016. doi: 10.1145/2872518.2889302. URL `http://arxiv.org/abs/1602.00975`.

M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz. Predicting depression via social media. In *Proc. 7th Intl. AAAI Conf. on Weblogs and Social Media (ICWSM)*, 2013.

D. Difranzo, J. S. Erickson, M. J. K. T. Gloria, J. S. Luciano, D. L. McGuinness, and J. Hendler. The web observatory extension: Facilitating web science collaboration through semantic markup. In *Proc. 23rd Intl. Conf. on World Wide Web Companion*, pages 475–480, 2014. doi: 10.1145/2567948.2576936.

J. DiGrazia, K. McKelvey, J. Bollen, and F. Rojas. More tweets, more votes: Social media as a quantitative indicator of political behavior. *PLoS ONE*, 8(11), 2013.

L. Einav and J. Levin. Economics in the age of big data. *Science*, 346(6210): 1243089–1243089, Nov 2014. ISSN 1095-9203. doi: 10.1126/science.1243089.

E. Ferrara, O. Varol, F. Menczer, and A. Flammini. Traveling Trends: Social Butterflies or Frequent Fliers? In *Proc. 1st ACM Conf. on Online Social Networks (COSN)*, pages 213–222, 2013.

E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini. The rise of social bots. *Commun. ACM*, in press. arXiv preprint arXiv:1407.5225.

S. T. Fiske and R. M. Hauser. Protecting human research participants in the age of big data. *Proceedings of the National Academy of Sciences*, 111(38):13675–13676, 2014. doi: 10.1073/pnas.1414626111.

G. C. Fox, S. Jha, J. Qiu, and A. Luckow. Towards an understanding of facets and exemplars of big data applications. In *Proceedings of 20 Years of Beowulf: Workshop to Honor Thomas Sterling's 65th Birthday*, pages 7–16, 2014.

X. Gao and J. Qiu. Supporting end-to-end social media data analysis with the Indexed-HBase platform. In *Proceedings of the 6th Workshop on Many-Task Computing on Clouds, Grids, and Supercomputers (MTAGS) at SC13*, 2013.

X. Gao and J. Qiu. Supporting queries and analyses of large-scale social media data with customizable and scalable indexing techniques over nosql databases. In *Proceedings of the 14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid 2014)*, pages 587–590, 2014.

X. Gao, V. Nachankar, and J. Qiu. Experimenting Lucene Index on HBase in an HPC Environment. In *Proceedings of ACM High Performance Computing meets Databases workshop (HPCDB'11) at SuperComputing 11*, pages 25–28, 2011.

X. Gao, E. Roth, K. McKelvey, C. Davis, A. Younge, E. Ferrara, F. Menczer, and J. Qiu. Supporting a social media observatory with customizable index structures: Architecture and performance. In *Cloud Computing for Data Intensive Applications*, pages 401–427. Springer, 2014.

X. Gao, E. Ferrara, and J. Qiu. Parallel clustering of high-dimensional social media data streams. In *Proc. 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid)*, pages 323–332, 2015.

P. A. Grabowicz, L. M. Aiello, and F. Menczer. Fast filtering and animation of large dynamic networks. *EPJ Data Science*, 3(1):27, 2014. doi: 10.1140/epjds/s13688-014-0027-8.

D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *Proceedings of the 13th International ACM Conference on World Wide Web*, WWW '04, pages 491–501, 2004. doi: 10.1145/988672.988739.

A. Guille, H. Hacid, C. Favre, and D. A. Zighed. Information diffusion in online social networks. *SIGMOD Rec.*, 42(1):17, 2013. doi: 10.1145/2503792.2503797.

E. Hargittai. Is Bigger Always Better? Potential Biases of Big Data Derived from Social Network Sites. *The Annals of the American Academy of Political and Social Science*, 659(1): 63—-76, 2015. doi: 10.1177/0002716215570866.

S. Harriman and J. Patel. The ethics and editorial challenges of internet-based research. *BMC Med*, 12(1), 2014. doi: 10.1186/s12916-014-0124-3.

K. Healy and J. Moody. Data visualization in sociology. *Annual review of sociology*, 40: 105—128, 2014. doi: 10.1146/annurev-soc-071312-145551.

S. Jha, J. Qiu, A. Luckow, P. Mantha, and G. C. Fox. A tale of two data-intensive paradigms: Applications, abstractions, and architectures. In *Proceedings of the 3rd International Congress on Big Data Conference (IEEE BigData)*, 2014.

J. P. Kahn, E. Vayena, and A. C. Mastroianni. Opinion: Learning as we go: Lessons from the publication of facebook's social-computing research. *Proceedings of the National Academy of Sciences*, 111(38):13677–13679, 2014. doi: 10.1073/pnas.1416405111.

T. Kamada and S. Kawai. An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31(1):7 – 15, 1989.

G. King. Ensuring the data-rich future of the social sciences. *Science*, 331(6018):719–721, 2011. doi: 10.1126/science.1197872.

A. D. Kramer, J. E. Guillory, and J. T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790, 2014.

D. Lazer, A. S. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, et al. Life in the network: the coming age of computational social science. *Science*, 323(5915):721, 2009.

B. G. Link, J. C. Phelan, M. Bresnahan, A. Stueve, and B. A. Pescosolido. Public conceptions of mental illness: labels, causes, dangerousness, and social distance. *Am J Public Health*, 89(9):1328–1333, 1999. doi: 10.2105/AJPH.89.9.1328.

K. McKelvey and F. Menczer. Design and prototyping of a social media observatory. In *Proc. 22nd Intl. Conf. on World Wide Web (WWW) Companion*, pages 1351–1358, 2013a.

K. McKelvey and F. Menczer. Truthy: Enabling the Study of Online Social Networks. In *Proc. 16th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion (CSCW)*, 2013b.

E. F. Moran, S. L. Hofferth, C. C. Eckel, D. Hamilton, B. Entwisle, J. L. Aber, H. E. Brady, D. Conley, S. L. Cutter, K. Hubacek, et al. Opinion: Building a 21st-century infrastructure for the social sciences. *Proceedings of the National Academy of Sciences*, 111(45):15855–15856, 2014.

F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In *Proc. 7th Intl. AAAI Conf. on Weblogs and Social Media (ICWSM)*, 2013.

A. Nematzadeh, E. Ferrara, A. Flammini, and Y.-Y. Ahn. Optimal network modularity for information diffusion. *Phys. Rev. Lett.*, 113:088701, 2014. doi: 10.1103/PhysRevLett.113.088701.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251), 2015. doi: 10.1126/science.aac4716.

Pivotal Software, Inc. RabbitMQ, 2016. URL `https://www.rabbitmq.com/`. Last accessed April 27, 2016.

J. Qiu, S. Jha, A. Luckow, and G. C. Fox. Towards hpc-abds: An initial high-performance big data stack. In *Proceedings of 1st ACM Big Data Interoperability Framework Workshop: Building Robust Big Data ecosystem*, 2014.

M. Raento, A. Oulasvirta, and N. Eagle. Smartphones: An emerging tool for social scientists. *Sociological Methods & Research*, 37(3):426–454, 2009. doi: 10.1177/0049124108330005.

S. Rafaeli. Interactivity: From new media to communication. *Sage annual review of communication research: Advancing communication science*, 16(CA):110–134, 1988.

J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, A. Flammini, and F. Menczer. Detecting and tracking political abuse in social media. In *Proc. 5th Intl. AAAI Conf. on Weblogs and Social Media (ICWSM)*, 2011a.

J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer. Truthy: Mapping the spread of astroturf in microblog streams. In *Proc. 20th Intl. World Wide Web Conf. Companion (WWW)*, 2011b.

D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter. In *Proc. 20th Intl. Conf. on World Wide Web (WWW)*, pages 695–704, 2011. doi: 10.1145/1963405.1963503.

D. Ruths and J. Pfeffer. Social media for large studies of behavior. *Science*, 346(6213):1063–1064, 2014. doi: 10.1126/science.346.6213.1063.

D. Selassie, B. Heller, and J. Heer. Divided edge bundling for directional network data. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2354–2363, 2011. doi: 10.1109/TVCG.2011.190.

M. Á. Serrano, M. Boguná, and A. Vespignani. Extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences*, 106(16):6483–6488, 2009.

A. Solem and Contributors. Celery, 2016. URL `http://www.celeryproject.org/`. Last accessed April 05, 2016.

V. Subrahmanian, A. Azaria, S. Durst, V. Kagan, A. Galstyan, K. Lerman, L. Zhu, E. Ferrara, A. Flammini, F. Menczer, et al. The DARPA Twitter Bot Challenge. *IEEE Computer*, 2016. Forthcoming. Preprint arXiv:1601.05140.

O. Tene and J. Polonetsky. Privacy in the age of big data: a time for big decisions. *Stanford Law Review Online*, 64:63, 2012.

P. Terna et al. Simulation tools for social scientists: Building agent based models with swarm. *Journal of artificial societies and social simulation*, 1(2):1–12, 1998.

The Apache Software Foundation. Apache HBase, 2016a. URL `http://hbase.apache.org/`. Last accessed April 05, 2016.

The Apache Software Foundation. Hadoop, 2016b. URL `http://hadoop.apache.org/`. Last accessed April 05, 2016.

The Apache Software Foundation. Apache Solr, 2016c. URL http://lucene.apache.org/solr/. Last accessed April 05, 2016.

The Apache Software Foundation. Apache Hadoop YARN, 2016d. URL http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html. Last accessed April 05, 2016.

J. Travers and S. Milgram. An experimental study of the small world problem. *Sociometry*, pages 425–443, 1969.

Twitter, Inc. Developer policy. Available at: https://dev.twitter.com/overview/terms/policy, Internet Archive: https://web.archive.org/web/20160311122344/https://dev.twitter.com/overview/terms/policy, 2016. Last accessed: 04/09/2016.

O. Varol, E. Ferrara, C. Ogan, F. Menczer, and A. Flammini. Evolution of online user behavior during a social upheaval. In *Proc. ACM Web Science Conference (WebSci)*, 2014.

E. Vayena, M. Salathé, L. C. Madoff, and J. S. Brownstein. Ethical challenges of big data in public health. *PLoS Comput Biol*, 11(2):e1003904, 2015. doi: 10.1371/journal.pcbi.1003904.

L. Weng and F. Menczer. Topicality and impact in social media: Diverse messages, focused messengers. *PLoS ONE*, 10(2):e0118410, 2015.

L. Weng, A. Flammini, A. Vespignani, and F. Menczer. Competition among memes in a world with limited attention. *Sci. Rep.*, 2(335), 2012.

L. Weng, F. Menczer, and Y.-Y. Ahn. Virality prediction and community structure in social networks. *Sci. Rep.*, 3(2522), 2013a.

L. Weng, J. Ratkiewicz, N. Perra, B. Gonçalves, C. Castillo, F. Bonchi, R. Schifanella, F. Menczer, and A. Flammini. The role of information diffusion in the evolution of social networks. In *Proc. 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2013b.

L. Weng, F. Menczer, and Y.-Y. Ahn. Predicting successful memes using network and community structure. In *Proc. Eighth International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2014.

T. B. Wiggins, X. Gao, and J. Qiu. IndexedHBase, 2016. URL http://salsaproj.indiana.edu/IndexedHBase. Last accessed April 05, 2016.

T.-L. Wu, B. Zhang, C. A. Davis, E. Ferrara, A. Flammini, F. Menczer, and J. Qiu. Scalable query and analysis for social networks: An integrated high-level dataflow system with pig and harp. In M. Thai, H. Xiong, and W. Wu, editors, *Big Data in Complex and Social Networks*. Chapman and Hall/CRC, 2016. Forthcoming.

M. Zimmer. The Twitter archive at the library of congress: Challenges for information practice and information policy. *First Monday*, 20(7), 2015.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2008v1 | CC-BY 4.0 Open Access | rec: 29 Apr 2016, publ: 29 Apr 2016