

A peer-reviewed version of this preprint was published in PeerJ on 3 October 2016.

[View the peer-reviewed version](https://peerj.com/articles/cs-87) (peerj.com/articles/cs-87), which is the preferred citable publication unless you specifically need to cite this preprint.

Davis CA, Ciampaglia GL, Aiello LM, Chung K, Conover MD, Ferrara E, Flammini A, Fox GC, Gao X, Gonçalves B, Grabowicz PA, Hong K, Hui P, McCaulay S, McKelvey K, Meiss MR, Patil S, Peli Kankanamalage C, Pentchev V, Qiu J, Ratkiewicz J, Rudnick A, Serrette B, Shiralkar P, Varol O, Weng L, Wu T, Younge AJ, Menczer F. (2016) OSoMe: the IUNI observatory on social media. PeerJ Computer Science 2:e87 <https://doi.org/10.7717/peerj-cs.87>

OSoMe: The IUNI Observatory on Social Media

Clayton A. Davis^{1,2}, Giovanni Luca Ciampaglia^{1,3}, Luca Maria Aiello⁴,
Keychul Chung², Michael Conover⁵, Emilio Ferrara⁶, Alessandro
Flammini^{1,2,3}, Geoffrey Fox², Xiaoming Gao⁷, Bruno Gonçalves⁸, Przemyslaw
Grabowicz⁹, Alex Hong², Pik-Mai Hui², Scott McCaulay³, Karissa
McKelvey¹⁰, Mark Meiss¹¹, Snehal Patil⁴, Chathuri Peli Kankanamalage³,
Valentin Pentchev³, Judy Qiu², Jacob Ratkiewicz¹¹, Alex Rudnick¹¹, Benjamin
Serrette³, Prashant Shiralkar^{1,2}, Onur Varol^{1,2}, Lilian Weng¹², Tak-Lon Wu¹³,
Andrew Younge², 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.

1

Abstract

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

11 Introduction

12 The collective processes of production, consumption, and diffusion of information on
13 social media are starting to reveal a significant portion of human social life, yet scien-
14 tists struggle to get access to data about it. Recent research has shown that social media
15 can perform as ‘sensors’ for collective activity at multiple scales (Lazer et al., 2009). As
16 a consequence, data extracted from social media platforms are increasingly used side-
17 by-side with — and sometimes even replacing — traditional methods to investigate
18 hard-pressing questions in the social, behavioral, and economic (SBE) sciences (King,
19 2011; Moran et al., 2014; Einav and Levin, 2014). For example, interpersonal connections
20 from Facebook have been used to replicate the famous experiment by Travers and Mil-
21 gram (1969) on a global scale (Backstrom et al., 2012); the emotional content of social
22 media streams has been used to estimate macroeconomic quantities in country-wide
23 economies (Bollen et al., 2011; Choi and Varian, 2012; Antenucci et al., 2014); and im-
24 agery from Instagram has been mined (De Choudhury et al., 2013; Andalibi et al., 2015)
25 to understand the spread of depression among teenagers (Link et al., 1999).

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

34 Examples of empirical works concerned with these questions include geographic
35 and temporal patterns in social movements (Conover et al., 2013b,a; Varol et al., 2014),
36 the polarization of online political discourse (Conover et al., 2011b,a, 2012), the use of
37 social media data to predict election outcomes (DiGrazia et al., 2013) and stock market
38 movements (Bollen et al., 2011), the geographic diffusion of trending topics (Ferrara
39 et al., 2013), and the lifecycle of information in the attention economy (Ciampaglia et al.,
40 2015).

41 On the more theoretical side, agent-based models have been proposed to explain
42 how limited individual attention affects what information we propagate (Weng et al.,
43 2012), what social connections we make (Weng et al., 2013b), and how the structure
44 of social and topical networks can help predict which memes are likely to become vi-
45 ral (Weng et al., 2013a, 2014; Nematzadeh et al., 2014; Weng and Menczer, 2015).

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

56 A second limitation is related to the ease of use of APIs, which are usually meant
57 for software developers, not researchers. While researchers in the SBE sciences are

58 increasingly acquiring software development skills (Terna et al., 1998; Raento et al., 2009;
59 Healy and Moody, 2014), and intuitive user interfaces are becoming more ubiquitous,
60 many tasks remain challenging enough to hinder research advances. This is especially
61 true for those tasks related to the application of fast visualization techniques.

62 A third, important limitation is related to user privacy. Unfettered access to sensitive,
63 private data about the choices, behaviors, and preferences of individuals is happening at
64 an increasing rate (Tene and Polonetsky, 2012). Coupled with the possibility to manip-
65 ulate the environment presented to users (Kramer et al., 2014), this has raised in more
66 than one occasion deep ethical concerns in both the public and the scientific commu-
67 nity (Kahn et al., 2014; Fiske and Hauser, 2014; Harriman and Patel, 2014; Vayena et al.,
68 2015).

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

75 Our long-term goal is to enable SBE researchers and the general public to openly
76 access relevant social media data. As a concrete milestone of our project, here we present
77 an *Observatory on Social Media* — an open infrastructure for sharing public data about
78 information that is spread and collected through online social networks. Our initial
79 focus has been on Twitter as a source of public microblogging posts. The infrastructure
80 takes care of storing, indexing, and analyzing public collections and historical archives
81 of big social data; it does so in an easy-to-use way, enabling broad access from scientists
82 and other stakeholders, like journalists and the general public. We envision that data
83 and analytics from social media will be integrated within a nation-wide network of
84 social observatories. These data centers would allow access to a broad range of data
85 about social, behavioral, and economic phenomena nationwide (King, 2011; Moran et al.,
86 2014; Difranzo et al., 2014).

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

94 Data Source

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

¹The 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.

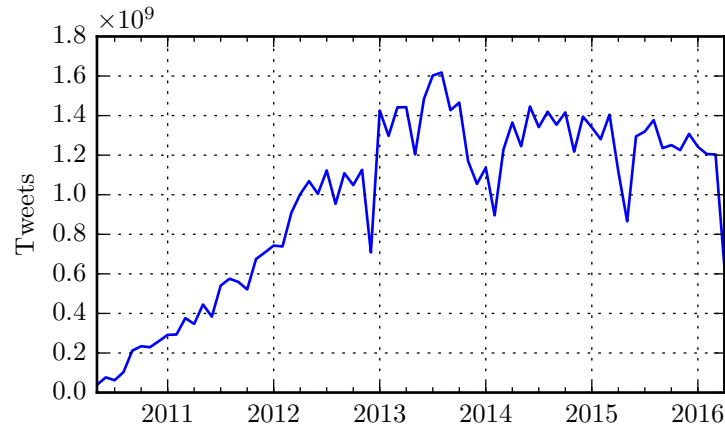


Figure 1: Number of monthly messages collected and indexed by OSoMe. System failures have caused occasional interruptions of the collection system.

99 is built around a Terabyte-scale historical (and ongoing) collection of approximately 70
 100 **billion public tweets** to date. The data has been collected from a random 10% stream
 101 sample of public Twitter posts and dates back to mid 2010.² The high-speed stream from
 102 which the data originates has a rate that ranges in the order of $10^6 - 10^8$ tweets/day.
 103 Figure 1 illustrates the growth of the Twitter collection over time.

104 System Architecture

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

108 The architecture the Observatory builds upon the Apache Big Data Stack (ABDS)
 109 framework (Jha et al., 2014; Qiu et al., 2014; Fox et al., 2014). Development has been
 110 driven over the years by the need for increasingly demanding social media analytics
 111 applications (Gao et al., 2011; Gao and Qiu, 2013, 2014; Gao et al., 2014, 2015; Wu et al.,
 112 2016). A key idea behind our enhancement of the ABDS architecture is the shift from
 113 standalone systems to modules; multiple modules can be used within existing software
 114 ecosystems. In particular, we have focused our efforts on enhancing two well-known
 115 Apache modules, Hadoop (The Apache Software Foundation, 2016b) and HBase (The
 116 Apache Software Foundation, 2016a).

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

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

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

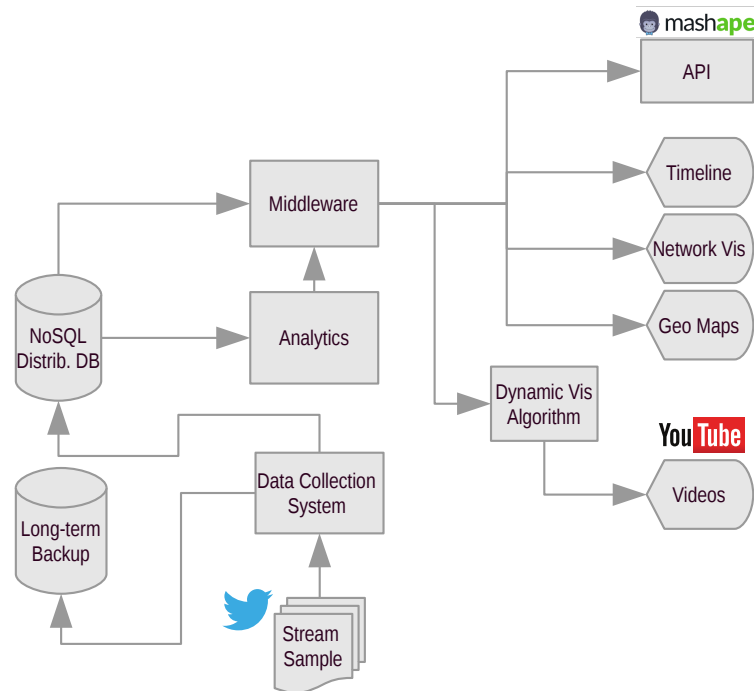


Figure 2: Flowchart diagram of the OSoMe architecture. Arrows indicate flow of data.

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

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

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

146 The Observatory user interface follows a modular architecture too, and is based on
 147 a number of apps, which we describe in greater detail in the following section. Three
 148 of the apps (*Timeline*, *Network visualization*, and *Geographic maps*) are directly accessible

149 within OSoMe through Web interfaces. We rely on the popular video-sharing service
150 YouTube for the fourth app, which generates meme diffusion movies (*Videos*) using a fast
151 *dynamic visualization algorithm* (Grabowicz et al., 2014) specifically designed for temporal
152 networks. Finally, the Observatory provides access to raw data via a programmatic
153 interface (*API*).

154 Applications

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

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

166 Temporal Trends

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

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

180 Diffusion and Co-occurrence Networks

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

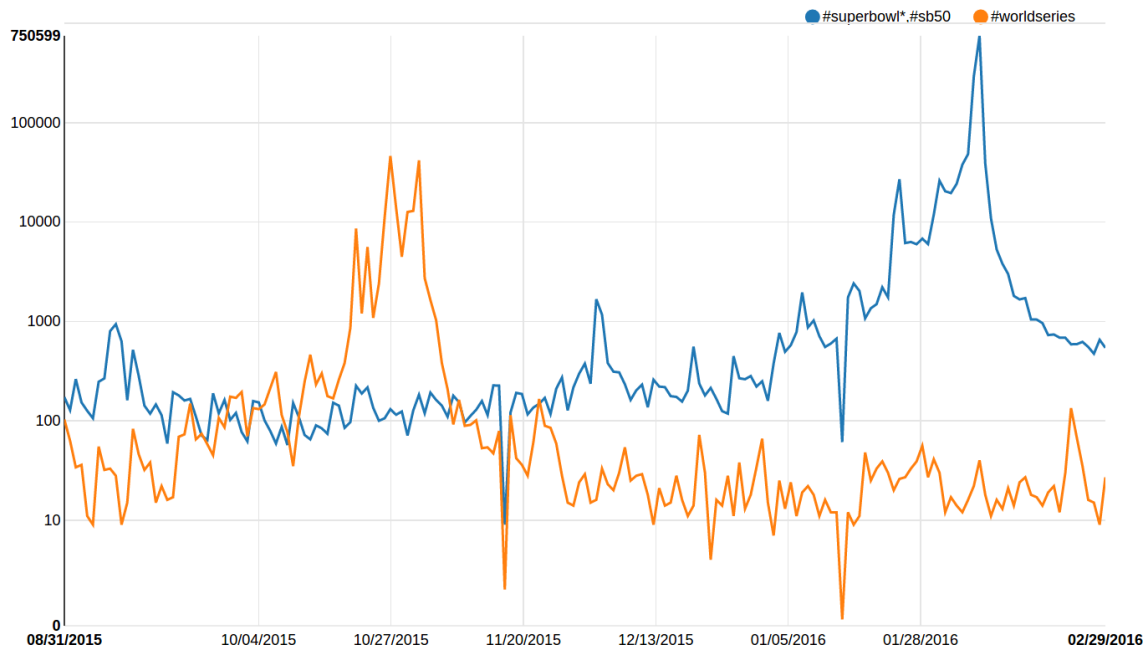


Figure 3: Number of tweets per day about the Super Bowl (in blue) and the World Series (in orange), from September 2015 through February 2016. The Y-axis is in logarithmic scale, shifted by one to account for null counts. The plot shows two outages in the data collection that occurred around mid-November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

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

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

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

200 Interactive Network Visualization

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

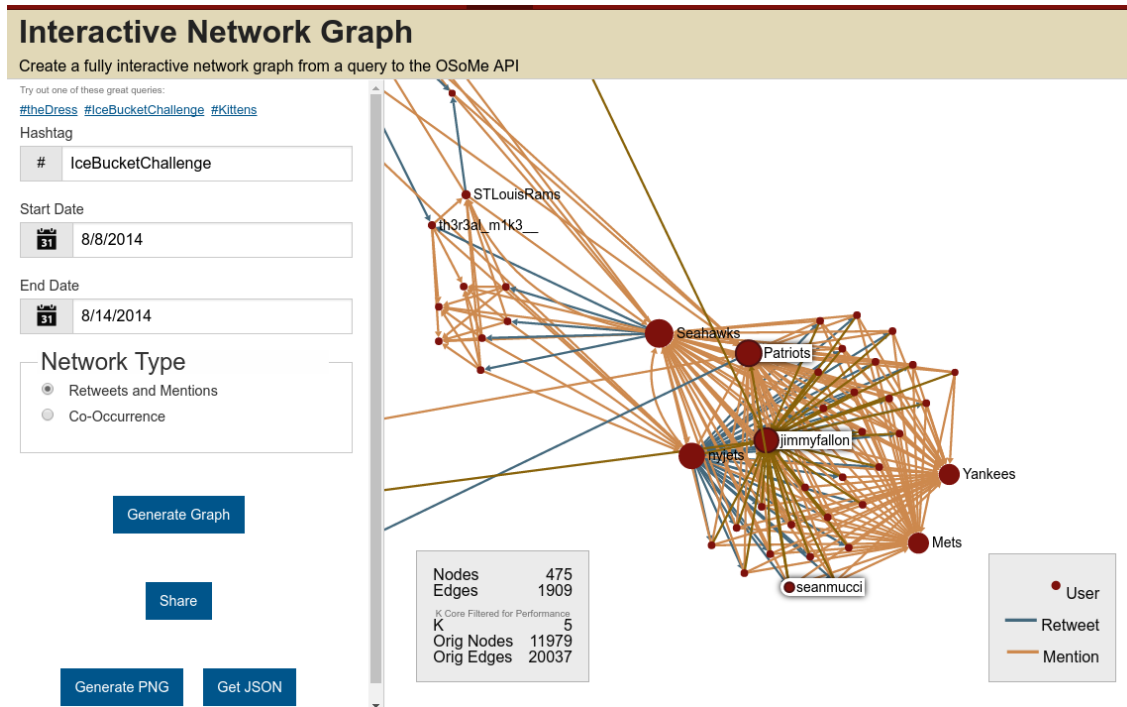


Figure 4: Interactive Network Visualization Tool. A detail of the network of retweets and mention for a hashtag commonly linked to “Ice Bucket Challenge,” a popular Internet phenomenon from 2014. The size of a node is proportional to its strength (weighted degree). For visualization purposes, the size 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.

209 Animations

210 Because tweet data are time resolved, the evolution of a diffusion or co-occurrence net-
 211 work can be also visualized over time. Currently the *Networks* tool visualizes only static
 212 networks aggregated over the entire search period specified by the user; we aim to add
 213 the ability to observe the network evolution over time, but in the meantime we also pro-
 214 vide the *Movies* tools, an alternative service that lets users generate animations of such
 215 processes (Figure 5). We have successfully experimented with fast visualization tech-
 216 niques in the past, and have found that edge filtering is the best approach for efficiently
 217 visualizing networks that undergo a rapid churn of both edges and nodes. We have
 218 therefore deployed a fast filtering algorithm developed by our team (Grabowicz et al.,
 219 2014). The user-generated videos are uploaded to YouTube, and we cache the videos in
 220 case multiple users try to visualize the same network.

221 Geographic maps

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

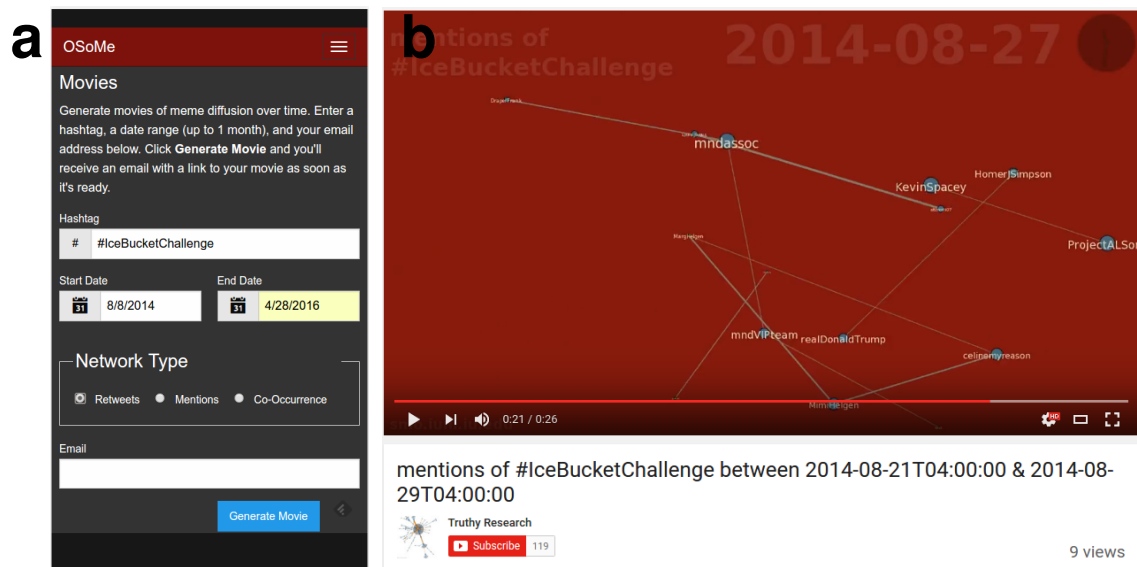


Figure 5: Temporal information diffusion movies. (a) The interface of the *Movies* tool let users specify a hashtag, a temporal interval, and the type of diffusion ties to visualize (retweets, mentions, or hashtag co-occurrence). (b) Example of a generated movie frame, showing a retweet network for the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

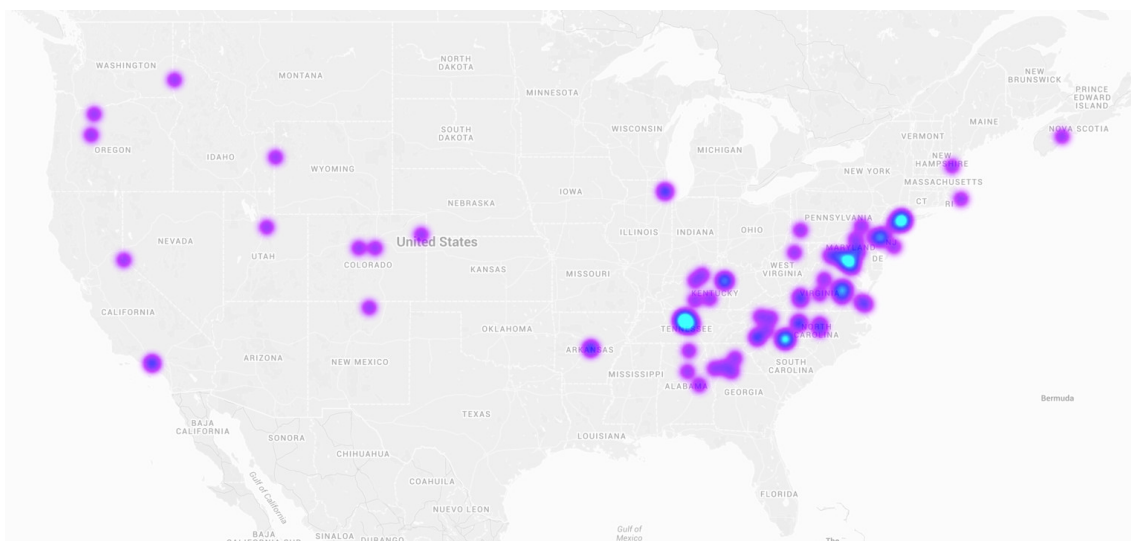


Figure 6: Heatmap of tweets containing the hashtag #snow on January 22, 2016, the day of a large snowstorm over the Eastern United States.

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

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

240 API

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

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

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

257 Conclusion

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

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

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

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

275 Acknowledgements

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

288 References

- 289 J. I. Alvarez-Hamelin, L. Dall’Asta, A. Barrat, and A. Vespignani. Large scale networks
290 fingerprinting and visualization using the k-core decomposition. In *Advances in Neural*
291 *Information Processing Systems 18 (NIPS)*, pages 41–50, 2005.
- 292 N. Andalibi, P. Ozturk, and A. Forte. Depression-related imagery on instagram. In *Proc.*
293 *18th ACM Conf. Companion on Computer Supported Cooperative Work & Social Computing*
294 *(CSCW)*, pages 231–234, 2015. doi: 10.1145/2685553.2699014.
- 295 D. Antenucci, M. Cafarella, M. Levenstein, C. Ré, and M. D. Shapiro. Using social media
296 to measure labor market flows. Working Paper 20010, National Bureau of Economic
297 Research, March 2014.
- 298 L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna. Four degrees of separation.
299 In *Proceedings of the 4th Annual ACM Web Science Conference, WebSci ’12*, pages 33–42,
300 2012. doi: 10.1145/2380718.2380723.
- 301 E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The role of social networks in infor-
302 mation diffusion. In *Proceedings of the 21st ACM International Conference on World Wide*
303 *Web*, pages 519–528, 2012.
- 304 J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of*
305 *Computational Science*, 2(1):1–8, 2011.

- 306 S. Card. Information visualization. In *Human-computer interaction: design issues, solutions,*
307 *and applications*, pages 181–216. CRC Press, 2009.
- 308 H. Choi and H. Varian. Predicting the present with google trends. *Economic Record*, 88
309 (s1):2–9, 2012.
- 310 G. L. Ciampaglia, A. Flammini, and F. Menczer. The production of information in the
311 attention economy. *Scientific Reports*, 5:9452, 2015. doi: 10.1038/srep09452.
- 312 M. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer. Predicting the
313 political alignment of Twitter users. In *Proc. 3rd IEEE Conference on Social Computing*
314 *(SocialCom)*, 2011a.
- 315 M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, A. Flammini, and F. Menczer.
316 Political polarization on Twitter. In *Proc. 5th International AAAI Conference on Weblogs*
317 *and Social Media (ICWSM)*, 2011b.
- 318 M. Conover, C. Davis, E. Ferrara, K. McKelvey, F. Menczer, and A. Flammini. The
319 geospatial characteristics of social movement communication networks. *PLoS ONE*, 8
320 (3):e55957, 2013a.
- 321 M. Conover, E. Ferrara, F. Menczer, and A. Flammini. The digital evolution of occupy
322 wall street. *PLoS ONE*, 8(3):e64679, 2013b.
- 323 M. D. Conover, B. Gonçalves, A. Flammini, and F. Menczer. Partisan asymmetries in
324 online political activity. *EPJ Data Science*, 1:6, 2012.
- 325 C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer. Botornot: A system
326 to evaluate social bots. In *Proc. WWW Developers Day Workshop*, 2016. doi: 10.1145/
327 2872518.2889302. URL <http://arxiv.org/abs/1602.00975>.
- 328 M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz. Predicting depression via
329 social media. In *Proc. 7th Intl. AAAI Conf. on Weblogs and Social Media (ICWSM)*, 2013.
- 330 D. Difranzo, J. S. Erickson, M. J. K. T. Gloria, J. S. Luciano, D. L. McGuinness, and
331 J. Hendler. The web observatory extension: Facilitating web science collaboration
332 through semantic markup. In *Proc. 23rd Intl. Conf. on World Wide Web Companion*,
333 pages 475–480, 2014. doi: 10.1145/2567948.2576936.
- 334 J. DiGrazia, K. McKelvey, J. Bollen, and F. Rojas. More tweets, more votes: Social media
335 as a quantitative indicator of political behavior. *PLoS ONE*, 8(11), 2013.
- 336 L. Einav and J. Levin. Economics in the age of big data. *Science*, 346(6210):
337 1243089–1243089, Nov 2014. ISSN 1095-9203. doi: 10.1126/science.1243089.
- 338 E. Ferrara, O. Varol, F. Menczer, and A. Flammini. Traveling Trends: Social Butterflies
339 or Frequent Fliers? In *Proc. 1st ACM Conf. on Online Social Networks (COSN)*, pages
340 213–222, 2013.
- 341 E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini. The rise of social bots.
342 *Commun. ACM*, in press. arXiv preprint arXiv:1407.5225.

- 343 S. T. Fiske and R. M. Hauser. Protecting human research participants in the age of big
344 data. *Proceedings of the National Academy of Sciences*, 111(38):13675–13676, 2014. doi:
345 10.1073/pnas.1414626111.
- 346 G. C. Fox, S. Jha, J. Qiu, and A. Luckow. Towards an understanding of facets and
347 exemplars of big data applications. In *Proceedings of 20 Years of Beowulf: Workshop to*
348 *Honor Thomas Sterling's 65th Birthday*, pages 7–16, 2014.
- 349 X. Gao and J. Qiu. Supporting end-to-end social media data analysis with the Indexed-
350 HBase platform. In *Proceedings of the 6th Workshop on Many-Task Computing on Clouds,*
351 *Grids, and Supercomputers (MTAGS) at SC13*, 2013.
- 352 X. Gao and J. Qiu. Supporting queries and analyses of large-scale social media data with
353 customizable and scalable indexing techniques over nosql databases. In *Proceedings*
354 *of the 14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*
355 *(CCGrid 2014)*, pages 587–590, 2014.
- 356 X. Gao, V. Nachankar, and J. Qiu. Experimenting Lucene Index on HBase in an HPC En-
357 vironment. In *Proceedings of ACM High Performance Computing meets Databases workshop*
358 *(HPCDB'11) at SuperComputing 11*, pages 25–28, 2011.
- 359 X. Gao, E. Roth, K. McKelvey, C. Davis, A. Younge, E. Ferrara, F. Menczer, and J. Qiu.
360 Supporting a social media observatory with customizable index structures: Archi-
361 tecture and performance. In *Cloud Computing for Data Intensive Applications*, pages
362 401–427. Springer, 2014.
- 363 X. Gao, E. Ferrara, and J. Qiu. Parallel clustering of high-dimensional social media data
364 streams. In *Proc. 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid*
365 *Computing (CCGrid)*, pages 323–332, 2015.
- 366 P. A. Grabowicz, L. M. Aiello, and F. Menczer. Fast filtering and animation of
367 large dynamic networks. *EPJ Data Science*, 3(1):27, 2014. doi: 10.1140/epjds/
368 s13688-014-0027-8.
- 369 D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through
370 blogspace. In *Proceedings of the 13th International ACM Conference on World Wide Web,*
371 *WWW '04*, pages 491–501, 2004. doi: 10.1145/988672.988739.
- 372 A. Guille, H. Hacid, C. Favre, and D. A. Zighed. Information diffusion in online social
373 networks. *SIGMOD Rec.*, 42(1):17, 2013. doi: 10.1145/2503792.2503797.
- 374 E. Hargittai. Is Bigger Always Better? Potential Biases of Big Data Derived from Social
375 Network Sites. *The Annals of the American Academy of Political and Social Science*, 659(1):
376 63—76, 2015. doi: 10.1177/0002716215570866.
- 377 S. Harriman and J. Patel. The ethics and editorial challenges of internet-based research.
378 *BMC Med*, 12(1), 2014. doi: 10.1186/s12916-014-0124-3.
- 379 K. Healy and J. Moody. Data visualization in sociology. *Annual review of sociology*, 40:
380 105—128, 2014. doi: 10.1146/annurev-soc-071312-145551.

- 381 S. Jha, J. Qiu, A. Luckow, P. Mantha, and G. C. Fox. A tale of two data-intensive
382 paradigms: Applications, abstractions, and architectures. In *Proceedings of the 3rd*
383 *International Congress on Big Data Conference (IEEE BigData)*, 2014.
- 384 J. P. Kahn, E. Vayena, and A. C. Mastroianni. Opinion: Learning as we go: Lessons from
385 the publication of facebook’s social-computing research. *Proceedings of the National*
386 *Academy of Sciences*, 111(38):13677–13679, 2014. doi: 10.1073/pnas.1416405111.
- 387 T. Kamada and S. Kawai. An algorithm for drawing general undirected graphs. *Infor-*
388 *mation Processing Letters*, 31(1):7 – 15, 1989.
- 389 G. King. Ensuring the data-rich future of the social sciences. *Science*, 331(6018):719–721,
390 2011. doi: 10.1126/science.1197872.
- 391 A. D. Kramer, J. E. Guillory, and J. T. Hancock. Experimental evidence of massive-scale
392 emotional contagion through social networks. *Proceedings of the National Academy of*
393 *Sciences*, 111(24):8788–8790, 2014.
- 394 D. Lazer, A. S. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis,
395 N. Contractor, J. Fowler, M. Gutmann, et al. Life in the network: the coming age of
396 computational social science. *Science*, 323(5915):721, 2009.
- 397 B. G. Link, J. C. Phelan, M. Bresnahan, A. Stueve, and B. A. Pescosolido. Public concep-
398 tions of mental illness: labels, causes, dangerousness, and social distance. *Am J Public*
399 *Health*, 89(9):1328–1333, 1999. doi: 10.2105/AJPH.89.9.1328.
- 400 K. McKelvey and F. Menczer. Design and prototyping of a social media observatory. In
401 *Proc. 22nd Intl. Conf. on World Wide Web (WWW) Companion*, pages 1351–1358, 2013a.
- 402 K. McKelvey and F. Menczer. Truthy: Enabling the Study of Online Social Networks. In
403 *Proc. 16th ACM Conference on Computer Supported Cooperative Work and Social Computing*
404 *Companion (CSCW)*, 2013b.
- 405 E. F. Moran, S. L. Hofferth, C. C. Eckel, D. Hamilton, B. Entwisle, J. L. Aber, H. E.
406 Brady, D. Conley, S. L. Cutter, K. Hubacek, et al. Opinion: Building a 21st-century
407 infrastructure for the social sciences. *Proceedings of the National Academy of Sciences*,
408 111(45):15855–15856, 2014.
- 409 F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley. Is the Sample Good Enough? Com-
410 paring Data from Twitter’s Streaming API with Twitter’s Firehose. In *Proc. 7th Intl.*
411 *AAAI Conf. on Weblogs and Social Media (ICWSM)*, 2013.
- 412 A. Nematzadeh, E. Ferrara, A. Flammini, and Y.-Y. Ahn. Optimal network modularity
413 for information diffusion. *Phys. Rev. Lett.*, 113:088701, 2014. doi: 10.1103/PhysRevLett.
414 113.088701.
- 415 Open Science Collaboration. Estimating the reproducibility of psychological science.
416 *Science*, 349(6251), 2015. doi: 10.1126/science.aac4716.
- 417 Pivotal Software, Inc. RabbitMQ, 2016. URL <https://www.rabbitmq.com/>. Last accessed
418 April 27, 2016.

- 419 J. Qiu, S. Jha, A. Luckow, and G. C. Fox. Towards hpc-abds: An initial high-performance
420 big data stack. In *Proceedings of 1st ACM Big Data Interoperability Framework Workshop:
421 Building Robust Big Data ecosystem*, 2014.
- 422 M. Raento, A. Oulasvirta, and N. Eagle. Smartphones: An emerging tool for so-
423 cial scientists. *Sociological Methods & Research*, 37(3):426–454, 2009. doi: 10.1177/
424 0049124108330005.
- 425 S. Rafaeli. Interactivity: From new media to communication. *Sage annual review of
426 communication research: Advancing communication science*, 16(CA):110–134, 1988.
- 427 J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, A. Flammini, and F. Menczer. De-
428 tecting and tracking political abuse in social media. In *Proc. 5th Intl. AAAI Conf. on
429 Weblogs and Social Media (ICWSM)*, 2011a.
- 430 J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer.
431 Truthy: Mapping the spread of astroturf in microblog streams. In *Proc. 20th Intl. World
432 Wide Web Conf. Companion (WWW)*, 2011b.
- 433 D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information
434 diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter.
435 In *Proc. 20th Intl. Conf. on World Wide Web (WWW)*, pages 695–704, 2011. doi: 10.1145/
436 1963405.1963503.
- 437 D. Ruths and J. Pfeffer. Social media for large studies of behavior. *Science*, 346(6213):
438 1063–1064, 2014. doi: 10.1126/science.346.6213.1063.
- 439 D. Selassie, B. Heller, and J. Heer. Divided edge bundling for directional network data.
440 *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2354–2363, 2011. doi:
441 10.1109/TVCG.2011.190.
- 442 M. Á. Serrano, M. Boguná, and A. Vespignani. Extracting the multiscale backbone of
443 complex weighted networks. *Proceedings of the National Academy of Sciences*, 106(16):
444 6483–6488, 2009.
- 445 A. Solem and Contributors. Celery, 2016. URL <http://www.celeryproject.org/>. Last
446 accessed April 05, 2016.
- 447 V. Subrahmanian, A. Azaria, S. Durst, V. Kagan, A. Galstyan, K. Lerman, L. Zhu, E. Fer-
448 rera, A. Flammini, F. Menczer, et al. The DARPA Twitter Bot Challenge. *IEEE Com-
449 puter*, 2016. Forthcoming. Preprint arXiv:1601.05140.
- 450 O. Tene and J. Polonetsky. Privacy in the age of big data: a time for big decisions.
451 *Stanford Law Review Online*, 64:63, 2012.
- 452 P. Terna et al. Simulation tools for social scientists: Building agent based models with
453 swarm. *Journal of artificial societies and social simulation*, 1(2):1–12, 1998.
- 454 The Apache Software Foundation. Apache HBase, 2016a. URL [http://hbase.apache.
455 org/](http://hbase.apache.org/). Last accessed April 05, 2016.
- 456 The Apache Software Foundation. Hadoop, 2016b. URL <http://hadoop.apache.org/>.
457 Last accessed April 05, 2016.

- 458 The Apache Software Foundation. Apache Solr, 2016c. URL <http://lucene.apache.org/solr/>. Last accessed April 05, 2016.
- 460 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.
- 463 J. Travers and S. Milgram. An experimental study of the small world problem. *Sociometry*, pages 425–443, 1969.
- 465 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.
- 469 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.
- 471 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.
- 474 L. Weng and F. Menczer. Topicality and impact in social media: Diverse messages, focused messengers. *PLoS ONE*, 10(2):e0118410, 2015.
- 476 L. Weng, A. Flammini, A. Vespignani, and F. Menczer. Competition among memes in a world with limited attention. *Sci. Rep.*, 2(335), 2012.
- 478 L. Weng, F. Menczer, and Y.-Y. Ahn. Virality prediction and community structure in social networks. *Sci. Rep.*, 3(2522), 2013a.
- 480 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.
- 484 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.
- 487 T. B. Wiggins, X. Gao, and J. Qiu. IndexedHBase, 2016. URL <http://salsaproj.indiana.edu/IndexedHBase>. Last accessed April 05, 2016.
- 489 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.
- 493 M. Zimmer. The Twitter archive at the library of congress: Challenges for information practice and information policy. *First Monday*, 20(7), 2015.