Design and Prototyping of a Social Media Observatory

Karissa McKelvey and Filippo Menczer
Center for Complex Networks and Systems Research
School of Informatics and Computing
Indiana University, Bloomington
cnets.indiana.edu

ABSTRACT

The broad adoption of online social networking platforms has made it possible to study communication networks at an unprecedented scale. With social media and micro-blogging platforms such as Twitter, we can observe high-volume data streams of online discourse. However, it is a challenge to collect, manage, analyze, visualize, and deliver large amounts of data, even by experts in the computational sciences. In this paper, we describe our recent extensions to Truthy, a social media observatory that collects and analyzes discourse on Twitter dating from August 2010. We introduce several interactive visualizations and analytical tools with the goal of enabling researchers to study online social networks with mixed methods at multiple scales. We present design considerations and a prototype for integrating social media observatories as important components of a web observatory framework.

Keywords

Visualization; Resource Data Management; Social Media Observatory; Web Observatory; API.

Categories and Subject Descriptors

H.3.7 [**Information systems**]: Digital libraries and archives; Data stream mining; H.5 [**Human-centered computing**]: Information visualization

1. INTRODUCTION

The World Wide Web is a new frontier in the study of human behavior. People volunteer information about themselves which has been historically expensive and time consuming to collect. From social media to mobile phone networks, "computational social science" has advanced knowledge and understanding in the behavioral [?], political [?], and economic [?] sciences [?]. This has resulted in a clamor for access to "big data" about people and their interactions as a primary source for interesting questions about human social organization [?].

Despite the promises of these advances, there are various limitations to using social networks as a primary data source. It is often difficult or expensive for researchers trained in the social sciences to utilize the technological expertise required to collect, manage, filter, visualize, and analyze large amounts of data. Research is also hard to reproduce when performed on a wide variety of datasets gathered with custom toolkits. Thus, it is beneficial for researchers to utilize a centralized platform. It is also important that any such platform be free or inexpensive, unlike for-profit social media analytics services such as GNIP (`gnip.com`), as researchers often operate within limited budgets.

We argue that "social media observatories" can leverage information visualization as well as open access to data and statistics to provide a natural solution to these problems. In this paper, we introduce several key design concerns for social media observatories. We then describe extensions of a currently operational social media observatory prototype, the *Truthy* project (`truthy.indiana.edu`) [?]. We deploy real-time, interactive visualizations of information diffusion processes on Twitter. We have created an interactive dashboard to address the needs of researchers, journalists, and others interested in qualitative and quantitative inquiry. Finally, we introduce the Truthy API, which allows users to download derived statistical data for the last 90 days of Twitter posts related to politics, social movements, and news.

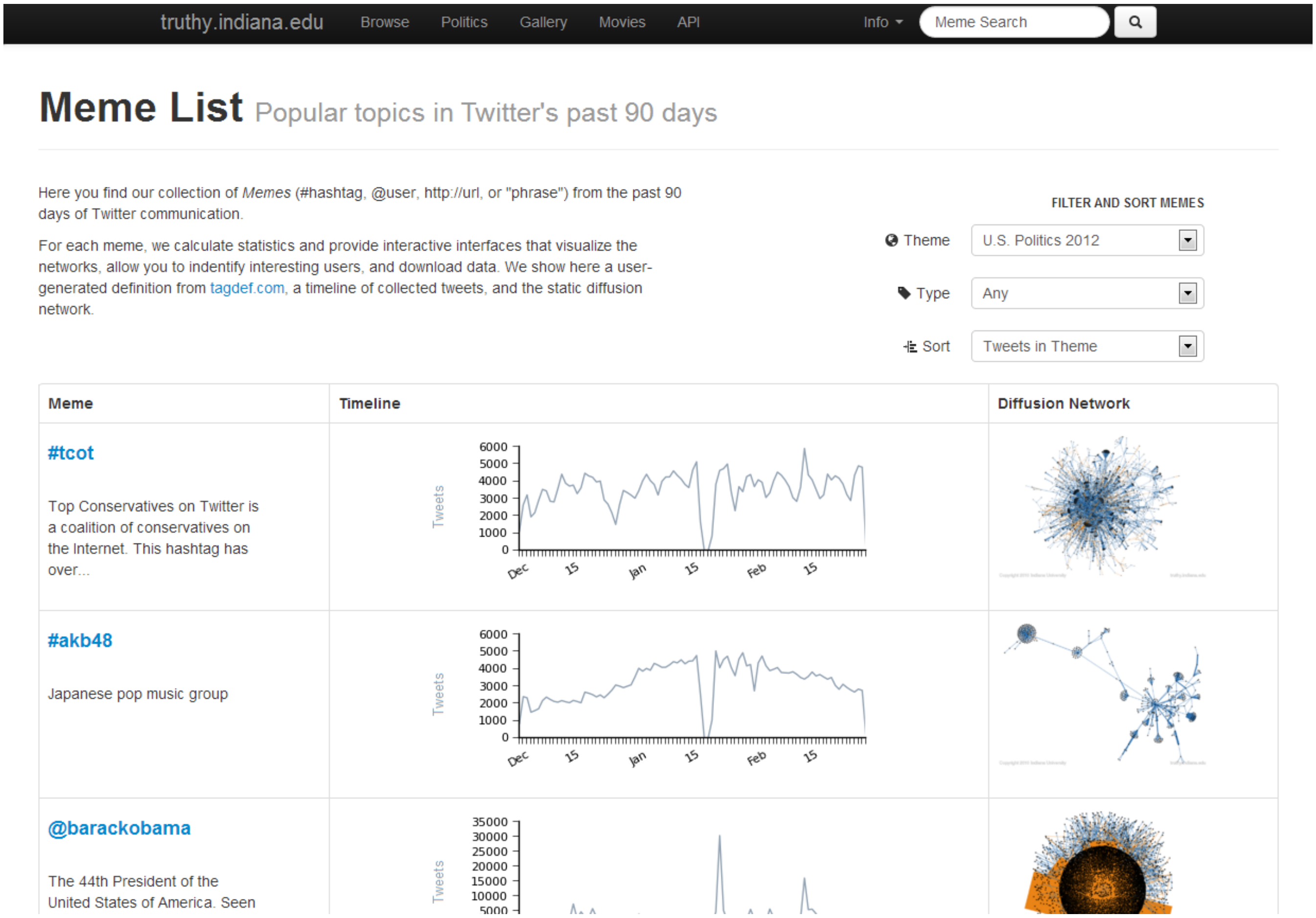
2. RELATED WORK

140kit (`140kit.com`) is a platform that aims to make the collection and analysis of Twitter data accessible to researchers. Users can track streams by terms, users, or locations, within a time frame of up to one week. This does not allow one to track a stream over a long period of time, nor to compare different streams. The Twitter datasets are available for analysis using a set of predefined offerings, such as a basic histogram, interaction list, and conversational network graph. Users can browse these visualizations publicly, yet do not have access to download the raw data. At the time of this writing, 140kit is in the process of shutting down due to lack of funding.

VisPolitics (`vispolitics.com`) contains a set of visualization projects, the most relevant of which are the so-called "debate and election tweet meters." These projects collect data in real time from the Twitter API and provide visualizations of tweeting activity as well as a computed "winning index" and sentiment over time. Although these interactive charts are only active during political debates and elections, they provide an example of how an observational lens into social media activity can be tailored for a specific use case.

Several commercial entities, including PeopleBrowsr (`peoplebrowsr.com`), Datasift (`datasift.com`), and SocialFlow (`socialflow.com`), provide data and analytics

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.



Meme List Popular topics in Twitter's past 90 days

Here you find our collection of *Memes* (#hashtag, @user, http://url, or "phrase") from the past 90 days of Twitter communication.

For each meme, we calculate statistics and provide interactive interfaces that visualize the networks, allow you to indentify interesting users, and download data. We show here a user-generated definition from tagdef.com, a timeline of collected tweets, and the static diffusion network.

FILTER AND SORT MEMES

Theme: U.S. Politics 2012

Type: Any

Sort: Tweets in Theme

Meme	Timeline	Diffusion Network
#tcot Top Conservatives on Twitter is a coalition of conservatives on the Internet. This hashtag has over...		
#akb48 Japanese pop music group		
@barackobama The 44th President of the United States of America. Seen		

Figure 1: Theme detail page for *U.S. Politics*. One can click a meme to reach its detail page.

from multiple social media platforms. Their primary audiences include enterprise ventures, financial industries, business intelligence, and government. The methods of delivery vary from web-based access and export of data to consulting services. These services include raw data access and computation of various measures, such as influence, sentiment, and information flows [?].

TwitInfo (`twitinfo.csail.mit.edu`) is a website displaying network analysis and visualizations of Twitter data. Its content is collected in automatically identified "bursts" of tweets [?]. Twitinfo also calculates the top tweeted URLs in each burst, and plots each tweet on a map, colored according to sentiment. Twitinfo focuses on specific memes identified by the researchers, and is thus somewhat limited for users who might wish to investigate arbitrary topics.

Ripples (`plus.google.com/ripple/details`) is a feature of the Google+ social network, useful for visualizing the spread of reposts among users (similar to Twitter's retweets). When a user reposts content, Ripples tracks the intermediate users along the diffusion path, information which is not available via the Twitter API. Users are represented as colorful bubbles, which recursively contain the intermediate users who have also shared the post, and are scaled according to estimated influence. This visualization is similar to the diffusion networks in the Truthy system [?, ?].

3. SOCIAL MEDIA OBSERVATORIES

Our proposed goal for a social media observatory is to collect large-scale online social network data and organize it for easy access. Bearing qualitative resemblance to observatories used in the natural and physical sciences, social media observatories can stimulate the rigorous use of large-scale data for studies of human behavior. The target users include researchers, journalists, educators, analysts, and the general public.

A proper social media observatory should facilitate and support the use of mixed methods approaches. Users proficient in qualitative analysis should benefit from interactive visualizations that allow them to identify key topics, users, and events in the social media stream. Those interested in quantitative data analysis may want access to statistics from real-time and historical data streams.

We have identified a series of design goals and concerns for social observatories, outlined below.

Reliablity. Even when large-scale data sets are collected, the reliability and stability of that data is often varied or unknown. Spam and misinformation may add noise, often originating from compromised accounts of otherwise legitimate users [?, ?, ?]. A social observatory should strive to legitimize data through cleansing and tagging by social, algorithmic, or other means. This is an ongoing research area and will require exploratory approaches. The observatory should also expose the consequences (e.g., bias) of any sampling method.

Reproducibility. Research is often difficult to reproduce when performed on a wide variety of possibly proprietary datasets, gathered with custom toolkits. Social observatories should attempt to mitigate this problem by offering a standard ontology for the storage, reference, and transfer of these datasets between users. This will improve the quality of research work and encourage the investigation of behaviors and events from multiple interdisciplinary and methodological perspectives.

Topic Filtering. To analyze the social network surrounding a specific event or topic, users may initially employ a method such as keyword search. Problems with such approaches include the need for prior knowledge and the risk of selection bias, whereby a query may be difficult to produce and/or reflect the preconceptions of the user. It is therefore desirable for a social observatory to provide data-driven views of topically consistent information clusters [?]. Users should be able to browse these topical communities. This will facilitate the observation of naturally emergent patterms.

Visualization. Visualizations can provide an intuitive lens to help users sift through large volumes of data for understanding observations at multiple scales through human vision. Information visualization pioneer Ben Schneiderman identifies several key features of an information visualization tool [?]. Such a tool should allow users to gain an *overview* of the data under study, provide *zoom* & *filtering* capabilities, item-level *details-on-demand*, allow users to see *relationships* among items in a collection, and *extract* target data about specific subsets within the collection. For social observatories to be useful, the interface between the user and data should benefit from these same design concerns.

Open Access. Even for those in the computational sciences, the sheer cost of technological infrastructure makes it difficult to deploy large-scale, scalable data collection and management systems. The start-up costs are often insurmountable for smaller labs. Commercial social media analytics companies offer such services; however, they are not always accessible as researchers often operate within limited budgets. A public and free — or low-cost — social observatory enables access to large-scale social media data analytics for non-profit endeavors.

Legal Compliance. Social media observatories will need to consider the risks of providing sensitive or protected content, such as profile information. They must ensure compliance with the terms of service of online social media platforms by only providing derived data, such as social network structure features and vector representations of textual content.

4. THE TRUTHY PLATFORM

The *Truthy* system was originally designed to analyze and detect the emergence of coordinated misinformation campaigns on Twitter [?, ?]. Now tasked with the study of information diffusion in general, Truthy monitors a real-time, high-throughput feed of 140-character messages known as *tweets*. The data source is a sample of approximately 10% of public tweets obtained from the Twitter streaming API (`dev.twitter.com`).

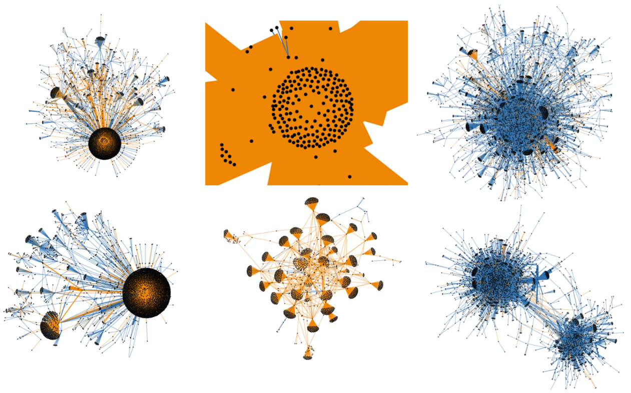


Figure 2: Diffusion networks associated with different Twitter memes. Nodes represent individual users and edges show how these memes spread from user to user by way of mentions (orange) and retweets (blue). In clockwise order: `@whitehouse`, `#nightclub`, `#tcot`, `#syria`, `#rsvp`, and `@michelleobama`.

In the current prototype, Truthy first filters content based on broad themes (*politics*, *social movements*, and *news* as of this writing). This filter is implemented as a run-time, asynchronous keyword-matching process. Further details can be found in previous work [?]. The system then clusters the selected tweets into groups of related messaged called ‘memes.’ Memes typically correspond to discussion topics, communication channels, or information resources shared among Twitter users. Memes are useful to provide users with an identifiable unit of information transfer for filtering content. Here we outline the criteria for the grouping of content into memes:

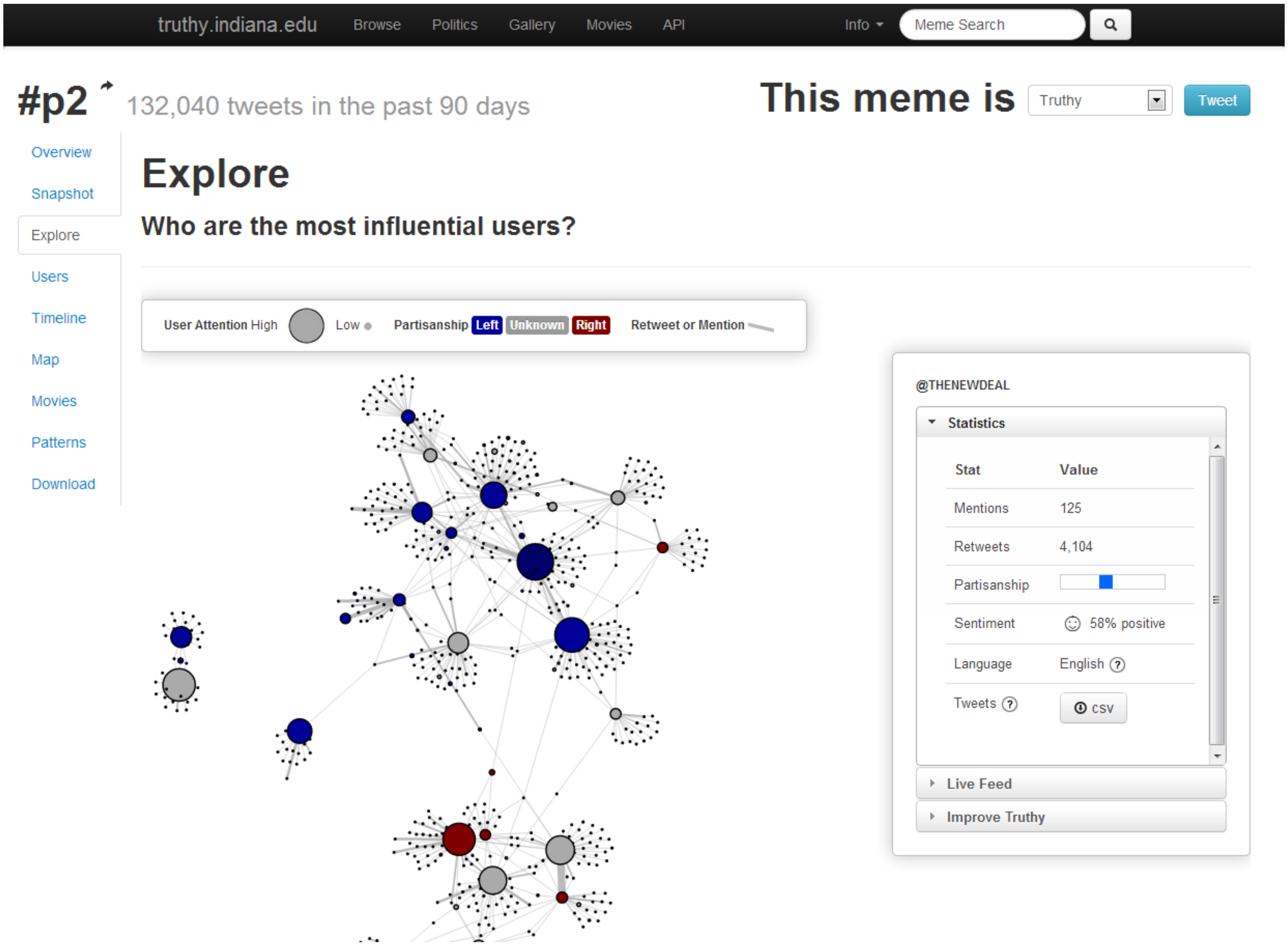
Hashtags. Hashtags are tokens used to identify the topic or intended audience of a tweet. Examples are `#taxes` and `#news`.

Mentions. A Twitter user can include another user’s screen name in a post, prefixed with the ‘@’ symbol, for example `@barackobama`. These mentions are used to denote that a particular Twitter user is being discussed or to address a post to that user.

Hyperlinks. We extract URLs from tweets by matching strings of valid URL characters that begin with `http://`.

Phrases. Finally, we consider the entire text of the tweet itself to be a meme once all Twitter metadata (hashtags and mentions), punctuation, and URLs have been removed. Substrings of tweets may also be matched.

Let us operationally define a meme as the set of all tweets containing a common hashtag, mentioned username, hyperlink, or phrase. Using memes as the atomic units of information transfer, we are able to create large-scale, high-resolution models of information propagation dynamics.



Stat	Value
Mentions	125
Retweets	4,104
Partisanship	
Sentiment	☺ 58% positive
Language	English ?
Tweets ?	csv

Figure 3: Interactive diffusion network visualization and user data exploration interface for content associated with the #p2 meme, or "Progressives 2.0." On the top right, the push-button tool allows one to tweet about a particular meme. Options include "Truthy" and "Spam."

On the Truthy website, users can browse and search through collections of theme-related memes. Users can sort memes based on a variety of statistical features (Fig. ??). For each meme, the user is presented with an interactive dashboard containing a crowdsourced definition (from `tagdef.com`), a high-resolution image of the meme's information diffusion network (Fig. ??), and various interactive visualizations and statistics. In Section ?? we introduce an API that allows one to access user- and meme-level statistics.

In an attempt to improve the reliability of the data, Truthy facilitates social tagging of suspicious content, such as spam and astroturf. Through the meme detail interface (Fig. ??), one can tweet about a meme or user (Fig. ??) in a syntax that can be automatically parsed by our system. We collect these posts for future analysis of the reliability of crowdsourced data in a social media observatory.

5. SOCIAL OBSERVATORY ELEMENTS

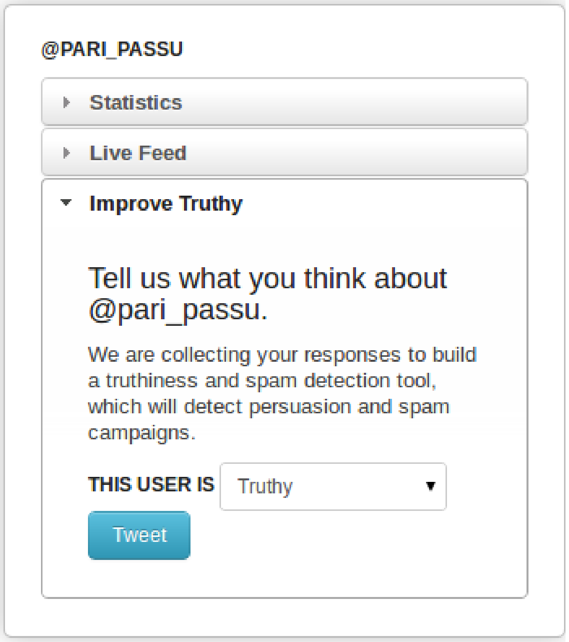
5.1 Computed Statistics

We compute a number of descriptive statistics for each meme and user. For memes, we compute the number of tweets and users, as well as sentiment and network properties. For users, we compute their total tweets, retweets, mentions, probable language, date of most recent activity, and account creation date. For users with tweets in the *U.S. Politics* theme, we also predict their political partisanship. Finally, we compute the sentiment associated with each meme tweeted by a user.

Sentiment is calculated using OpinionFinder, a system that performs emotional valence analysis by searching for substrings in a text to identify phrases that express positive or negative sentiment [?]. We attempt to identify the dominant language of each Twitter user with the Compact Language Detector library, developed by Google for use in the Chrome web browser.[1] This library makes use of character-level n-grams.

To infer the partisan leanings of individual users we apply a machine-learning algorithm that leverages network and text features to make predictions of political ideology [?]. This technique relies on the highly-clustered structure and hashtag usage of a political retweet network. Using a data

[1] `code.google.com/p/chromium-compact-language-detector`



Tell us what you think about @pari_passu.

We are collecting your responses to build a truthiness and spam detection tool, which will detect persuasion and spam campaigns.

THIS USER IS Truthy

Figure 4: Interface one sees after clicking on a user in the interactive network.

set of 1,000 manually-annotated users, the authors found that membership in a specific network cluster predicted political affiliation with 87.3% accuracy.

5.2 Interactive Diffusion Network

The interactive diffusion network, designed for qualitative inquiry, allows users to identify the accounts associated with interesting nodes. To allow users to investigate the most important users, we filter the complete network layout to include only the 25 most retweeted users and their neighbors. This allows for a detailed examination of the relationships between the most prominent individuals in the network (see Fig. **??**).

A directed edge from node i to node j in the interactive diffusion network represents at least one event in which j retweets i, and the width of the link increases with the number of retweets between the users. Nodes have an area that scales logarithmically with out-degree and are colored according to inferred political partisanship (for memes about *U.S. Politics*). A user can hover over any node in the network to see the associated account name, and clicking a node brings up an interface containing computed statistics about that user (depicted on the right-hand side of Fig. **??**).

The interactive diffusion network is created using *d3.js*, a JavaScript library that binds arbitrary data to a Document Object Model (DOM), and then applies transformations using Scalable Vector Graphics (`mbostock.github.com/d3`).

5.3 User Exploration Interface

In addition to the data exploration framework described above, we provide a filterable, sortable, and exportable data table built with the Google Chart Tools API (`code.google.com/apis/chart/interactive/docs/gallery/table.html`). In this interface (Fig. **??**), a data table is presented containing fields described in Section **??**. This interface allows users to investigate the behavioral and demographic characteristics of larger collections of individuals compared to the interactive diffusion network. This data can be exported in CSV (comma-separated values) format for further analysis, promoting reproducibility.

5.4 Application Programming Interface (API)

The Twitter Terms of Service prohibit third-party services from providing direct access to historical tweet content; requests for tweet content must be made through the official Twitter API. One of the key demands of our users is the ability to access and export historical social media data. To this end, we offer a public `REST` API to access our meme and user statistics, as well as static diffusion network images. We currently offer our API through Mashape (`mashape.com/truthy/public`). Mashape offers modules in various languages that one can use to facilitate programmatic access the API, especially for non-technical users.

The API provides the following endpoints: meme communication networks in gexf, graphml, edgelist, and adjlist; meme statistics including number of events, users, and component size; and user statistics such as political partisanship, sentiment, and activity levels. Each meme and user is identified by a unique, anonymous ID. One can query by meme text and sort by various statistics. We allow 150 requests per hour at 1,000 results per request.

6. CONCLUSION

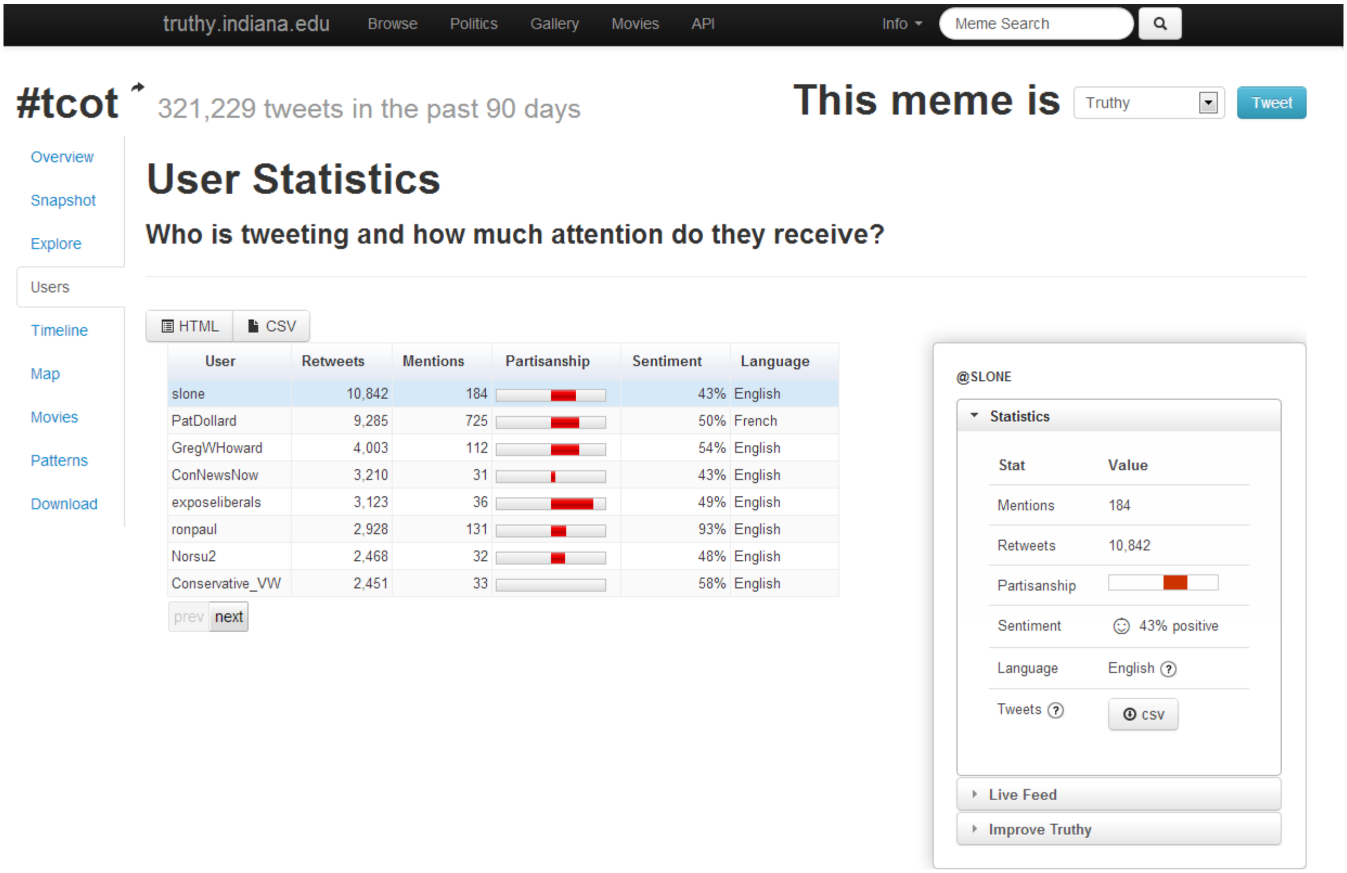
We would like to provide a broader set of historical data while improving our visualizations. These initiatives include expanding the API to include all of the tweets we have collected since August 2010. Our current infrastructure does not scale to this expanded collection. We plan to build a new distributed storage infrastructure to handle flexible and fast search capability. To this end we are leveraging generous support from the scalable infrastructure offered by FutureGrid [**?**].

Regarding the user interface and API, we plan to enable user-defined themes, thus expanding visualizations to customized content. Furthermore, we will provide more statistics on commonly occurring words and n-grams, and improve the interactive network visualizations to highlight meme diffusion dynamics. Finally, we plan to conduct interviews to measure the impact of social media observatories.

We have outlined design approaches and a prototype of a social media observatory. As increasing amounts of data on online discourse and deliberation are collected, we see a rising demand for technologies that make these data accessible to the broader community. The observatory elements described in this paper are all currently available on the Truthy website, and we encourage readers to explore them and provide feedback.

Acknowledgements.

We would like to thank Clayton Davis, Michael Conover, Jacob Ratkiewicz, Bruno Goncalves, Mark Meiss, Alessandro Flammini, Johan Bollen, Alessandro Vespignani, and other current and past members of the Center for Complex Networks and Systems Research (`cnets.indiana.edu`) at Indiana University for helpful discussions and contributions



#tcot 321,229 tweets in the past 90 days

This meme is Truthy

User Statistics

Who is tweeting and how much attention do they receive?

User	Retweets	Mentions	Partisanship	Sentiment	Language
slone	10,842	184	🟥	43%	English
PatDollard	9,285	725	🟥	50%	French
GregWHoward	4,003	112	🟥	54%	English
ConNewsNow	3,210	31	🟥	43%	English
exposeliberals	3,123	36	🟥	49%	English
ronpaul	2,928	131	🟥	93%	English
Norsu2	2,468	32	🟥	48%	English
Conservative_VW	2,451	33		58%	English

@SLONE

Statistics

Stat	Value
Mentions	184
Retweets	10,842
Partisanship	🟥
Sentiment	☺ 43% positive
Language	English ?
Tweets ?	csv

Figure 5: Filterable, sortable, and exportable data table for users associated with the #tcot meme, a conservative communication channel.

to the Truthy Project. We gratefully acknowledge partial support from the National Science Foundation (grant CCF-1101743) and the J. S. McDonnell Foundation.

7. REFERENCES

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

[2] D. Boyd and K. Crawford. Six Provocations for Big Data. *SSRN Electronic Journal*, pages 1–17, 2011.

[3] M. D. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer. Predicting the political alignment of twitter users. In *Proceedings of the IEEE Third International Conference on Social Computing (SocialCom)*, pages 192–199. IEEE, Oct. 2011.

[4] M. D. Conover, J. Ratkiewicz, B. Gonçalves, A. Flammini, and F. Menczer. Political polarization on twitter. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media 2011*, 2011.

[5] N. Eagle, A. S. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278, 2009.

[6] G. Fox, G. von Laszewski, J. Diaz, K. Keahey, J. Fortes, R. Figueiredo, S. Smallen, W. Smith, and A. Grimshaw. FutureGrid — A reconfigurable testbed for Cloud, HPC and Grid Computing. In J. S. Vetter, editor, *Contemporary High Performance Computing: From Petascale toward Exascale*, Computational Science. Chapman and Hall/CRC, 2013.

[7] B. Gonçalves, M. Conover, and F. Menczer. Abuse of social media and political manipulation. In M. Jakobsson, editor, *The Death of The Internet*. Wiley, 2012.

[8] C. Grier, K. Thomas, V. Paxson, and M. Zhang. @ spam : The Underground on 140 Characters or Less. In *Proceedings of the 17th ACM conference on Computer and communications security*, pages 27–37, Chicago, IL, USA, 2010. ACM.

[9] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne. Computational social science. *Science*, 323(5915):721–723, 2009.

[10] G. Lotan. Mapping information flows on twitter. In *Fifth International AAAI Conference on Weblogs and Social Media*, 2011.

[11] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller. Twitinfo: Aggregating and visualizing microblogs for event exploration. In *ACM CHI Conference on Human Factors in Computing Systems*, 2011.

[12] 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 Companiono*, 2013.

[13] K. McKelvey, A. Rudnick, M. Conover, and F. Menczer. Visualizing Communication on Social Media: Making Big Data Accessible. In *Proc. 15th ACM Conference on Computer Supported Cooperative Work, Workshop on Collective Intelligence as Community Discourse and Action*, 2012.

[14] 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)*, 2011.

[15] J. Ratkiewicz, M. D. Conover, M. Meiss, B. Goncalves, A. Flammini, and F. Menczer. Detecting and tracking political abuse in social media. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, 2011.

[16] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on*, pages 336–343. IEEE, 1996.

[17] L. Weng, A. Flammini, A. Vespignani, and F. Menczer. Competition among memes in a world with limited attention. *Scientific Reports*, 2(335), 2012.

[18] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan. Opinionfinder: a system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, HLT-Demo '05, pages 34–35, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.