Truthy: Mapping the Spread of Astroturf in Microblog Streams

Jacob Ratkiewicz,* Michael Conover, Mark Meiss, Bruno Gonçalves,
Snehal Patil, Alessandro Flammini, Filippo Menczer

Center for Complex Networks and Systems Research
Pervasive Technology Institute
School of Informatics and Computing, Indiana University, Bloomington, IN, USA

ABSTRACT

Online social media are complementing and in some cases replacing person-to-person social interaction and redefining the diffusion of information. In particular, microblogs have become crucial grounds on which public relations, marketing, and political battles are fought. We demonstrate a web service that tracks political *memes* in Twitter and helps detect astroturfing, smear campaigns, and other misinformation in the context of U.S. political elections. We also present some cases of abusive behaviors uncovered by our service. Our web service is based on an extensible framework that will enable the real-time analysis of meme diffusion in social media by mining, visualizing, mapping, classifying, and modeling massive streams of public microblogging events.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database applications—*Data mining*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Information networks*; H.4 [**Information Systems Applications**]: Miscellaneous; K.4.1 [**Computers and Society**]: Public Policy Issues

General Terms

Measurement

Keywords

Twitter, Truthy, Memes, Information Diffusion, Social Media, Microblogs, Classification, Politics

1. INTRODUCTION

Social networking and microblogging services reach hundreds of million users and have become fertile ground for a variety of research efforts, since they offer an opportunity to study patterns of social interaction among a population larger than ever before. In particular, Twitter has recently generated much attention in the research community due to its enormous popularity and open policy on data sharing.

The growth in reach of microblogs is accompained by a surge in the amount of potentially useful information that can be mined from their data streams. However, as microblogs become valuable media to spread information, e.g., for marketeers and politicians, it is natural that people find ways to abuse them. As a result, we observe various types of illegitimate use, such as spam [7, 2, 6]. In this paper we focus on one particular type of abuse, namely *political astroturf* — campaigns disguised as spontaneous, popular "grassroots" behavior that are in reality carried out by a single person or organization. This is related to spam but with a more specific domain context, and with potentially larger consequences.

A recent case study provides an example of the kind of abuse we wish to detect [4]. This work describes a concerted, deceitful attempt to cause a link to a smear site to gain wide distribution on Twitter. Using a small number of fake accounts, the perpetrators of the attack were able to cause a link to be retweeted (by legitimate users) enough to appear in Google real-time search results for a political candidate name. This type of attack, which can be mounted very cheaply and could potentially reach an even larger audience than traditional advertisements, will certainly be used again.

While some of the techniques associated with spam (such as the mass creation of accounts meant to look like real users) are shared with political astroturfing, spam and astroturf differ in several ways. Spammers are often interested in causing users to click a link; in contrast, astroturfers want a particular tweet or idea to have a false sense of group consensus. Further, many of the users involved in propagating a successfully astroturfed message may in fact be legitimate users, unwittingly complicit in the deception, having been themselves deceived by the original core of automated accounts. Thus, astroturf detection methods cannot only rely on a message's content, or on the accounts of the users who propagate it.

In light of these characteristics of political astroturf, we need a definition that allows us to discriminate such falsely-propagated information from grassroot generated and organically diffused information. We decided to borrow a term, *truthy,* to describe political astroturf memes. The term was coined by comedian Stephen Colbert to describe something that a person claims to know based on emotion rather than evidence or facts. Here we introduce a software system and associated website (`truthy.indiana.edu`), which we collectively call 'Truthy,' designed to aid in the identification and analysis of truthy memes in the Twitter stream. For further details see [5].

2. ANALYTICAL FRAMEWORK

Let us first describe the analytical framework at the base of our web service. A main motivation beyond the development of such framework is that social media analysis presents major challenges in the area of data management, particularly when it comes to interoperability, curation, and consistency of process. Due to diversity among site designs, data models, and APIs, any analytical tools

*Corresponding author. Email: `jpr@cs.indiana.edu`

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

written by researchers to address one site are not easily portable to another. To focus on the common features of all social media and microblogging sites, we developed a unified framework, which we call *Klatsch,* that makes it possible to analyze the behavior of users and diffusion of ideas in a broad variety of data feeds. The Klatsch framework is designed to provide data interoperability for the real-time analysis of massive social media data streams (millions of posts per day) from sites with diverse structures and interfaces. To support this, we developed a new, unified model for social networking data as a series of interactions between *actors* and *memes.*

In the Klatsch model, social networking sites are sources of a timestamped series of events. Each event involves some number of actors (representing individual users), some number of memes (representing units of information), and interactions among those actors and memes. For example, a single Twitter post might constitute an event involving three or more actors: the poster, the user she is retweeting, and the people she is addressing. The post might also involve a set of memes consisting of hashtags and URLs referenced in the tweet. Each event can be thought of as contributing a unit of weight to edges in a network structure, where nodes represent either actors or memes. This is not a strictly bipartite network: actors can be linked through replying or mentioning, and memes by concurrent discussion or semantic similarity. The timestamps associated with the events allow us to observe the changing structure of this network over time.

2.1 Meme Types

To study the diffusion of information on Twitter it is necessary to single out features that can be used to identify a specific topic as it propagates. We chose to forgo sophisticated topic modeling techniques and focus on features unique to Twitter data which can be used as topic markers — *hashtags* and *mentions*. Hashtags are tokens, included in the text of a tweet and prefixed by a hash (#), that are used to label the topical content of tweets. Some examples of popular tags are `#gop` and `#obama`, marking discussion about the Republican party and President Obama, respectively. A Twitter user can call another user's attention to a particular post by including that user's screen name in the post, prepended by the `@` symbol. These *mentions* can be used as a way to carry on conversations between users, or to denote that a particular Twitter user is being discussed. Besides hashtags and mentions, we also use URLs as topical markers, as well as the text of the tweet itself when all URLs and metadata markup have been removed. Information about the propagation of each of these types of memes is used to build networks representing the diffusion of information among users.

2.2 Network Edges

To represent the flow of information through the Twitter community we construct a directed graph in which nodes are individual user accounts. An edge is drawn from node A to B when either B retweets a message from A, or A mentions B in a tweet, with the weight of the edge representing the number of occurrences of the associated event. In both cases we can infer that some information has flowed along the direction of the edge.

We rely on Twitter metadata provided along with the tweet to determine the users mentioned or retweeted, rather than parsing the text of the tweet itself. Thus, while the text of the tweet may contain several mentions, we consider the user identified in the metadata to be the user to whom an edge is drawn in the network. Note that this is separate from our use of mentions as memes (§ 2.1), which we parse from the text of the tweet.

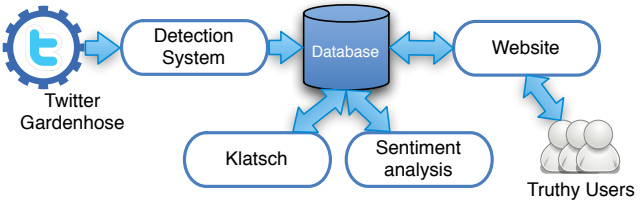


Figure 1: The Truthy system architecture.

3. TRUTHY SYSTEM ARCHITECTURE

A general overview of the components of the Truthy system is shown in Fig. 1. These are in several major parts: a low-level system overseeing the collecting and processing of the raw data feeds from the Twitter API, the meme detection framework, the Klatsch framework responsible for computing key network statistics and layouts, and a web based presentation framework that allows us to collect user input on which memes the community deems most suspicious. These components are described next.

3.1 Streaming Data Collection

To collect meme diffusion data we rely on access to the Twitter 'gardenhose.' This streaming API provides detailed data on a sample of the Twitter corpus at a rate that varied between roughly 4 million tweets a day near the beginning of our study, to around 8 million tweets per day at the time of this writing. We note that the process of sampling edges from a network to determine its structure has its drawbacks [3]; however, we found that this sampling is still useful for our purposes. All collected tweets are stored in files at a daily time resolution. We maintain files both in a verbose JSON format containing all the features provided by Twitter, and in a more compact format that contains only the features used in our analysis. This collection is accomplished by a component of our system that operates asynchronously from the others.

3.2 Meme Detection

A second component of our system is devoted to scanning the tweets we collect in real time. The task of this meme detection component is to identify the tweets (for further processing) that (a) have content related to the political elections, and (b) are of sufficiently general interest. We implemented a filtering step for each of these criteria, as described below.

To identify politically relevant tweets, we used a hand-curated collection of approximately 2500 keywords relating to the 2010 U.S. midterm elections. This keyword list contains the names of all candidates running for federal office, as well as any common variations, known Twitter account usernames, and some popular political hashtags. This *tweet_filter* component also operates asynchronously.

To identify memes that were used often enough to be considered of general interest, we implemented a second stage of filtering — the *meme_filter*. This filter first extracts all memes (of the types described in § 2.1) from each incoming tweet, and tracks the activation over the past hour of each meme, in real time. If any meme exceeds a rate threshold of five mentions in a given hour it is considered 'activated;' any tweets containing that meme are then stored. If a tweet contains a meme that is already considered activated due to its presence in previous tweets, it is stored immediately. When the mention rate of the meme drops below the activation limit, it is no longer considered activated and tweets containing the meme are no longer automatically stored. Note that a tweet can contain

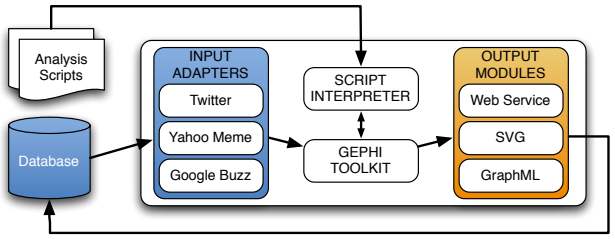


Figure 2: The Klatsch framework architecture.

more than one meme, and thus the activation of multiple memes can be triggered by the arrival of a single tweet. We chose a low rate threshold with the understanding that if a meme is observed five times in our sample it is likely mentioned many more times in Twitter at large.

Our system tracked approximately 305 million tweets collected from September 14 until October 27, 2010. Of these, 1.2 million contained one or more of our political keywords; detection of general interest memes further reduced this set to 600,000 tweets actually entered in our database for analysis.

3.3 Network Analysis

The Klatsch framework is responsible for network analysis and layout for visualization of the meme diffusion patterns. It consists of several components (Fig. 2): a set of input adapters for importing external social network data into the Klatsch data model; support for a variety of standard graph layout and visualization algorithms; a flexible scripting language for coding site-agnostic analysis modules; and a set of export modules, including an embedded lightweight web server, for visualizing analysis, saving statistical results, supporting interactive web tools, and producing publication-ready graphs and tables.

Klatsch includes a domain-specific scripting language with advanced features such as first-order functions, streams, and map/filter/reduce primitives. For instance, the inclusion of streams as a first-order data type supports the lazy evaluation of algorithms that operate on the nodes and edges of large graphs. Our graph analysis and visualization algorithms are implemented in this language.

To characterize the structure of the diffusion network we compute several statistics based on the topology of its largest connected component. These include the number of nodes and edges in the graph, the mean degree and strength of nodes in the graph, mean edge weight, clustering coefficient of the largest connected components, and the standard deviation and skew of each network's in-degree, out-degree and strength distributions. Additionally we track the out-degree and out-strength of the most prolific broadcaster, as well as the in-degree and in-strength of the most focused-upon user. We also monitor the number of unique injection points of the meme in the largest connected component, reasoning that organic memes (such as those relating to news events) will be associated with larger number of originating users. Finally, we track the mood associated with each meme using a modified version of the Google-based Profile of Mood States (GPOMS) sentiment analysis method [1].

3.4 Web Interface

The final component of our analytical framework includes a dynamic web interface to allow users to inspect memes through various views, and annotate those they consider to be truthy. The site provides a mixed presentation of statistical information and interactive visualization elements. Fig. 3 provides a snapshot of the summary view for several memes.

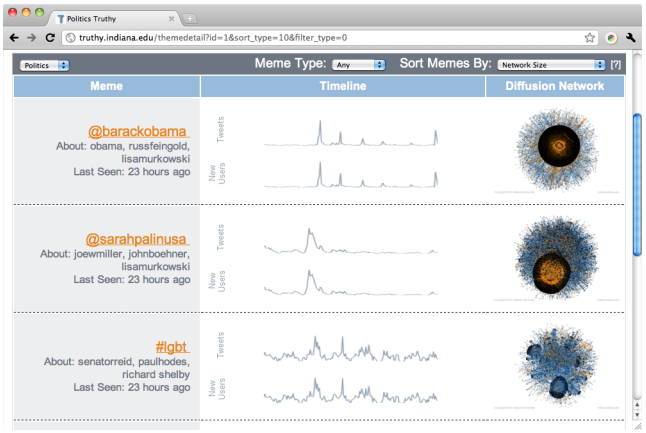


Figure 3: Screenshots of the Truthy meme overview page.

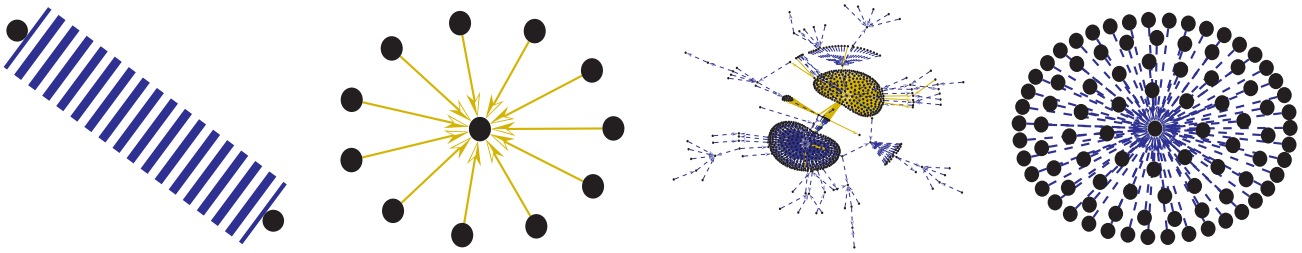
Users who wish to explore the Truthy database using the web interface can sort memes according to a variety of ranking criteria, including the size of the largest connected component, number of user annotations, number of users, number of tweets, number of tweets per user, number of retweets, and number of meme injection points. This list-based presentation of memes functions as a concise, high-level view of the data, allowing users to examine related keywords, time of most recent activity, tweet volume sparklines and thumbnails of the information diffusion network. At this high level users can examine a large number of memes quickly and subsequently drill down into those that exhibit interesting behavior.

Once a user has selected an individual meme for exploration, she is presented with a more detailed presentation of statistical data and interactive visualizations. Here the user can examine the statistical data described above, tweets relating the meme of interest, and sentiment analysis data. Additionally users can explore the temporal data through an interactive annotated timeline, inspect a force-directed layout of the meme diffusion network, and view a map of the tweet geo-locations. Upon examining these features, the user can choose to mark the meme as 'truthy.' These crowdsourced annotations are used, along with the other features listed above, to train a classifier for automatic detection of truthy memes [5].

4. EXAMPLES OF TRUTHY MEMES

Truthy allowed us to identify several suspicious memes. Some of these cases caught the attention of the popular press due to the sensitivity of the topic in the run up to the midterm political elections, and subsequently many of the accounts involved were suspended by Twitter. Below are a few representative examples.

`@PeaceKaren_25` This account has generated a very large number of tweets (over 10,000 in four months), almost all supporting one of a few candidates. Another account, `@HopeMarie_25`, behaved similarly, supporting the same candidates and boosting the same websites. It did not produce any original tweets, instead retweeting all those of `@PeaceKaren_25`. A visualization of the interaction between these two accounts can be see in Fig. 4(a). Both accounts were suspended by Twitter by the time of this writing.



(a) @PeaceKaren_25

(b) "How Chris Coons budget works- uses tax $ 2 attend dinners and fashion shows"

(c) @senjohnmccain

(d) "Obama said taxes have gone down during his administration. That's ONE way to get rid of income tax — getting rid of income."

Figure 4: Diffusion networks of sample memes from our dataset. Blue dashed edges represent retweets, while orange solid edges represent mentions. (a), (b): Truthy memes. (c), (d): Legitimate memes.

How Chris Coons budget works- uses tax $ 2 attend dinners and fashion shows

> This is one of a set of truthy memes smearing the Democratic candidate for U.S. Senate from Delaware. Looking at the injection points of these memes, we uncovered a network of about ten bot accounts. They inject thousands of tweets with links to posts from the `freedomist.com` website (to which the ownership of many of the accounts can be traced). To avoid detection by Twitter and increase visibility to different users, duplicate tweets are disguised by adding different hashtags and appending junk query parameters (ignored by many URL shorteners) to the URLs. The diffusion network corresponding to this case is illustrated in Fig. 4(b).

In addition to these, two other networks of bots were shut down by Twitter after being detected by Truthy. In one case we observed the automated accounts using text segments drawn from newswire services to produce multiple legitimate-looking tweets in between the injection of URLs. These instances highlight several of the more general properties of truthy memes detected by our system.

Fig. 4 also shows the diffusion networks for two legitimate memes. One, `@senjohnmccain`, displays two different communities in which the meme was propagated: one by retweets from `@ladygaga` in the context of discussion on the repeal of the "Don't ask, don't tell" policy on gays in the military, and the other by mentions of `@senjohnmccain` (Fig. 4(c)). Our last legitimate example is a network of retweets of a single popular user (Fig. 4(d)). A gallery with details about various truthy and legitimate memes can be found online (`truthy.indiana.edu/gallery`).

5. DISCUSSION

We demonstrated a system for the real-time analysis of meme diffusion from microblog streams. The Klatsch framework will soon be released as open source. We described the Truthy system and website, which leverage this framework to track political memes in Twitter and help detect astroturfing campaigns in the context of U.S. political elections.

Despite the fact that some of the memes mentioned here are characterized by small diffusion networks, it is important to note that this is the stage at which such attempts at deception must be identified. Once one of these attempts is successful at gaining the attention of the community, it will quickly become indistinguishable from an organic meme. Therefore, the early identification and termination of accounts associated with astroturf memes is critical. For this reason we are developing a data mining framework for the automatic detection of astroturf, with promising preliminary results that are outside the scope of this demonstration [5].

In the future we intend to add more views to the website, including views on the users, such as the ages of the accounts, and tag clouds to interpret the sentiment analysis scores. Another important area to address is that of sampling bias, since the properties of the sample made available in the Twitter gardenhose are currently unknown. To explore this, we intend to track injected memes of various sizes and with different topological properties of their diffusion graphs.

Acknowledgments. We are grateful to A. Vespignani, C. Catutto, J. Ramasco, and J. Lehmann for helpful discussions, J. Bollen for his GPOMS code, T. Metaxas and E. Mustafaraj for inspiration and advice, and Y. Wang for web design support. We thank the Gephi toolkit for aid in our visualizations and the many users who have provided feedback and annotations. We acknowledge support from NSF (grant No. IIS-0811994), Lilly Foundation (Data to Insight Center Research Grant), the Center for Complex Networks and Systems Research, and the IU School of Informatics and Computing.

6. REFERENCES

[1] J. Bollen, H. Mao, and A. Pepe. Determining the public mood state by analysis of microblogging posts. In *Proc. of the Alife XII Conf.* MIT Press, 2010.

[2] C. Grier, K. Thomas, V. Paxson, and M. Zhang. @spam: the underground on 140 characters or less. In *Proc. of the 17th ACM Conf. on Computer and Communications Security (CCS)*, pages 27–37, 2010.

[3] J. Leskovec and C. Faloutsos. Sampling from large graphs. In *KDD '06: Proc. of the 12th ACM SIGKDD international Conf. on Knowledge discovery and data mining*, pages 631–636, New York, NY, USA, 2006. ACM.

[4] E. Mustafaraj and P. Metaxas. From obscurity to prominence in minutes: Political speech and real-time search. In *WebSci10: Extending the Frontiers of Society On-Line*, page 317, 2010.

[5] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer. Detecting and tracking the spread of astroturf memes in microblog streams. Technical Report arXiv:1011.3768 [cs.SI], CoRR, 2010.

[6] A. H. Wang. Don't follow me: Twitter spam detection. In *Proc. 5th International Conf. on Security and Cryptography (SECRYPT)*, 2010.

[7] S. Yardi, D. Romero, G. Schoenebeck, and danah boyd. Detecting spam in a Twitter network. *First Monday*, 15(1), 2009.