Detecting and Tracking Political Abuse in Social Media

J. Ratkiewicz, M. D. Conover, M. Meiss, B. Gonçalves, A. Flammini, F. Menczer
Center for Complex Networks and Systems Research
School of Informatics and Computing
Indiana University, Bloomington, IN, USA

Abstract

We study *astroturf* political campaigns on microblogging platforms: politically-motivated individuals and organizations that use multiple centrally-controlled accounts to create the appearance of widespread support for a candidate or opinion. We describe a machine learning framework that combines topological, content-based and crowdsourced features of information diffusion networks on Twitter to detect the early stages of viral spreading of political misinformation. We present promising preliminary results with better than 96% accuracy in the detection of astroturf content in the run-up to the 2010 U.S. midterm elections.

1 Introduction

Social networking and microblogging services reach hundreds of millions of users and have become fertile ground for a variety of research efforts. They offer a unique opportunity to study patterns of social interaction among far larger populations than ever before. In particular, Twitter has recently generated much attention in the research community due to its peculiar features, open policy on data sharing, and enormous popularity. The popularity of Twitter, and of social media in general, is further enhanced by the fact that traditional media pay close attention to the ebb and flow of the communication that they support. With this scrutiny comes the potential for the hosted discussions to reach a far larger audience than simply the original social media users. Along with the recent growth of social media popularity, we are witnessing an increased usage of these platforms to discuss issues of public interest, as they offer unprecedented opportunities for increased participation and information awareness among the Internet-connected public (Adamic and Glance 2005). While some of the discussions taking place on social media may seem banal and superficial, the attention is not without merit. Social media often enjoy substantial user bases with participants drawn from diverse geographic, social, and political backgrounds (Java et al. 2007). Moreover, the user-as-information-producer model provides researchers and news organizations alike with a means of instrumenting and observing a representative sample of the population in real time. Indeed, it has been recently demonstrated that useful information can be mined from Twitter data streams(Asur and Huberman 2010; Tumasjan et al. 2010; Bollen, Mao, and Zeng 2011).

With this increasing popularity, however, comes a dark side — as social media grows in prominence, it is natural that people find ways to abuse it. As a result, we observe various types of illegitimate use; spam is a common example (Grier et al. 2010; Wang 2010). Here we focus on a particular social media platform, Twitter, and on one particular type of abuse, namely *political astroturf* — political campaigns disguised as spontaneous "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 potentially larger consequences.

Online social media tools play a crucial role in the successes and failures of numerous political campaigns and causes. Examples range from the grassroots organizing power of Barack Obama's 2008 presidential campaign, to Howard Dean's failed 2004 presidential bid and the first-ever Tea Party rally (Rasmussen and Schoen 2010; Wiese and Gronbeck 2005).

The same structural and systemic properties that enable social media such as Twitter to boost grassroots political organization can also be leveraged, even inadvertently, to spread less constructive information. For example, during the political campaign for the 2010 midterm election, several major news organizations picked up on the messaging frame of a viral tweet relating to the allocation of stimulus funds, succinctly describing a study of decision making in drug-addicted macaques as "Stimulus $ for coke monkeys" (The Fox Nation 2010).

While the "coke monkeys" meme developed organically from the attention dynamics of thousands of users, it illustrates the powerful and potentially detrimental role that social media can play in shaping public discourse. As we will demonstrate, a motivated attacker can easily orchestrate a distributed effort to mimic or initiate this kind of organic spreading behavior, and with the right choice of inflammatory wording, influence a public well beyond the confines of his or her own social network.

Unlike traditional news sources, social media provide little in the way of individual accountability or fact-checking mechanisms. Catchiness and repeatability, rather than truthfulness, can function as the primary drivers of information diffusion. While flame wars and hyperbole are hardly new phenomena online, Twitter's 140-character sound bytes are ready-made headline fodder for the 24-hour news cycle.

In the remainder of this paper we describe a system to analyze the diffusion of information in social media, and, in particular, to automatically identify and track orchestrated, deceptive efforts to mimic the organic spread of information through the Twitter network. The main contributions of this paper are very encouraging preliminary results on the detection of suspicious memes via supervised learning (96% accuracy) based on features extracted from the topology of the diffusion networks, sentiment analysis, and crowdsourced annotations. Because what distinguishes astoturf from true political dialogue includes the way they are spread, our approach explicitly takes into account the diffusion patterns of messages across the social network.

2 Background and Related Work

2.1 Information Diffusion

The study of opinion dynamics and information diffusion in social networks has a long tradition in the social, physical, and computational sciences (Castellano, Fortunato, and Loreto 2009; Barrat, Barthelemy, and Vespignani 2008; Leskovec, Adamic, and Huberman 2006; Leskovec, Backstrom, and Kleinberg 2009). Twitter has recently been considered as case study for information diffusion. For example, Galuba et al. (2010) take into account user behavior, user-user influence, and resource virulence to predict the spread of URLs through the social network. While usually referred to as 'viral,' the way in which information or rumors diffuse in a network has important differences with respect to infectious diseases (Morris 2000). Rumors gradually acquire more credibility as more and more network neighbors acquire them. After some time, a threshold is crossed and the rumor is believed to be true within a community.

A serious obstacle in the modeling of information propagation in the real world as well as in the blogosphere is the fact that the structure of the underlying social network is often unknown. When explicit information on the social network is available (e.g. the Twitter's follower relations) the strength of the social links are hardly known and their importance cannot be deemed uniform across the network (Huberman, Romero, and Wu 2008). Heuristic methods are being developed to face this issue. Gomez-Rodriguez, Leskovec, and Krause (2010) propose an algorithm that can efficiently approximate linkage information based on the times at which specific URLs appear in a network of news sites. For the purposes of our study such problem can be, at least partially, ignored. Twitter provides an explicit way to follow the diffusion of information via the tracking of *retweets*. This metadata tells us which links in the social network have actually played a role in the diffusion of information. Retweets have already been considered, e.g., to highlight the conversational aspects of online social interaction (Honeycutt and Herring 2008). and because it is not published or accessible yet The reliability of retweeted information has also been investigated. Mendoza, Poblete, and Castillo (2010) found that false information is more likely to be questioned by users than reliable accounts of an event. Their work is distinct from our own in that it does not investigate the dynamics of misinformation propagation.

2.2 Mining Microblog Data

Several studies have demonstrated that information shared on Twitter has some intrinsic value, facilitating, e.g., predictions of box office success (Asur and Huberman 2010) and the results of political elections (Tumasjan et al. 2010). Content has been further analyzed to study consumer reactions to specific brands (Jansen et al. 2009), the use of tags to alter content (Huang, Thornton, and Efthimiadis 2010), its relation to headline news (Kwak et al. 2010), and the factors that influence the probability of a meme to be retweeted (Suh et al. 2010). Romero et al. (2010) have focused on how passive and active users influence the spreading paths.

Recent work has leveraged the collective behavior of Twitter users to gain insight into a number of diverse phenomena. Analysis of tweet content has shown that some correlation exists between the global mood of its users and important worldwide events, including stock market fluctuations (Bollen, Mao, and Pepe 2010; Bollen, Mao, and Zeng 2011). Similar techniques have been applied to infer relationships between media events such as presidential debates and affective responses among social media users (Diakopoulos and Shamma 2010). Sankaranarayanan et al. (2009) developed an automated breaking news detection system based on the linking behavior of Twitter users, while Heer and boyd (2005) describe a system for visualizing and exploring the relationships between users in large-scale social media systems. Driven by practical concerns, others have successfully approximated the epicenter of earthquakes in Japan by treating Twitter users as a geographically-distributed sensor network (Sakaki, Okazaki, and Matsuo 2010).

2.3 Political Astroturf and Truthiness

In the remainder of this paper we describe the analysis of data obtained by a system designed to detect astroturfing campaigns on Twitter (Ratkiewicz et al. 2011). An illustrative example of such campaign has been recently documented by Mustafaraj and Metaxas (2010). They described a concerted, deceitful attempt to cause a specific URL to rise to prominence on Twitter through the use of a network of nine fake user accounts. These accounts produced 929 tweets over the course of 138 minutes, all of which included a link to a website smearing one of the candidates in the 2009 Massachusetts special election. The tweets injecting this meme mentioned users who had previously expressed interest in the election. The initiators sought not just to expose a finite audience to a specific URL, but to trigger an information cascade that would lend a sense of credibility and grassroots enthusiasm to a specific political message. Within hours, a substantial portion of the targeted users retweeted the link, resulting in a rapid spread detected by Google's real-time search engine. This caused the URL in question to be promoted to the top of the Google results page for a query on the candidate's name — a so-called *Twitter bomb.* This case study demonstrates the ease with which a focused

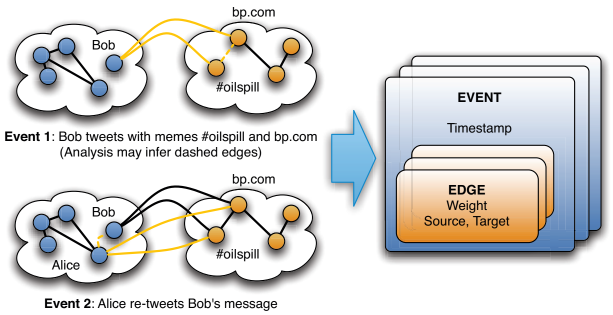

bp.com
Bob
#oilspill
Event 1: Bob tweets with memes #oilspill and bp.com
(Analysis may infer dashed edges)
bp.com
Bob
Alice
#oilspill
Event 2: Alice re-tweets Bob's message
EVENT
Timestamp
EDGE
Weight
Source, Target

Figure 1: Model of streaming social media events.

effort can initiate the viral spread of information on Twitter, and the serious consequences of such abuse.

Mass creation of accounts, impersonation of users, and the posting of deceptive content are behaviors that are likely common to both spam and political astroturfing. However, political astroturf is not exactly the same as spam. While the primary objective of a spammer is often to persuade users to click a link, someone interested in promoting an astroturf message wants to establish a false sense of group consensus about a particular idea. Related to this process is the fact that users are more likely to believe a message that they perceive as coming from several independent sources, or from an acquaintance (Jagatic et al. 2007). Spam detection systems often focus on the content of a potential spam message — for instance, to see if the message contains a certain link or set of tags. In detecting political astroturf, we focus on *how* the message is delivered rather than on its content. Further, many legitimate users may be unwittingly complicit in the propagation of astroturf, having been themselves deceived. Spam detection methods that focus solely on properties of user accounts, such as the number of URLs in tweets from an account or the interval between successive tweets, may therefore be unsuccessful in finding such abuse.

We adopt the term *truthy* to discriminate falsely-propagated information from organic grassroots memes. The term was coined by comedian Stephen Colbert to describe something that a person believes based on emotion rather than facts. We can then define our task as the detection of truthy memes in the Twitter stream. Not every truthy meme will result in a viral cascade like the one documented by Mustafaraj and Metaxas, but we wish to test the hypothesis that the initial stages exhibit identifiable signatures.

3 Analytical Framework

We developed a unified framework, which we call *Klatsch*, that analyzes the behavior of users and diffusion of ideas in a broad variety of data feeds. This 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 this end, we model a generic stream of social networking data as a series of events that represent interactions between *actors* and *memes*, as shown in Fig. 1. Each event involves some number of actors (entities that represent users), some number of memes (entities that represent units of information at the desired level of detail), and interactions among them. For example, a single tweet event might involve 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 are associated with either actors or memes. The timestamps associated with the events allow us to observe the changing structure of this network over time.

3.1 Meme Types

To study the diffusion of information on Twitter it is necessary to identify a specific topic as it propagates through the social substrate. While there exist sophisticated statistical techniques for modeling the topics underlying bodies of text, the small size of each tweet and the contextual drift present in streaming data create significant complications (Wang et al. 2003). Fortunately, several conventions shared by Twitter users allow us to sidestep these issues. We focus on the following features to identify different types of memes:

Hashtags The Twitter community uses tokens prefixed by a hashmark (#) to label the topical content of tweets. Some examples of popular tags are `#gop`, `#obama`, and `#desen`, marking discussion about the Republican party, President Obama, and the Delaware race for U.S. Senate, respectively. These are often called *hashtags*.

Mentions A Twitter user can include another user's screen name in a post, prepended by the `@` symbol. These *mentions* can be used to denote that a particular Twitter user is being discussed.

URLs We extract URLs from tweets by matching strings of valid URL characters that begin with '`http://`.' Honeycutt and Herring (2008) suggest that URLs are associated with the transmission of information on Twitter.

Phrases Finally, we consider the entire text of the tweet itself to be a meme, once all Twitter metadata, punctuation, and URLs have been removed.

Relying on these conventions we are able to focus on the ways in which a large number of memes propagate through the Twitter social network. Note that a tweet may be included in several of these categories. A tweet containing (for instance) two hashtags and a URL would count as a member of each of the three resulting memes.

3.2 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 example diffusion network involving three users is shown in Fig. 2. An edge is drawn from node A to B when either B is observed to retweet a message from A, or A mentions B in a tweet. The weight of an edge is incremented each time we observe an event connecting two users. In this way, either type of edge can be understood to represent a flow of information from A to B.

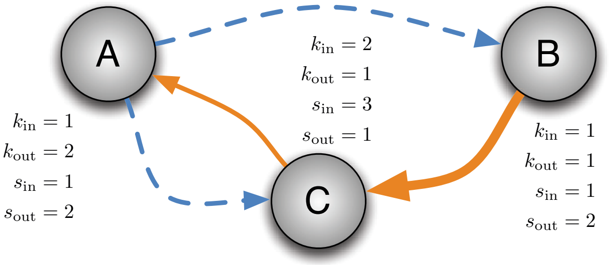


Figure 2: Example of a meme diffusion network involving three users mentioning and retweeting each other. The values of various node statistics are shown next to each node. The strength s refers to weighted degree, k stands for degree.

Observing a retweet at node B provides implicit confirmation that information from A appeared in B's Twitter feed, while a mention of B originating at node A explicitly confirms that A's message appeared in B's Twitter feed. This may or may not be noticed by B, therefore mention edges are less reliable indicators of information flow compared to retweet edges.

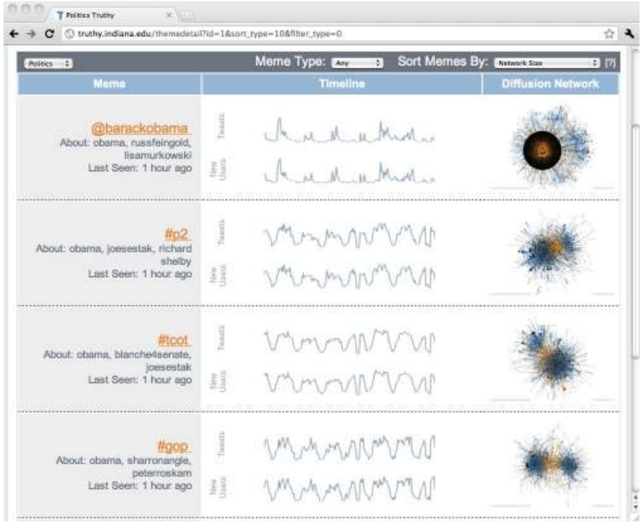
Retweet and reply/mention information parsed from the text can be ambiguous, as in the case when a tweet is marked as being a 'retweet' of multiple people. Rather, we rely on Twitter metadata, which designates users replied to or retweeted by each message. Thus, while the text of a tweet may contain several mentions, we only draw an edge to the user explicitly designated as the mentioned user by the metadata. In so doing, we may miss retweets that do not use the explicit retweet feature and thus are not captured in the metadata. Note that this is separate from our use of mentions as memes (§ 3.1), which we parse from the text of the tweet.

4 System Architecture

We implemented a system based on the data representation described above to automatically monitor the data stream from Twitter, detect relevant memes, collect the tweets that match themes of interest, and produce basic statistical features relative to patterns of diffusion. These features are then passed to our meme classifier and/or visualized. We called this system "Truthy." The different stages that lead to the identification of the truthy memes are described in the following subsections. A screenshot of the meme overview page of our website (`truthy.indiana.edu`) is shown in Fig. 3. Upon clicking on any meme, the user is taken to another page with more detailed statistics about that meme. They are also given an opportunity to label the meme as 'truthy;' the idea is to crowdsource the identification of truthy memes, as an input to the classifier described in § 5.

4.1 Data Collection

To collect meme diffusion data we rely on whitelisted access to the Twitter 'Gardenhose' streaming API (`dev.twitter.com/pages/streaming_api`). The Gardenhose provides detailed data on a sample of the Twitter corpus at a rate that varied between roughly 4 million tweets per day near the beginning of our study, to around 8 million tweets per day at the time of this writing. While the process of sampling edges (tweets between users) from a network to investigate structural properties has been shown to produce suboptimal approximations of true network characteristics (Leskovec and Faloutsos 2006), we find that the analyses described below are able to produce accurate classifications of truthy memes even in light of this shortcoming.



truthy.indiana.edu/themedetail?id=1&sort_type=10&filter_type=0

Meme Type: Any
Sort Memes By: Network Size [?]

Meme	Timeline	Diffusion Network
@barackobama About: obama, russfeingold, lisamurkowski Last Seen: 1 hour ago		
#p2 About: obama, joesestak, richard shelby Last Seen: 1 hour ago		
#tcot About: obama, blanche4senate, joesestak Last Seen: 1 hour ago		
#gop About: obama, sharronangle, peterroskam Last Seen: 1 hour ago		

Figure 3: Screenshot of the Meme Overview page of our website, displaying a number of vital statistics about tracked memes. Users can then select a particular meme for more detailed information.

4.2 Meme Detection

A second component of our system is devoted to scanning the collected tweets in real time. The task of this meme detection component is to determine which of the collected tweets are to be stored in our database for further analysis. Our goal is to collect only tweets *(a)* with content related to U.S. politics, and *(b)* of sufficiently general interest in that context. Political relevance is determined by matching against a manually compiled list of keywords. We consider a meme to be of general interest if the number of tweets with that meme observed in a sliding window of time exceeds a given threshold. We implemented a filtering step for each of these criteria, described elsewhere (Ratkiewicz et al. 2011).

Our system has tracked a total of approximately 305 million tweets collected from September 14 until October 27, 2010. Of these, 1.2 million contain one or more of our political keywords; the meme filtering step further reduced this number to 600,000. Note that this number of tweets does not directly correspond to the number of tracked memes, as each tweet might contribute to several memes.

4.3 Network Analysis

To characterize the structure of each meme's diffusion network we compute several statistics based on the topology of the largest connected component of the retweet/mention

Table 1: Features used in truthy classification.

nodes	Number of nodes
edges	Number of edges
mean_k	Mean degree
mean_s	Mean strength
mean_w	Mean edge weight in largest connected component
max_k(i,o)	Maximum (in,out)-degree
max_k(i,o)_user	User with max. (in,out)-degree
max_s(i,o)	Maximum (in,out)-strength
max_s(i,o)_user	User with max. (in,out)-strength
std_k(i,o)	Std. dev. of (in,out)-degree
std_s(i,o)	Std. dev. of (in,out)-strength
skew_k(i,o)	Skew of (in,out)-degree distribution
skew_s(i,o)	Skew of (in,out)-strength distribution
mean_cc	Mean size of connected components
max_cc	Size of largest connected component
entry_nodes	Number of unique injections
num_truthy	Number of times 'truthy' button was clicked
sentiment scores	Six GPOMS sentiment dimensions

Table 2: Performance of two classifiers with and without resampling training data to equalize class sizes. All results are averaged based on 10-fold cross-validation.

Classifier	Resampling?	Accuracy	AUC
AdaBoost	No	92.6%	0.91
AdaBoost	Yes	96.4%	0.99
SVM	No	88.3%	0.77
SVM	Yes	95.6%	0.95

Table 3: Confusion matrices for a boosted decision stump classifier with and without resampling. The labels on the rows refer to true class assignments; the labels on the columns are those predicted.

	No resampling		With resampling	
	Truthy	Legitimate	Truthy	Legitimate
T	45 (12%)	16 (4%)	165 (45%)	6 (1%)
L	11 (3%)	294 (80%)	7 (2%)	188 (51%)

graph. These include the number of nodes and edges in the graph, the mean degree and strength of nodes in the graph, mean edge weight, mean clustering coefficient across nodes in the largest connected component, and the standard deviation and skew of each network's in-degree, out-degree and strength distributions (see Fig. 2). 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, reasoning that organic memes (such as those relating to news events) will be associated with larger number of originating users.

4.4 Sentiment Analysis

We also utilize a modified version of the Google-based Profile of Mood States (GPOMS) sentiment analysis method (Bollen, Mao, and Pepe 2010) in the analysis of meme-specific sentiment on Twitter. The GPOMS tool assigns to a body of text a six-dimensional vector with bases corresponding to different mood attributes (*Calm, Alert, Sure, Vital, Kind,* and *Happy*). To produce scores for a meme along each of the six dimensions, GPOMS relies on a vocabulary taken from an established psychometric evaluation instrument extended with co-occurring terms from the Google n-gram corpus. We applied the GPOMS methodology to the collection of tweets, obtaining a six-dimensional mood vector for each meme.

5 Automatic Classification

As an application of the analyses performed by the Truthy system, we trained a binary classifier to automatically label legitimate and truthy memes.

We began by producing a hand-labeled corpus of training examples in three classes — 'truthy,' 'legitimate,' and 'remove.' We labeled these by presenting random memes to several human reviewers (the authors of the paper and a few additional volunteers), and asking them to place each meme in one of the three categories. A meme was to be classified as 'truthy' if a significant portion of the users involved in that meme appeared to be spreading it in misleading ways — e.g., if a number of the accounts tweeting about the meme appeared to be robots or sock puppets, the accounts appeared to follow only other propagators of the meme (clique behavior), or the users engaged in repeated reply/retweet exclusively with other users who had tweeted the meme. 'Legitimate' memes were described as memes representing normal use of Twitter — several non-automated users conversing about a topic. The final category, 'remove,' was used for memes in a non-English language or otherwise unrelated to U.S. politics (#youth, for example). These memes were not used in the training or evaluation of classifiers.

Upon gathering 252 annotated memes, we observed an imbalance in our labeled data (231 legitimate and only 21 truthy). Rather than simply resampling from the smaller class, as is common practice in the case of class imbalance, we performed a second round of human annotations on previously-unlabeled memes predicted to be 'truthy' by the classifier trained in the previous round, gaining 103 more annotations (74 legitimate and 40 truthy). We note that the human classifiers knew that the additional memes were possibly more likely to be truthy, but that the classifier was not very good at this point due to the paucity of training data and indeed was often contradicted by the human classification. This bootstrapping procedure allowed us to manually label a larger portion of truthy memes with less bias than resampling. Our final training dataset consisted of 366 training examples — 61 'truthy' memes and 305 legitimate ones. In a few cases where multiple reviewers disagreed on the labeling of a meme, we determined the final label by reaching consensus in a group discussion among all reviewers. The dataset is available online.[1]

We experimented with several classifiers, as implemented

[1] cnets.indiana.edu/groups/nan/truthy

Table 4: Top 10 most discriminative features, according to a χ^2 analysis under 10-fold cross validation. Intervals represent the variation of the χ^2 or rank across the folds.

Feature	χ^2	Rank
mean_w	230 ± 4	1.0 ± 0.0
mean_s	204 ± 6	2.0 ± 0.0
edges	188 ± 4	4.3 ± 1.9
skew_ko	185 ± 4	4.4 ± 1.1
std_si	183 ± 5	5.1 ± 1.3
skew_so	184 ± 4	5.1 ± 0.9
skew_si	180 ± 4	6.7 ± 1.3
max_cc	177 ± 4	8.1 ± 1.0
skew_ki	174 ± 4	9.6 ± 0.9
std_ko	168 ± 5	11.5 ± 0.9

by Hall et al. (2009). Since comparing different learning algorithms is not our goal, we report on the results obtained with just two well-known classifiers: AdaBoost with DecisionStump, and SVM. We provided each classifier with 31 features about each meme, as shown in Table 1. A few of these features bear further explanation. Measures relating to 'degree' and 'strength' refer to the nodes in the diffusion network of the meme in question — that is, the number of people that each user retweeted or mentioned, and the number of times these connections were made, respectively. We defined an 'injection point' as a tweet containing the meme which was not itself a retweet; our intuition was that memes with a larger number of injection points were more likely to be legitimate. No features were normalized.

As the number of instances of truthy memes was still less than instances of legitimate ones, we also experimented with resampling the training data to balance the classes prior to classification. The performance of the classifiers is shown in Table 2, as evaluated by their accuracy and the area under their ROC curves (AUC). The latter is an appropriate evaluation measure in the presence of class imbalance. In all cases these preliminary results are quite encouraging, with accuracy around or above 90%. The best results are obtained by AdaBoost with resampling: better than 96% accuracy and 0.99 AUC. Table 3 further shows the confusion matrices for AdaBoost. In this task, false negatives (truthy memes incorrectly classified as legitimate, in the upper-right quadrant of each matrix) are less desirable than false positives (the lower-left quadrant). In the worst case, the false negative rate is 4%. We did not perform any feature selection or other optimization; the classifiers were provided with all the features computed for each meme (Table 1). Table 4 shows the 10 most discriminative features, as determined by χ^2 analysis. Network features appear to be more discriminative than sentiment scores or the few user annotations that we collected.

6 Examples of Astroturf

The Truthy system allowed us to identify several egregious instances of astroturf 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 2010 U.S. midterm political elections, and subsequently many of the accounts involved were suspended by Twitter. Let us illustrate a few representative examples.

#ampat The #ampat hashtag is used by many conservative users. What makes this meme suspicious is that the bursts of activity are driven by two accounts, @CSteven and @CStevenTucker, which are controlled by the same user, in an apparent effort to give the impression that more people are tweeting about the same topics. This user posts the same tweets using the two accounts and has generated a total of over $41,000$ tweets in this fashion. See Fig. 4(A) for the #ampat diffusion network.

@PeaceKaren_25 This account did not disclose information about the identity of its owner, and generated a very large number of tweets (over 10,000 in four months). Almost all of these tweets supported several Republican candidates. Another account, @HopeMarie_25, had a similar behavior to @PeaceKaren_25 in retweeting the accounts of the same candidates and boosting the same websites. It did not produce any original tweets, and in addition it retweeted all of @PeaceKaren_25's tweets, promoting that account. These accounts had also succeeded at creating a 'twitter bomb:' for a time, Google searches for "gopleader" returned these tweets in the first page of results. A visualization of the interaction between these two accounts can be seen in Fig. 4(B). Both accounts were suspended by Twitter by the time of this writing.

gopleader.gov This meme is the website of the Republican Leader John Boehner. It looks truthy because it is promoted by the two suspicious accounts described above. The diffusion of this URL is shown in Fig. 4(C).

How Chris Coons budget works- uses tax $ 2 attend dinners and fashion shows
This is one of a set of truthy memes smearing Chris Coons, 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 avoid detection by Twitter and increase visibility to different users, duplicate tweets are disguised by adding different hashtags and appending junk query parameters to the URLs. To generate retweeting cascades, the bots also coordinate mentioning a few popular users. When these targets perceive receiving the same news from several people, they are more likely to think it is true and spread it to their followers. Most bot accounts in this network can be traced back to a single person who runs the freedomist.com website. The diffusion network corresponding to this case is illustrated in Fig. 4(D).

These are just a few examples of truthy memes that our system was able to identify. Two other networks of bots were shut down by Twitter after being detected by Truthy.

Fig. 4 also shows the diffusion networks for four legitimate memes. One, #Truthy, was injected as an experiment by the NPR Science Friday radio program. Another, @senjohnmccain, displays two different communities in which the meme was propagated: one by retweets

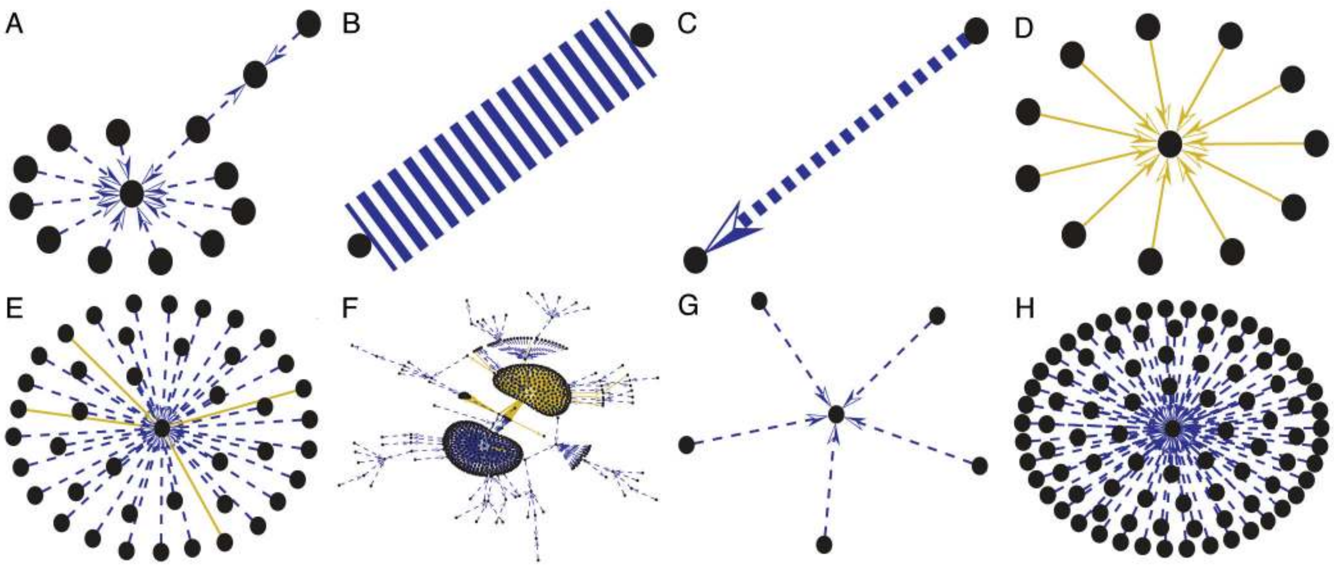


Figure 4: Diffusion networks of sample memes from our dataset. Edges are represented using the same notation as in Fig. 2. Four truthy memes are shown in the top row and four legitimate ones in the bottom row. (A) `#ampat` (B) `@PeaceKaren_25` (C) `gopleader.gov` (D) "How Chris Coons budget works- uses tax $ 2 attend dinners and fashion shows" (E) `#Truthy` (F) `@senjohnmccain` (G) `on.cnn.com/aVMu5y` (H) "Obama said taxes have gone down during his administration. That's ONE way to get rid of income tax — getting rid of income"

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`. A gallery with detailed explanations about various truthy and legitimate memes can be found on our website (`truthy.indiana.edu/gallery`).

7 Discussion

Our simple classification system was able to accurately detect 'truthy' memes based on features extracted from the topology of the diffusion networks. Using this system we have been able to identify a number of 'truthy' memes. Though few of these exhibit the explosive growth characteristic of true viral memes, they are nonetheless clear examples of coordinated attempts to deceive Twitter users. Truthy memes are often spread initially by bots, causing them to exhibit, when compared with organic memes, pathological diffusion graphs. These graphs show a number of peculiar features, including high numbers of unique injection points with few or no connected components, strong star-like topologies characterized by high average degree, and most tellingly large edge weights between dyads.

In addition, we observed several other approaches to deception that were not discoverable using graph-based properties only. One case was that of a bot network using unique query string suffixes on otherwise identical URLs in an effort to make them look distinct. This works because many URL-shortening services ignore query strings when processing redirect requests. In another case we observed a number of automated accounts that use 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.

The accuracy scores we obtain in the classification task are surprisingly high. We hypothesize that this performance is partially explained by the fact that a consistent proportion of the memes were failed attempts of starting a cascade. In these cases the networks reduced to isolated injection points or small components, resulting in network properties amenable to easy classification.

Despite the fact that many of the memes discussed in this paper 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, the meme spreading pattern becomes indistinguishable from an organic one. Therefore, the early identification and termination of accounts associated with astroturf memes is critical.

Future work could explore further crowdsourcing the annotation of truthy memes. In our present system, we were not able to collect sufficient crowdsourcing data (only 304 clicks of the 'truthy' button, and mostly correlated with meme popularity), but these annotations may well prove useful with more data. Several other promising features could be used as input to a classifier, such as the age of the accounts involved in spreading a meme, the reputation of users based on other memes they have contributed, and other features from bot detection methods (Chu et al. 2010).

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 IUB School of Informatics and Computing.

References

Adamic, L., and Glance, N. 2005. The political blogosphere and the 2004 U.S. election: Divided they blog. In *Proc. 3rd Intl. Workshop on Link Discovery (LinkKDD)*, 36–43.

Asur, S., and Huberman, B. A. 2010. Predicting the future with social media. Technical Report arXiv:1003.5699, CoRR.

Barrat, A.; Barthelemy, M.; and Vespignani, A. 2008. *Dynamical Processes on Complex Networks*. Cambridge University Press.

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

Bollen, J.; Mao, H.; and Zeng, X. 2011. Twitter mood predicts the stock market. *J. of Computational Science* In Press.

Castellano, C.; Fortunato, S.; and Loreto, V. 2009. Statistical physics of social dynamics. *Rev. Mod. Phys.* 81(2):591–646.

Chu, Z.; Gianvecchio, S.; Wang, H.; and Jajodia, S. 2010. Who is tweeting on twitter: human, bot, or cyborg? In *Proc. 26th Annual Computer Security Applications Conf. (ASAC)*, 21–30.

Diakopoulos, N. A., and Shamma, D. A. 2010. Characterizing debate performance via aggregated twitter sentiment. In *Proc. 28th Intl. Conf. on Human Factors in Computing Systems (CHI)*, 1195–1198.

Galuba, W.; Aberer, K.; Chakraborty, D.; Despotovic, Z.; and Kellerer, W. 2010. Outtweeting the Twitterers - Predicting Information Cascades in Microblogs. In *3rd Workshop on Online Social Networks (WOSN)*.

Gomez-Rodriguez, M.; Leskovec, J.; and Krause, A. 2010. Inferring networks of diffusion and influence. In *Proc. 16th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (KDD)*, 1019–1028.

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

Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; and Witten, I. H. 2009. The WEKA data mining software: An update. *ACM SIGKDD Explorations* 11(1):10–18.

Heer, J., and boyd, d. 2005. Vizster: Visualizing online social networks. In *Proc. IEEE Symp. on Information Visualization (InfoVis)*.

Honeycutt, C., and Herring, S. C. 2008. Beyond microblogging: Conversation and collaboration via Twitter. In *Proc. 42nd Hawaii Intl. Conf. on System Sciences*.

Huang, J.; Thornton, K. M.; and Efthimiadis, E. N. 2010. Conversational tagging in Twitter. In *Proc. 21st ACM Conf. on Hypertext and Hypermedia (HT)*.

Huberman, B. A.; Romero, D. M.; and Wu, F. 2008. Social networks that matter: Twitter under the microscope. Technical Report arXiv:0812.1045, CoRR.

Jagatic, T.; Johnson, N.; Jakobsson, M.; and Menczer, F. 2007. Social phishing. *Communications of the ACM* 50(10):94–100.

Jansen, B. J.; Zhang, M.; Sobel, K.; and Chowdury, A. 2009. Twitter power: Tweets as electronic word of mouth. *J. of the American Society for Information Science* 60:2169–2188.

Java, A.; Song, X.; Finin, T.; and Tseng, B. 2007. Why we Twitter: understanding microblogging usage and communities. In *Proc. 9th WebKDD and 1st SNA-KDD Workshop on Web mining and social network analysis*, 56–65.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is Twitter, a social network or a news media? In *Proc. 19th Intl. World Wide Web Conf. (WWW)*, 591–600.

Leskovec, J.; Adamic, L. A.; and Huberman, B. A. 2006. Dynamics of viral marketing. *ACM Trans. Web* 1(1):5.

Leskovec, J., and Faloutsos, C. 2006. Sampling from large graphs. In *Proc. 12th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (KDD)*, 631–636.

Leskovec, J.; Backstrom, L.; and Kleinberg, J. 2009. Meme-tracking and the dynamics of the news cycle. In *Proc. 15th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (KDD)*, 497–506.

Mendoza, M.; Poblete, B.; and Castillo, C. 2010. Twitter under crisis: Can we trust what we RT? In *Proc. 1st Workshop on Social Media Analytics (SOMA)*.

Morris, S. 2000. Contagion. *Rev. Economic Studies* 67(1):57–78.

Mustafaraj, E., and Metaxas, P. 2010. From obscurity to prominence in minutes: Political speech and real-time search. In *Proc. Web Science: Extending the Frontiers of Society On-Line (WebSci)*, 317.

Rasmussen, S., and Schoen, D. 2010. *Mad as Hell: How the Tea Party Movement Is Fundamentally Remaking Our Two-Party System*. HarperCollins.

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

Romero, D. M.; Galuba, W.; Asur, S.; and Huberman, B. A. 2010. Influence and passivity in social media. Technical Report arXiv:1008.1253, CoRR.

Sakaki, T.; Okazaki, M.; and Matsuo, Y. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proc. 19th Intl. World Wide Web Conf. (WWW)*, 851–860.

Sankaranarayanan, J.; Samet, H.; Teitler, B.; Lieberman, M.; and Sperling, J. 2009. Twitterstand: news in tweets. In *Proc. 17th ACM SIGSPATIAL Intl. Conf. on Advances in Geographic Information Systems (GIS)*, 42–51.

Suh, B.; Hong, L.; Pirolli, P.; and Chi, E. H. 2010. Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. In *Proc. IEEE Intl. Conf. on Social Computing*.

The Fox Nation. 2010. Stimulus $ for coke monkeys. `politifi.com/news/Stimulus-for-Coke-Monkeys-267998.html`.

Tumasjan, A.; Sprenger, T. O.; Sandner, P. G.; and Welpe, I. M. 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In *Proc.4th Intl. AAAI Conf. on Weblogs and Social Media (ICWSM)*.

Wang, H.; Fan, W.; Yu, P. S.; and Han, J. 2003. Mining concept-drifting data streams using ensemble classifiers. In *Proc.9th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (KDD)*, 226–235.

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

Wiese, D. R., and Gronbeck, B. E. 2005. Campaign 2004: Developments in cyberpolitics. In Denton, R. E., ed., *The 2004 Presidential Campaign: A Communication Perspective*. Rowman & Littlefield. 217–240.