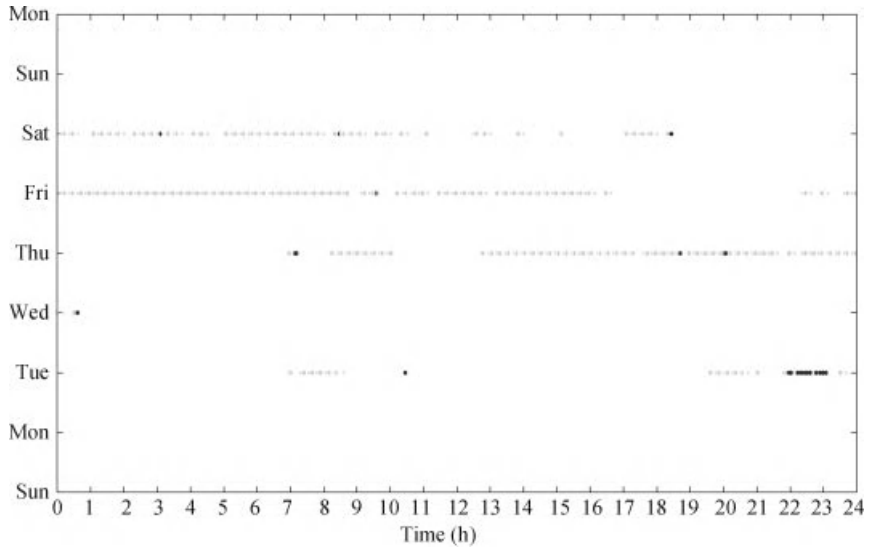

Figure 6.19 A scatter plot for 1 week of a potential bot. The state changes of this user were extremely precise—beyond what a human can achieve. Information posted in this profile like, wall posts and pictures, also point to the conclusion that it is a fake profile. Any click from this account must be considered invalid for purposes of billing, and the IP address of such an account should be blacklisted or at least be put on bot and spiders lists.

6.3 ABUSE OF SOCIAL MEDIA AND POLITICAL MANIPULATION

Bruno Gonçalves, Michael Conover, and Filippo Menczer

Abstract. With the exploding popularity of online social networks and microblogging platforms, social media have become the turf on which battles of opinion are fought. This section discusses a particularly insidious type of abuse of social media, aimed at manipulation of political discourse online. Grassroots campaigns can be simulated using techniques of what has come to be known as astroturf with the goal of promoting a certain view or candidate, slandering an opponent, or simply inciting or suppressing the vote. Such deception threatens the democratic process. We describe various attacks of this kind and a system designed to detect them.

6.3.1 The Rise of Online Grassroots Political Movements

The 2008 presidential election will go down in history as the first to be dominated by grassroots movements organized and coordinated online. The ultimate success of Senator Obama's campaign was due in no small part to its pioneering use of social media. An approach of direct dialog with his grassroots supporters captivated and connected with untapped layers of society and initiated a new era of political participation in American politics. On the other side of the aisle, the aftermath of the election brought about a reaction

culminating in the Tea Party movement [397]. In both cases, it was clear that citizens would no longer be content as passive targets of political messages. They demanded an increased role in defining political discourse.

As individuals gradually turn to the Internet in search of political and economic information, they naturally use existing social networks and platforms to discuss their views and ideals with their peers. Microblogging tools such as Twitter play an important role in this movement by allowing individuals to act as media aggregators and curators who are able to influence their followers and who are, in turn, influenced by the people that they elect to follow. Over time, trust develops between followers and followees making the latter more likely to accept content and information provided by the former.

Perhaps the most striking demonstration of the relevance of this type of discourse, and of how aligned it is with public opinion at large, can be found in a 2010 paper by Tumasjan et al. [470]. By analyzing over 100, 000 tweets containing direct political references to parties or politicians in the ramp-up to the German Federal Election in 2009, they found that the fraction of activity within Twitter corresponding to each party closely matched the vote shares in the final election results. If this result could be generalized, this would imply that Twitter can be used as a real-time public sentiment monitoring tool. Based on this finding, Tumasjan et al. proposed that Twitter be used as a distributed real-time campaign monitoring tool.

If it is true that Twitter truly mirrors public perception, then perhaps it is also true that by manipulating perception within Twitter one is also able to manipulate it in the real world. This inference has not escaped the attention of institutions and groups interested in promoting specific topics or actions. Mustafaraj and Metaxas [364] studied in detail one such case that occurred during the 2010 special Massachusetts senate election. They observed how a network of fake, computer controlled, accounts produced almost 1000 tweets in just over 2 h, containing a specific URL smearing one of the candidates. The goal of the perpetrators was to generate as much traffic as possible to reach a wide audience and thus influence the outcome of the election. To achieve this goal, specific users perceived as influential were they targeted in hopes that they would retweet the URL, thus bestowing upon it an added layer of credibility. Blunt as it was, this attempt was extremely successful, generating such a large number of retweets to briefly place the URL in the first page of Google results for the query *Martha Coakley*—the name of the smeared candidate. Coordinated deception of this sort, where a single agent forges the appearance of widespread support for an idea or position, is known as *astroturfing*, a name that stems from the parallel between fake grassroots movement and the common type of artificial grass used in sports stadiums.

In this section, we look in detail at several tactics being used to promote misinformation in a covert way and a system that aims to automatically detect and track such attempts.

6.3.2 Spam and Astroturfing

As anyone with an email inbox is well aware, spammers have decades of experience in reaching huge audiences. Their techniques range from simple mass email campaigns to sophisticated techniques that automatically customize each message to avoid detection by automated countermeasures. As with other communication media in the past, spammers have descended upon Twitter and adapted their toolbox to this new medium. Many of these techniques and potential countermeasures have been analyzed in detail [113,223,479,499].

Although there seems to be limited amounts of collusion between spammer accounts [223] in the form of spam campaigns designed to make users click a specific URL, there are specific characteristics that can identify spammer accounts. Defining features include the frequency of tweets, the age of the accounts,

and their periphery in the social graph [499]. The combination of content and user behavior attributes makes it possible to train machine learning algorithms to automatically detect spam accounts with a large accuracy [113]. This is likely due to the fact that spam relies on large numbers of accounts controlled by a small number of spammers.

At first glance the goals of spammers and astroturfers might seem similar. They both want to communicate a message to a large audience of users, and both want to effect action (clicks, votes, changes of opinion) in the targeted users. However, there are several fundamental differences between the two types of attacks. Astroturfers, to create the illusion of widespread autonomous support, must retain some degree of credibility and appear independent with respect to commercial or political interests. Likewise, while spammers can use a single account to target many users, astroturfers rely on the fact that users are more receptive to messages they perceive as coming from multiple independent sources. These different techniques necessitate distinct approaches to the detection problem. Spam detection systems often focus on the content of messages—for instance, determining whether the message contains a certain link or set of tags. In detecting astroturf, the focus must be on how the message is delivered rather than its content. The fact that the message is delivered in the guise of legitimate online chatter instead of an organized campaign is more relevant than its veracity. Content may be a legitimate opinion or information resource; the fraud is not a product of the content but rather the distribution mechanism. Further, many of the users involved in propagating a successful astroturf message may in fact be legitimate users who are unwittingly complicit in the deception, having been deceived themselves. Thus, methods for detecting spam that focus on properties of user accounts, such as the number of URLs in tweets originating from an account or the interval between successive tweets, are likely to be unsuccessful in the detection of astroturf. A normal user may come to believe and disseminate a piece of information that had its origins on a campaign of this type. As more and more normal users join the dissemination of this message, any information that could potentially be extracted from analyzing the properties of the accounts spreading it will become increasingly muddled.

6.3.3 Deceptive Tactics

Anyone trying to increase their visibility on Twitter has an obvious strategy: create an account, start tweeting and gradually accumulate followers. However, the egalitarian nature of the platform means that they are just one voice in a crowd of millions. When the goal is to have your voice heard no matter what the cost, several deceptive tactics can be used to quickly gain a large number of followers and obtain an aura of influence or importance within the community [141].

6.3.3.1 Centrally and computer controlled accounts

The old tenet, “nothing attracts a crowd like a crowd” holds true online. Astroturfers take advantage of this fact to catalyze faux grassroots activity by creating the illusion that a large number of people are behind a message or movement. The simplest way to achieve this effect is the creation of multiple centrally controlled accounts, known as *sockpuppets*, which are used to simulate several independent actors promoting a coherent message. These accounts can then be used to broadcast a message seemingly independent of one another, or be manipulated to appear as though they are engaged socially with one another. One advantage of the first approach is that it creates the appearance of independent actors responding to an exogenous influence at the expense of the credibility that comes with a rich social circle. The second approach relies on social

expectations to create the appearance of authenticity at the expense of appearing independent. Common to both of these approaches is a reliance on a large number of centrally coordinated accounts.

To effectively astroturf at scale requires automation, and Chu et al. [153] studied the behavioral differences between real users and *bots* on Twitter. They distinguish between two types of bots: "benign" bots, which often self-identify as automated processes and simply relay information from RSS feeds or other automated sources; and "malicious" bots, which spread spam or malicious content while acting as real users. One of the key distinguishing features between humans and bots is that bots tend to generate a large number of tweets over the course of a short period of time and then hibernate for extended periods, presumably to avoid detection. Humans, on the other hand, tend to follow more regular patterns of online activity [218,219]. While bots can gradually improve their behavior to more closely mimic that of a human, it is not clear if they will ever be able to be completely successful in this task. At the same time, our understanding of human behavior also undergoes a process of continuous refining that further complicates the task of any bot creator.

A large number of emails from HBGary Federal, a security consulting firm, were leaked in early 2011 by a loosely organized group of Internet activists, known as *Anonymous*. Included in these emails are accounts of the development of commercial software to manage large numbers of online personas [459]. The software would keep track of every account associated with each persona as well as its main characteristics. The user of a persona could then rely on a dashboard of personal information necessary to allow him or her to adopt each persona on various social media and easily switch among them. The software package was marketed to companies looking to deceptively manage their online identities as well as government agencies trying to covertly monitor or help shape public opinion [187].

In a related scheme, organizations may recruit volunteers and ask them to donate access to their social media accounts. In doing so, volunteers allow the organization to post on their behalf, making detection much harder while amplifying the potential reach of the message. As these accounts are often unrelated to one another, it's possible to reach separate regions of the social network and potentially expose large portions of the population to a single unified message. Recently this tactic has been employed by groups such as the Christian Coalition of America (CCA) and the Human Rights Campaign (HRC) [251,378]. CCA claims that their "Educate Voters" campaign leading up to the 2008 presidential election involved 1111 Facebook users and 147 Twitter users resulting in 4339 posts and 458 tweets, respectively. HRC's "National Coming Out Day" Facebook campaign involved more than 125, 000 people and generated over 16.3 million posts on National Coming Out Day.

6.3.3.2 Content injection

Twitter is famous for limiting messages to 140 characters, and the need for users to compress messages has led the community to create ways of squeezing as much information in as little space as possible. As a result, users created "hashtag" annotations, terms prefixed by a # (hash) sign, that serve to identify a stream of information associated with a topic or intended audience, such as #dadt for "Don't Ask Don't Tell" or #gop for "Grand Old Party." Their quick adoption by the Twitter community eventually forced Twitter to officially support them and as a result, tweets associated with a given hashtag can be accessed by clicking on the hyperlink embedded in any tweet annotated with the tag or through the official Twitter search tool. For example, clicking on the #gop hashtag in the Twitter website will take the users to https://twitter.com/#!/search?q=%23gop, a constantly updating page containing all recent tweets marked with this hashtag.

Since hashtag streams are not centrally controlled there is nothing stopping any individual user from contributing to a content stream by including the corresponding hashtag in his or her tweets. The potential for abuse is clear: if I'm a political activist trying to reach a large audience of users interested in the Republican

party, appending the #gop hashtag to my tweets is a straightforward way to accomplish this. Using multiple centrally controlled sockpuppet accounts in this way amplifies the effect.

As for hashtags, the use of URL shortening services became widespread in response to the space constraints of a communication medium originally designed to be accessed via SMS. Services such as bit.ly and Twitter's t.co provide URLs containing unique hashes that redirect to the original target. For example, www.example.com would be represented by http://bit.ly/3hDSUb making it difficult to rapidly inspect tweets for suspicious domains for both end users and filtering tools alike.

Twitter has developed spam detection mechanisms that will prevent a single user from simply reposting the same URL over and over again. A simple countermeasure is to slightly modify the URL that is posted by adding meaningless query-string parameters to a shortened URL. Any query parameters in the original URL are already contained within the shortened URL. Extra query parameters are ignored by the shortening service, while providing a simple way of averting detection by Twitter. The predominance of this tactic is demonstrated in Kandylas and Dasdan [292]. They studied the quality of tweeted URLs and found that they are bimodally distributed between high-quality URLs and spam. This highlights the volume of spam that flows through Twitter daily.

The combination of URL obfuscation and content injection makes it possible for spammers and astroturfers alike to target well-defined populations with content that may be, at first glance, difficult to distinguish as being fraudulent.

6.3.3.3 Followback groups

Classic sociological theory holds that public opinion and perception is shaped by leaders that are able to influence a large fraction of the remaining population (see [407] for an in depth description). In social media in general and on Twitter in particular, having a large number of followers is often associated with reputability or importance. Although the applicability of this idea in modern social networks has recently been questioned [141,482], there is still a widespread belief that popular people are necessarily more influential or reliable than others. This perception, fuelled by the fact that well known or influential individuals such as Barack Obama or Oprah Winfrey have several million followers, has led to the development of numerous strategies to increase the numbers of followers.

One of the better known techniques to increase the number of followers is identified by the #FollowFriday hastag. Micah Baldwin in credited with having created this trend in 2009 when he suggested to his followers that they should follow two other users. The idea of promoting other users quickly caught on and led Mykl Roventine to coin the hashtag #FollowFriday. It, and its variation #FF, have endured as a regular trend ever since [97]. This was only one of the first among several hashtags that have been created for this same end.

Another notorious example is that of #TeamFollowBack, a hashtag meant to identify anyone who will reciprocally follow any new follower. Users will add #teamfollowback to a tweet or their profile information as a way of requesting new followers and demonstrating that they will return the favor (follow back). In this way, Bob upon seeing a #teamfollowback tweet by Alice is assured that he will also receive a follow link and have his own number of followers increased if he chooses to follow Alice. Several derivative tags, such as #FollowMe and #InstantFollow, have also been created failed to achieve the same levels of popularity. This trend is the Twitter equivalent of link exchange programs that were developed to influence the PageRank algorithm [241]. PageRank relies on the number of links that a page receives to evaluate its quality, under the assumption that the more pages link to it, the higher its quality is. As a result, webmasters started to exchange links among themselves as a way of boosting their PageRank score.

These mechanisms for developing a large though perhaps fake social circle have clear appeal for spammers and astroturfers. A spammer can draw attention to a particular website by following a large numbers of users and placing a short message along with a URL in the biographical information section of their profile. By default, Twitter notifies users by email whenever they acquire a new follower, and these messages not only include the biographical details of the user but also easily get through automatic spam detection filters as coming from a trusted source. This task is usually carried out by bots, which will stop following the target person after a certain amount of time if the person does not follow them back. This is to avoid having too high a follower to followee ratio, a feature that has been used to detect this type of spam attack in the past [153]. These idiosyncrasies of typical spam bots result in measurable network differences between spammers and nonspammers [499].

6.3.4 The Truthy System for Astroturf Detection

In the previous section, we illustrated some of the tools and techniques commonly employed by malicious users to acquire an aura of respectability and to inject spam and astroturf in online discussions. In this section we describe a system designed to track the spread of political information on Twitter, with the intent of detecting organized astroturf attempts (Figure 6.20). The system is called Truthy, a nod to a term coined

Figure 6.20 Screenshot of a meme detail page on the Truthy website.

by comedian Stephen Colbert to describe something that a person claims to know based on emotion rather than evidence or facts.

Snippets of information that are passed from person to person by word of mouth or common usage are dubbed *memes*, a term originally introduced by Richard Dawkins as the cultural transmission analogous of the gene [169]. In the context of the Internet, a meme is usually meant to refer to a small element of information that has become popular, such as a specific video that has gone viral or a sound byte that has became prominent in online forums.

In light of the characteristics of astroturf discussed above, we need a definition that allows us to discriminate falsely propagated information from organically propagated information originating at the real grassroots. We then define our task as the detection of truthy memes in the Twitter stream. Not every truthy meme will result in a viral information cascade like the one documented above [364], but we wish to test the hypothesis that the initial stages exhibit common signatures that can help us identify this behavior. Of particular interest to this end is the identification of differences in the diffusion patterns of organic and injected memes. For example, a comedians sound byte might become popular and be used by various sources, thus starting many independent cascades. On the other hand, two accounts can try to initiate a cascade by producing many identical tweets in hopes of starting a trend with varying degrees of success. These two examples illustrate some of the obvious differences that we might look for.

Such fingerprints are also likely to concern the way in which Twitter is used as a conversational medium [120,249], but emphasis must be given to the way these conversations lead to the diffusion of information. While usually referred to as "viral" [94,102,138,183,323,324,396], the way in which information or rumors diffuse in a network has important differences with respect to infections diseases. Rumors gradually acquire more credibility and appeal as more and more network neighbors acquire them. After some time, a threshold is crossed and the rumor becomes so widespread that it is considered as "common-knowledge" within a given community and hence, true [199].

We should note that, unlike other systems [139], Truthy does not attempt to make any claims with respect to the factuality of the content in the memes that it tracks. Our goal is simply to be able to distinguish between naturally occurring and injected memes regardless of the validity of the information they might contain. For example, if the sentence "The moon is made of cheese" were to become a meme, it could be considered to be "Truthy" based solely on whether such a surge of popularity was orchestrated by a third party.

Truthy was originally deployed in the run up to the 2010 US midterm elections with the explicit purpose of detecting and tracking political astroturfing attempts in real time. However, the system is general in scope and can easily track multiple themes. An in-depth description of the system and of several analyses performed on the data it generated are available elsewhere [158,399,400]. In the remainder of this section we focus on an overview of the system and some of the main results obtained.

6.3.4.1 Data collection

When a user visits the Truthy website[1], he finds both a list of all the memes detected for each theme, and in-depth information on each meme (Figure 6.20). This presentation allows users to use statistical and visual information to make judgments about the provenance of each meme. The number of nodes and the thickness of edges connecting them in the representation of the diffusion network give a quick indication of how many users are involved and how much traffic each of them is generating. The streaming box of tweets containing that meme that is displayed on the right hand side of the page illustrates the latest contributions and help to

[1] http://truthy.indiana.edu.

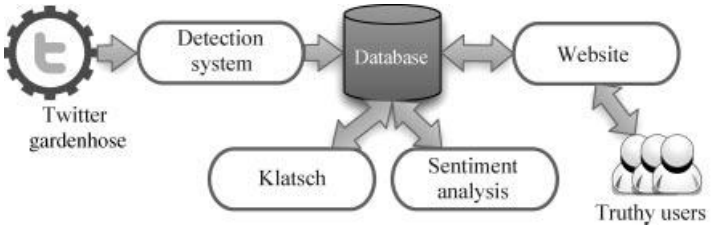

Figure 6.21 Architecture of the Truthy data collection and analysis system.

convey a better understanding of what the meme is about. A more in-depth exploration can also be made by analyzing the information contained in the other tabs, such as the geographical location of users contributing to this meme, statistics on the structure of the network and of more significant users and on how the traffic has changed over time. The website also provides a "Truthy" button that allows any visitor to mark a meme as being truthy, in hopes that these crowdsourced annotations may provide a useful feature in the automatic classification of truthy memes.

The Truthy systems relies on Twitter's streaming API to collect relevant tweets. The overall architecture of the system is illustrated in Figure 6.21. As tweets arrive through the stream, they are processed by a filtering system that looks for political keywords such as the names of all candidates running for US federal office, as well as any common variations of their names and known Twitter account usernames. We also include the top 100 hashtags that co-occurred with the tags #tcot (Top Conservatives On Twitter) and #p2 (Progressives 2.0), the top conservative and liberal tags, respectively, during the last 10 days of August 2010. Any tweet that matches one of more of these keywords it is added to the database [400].

We also consider as a meme any hashtag, URL, username, or phrase that co-occurs with any of our keywords at least one time. These memes are further filtered to extract only those tweets that pertain to memes of significant general interest. To this end, we extract all memes from each incoming tweet, and track activity over the past hour. If any meme exceeds a threshold of traffic in a given hour it is considered "active" and any tweets containing that meme are then stored in the database, until the meme becomes inactive. Note that a tweet can contain more than one meme, and thus the activation of multiple memes can be triggered by the arrival of a single tweet.

Properties of the diffusion network are computed by a system referred to as "Klatsch," the German term for gossip (Figure 6.21), while the sentiment analysis is performed using the GPOMS systems [118]. The end result of these analyses is further added to the database that is also used by the web frontend to produce the page presented to visitors.

The streams provided our system with up to 8 million tweets per day during the course of the study. These were scanned in real time by our system. In total, our analysis considered over 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 interesting memes further reduced this set to 600, 000 tweets actually entered in our database for analysis.

6.3.4.2 Detection of astroturf

We focused on the analysis of three sets of features in our effort to detect truthy memes. The first set of features originates from the network properties of each meme, computed by Klatsch. Considering the nodes corresponding to users, we can observe how a particular meme spreads among users by way of retweets and

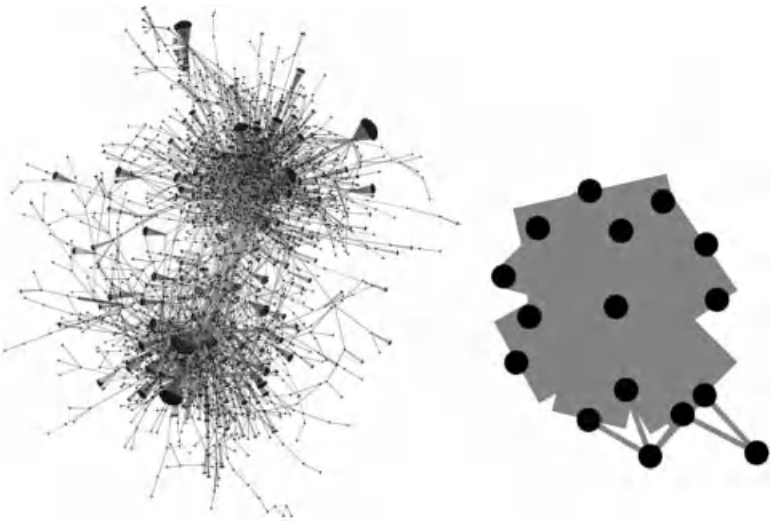

Figure 6.22 Diffusion networks of sample memes from our dataset. Edges represent propagation of memes by way of retweets as well as mentions. Each edge has a weight (width) determined by the number of tweets related to a meme exchanged between two users. Left: a legitimate meme corresponding to a popular, grassroots tag (#gop). We can observe two clusters representing the conservative and liberal communities. Right: a truthy meme boosted by accounts that collude by retweeting each other to promote a particular website. Different tweets include links to different pages from the promoted site, as well as popular but irrelevant hashtags in an effort to catch public attention.

mentions, that is, actions by which a meme can be transmitted from a user to another. For example, when a user Bob retweets a post from Alice containing the meme Charlie, we can see that Charlie has propagated from Alice to Bob. We extract a number of features from the topology of the largest connected component of each meme's diffusion 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, 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, the in-degree and in-strength of the most focused-upon user, and the number of unique injection points of the meme. Figure 6.22 illustrates the differences between the diffusion networks of a legitimate and a truthy meme.

A second set of meme features comes from a sentiment analysis system (Figure 6.21) that extracts a six-dimensional vector of mood attributes from the content of the tweets corresponding to a meme [118]. The crowdsourced truthy annotations provide a final feature.

A binary classifier was trained to automatically label legitimate and truthy memes, based on a training set of hand-labeled memes [399]. We used semisupervised learning (bootstrapping) and resampling to deal with class imbalance between truthy and legitimate memes in the labeled examples. AdaBoost, an ensemble classifier as implemented by Hall et al. [235], yielded excellent results as shown in Table 6.2. Our system identified several truthy memes, resulting in many of the accounts involved being suspended by Twitter. Below we elaborate on a few representative examples of particular relevance.

Example 1: #ampat This hashtag, meaning "American Patriots" is widely used by conservatives on Twitter, however we observed bursts of activity driven by just two accounts (@CSteven and

Table 6.2 Average Performance of the AdaBoost Classifier in the Detection of Truthy Memes, Based on 10-fold Cross-Validation

Accuracy	96.4%
Area under ROC curve	0.99
False negative rate	1%
False positive rate	2%

@CStevenTucker) that belong to the same person. This activity generated traffic around this hashtag and gave the impression that more people were tweeting about it. These two accounts had generated a total of over 41, 000 tweets.

Example 2: @PeaceKaren_25 This account generated over 10, 000 tweets in just 4 months in support of several Republican candidates, boosting for example the popularity of the site gopleader.gov. A separate colluding account @HopeMarie_25 retweeted all the tweets generated by @PeaceKaren_25 supporting the same candidates and boosting the same websites. This is an example of a successful Twitter bomb similar to the one observed by Mustafaraj and Metaxas [364]. For a short period, Google searches for "gopleader" returned these tweets in the first page of results. Both accounts were subsequently suspended by Twitter.

Example 3: How Chris Coons budget works From an analysis of the injection points of this meme we uncovered a network of about 10 bot accounts that injected thousands of tweets with links to posts from the freedomist.com website. These accounts also used several of the tactics described above, such as adding different hashtags to tap into different content streams and appending junk query parameters to the URLs to avoid automatic detection by Twitter. This particular meme was part of a campaign smearing a Democratic candidate for US Senate from Delaware. After the scheme was uncovered by our system, the website administrator admitted to the astroturf behavior in response to a reporter [212].

These are just a few instructive examples of characteristically truthy memes our system was able to identify. 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. A gallery with detailed explanations about various truthy and legitimate memes can be found on the Truthy website.

6.3.5 Discussion

Our social nature makes us vulnerable to attacks that target deep-seated expectations about group consensus and perceptions about source objectivity. In this section, we have explored a variety of techniques malicious attackers can employ to prey on these vulnerabilities, and highlighted several characteristics that make these attacks particularly difficult to defend against.

The ease with which dummy accounts can be created across different social media platforms and the existence of low-cost mechanisms for quickly developing an aura of credibility all but ensure that the scope of this problem is not likely to decrease in coming years. Likewise, as more users turn to social media for

information about products, politics, and public opinion, the potential impact of such deceptive attacks may continue to increase.

Short of the undesirable outcome of unique identifiers for digital content creators, the challenge going forward will be the development of new techniques for the real-time identification of coordinated deception as discussed here. While no one solution can be expected to address these attacks entirely, improved reputation systems, crowdsourced detection mechanisms, and traditional machine learning approaches will all likely play a role in reducing the negative impact of astroturf attacks on the digital ecosystem.

ACKNOWLEDGMENTS

Jacob Ratkiewicz, Mark Meiss, and Snehal Patil have contributed to the development of the Truthy system. We are also grateful to Alessandro Flammini, Johan Bollen, Alessandro Vespignani, Takis Metaxas, Eni Mustafaraj, Ciro Cattuto, Jos Ramasco, and Matt Francisco for critical discussions of the material presented here. We acknowledge support from the Lilly Foundation (Data to Insight Center Research Grant) and the Center for Complex Networks and Systems Research at the IUB School of Informatics and Computing.