Detecting Emerging Topics and Trends Via Predictive Analysis of ‘Meme’ Dynamics

Richard Colbaugh
Sandia National Laboratories
Albuquerque, NM USA
colbaugh@comcast.net

Kristin Glass
New Mexico Institute of Mining and Technology
Socorro, NM USA
kglass@icasa.nmt.edu

***Abstract*—Discovering and characterizing emerging topics and trends through analysis of Web data is of great interest to security analysts and policy makers. This paper considers the problem of monitoring social media to spot emerging *memes* – distinctive phrases which act as “tracers” for discrete cultural units – as a means of rapidly detecting new topics and trends. We have recently developed a method for predicting which memes will propagate widely and which will not, thereby enabling the discovery of *significant* topics. Here we demonstrate the efficacy of this approach through case studies involving political memes and memes associated with an emerging cyber threat.**

***Keywords*—**emerging topics, social media, security informatics.

I. Introduction

The enormous popularity of “social media”, such as blogs, forums, and social networking sites, represents both a significant opportunity and a daunting challenge for security analysts and policy makers [e.g., 1-4]. A vast volume of security-relevant information is generated each day by bloggers and other content producers worldwide, thereby providing an essentially real-time view of opinions, intentions, activities, and trends of individuals and groups across the globe. These data may, for instance, enable early detection of emerging issues, topics, and trends in regions of interest, which could be of considerable value. However, the signatures of emerging topics and trends are buried in the massive, and largely irrelevant, output of millions of online content generators, so that discovering them rapidly enough to be useful is extremely difficult.

We have recently developed an *automated* method for detecting and characterizing emerging topics and trends in social media [5]. Briefly, the proposed approach consists of two steps: 1.) monitoring of social media to spot emerging *memes* – distinctive phrases which act as “tracers” for discrete cultural units – which enables early discovery of new topics and trends [6]; and 2.) predictive analysis of early meme dynamics to identify those which will go on to attract substantial attention, thereby uncovering the memes which correspond to significant, interesting topics. The present paper investigates the performance of the proposed methodology for the task of predicting successful and unsuccessful memes associated with the 2008 U.S. presidential election campaign. The applicability of the approach for security informatics is also illustrated through a preliminary analysis of the emergence in late 2008/early 2009 of a particular cyber threat against Israel.

II. Meme Prediction

In this section we begin by briefly summarizing the approach to meme prediction proposed in [5] and then illustrate the performance of this methodology through case studies involving prediction of successful and unsuccessful political memes and detection of security-relevant memes associated with an emerging cyber threat.

A. Prediction Method

Recently the paper [6] proposed that monitoring social media to spot emerging memes can enable early discovery of new topics and trends, and presented an effective and scalable algorithm for detecting memes. However, a challenge with this method is the fact that the vast majority of online memes attract very little attention, and in most applications we are interested in those memes, and the underlying topics, that reach a nontrivial fraction of the population. Practical emerging topic discovery therefore requires the ability to identify, early in the meme lifecycle, those memes which will go on to attract substantial attention, as this capability would enable the early detection of *significant* emerging topics and trends.

In [5] we present a new predictive methodology which exploits information about network topology and dynamics to accurately forecast which memes will propagate extensively, appearing in hundreds or thousands of blog posts, and which will not. The particular network features used by the prediction algorithm are those identified as likely to be predictive of meme success by our recently developed predictability analysis procedure [7]. Interestingly, the metrics nominated by this theoretical analysis turn out to be fairly subtle measures of the network dynamics associated with early meme diffusion. Meme prediction is accomplished with a machine learning algorithm that, based upon these features of the early network dynamics, is able to accurately distinguish memes which will ultimately diffuse widely from those that will not.

B. Case Study One: Political Memes

To support an empirical evaluation of the proposed meme prediction method, we downloaded from [8] the time series data for approximately 70 000 memes. These data contain, for each meme M, a sequence of pairs $(t_1, \mathrm{URL}_1)_M$, $(t_2, \mathrm{URL}_2)_M$, …, $(t_T, \mathrm{URL}_T)_M$, where t_k is the time of appearance of the kth blog post or news article that contains at least one mention of

978-1-61284-4577-0085-9/11/$26.00 ©2011 IEEE

meme M, URL_k is the URL of the blog or news site on which that post/article was published, and T is the total number of posts that mention meme M. From this set of time series we randomly selected 100 "successful" meme trajectories, defined as those corresponding to memes which attracted at least 1000 posts during their lifetimes, and 100 "unsuccessful" meme trajectories, defined as those whose memes acquired no more than 100 total posts. Note that, in assembling the data in [8], all memes which received fewer than 15 total posts were deleted, and that ~50% of the remaining memes have <50 posts; thus the large majority of memes are unsuccessful by our definition (as well as according to the criteria of most applications [5]).

We collected two additional forms of data associated with these meme trajectories: 1.) a large Web graph which includes the websites that appear in the meme time series, and 2.) samples of the text surrounding the memes in the posts which contain them. The Web graph, denoted G_{web}, was obtained via Web crawling and consists of approximately 550 000 vertices (websites) and 1.4 million edges (hyperlinks). Samples of text surrounding memes in posts were assembled by selecting ten posts at random for each meme and then extracting from each post the paragraph that contains the first mention of the meme.

We now turn to a description of a machine learning-based classifier which is capable of accurately predicting, very early in the lifecycle of a meme of interest, whether that meme will propagate widely. The Avatar ensembles of decision trees algorithm, denoted A-EDT, is the classifier adopted for this study [9]. Our goal is to learn a classifier which takes as input some combination of relevant post content and meme dynamics and accurately predicts whether a given meme will ultimately be successful (acquire ≥1000 posts during its lifetime) or unsuccessful (attract ≤100 total posts). We employ standard ten-fold cross-validation to estimate the accuracy of our classifier. More specifically, our set of 200 memes (100 successful and 100 unsuccessful) is randomly partitioned into ten subsets of equal size, and the A-EDT algorithm is successively "trained" on nine of the subsets and "tested" on the held-out subset in such a way that each subset is used as the test set exactly once.

A key aspect of the analysis is determining which characteristics of memes and their dynamics, if any, possess exploitable predictive power. We consider three classes of features:

- *language-based* measures, such as the sentiment and emotion expressed in the text surrounding memes in posts;
- *simple dynamics-based* metrics, capturing the early volume of posts mentioning the meme of interest and the rate at which this volume is increasing;
- *network dynamics-based* features, such as those identified through the predictability analysis summarized above.

We now describe each of these feature classes. Consider first language-based measures. Each "document" of text surrounding a meme in its (sample) posts is represented by a simple "bag of words" feature vector $x \in \Re^{|V|}$, where the entries of x are the frequencies with which the words in the vocabulary set V appear in the document. The sentiment and emotion of a document may be quantified very simply through the use of appropriate lexicons. Let $s \in \Re^{|V|}$ denote a lexicon vector, in which each entry of s is a numerical "score" quantifying the sentiment or emotion intensity of the corresponding word in the vocabulary V. The sentiment or emotion score of the document x can then be computed as $score(x) = s^T x / s^T 1$, where 1 a vector of ones. Note that this simple formula estimates the sentiment or emotion of a document as a weighted average of the sentiment or emotion scores for the words comprising the document.

To characterize the emotion content of a document we use the Affective Norms for English Words (ANEW) lexicon [10]; this lexicon consists of 1034 words to which human subjects assigned numerical scores with respect to three emotion "axes" – happiness, arousal, and dominance. Previous work had identified this set of words to bear meaningful emotional content [10]. Positive or negative sentiment is quantified by employing the "IBM lexicon", a collection of 2968 words that were assigned {positive, negative} sentiment labels by human subjects [11]. This simple approach generates four language features for each meme: the happiness, arousal, dominance, and positive/negative sentiment of the text surrounding that meme in the (sample) posts containing it.

As a preliminary test, we estimated the mean sentiment and affect of content surrounding the 100 successful and 100 unsuccessful memes in our dataset. On average the text surrounding successful memes is happier, more active, more dominant, and more positive than that surrounding unsuccessful memes, and this difference is statistically significant ($p<0.0001$). Thus it is plausible that these language features may be predictive of meme success.

Consider next two simple dynamics-based features, defined to capture the basic characteristics of the early evolution of a meme's post volume:

- #posts(τ) – the cumulative number of posts mentioning the given meme by time τ;
- post rate(τ) – an estimate of the rate of accumulation of such posts at time τ.

Here we adopt a simple finite difference definition for post rate given by post rate(τ) = (#posts(τ) – #posts(τ/2)) / (τ/2).

The dynamics-based measures of early meme diffusion defined above, while potentially useful, do not characterize the manner in which a meme propagates over the underlying social or information networks. Recall that our predictability analysis suggests that both early dispersion of diffusion activity across network communities and early diffusion activity within the network core ought to be predictive of meme success [5]. This insight motivates the definition of two network dynamics-based features for predicting meme success:

- community dispersion(τ) – the cumulative number of network communities in Web graph G_{web} that, by time τ, contain at least one post which mentions the meme (see [5,12] for background material on graph communities);
- #k-core blogs(τ) – the cumulative number of blogs in the k_{max}-shell of Web graph G_{web} that, by time τ, contain at least one post which mentions the meme (see [5,13] for background material on graph k_{max}-shells).

We now summarize the main results of the study. First, using the four language features with the A-EDT algorithm to predict which memes will be successful yields a prediction accuracy of 66.5%. Since simply guessing 'successful' for all memes gives an accuracy of 50%, it can be seen that these simple language "intrinsics" are not very predictive. In contrast, the features characterizing the early network dynamics of memes possess significant predictive power, and in fact are useful even if only very limited early time series is available to the prediction strategy. More quantitatively, applying the A-EDT algorithm together with the five meme dynamics features produces the following results (based upon standard ten-fold cross-validation):

- τ = 12hr, accuracy = 84.0%, most predictive features (rank ordered): 1.) community dispersion, 2.) #k-core blogs, 3.) #posts.
- τ = 24hr, accuracy = 91.5%, most predictive features: 1.) community dispersion, 2.) post rate, 3.) #posts.
- τ = 48hr, accuracy = 92.8%, most predictive features: 1.) community dispersion, 2.) post rate, 3.) #posts.

C. *Case Study Two: Cyber Attack Memes*

We now briefly summarize a preliminary examination of the utility of meme-based emerging topic detection for security informatics applications. On 27 December 2008 Israel initiated an air strike against the Gaza Strip, triggering outrage in significant portions of the Muslim world. At the time, national security analysts and officials expressed interest in discovering and characterizing social media discussions which called for retaliations against Israel, particularly those likely to "resonate" with individuals and groups.

To enable a preliminary investigation along these lines, we employed the two-step meme detection/meme prediction technique proposed in [5]. We wrote a Perl program implementing the meme detection algorithm presented in [6] and used this program to identify memes associated with 'Israel' and 'attack' in a broad range of languages, including Arabic, English, Farsi, French, German, Indonesian, and Turkish; this effort returned a large number of topically-relevant memes. These memes were then subjected to the prediction algorithm developed in [5] and summarized above. This analysis classified a few of these memes as being likely to attract significant attention. Interestingly, most of the memes predicted to propagate widely involved *cyber* attacks on Israel, for instance exhorting Muslim hackers to attack Israeli government and commercial web sites. Example memes detected and analyzed in this way include 'harrased [sic] by Denmark' and '2485 (web)sites' (the latter meme is a reference to a much repeated claim by one hacker group that it had defaced 2485 Israeli websites).

A focused manual examination of the URLs which mention these "cyber attack" memes produced some interesting findings. For example, this analysis led to the discovery of Arabic and Indonesian websites which contain downloadable hacking tools and detailed instructions on the use of these tools. Additionally, we discovered blog posts calling for particular classes of cyber attacks; it is noted that attacks belonging to these classes were subsequently reported by news sources to have taken place.

Acknowledgements

This work was supported by the U.S. Department of Defense and the Laboratory Directed Research and Development Program at Sandia National Laboratories.

References

[1] US Committee on Homeland Security and Government Affairs, Violent Extremism, the Internet, and the Homegrown Terrorism Threat, 2008.

[2] Bergin, A., S. Osman, C. Ungerer, and N. Yasin, "Countering Internet Radicalization in Southeast Asia", ASPI Special Report, March 2009.

[3] Chen, H., C. Yang, M. Chau, and S. Li (Editors), *Intelligence and Security Informatics*, Lecture Notes in Computer Science, Springer, Berlin, 2009.

[4] *Proc. 2010 IEEE International Conference on Intelligence and Security Informatics*, Vancouver, BC, Canada, May 2010.

[5] Colbaugh, R. and K. Glass, "Emerging topic detection for business intelligence via predictive analysis of 'meme' dynamics", *Proc. 2011 AAAI Spring Symposium Series*, Palo Alto, CA., March 2011.

[6] Leskovec, J., L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle", *Proc. 15th ACM International Conference on Knowledge Discovery and Data Mining*, Paris, France, June 2009.

[7] Colbaugh, R. and K. Glass, "Predictive analysis for social processes I: Multi-scale hybrid system modeling, and II: Predictability and warning analysis", *Proc. 2009 IEEE Multi-Conference on Systems and Control*, Saint Petersburg, Russia, July 2009.

[8] http://memetracker.org, accessed January 2010.

[9] http://www.sandia.gov/avatar/, accessed July 2010.

[10] Bradley, M. and P. Lang, "Affective norms for English words (ANEW): Stimuli, instruction manual, and affective ratings", Technical Report C1, University of Florida, 1999.

[11] Ramakrishnan, G., A. Jadhav, A. Joshi, S. Chakrabarti, and P. Bhattacharyya, "Question answering via Bayesian inference on lexical relations", *Proc. Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, July 2003.Newman, M., "Modularity and community structure in networks", *Proc. National Academy of Sciences USA*, Vol. 103, pp. 8577-8582, 2006.

[12] Newman, M., "Modularity and community structure in networks", *Proc. National Academy of Sciences USA*, Vol. 103, pp. 8577-8582, 2006.

[13] Carmi, S., S. Havlin, S. Kirkpatrick, Y. Shavitt, and E. Shir, "A model of Internet topology using the k-shell decomposition", *Proc. National Academy of Sciences USA,* Vol. 104, pp. 11150-11154, 2007.