SAND2011-4826C

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 present a novel methodology for predicting which memes will propagate widely, appearing in hundreds or thousands of blog posts, and which will not, thereby enabling the discovery of *significant* topics. We begin by identifying measurables which should be predictive of meme success. Interestingly, these metrics are not those normally used for such prediction, but instead are subtle measures of meme dynamics. These metrics form the basis for learning a classifier which predicts, for a given meme, whether or not it will diffuse widely. The efficacy of the proposed methodology is demonstrated through an analysis of successful and unsuccessful memes associated with the 2008 U.S. presidential election campaign. The applicability of the approach to security informatics tasks is illustrated via a case study involving analysis of the emergence in late 2008 of a particular cyber threat against Israel.**

***Keywords*—**emerging topics, social media, predictive analysis, graph analysis, 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. For example, negative information, grievances, and contentious situations are much easier to address if discovered in their early stages, while nascent positive sentiments and activities can often be leveraged and amplified. Of course, 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.

This paper considers the problem of *automated* detection and characterization of emerging topics and trends in social media. Recently [5] proposed that monitoring social media to spot emerging *memes* – distinctive phrases which act as "tracers" for discrete cultural units – 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 security-related applications we are interested in those memes, and the underlying topics, that reach a nontrivial fraction of the population.

This consideration motivates our interest in *predictive* analysis of meme dynamics: we wish to identify those memes which will go on to attract substantial attention, and to do so early in the meme lifecycle. This capability is essential for practical emerging topic discovery, as it would enable early detection of the emergence of *significant* topics and trends. Standard approaches to predictive analysis of social diffusion phenomena like meme propagation assume, either explicitly or implicitly, that diffusion events which propagate widely possess more appealing "intrinsics" than those which don't, and focus attention on identifying these intrinsics [6]. Recent research calls into question this premise, indicating that intrinsic attributes typically don't have much predictive power [e.g., 6-8].

This paper proposes that generating useful predictions about social diffusion requires careful consideration of the way individuals influence one another through their social networks. We present a new predictive methodology which exploits information about network topology and dynamics to accurately forecast which memes will propagate widely, 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 [7,8]. 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 very early network dynamics, is able to accurately distinguish memes which will ultimately diffuse widely from those that will not. The utility of the proposed algorithm is demonstrated through a study of successful and unsuccessful memes associated with the 2008 U.S. presidential election campaign. The applicability of the approach to security informatics tasks is illustrated via a case study involving analysis of the emergence in late 2008 of a particular cyber threat against Israel.

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

II. PRELIMINARIES

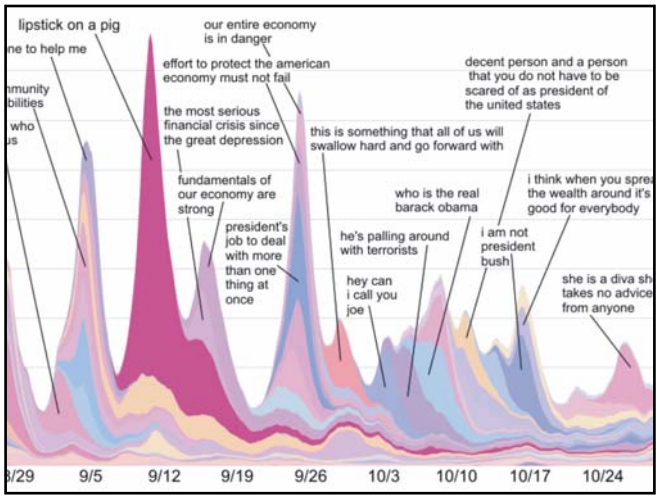
The goal of this paper is to develop a methodology for early and accurate identification of 'memes' which will propagate widely, thereby enabling the discovery of emerging topics and trends which are likely to attract significant attention. This objective leads naturally to two predictive analysis tasks: 1.) identify measurables which are predictive of meme success, and 2.) use these predictive measurables as the basis for classifying memes into two groups – those which will acquire many posts and those that won't – very early in the meme lifecycle. To support an empirical evaluation of our proposed solutions to the above tasks, we downloaded from [9] the time series data for slightly more than 70 000 memes. These data contain, for each meme M, a sequence of pairs $(t_1, URL_1)_M$, $(t_2, URL_2)_M$, …, $(t_T, URL_T)_M$, where t_k is the time of appearance of the kth blog post or news article that contains at least one mention of 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 [9], 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.

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.

Meme dynamics possess several characteristics which are likely to make predictive analysis challenging. For example, the distribution for meme success is strongly right-skewed, with most memes receiving relatively little attention and a few attracting considerable interest [10]; it is known that predicting the evolution of such phenomena can be quite difficult [6-8]. Memes also exhibit highly variable times to acquire their first few posts and to accumulate their final tally of posts. Figure 1 reports the mean and median times required for successful and unsuccessful memes to attract five, ten, and their total number of posts and depicts the evolution of several successful memes.

III. PREDICTIVE ANALYSIS

In this section we begin by summarizing the application of the predictability assessment process [7,8] to a simple model of meme diffusion. This procedure reveals two features of meme network dynamics which should enable early differentiation of successful and unsuccessful memes. We then develop a learning-based algorithm that employs our new network dynamics metrics to accurately predict, very early in a meme's lifecycle, whether that meme will propagate widely or not. The performance of the prediction algorithm is illustrated through empirical studies involving successful and unsuccessful political memes as well as security-relevant memes associated with an emerging cyber threat.



Early Meme Dynamics

Successful Memes (>1000 posts)

#posts	mean (hr)	median (hr)
5	108	18.5
10	171	41.5
total	~4400	~4400

Unsuccessful Memes (<100 posts)

#posts	mean (hr)	median (hr)
5	375	10.1
10	765	30.5
total	~1010	~410

Figure 1. Meme dynamics. In the "stacked" plot at top, thread thickness corresponds to number of posts/articles mentioning the specific meme during that time period (horizontal axis) [5]. The table at bottom reports the mean and median time (in hours) required for successful and unsuccessful memes to acquire five posts, ten posts, and their total number of posts.

A. *Predictability Assessment*

Here we briefly describe the results of applying the predictability assessment procedure presented in [7,8] to the task of identifying measurables which should be predictive of meme success. The discussion begins with short, intuitive reviews of our predictability assessment process and social diffusion modeling framework, and then summarizes the main results obtained through this theoretical analysis.

Predictability. Consider a simple model of information diffusion, in which individuals combine their own beliefs and opinions regarding a new piece of information with their observations of the actions of others to arrive at their decisions about whether to pass along the information. In such situations it can be quite difficult to determine which characteristics of the diffusion process, if any, are predictive of things like the speed or ultimate reach of the diffusion [6-8]. In [7,8] we propose a mathematically rigorous approach to predictability assessment

which, among other things, permits identification of features of social dynamics which should have predictive power; we now summarize this assessment methodology.

The basic idea behind the proposed approach to predictability analysis is simple and natural: we assess predictability by answering questions about the reachability of diffusion events. To obtain a mathematical formulation of this strategy, the behavior about which predictions are to be made is used to define the system *state space subsets of interest* (SSI), while the particular set of candidate measurables under consideration allows identification of the *candidate starting set* (CSS), that is, the set of states and system parameter values which represent initializations that are consistent with, and equivalent under, the presumed observational capability. As a simple example, consider an online market with two products, A and B, and suppose the system state variables consist of the current market share for A, ms(A), and the rate of change of this market share, r(A) (ms(B) and r(B) are not independent state variables because ms(A) + ms(B) = 1 and r(A) + r(B) = 0); let the parameters be the advertising budgets for the products, b(A) and b(B). The producer of A might find it useful to define the SSI to reflect market share dominance by A, that is, the subset of the two-dimensional state space where ms(A) exceeds a specified threshold (and r(A) can take any value). If only market share and advertising budgets can be measured then the CSS is the one-dimensional subset of state-parameter space consisting of the initial magnitudes for ms(A), b(A), and b(B), with r(A) unspecified.

Roughly speaking, the approach to predictability assessment proposed in [7,8] involves determining how probable it is to reach the SSI from a CSS and deciding if these reachability properties are compatible with the prediction goals. If a system's reachability characteristics are incompatible with the given prediction question – if, say, "hit" and "flop" states in the online market example are both fairly likely to be reached from the CSS – then the situation is deemed unpredictable. This setup permits the identification of candidate predictive measurables: these are the measurable states and/or parameters for which predictability is most sensitive [7]. Continuing with the online market example, if trajectories with positive early market share rates r(A) are much more likely to yield market share dominance for A than are trajectories with negative early r(A), then the situation is unpredictable (because the outcome depends strongly upon r(A) and this quantity is not measured). Moreover, this analysis suggests that market share rate is likely to possess predictive power, so it may be possible to increase predictability by adding the capacity to measure this quantity.

Model. In social diffusion, people are affected by what others do. This is easy to visualize in the case of disease transmission, with infections being passed from person to person. Information, such as that in the topics of discussion underlying memes, can also propagate through a population, as individuals become aware of information and persuaded of its relevance through their social and information networks. The dynamics of information diffusion can therefore depend upon the topological features of the pertinent networks. This dependence suggests that, in order to identify features of social diffusion which have predictive power, it is necessary to assess predictability using social and information network models with realistic topologies.

Specifically, the social diffusion models examined in this study possess networks with four topological properties:

- *right-skewed degree distribution* – most vertices have few network neighbors while a few have many neighbors;
- *transitivity* – the network neighbors of a given vertex have an increased probability of being connected to one another;
- *community structure* – the presence of densely connected groupings of vertices which have only relatively few links to other groups;
- *core-periphery structure* – the presence of a small group of "core" vertices which are densely connected to each other and are also close to the other vertices in the network.

Note that these properties are ubiquitous in real world networks and have the potential to impact diffusion dynamics.

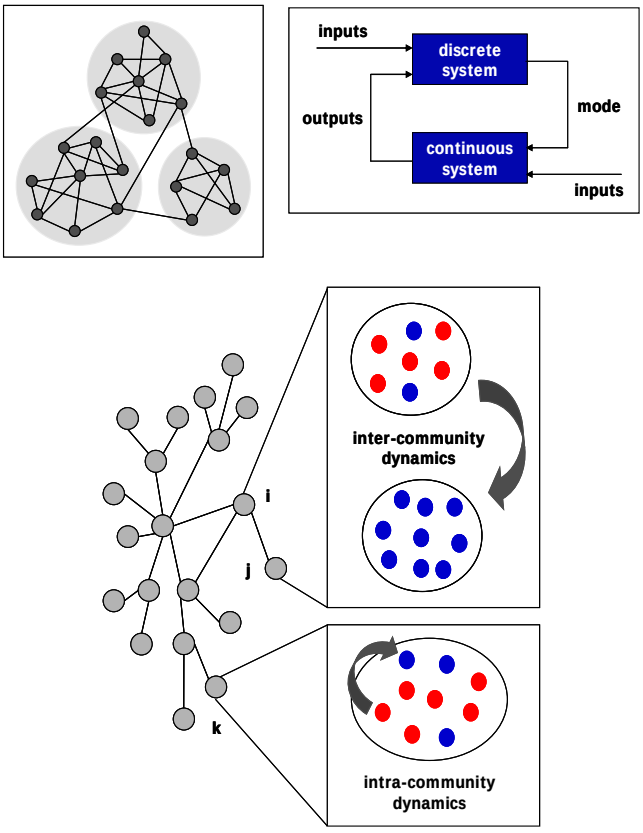


Figure 2. Modeling diffusion on networks with community structure via S-HDS. The cartoon at top left depicts a network with three communities. The cartoon at bottom illustrates diffusion *within* a community k and *between* communities i and j. The schematic at top right shows the basic S-HDS feedback structure; the discrete and continuous systems in this framework model the inter-community and intra-community diffusion dynamics, respectively.

It is shown in [7] that *stochastic hybrid dynamical systems* (S-HDS) provide a useful mathematical formalism with which to represent social diffusion on realistic networks (see Figure 2). An S-HDS is a feedback interconnection of a discrete-state stochastic process, such as a Markov chain, with a family of

continuous-state stochastic dynamical systems [7]. Combining discrete and continuous dynamics within a unified, computationally tractable framework offers an expressive, scalable modeling environment that is amenable to formal mathematical analysis. In particular, S-HDS models can be used to efficiently represent and analyze social diffusion on large-scale networks with the four topological properties listed above [10].

As an intuitive illustration of the way S-HDS enable effective, tractable modeling of complex network phenomena, consider the task of modeling diffusion on a network that possesses community structure. As shown in Figure 2, this diffusion consists of two components: 1.) *intra-community dynamics*, involving frequent interactions between individuals within the same community and the resulting gradual change in the concentrations of "infected" (red) individuals, and 2.) *inter-community dynamics*, in which the "infection" jumps from one community to another, for instance because an infected individual "visits" a new community. S-HDS models offer a natural framework for representing these dynamics, with the S-HDS continuous system modeling the intra-community dynamics (e.g., via stochastic differential equations), the discrete system capturing the inter-community dynamics (e.g., using a Markov chain), and the interplay between these dynamics being encoded in the S-HDS feedback structure.

Results. We have applied the predictability assessment methodology summarized above to a class of empirically-grounded S-HDS models for social diffusion, thereby obtaining a fairly comprehensive theoretical characterization of the predictability of meme propagation on networks with realistic topologies. The main finding of the study, from the perspective of this paper, is a demonstration that the predictability of meme diffusion depends crucially upon social and information network topology, and in particular on a network's community and core-periphery structures. We now summarize the main conclusions of this study; a more complete discussion of the results of this investigation is given in [10].

Community structure is widely recognized to be important in real-world networks, and there exists a range of qualitative and quantitative definitions for this concept. Here we adopt the *modularity-based* definition proposed in [11], whereby a good partitioning of a network's vertices into communities is one for which the number of edges between putative communities is smaller than would be expected in a random partitioning. To be concrete, a modularity-based partitioning of a network into two communities maximizes the modularity $Q = s^T B s / 4m$, where m is the total number of edges in the network, the partition is specified with the elements of vector s by setting $s_i = 1$ if vertex i belongs to community 1 and $s_i = -1$ if it belongs to community 2, and the matrix B has elements $B_{ij} = A_{ij} - k_i k_j / 2m$, with A_{ij} and k_i denoting the network adjacency matrix and degree of vertex i, respectively. Partitions of the network into more than two communities can be constructed recursively [11]. With this definition in hand, we are in a position to present the first candidate predictive feature nominated by our predictability assessment [10]: early dispersion of the diffusion process across network communities should be a reliable predictor that the ultimate reach of diffusion will be significant (see Figure 3).

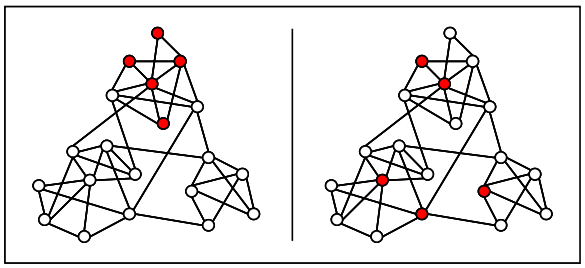


Figure 3. Early dispersion across communities is predictive. The cartoon illustrates the predictive feature associated with community structure: social diffusion initiated with five "seed" individuals is much more likely to propagate widely if these seeds are dispersed across multiple communities (right) rather than concentrated within a single community (left). Note that in [10] this result is established for networks of realistic scale and not simply for "toy" networks.

Analogously to the situation with network communities, there exist several ways to describe core-periphery structure in networks. Here we adopt the characterization of network core-periphery which results from *k-shell decomposition*, a well-established technique in graph theory which is summarized, for instance, in [12]. To partition a network into its k-shells, one first removes all vertices with degree one, repeating this step if necessary until all remaining vertices have degree two or higher; the removed vertices constitute the 1-shell. Continuing in the same way, all vertices with degree two (or less) are recursively removed, creating the 2-shell. This process is repeated until all vertices have been assigned to a k-shell, and the shell with the highest index, the k_{max}-shell, is deemed to be the core of the network. Given this definition, we are in a position to report the second candidate predictive feature nominated by the theoretical predictability assessment [10]: early diffusion activity within the network k_{max}-shell should be a reliable predictor that the ultimate reach of the diffusion will be significant (see Figure 4).

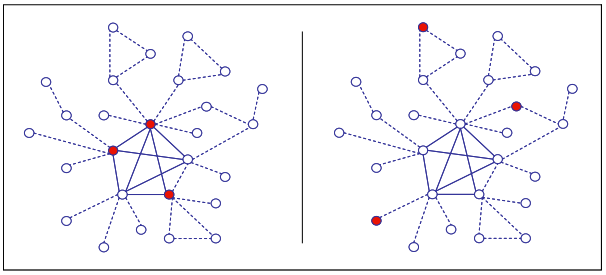


Figure 4. Early diffusion within the core is predictive. The cartoon illustrates the predictive feature associated with k-shell structure: social diffusion initiated with three "seed" individuals is much more likely to propagate widely if these seeds reside within the network's core (left) rather than at its periphery (right). Note that in [10] this result is established for networks of realistic scale and not simply for "toy" networks.

B. Prediction

We now turn to the task of developing 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 [13]. 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 the 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 [14]; 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 [14]. 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 [15]. 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. 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;
- #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.

Note that these quantities can be efficiently computed using fast algorithms for partitioning a graph into its communities and for identifying a graph's k_{max}-shell [10].

We now summarize the main results of the prediction study (see [10] for a more complete description of the results). 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 for use in prediction. More quantitatively, applying the A-EDT algorithm together with the five meme dynamics features produces the following results (ten-fold cross-validation):

- τ = 12hr, accuracy = 84.0%, most predictive features: 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. Security Informatics Example

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, interest was expressed to discover and characterize social media discussions which called for retaliations against Israel, particularly those involving "influential" individuals and groups [10].

To enable a preliminary investigation along these lines, we wrote a Perl program implementing the meme detection algorithm presented in [5] 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 data identification and collection effort returned a large number of memes, a few of which were classified by our prediction algorithm as being likely to attract significant attention. Interestingly, most of the memes predicted to be "successful" 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 (and other) 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.

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] 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.

[6] Salganik, M., P. Dodds, and D. Watts, "Experimental study of inequality and unpredictability in an artificial cultural market", *Science*, Vol. 311, pp. 854-856, 2006.

[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] Colbaugh, R., K. Glass, and P. Ormerod, "Predictability of 'unpredictable' cultural markets", *Proc. 105th Annual Meeting of the American Sociological Association*, Atlanta, GA, August 2010.

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

[10] Colbaugh, R. and K. Glass, "Prediction of social dynamics via Web analytics", ICASA Technical Report, New Mexico Institute of Mining and Technology, in preparation.

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

[12] 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.

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

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

[15] 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.