Meme diffusion through mass social media

1 Motivation and Objectives

Online social media are rapidly complementing and even replacing person-to-person social contacts. The broadcast model of traditional mass media as channels for the diffusion of information is equally losing traction. As a result, online social networks and media have become a crucial turf on which public relations, marketing, and political battles are fought. Viral messages and political smear campaigns are increasingly designed specifically for social media.

Here we propose to build a computational framework that will enable the study of meme diffusion in large-scale social media by collecting, analyzing, classifying, visualizing, and modeling massive streams of public micro-blogging data. While a number of websites have recently emerged that track popular and trending memes, such end-user services are not designed for scientific purposes. We aim to fill this gap by providing for the first time a large-scale infrastructure to delve deeply into a broad set of questions about how and why information spreads online.

To enable this empirical analysis of meme diffusion, we will develop a unified framework that makes it possible to apply the same analytical methods to a broad variety of data feeds, e.g., Twitter, Google Buzz, Yahoo! Meme, and Facebook. In particular, the framework will model streams of social networking data as a series of events that represent interactions between actors and memes. As such it will facilitate comparison of the high-level statistical features across different Web 2.0 communities, the development of general models for the behavior of users, and models for the diffusion of ideas in a social network.

The proposed project will thus span both technical and broad scientific challenges that lie at the intersection of computing and the social sciences. On the technical side, the framework will provide an unprecedented level of data interoperability for the real-time analysis of massive social media data streams on the order of millions of posts per day. More broadly, our framework will have tremendous impact on the emerging field of computational social science, greatly facilitating the study of the growing number of social phenomena that are mirrored online. Below are some key goals that our project will achieve.

- Our framework will analyze information diffusion in the Twitter network and assess the reliability of information by detecting spam, robot-generated content, and astroturfing campaigns as well as inferring the nature of content (e.g., news, promotion, and conversation), identifying controversial content, and predicting popularity trends.
- We will extend existing epidemiological and diffusion models in order to describe and reproduce the empirical patterns captured by our framework.
- We will focus our efforts on developing scientific models of how the observed semantic and emotional features of a meme influence its ability to propagate and become more or less prevalent among particular communities.

The PIs combine expertise in computer science, cognitive science, and complex networks analysis and modeling, and are thus uniquely positioned to achieve the interdisciplinary synergy crucial to address both the scientific and broader impacts of the proposed work. Menczer's team has performed ground-breaking research on meme and burst detection in online media, including the engineering of advanced systems to model and track the longitudinal features of such bursts. Vespignani's team is world-renowned for its work on epidemiological models for the spread of disease through complex networks. Flammini's team brings world-class expertise in complex network analysis, in particular as applied to online text and social media, to our project. His work in computational models of urbanization, information networks and online traffic will be of central

importance to the project's ability to address issues of online meme propagation in socio-technical networks. Bollen's team has developed unique expertise in the domain of sentiment and mood analysis from online text. His work on the cognitive and emotional features of Twitter and email feeds, their relation to the underlying social network substrate and the occurrence of mood contagion will be key to the proposed modeling work on meme propagation. The PIs are members of the Center for Complex Networks and Systems Research (`cnets.indiana.edu`) at IU's School of Informatics and Computing (`soic.indiana.edu`), which provides access to world-class computational resources, outstanding graduate and undergraduate students, and an amazing array of opportunities for interdisciplinary collaboration with colleagues in Cognitive Science, Information Science, Physics, Statistics, and the Social Sciences — both within IU and with collaborators at Wellesley College in Boston, Yahoo! Labs, and the Institute for Scientific Interchange in Torino, Italy. All of these assets will greatly contribute to the success of the project.

2 Background

The process of information diffusion bears a qualitative resemblance to the diffusion of epidemics [7] an idea dating back to Goffman and Newill [27] that was immediately formalized by Dailey [20]. A piece of information can pass from one individual to another through social contact and 'infected' individuals can, in turn, propagate the information to others, possibly generating a full-scale contagion. Once recovered, individuals cannot be infected again. In this scenario, after an initial phase in which the infected population grows exponentially, epidemics reach a maximum where the rate of infection is limited by the decreasing number of susceptible individuals and is balanced by the rate of recovery. The contagion finally dies off when the recovery rate prevails.

Since Goffman and Newill's seminal work [27], scholars have expanded the SIR model [5] to include: phenomena peculiar to information diffusion such as paradigms of thresholds [31] and cascades [28]; factors that influence the speed of spreading such as the age of the information in question [62]; the inhomogeneity in connectivity patterns as studied by our group [70]; and other factors such as probability of transmission along specific links, clustering, social influence, learning, user-created content [8] and norm diffusion [16]. Despite the indisputable progress of the last few years, the field of information diffusion still lacks the hierarchy of commonly-accepted, empirically-validated, and increasingly-detailed models that allow for reliable predictions on the size and duration of specific diffusion processes. On the empirical side, difficulties related to privacy concerns in collecting data and the massive size of relevant data sets have hindered faster progress. The empirical investigation of information diffusion has started in earnest with the availability of data relative to online social networks. The possibility to track large amount of digital data has inspired important work about diffusion in email networks [41], the blogosphere [2, 1], online social networks [29], online recommendation systems [38], and online games [88].

Social networking and micro-blogging services reach hundreds of million of users and have become fertile ground for a variety of research efforts, since they offer an opportunity to study patterns of social interaction among far larger populations than ever before. In particular, Twitter has recently generated much attention in the community of information diffusion due to its peculiar features, enormous popularity, and open policy on data sharing. Indeed, it has even been shown that information shared on Twitter has some intrinsic value, facilitating, e.g., predictions of box office successes [6], and the results of political elections [87]. Twitter works as a micro-blogging platform that allows users to post short messages (tweets) from computers and mobile devices. These messages are automatically forwarded to other users who have declared themselves followers of the sender. Although messages are publicly broadcast in principle, they are generally not noticed by non-followers. One can also address a tweet to specific other users by mentioning them or replying to their tweets. A user will then see the tweet even if it is not from a followed user. Finally users can

retweet messages by forwarding them to their own followers. Many users include special keywords (the hashtags mentioned above) to describe the topics of their tweets. As a communication medium, Twitter presents an original mix of public and private like features.

Analysis of tweet content has shown that some correlation exists between the global mood of its users and important worldwide events [13]. Content has been further analyzed to study consumer reactions to specific brands [34], the use of tags to filter content [33], its relation to headline news [37], and on the factors that influence the probability of a meme to be retweeted [85]. Other authors have focused on how passive and active users influence the spreading paths [80].

Additional study by our group and others has addressed time series data representing social attention [77, 39, 19, 86, 90, 21, 9, 36, 35]. This work has revealed a large variety of patterns, including spike-dominated activity of relatively short duration, and more uniform and long-lasting signals. The former are of particular relevance to this proposal as they more closely resemble the spread of epidemics.

In the domain of sentiment analysis, existing approaches have used text analysis methods [22, 23, 15, 95, 65] in which either topic detection models or pre-defined lexicons are used to assess text sentiment. These methods are however difficult to apply to online memes given their frequent lack of text content and deliberate engineering for brevity, e.g., hashtags and shortened links.

3 Proposed Project

Social media analysis presents some major challenges in the area of data management, particularly when it comes to interoperability, curation, and consistency of process. Although social networking sites share a set of broadly-similar features, their structure is different in each case; Twitter delivers information packaged into 140-character chunks, YouTube attaches streams of comments to a vast network of interlinked videos, and LinkedIn focuses on professional relationships. The consequence of this diversity among site designs, data models, and APIs has been a patchwork of home-grown analytical tools written by researchers to address specific sites. This has resulted in many missed opportunities: the same analysis that has been applied to Twitter data might be fascinating when applied to Facebook updates, but a set of ad-hoc scripts that assumes one model may be difficult to adapt to other sites.

We thus propose the development of a unified framework that we call *Klatsch* for the analysis of social media data that brings together the common features of all social media sites and makes it possible to apply the same analytical methods to a broad variety of data feeds. The Klatsch framework will facilitate comparison of the high-level statistical features of different Web 2.0 communities and the development of general models for the behavior of users and diffusion of ideas in a social network.

The Klatsch framework will provide an unprecedented level of 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. Such an effort necessitates a common data model that captures the shared traits of social networking sites while abstracting over their differences. We achieve this by modeling a stream of social networking data as a series of events that represent interactions between actors and memes, as shown in Figure 1.

In the Klatsch model, social networking sites are sources of a timestamped series of events. Each event involves some number of actors (entities that represent individual users), some number of memes (entities that represent units of information at the desired level of detail), and interactions among those actors and memes. For example, a single Twitter post might constitute an event involving three or more actors: the poster, the user she is retweeting, and the people she is addressing; and a set of memes consisting of 'hashtags' and URLs referenced in the tweet. Each

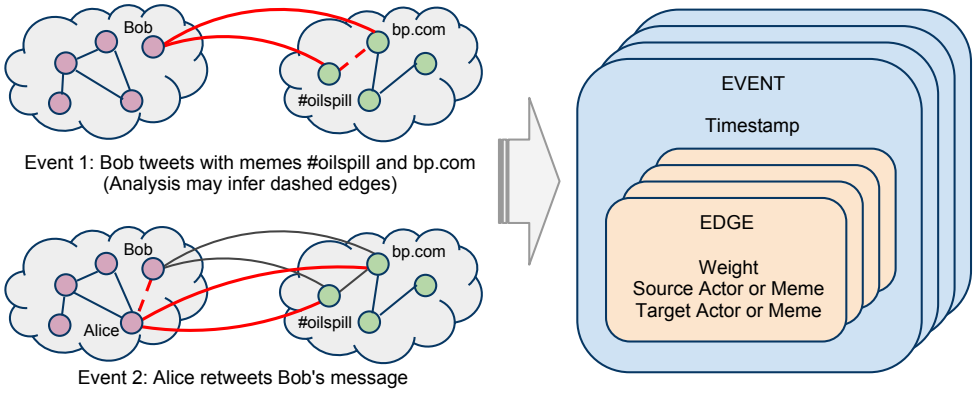

Figure 1: The Klatsch model of streaming social media events.

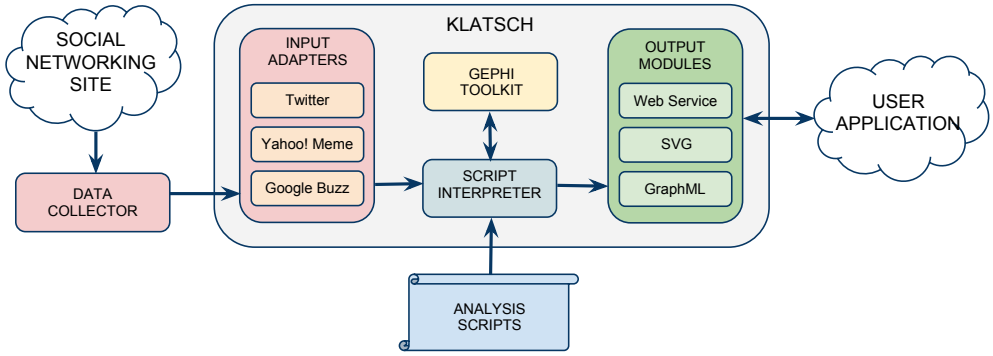

Figure 2: The Klatsch framework architecture.

event can be thought of as contributing weight to a network structure in which nodes are associated with either actors or memes. This is not a strictly bipartite network: actors can be linked through replying or following, and memes by concurrent discussion or semantic similarity. The timestamps associated with the events allow us to observe the changing structure of this network over time.

The Klatsch framework itself will consist of several components, as depicted in Figure 2. Some of these components are already under development. The key components of the system are: a set of input adapters for importing external social network data into the Klatsch data model; support for a variety of standard graph layout and visualization algorithms; a flexible scripting language for coding site-agnostic analysis modules; and a set of export modules, including an embedded light-weight Web server, for visualizing analysis, saving statistical results, supporting interactive Web tools, and producing publication-ready graphs and tables.

The process for developing a research project that deals with a new social networking site will be reduced largely to recasting the data in terms of the Klatsch model. Analysis scripts based on the model will be used without modification, or adapted with minimal effort. Furthermore, visualizations and Web sites built to support investigation of other social media data streams will be reusable with only domain-appropriate changes to formatting and presentation. Users of the Klatsch language will be able to write analytical modules without worrying about database interaction, network programming, or XML parsing.

The Klatsch scripting language embedded in Klatsch will be a high-level domain-specific language with advanced features such as garbage collection, dynamic typing, first-order functions, streams, and map/filter/reduce primitives. Its design will address several computational challenges

particular to the analysis of large graph structures and data sets. The inclusion of streams as a first-order data type will support the lazy evaluation of algorithms that operate on the nodes and edges of large graphs. Streams will provide a clean programmatic interface to potentially infinite sources of data such as those generated by social media. The inclusion of procedures as a first-order data type will additionally make it possible to apply filters, reductions, and mapping procedures to these streams, facilitating the reuse of existing analysis code on a variety of derived graphs without modification. Finally, the internal design will make it easy to add new primitive functions written in Java for cases in which the speed of the interpreter becomes an issue.

The Klatsch framework will help social networking researchers to move away from a series of compelling anecdotes and build a true science of online social interaction. To evaluate this benefit of our framework, it will be necessary to analyze multiple streaming data sources from different social media sites. We will leverage access we have already gained to Twitter and Google Buzz through their streaming APIs. Additionally, we will collaborate with Yahoo! Research Barcelona to gain access to the entire history of posts in the Yahoo! Meme system, as well as to the entire history of its underlying social network.

Below we outline three main research directions that will be enabled by the Klatsch framework and pursued as part of the proposed project. We focus on Twitter as an initial social platform to study, owing to its established popularity, but the same questions can be pursued in the context of other online social media.

3.1 Empirical Analysis of Meme Diffusion

We will apply our proposed framework to the study of information diffusion in the Twitter network, introducing a number of network-centric features to describe the diffusion patterns. This analysis will inform the modeling of information epidemics (Section 3.2) and inspire a classification of the patterns observed online. The empirical analysis discussed in this section is centered around three sub-themes (discussed below): network analysis and evolution, meme classification, and time series analysis.

Network analysis and evolution: Two dynamic networks result from the Twitter platform: the social network of followers and following relations, and the usage network of messages. We are particularly interested in the latter, as it allows us to study the diffusion of information as an epidemic process based on contagion. We will infer the probability that a user i will retweet a message from user j (i.e. that j infects i) within a time interval t by counting the portion of i's messages that are retweeted by j before time t has elapsed. Our assumption is that this probability is a fast-decreasing function of time, i.e. i is more likely to pass j's message early. For example if $P(t) = \alpha e^{-t/\tau}$, the probability that the message will be retweeted is $\int_0^\infty P(t)dt = \alpha\tau < 1$ to account for the possibility that the message is not retweeted at all. This information, when used in a diffusion model, will allow us to test the relevance of local contagion inhomogeneity in the global propagation patterns. Furthermore, our preliminary analysis suggests that the size of an information epidemics (number of infected users) is often smaller than predicted by standard models. A second reason to measure the temporal decay in the local contagion probabilities is to test whether the infectiousness of a piece of information decays with time, thus accounting for these empirical observations.

Our framework will allow us to study for the first time, in a quantitative and data-driven way, the critical phenomenon of competition for attention first hypothesized by Simon [82]. The hypothesis we will test is that the probability or retweeting a message decreases with the number of competing messages a user receives in a given time interval, modulated by their respective ages.

On longer time scales we will be able to monitor whether the above probabilities change over

time. We will test the hypothesis that con- nections are reinforced by usage, as predicted by Hebbian models. If our hypothesis is confirmed we could see the emergence of communities as a consequence of the Hebbian dynamics and ob- serve the diffusion of memes in relatively closed groups of users. Such dynamics, if confirmed, will be folded into our models.

We will extract features, that we call memes, from the content of messages. For instance, in Twitter, a meme can be a hashtag, user mention, URL, or phrase. We will construct diffusion networks for all observed memes, and study their diffusion patterns and properties. One of the simplest measures we will perform is the distribution of meme duration and size: how long does a meme persist in the twittersphere, and how many people does it affect? Such distributions will provide preliminary information about the diffusion process. For example, a power-law decay would suggest an underlying random branching process, often used to describe epidemic spreading.

Figure 3: Diffusion network for the hashtag `#GOP` in Twitter.

Meme classification: Our analysis of meme diffusion patterns will also lead to concrete applications. These will include data mining based on both unsupervised and supervised learning and focus on the construction of classifiers to detect, categorize, or predict various aspects of online discourse. One aspect of this will be determining the reliability of information. In Section 6 we discuss a practical application of our classification approach to political discourse, with potential for broad societal impact. A second application concerns the nature of content: news, (self) promotion, babble, conversation, and so on. Thirdly, we will attempt to identify controversial content. As an illustrative example, we see in Figure 3 that controversial political content is often reflected in two communities with opposing views. Finally, we will apply our classification framework to the early detection of long-term trends in an attempt to, e.g., predict who will be the most influential users or most popular memes.

As input to our classifiers, we will explore a number of network and time series features along with more traditional aggregate ones. The diffusion networks provide us with natural candidates, relative to both nodes and links. Features for actor nodes may include the number of tweets, retweets, memes, mentions, and so on. Similarly we will consider meme node features such as number of actors, tweets, retweets, and so on. Link features include reliability (e.g., replies vs. retweets) and weights (e.g., number of messages between two users or mentioning two memes). It is important to underscore that the diffusion of a particular meme in our framework will be represented as a network of users, and conversely the activity of a user will be represented as a network of memes (cf. Figure 1). If, for instance, we consider a meme diffusion network, then we will have distributional properties across an entire population of actor nodes and edges. We will therefore consider means, modes, standard deviations, and various skew measures to capture broad (e.g., scale free) distributions commonly observed in social activity.

We will explore various classification algorithms in the literature, focusing on the identification of predictive features. For instance, ensemble approaches have proven to perform well in the presence

of many irrelevant features. To provide our classifiers with labeled data for training and evaluation, we will turn to crowdsourcing techniques, which have been successful in citizen science contexts. We will engage users by exposing information about memes and soliciting their participation as a public service. An example is discussed in Section 6.

Time series analysis: A further goal of this proposal is to develop a classification scheme for the meme time series we will be collecting. We will initially focus on bursty behavior because it is characterized by patterns that most closely resemble those of epidemic spreading and are therefore of interest for the work described in Section 3.2. The lack of a shared consensus on a set of representative time patterns that need to be reproduced by models has hindered progress on the modeling of information diffusion. We will therefore develop a time series classification scheme that could lead to such a representative set. The initial strategy we will employ will be to study the Fourier transform of the time series rather than the series themselves. The Fourier transform can readily address some common problems that trouble the task of time series classification: (*i*) it can naturally recognize intrinsic periodicities in the data; (*ii*) it can easily recognize time translation and differences in the overall intensity of the signal, that result in a common (complex) factor in the Fourier series; and (*iii*) it offers an efficient method to filter out noisy signals. We plan to apply standard clustering techniques to the high-frequency filtered Fourier series. As illustration, Figure 4 shows three artificial time patterns characterized by similar looking shapes. These are hardly machine matchable because they differ in the overall intensity of the signal, the large amount of noise superimposed, and the different peak times. The figure shows how well the properly-normalized spectra of the three curves coincide for the most relevant components. By controlling the granularity of the clustering algorithm we hope to find a limited number of representative classes that can drive our subsequent modeling effort. An alternative path toward a significant representation of time series to be clustered, could be through hidden Markov models. Drawing inspiration from Kleinberg's model of bursts of activity [35] we could assume the rate of change of a series in a time step to be drawn from independent Gaussians whose parameters are determined by the current state of the hidden model. Both the parameters of the Gaussians and the transition probabilities of the Markov model could be learned by data with maximum likelihood estimators. Progress on the two challenging problems of burst identification and time series classification could be highly relevant for a large and diverse community of scholars.

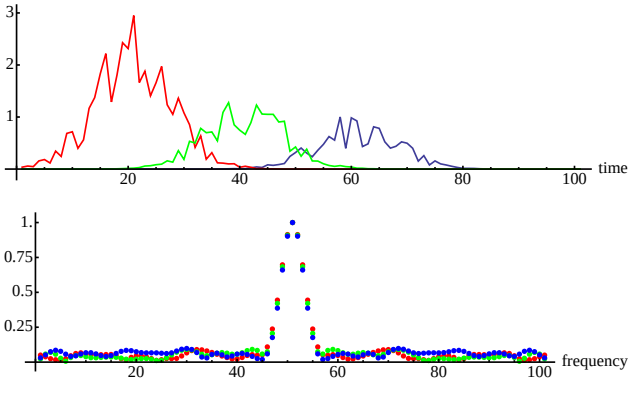

Figure 4: Apparently different time signals (top) show remarkably similar Fourier components (bottom).

3.2 Epidemic Modeling

Although noticeable progress has been made in the last decade in the study of information diffusion, most of the advances are rather speculative and borrow from the vast literature on the spread of infectious diseases. Since the works of Rapoport [73], Goffman and Newill [27] and Granovetter [31], the field has struggled with the lack of large-scale data and the intrinsic difficulties in modeling any process of social contagion quantitatively. For, while the analogy with biological epidemics is conceptually appealing, the information contagion process encompasses many more facets than the biological one, including, e.g. the mixture of endogenous and exogenous factor that shape the behavior of information spreading, since almost no social network can be considered to be a closed

system. For example, the widespread availability of news broadcasted by traditional media may have enormous influence on the speed and duration of diffusion processes and the mutual interaction and possible feedback loops between traditional and online media only further complicate the modeling of exogenous factors that influence the diffusion process. Additionally, information concerning the microscopic process of diffusion (i.e., the single events by which information is passed along the network) is often unavailable. We may assume that a rise in the popularity of a meme derives from the spreading of information over one or more overlapping social networks, but these networks have traditionally been difficult to observe.

The above complications have hindered the understanding of information processes, relegating models to a conceptual role focused on the ability to reproduce coarse-grained, large-scale features of the process under scrutiny [86]. This has made it impossible to discriminate between competing models of the microscopic processes that drive the diffusion. The dynamical behaviors observed in information spreading are obviously much more varied and heterogenous than the biological analogs, but we lack even the basic classification and categorization achieved in biology long ago.

The empirical infrastructure built in this project allows us for the first time to put forward a research plan aimed at overcoming the above problems. The sheer size of data that we will be able to analyze and the possibility of gathering data from both the diffusion process and the network evolution marks an unprecedented opportunity for the understanding of information contagion processes. We will begin our research with the following simple hypotheses: *The arrival of new information is accompanied by predictable changes in communication patterns, network structure and macroscopic dynamical behavior of the diffusion process. Social and communication network structure affects the speed and depth of information spread.* Around these hypotheses we want to articulate the following key research questions:

- Is it possible to define general types of information spreading behavior?
- How can different emerging classes of information spreading behavior be captured by formal mathematical models?
- How can the relationship between interactions and network structure be quantified? What is the appropriate time scale for representing interactions of co-evolution mechanisms that can be mapped into specific particle-network equation classes?

The technical approach we have identified to tackle the above question is the particle-network framework in the case of multi-scale networks. The key idea builds on the proven success of reaction-diffusion processes for network flow modeling. In their simplest formulation, communication processes, and social and biological contagion are equivalent to classic reaction-diffusion processes used in many physical, chemical, and biological systems [17, 18]. The particle-network framework is an ideal framework to undertake the study of spreading, mobility, and diffusion processes in a wide range of problems. Each particle diffuses along edges connecting nodes with a diffusion coefficient d_{ij} that depends on the node degree, node attributes and/or the mobility matrix. Within each node particles may react according to various schemes that represent possible interactions among information/individuals. This framework has the advantage of dealing with arbitrary network structures (heavy-tailed, homogeneous, etc.) as well as combination of them. This basic reaction-diffusion framework can be used to study the propagation of information as well as individuals' interactions, the dynamics of which depend on or are modulated by the structural properties (e.g., node degrees) of the underlying network. The above basic reaction-diffusion framework can be further extended by including multiple particle types, changes of state or awareness, and birth-death processes to model the injection/absorption of information. Such an extended framework is particularly useful for modeling networks under critical conditions or stress. In this task the reaction-diffusion formulation on networks will allow us to explore the correlation between

feedback and co-evolution of the network structure and the equilibrium stationary distribution of particles and their flows.

Within this framework we intend to classify the various processes emerging from real data analysis. In particular we aim to identify the general non-linear coupling mechanisms able to reproduce the different behavioral types identified in the data analysis. At the same time, the particle-network framework will allow us to deal with the co-evolution of network and diffusion processes once the relevant time scales of both processes have been identified. We will develop and employ models capturing the co-evolution of opinion clusters with the network structure. We will consider other drivers of social tie formation and dissolution for inclusion in the modeling framework by using metapopulation approaches [18], in which subpopulations are defined in the abstract spaces of topics and opinions.

3.3 Sentiment Analysis

An integrated research program to evaluate the effects of semantics and emotional valence on meme propagation must aim to make fundamental advances in both the tools that are presently available for sentiment tracking and topic detection, and scientific models of how those various attributes of meme content interact with online context. In the proposed work we will leverage the Klatsch framework to undertake scientific investigations with the objective of modeling the interactions between cognitive, affective, and network factors, and their effects on enhancing or preventing the prevalence of certain memes over time. These efforts will shed light on the role that semantics and sentiment play in the propagation of memes. Our efforts will focus on 3 related tracks of activities:

(1) Improving sentiment tracking tools for online memes. The study of the effects of cognitive and emotional factors on meme diffusion requires the availability of tools that can accurately measure the affective and semantic content of memes which are characterized by: a lack of textual content; presence of non-standard grammar and vocabulary; and idiomatic, community-defined expressions. Generally speaking, most research in sentiment analysis relies on the classification of sentiment into either positive or negative categories [64, 42, 66, 94]. These approaches ignore the complex nature of human mood states, and hinge on a large amount of manually labeled training data, which are costly, time-consuming and inherently subjective to create. Starting from our existing research program we propose to conduct a thorough, scientific investigation of online sentiment tracking tools that are suitable to study the diffusion of online memes [14]. This work will involve a program for the automated creation and fine-tuning of large-scale training sets suitable for online meme characterization, potentially leveraging crowd-sourcing methods such as Mechanical Turk, and the creation of online lexicons potentially derived from available linguistic data sets such as Google's LDC n-grams and online sources such as the records of publicly-available Twitter data. We will conduct a scientific program to design a variety of novel sentiment and mood tracking methods with a particular focus on their empirical cross-validation against other socio-economic indicators such as stock market data, weather data, and news event data.

(2) Modeling meme bursts in relation to affective and cognitive features. Given the occurrence of particular meme bursts, i.e. memes that undergo sudden and consistent changes in their diffusion and prevalence in the online community, sentiment tracking tools can be leveraged to determine the affective and cognitive correlates of these memes. The outcome of track (1) will provide a foundation for work done in this area. This study wil investigate the following scientific questions:

1. What are the cognitive and affective correlates of meme bursts vs. memes that do not propagate as efficiently or 'virally'?

2. Can the observed cognitive and affective features of memes be related to (i) their magnitude and temporal distributions, (ii) prevalence among certain online communities, and (iii) occurrence along with other memes, or in the presence of certain beliefs?
3. What are the affective and cognitive differences between memes 'engineered' to evoke particular emotional responses vs. 'naturally-occurring' memes?

(3) Expansion of tracks (1) and (2) to study the impact of network topology and social environment on meme diffusion. The affective and cognitive features of a meme are expected to interact. In addition a number of other factors such as network topology and the community structure of online users may attenuate or strengthen its effects. For example, a particular meme may be engineered to provoke an emotional response among social conservatives, but could be released in a manner such that its distribution remains limited to a small sub-network of well-connected groups. In addition, a meme may be engineered to provoke an emotional response that is at odds with its semantic features (content), and therefore becomes very popular among political adversaries, cf. the "This is Great news! For John McCain!" meme. Existing research has focused on the mechanics of diffusion and percolation of ideas through online social networks, but the interaction between meme features on the cognitive and affective levels with network features has not yet been adequately modeled. In this portion of our research efforts we will focus our efforts on answering the following questions:

1. With respect to features and network topology, do memes with certain affective and semantic features propagate better in certain network topologies?
2. Do memes' features interact with the nature of certain user clusters such as profiles and topic detection?

4 Preliminary Work

We have begun designing the framework and language (*Klatsch*) described in the previous section. In Section 6 we describe a prototype system, called *Truthy,* that is being developed as a proof of concept for the framework.

Previously, we have performed a quantitative, large-scale, longitudinal analysis of the dynamics of online content popularity in two massive model systems: Wikipedia and the Chilean Web space. In these systems, we tracked the change in the number of links to pages, and the number of times pages were visited. We found that these changes occur in bursts, the magnitude and time separation of which are very broadly distributed. To make sense of these empirical results, we offered a simple model that mimics the exogenous shifts of user attention and the ensuing non-linear perturbations in popularity rankings. While established models based on preferential attachment are insufficient to explain the observed dynamics, our stylized model was successful in recovering the key features observed in the empirical analysis of our systems [77, 76, 75]. The burst analysis in this prior work, and in particular the logarithmic derivative as a signature of very large shifts in user attention has direct applicability to the detection of meme outbreaks.

We have also carried out extensive work on modeling Web traffic [49, 50, 48, 30, 51]. For example, our analysis of aggregate and individual Web requests has shown that PageRank is a poor predictor of traffic. We used empirical data to characterize properties of Web traffic not reproduced by Markovian models, including both aggregate statistics such as page and link traffic, and individual statistics such as entropy and session size. As no existing model reconciled all of these observations, we introduced an agent-based model that explains them through realistic browsing behaviors: revisiting bookmarked pages; backtracking; and seeking out novel pages of topical interest. The resulting model was capable of reproducing the behaviors observed in empirical

data, especially heterogeneous session lengths, reconciling the narrowly-focused browsing patterns of individual users with the extreme variance in aggregate traffic measurements. This work identified a few salient features that are necessary and sufficient to interpret Web traffic data. Beyond the descriptive and explanatory power of our model, these results will provide us with a solid theoretical basis to model the attention processes that drive the spread of information in social media.

Also relevant is our prior work on the relationship between online traffic and network growth models [57, 26]. Since search engines bias the traffic of users according to their page ranking strategies, it has been argued that they create a vicious cycle that amplifies the dominance of established and already-popular sites. This practice could lead to a dangerous monopoly of information. We showed that, contrary to intuition, empirical data do not support this conclusion. We introduced a model that takes into consideration the topical interests of users and their searching behaviors in addition to the way search engines rank pages. Such a model accurately predicts traffic data patterns. The heterogeneity of user interests explained the observed mitigation of search engines' popularity bias. In the proposed project we will explore how such feedback loops, which can now occur instantaneously, can affect the vulnerability of search engines to manipulation in social media — the 'Twitter Bomb' phenomenon [63].

Our group has considerable experience in epidemic modeling as well, having substantially advanced understanding of epidemic spreading and forecasting in both the biological and information technology domains [71, 11]. We have shown that heavy-tailed networks consistently lower the epidemic/invasion threshold, which vanishes in the case of increasingly large networks [70]. Our work has brought the potential implications of highly heterogeneous network topologies in the description of diffusion processes to the attention of the research community. By analyzing the behavior of 800 computer viruses in the wild, we provided one of the first mathematical characterizations of spreading dynamics in complex information networks [70]. We have also included traffic flows into our modeling of the process governing the spreading of epidemics [10]. Indeed, these papers are among the most cited in the field of complex networks. More recently we developed a computational platform to simulate the spatio-temporal spread of emerging disease on a world-wide scale and in complex structured populations. Our metapopulation model includes 4,000 urban areas and travel fluxes among them. The platform integrates the International Air Transport Association (IATA) database and worldwide census data. The model and computational platform have been used to quantify the accuracy and reliability of predictions in cases of an emerging infectious diseases.

Over the past several years our team has also accumulated extensive experience in the analysis of large-scale social media data for online sentiment tracking and public mood analysis. Earlier we tracked the public's moods toward the future by analyzing the the collection at futureme.com, which consists of emails that users direct at themselves to be sent at a future date [72]. The confessional nature of these emails and the fact that they were directed at a future date allowed a mood analysis of present sentiments towards the future. For this work we extended an existing psychometric instrument [40] for application to online resources using various linguistic and semantic databases [24]. Our results identified sharp increases in public apprehension about the near future similar to the inversion of yield curves in finance, shortly before the start of the recession of 2007. This work was recently elaborated to models of online, public mood states derived from very large-scale social networking data such as Twitter [14]. In this work we have started to develop novel sentiment tracking tools based on large-scale linguistic data sets such as Google's n-gram data set [12] that is derived from nearly 1 trillion word tokens extracted from publicly-available web pages. The results of our research indicate that along with semantics, sentiment plays a crucial part in the socio-cognitive features of information diffusion and social contagion [78, 74].

5 Project Management and Collaborations

We plan to organize the proposed work into various major tasks, as shown in Figure 5 along with a timeline. The figure also illustrates how tasks will be assigned to each of three graduate students, and which tasks/students will be supervised or co-supervised by which PIs.

In addition to the PIs and graduate students supported by the project, we will be fortunate to rely on an array of partners. Mark Meiss will be a key collaborator. Mark recently obtained his PhD in Computer Science at IU, under the supervision of PI Menczer, and is about to take a position at Google. Dr. Meiss has been instrumental in laying the foundations of the Klatsch design proposed here, and will continue to serve in an advisory role. We also hope that he will facilitate a collaboration with Google, on an open-source basis, once the platform has proven its value to the community.

A second critical collaboration is with colleagues at the ISI Foundation in Torino, Italy (`www.isi.it`). PIs Menczer and Vespignani maintain active research relationships with ISI, as senior fellow and scientific director, respectively. In particular we will collaborate with the group led by Dr. Ciro Cattuto in the Complex Networks Lagrange Lab at ISI. The group will provide valuable support in complex networks analysis, social media data management, and epidemic modeling.

Thirdly, we are fortunate to collaborate with two colleagues at Wellesley College, Drs. Takis Metaxas and Eni Mustafaraj, who first discovered how Twitter can be abused for political smear campaigns. They demonstrated that due to real-time search based on micro-blogging, search engines can be vulnerable to Twitter Bombs [63]. With their assistance, we hope to detect such abuses in real time, as illustrated in Section 6.

The project has direct relevance to the social sciences, as the process we propose to study — diffusion of ideas and information through social media — is a communication phenomenon. The PIs are not social scientists, however the School of Information has a social informatics group with whom we will collaborate, especially drawing on their expertise in ethnographic methods to understand the patterns we find and inform our models. Additionally, we will consult with Emily Metzgar, a faculty in the School of Journalism who has tracked the development of memes in online discussion surrounding presidential elections. We will consult with Prof. Metzgar in the application of our platform described in Section 6, to create a useful service for journalists investigating the emergence of memes in online discussion of elections and other newsworthy events.

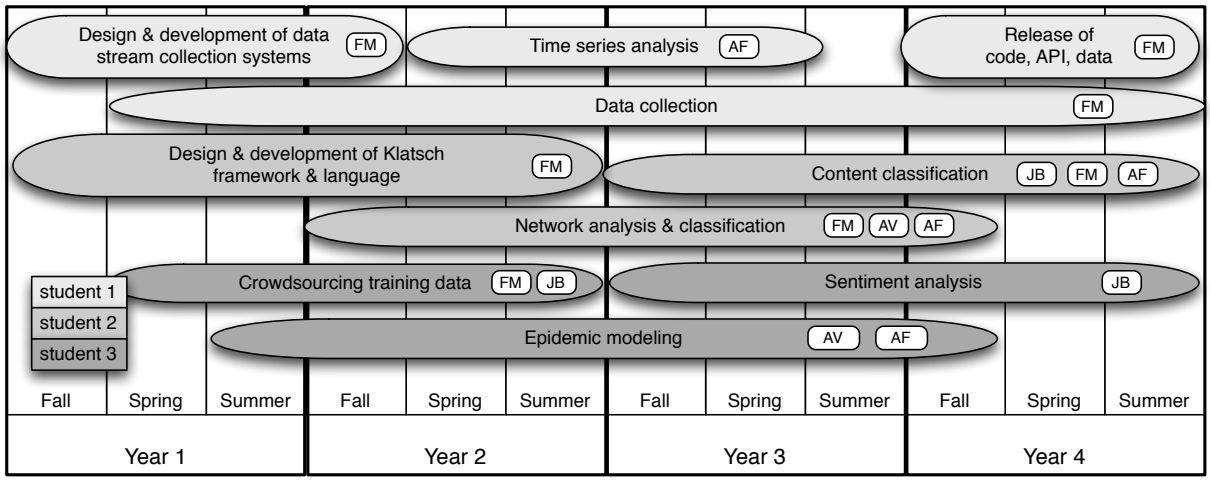

Figure 5: Timeline of the proposed project. Gray scales represent students assigned to tasks, and initials represent PI supervision.

Last but not least, we will take advantage of a collaboration with Yahoo! Labs in Barcelona, Spain, led by Prof. Ricardo Baeza-Yates. In particular we will apply our framework to study and model the evolution of Yahoo! Meme. Yahoo! Labs will be providing us access to this micro-blogging platform's entire history of posts and social links since the inception of the system. This will give our team a unique opportunity to analyze the interactions between dynamics of the network (its evolution) and of the network (messages).

On the infrastructure side, we will avail ourselves of excellent existing resources, both at the university level and in our Center for Complex Networks and Systems Research. These resources include both supercomputers and mass storage, and several 16-core servers with large amounts of memory and in excess of 20TB of shared SAN storage. Along with this infrastructure we have access to excellent computing support within the School of Informatics and Computing.

We will coordinate the proposed research through weekly meetings of the existing research groups of each PI, plus bi-weekly meetings of the PIs and graduate students dedicated to the project. These will be used for progress reports, problem solving, brainstorming, and planning.

Finally, our dissemination plans include targeting various key conferences and meetings that will allow us to present our results to the various communities relevant to our interdisciplinary project. These include Web and data mining (WSDM, WWW, CIKM, KDD), social media (ICWSM, Hypertext), complex systems (ECCS), and social networks (Sunbelt).

6 Broader Impacts of Research

6.1 Community and Societal Impact

The research proposed here has immediate applications of broad societal impact. The Klatsch platform to monitor, analyze, model, and visualize the diffusion of memes will be developed as open source and made available to the research community. It will be designed in a flexible way, making it easy to adapt to different themes and other research foci. The sentiment analysis algorithms will also be shared with the community. Our data will be made publicly available to researchers as well as the wide public via APIs. This data will include meme propagation networks and their statistical features, as well as user and content features.

With respect to the public, we will develop a web service allowing people to follow trending, bursty, and suspicious memes, and see how they have emerged, providing data on temporal evolution, diffusion through the social network, and sentiment content. As a preliminary test of our ideas, we are developing prototypes to monitor chatter during upcoming political elections in the US and Italy (`truthy.indiana.edu`), and track changes in public mood states over time (`terramood.informatics.indiana.edu`). One goal is to explain and mitigate the diffusion of false or deliberately confounding ideas on the basis of epidemiological models that take into account a meme's cognitive and emotional features, as well as network data mining and crowdsourcing techniques. We have already observed successful instances of such astroturfing campaigns in recent elections, via Twitter and with the unintentional amplification by Google's real-time search [63]. Our platform will ultimately allow the general public to check whether assertions propagated via mass social media are genuine grass-roots efforts or engineered manipulative operations. Longer-term implications with broad societal impact include the detection of hate speech by content analysis and sentiment tracking.

6.2 Integration of Teaching, Mentoring, and Research

The PIs are actively engaged in teaching and training both graduate and undergraduate students in the Computer Science and Informatics programs. PI Flammini, as director of the Informatics Undergraduate program, is leading an effort to modernize the curriculum along multiple dimen-

sions. First, we want to create synergies between students with stronger technical backgrounds and interests (a profile that is more prevalent among computer science students) and students with applied interests and backgrounds in the social sciences as is more common among our informatics students. Second, we seek a better alignment between the informatics curriculum and the research expertise and efforts of our faculty. This has the potential to increase the involvement of undergraduate students in cutting-edge research. PIs Menczer and Vespignani are currently developing a novel undergraduate course focused on Socio-Technical Complex Networks, which achieved full capacity in its first offering. We are confident that we have identified a path towards a mutually-beneficial integration of our research efforts with the curriculum. In addition, this effort brings forth an exciting opportunity to attract underrepresented minorities and women, who have been shown to find greater appeal in socially relevant applications of technology.

At the graduate level the PIs are all involved in the Complex Systems track of the Informatics PhD as well as the Computer Science PhD. Indeed, our research groups provide an ideal interdisciplinary environment in which these populations of students can successfully collaborate. An initiative is currently underway to extend this model beyond the School of Informatics and Computing with an IGERT proposal on network science that is under preparation in collaboration with colleagues in Sociology, Cognitive Science, Library and Information Science, Economics, and Medicine.

6.3 Diversity Recruitment

We intend to partner with the NSF-funded Alliance for the Advancement of African-American Researchers in Computing (A4RC), whose Program Manager lies within the School of Informatics and Computing. The A4RC mission is to be A Force for Change, by increasing the number of African-American recipients of advanced degrees in Computing via collaborative partnerships between Historically Black Colleges and Universities (HBCUs) and research universities. A4RC has agreed to partner with our project and to provide one MS-level African-American HBCU student researcher for each of the first three summers of this proposal. Prior to the first summer of this grant we will visit the HBCU campus to give a research talk and to meet with the intern assigned to us to kick-off the project in advance and to meet with faculty there who are interested in this research area. We will support the A4RC mission by encouraging the intern to pursue the PhD path. We will evaluate the success of our program by an in-depth pre- and post-experience assessment that will be used to guide us in providing quality experiences for our student visitors. During the summer of 2010, the School of Informatics and Computing hosted 17 minority student researchers and successfully adapted the Affinity Research Group model for use with our research teams. Assessment results were very positive and we plan to continue using this model. Finally, to further ensure an optimal experience for our student researchers they will participate as members of the IU Summer STEM Scholars initiative, which is an organized program to maximize the sense of community and thus increase the chances of a successful research and career-building experience.

7 Results from Prior NSF Support

Menczer Dr. Menczer's funding from NSF Career award IIS-0133124/0348940 ran out in 2007. The project studied *Scalable search engines via adaptive topic-driven crawlers.* It generated a solid evaluation framework [84], which was made publicly available [58] and was used to compare several crawlers in both general and specialized domains [83, 69], including adaptive crawling algorithms that are currently state-of-the-art [60]. We released an open-source Java library for topical Web crawlers [67] and a public demo of query-driven, client-based, adaptive Web crawling [68, 59]. The project also studied the relationships between Web page similarity measures based on content, links, and meaning [52, 55, 53, 56, 44, 43] leading to models of Web growth that help us understand

how content and links coevolve [52, 54] and how search affects the dynamics of the Web [26, 25]. In the last phase of the project, we began to address the integration between topical crawlers and distributed search systems, with promising preliminary results and a working prototype called `Sixearch.org` [4, 92, 91, 3, 61, 93]. Dr. Menczer's current NSF grant IIS-0811994 ($449,995 from 9/1/2008 to 8/31/2011) supports the study of *Social Integration of Semantic Annotation Networks for Web Applications.* As part of this research we actively develop `GiveALink.org`, a social tagging site sustained through crowdsourcing services and applications, such as games [79, 45, 89, 32]. We design information-theoretic similarity measures based on the annotation data generated by these systems, giving rise to semantic networks among resources, tags, and people [81, 47, 46].

Vespignani Dr. Vespignani is one of the co-PIs of the NSF IIS-0513650 award "Net-Work-Bench: A Large-Scale Network Analysis, Modeling and Visualization Toolkit for Biomedical, Social Science and Physics Research" ($1.2M, 2005-2009). Dr. Vespignani has been supervising the integration of network analysis algorithms. The NWB tool is publicly available as the NWB 1.0.0 Official Release (`nwb.slis.indiana.edu`). This project has also generated results on the role of transportation networks in the diffusion of large-scale epidemics and the analysis of reaction diffusion processes in heterogeneous networks that are relevant to the present proposal. Dr. Vespignani has also been a co-PI for NSF. Grant No. SES-0527638 supporting the organization of the International Workshop/School and Conference on Network Science. The primary objective of the Conference was to facilitate interactions between social and behavioral scientists and the many other disciplines interested in and utilizing network science. This conference established the NetSci conference series that is now at its fourth edition (Venice 2009) and is attended regularly by more than 300 network science researchers.

Bollen Dr. Bollen's work is supported in 2010-2011 by an NSF grant BCS-1032101, "RAPID: Models of social contagion of charitable sentiment towards Haiti on Twitter". It is well-known that the magnitude of the charitable response to the Haiti disaster was in large part modulated by grass-roots organizing and "viral" social networking on Twitter.com, often through emotional appeals of a personal nature. This project performs a large-scale analysis of Twitter submissions during the unfolding humanitarian disaster in Haiti to study how fluctuations of public sentiment modulate charitable behavior, and whether models of mood contagion can be developed and leveraged to optimize charitable responses in future disasters. This ongoing research is pointing the way towards more sophisticated psychological models of how emotions and their contagion in social network environments shapes pro-social behavior. In addition, Bollen is supported from 2009 to 2012 by the NSF grant SBE-0914939 "COLLABORATIVE RESEARCH: Tracking Scientific Innovation from Usage Data: Models and Tools to Support a Science of Science" which seeks to leverage very large-scale scholarly usage data, collected from some of the world's most significant publishers, aggregators and institutions, to study how shifts in collective attention within the scientific community can underpin efforts to develop predictive models of scientific innovation. In 2009 Dr. Bollen received NSF grant IIS-0936204 "Scholarly Evaluation Metrics: Opportunities and Challenges" to organize a workshop to bring together the world's leading experts in the domain of scholarly assessment to determine ways to shift the incentives structure of scientific research to more open, innovative and fair models (`http://informatics.indiana.edu/scholmet09/announcement.html`).

References

[1] E. Adar and L. A. Adamic. Tracking information epidemics in blogspace. In *Conference on Web Intelligence*, 2005.

[2] E. Adar, L. Zhang, L. Adamic, and R. M. Lukose. Implicit structure and the dynamics of blogspace. In *Workshop on the Weblogging Ecosystem, 13th International World Wide Web Conference*, 2004.

[3] R. Akavipat, L.-S. Wu, A. G. Maguitman, and F. Menczer. Emerging semantic communities in peer web search. In *Proc. ACM CIKM Workshop on Information Retrieval in Peer-to-Peer Networks (P2PIR)*, 2006.

[4] R. Akavipat, L.-S. Wu, and F. Menczer. Small world peer networks in distributed Web search. In *Alt. Track Papers and Posters Proc. 13th International World Wide Web Conference*, pages 396–397, 2004.

[5] R. M. Anderson and R. M. May. *Infectious Diseases in Humans.* Oxford University Press, 1992.

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

[7] N. Bailey. *The Mathematical Theory of Infectious Diseases and its Applications.* Griffin, London, 2nd edition, 1975.

[8] E. Bakshy, K. Brian, and A. Lada. Social influence and the diffusion of user created content. In *Proc. 10th ACM Conference on Electronic commerce*, 2009.

[9] A.-L. Barabási. The origin of bursts and heavy tails in human dynamics. *Nature*, 435:207–211, 2005.

[10] A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. *Proc. Natl. Acad. Sci. USA*, 101:3747–3752, 2004.

[11] A. Barrat, M. Barthélemy, and A. Vespignani. *Dynamical Processes on Complex Networks.* Cambridge University Press, 2008.

[12] S. Bergsma, D. Lin, and R. Goebel. *Web-scale N-gram models for lexical disambiguation*, pages 1507–1512. Morgan Kaufmann Publishers, 2009.

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

[14] J. Bollen, A. Pepe, and H. Mao. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *http://arxiv.org/abs/0911.1583*, 2010.

[15] M. M. Bradley and P. J. Lang. Affective norms for English words (ANEW): Stimuli, instruction manual, and affective ratings. Technical report, Center for Research in Psychophysiology, University of Florida, Gainesville, Florida, 1999.

[16] D. Centola. The spread of behavior in an online social network experiment. *Science*, 329(5996):1194–1197, 2010.

[17] V. Colizza, P.-S. R., and A. Vespignani. Reaction-diffusion processes and metapopulation models in heterogeneous networks. *Nat. Phys.*, 3:276–282, 2007.

[18] V. Colizza and A. Vespignani. Epidemic modeling in metapopulation systems with heterogeneous coupling pattern: Theory and simulations. *J. Theor. Biol.*, 251:450–467, 2008.

[19] R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proc. Natl. Acad. Sci. USA*, 105(41):15649–15653, 2008.

[20] D. J. Daley and D. G. Kendall. Epidemics and rumours. *Nature*, 204(4963):1118, 1964.

[21] Z. Dezso, E. Almaas, A. Lukacs, B. Racz, I. Szakadat, and A. Barabasi. Dynamics of information access on the Web. *Phys. Rev. E*, 73:066132, 2006.

[22] P. S. Dodds and C. M. Danforth. Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness*, 2009.

[23] A. Esuli and F. Sebastiani. {SENTIWORDNET}: A Publicly Available Lexical Resource for Opinion Mining, 2006.

[24] C. Fellbaum. *WordNet - An Electronic Lexical Database*. The MIT Press.

[25] S. Fortunato, A. Flammini, and F. Menczer. Scale-free network growth by ranking. *Phys. Rev. Lett.*, 96(21):218701, 2006.

[26] S. Fortunato, A. Flammini, F. Menczer, and A. Vespignani. Topical interests and the mitigation of search engine bias. *Proc. Natl. Acad. Sci. USA*, 103(34):12684–12689, 2006.

[27] W. Goffman and V. A. Newill. Generalization of epidemic theory: an application to the transmission of ideas. *Nature*, 204(4955):225–228, 1964.

[28] J. Goldenberg, B. Libai, and E. Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Market. Lett.*, 3(12):211—223, 2001.

[29] S. A. Golder, D. Wilkinson, and B. A. Huberman. Rhythms of social interaction: Messaging within a massive online network. In *Proc. of Third International Conference on Communities and Technologies*, pages 41–66, 2007.

[30] B. Goncalves, M. Meiss, J. Ramasco, A. Flammini, and F. Menczer. Remembering what we like: Toward an agent-based model of web traffic. In *Late-breaking result at WSDM*, 2009.

[31] M. Granovetter. Threshold models of collective behavior. *Ameri. J. Sociol.*, 83(6):1420–1433, 1978.

[32] D. T. Hoang, J. Kaur, and F. Menczer. Crowdsourcing scholarly data. In *Proc. Web Science Conference: Extending the Frontiers of Society On-Line (WebSci)*, 2010.

[33] J. Huang, K. M. Thornton, and E. N. Efthimiadis. Conversational tagging in twitter. In *Proc. of the 21st ACM conference on hypertext and hypermedia*, 2010.

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

[35] J. Kleinberg. Bursty and hierarchical structure in streams. In *Proc. 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (KDD)*, 2002.

[36] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *Proc. 12th International World Wide Web Conference*, pages 568–576, 2003.

[37] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 591–600, New York, NY, USA, 2010. ACM.

[38] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM Transactions on the Web*, 1, 2007.

[39] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506. ACM, 2009.

[40] A. LeUnes. Updated bibliography on the profile of mood states in sport and exercise psychology research. *Journal of Applied Sport Psychology*, 12:110–113, 2000.

[41] D. Liben-Nowell and J. Kleinberg. Tracing information flow on a global scale using Internet chain-letter data. *Proc. Natl. Acad. Sci. USA*, 105(12):4633–4638, 2008.

[42] Y. Liu, X. Huang, A. An, and X. Yu. Arsa: a sentiment-aware model for predicting sales performance using blogs. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 607–614. ACM, 2007.

[43] A. G. Maguitman, F. Menczer, F. Erdinc, H. Roinestad, and A. Vespignani. Algorithmic computation and approximation of semantic similarity. *World Wide Web*, 9(4):431–456, December 2006.

[44] A. G. Maguitman, F. Menczer, H. Roinestad, and A. Vespignani. Algorithmic detection of semantic similarity. In *Proc. 14th International World Wide Web Conference*, pages 107–116, 2005.

[45] B. Markines, C. Cattuto, and F. Menczer. Social spam detection. In *Proc. 5th International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2009.

[46] B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and G. Stumme. Evaluating similarity measures for emergent semantics of social tagging. In *Proc. WWW*, pages 641–650, 2009.

[47] B. Markines and F. Menczer. A scalable, collaborative similarity measure for social annotation systems. In *Proc. 20th ACM Conf. on Hypertext and Hypermedia (HT)*, pages 347–348, 2009.

[48] M. Meiss, J. Duncan, B. Goncalves, J. Ramasco, and F. Menczer. What's in a session: Tracking individual behavior on the web. In *Proc. 20th ACM Conf. on Hypertext and Hypermedia (HT)*, pages 173–182, 2009.

[49] M. Meiss, B. Goncalves, J. Ramasco, A. Flammini, and F. Menczer. Modeling traffic on the web graph. Under review.

[50] M. Meiss, B. Goncalves, J. Ramasco, A. Flammini, and F. Menczer. Agents, Bookmarks and Clicks: A topical model of Web navigation. In *Proc. 21th ACM Conf. on Hypertext and Hypermedia (HT)*, 2010.

[51] M. Meiss, F. Menczer, S. Fortunato, A. Flammini, and A. Vespignani. Ranking web sites with real user traffic. In *Proc. First ACM International Conference on Web Search and Data Mining (WSDM)*, pages 65–75, 2008.

[52] F. Menczer. Growing and navigating the small world Web by local content. *Proc. Natl. Acad. Sci. USA*, 99(22):14014–14019, 2002.

[53] F. Menczer. Correlated topologies in citation networks and the web. *European Physical Journal B*, 38(2):211–221, 2004.

[54] F. Menczer. The evolution of document networks. *Proc. Natl. Acad. Sci. USA*, 101:5261–5265, 2004.

[55] F. Menczer. Lexical and semantic clustering by Web links. *Journal of the American Society for Information Science and Technology*, 55(14):1261–1269, 2004.

[56] F. Menczer. Mapping the semantics of web text and links. *IEEE Internet Computing*, 9(3):27–36, May/June 2005.

[57] F. Menczer, S. Fortunato, A. Flammini, and A. Vespignani. Googlearchy or googlocracy? *IEEE Spectrum Online*, February 2006.

[58] F. Menczer and G. Pant. A general evaluation framework for topical crawlers: Support data and script. http://informatics.indiana.edu/fil/IS/Framework/, 2002.

[59] F. Menczer, G. Pant, and M. Degeratu. Myspiders applet. http://myspiders.informatics.indiana.edu, 2002.

[60] F. Menczer, G. Pant, and P. Srinivasan. Topical web crawlers: Evaluating adaptive algorithms. *ACM Transactions on Internet Technology*, 4(4):378–419, 2004.

[61] F. Menczer, L.-S. Wu, and R. Akavipat. Intelligent peer networks for collaborative web search. *AI Magazine*, 29(3):35, 2008.

[62] Y. Moreno, M. Nekovee, and A. Vespignani. Efficiency and reliability of epidemic data dissemination in complex networks. *Phys. Rev. E*, 69:055101(R), 2004.

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

[64] A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), 2010.

[65] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*, pages 271–278, 2004.

[66] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86, 2002.

[67] G. Pant and F. Menczer. Javacrawlers. http://informatics.indiana.edu/fil/IS/JavaCrawlers/, 2002.

[68] G. Pant and F. Menczer. MySpiders: Evolve your own intelligent Web crawlers. *Autonomous Agents and Multi-Agent Systems*, 5(2):221–229, 2002.

[69] G. Pant and F. Menczer. Topical crawling for business intelligence. In T. Koch and I. Solvberg, editors, *Proc. 7th European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*, Lecture Notes in Computer Science, Vol. 2769, Berlin, 2003.

[70] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86:3200–3203, 2001.

[71] R. Pastor-Satorras and A. Vespignani. *Evolution and structure of the Internet: A Statistical Physics Approach*. Cambridge University Press, 2004.

[72] A. Pepe and J. Bollen. Between conjecture and memento: shaping a collective emotional perception of the future. In *Proceedings of the AAAI 2008 Spring Symposium on*, 2008.

[73] A. Rapoport. Spread of information through a population with socio-structural bias: I. assumption of transitivity. *Bull. Math. Biol.*, 15(523–533), 1953.

[74] S. Rasmussen, D. Mangalagiu, H. Ziock, J. Bollen, and G. Keating. *Collective intelligence for decision support in very large stakeholder networks: The future US energy system.*, pages 468–475. IEEE, 2007.

[75] J. Ratkiewicz, A. Flammini, and F. Menczer. Traffic in Social Media I: Paths Through Information Networks. In *Proc. International Symposium on Social Intelligence and Networking (SIN-10)*. IEEE, 2010.

[76] J. Ratkiewicz, S. Fortunato, A. Flammini, F. Menczer, and A. Vespignani. Characterizing and modeling the dynamics of online popularity. *Phys. Rev. Lett.*, 2010. In press.

[77] J. Ratkiewicz, F. Menczer, S. Fortunato, A. Flammini, and A. Vespignani. Traffic in Social Media II: Modeling Bursty Popularity. In *Proc. International Symposium on Social Intelligence and Networking (SIN-10)*. IEEE, 2010.

[78] M. A. Rodriguez, D. J. Steinbock, J. H. Watkins, C. Gershenson, J. Bollen, V. Grey, and B. deGraf. Smartocracy: Social networks for collective decision making. In *Proceedings of the International Conference on Systems Science (HICSS)*, 2007.

[79] H. Roinestad, J. Burgoon, B. Markines, and F. Menczer. Incentives for social annotation. In *Proc. 32nd Annual ACM SIGIR Intl. Conf. on Research and Development in Information Retrieval*, page 838, 2009.

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

[81] R. Schifanella, A. Barrat, C. Cattuto, B. Markines, and F. Menczer. Folks in folksonomies: Social link prediction from shared metadata. In *Proc. 3rd ACM International Conference on Web Search and Data Mining (WSDM)*, 2010.

[82] H. A. Simon. Designing organizations for an information-rich world. In M. Greenberger, editor, *Computers, Communication, and the Public Interest*, pages 37–72. The Johns Hopkins Press, Baltimore, 1971.

[83] P. Srinivasan, J. Mitchell, O. Bodenreider, G. Pant, and F. Menczer. Web crawling agents for retrieving biomedical information. In *Proc. Int. Workshop on Agents in Bioinformatics (NETTAB-02)*, 2002.

[84] P. Srinivasan, G. Pant, and F. Menczer. A general evaluation framework for topical crawlers. *Information Retrieval*, 8(3):417–447, 2005.

[85] B. Suh, L. Hong, P. Pirolli, and E. H. Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Proc. 2010 IEEE International Conference on Social Computing*, 2010.

[86] G. Szabo and B. A. Huberman. Predicting the popularity of online content. *Commun. ACM*, 53(8):80–88, 2010.

[87] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2010.

[88] X. Wei, J. Yang, L. Adamic, R. de Araújo, and M. Rekhi. Diffusion dynamics of games on online social networks. In *Proc. 3rd Workshop on Online Social Networks (WOSN)*, 2010.

[89] L. Weng and F. Menczer. GiveALink Tagging Game: An Incentive for Social Annotation. In *Proc. KDD Human Computation Workshop (HComp)*, 2010.

[90] F. Wu and B. A. Huberman. Novelty and collective attention. *Proc. Natl. Acad. Sci. USA*, 104(45):17599–17601, 2007.

[91] L.-S. Wu, R. Akavipat, and F. Menczer. 6S: Distributing crawling and searching across Web peers. In *Proc. IASTED Int. Conf. on Web Technologies, Applications, and Services (WTAS)*, 2005.

[92] L.-S. Wu, R. Akavipat, and F. Menczer. Adaptive query routing in peer Web search. In *Proc. 14th International World Wide Web Conference*, pages 1074–1075, 2005.

[93] L.-S. Wu and F. Menczer. Diverse peer selection in collaborative web search. In *Proc. ACM Symposium on Applied Computing, Special Track on Information Access and Retrieval (SAC-IAR)*, pages 1005–1009, 2009.

[94] C. Yang, K. H.-Y. Lin, and H.-H. Chen. Emotion classification using web blog corpora. In *WI '07: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 275–278. IEEE Computer Society, 2007.

[95] Y.-H. Yang, C.-C. Liu, and H. H. Chen. Music emotion classification: a fuzzy approach. In {*MULTIMEDIA*} *'06: Proceedings of the 14th annual ACM international conference on Multimedia*, pages 81–84, New York, NY, USA, 2006. ACM.