Social Mechanics: An Empirically Grounded Science of Social Media

Kristina Lerman, Aram Galstyan, Greg Ver Steeg
USC Information Sciences Institute
Marina del Rey, CA

Tad Hogg
Institute for Molecular Manufacturing
Palo Alto, CA

Abstract

What will social media sites of tomorrow look like? What behaviors will their interfaces enable? A major challenge for designing new sites that allow a broader range of user actions is the difficulty of extrapolating from experience with current sites without first distinguishing correlations from underlying causal mechanisms. The growing availability of data on user activities provides new opportunities to uncover correlations among user activity, contributed content and the structure of links among users. However, such correlations do not necessarily translate into predictive models. Instead, empirically grounded mechanistic models provide a stronger basis for establishing causal mechanisms and discovering the underlying statistical laws governing social behavior. We describe a statistical physics-based framework for modeling and analyzing social media and illustrate its application to the problems of prediction and inference. We hope these examples will inspire the research community to explore these methods to look for empirically valid causal mechanisms for the observed correlations.

1 Introduction

The study of social systems is in the midst of a transformation into a hard science. While a confluence of factors is responsible for this transformation, at the core, it is driven by the widespread availability of detailed data about human behavior enabled by social media. People are joining sites such Twitter, Digg, Flickr, Delicious, and YouTube among others, to post or find interesting content, talk about themselves and things they find interesting, connect to and interact with friends and like-minded people through online social networks. Traces of human activity are exposed by the sites themselves, often for 3rd parties to build value-added applications on top of the data, with researchers becoming unintended beneficiaries. While the availability of data affects the kinds of questions social scientists can now ask about individual and group behavior (Lazer 2009), we focus on an equally dramatic data-driven transformation in the *practice* of research. The goal of this practice is to uncover basic statistical laws of social interactions that generalize across different social media data sets. We describe how these discoveries can be used to empirically ground models that help explain and predict social behavior. Understanding these essential ingredients of social interactions and behavior will allow us to design future social media sites to effectively achieve specific goals, such as optimizing a public good or acting as a computational system.

In this position paper we describe the statistical mechanics-inspired approach we use to study two problems commonly encountered in social data mining: *prediction* and *inference* in networks. A common solution used by the machine learning and data mining communities involves applying statistical regression-based methods to classify large available data sets according to features in the data. Such methods can identify correlations among sets of features or behaviors, which are then used to predict outcomes in new cases. However, these approaches are limited in their ability to identify causal mechanisms. Experiments, especially with multiple randomly-selected groups (Salganik, Dodds, and Watts 2006), are a more powerful approach, but they are seldom practical in the social media domain.

Statistical mechanics provides an alternative framework for studying social media. We explain the approach and illustrate with applications to social media analysis. Stochastic models can be used to identify key mechanisms relating the design choices of social media sites to the collective behavior observed on them (Lerman 2007a; Hogg and Lerman 2009; Hogg and Szabo 2009; Iribarren and Moro 2009; Castellano, Fortunato, and Loreto 2009). These models consider a few key features of the social media web site to define a set of states among which users and content transition probabilistically. By comparing predictions of the models to observed user behavior, such models could aid development of future social media services by identifying key mechanisms leading to successful outcomes. In Section 2 we show how we used stochastic modeling to study social dynamics of news aggregator Digg and used these models to predict popularity of content on this site (Lerman and Hogg 2010).

As another illustration, we study dynamics of social contagion (e.g., information spread) on online social networks. Understanding this process is crucial to identifying influential users, predicting how far contagion will spread, and identifying methods to enhance or impede its progress. Existing works in this area correlate user features with observed outcomes (e.g., cascade size) (Cha et al. 2010; Bakshy et al. 2011) without investigating the mechanisms

underlying social contagion, or they study the contagion process (Leskovec, Adamic, and Huberman 2007; Romero, Meeder, and Kleinberg 2011) without examining its effect on macroscopic properties of the system (e.g., cascade size). However, details of the underlying dynamic process, specifically, the mechanism for social contagion, dramatically impacts our understanding of network structure and behavior, as well as the choice of metrics used for network analysis (Ghosh and Lerman 2010; Ghosh et al. 2011). In Section 3 we show how we used (Ver Steeg, Ghosh, and Lerman 2011) empirical analysis and simulations to investigate the effect network structure and contagion mechanism have on the macroscopic properties of social contagion.

For the inference task, we consider the problem of detecting *communities* in social networks, which can be loosely defined as a group, or *cluster* of well connected nodes, with relatively lower density of links across different clusters (Newman 2006). Understanding community structure is important for adequately describing various dynamical processes unfolding on networks (Arenas, Díaz-Guilera, and Pérez-Vicente 2006; Galstyan and Cohen 2007; Gleeson 2008; Dorogovtsev et al. 2008), and a large number of methods for finding communities have been proposed (for a recent survey of existing approaches see (Fortunato 2010)). Recent research has also focused on the feasibility of detecting such structures, assuming they are present in the network. Results for the *planted partition* graph models suggest that clusters can be recovered with arbitrary accuracy if sufficient data (link density) is available (Condon and Karp 2001). More recently, this problem of cluster delectability in *sparse* graphs has been addressed within a statistical mechanics framework (Reichardt and Leone 2008). In particular, it was shown that clustering in the sparse planted partition model is characterized by a phase transition from detectable to undetectable regimes as one increases the overlap between the clusters (Reichardt and Leone 2008). We examine the cluster-detection problem in semi-supervised settings, showing how different types of background knowledge about the clusters alter the detection threshold.

Statistical physics-based approaches will not replace traditional machine learning and data mining algorithms. The former are reductionist at their core. Guided by empirical analysis, they look for simple mechanistic (causal) models and underlying statistical principles that allow for a deeper understanding and predictability. As a tradeoff, the models lose specificity, which means they cannot predict outcomes for any specific individual, only for a population on average. Despite these drawbacks, statistical physics-based methods complement existing machine learning and data mining approaches, and inspire this community to look for empirically valid causal mechanisms for the observed correlations.

2 Stochastic Models of Social Dynamics

Descriptions of social media typically focus on aggregate behavior of the large numbers of users that is captured by *average* quantities. Such quantities include average rate at which users contribute and rate content, and explicitly link to other users. Stochastic models provide a useful approach to understanding the dynamics of aggregate behaviors. These models are similar to approaches used in statistical physics, demographics, epidemiology (Ellner and Guckenheimer 2006) and macroeconomics, where the focus is not to reproduce the results of a single observation, but rather to describe the typical behaviors and relations among aggregate quantities, such as vaccination policy and fraction of infected population or interest rates and employment.

We represent an individual entity, whether a user or contributed content, as a stochastic process with a few states. This abstraction captures much of the individual complexity and environmental variability by casting individual's actions as inducing probabilistic transitions between states. While this modeling framework applies to stochastic processes of varying complexity, we focus on processes that obey the Markov property, namely, a user whose future state depends only on her present state and the input she receives. A Markov process is captured by a *state diagram* showing the possible states of the user and conditions for transition between those states. This approach is similar to compartmental models in biology (Ellner and Guckenheimer 2006). For instance, in epidemiology such models track the progress of a disease as shifting individuals between states, or compartments, such as susceptible and infected.

We assume that all users have the same set of states, and that transitions between states depend only on the state and not the individual user. That is, the state captures the key relevant properties determining subsequent user actions. A choice of states to describe users results in grouping users in the same state into the same compartment for modeling. The aggregate state of the system can then be described simply by the *number* of individuals in each state at a given time. That is, the system configuration at this time is defined by the occupation vector: $\vec{n} = (n_1, n_2, \ldots)$ where n_k is the number of individuals in state k.

A key requirement for designing stochastic models is to ensure the state captures enough of the large variation in individual behavior to give a useful description of aggregate system properties. This is particularly challenging when individual activity follows a long-tail distribution, such as seen in some epidemics (Lloyd-Smith et al. 2005), as well as in social media web sites (Wilkinson 2008). In our case, including user link information as part of the state accounts for enough of this variation to provide reasonable accuracy, in particular significantly improving predictions compared to direct extrapolation of voting rates without accounting for the properties of the web site user interface.

The next step in developing the stochastic model is to summarize the variation within the collection of histories of changing occupation vectors with a probabilistic description. That is, we characterize the possible occupation vectors by the probability, $P(\vec{n}, t)$, the system is in configuration $\vec{n}$ at time t. The evolution of $P(\vec{n}, t)$, governed by the Stochastic Master Equation (Kampen 1992), is almost always too complex to be analytically tractable. We can simplify the problem by working with the average occupation number, whose evolution is given by the Rate Equation

$$\frac{d\langle n_k \rangle}{dt} = \sum_j w_{jk}(\langle \vec{n} \rangle)\langle n_j \rangle - \langle n_k \rangle \sum_j w_{kj}(\langle \vec{n} \rangle) \qquad (1)$$

where $\langle n_k \rangle$ denotes the average number of users in state k at time t, i.e., $\sum_{\vec{n}} n_k P(\vec{n}, t)$ and $w_{jk}(\langle \vec{n} \rangle)$ is the transition rate from configuration j to configuration k when the occupation vector is $\langle \vec{n} \rangle$.

Using the average of the occupation vector in the transition rates is a common simplifying technique for stochastic models. A sufficient condition for the accuracy of this approximation is that variations around the average are relatively small. In many stochastic models of systems with large numbers of components, variations are indeed small due to many independent interactions among the components and the short tails of the distributions of these component behaviors. More elaborate versions of the stochastic approach give improved approximations when variations are not small, particularly due to correlated interactions (Opper and Saad 2001) or large individual heterogeneity (Moreno, Pastor-Satorras, and Vespignani 2002). User behavior on the web, however, often involves distributions with long tails, whose typical behaviors differ significantly from the average. In this case we have no guarantee that the averaged approximation is adequate, even when aggregating the behavior of many users . Instead we must test its accuracy for particular aggregate behaviors by comparing model predictions with observations of actual behavior, as we report below.

In the Rate Equation, occupation number n_k increases due to users' transitions from other states to state k, and decreases due to transitions from the state k to other states. The equations can be easily written down from the user state diagram. Each state corresponds to a dynamic variable in the mathematical model — the average number of users in that state — and it is coupled to other variables via transitions between states. Every transition must be accounted for by a term in the equation, with transition rates specified by the details of the interactions between users.

In summary, the stochastic modeling framework is quite general and requires only specifying the aggregate states of interest for describing the system and how individual user behaviors create transitions among these states. The modeling approach is best suited to cases where the users' decisions are mainly determined by a few characteristics of the user and the information they have about the system. These system states and transitions give the rate equations. Solutions to these equations then give estimates of how aggregate behavior varies in time and depends on the characteristics of the users involved.

An Example: Digg

With over 6 million registered users, the social news aggregator Digg is one of the more popular news portals on the Web. Digg allows users to submit and rate news stories by voting on, or 'digging', them. There are many new submissions every minute, over 16,000 a day. Every day Digg picks about a hundred stories that it believes will be most interesting to the community and promotes them to the front page. Digg's front page is created by the collective decision of its many users. Digg's user interface defines how users post or discover new stories and interact with other users. A model of social dynamics has to take these elements into account when describing the evolution of story popularity.

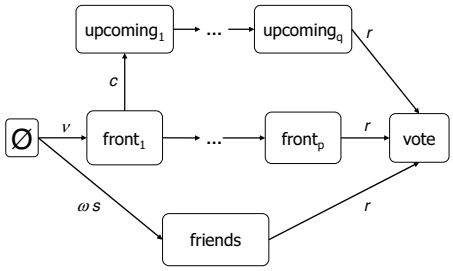

Figure 1: State diagram of user behavior for a single story. A user starts in the $\emptyset$ state at the left, may find the story through one of the three interfaces and may then vote on it. At a given time, the story is located on a particular page of either the upcoming or front page lists, not both. This diagram shows votes for a story on either page p of the front pages or page q of the upcoming pages. Only fans of previous voters can see the story through the friends interface. Users in the friends, front or upcoming states may choose to leave Digg, thereby returning to the $\emptyset$ state (with those transitions not shown in the figure). Users reaching the "vote" state remain there indefinitely and can not vote on the story again. Parameters next to the arrows characterize state transitions.

A newly submitted story goes on the *upcoming* stories list, where it remains for a period of time, typically 24 hours, or until it is promoted to the front page, whichever comes first. The default view shows newly submitted stories as a chronologically ordered list, with the most recently submitted story at the top of the list, 15 stories to a page. To see older stories, a user must navigate to page 2, 3, *etc.* of the upcoming stories list. Promoted stories (Digg calls them 'popular') are also displayed as a chronologically ordered list on the *front pages*, 15 stories to a page, with the most recently promoted story at the top of the list. To see older promoted stories, user must navigate to page 2, 3, *etc.* of the front page. Users vote for the stories they like by 'digging' them.

Digg allows users to designate friends and track their activities, i.e., see the stories friends recently submitted or voted for. The friend relationship is asymmetric. When user A lists user B as a *friend*, A can watch the activities of B but not vice versa. We call A the *fan* of B. A newly submitted story is visible in the upcoming stories list, as well as to submitter's fans through the friends interface. With each vote, a story becomes visible to the voter's fans through the friends interface, which shows the newly submitted stories that user's friends voted for.

A prior study of social dynamics of Digg (Hogg and Lerman 2009) used a simple behavioral model that viewed each Digg user as a stochastic Markov process, whose state diagram with respect to a single story is shown in Fig. 1. According to this model, a user visiting Digg can choose to browse the *front* pages to see the recently promoted stories, *upcoming* stories pages for the recently submitted stories, or use the *friends* interface to see the stories her friends have recently submitted or voted for. She can select a story to read from one of these pages and, if she considers it interesting, *vote* for it. The user's environment, the stories she is seeing, changes in time due to the actions of all the users. We cre-

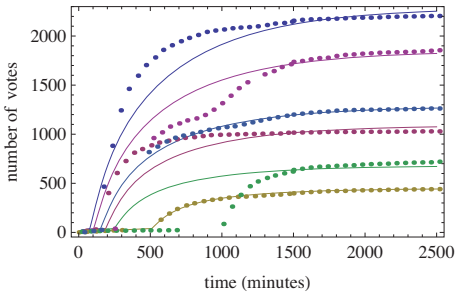

Figure 2: Evolution of the number of votes received by six stories compared with model solution.

ated a model based on these states and transitions, matching the way Digg shows stories to users, how those users browse the web site, and the diversity of how stories appeal to the user community.

We evaluated this model using voting activity and a snapshot of the social network on Digg in June 2006. Figure 2 shows the behavior of six stories that were promoted to the front page and the corresponding solution of the rate equations from the model. In the model, a story has two characteristics: the number of fans of the story's submitter and the probability a user seeing the story will vote for it (the story's "interestingness"). The number of fans is given in our data, but the interestingness must be estimated by matching the rate equation solution to the observed data. Thus each story has a single adjustable parameter for the model. Overall there is qualitative agreement between the data and the model, indicating that the features of the Digg user interface we considered can explain the patterns of collective voting. Specifically, the model reproduces three generic behaviors of Digg stories: (1) slow initial growth in votes of upcoming stories; (2) more interesting stories are promoted to the front page (inflection point in the curve) faster and receive more votes than less interesting stories; (3) however, as first described in (Lerman 2007b), better connected users are more successful in getting their less interesting stories promoted to the front page than poorly-connected users. These observations highlight a benefit of the stochastic approach: identifying simple models of user behavior that are sufficient to produce the aggregate properties of interest. Thus while the stochastic model primarily describes typical story behavior, we see it gives a reasonable match to the actual vote history of individual stories. Nevertheless, there are some cases where individual stories differ considerably from the model, particularly where an early voter happens to have an exceptionally large number of fans, thereby increasing the story's visibility to other users far more than expected. This variation, a consequence of the long-tail distributions involved in social media, is considerably larger than seen, for example, in most statistical physics applications of stochastic models. The effect of such large variations is an important issue to address when using stochastic models to predict the behavior of individual stories in social media.

Predicting Popularity of Stories

By separating the impact of story quality and social influence on the popularity of stories on Digg, a stochastic model of social dynamics enables two novel applications: (1) estimating inherent story quality from the evolution of its observed popularity, and (2) predicting its eventual popularity based on users' early reactions to the story. To predict how popular a story will become, we use early votes, even those cast before the story is promoted, to estimate how interesting it is to the community. With this estimate, the model then determines, on average, the story's subsequent evolution. We showed (Lerman and Hogg 2010) that this method significantly outperforms an alternative prediction method (Szabo and Huberman 2010) that extrapolates the final number of votes from votes immediately after promotion.

By estimating story interestingness from the evolution of its popularity, the model identifies a lognormal distribution of interestingness among the stories (Hogg and Lerman 2009), which is not directly apparent from the data itself which confounds effects of changing visibility of the stories with their appeal to the user community. The observation of this distribution suggests there is an underlying multiplicative process giving rise to user interest. The nature of this process is a significant open question that the stochastic modeling approach helped to identify.

Discussion

The ability of the stochastic approach to incorporate details of user behaviors based on information available on the web site illustrates its value in providing insights into how aggregate behavior arises from the users, in contrast to models that evaluate regularities in the aggregate behaviors only (Wu and Huberman 2007). In particular, user models can help distinguish aggregate behaviors that arise from intrinsic properties of the stories (e.g., their interestingness) from behavior due to the information the web site provides, such as ratings of other users and how stories are placed in the site, i.e., their visibility (Hogg and Lerman 2009). In addition to explaining empirically observed phenomena (e.g., it is easier for submitters with more fans to get a story promoted to the front page, even when the story is less interesting), stochastic models also have predictive power.

Social media sites often allow users to link to others whose activity they find particularly interesting. These links can be a significant factor in how users find and react to content. Thus we can expect that extending stochastic models to account for this community structure will improve their accuracy. One way to partially account for this structure in Digg is distinguishing behavior of users who are fans of prior voters from those who are not. While such a model can faithfully predict the evolution of story's popularity among voters who are not fans of any of the prior voters, and even its popularity among submitter's fans, it does not do a great job predicting votes from other voters' fans. This is the consequence of the simplification made by the mean field approach which models growth in visibility through the social network by an average rate parameter. In order to model information spread through a community, we need to better

understand dynamics of cascades. The following section describes our attempts in that direction.

3 Dynamics of Information Cascades

While the previous section studied the relationship between aggregate quantities like story visibility and popularity, this section considers a stochastic description of the dynamics of individual nodes. In statistical mechanics, these correspond to macroscopic versus microscopic approaches, and the connection between the two has historically played an important role. For instance, the ideal gas law, which predicts how changes in one global property of a gas like temperature will affect the other properties like pressure and volume, was derived first from empirical laws. Later, it was shown that the ideal gas law could be derived from the kinetic theory of a gases, a microscopic modeling approach describing dynamics of individual gas molecules. This confirmation of the kinetic theory provided an intuitive framework that allowed theorists to confidently make many new predictions.

The microscopic approach is embodied in the study of contact processes in graphs. We consider nodes to be the fundamental dynamic entities, and links represent interactions between nodes. We have a stochastic description of the node's dynamics in terms of transition probabilities that result from interactions with other nodes. A contact process is simply a diffusion of activation on a graph, where each activated, or "infected," node can infect its neighbors with some probability given by the $transmissibility$. Given their prevalence, contact processes and the effect of network topology on their dynamics have been widely studied, see, e.g., references in (Ver Steeg, Ghosh, and Lerman 2011).

Theoretical progress in understanding the dynamics of spreading processes on graphs suggests the existence of an *epidemic threshold* (Wang et al. 2003; Chakrabarti et al. 2008; Castellano and Pastor-Satorras 2010) below which no epidemics form and above which epidemics spread to a significant fraction of the graph. At odds with these theoretical results, we have observed information cascades on Digg that spread fast enough for one initial spreader to infect hundreds of people, yet end up affecting only 0.1% of the entire network. We demonstrate with a microscopic modeling approach that two complementary effects conspire to dramatically reduce the size of epidemics. First, because of the highly clustered structure of the Digg network, most people who are aware of a story have been exposed to it via multiple friends. This structure lowers the epidemic threshold while also slowing the overall growth of cascades. We also find that the mechanism for social contagion on Digg deviates from standard social contagion models and this severely curtails the size of social epidemics on Digg. These findings underscore the fundamental difference between information spread and other contagion processes: despite multiple opportunities for infection within a social group, people are less likely to become spreaders of information with repeated exposure. The more clustered a graph is, the more pronounced this effect becomes.

We collected data from the social news aggregator Digg detailing how interest in a story spreads through Digg's social network (Lerman and Ghosh 2010). A user becomes $infected$ by $digging$ (i.e., voting for) a story and exposes her network neighbors to it. Each neighbor may in turn become infected (i.e., vote), exposing her own neighbors to it, and so on. This way interest in a story cascades through Digg's network. This data enables us to trace the flow of information along social links and quantitatively study dynamics of information spread on a network.

We find that the vast majority of cascades grow far slower than expected and fail to reach "epidemic" proportions. To understand why, we simulate information cascades on the Digg graph and on a synthetic graph constructed to have similar properties. We compare results to theoretical predictions and properties of real cascades on Digg. We find that while network structure somewhat limits the growth of cascades, a far more dramatic effect comes from the social contagion mechanism. Unlike the standard cascade models used in previous works on the spread of epidemics, repeated exposure to the same story on Digg does not make the user more likely to vote for it. This effect becomes significant due to the structure of the Digg graph which results in repeated exposure for most users. We define an alternate cascade model that fits empirical observations and show that in simulation it reproduces the observed properties of real information cascades on Digg.

Information cascades on Digg

In this section we primarily focus on the spread of stories through directed friend network of Digg. With each vote, a story becomes visible to the voter's fans. In the event that a user has n friends who have voted for a story, the story appears in their interface along with how many friends had voted on it.

We used Digg API to collect data[1] about 3,553 stories promoted to the front page in June 2009. The data associated with each story contains story title, story id, link, submitter's name, submission time, list of voters and the time of each vote, the time the story was promoted to the front page. In addition, we collected the list of voters' friends. We define an *active user* as any user who voted for at least one story on Digg during the data collection period. There were 139,409 active users, of which 71,367 designated at least one other user as a friend. We extracted the friends of these users and reconstructed the fan network of active users, i.e., a directed graph of active users who are watching activities of other users. There were 279,634 nodes in the fan network, with 1,731,658 links.

As interest in a story spreads, it may generate many cascades from independent seeds. For each story, using the methodology proposed in (Ghosh and Lerman 2011), we extracted the cascade that starts with the submitter and includes all voters who are connected to the submitter either directly or indirectly via the fans network. We call this the *principal cascade* of the story. The distribution of principal cascade size is well described by a log-normal function with the mean of 156. Most of the cascades are smaller than 500, and only three are bigger than 1,000.

[1]The data set is available at http://www.isi.edu/~lerman/downloads/digg2009.html

What Limits Cascades on Digg?

These observations present a puzzle: why are information cascades on Digg so small? In our sample, only one cascade, about Michael Jackson's death, can be said to have reached epidemic proportions, reaching about 5% of active Digg users. The majority of the cascades for the remaining stories reached fewer than 1% of active users. This observation becomes more striking in Fig. 3 which shows that typical epidemic models predict that stories will reach an order of magnitude more voters than we observe on Digg (Moreno, Pastor-Satorras, and Vespignani 2002).

There are a number of factors that could explain why information cascades on Digg are so small. Perhaps Digg users modulate transmissibility of stories and keep them small to prevent information overload. On the other hand, transmissibility could diminish in time, either because of novelty decay (Wu and Huberman 2007) or decrease in visibility of stories as new stories are submitted to Digg (Hogg and Lerman 2009). Perhaps the structure of the network (e.g., clustering or communities) limits the spread of information. Or it could be that the mechanism of social contagion, i.e., how people decide to vote for a story once their friends vote for it, prevents stories from growing on Digg. In addition, users are active at different times, and heterogeneity of their activity could be another explanation.

We examined some of these alternate hypotheses through simulations of contact processes on networks and empirical study of real cascades on Digg. Our basic approach is to compare the predictions of a microscopic model on aggregate quantities like cascade size. Ultimately, we were able to identify a critical combination of factors that allow us to closely reproduce the observed behavior on Digg. As we point out in the introduction, simple mechanisms that are capable of describing aggregate behavior also suggest casual implications that feature-based machine learning approaches miss.

In particular, we found two complementary effects that severely limit the size of cascades on Digg. First, due to the highly clustered structure of the Digg graph, more than half of people exposed to a story on Digg by their friends are exposed by multiple friends. Therefore, it becomes important to understand how people decide to vote on a story based on the number of recommendations. Second, we observe that, contrary to many contagion models, repeated exposure to a story does not make a user more likely to vote on it.

Therefore, we proposed a simple mechanism for story spread, the friend saturation model (FSM). We say the transmissibility of a story is the probability to vote given that some of your friends have voted. Based on our observations, and in contrast to standard epidemic models like the independent cascade model, we say that this probability is the same as long as at least one friend has voted for a story. More sophisticated mechanisms could certainly be proposed, but simulations of cascade processes obeying this simple mechanism, coupled with the structure of the Digg graph, suffice to closely reproduce the observed dynamics of cascade size. In Fig. 3, we see that the predictions of standard epidemiological models are orders of magnitude too big. Our model, on the other hand exactly reproduces the vast reduction in cascade size observed on Digg.

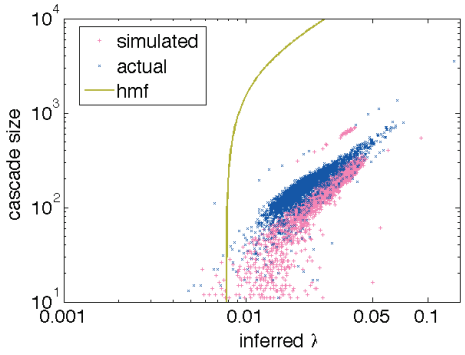

Figure 3: Cascade size vs inferred transmissibility for simulated and real cascades on the Digg graph, plotted on a log-log scale to highlight the order of magnitude difference between these cascade sizes and predictions of a standard epidemic model (HMF, or heterogeneous mean field theory (Moreno, Pastor-Satorras, and Vespignani 2002)).

Discussion

If one assumes Digg's graph structure consists of dense clusters, the effects on cascades in the independent cascade model are quite intuitive. It is easier for a story to take off within a smaller, more tightly connected community. This also explains why the majority of people exposed to story are exposed to it from multiple sources. On the other hand, for cascades to grow very large it is better to have a more homogeneous link structure to reach all parts of the graph quickly. Ultimately, clusters have the effect of marginally decreasing the size of cascades by sequestering an infection in one part of the graph.

In epidemic models, population models and other branching processes, the principal quantity of interest is the reproductive number, R_0. Intuitively, the reproductive number is just the average number of people infected by a single infected person. If $R_0 > 1$, each infection leads to another indefinitely, an epidemic. Whereas, if $R_0 < 1$, the infection will die out eventually. Naively, the reproductive number should just be the average fanout, i.e., the average number of fans, times the transmissibility. For Digg, we have $\langle k \rangle \approx 6$ so $R_0 \approx 6\lambda$. In that case, an epidemic threshold at $R_0 = 1 \rightarrow \lambda_c \approx 1/6$, much higher than we observe. It is well known, however, that heterogeneous degree distributions lower the threshold compared to this prediction(Barrat, Barthélemy, and Vespignani 2008). If we take R_0 to be a constant larger than 1, then it is easy to see that the epidemic must eventually spread to the whole network.

However, we can gain some intuition from this quantity if we view it as a dynamic quantity. FSM implies that the true fanout only includes the number of new fans (those that have not already been exposed to a story) and changes with time. With this definition the fanout is steadily decaying, both for actual and simulated cascades on the Digg graph. Effectively, this leads to a decrease in the reproductive number as well, so that a cascade that initially starts above the

epidemic threshold may fall below it with time.

We have demonstrated a simple behavior model that strikingly reproduces the behavior of Digg cascades, while standard methods go awry. Many network studies assume that graphs with locally tree-like behavior give a good approximation to real networks. In this case, we find that such methods wildly overestimate the size of cascades. If most of the people exposed to a story are exposed repeatedly, understanding how they are affected by repeat exposures is of paramount importance. On Digg, subsequent exposures to a story have almost no effect on the probability of voting. Much remains to be studied: whether these results hold on other social networks, more sophisticated models of response to friends, the time dependence of transmissibility, and more detailed analysis of the effect of graph structure on cascades. However, a simple causal mechanism that reproduces the system's aggregate behavior is a valuable tool for understanding dynamics.

4 Inferring Latent Structures in Networks

Turning from statistical models that predict user behavior, this section employs statistical mechanics to analyze the significance of our inferences about network structure. Most real–world networks are composed of *clusters* of well connected nodes, with relatively lower density of links across different clusters (Newman 2006). As we mentioned in Section 1, the problem of community detection has been studied actively in recent years (Fortunato 2010). In addition to algorithmic development, recent research has focused on characterizing statistical significance of clusters detected by different methods (Fortunato and Barthelemy 2007; Karrer, Levina, and Newman 2008; Lancichinetti, Radicchi, and Ramasco 2010). A related issue is the *feasibility* of detecting clusters, assuming they are present in the network. To be specific, consider the so called *planted bi–partition* model (Condon and Karp 2001), which is a special case of more general family of generative models known as *stochastic block-models* (Holland, Laskey, and Leinhardt 1983; Nowicki and Snijders 2001). In this model the nodes are partitioned into two equal–sized groups, and each link within a group and between the groups is present with probabilities p and r, respectively, so that $p > r$ corresponds to denser connections within each group. An important question is how well one can recover the latent cluster structure in the limit of large network sizes. It is known that in dense networks where the average connectivity scales linearly with the number of nodes N (e.g., p and r are constants), the clusters in the planted partition model can be recovered with asymptotically *perfect* accuracy for any $p - r > N^{-1/2+\epsilon}$ (Condon and Karp 2001). Recently, a more general result obtained for a larger class of stochastic block–models states that certain statistical inference methods are asymptotically consistent provided that the average connectivity grows faster than $\log N$ (Bickel and Chen 2009).

The situation is significantly different for sparse graphs, where the average connectivity remains finite in the asymptotic limit $N \to \infty$. Recently it was shown (Reichardt and Leone 2008) that planted partition models (of arbitrary topology) with finite connectivities are characterized by a phase transition from *detectable* to *undetectable* regimes as one increases the overlap between the clusters, with the transition point depending on the actual degree distribution of the partitions. In particular, let $p = \alpha/N$, $r = \gamma/N$, where α and γ are finite average connectivities *within* and *across* the clusters, and let $p_{in} = \alpha/(\alpha+\gamma)$ be the fraction of links that fall within the clusters, so that $p_{in} = 1$ and $p_{in} = \frac{1}{2}$ correspond to perfectly separated and perfectly mixed clusters, respectively. Then there is a critical value $p_{in}^c = \frac{1}{2} + \Delta$, $\Delta > 0$ such that for $p_{in} < p_{in}^c$ the clusters cannot be recovered with *better than random accuracy* in the asymptotic limit. When the planted clusters have Erdos–Renyi topology, one can show that $\Delta \propto 1/\sqrt{\alpha+\gamma}$ for large $(\alpha+\gamma)$ (Allahverdyan, Ver Steeg, and Galstyan 2010).

From the perspective of statistical inference, this type of phase transition between detectable and undetectable regimes is undesirable, as it signals inference instabilities – a small change in parameters causes a large change in accuracy. In (Allahverdyan, Ver Steeg, and Galstyan 2010) it was shown that this instability can be avoided if one uses prior knowledge about the underlying group structure. Namely, it was demonstrated that knowing the correct cluster assignments for arbitrarily small but *finite* fraction of nodes destroys the criticality and moves the detection threshold to its intuitive (dense–network limit) value $p_{in} = \frac{1}{2}$, or $\alpha = \gamma$. This can be viewed as a *semi–supervised* version of the problem, as opposed to an unsupervised version where the only available information is the observed graph structure.

In practice, semi–supervised graph clustering methods can utilize two types of background knowledge – cluster assignment for a subset of nodes (Zhu and Goldberg 2009; Getz, Shental, and Domany 2005), or pair–wise constraints in the form of *must-link* (*cannot-links*), which imply that a pair of nodes must be (cannot be) assigned to the same group (Basu, Bilenko, and Mooney 2004; Kulis et al. 2009). In fact, the latter type of constraints are more realistic in scenarios where it is easier to assess similarity of two objects rather than to label them individually. Below we examine the impact of semi-supervision on clustering accuracy using a statistical mechanics framework. We focus on a random network composed of two equal–sized clusters, where the clustering objective can mapped to an appropriately defined Ising model defined on the planted partition graph.

Community Detection

Let us consider a graph with two clusters containing N–nodes each. Each pair of nodes within the same cluster is linked with probability $p = \alpha/N$, where α is the average within–cluster connectivity. Also, each pair of nodes in different clusters is linked with probability $r = \gamma/N$, where γ is inter–cluster connectivity. Let a *spin* variable $s_i = \pm 1$ denote the cluster assignment of node i, and let $\mathbf{s} = (s_1, \ldots, s_{2N})$ denote a cluster assignment for all the nodes in the network. Further, let $\mathbf{A}$ be the observed adjacency matrix of interaction graph of $2N$ nodes so that $A_{ij} = 1$ if we have observed a link between nodes i and j, and $A_{ij} = 0$ otherwise.

Using $\delta_{s_i,s_j} \equiv [1 + s_i s_j]/2$, we define a Hamiltonian.

$$H(\mathbf{s}, \mathbf{A}) = -\frac{1}{2} \sum_{i<j}^{2N} A_{ij} s_i s_j + H_\pi(\mathbf{s}) \quad (2)$$

The statistical physics meaning of the first term in Eq. 2 is the following: If two spins s_i and s_j are linked ($A_{ij} = 1$), then they tend to align together.

Furthermore, we assume that the background information is encoded via a matrix Θ with elements θ_{ij}, so that $\theta_{ij} = A_{ij}$ and $\theta_{ij} = -A_{ij}$ correspond to the presence of a must link and cannot link ($\theta_{ij} = 0$ means no constraint at all)[2]. For the sake of simplicity, we will assume that violating either type of constraint carries the same cost w, yielding

$$H_\pi(\mathbf{s}) = -\frac{w}{2} \sum_{i<j}^{2N} \theta_{ij} s_i s_j \ . \quad (3)$$

Equation 2, 3 define a Hamiltonian of an Ising model, which is a well-studied example of a Markov Random Field (MRF). Minimizing the Hamiltonian then corresponds to finding the Maximum a Posteriori (MAP) estimation of the MRF(Allahverdyan, Ver Steeg, and Galstyan 2010). This problem is known to be computationally hard for general graphs, and one usually has to resort to approximate techniques, such as linear relaxation, message passing, and so on (e.g., see (Wainwright and Jordan 2008)).

Instead of focusing on MAP assignment of a particular instance of a MRF, our goal is to understand average (typical) behavior and characteristics of MAP estimation for an *ensemble* of such problems drawn from some distribution. In this approach, the parameters of the problem are treated as random (quenched) variables. While we consider statistical properties of random instances of graphs, the overall ensemble of graphs is defined according to a tunable cluster structure. In our case, the quenched variables are A_{ij} and θ_{ij}. The statistics of A_{ij} is defined by the stochastic block-structure. We also assume that the constraints are imposed on each link randomly and independently, with probability f_m and f_c for must-link and cannot-link constraints, respectively.

We are interested in the properties of the above Hamiltonian in the limit of large N. In particular, we would like to know whether the configuration that minimizes Eq. 2 carries information about the underlying cluster structure. This problem can be studied within the so called zero temperature cavity method (Mezard and Parisi 1987; 2001) which is related to the max–product message passing algorithm. In this mean field approximation, the solution is characterized via a cavity field distribution in each clusters $P(h)$ and $\bar{P}(h)$, that denotes the probability of an internal (*cavity*) field acting on an randomly chosen spin s in the respective cluster. This distribution is found from the cavity equation, which is like the fixed point of a belief propagation algorithm.

The details of the approach are irrelevant for the purposes of this paper. We simply note that once $P(h)$ is found we can obtain the so called magnetization of the spins in one of the clusters (say, the first one),

$$m = \int P(h) \operatorname{sign}(h), \quad (4)$$

m is the *average* value of a randomly selected *spin*, where the average is taken over the graph structure and the constraints (i.e., A_{ij}, θ_{ij}), and over all configurations of s_i that in the thermodynamic limit $N \to \infty$ have— the same (minimal) values of the Hamiltonian H. Note that the accuracy of the clustering (i.e., probability that a node has been assigned to the correct cluster) is simply $\frac{1+|m|}{2}$. Thus, $|m| = 1$ corresponds to perfect clustering, whereas $m = 0$ means that discovered clusters have only random overlap with the true cluster assignments.

Impact of pairwise constraint

We first consider the case when violating a constraint carries a finite cost. The most trivial such case is when $w = 1$. In this case, the must–link constraints do not yield any additional information. The cannot-link constraints, however, help clustering by "flipping" the sign of the corresponding edge, thus favoring anti–ferromagnetic interactions between the nodes across different clusters. In fact, it can be shown that the only impact of the constraints with $w = 1$ is to renormalize within and across cluster connectivities, $(\alpha, \gamma) \to (\alpha + \rho\gamma, \gamma - \rho\gamma)$. Recall that the *mixing* parameter is defined as $p_{in} = \frac{\alpha}{\alpha+\gamma}$. Thus, this situation is identical to the un–supervised clustering scenario (Reichardt and Leone 2008; Allahverdyan, Ver Steeg, and Galstyan 2010) with renormalized mixing parameter $p_{in} \to p_{in} + \rho(1 - p_{in})/(\alpha + \gamma)$. The sole impact of constraints is to shift the detection below which clusters cannot be detected with better than random accuracy. In particular, the modified threshold coincides with its dense network limit $\frac{1}{2}$ for $\rho = \frac{\alpha+\gamma}{2} \frac{1-2p_{in}}{1-p_{in}}$.

Now let us focus on the case where violating a constraints incurs an infinite cost, $w = \infty$. In this case, the cavity equation cannot be solved analytically. Instead, we address this case by solving the cavity equation using population dynamics. The goal of population dynamics is to estimate the distribution of fields $P(h)$ whose fixed point is the cavity equation (see (Mezard and Parisi 2001) for a detailed description). We also compare our results to simulations using synthetic data. After generating random graphs of size $N = 100,000$, we find the ground state of the Hamiltonian 2 using simulated annealing.

Looking at Fig. 4, we see that for small amounts of supervision, $\rho < 1$, the impact of the constraints is to shift the detection threshold to smaller values of p_{in}. This behavior is expected, since adding hard constraints is equivalent to studying the same unsupervised clustering problem on a *renormalized* graph (e.g., merging two nodes that are connected via constraints). This is in contrast to results for prior information on nodes in (Allahverdyan, Ver Steeg, and Galstyan 2010), which showed that even small amounts of node supervision shifted the detection threshold to its lowest possible value $p_{in} = 1/2$.

As $\rho \to 1$, there is a qualitative change in our ability to

[2]This notation assumes that the constraints are available only for the nodes that are already connected in the graph. Relaxing this assumption simply re-normalizes the graph parameters

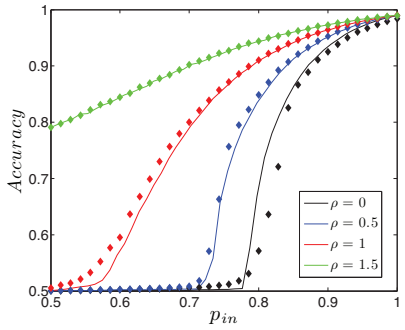

Figure 4: Magnetization plotted against the mixing parameter p_{in} for $\alpha + \gamma = 4$ and different ρ. Lines are generated from population dynamics and points are generated from simulated annealing. From bottom to top we have $\rho = 0, 0.5, 1, 1.5$

detect clusters. A large number of nodes, $O(N^{2/3})$, are connected by labeled edges. If we take the relative labeling of nodes in this largest group as the "correct" one, than we have a situation similar to node supervision, which, as discussed, moves the detection threshold to $\alpha = \gamma$. While this large labeled component suffices to create non-zero magnetization in finite graphs (as seen from the simulated annealing results), as N gets large, the effect of this component diminishes. For $\rho > 1$, we see that the fraction of nodes contained in the largest labeled component suffice to produce non-zero magnetization even at the group-defining threshold $\alpha = \gamma$.

Discussion

In this section we have presented a statistical mechanics analysis of semi-supervised clustering in sparse graphs in the presence of pair–wise constraints on node cluster assignments. Our results show that addition of constraints does not provide automatic improvement over the unsupervised case. This is in sharp contrast with the semi–supervised clustering scenario considered in (Allahverdyan, Ver Steeg, and Galstyan 2010), where any generic fraction of labeled nodes improves clustering accuracy.

When the cost of violating constraints is finite, the only impact of adding pair–wise constraints is *lowering* the detection boundary. Thus, whether adding constraints is beneficial depends on the network parameters. For semi–supervised clustering with hard pair–wise constraints, the situation is similar if the number of added constraints is small. For small density of constraints the subgraph induced by the must– and cannot links consists mostly of isolated small components, and the only impact of the added constraints is to lower the detection boundary. The situation changes drastically when the constraint density reaches the percolation threshold. Due to transitivity of constraints, this induces a non–vanishing subset of nodes (transitive closure) that belong to the same cluster, a scenario that is similar to one studied in (Allahverdyan, Ver Steeg, and Galstyan 2010). In this case, the detection boundary disappears for any α, γ.

In the study presented here, we assume that the edges are labeled randomly. One can ask whether other, non–random edge–selection strategies will lead to better results. Intuition tells us that the answer is affirmative. Indeed, in the random case one needs to add $\rho = 1$ additional edges per node in order to have the benefit of transitivity. For a given ρ, a much better strategy would be to choose $\rho N + 1$ random nodes (rather than edges), and connect them into a chain using labeled edges. This would guarantee the existence of a finite fraction of labeled nodes for any ρ.

Finally, it is possible to envision a situation where one has access to two types of information – about cluster assignment of specific nodes and pairwise constraints. Furthermore, this information might be available at a cost that, generally speaking, will be different for either type of information. An interesting problem then is to find an optimal allocation of a limited budget to achieve the best possible clustering accuracy.

5 Conclusion

Social media has transformed the Web into a participatory medium and potentially a powerful new computational platform. As people interact online, their collective activity and the structure of the Web itself are becoming increasingly more complex and dynamic. Complex feedback between individual decisions and collective actions often leads to qualitatively new behaviors. Statistical physics provides a framework to model emergent behaviors in social media. This framework represents individual dynamic entities as stochastic processes and allows the modeler to relate aggregate behaviors to these descriptions. We presented several examples where we used this framework to understand the underlying statistical laws of information diffusion and the structure of social networks.

References

Allahverdyan, A. E.; Ver Steeg, G.; and Galstyan, A. 2010. Community detection with and without prior information. *EPL (Europhysics Letters)* 90(1):18002.

Arenas, A.; Díaz-Guilera, A.; and Pérez-Vicente, C. J. 2006. Synchronization reveals topological scales in complex networks. *Phys. Rev. Lett.* 96(11):114102.

Bakshy, E.; Hofman, J. M.; Mason, W. A.; and Watts, D. J. 2011. Everyone's an influencer: quantifying influence on twitter. In *Proc. fourth ACM international conference on Web search and data mining*, WSDM '11, 65–74. New York, NY, USA: ACM.

Barrat, A.; Barthélemy, M.; and Vespignani, A. 2008. *Dynamical Processes on Complex Networks*. Cambridge, England: Cambridge University Press, 1st edition.

Basu, S.; Bilenko, M.; and Mooney, R. J. 2004. A probabilistic framework for semi-supervised clustering. In *Proc. of the 10th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, 59–68. New York, NY, USA: ACM.

Bickel, P. J., and Chen, A. 2009. A nonparametric view of network models and newmangirvan and other modularities. *Proceedings of the National Academy of Sciences*.

Castellano, C., and Pastor-Satorras, R. 2010. Thresholds for epidemic spreading in networks. *Physical Review Letters* 105:218701+.

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

Cha, M.; Haddadiy, H.; Benevenutoz, F.; and Gummadi, K. P. 2010. Measuring user influence in twitter: The million follower fallacy. In *Proc. 4th Int. Conf. on Weblogs and Social Media (ICWSM)*.

Chakrabarti, D.; Wang, Y.; Wang, C.; Leskovec, J.; and Faloutsos, C. 2008. Epidemic Thresholds in Real Networks. *ACM Transactions on Information and System Security,* 10(4):13+.

Condon, A., and Karp, R. M. 2001. Algorithms for graph partitioning on the planted partition model. *Random Struct. Algorithms* 18:116–140.

Dorogovtsev, S. N.; Mendes, J. F. F.; Samukhin, A. N.; and Zyuzin, A. Y. 2008. Organization of modular networks. *Phys. Rev. E* 78(5):056106.

Ellner, S. P., and Guckenheimer, J. 2006. *Dynamic Models in Biology*. Princeton, NJ: Princeton Univ. Press.

Fortunato, S., and Barthelemy, M. 2007. Resolution limit in community detection. *PNAS* 104(1):36–41.

Fortunato, S. 2010. Community detection in graphs. *Physics Reports* 486:75–174.

Galstyan, A., and Cohen, P. 2007. Cascading dynamics in modular networks. *Phys. Rev. E* 75(3):036109.

Getz, G.; Shental, N.; and Domany, E. 2005. Semi-supervised learning a statistical physics approach. In *Proc. of the 22nd ICML Workshop on Learning*.

Ghosh, R., and Lerman, K. 2010. Predicting influential users in online social networks. In *Proc. KDD workshop on Social Network Analysis (SNAKDD)*.

Ghosh, R., and Lerman, K. 2011. A Framework for Quantitative Analysis of Cascades on Networks. In *Proc. Web Search and Data Mining Conf. (WSDM)*.

Ghosh, R.; Lerman, K.; Surachawala, T.; Voevodski, K.; and Teng, S. 2011. Non-conservative diffusion and its application to social network analysis. In *submitted to KDD*. submitted.

Gleeson, J. P. 2008. Cascades on correlated and modular random networks. *Phys. Rev. E* 77(4):046117.

Hogg, T., and Lerman, K. 2009. Stochastic models of user-contributory web sites. In *Proc. Third Int. Conf. on Weblogs and Social Media (ICWSM2009)*, 50–57.

Hogg, T., and Szabo, G. 2009. Diversity of user activity and content quality in online communities. In *Proc. of the Third Int. Conf. on Weblogs and Social Media (ICWSM2009)*, 58–65. AAAI.

Holland, P. W.; Laskey, K. B.; and Leinhardt, S. 1983. Stochastic blockmodels: First steps. *Social Networks* 5(2):109 – 137.

Iribarren, J. L., and Moro, E. 2009. Impact of Human Activity Patterns on the Dynamics of Information Diffusion. *Physical Review Letters* 103(3):038702+.

Kampen, N. G. V. 1992. *Stochastic Processes in Physics and Chemistry*. Amsterdam: Elsevier Science, revised edition.

Karrer, B.; Levina, E.; and Newman, M. E. J. 2008. Robustness of community structure in networks. *Phys. Rev. E* 77(4):046119.

Kulis, B.; Basu, S.; Dhillon, I.; and Mooney, R. 2009. Semi-supervised graph clustering: a kernel approach. *Mach. Learn.* 74:1–22.

Lancichinetti, A.; Radicchi, F.; and Ramasco, J. J. 2010. Statistical significance of communities in networks. *Phys. Rev. E* 81(4):046110.

Lazer, D. e. a. 2009. Computational social science. *Science* 323:721–723.

Lerman, K., and Ghosh, R. 2010. Information contagion: an empirical study of spread of news on digg and twitter social networks. In *Proc. 4th Int. Conf. on Weblogs and Social Media (ICWSM)*.

Lerman, K., and Hogg, T. 2010. Using a model of social dynamics to predict popularity of news. In *Proc. 19th Int. World Wide Web Conf. (WWW)*.

Lerman, K. 2007a. Social information processing in social news aggregation. *IEEE Internet Computing: special issue on Social Search* 11(6):16–28.

Lerman, K. 2007b. Social networks and social information filtering on digg. In *Proc. of Int. Conf. on Weblogs and Social Media (ICWSM-07)*.

Leskovec, J.; Adamic, L.; and Huberman, B. 2007. The dynamics of viral marketing. *ACM Transactions on the Web* 1(1).

Lloyd-Smith, J. O.; Schreiber, S. J.; Kopp, P. E.; and Getz, W. M. 2005. Superspreading and the effect of individual variation on disease emergence. *Nature* 438:355–359.

Mezard, M., and Parisi, G. 1987. Mean-field theory of randomly frustrated systems with finite connectivity. *EPL (Europhysics Letters)* 3(10):1067.

Mezard, M., and Parisi, G. 2001. The bethe lattice spin glass revisited. *The European Physical Journal B* 20:217.

Moreno, Y.; Pastor-Satorras, R.; and Vespignani, A. 2002. Epidemic outbreaks in complex heterogeneous networks. *The European Physical Journal B - Condensed Matter and Complex Systems* 26(4):521–529.

Newman, M. E. J. 2006. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences* 103(23):8577–8582.

Nowicki, K., and Snijders, T. A. B. 2001. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association* 96(455):pp. 1077–1087.

Opper, M., and Saad, D., eds. 2001. *Advanced Mean Field Methods: Theory and Practice*. Cambridge, MA: MIT Press.

Reichardt, J., and Leone, M. 2008. (un)detectable cluster structure in sparse networks. *Phys. Rev. Lett.* 101(7):078701.

Romero, D. M.; Meeder, B.; and Kleinberg, J. 2011. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter. In *Proc. World Wide Web Conf.*

Salganik, M.; Dodds, P.; and Watts, D. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *Science* 311:854.

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

Ver Steeg, G.; Ghosh, R.; and Lerman, K. 2011. What stops social epidemics? In *Proc. 5th Int. Conf. on Weblogs and Social Media*.

Wainwright, M. J., and Jordan, M. I. 2008. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* 1:1–305.

Wang, Y.; Chakrabarti, D.; Wang, C.; and Faloutsos, C. 2003. Epidemic Spreading in Real Networks: An Eigenvalue Viewpoint. *Reliable Distributed Systems, IEEE Symposium on* 0:25+.

Wilkinson, D. M. 2008. Strong regularities in online peer production. In *EC '08: Proc. 9th ACM conference on Electronic commerce*, 302–309. New York, NY, USA: ACM.

Wu, F., and Huberman, B. A. 2007. Novelty and collective attention. *Proc. National Academy of Sciences* 104(45):17599–17601.

Zhu, X., and Goldberg, A. B. 2009. Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 3(1):1–130.