Diffusion and Cascading Behavior in Networks

Jure Leskovec [a]
[a] *Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, USA*

Abstract. Information cascades are phenomena in which individuals adopt a new action or idea due to influence by others. As such a process spreads through an underlying social network, it can result in widespread adoption overall. Here we consider information cascades in the context of recommendations and information propagation on the blogosphere. In particular, we study the patterns of cascading recommendations that arise in large social networks. We review recent studies of cascading behavior in product recommendation networks, and information diffusion on the blogosphere. Next, we examine theoretical models of models of information, virus and influence propagation. Last, we present developments on selecting and targeting nodes in networks to maximize the influence or detect cascades and disease/information outbreaks effectively.

Introduction

Diffusion is a process by which information, viruses, ideas and new behavior spread over the network. For example, adoption of a new technology begins on a small scale with a few "early adopters", then more and more people adopt it as they observe friends and neighbors using it. Eventually the adoption of the technology may spread through the social network as an epidemic "infecting" most of the network. As it spreads over the network it creates a cascade. Cascades have been studied for many years by sociologists concerned with the *diffusion of innovation* [31]; more recently, researchers have investigated cascades for selecting trendsetters for viral marketing, finding inoculation targets in epidemiology, and explaining trends in blogspace.

There are three aspects of studies on diffusion and cascading behavior in networks: (a) mathematical models of information, virus and influence propagation, (b) empirical studies of diffusion in social and information networks, and (c) algorithms for detecting cascades and selecting influential nodes.

(a) Mathematical models. Most of the research on the flow of information and influence through the networks has been done in the context of epidemiology and the spread of diseases over the network [4].

Classical disease propagation models are based on the stages of a disease in a host: a person is first *susceptible* to a disease, then if she is exposed to an infectious contact she can become *infected* and thus *infectious*. After the disease ceases the person is *recovered*. Person is then *immune* for some period. The immunity can also wear off and the

person becomes again susceptible. Thus SIR (susceptible – infected – recovered) models diseases where a recovered person never again becomes susceptible, while SIRS (SIS, susceptible – infected – (recovered) – susceptible) models population in which recovered host can become susceptible again. Given a network and a set of infected nodes the *epidemic threshold* is studied, *i.e.*, condition under which the disease will either dominate or die out. Interestingly, the largest eigenvalue of a graph adjacency matrix plays a fundamental role in deciding whether the disease will take over the network. Related are the diffusion models that try to model the process of adoption of an idea or a product. They can generally be divided into two groups:

Threshold model: [11] A node adopts the behavior (*e.g.*, purchases a product) if a sum of the connection weights of its neighbors that already adopted the behavior is greater than the threshold.

Independent cascade model [15] where whenever a neighbor v of node u adopts, then node u also adopts with probability $p_{u,v}$, *i.e.*, every time a neighbor of u purchases a product, there is a chance that u will decide to purchase as well.

(b) Empirical studies of cascading behavior. While the above models address the question of how processes spread in a network, they are based on *assumed* rather than *measured* influence effects.

Most work on measuring cascading behavior has been done in the blog domain. Blog posts refer to each other using hyper-links. Since posts are time-stamped, we can trace their linking patterns all the way to the source, and so identify the flow of information from the source post to the followers and followers of the followers [22]. Similarly, viral marketing can be thought of as a diffusion of information about the product and its adoption over the network [20]. Here the cascades are formed by people recommending products to each other and so the product recommendations (and purchases) spread over the network.

In our work [20,22] we observed rich cascading behavior on the blogosphere and in the viral marketing and investigated several interesting questions: What kinds of cascades arise frequently in real life? Are they like trees, stars, or something else? And how do they reflect properties of their underlying network environment? Do certain nodes have specific propagation patterns?

(c) Detecting cascades and finding influential nodes. Exploiting cascades could lead to important insights. For example, in viral marketing where a company wants to use word-of-mouth effects to market a product, exploiting the fact that early adopters may convince their friends to buy the product is crucial. So, the company wants to identify the most important nodes to target to spread the information about the product over the network [15]? A similar problem is of detecting outbreaks in networks [21], where we are given a network and a dynamic process spreading over it, and we want to select a set of nodes to detect the process as effectively as possible. For example, consider a city water distribution network, delivering water to households via pipes and junctions. Contaminants may spread over the network, and so we want to select a few locations (pipe junctions) to install sensors to effectively detect the contaminations.

One can formulate above tasks as optimization over sets of nodes, which turns out to be hard computational problem. However, it turns out that influence functions exhibit a diminishing returns property called *submodularity*. Exploiting submodularity we design near-optimal algorithms [21,15] for finding influential nodes and effectively detecting outbreaks in networks.

Cascades in networks

Information cascades are phenomena in which an action or idea becomes widely adopted due to influence by others [5]. Cascades are also known as "fads" or "resonance." Cascades have been studied for many years by sociologists concerned with the *diffusion of innovation* [31]; more recently, researchers in several fields have investigated cascades for the purpose of selecting trendsetters for viral marketing [8], finding inoculation targets in epidemiology [26], and explaining trends in blogspace [17]. Despite much empirical work in the social sciences on datasets of moderate size, the difficulty in obtaining data has limited the extent of analysis on very large-scale, complete datasets representing cascades. Later, we look at the patterns of influence in a large-scale, real recommendation network and examine the topological structure of cascades.

Most of the previous research on the flow of information and influence through the networks has been done in the context of epidemiology and the spread of diseases over the network [4,3]. Classical disease propagation models are based on the stages of a disease in a host: a person is first *susceptible* to a disease, then if she is exposed to an infectious contact she can become *infected* and thus *infectious*. After the disease ceases the person is *recovered* or *removed*. Person is then *immune* for some period. The immunity can also wear off and the person becomes again susceptible. Thus SIR (susceptible – infected – recovered) models diseases where a recovered person never again becomes susceptible, while SIRS (SIS, susceptible – infected – (recovered) – susceptible) models population in which recovered host can become susceptible again. Given a network and a set of infected nodes the *epidemic threshold* is studied, *i.e.*, conditions under which the disease will either dominate or die out.

Diffusion models that try to model the process of adoption of an idea or a product can generally be divided into two groups:

- *Threshold model* [11] where each node in the network has a threshold $t \in [0, 1]$, typically drawn from some probability distribution. We also assign *connection weights* $w_{u,v}$ on the edges of the network. A node adopts the behavior if a sum of the connection weights of its neighbors that already adopted the behavior (purchased a product in our case) is greater than the threshold: $t \leq \sum_{\text{adopters}(u)} w_{u,v}$.
- *Independent cascade model* [10] where whenever a neighbor v of node u adopts, then node u also adopts with probability $p_{u,v}$. In other words, every time a neighbor of u purchases a product, there is a chance that u will decide to purchase as well.

While these models address the question of how influence spreads in a network, they are based on *assumed* rather than *measured* influence effects. In contrast, our study tracks the actual diffusion of recommendations through email, allowing us to quantify the importance of factors such as the presence of highly connected individuals, or the effect of receiving recommendations from multiple contacts. Compared to previous empirical studies which tracked the adoption of a single innovation or product, our data encompasses over half a million different products, allowing us to model a product's suitability for viral marketing in terms of both the properties of the network and the product itself.

Information cascades in blogosphere

Most work on extracting cascades has been done in the blog domain [1,2,13]. The authors in this domain noted that, while information propagates between blogs, examples of genuine cascading behavior appeared relatively rarely. This is possibly due to bias in the web-crawling and text analysis techniques used to collect pages and infer relationships. In our dataset, all the recommendations are stored as database transactions, and we know that no records are missing. Associated with each recommendation is the product involved, and the time the recommendation was made. Studies of blogspace either spend a lot of effort mining topics from posts [2,13] or consider only the properties of blogspace as a graph of unlabeled URLs [1].

There are several potential models to capture the structure of the blogosphere. Work on information diffusion based on topics [13] showed that for some topics, their popularity remains constant in time ("chatter") while for other topics the popularity is more volatile ("spikes"). [17] analyze community-level behavior as inferred from blog-rolls – permanent links between "friend" blogs. In their extension [18] performed analysis of several topological properties of link graphs in communities, finding that much behavior was characterized by "stars".

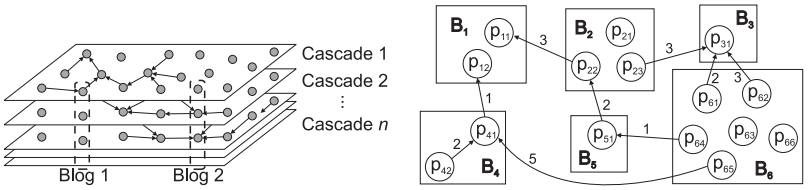
Cascades in viral marketing

Viral marketing can be thought of as a diffusion of information about the product and its adoption over the network. Primarily in social sciences there is a long history of research on the influence of social networks on innovation and product diffusion. However, such studies have been typically limited to small networks and typically a single product or service. For example, [6] interviewed the families of students being instructed by three piano teachers, in order to find out the network of referrals. They found that strong ties, those between family or friends, were more likely to be activated for information flow and were also more influential than weak ties [12] between acquaintances.

In the context of the internet, word-of-mouth advertising is not restricted to pairwise or small-group interactions between individuals. Rather, customers can share their experiences and opinions regarding a product with everyone. Quantitative marketing techniques have been proposed [24] to describe product information flow online, and the rating of products and merchants has been shown to effect the likelihood of an item being bought [29,7]. More sophisticated online recommendation systems allow users to rate others' reviews, or directly rate other reviewers to implicitly form a trusted reviewer network that may have very little overlap with a person's actual social circle. [30] used Epinions' trusted reviewer network to construct an algorithm to maximize viral marketing efficiency assuming that individuals' probability of purchasing a product depends on the opinions on the trusted peers in their network. [15] have followed up on the challenge of maximizing viral information spread by evaluating several algorithms given various models of adoption we discuss next.

Empirical observations of cascading behavior

We formally define a cascade as a graph where the nodes are agents and a directed edge (i, j, t) indicates that a node i influenced a node j at time t.



(a) Cascades as layers (b) Cascades as graphs

Figure 1. Two views on the formation of information cascades on the blogosphere.

Consider three examples of cascade formation and propagation in networks:

- First, we present results on cascades in a large viral marketing network, where people recommend products to each other and we study the spread and success of recommendations over the network.
- Second, we consider the tracking of a large population of blogs over a long period of time and observe the propagation of information between the blogs.
- Third, we present the propagation of infectious water in large real water distribution networks, and ask the question of where to place a limited number of sensors so the disease outbreaks will be detected early.

Blogs (weblogs) are web sites that are updated on a regular basis. Often times individuals use them for online diaries and social networking; other times news sites have blogs for timely stories. Blogs are composed of time-stamped posts, and posts typically link each other, as well as other resources on the Web.

For example, figure 1 shows two alternative views of information cascades that may occur on the blogosphere. In figure 1(a) each circle represents a blog post, and all circles at the same vertical position belong to the same blog. Often blog posts refer to each other using hyper-links. Given that the posts are time-stamped and usually not updated, we can trace their linking patterns all the way to the source. It is easy to identify the flow if information from the source post to the followers and followers of the followers. So, each layer represents a different information cascade (information propagation graph). Figure 1(b) gives an alternative view. Here posts (represented as circles) inside a rectangle belong to the same blog. Similarly, the information cascades correspond to connected components of the posts in the graph, *e.g.* posts p_{12}, p_{41}, p_{42} and p_{65} all form a cascade, where p_{12} is the *cascade initiator*.

Observing such behavior on the blogosphere or in the viral marketing poses several interesting questions: What kinds of cascades arise frequently in real life? Are they like trees, stars, or something else? And how do they reflect properties of their underlying network environment? How fast does the information spread? Do certain nodes have specific propagation patterns? What are the most important nodes to target if we want to spread the information over the network?

In addition to observing rich cascades and propagation [23] one can make a step further and analyze the effectiveness and dynamics of product recommendations in causing purchases [19,20]. To our knowledge this was the first study to directly observe the effectiveness of person to person word of mouth advertising for hundreds of thousands of products. Similarly, for blogs [22] is the first to perform a large study of cascading behavior in large blog networks.

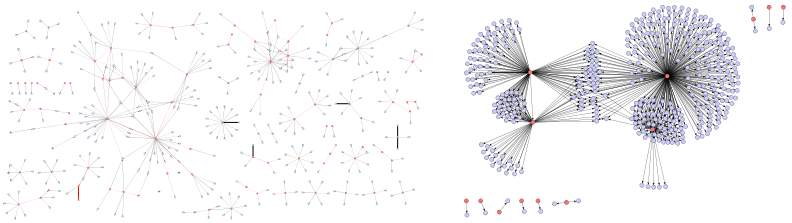


Figure 2. Examples of two product recommendation networks. Left: First aid study guide. Notice many small disconnected cascades. Right: Japanese graphic novel (manga). Notice a large, tight community.

Cascades in viral marketing

A recent study [19] examined a recommendation network consisting of 4 million people who made 16 million recommendations on half a million products from a large on-line retailer. Each time a person purchases a book, music, DVD, or video tape she is given the option to send an email recommending the item to her friends. The first recipient to purchase the item receives a discount and the sender of the recommendation receives a referral credit.

Figure 2 shows two typical product recommendation networks. Most product recommendation networks consist of a large number of small disconnected components where we do not observe cascades. Then there is usually a small number of relatively small components where we observe recommendations propagating. Also notice bursts of recommendations and collisions (figure 2(b)). Some individuals send recommendations to many friends which results in star-like patterns in the graph.

Cascading patterns

Consider the problem of finding patterns of recommendations in a large social network. One can ask the following questions: How does the influence propagate? What does it look like?

In order to analyze the data, new methods and algorithms had to be developed. First, to identify cascades, *i.e.* graphs where incoming recommendations influenced purchases and further recommendations. Next, to enumerate and count the cascade subgraphs. Graph isomorphism and enumeration are both computationally very expensive, so new algorithms for approximate graph isomorphism resolution were developed [23]. In a multi-level approach the computational complexity (and accuracy) of the graph isomorphism resolution depends on the size of the graph. This property makes the algorithm scale nicely to large datasets.

It has been found [**?**] that the distribution of sizes and depths of cascades follows a power law. Generally, cascades tend to be shallow, but occasional large bursts can occur. Cascades are mainly tree-like, but variability in connectivity and branching across different products groups was also observed. Figure 3 shows some typical examples of how the influence propagates over the recommendation network.

In addition to observing rich cascades and propagation one can make a step further and analyze the effectiveness and dynamics of product recommendations in causing purchases.

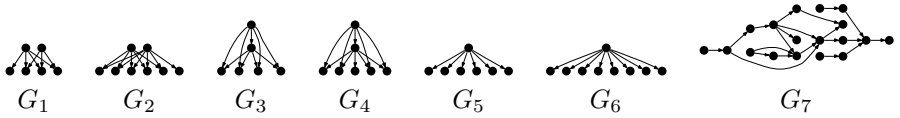
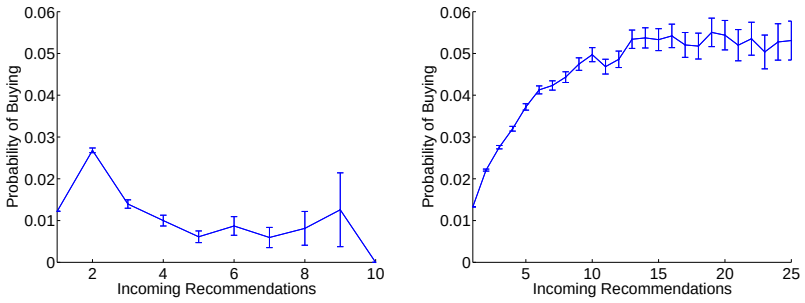


Figure 3. Typical classes of cascades. G_1, G_2: nodes recommending to the same set of people, but not each other. G_3, G_4: nodes recommending to same community. G_5, G_6: a flat cascade. G_7: a large propagation of recommendations.



(a) Book (b) DVD

Figure 4. Probability of purchasing a product given the number of received recommendations. Notice the decrease in purchasing probability for books and saturation for DVDs.

Implications for viral marketing

A study of Leskovec et al. [19] established how the recommendation network grows over time and how effective it is from the viewpoint of the sender and receiver of the recommendations. The examine what kind of product is more likely to be bought as a result of recommendation, and describe the size of the cascade that results from recommendations and purchases. While on average recommendations are not very effective at inducing purchases and do not spread very far, there are product and pricing categories for which viral marketing seems to be very effective.

Figure 4 presents an example of our findings. We plot the probability of purchasing a product given the number of received recommendations. Surprisingly, as more book recommendations are received their success *decreases*. Success of DVD recommendations saturates around 10 incoming recommendations. This means that after a person gets 10 recommendations they become immune to them – their probability of buying does not increase anymore. Traditional innovation diffusion models assume that an increasing number of infected contacts results in an increased likelihood of infection. Instead, it was shown that the probability of purchasing a product increases with the number of recommendations received, but then it quickly saturates. The result has important implications for viral marketing because providing too much incentive for people to recommend to one another can weaken the very social network links that the marketer is intending to exploit.

What determines the product's viral marketing success? A study [20] presents a model which characterizes product categories for which recommendations are more likely to be accepted, and find that the numbers of nodes and receivers have negative

coefficients, showing that successfully recommended products are actually more likely to be not so widely popular. It shows that more expensive and more recommended products have a higher success rate. These recommendations should occur between a small number of senders and receivers, which suggests a very dense recommendation network where lots of recommendations are exchanged between a small community of people. These insights could be of use to marketers — personal recommendations are most effective in small, densely connected communities enjoying expensive products. Refer to [20] for more details.

Cascades on the blogosphere

Similarly to the viral marketing setting we also analyze cascades on the blogosphere. We address a set of related questions: What kinds of cascades arise frequently in real life? Are they like trees, stars, or something else? And how do they reflect properties of their underlying network environment?

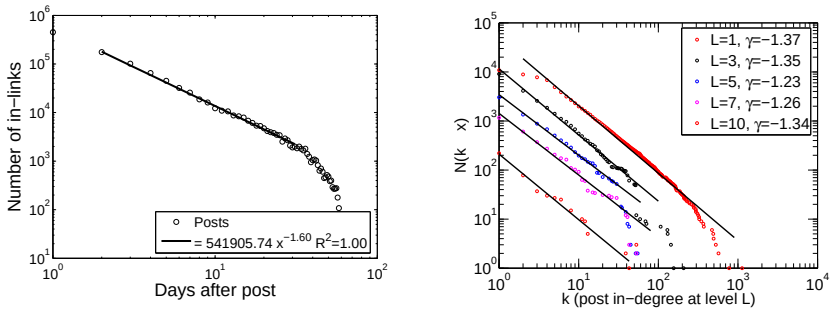
Shape of information cascades

We extracted our dataset from a larger set of blogs and posts from August and September 2005 [9]. We were interested in blogs and posts that actively participate in discussions, so we biased our dataset towards the more active part of the blogosphere. We focused on the most-cited blogs and traced forward and backward conversation trees containing these blogs. This process produced a dataset of 2.5 million posts from $45,000$ blogs gathered over the three-month period. To analyze the data, we first create graphs of time-obeying propagation of links. Then, we enumerate and count all possible cascade subgraphs.

We find novel patterns, and the analysis of the results gives us insight into the cascade formation process. Most surprisingly, the popularity of posts drops with a *power law*, instead of exponentially, that one may have expected. We collect all in-links to a post and plot the number of links occurring after each day following the post. This creates a curve that indicates the rise and fall of popularity. Figure 5(a) shows number of in-links for each day following a post for all posts in the dataset The exponent of the power law is -1.5, which is exactly the value predicted by the model where the bursty nature of human behavior is a consequence of a decision based queuing process [27,32] – when individuals execute tasks based on some perceived priority, the timing of the tasks is heavy tailed, with most tasks being rapidly executed, whereas a few experience very long waiting times.

We also find that probability of observing a cascade on n nodes follows a Zipf distribution: $p(n) \propto n^{-2}$. Figure 5(b) plots the in-degree distribution of nodes at level L of the cascade. A node is at level L if it is L hops away from the root (cascade initiator) node. Notice that the in-degree exponent is stable and does not change much given the level in the cascade. This means that posts still attract attention (get linked) even if they are somewhat late in the cascade and appear towards the bottom of it.

We also found rich cascade patterns. Generally cascades are shallow but occasional large bursts also occur. The cascade sub-patterns shown on figure 6 reveal mostly small tree-like subgraphs; however we observe differences in connectivity, density, and the shape of cascades. Indeed, the frequency of different cascade subgraphs is not a simple consequence of differences in size or density; rather, we find instances where denser



Power-law popularity decay In-degree at level L

Figure 5. Number of in-links vs. the days after the post in log-linear scale, after removing the day-of-the week effects. The power law fit has the exponent -1.5.

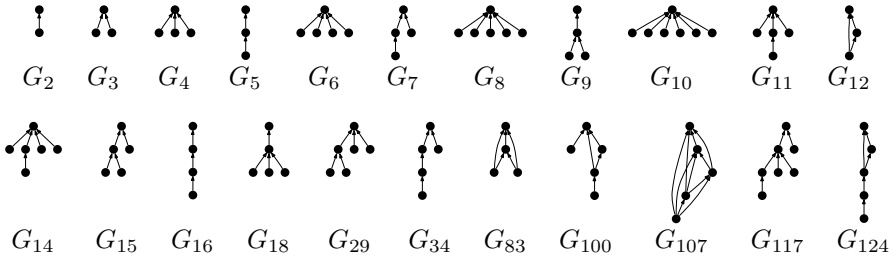


Figure 6. Common blog cascade shapes, ordered by the frequency of appearance.

subgraphs are more frequent than sparser ones, in a manner suggestive of properties in the underlying social network and propagation process.

For example, we found that BoingBoing, which a very popular blog about amusing things, is engaged in many cascades. Actually, 85% of all BoingBoing posts were cascade initiators. The cascades generally did not spread very far but were wide (*e.g.*, G_{10} and G_{14} in Figure 6). On the other hand 53% of the posts from an influential political blog MichelleMalkin were cascade initiators, but the cascades here were deeper and generally larger (*e.g.*, G_{117} in Figure 6) than those of BoingBoing.

Simple model of information cascades

We also developed a conceptual model for generating information cascades that produces cascade graphs matching several properties of real cascades. The model builds on independent cascade model [15]. Our model is intuitive and requires only a single parameter that corresponds to how interesting (easy spreading) the conversations in general on the blogosphere are.

Intuitively, cascades are generated by the following principle. A post is posted at some blog, other bloggers read the post, some create new posts, and link the source post. This process continues and creates a cascade. One can think of cascades as graphs created by the spread of a virus over the Blog Network. This means that the initial post

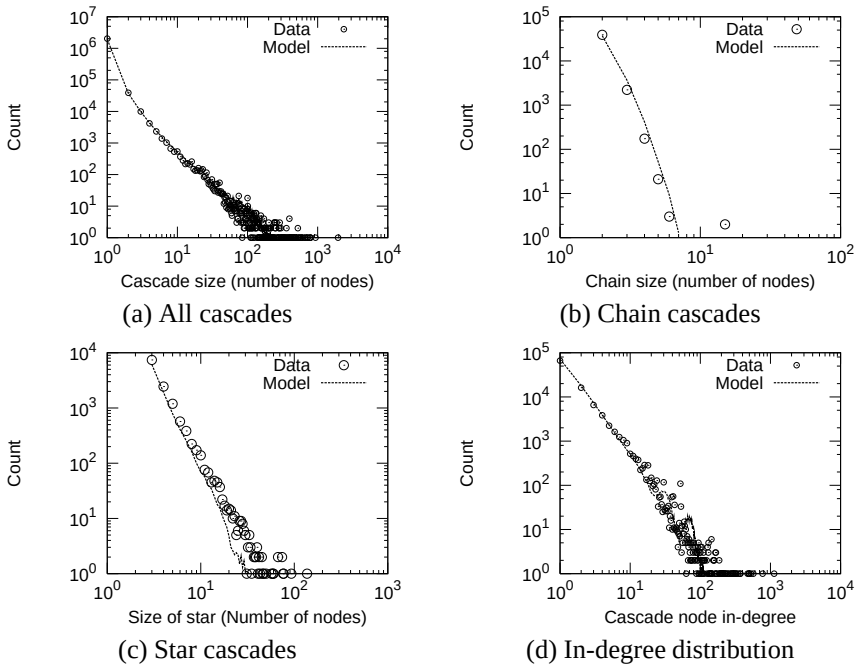


Figure 7. Comparison of the true data and the model. We plotted the distribution of the true cascades with circles and the estimate of our model with dashed line. Notice remarkable agreement between the data and the prediction of our simple model.

corresponds to infecting a blog. As the cascade unveils, the virus (information) spreads over the network and leaves a trail. To model this process we use a single parameter β that measures how infectiousness of the posts on the blogosphere. Our model is very similar to the SIS (susceptible – infected – susceptible) model from the epidemiology [14].

Figure 7 compares the cascades generated by the model with the ones found in the real blog network. Notice a very good agreement between the reality and simulated cascades in all plots. The distribution over cascade sizes is matched best. Chains and stars are slightly under-represented, especially in the tail of the distribution where the variance is high. The in-degree distribution is also matched nicely, with an exception for a spike that can be attributed to a set of outlier blogs all with in-degree 52.

Node selection for early cascade detection

Next, we explore the general problem of detecting outbreaks in networks, where we are given a network and a dynamic process spreading over this network, and we want to select a set of nodes to detect the process as effectively as possible.

Many real-world problems can be modeled under this setting. Consider a city water distribution network, delivering water to households via pipes and junctions. Accidental or malicious intrusions can cause contaminants to spread over the network, and we want to select a few locations (pipe junctions) to install sensors, in order to detect these contaminations as quickly as possible.

Similarly with blogs we want to select a set of blogs to read (or retrieve) which are most up to date, *i.e.*, catch (link to) most of the stories that propagate over the blogosphere. Our goal is to select a small set of blogs (two in case of Figure 1) which "catch" as many cascades (stories) as possible. A naive, intuitive solution would be to select the big, well-known blogs. However, these usually have a large number of posts, and are time-consuming to read. We show, that, perhaps counter-intuitively, a more cost-effective solution can be obtained, by reading smaller, but higher quality, blogs, which our algorithm can find.

Node selection criteria

There are several possible criteria one may want to optimize in outbreak detection. For example, one criterion seeks to minimize *detection time* (*i.e.*, to know about a cascade as soon as possible, or avoid spreading of contaminated water). Similarly, another criterion seeks to minimize the *population affected* by an undetected outbreak (*i.e.*, the number of blogs referring to the story we just missed, or the population consuming the contamination we cannot detect). Optimizing these objective functions is NP-hard [16], so for large, real-world problems, we cannot expect to find the optimal solution.

Exploiting submodularity

In our work [21] we show that these and many other realistic outbreak detection objectives are *submodular* [25], *i.e.*, they exhibit a diminishing returns property: Reading a blog (or placing a sensor) when we have only read a few blogs provides more new information, than reading it after we have read many blogs (placed many sensors). We find ways to exploit this submodularity property to *efficiently obtain* solutions which are *provably close* to the optimal solution. These guarantees are important in practice, since selecting nodes is expensive (reading blogs is time-consuming, sensors have high cost), and we desire solutions which are not too far from the optimal solution.

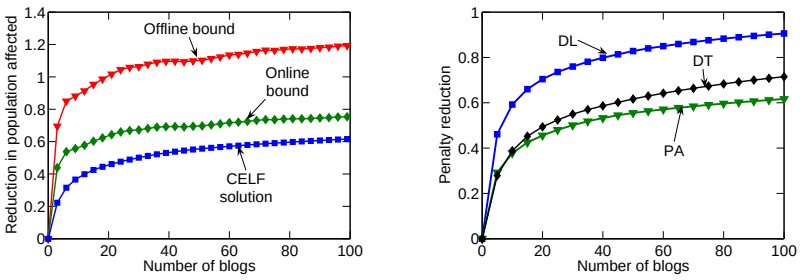
We also show that many objective functions for detecting outbreaks in networks are submodular, including detection time and population affected in the blogosphere and water distribution monitoring problems. We show that our approach also generalizes work by [15] on selecting nodes maximizing influence in a social network.

We also exploit the submodularity of the objective (*e.g.*, detection time) to develop an efficient approximation algorithm, CELF, which achieves near-optimal placements (guaranteeing at least a constant fraction of the optimal solution), providing a novel theoretical result for non-constant node cost functions. CELF is up to 700 times faster than simple greedy algorithm. We also derive novel online bounds on the quality of the placements obtained by *any* algorithm.

Evaluation on water distribution and blog networks

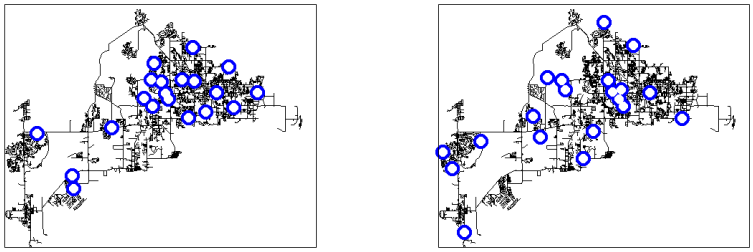
We extensively evaluate our methodology on the applications introduced above – water quality and blogosphere monitoring. These are large real-world problems, involving a model of a water distribution network from the EPA with millions of contamination scenarios, and real blog data with millions of posts.

First, we evaluate the performance of CELF, and estimate how far from optimal the solution could be. Obtaining the optimal solution would require enumeration of $2^{45,000}$



(a) Solution quality (b) Various Objective functions

Figure 8. Both plots show the solution quality vs. the number of selected sensors (blogs). (a) Performance of CELF algorithm and off-line and on-line bounds. Notice on-line bound is much tighter. (b) Compares different objective functions: detection likelihood (DL), detection time (DT) and population affected (PA).



(a) Population Affected (b) Detection Likelihood

Figure 9. Water network sensor placements: (a) when optimizing Population Affected, sensors are concentrated in high population areas. (b) when optimizing Detection Likelihood, sensors are uniformly spread out.

subsets. Since this is impractical, we compare our algorithm to the bounds we developed. Figure 8(a) shows scores for increasing budgets when optimized the Population affected criterion. As we select more blogs to read, the proportion of cascades we catch increases (bottom line). We also plot the two bounds. Notice the off-line bound (top line) is very loose. On the other hand, our on-line bound is much tighter than the traditional off-line bound.

In contrast to the off-line bound, our on-line bound is *algorithm independent*, and thus can be computed regardless of the algorithm used to obtain the solution. Since it is tighter, it gives a much better worst case estimate of the solution quality. For this particular experiment, we see that CELF works very well: after selecting 100 blogs, we are at most 13.8% away from the optimal solution. Similarly, figure 8(b) shows the performance using various objective functions. By using the on-line bound we also calculated that our results for all objective functions are at most 5% to 15% from optimal.

In August 2006, the Battle of Water Sensor Networks (BWSN) [28] was organized as an international challenge to find the best sensor placements for a real metropolitan area water distribution network. In Figure 9 we show two 20 sensor placements obtained by our algorithm after optimizing Detection Likelihood and Population Affected, respec-

tively. When optimizing the population affected, the placed sensors are concentrated in the dense high-population areas, since the goal is to detect outbreaks which affect the population the most. When optimizing the detection likelihood, the sensors are uniformly spread out over the network. Intuitively this makes sense, since according to BWSN challenge, outbreaks happen with same probability at every node. So, for Detection Likelihood, the placed sensors should be as close to all nodes as possible.

References

[1] L. A. Adamic and N. Glance. The political blogosphere and the 2004 u.s. election: divided they blog. In *LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery*, pages 36–43, 2005.

[2] E. Adar and L. A. Adamic. Tracking information epidemics in blogspace. In *Web Intelligence*, pages 207–214, 2005.

[3] R. M. Anderson and R. M. May. *Infectious diseases of humans: Dynamics and control*. Oxford Press, 2002.

[4] N. T. J. Bailey. *The Mathematical Theory of Infectious Diseases and its Applications*. Hafner Press, 2nd edition, 1975.

[5] S. Bikhchandani, D. Hirshleifer, and I. Welch. A theory of fads, fashion, custom, and cultural change in informational cascades. *Journal of Political Economy*, 100(5):992–1026, October 1992.

[6] J. J. Brown and P. H. Reingen. Social ties and word-of-mouth referral behavior. *The Journal of Consumer Research*, 14(3):350–362, 1987.

[7] J. Chevalier and D. Mayzlin. The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, 43(3):345, 2006.

[8] P. Domingos and M. Richardson. Mining the network value of customers. In *KDD '01: Proceedings of the 7th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2001.

[9] N. Glance, M. Hurst, K. Nigam, M. Siegler, R. Stockton, and T. Tomokiyo. Deriving marketing intelligence from online discussion. In *KDD '05: Proceeding of the 11th ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 419–428, 2005.

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

[11] M. Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 83(6):1420–1443, 1978.

[12] M. S. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78:1360–1380, 1973.

[13] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 491–501, 2004.

[14] H. W. Hethcote. The mathematics of infectious diseases. *SIAM Rev.*, 42(4):599–653, 2000.

[15] D. Kempe, J. M. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *KDD '03: Proceedings of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146, 2003.

[16] S. Khuller, A. Moss, and J. Naor. The budgeted maximum coverage problem. *Information Processing Letters*, 70(1):39–45, 1999.

[17] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *WWW '02: Proceedings of the 11th international conference on World Wide Web*, pages 568–576, 2003.

[18] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 611–617, 2006.

[19] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. In *EC '06: Proceedings of the 7th ACM conference on Electronic commerce*, pages 228–237, 2006.

[20] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*, 1(1):2, 2007.

[21] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *KDD '07: Proceeding of the 13th ACM SIGKDD international conference on Knowledge discovery in data mining*, 2007.

[22] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs. In *SDM '07: Proceedings of the SIAM Conference on Data Mining*, 2007.

[23] J. Leskovec, A. Singh, and J. M. Kleinberg. Patterns of influence in a recommendation network. In *PAKDD '06: Proceedings of the 10th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 380–389, 2006.

[24] A. L. Montgomery. Applying quantitative marketing techniques to the internet. *Interfaces*, 30:90–108, 2001.

[25] G. Nemhauser, L. Wolsey, and M. Fisher. An analysis of the approximations for maximizing submodular set functions. *Mathematical Programming*, 14:265–294, 1978.

[26] M. E. J. Newman, S. Forrest, and J. Balthrop. Email networks and the spread of computer viruses. *Physical Review E*, 66(3):035101, 2002.

[27] J. G. Oliveira and A. L. Barabasi. Human dynamics: The correspondence patterns of darwin and einstein. *Nature*, 437:1251, 2005.

[28] A. Ostfeld, J. G. Uber, and E. Salomons. Battle of water sensor networks: A design challenge for engineers and algorithms. In *8th Symposium on Water Distribution Systems Analysis*, 2006.

[29] P. Resnick and R. Zeckhauser. Trust among strangers in internet transactions: Empirical analysis of ebay's reputation system. In *The Economics of the Internet and E-Commerce*. Elsevier Science, 2002.

[30] M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. In *KDD '02: Proceedings of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 61–70, 2002.

[31] E. M. Rogers. *Diffusion of Innovations*. Free Press, New York, fourth edition, 1995.

[32] A. Vazquez, J. G. Oliveira, Z. Dezso, K.-I. Goh, I. Kondor, and A.-L. Barabasi. Modeling bursts and heavy tails in human dynamics. *Physical Review E*, 73(3):036127, 2006.