Where The Blogs Tip: Connectors, Mavens, Salesmen and Translators of the Blogosphere

Ceren Budak
Department of Computer Science, UCSB
Santa Barbara, USA
cbudak@cs.ucsb.edu

Divyakant Agrawal
Department of Computer Science, UCSB
Santa Barbara, USA
agrawal@cs.ucsb.edu

Amr El Abbadi
Department of Computer Science, UCSB
Santa Barbara, USA
amr@cs.ucsb.edu

ABSTRACT

Why is it that some ideas or products become unusually successful and get adopted widely while others don't? This question has been puzzling many social scientists, economists, politicians and educators for a long time. Knowing the answer to this question can help deliberately start such successful cascades. Many theories have been introduced in this topic by economists and social scientists and these theories have been backed by small numbers of case studies. In this paper, we will focus on the popular theories introduced in "The Tipping Point" by Malcolm Gladwell. The basic idea is the crucial effect of three types of "fascinating" people that the author calls *mavens, connectors and salesmen* on the effectiveness of a cascade. Those people are claimed to "play a critical role in the word-of-mouth epidemics that dictate our tastes, trends and fashions". In this work, we investigate existence of *mavens, connectors and salesmen* in the blogosphere. We formally define what it means to be a *maven, connector* or a *salesman* and study their possible effect on the success of cascades in the blogosphere. We also study a fourth type of interesting actor that we call *translator*, an actor that acts as a bridge between different interest groups and communities. We observe that these four types of important players do in fact exist in the blogosphere and they have high correlation with successful cascades. More interestingly, we show that the cascades where these actors act as intermediaries rather than initiators are more likely to reach a larger size.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications - Data Mining

General Terms

Algorithms; Measurement

Keywords

social networks, information cascades, heuristics, connectors, mavens, salesmen, translators

1st Workshop on Social Media Analytics (SOMA '10), July 25, 2010, Washington, DC, USA.

1. INTRODUCTION

How and why a society adopts new ideas and how information cascades happen are questions that have been puzzling social scientists, economists and politicians for decades. Information cascades have been studied by social scientists for a long time, mostly under diffusion of innovations, but these studies have the disadvantage of lacking large-scale data to support the proposed theories. The mainstream adoption of the Internet, Web and recently online social networks (OSNs) have provided great opportunities for studying information cascades at a very large-scale. Therefore, recently there has been a great deal of interest in information diffusion in computer science research. Various studies have studied information diffusion from a variety of perspectives [25, 18, 19, 20, 23]. The interest in identifying influential people or what makes an idea successful is not exclusive to the academic world. Klout, a startup company that measures the influence of individuals and topics around the web, just secured 1.5 million dollars of investment [21].

An important problem that has attracted attention is how to start a successful information cascade or influence propagation on a large scale [20, 24, 7, 9, 31, 36]. Most of these works suggest using expensive algorithms that are not applicable to very large-scale networks compared to the simple heuristics used in earlier work [34]. Those algorithms assume diffusion models that do not fully capture the behavior observed in real social networks. Leskovec et al. show the existence of a saturation point for recommendations given to a person after which additional recommendations seem to have negative effects [23]. Such behavior has been overlooked by many models introduced. Another example is the existence of multiple conflicting campaigns that are concurrently happening in a network. This characteristic has been largely ignored with a few notable exceptions [4, 3, 10, 6, 22]. In addition to the shortcomings of current models that information propagation algorithms are based on, their robustness with respect to the parameters of the model remain an open problem. The correctness of the algorithms critically depend on the assumption that we are fully aware of the exact real world, i.e. we know the *exact* influence each node has on another. It is not clear if the performance of these solutions will still be within a known bound of the optimal solution if the assumptions about the real world are inaccurate even by minor perturbations. In light of such observations, we claim that a good alternative to applying those complex algorithms that provide approximation bounds for a world that is not necessarily representative of the real one, is to study social networks and the cascading behavior to extract properties or heuristics that would help create successful information cascades. Such new findings, based on real world observations, can also be used to develop a better model for information diffusion which can be used to design optimization algorithms. There have been many heuristics that have been studied in online social

networks, but most of these heuristics are static and do not reveal information about the dynamics of cascades. Degree centrality and distance centrality are two such examples [20]. In this paper we study the correlation between these actors and success of the cascades and study "possible effectiveness" of such heuristics. Different from prior works, we do not use a model on which to evaluate these heuristics. The evaluation is done on real data sets of real information cascades. Therefore, the findings we present do not have the disadvantage of being based on a model that does not capture real-world behavior. We also introduce several new heuristics that capture the dynamic behavior of OSNs and how people interact with information and each other. We show that there are various different types of "actors" that can be identified in a social network that are likely to have a positive effect on the success of a cascade. The notion of people in a social network having characteristics that distinguish them from the rest and these characteristics having a notable impact on the success of diffusion of information has also been studied in social sciences under diffusion of innovations [32]. Similar phenomena attracted attention from the computer scientists as well. Mathioudakis et al. study certain types of actors in a social network called "starters" and "followers" [26] and introduce efficient algorithms to detect such actors in a social network whereas Ghosh et al. study the notion of *leaders* and *negotiators* [12].

Although most of recent work dismisses simple heuristics as a means of cascade strategy [20], there are still communities that claim certain simple heuristics can be used to explain why certain ideas or products are highly adopted while others are not. Gladwell [13] identifies three types of people that are crucial to the success of an information cascade. He uses ideas from epidemiology while reasoning about various social phenomena such as the growth of AIDS in the United States in the late 1980s and the sudden increase in the popularity of the Hush Puppies as a fashion trend. The basic idea is simple, that the process that creates such trends is very similar to epidemics, that ideas and products behave like viruses and that there is a dramatic point beyond which the idea or the epidemic rises or falls dramatically. Reaching that "tipping point" is in no way coincidental according to Gladwell. What makes an idea or a product reach the tipping point? There are three rules of epidemics introduced in this study: the Law of the Few, the Stickiness Factor and the Power of Context. The Law of the Few, which is the central idea studied in our work, identifies three types of people that are crucial to the success of a cascade. Those three types of actors are: *Connectors, Salesman* and *Mavens*. In this paper, we will first formally define what these three types of "actors" correspond to in a social network and study their correlation with the success of information cascades. In addition to those actors, we define another class of actors called *Translators*, those that act as bridges among different interest groups or communities, and study them in detail as well. Similar actors have been studied in various contexts [16, 5, 12] and have been referred to as *structural holes* [16], *weak ties* [5] or *negotiators* [12]. Although these terms all aim to capture those that *act as bridges between close-knit groups*, their definition of what makes such an actor is slightly different. *Translators*, which we are interested in identifying, are similar to these notions in the sense that we are also interested in identifying "bridges" between communities, but we use a more behavioral approach than a structural approach, i.e. we use cascading behavior to determine the communities and translators of the network rather than depending solely on the structure of the network.

We show that involvement of these actors in an information cascade has high correlation with the success of the cascade. More interestingly, we show that cascades that include these actors as *intermediaries* rather than *initiators* are more likely to be successful. These findings have various implications: 1) simple heuristics can be employed to improve the success of information cascades; 2) cascades that involve these actors as *intermediaries* rather than *initiators* are more likely to be successful so algorithms that aim to optimally reach those actors rather than start from them can increase the effectiveness of the solution; 3) these actors can be used to augment the models of diffusion to capture real world behavior.

2. THE LAW OF THE FEW

Consider the problem of maximizing the spread of influence in social networks. Choosing the right set of people to first influence by a new idea to achieve this goal is computationally very expensive. Therefore, researchers as well as marketeers and politicians have been looking for "types of people" to identify in a much easier manner that will improve the effectiveness of their campaigns. Kempe et al. study one notion, degree centrality, and claim that their influence is far from the optimal [20], the same notion is shown to be effective for the problem of minimizing the spread of influence in some circumstances and not so effective in others [4]. Can we conclude that we can never find simple heuristics that will lead to successful cascades? Gladwell does not agree with this. He identifies three types of important people that make an idea tip. So far the computer science research community studied one of those people, the *Connector*. But is it possible that we are missing something? Gladwell does not claim that the *Connectors* are sufficient for the success of a cascade. Have we been looking at wrong or incomplete heuristics and settling with expensive algorithms?

We investigate the effectiveness of the three heuristics introduced in [13], as well as another heuristic that we call *translators*. We study each in isolation as well as in combination. We will first explain the *connectors, mavens, salesmen* and *translators* in more detail and extract mathematical properties of such actors to identify them in a graph-theoretical manner. In Section 2.1, we first introduce the notation used in the rest of this paper. In Section 2.2, we will formally introduce the heuristics to be investigated.

2.1 Preliminaries

A social network can be modeled as a directed graph $G = (N, E)$ consisting of nodes *N* and edges *E*. A node n_i is a *neighbor* of n_j if and only if there is an edge $e_{i,j}$ from n_j to n_i in *G*. Each edge $e_{i,j}$ has a weight $p_{i,j}$ assigned to it that models the direct influence n_i has on n_j. The historical data H of the set of cascades observed on graph G is represented as a set $H = \{c_1, c_2, ..., c_m\}$ where c_i is an ordered list of nodes n_j s.t. $n_j \in N$ where the nodes are ordered w.r.t. the time they *adopt* or *advocate* cascade c_i. For the blogosphere, which consists of blogs that post various posts which can link to each other, *advocating* can be a blog posting a post that relates to a specific cascade. Note that a blog can post multiple posts on a cascade and therefore a node n_j can appear in cascade c_i multiple times. We denote the first time n_j appears on the list of c_i as the time it *adopted* cascade c_i and all the occurrences of n_j in c_i as the times it *advocates* cascade c_i. Note that this type of information diffusion is different from the commonly used models of Independent Cascade or Linear Threshold [14, 15], where a node has only one chance to activate each of its neighbors in a cascade. We also denote the time n_j *adopts* c_i as $t_{i,j}$ and this refers to the index of first occurrence of n_j in c_i.

2.2 Mavens, Connectors, Salesmen and Translators

2.2.1 Connectors

The first type of actor introduced by Gladwell is the *Connector*. The *Connector*, translated into a graph, is a node that has high

degree centrality. This type of actor and its effect on information propagation have been studied for a long time [34, 20, 4]. Consider the graph G, let $degree(n_i)$ denote the number of outgoing edges originating from n_i. W.l.o.g. let $\langle n_1, n_2, n_3, ..., n_n \rangle$ denote the list of all the nodes in N ordered by $degree(.)$. The top-k Connectors list CON_k consists of nodes $\langle n_1, n_2, n_3, ..., n_k \rangle$.

2.2.2 Mavens

The second type of important actor in information propagation is the *Maven*. The word "maven" comes from Yiddish and means one who accumulates knowledge. While describing mavens, Gladwell lists three important characteristics: 1) they seek new knowledge, 2) they cannot help but help others and therefore share the knowledge they acquire with others and 3) an individual hearing something from a maven is very likely to believe the correctness and importance of this piece of information. Translating those features into graph theory, we define *Mavens* as those that start a large number of cascades (they are the original source of new information) and have high influence on their neighbors.

In order to locate mavens in a graph G, we first need to define *influence* formally. We define $Inf_{i,j}$, influence of n_i on n_j as the ratio of the number of cascades that n_i successfully activated n_j to all the cascades n_i tried activating n_j.

$$Inf_{i,j} = \begin{cases} \perp & Success_{i,j} + Fail_{i,j} \leq \tau \\ \frac{Success_{i,j}}{Success_{i,j}+Fail_{i,j}} & \text{otherwise} \end{cases}$$

where

$$Success_{i,j} = |\{c_k | n_i \in c_k \wedge n_j \in c_k \wedge t_{k,i} \leq t_{k,j}\}|$$
$$Fail_{i,j} = |\{c_k | n_i \in c_k \wedge n_j \notin c_k\}|$$

Threshold τ is used to avoid deciding on influence of nodes for which there are a small number of datapoints. Given this definition of *node-to-node influence*, the influence set of a node n_i can be defined as $InfSet_i = \{n_j | e_{i,j} \in E \wedge Inf_{i,j}! = \perp\}$ and the aggregate influence of a node n_i as:

$$Inf(n_i) = \sum_{n_j \in InfSet_i} Inf_{i,j} / |\{InfSet_i\}| \quad (1)$$

Defining influence of a node using Equation 1, we sort the nodes N of graph G w.r.t. their influence. W.l.o.g. let $CM = \langle n_1, n_2, n_3, ..., n_n \rangle$ denote the list of all the nodes in N ordered by $Inf(.)$. We define the set *Candidate Mavens* as the first $\lceil n/k \rceil$ nodes in list CM, i.e. the top $100/k$-percentile *local inluentials*.

We then further filter the list CM to only include those nodes that are *original sources* of information. For a cascade c_i, let n_j denote the first blog in the list c_i. We call n_j the *creator* of cascade c_i. The *maven score* of a node n_j can be computed as:

$$MS(n_j) = |\{c_k | t_{k,j} = 1\}| \quad (2)$$

In order to find out the final set of mavens, we re-sort CM using the score computed with Equation 2. W.l.o.g. let the list created with this method consist of nodes $\langle n_1\prime, n_2\prime, n_3\prime, ..., n_{\lceil n/k \rceil}\prime \rangle$. The top-k Mavens list M_k consists of nodes $\langle n_1\prime, n_2\prime, n_3\prime, ..., n_k\prime \rangle$.

2.2.3 Salesmen

The third kind of actor that Gladwell introduces is the *salesman*, a person with high charisma who can sell ideas to almost anyone since he never gives up. When reaching out to a person with an idea and being declined, a *connector* or a *maven* would give up, but a *salesman* tries different ways of persuasion [13]. We capture the notion of a *salesman* as nodes that have a large number of trials to activate its neighbors for cascades that the node itself is a part of.

Note that, we assume a broadcast system, i.e. when n_j adopts or advocates c_i all its neighbors are notified of the adoption/advocation, which also means that n_j tries to activate all its inactive neighbors in c_i. Therefore the number of times n_j advocates c_i is equivalent to the number of times n_j tries to activate its neighbors in c_i. Therefore we compute the *salesman score* of each node as:

$$SalesScore(n_i) = \sum_{c_k, s.t. n_i \in c_k} SalesScore_{i,k} / |\{c_k | n_i \in c_k\}| \quad (3)$$

where $SalesScore_{i,k}$ is defined as the number of times n_i appears in list c_k. Having defined *salesman score* of each node using Equation 3, we sort the nodes of the graph G using this measure. W.l.o.g. let $\langle n_1, n_2, n_3, ..., n_n \rangle$ denote the list of all the nodes in N ordered by $SalesScore(.)$. The top-k Salesmen list S_k consists of nodes $\langle n_1, n_2, n_3, ..., n_k \rangle$.

2.2.4 Translators

We also study another class of actors we call *translators*. These actors act as bridges or translators among different communities and therefore have the power of changing the context in which to present an idea. For instance consider a blogger that is interested in economy and political issues. This blogger will have the ability to read another blog that is written in the context of politics and be able to extract the economical issues of the same problem and restate the same piece of information in the context of economics. This way, this piece of information could reach the community that is interested in economics.

Identifying translators in a social network is a non-trivial problem. In order to identify the translators in the blogosphere, first we need to detect the communities. Even though the problem of community detection has been studied extensively [8, 30, 28, 29, 11, 2, 37, 17], so far there is no agreement on exactly what defines a community. One method for detecting communities is hierarchical clustering, which assigns nodes to the same community if they are sufficiently similar to each other [33]. Similarity measures can be based on the set of neighbors, Euclidean distance and Pearson correlation or the number of paths between nodes. A community can also be defined as "a group of vertices in which there are more edges between vertices within the group than to vertices outside of it" [8], a notion that can be identified using graph partitioning [30] and modularity optimization techniques. A third way to define a community identifies set of nodes that have higher than expected number of edges within themselves and lower than expected edges between them [28, 29]. These algorithms maximize *modularity*, the fraction of all edges within communities minus the expected value of the same quantity. In addition to those somewhat static views on what makes a community Ghosh et al define a community as being composed of individuals who have more influence on individuals within the community than on those outside of it [11]. A structure-based view of influence is used that is defined as the number of paths, of any length, that exist between two nodes. Inspired by the notions presented in [11], we believe static properties such as the sole existence of links is not enough to define what makes a community. We claim that existence of flow of influence between nodes is a better indication of community. Different from [11], we do not solely depend on a structural view of influence. Instead, we use the cascade history H to capture "real" influence cascade in the network and use H to detect communities. Considering a cascade $c_i \in H$, the nodes that are a part of this cascade influence each other in a way and the fact that they belong to the same cascade also indicates similar interests.

Since we are looking for translators, i.e. the nodes of a networks that belong to various communities, we need to employ methods to

detect overlapping communities. Only a few clustering algorithms introduced so far can detect overlapping communities [2, 37, 17]. In this work, we implemented the one introduced in [2] which defines a cluster as a locally optimal subgraph with respect to a given *density metric*. The algorithm consists of two parts: an initialization phase which creates seed clusters; and an improvement phase which repeatedly scans the nodes in order to improve the current clusters until reaching a locally optimal collection of clusters. Next we present the *density metric* that will be used by the algorithm presented in [2] to cluster the nodes into overlapping communities. The *density metric* we used is based on the influence notion presented above.

The *density metric* used is based on the co-occurrences of nodes in cascades. To this end, we construct a hashtable T to keep track of the co-occurrences of nodes in cascades. The keys in T are of the form (n_i, n_j). Let $V_{T,i,j}$ denote the value of the key (n_i, n_j) in T, i.e. the number of cascades that n_i and n_j both *advocated*. The *density metric* was chosen as W_{ai}, called the average influence, which is defined for a set of nodes Set_k as:

$$W_{ai}(Set_k) = \sum_{n_i \in Set_k} \sum_{n_j \in Set_k} Value_{i,j} / |Set_k| \qquad (4)$$

where

$$Value_{i,k} \begin{cases} V_{T,i,j} & (n_i, n_j) \in T \\ 0 & \text{otherwise} \end{cases}$$

This will assign a high *density* to a set of nodes that occur frequently in the same cascade and will also avoid assigning too many nodes to one set by offsetting the weight by the number of nodes in the cascade. Having discovered communities this way, the next step is to detect the translators between communities. Let the set of communities detected by employing the algorithm presented in [2] and using the *density metric* W_{ai} be $\{Set_1, Set_2, ..., Set_m\}$. We simply define *translator score* of a node as:

$$TranslatorScore(n_i) = |\{Set_j | n_i \in Set_j\}| \qquad (5)$$

Having defined the *translator score* of each node using Equation 5, we sort the nodes of the graph G using this measure. W.l.o.g. let $\langle n_1, n_2, n_3, ..., n_n \rangle$ denote the list of all the nodes in N ordered by $TranslatorScore(.)$. The top-k Translators list B_k consists of nodes $\langle n_1, n_2, n_3, ..., n_k \rangle$.

3. THE BLOGOSPHERE

3.1 Experimental Setup

In this section we will present some basic terminology about the blogosphere and information cascades for this context. We will explain how the social network graph and the cascade information were extracted from the blogosphere data.

Weblogs have become a predominant way of sharing data online. The blogosphere has considerable influence on how people make decisions in areas such as politics or technology. [1]. As shown in [25], there is non-trivial influence cascades in the blogosphere when cascades are defined purely on the propagation of links.

Following the same ideas presented in [25], we use two different graph structures that result from interactions in the blogosphere, *Blog Network* and *Post Network*. Figure 2 summarizes the graphs created for the blogosphere presented in Figure 1. As shown in Figure 1, the blogosphere consists of blogs that have one or more posts in them. Those posts can have links to other posts that reside in either the same blog or some other blog (in addition to links to other websites). Note also that the edges in Figure 1 are given as

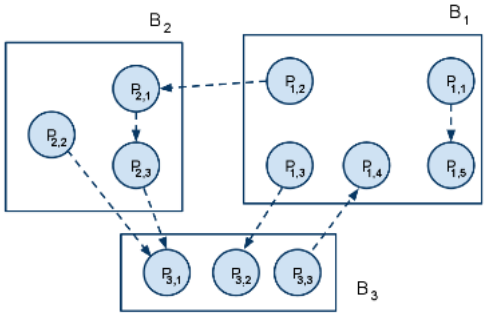
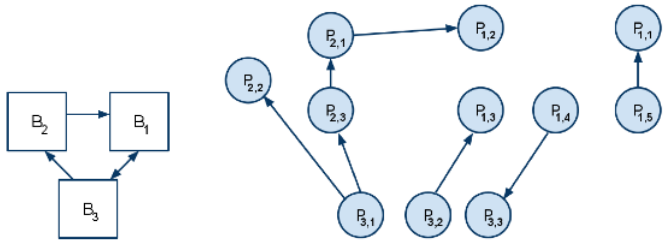


Figure 1: Blogosphere data

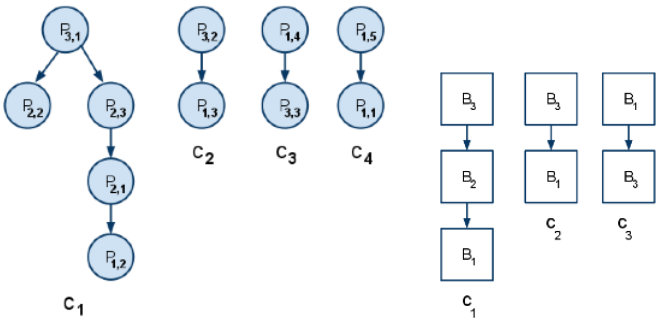
dotted lines whereas for all the other graphs we use solid lines. This notation aims to highlight the difference between *linking* behavior which is demonstrated in Figure 1 and *influencing* behavior which is demonstrated in the *Blog Network* and the *Post Network*. A post P_i linking to another post P_j will be represented by a dotted line from P_i to P_j in the blogosphere graph, whereas the same behavior translates to a solid line from P_j to P_i in the *Post Network*. The *Blog Network* given in Figure 2(a) can be created by processing the blogosphere data and creating a directed edge from a blog B_i to B_j if there is at least one post in B_j that links to another post in B_i. The *Blog Network* basically represents the blogs that are *aware* of each other and therefore can *influence* each other. The *Post Network* that is presented in Figure 2(b) can be obtained by ignoring the blog information and only considering the links between posts. From this data, we can extract the post cascades. In order to identify a cascade, we first find nodes (posts) that have no outgoing edges, i.e. do not link to any other post. These posts are *creators* of cascades. As shown in Figure 3(a), $P_{3,1}$ is such a post. Since $P_{2,2}$ links to $P_{3,1}$, we say $P_{2,2}$ is *influenced* by $P_{3,1}$. Another way to look at the cascades is to consider the flow of information between blogs rather than posts. Figure 3(b) represents the cascades from this perspective. Cascade c_1 induces a cascade from blog B_3 to B_2 and then to B_1. We refer to cascades that involve only one blog as *intra-blog* cascades. Note that c_4 is a *intra-blog* cascade and therefore is filtered out from *blog cascades*.



(a) Blog Network (b) Post Network

Figure 2: Graphs created for blogosphere data

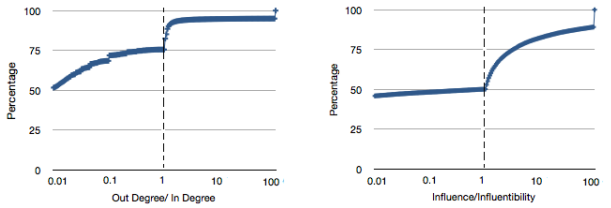
We used the August-October 2008 Memetracker data that contains timestamped phrase and link information for news media articles and blog posts from different blogs and news websites [27]. The data set consists of 53,744,349 posts and 2,115,449 sources of information (blogs, news media and sources that reside outside the blogosphere). 819,368 of those sources information are blogs. When we restrict the cascades to *non-trivial* cascades, i.e. cascades that involve at least two posts, and filtered out invalid data (e.g. posts that link to posts in the future) there are 744,189 cascades.



(a) Cascades from the post data perspective

(b) Cascades from the blog data perspective

Figure 3: Cascades extracted from Figure 2



(a) Degree distribution correlation

(b) Influence/Influentibility ratio

Figure 4: Degree and Influence correlation of the blogosphere

We have studied those cascades, as well as the unsuccessful trials to cascades, i.e. the *intra-blog* cascades that have only one post involved in them, to validate/invalidate theories as to what makes a cascade tip and reach a large number of blogs. Before we present our findings in this subject, we first present some interesting characteristics of the blogosphere data.

3.2 General Properties of the Blogosphere

Before conducting experiments for evaluating the effect of the four actors discussed, we studied certain structural and dynamic characteristics of the blogosphere to check if adhere to the findings on the characteristics of social networks so far. This section presents details about two classes of characteristics: degree distribution and correlation and influence distribution and correlation.

Similar to previous findings on degree distribution in social networks, the August-October blogosphere data set has a power-law degree distribution for both the in-degree and out-degree of the blogs. Note that, this agrees with the similar study done on blogosphere data collected from August-September 2005 [25]. Due to space limitations, we omit the graphs related to degree distribution of the blogosphere. Figure 4(a) shows correlation between the in-degree and out-degree of blogs. It gives the CDF (cumulative distribution function) of the ratio of out-degree to in-degree of the blogs. *Outgoing* edges of a blog are defined as the blogs that this particular blog influences, i.e. edges extracted from the *Blog Network*. As shown in Figure 4(a), there are a large number of blogs that have a very low out/in degree ratio, suggesting that these blogs have a "high potential" to be influenced by other blogs and not have a high potential to influence others. There are also a large number of blogs with a very high out/in degree that have "a high potential" to influence other blogs and not to be influenced themselves. The blogs for which the out/in ratio > 100 or indegree= 0 were mapped to out/in degree = 100. Therefore, the spike at 100 is more pronounced. Note that, the degree of a node indicates a "potential" rather than real influence, much like a user having 1000 friends on Facebook is not necessarily influencing all the 1000 friends. The spike at the ratio of 1 also suggests that there are a large number of blogs whose in-degree and out-degree are highly correlated.

Next, we present our findings on the influence propagation behavior which demonstrates high skew of influence in the blogosphere. The analysis is done on *local influence*, i.e. influence calculated using Equation 1. We define *influence* as how likely "neighboring" blogs of a certain blog are to link to some post in that specific blog. An "influential" blog will have many posts that are linked by posts in other blogs. Likewise, we define "influentibility" as how likely a blog is to be influenced by some other blog, i.e. have posts that link to posts in other blogs. An "easily influenceable" blog will have many outgoing links to posts in other blogs. The distribution of influence and influentibility in the blogosphere adheres to the power law distribution. There are a large number of blogs that are very difficult to influence and only a small number of blogs that are easily influenced. Similarly many of the blogs have very small influence. These findings indicate that there are many blogs that have very low effect on the flow of information in the blogosphere. These are usually isolated from the other blogs, not influencing or being influenced by other blogs. This indicates that there is room for improvement in terms of computation savings since many of the blogs can be filtered out during computations. Due to space limitations, we omit the graphs representing the influence and influentibility distribution of the data set. Next, we looked into the correlation between the influence and influenbility of each blog to see if the nodes that are influential are also the nodes that are easy to influence. As shown in Figure 4(b), a large proportion of the blogosphere population has very low influence compared to their influentibility and vice versa,hence the spikes around the values of 0 and 100. There are also a large number of blogs which have comparable influence and influentibility as can be observed from the spike around the value of 1.

4. THE LAW OF THE FEW IN THE BLOGOSPHERE

In order to evaluate the "possible" effect of the four types of actors presented in Section 2.2, we analyzed the blogosphere data to mine the correlation between the four types of actors and the success of cascades. Even though correlation does not guarantee causality, a high correlation can be a good indication of possible causality. In the rest of this paper, we will loosely use the term "possible effect" to refer to such high correlation. As explained in Section 3.1, there are two ways to examine the cascade data for the blogosphere: from the blog perspective as in the case of Figure 3(b) and from the post data perspective as in the case of 3(a). We analyzed the blogosphere data to identify *maven, connector, salesman* and *translator* blogs rather than posts. Therefore the data set used in these experiments are of the form presented in Figure 3(b). The *historical data* consists of a set of cascades, each of which consists of the blogs that *advocated* that specific cascade ordered by the time of advocation. Here *advocation* is equivalent to posting a new post in this cascade. Our definitions of the *actors* depend on this historical linking behavior rather than any other semantic information. This means that we do not leverage from the available data to verify for instance whether translators do in fact translate or mavens are actual sources of information. But this method enables us to *easily* identify types of actors which is the crucial idea behind our goal of identifying *simple heuristics*.

Using the equations presented in Section 2.2, we identify the blogs that are in the top-k *mavens, connectors, salesmen* or *transla-*

tors lists (for various k) and investigate if cascades involving those actors are in fact more likely to be successful. We measure the success of a cascade by the number of distinct blogs that *advocate* it. We address two questions: 1) Considering the *extraordinarily* successful cascades, can we extract certain properties that hold consistently for such cascades? 2)What are good measures and indications to reach *better than expected* success with high probability? Does the involvement of the four types of actors indicate *better than expected* success? In Section 4.1 we address the first question and in Section 4.2, we address the second question.

4.1 Extraordinarily successful cascades

In this section, we address the question of what makes a cascade reach far beyond the "tipping point" and become *extraordinarily* successful. Since cascade sizes follow the power-law distribution, our data set contains only a small subset of cascades that involve 500 or more blogs. Table 1 summarizes the top 10 cascades of the data set. The second column, *size*, refers to the size of the cascade, i.e. the number of distinct blogs that *advocate* this cascade. The third, fourth, fifth and sixth columns represent the first index of a node from the top-k maven, connector, salesmen and translator list in the cascade respectively. The seventh, eighth and ninth columns represent the number of top-k mavens, connectors, salesmen and translators in the cascade respectively. We chose to evaluate the top 0.1, 1 and 10 percentile of each top-k list and then choose the best of these three choices in terms of performance. This resulted in setting $k = 5561$ for top-k mavens, $k = 148$ for top-k connectors, $k = 211$ for top-k salesmen and $k = 29$ for top-k translators. The results indicate that the *extraordinarily successful* cascades almost always started from a *connector* and involved a large number of connectors. This explains the general convention of choosing highly connected people as starters of cascades, a strategy that has been employed for a long time [34]. Table 1 also demonstrates that those cascades involve a large number of *salesmen* and *translators* but do not necessarily start with one. It is worthwhile to note that the top 3 cascades involve *all* of the top-29 translators. This reflects the likely importance of having a translator between communities to make an idea attractive for a diverse set of people/blogs. The positive correlation observed for the *connectors, salesman* and *translators* does not exist for the *mavens*, Table 1 demonstrates no correlation between mavens and "extraordinarily" successful cascades.

Although it is temping to suggest, using Table 1, that starting a cascade from a *connector* and reaching out to a *translator* or *salesman* will "guarantee" an extraordinarily successful cascade, this would be an overly simplified conclusion. Basing methods of influence maximization on findings from a small number of success cases introduces the danger of investing in "accidental influentials" [35] rather than real influentials. In the next section we will study the success of cascades that involve the four actors to investigate if these cascades achieve *better than expected* success. Since we have a larger data set for such cascades, the findings give better indications of the effects of these actors. For instance, Table 1 shows that almost all the top cascades started from a *connector*, but for all cascades that start from a *connector*, what is the success, i.e. cascade size, distribution? We address such questions in the next section.

4.2 Cascades with better than expected success

In this section, we investigate the possible effect *connectors, mavens, salesmen* and *translators* have on a cascade being (several magnitudes) more successful than expected. The kind of questions we would like to address are: How much more likely is a cascade

Table 1: Top Cascades

Cascade	size	t_M	t_{CON}	t_S	t_T	$\#_M$	$\#_{CON}$	$\#_S$	$\#_T$
1	1114	601	0	168	65	2	17	5	29
2	1095	581	0	146	174	2	17	5	29
3	1095	581	0	146	174	2	17	5	29
4	1061	51	0	0	107	2	15	5	5
5	928	527	0	77	107	1	16	5	5
6	920	519	0	72	101	1	15	5	5
7	908	506	2	38	68	1	15	5	5
8	876	205	0	36	91	2	15	4	5
9	779	362	4	100	36	1	14	4	5
10	776	361	2	97	33	1	14	4	5

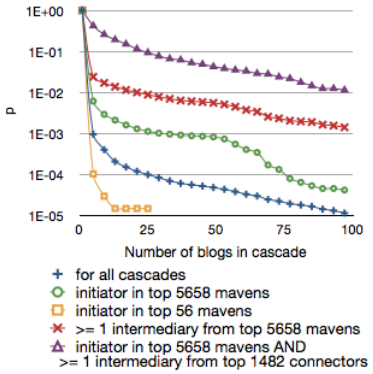


Figure 5: Cascades with Mavens

starting from a *connector* or involving a *connector* to reach out to more than k blogs (say 50 or 100) than an arbitrary cascade? Figures 5, 6, 7, 8 and 9 present our findings. All graphs are log-scale and show the cumulative distribution function (CDF) of the size of cascades. The X-axis represents the cascade size whereas the Y-axis represents the log of ratio of cascades that have at least that many blogs in it. Inspired by the findings presented in Table 1, we introduce the notion of *initiators* and *intermediaries*. An *initiator* is a blog that starts a cascade. For instance B_3 in Figure 3(b) is the *initiator* of cascade c_1. An *intermediary* is a blog that is a part of a cascade and therefore acts as an "intermediary" for spreading the cascade further. All the blogs that are part of a cascade are *intermediaries* of that cascade. B_1, B_2 and B_3 are *intermediaries* for cascade c_1. We study the possible effect of the four different actors when they act as *initiators* and *intermediaries* for cascades.

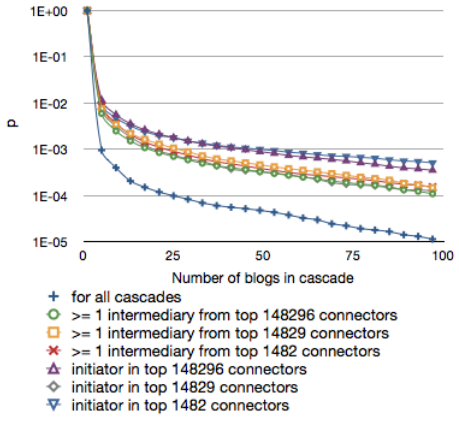


Figure 6: Cascades with Connectors

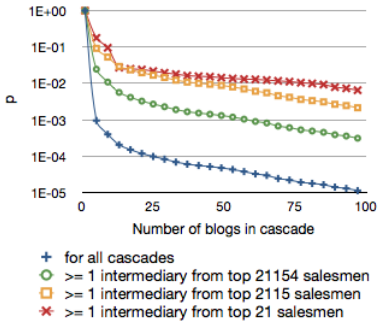


Figure 7: Cascades with Salesmen

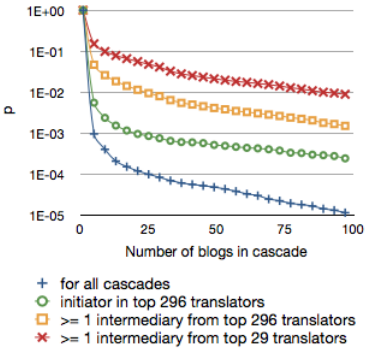


Figure 8: Cascades with Translators

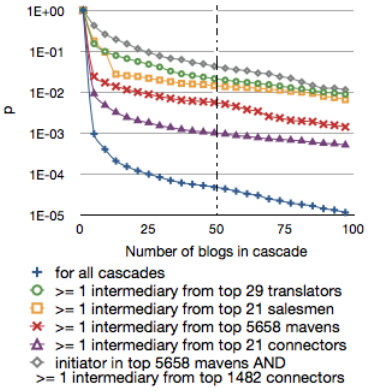


Figure 9: Comparison of All Actors

Figure 5 presents our findings on the possible effect of mavens on the success of cascades. We have set $\tau = 10$ to avoid using data based on interactions between two blogs that have interacted less than 10 times in the entire cascade history. We also used $k = 100$ to limit the mavens to 1-percentile. After sorting the *Candidate Mavens* w.r.t. the $MS(.)$ measure defined in Equation 2, we picked the top 10 and 0.1 percentile as the top-k maven list. These settings resulted in the three top-k maven lists presented in Figure 5. As Figure 5 demonstrates, cascade size follows a power-law distribution, a property that has been demonstrated in prior work [25]. Also note that the figures are in log-scale so the cascade size has even a heavier tail than presented in Figure 5. Figure 5 demonstrates the positive correlation of the success of a cascade and starting the cascade from a *maven*. A cascade starting from a maven is likely to be more successful than expected, i.e. a random cascade. However it is important to note that although a cascade starting from a maven is substantially more likely to reach 20-70 blogs, the same is not true for reaching 100 or more blogs. This indicates that *mavens* in the blogosphere can be good heuristics to use if the goal is to reach a large scale, but not necessarily the *maximum scale*. Figure 5 also demonstrates the case of starting a cascade from a *maven* and reaching a connector at some point. We can see that the success of cascades in this case is several magnitudes larger. Also, considering the drop of performance from top-5658 to top-56 *mavens*, we reason that the blogs that start *too many* cascades show spammer-like behavior and therefore are less effective in creating successful cascades. To further investigate this, we also evaluated blogs that start a large number of cascades without limiting the definition to those blogs that have high trust from their immediate neighborhood (use $k = 1$ to remove the restriction on *local influentials*) and observed that such blogs have very little effect on the performance of cascades. Such blogs that start a high number of cascades and have low influence on their neighbors can be identified as *spammers* rather than *mavens*. Due to space limitations we omit the detailed results for such blogs.

Figure 6 presents the possible effect of connectors as *initiators* and *intermediaries*. Although the findings presented in Section 4.1 showed that the top-10 cascades almost all started from a connector, Figure 6 demonstrates that starting a cascade from a *connector* is not a *sufficient* condition for creating a successful cascade. Interestingly, we find that when a cascade involves a connector rather than solely starting from it, it is more likely to be successful. Although, this might be counter-intuitive to some, it agrees with studies that claim it is harder to influence a connector and therefore it is more beneficial to reach out to a critical mass of easily influenced people before trying to influence those hard to influence connectors [36].

Figures 7 and 8 present our analysis for salesmen and translators respectively and demonstrate similar findings to those presented in Section 4.1, as they both suggest that *salesmen* and *translators* have a possible positive effect as *intermediaries* of cascades. Figure 9 gives an overview of our analysis for all the actors in a way we can qualitatively compare the possible effect of various actors. It indicates that the salesmen and the translators have a higher impact on the success of cascades than that of the connectors and the mavens. This figure also indicates that the combined effect of actors is much more pronounced then their effects in isolation. Consider the y values on the dotted vertical line in Figure 9. These values can be used to qualitatively compare how effective the actors are in creating cascades of size $>= 50$. Cascades involving any of the four actors have better performance than *expected* (the lowest y value). Cascades that start from a *maven* and have a *connector* as intermediary have the best performance. *Salesmen* and *translators* provide the next best performance, followed by the mavens and connectors respectively.

Our analysis so far demonstrates the high correlation between involvement of the four actors in cascades and the success of cascades. Although this is a good indication of the effects of these actors to increase the spread of influence in a social network, it does not *guarantee* that the cascades are successful *because of* the four actors. One natural question one can ask is: *Are the cascades more successful because of the four actors or are the four actors involved in such successful cascades because they are already successful?* Consider two cascades c_1 and c_2. Assume they consist of blogs $B_i, B_j, B_k, ...$ and $B_i, B_j, ...$ respectively. Let B_k be the first salesman in c_1 and let c_2 have no salesmen. We can see that c_1 and c_2 have similar behavior until c_1 reaches out to a salesman (same prefix until B_k). If the salesman were to be involved in a cascade *because the cascade is successful*, we would expect c_1 and

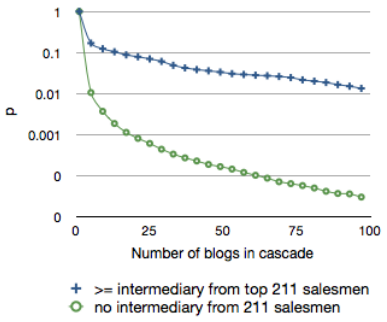


Figure 10: Comparison of cascades that involve a salesman and ones that do not and have "similar behavior" otherwise

c_2 to compare well, whereas we would expect c_1 to be substantially more successful than c_2 if a cascade becomes more successful *because a salesman is involved in it*. We analyzed the blogosphere data to extract such pairs of cascades and aggregated the results. Figure 10 presents our findings for the salesmen and they support our intuition that cascades are more successful *because of the salesmen*. We have done similar analysis on the other three actors and obtained similar results. We omit the details of these results due to space limitations. Although such analysis still does not guarantee the effect of the four actors, it gives us better confidence that the effect of the four actors exists.

Our findings show that the four types of actors are highly correlated with the success of cascades. Cascade size, even of the cascades that involve these actors follow a power-law distribution. However cascades that involve a *connector, maven, salesman* or *translator* have a much lighter tail. We observed that cascades with *translators* and *salesmen* reach a larger number of blogs than those with *connectors* and *mavens*. We also observe that the combination of those actors have a much higher possible effect than what they have in isolation. Another important observation is that the cascades that involve these four actors as *intermediaries* rather than *initiators* are likely to be more successful. This points us towards new algorithms to increase the effectiveness of information cascades. In future work, we plan to investigate heuristics that optimize how to reach out to these four (or possibly more) actors rather than optimizing how to reach out to the entire population which is computationally expensive. Although almost all the top-10 cascades started from a *connector*, the "ambitious" strategy of employing *connectors* as the starters of cascades has a worse *expected* performance than some other methods. The question then becomes: of two choices one with a high benefit but a low probability of delivering that benefit and another choice with a lower benefit but a higher probability of delivering that benefit, which one is more beneficial?

5. CONCLUSION

Understanding why some ideas or products become extraordinarily successful while others do not is an open problem. Getting closer to the answer of this question is crucial for making social networks useful beyond the obvious use of connecting with people online. Having a better grip of the real world behavior of how information cascades happen in social networks can provide us tools for optimally disseminating important piece of information. Prior studies aimed at targeting this problem by building models and solving expensive approximation algorithms with error bounds on those models. However the correctness of these models and their robustness to small errors in the parameters of the models are two main problems that warrant the question: "Do the optimization algorithms on these models achieve their goals?" We claim that by mining the behavior observed on a social network, we can extract certain heuristics that make an information cascade more likely to be successful.

We formally define four types of actors, namely *connectors, mavens, salesmen* and *translators* in the context of social networks and show that the cascades that involve such actors are significantly more likely to be successful. The first three of these actors were first introduced by Gladwell who claims that a few "fascinating" people decide our tastes and trends and they are crucial to reaching the tipping point beyond which an idea or a product becomes extraordinarily successful. Unfortunately, the nature of social networks prevents us from drawing definite conclusions about exactly what makes an idea tip since it is very hard, if not impossible, to find causality by doing controlled experiments. Therefore, we studied a very large historical data set that consistently shows high correlation which is a good indication of causality.

Although most of prior work has focused on optimizing the spread of influence in social networks, in the real world it is hard to pinpoint what exactly makes the spread of influence reach maximum. Therefore, we studied the characteristics related to what makes a cascade *more successful than expected*. We showed that all four actors have a possible effect on such behavior. More interestingly, we showed that cascades that involve at least one *connector, salesman, maven* or *translator* blog are likely to be even more successful than cascades that start with a *connector, salesman, maven* or *translator* blog respectively. These findings provide opportunities for new algorithms aimed at optimizing the diffusion of information in social networks. They indicate that algorithms for finding the best method of reaching out to those actors, rather than the entire network, can be a good heuristic. The types of actors identified can also be used to augment the current models of diffusion to capture real world behavior. As part of future work, we also plan to augment our analysis on the *intermediaries* to investigate if there exists an optimal timing to reach out to a *connector, maven, salesman* or *translator*. Are these actors more useful if they adopt and advocate a cascade early or later on?

We analyzed the blogosphere data to investigate the validity of the heuristics introduced but the same heuristics can be evaluated on other social networks. We believe that different social networks provide different ways of interacting which means that certain actors, while not so significant in certain networks, can be highly influential in others. Although among the four actors the *translators* and *salesmen* were observed to have the highest possible effect in the blogosphere, for other networks, the relative importance of the actors can be different. Using the formalization introduced in this work, which is generic and can be applied to any social network for which there exists historical data on information cascades, these actors and their possible effect can be easily mined and this information can be used to construct algorithms that aim at improving information propagation.

6. ACKNOWLEDGMENTS

This work is partially supported by NSF Grant IIS-0847925. The authors would also like to thank OIT at UCSB for providing access to TRITON resources.

7. 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, New York, NY, USA, 2005. ACM.

[2] J. Baumes, M. Goldberg, and M. Magdon-Ismail. Efficient identification of overlapping communities. In *In IEEE*

International Conference on Intelligence and Security Informatics (ISI), pages 27–36, 2005.

[3] S. Bharathi, D. Kempe, and M. Salek. Competitive influence maximization in social networks. In *In WINE*, pages 306–311, 2007.

[4] C. Budak, D. Agrawal, and A. El Abbadi. Limiting the spread of misinformation in social networks. Technical Report UCSB/CS-2008-02, CS Department, University of California, Santa Barbara, February 2010.

[5] R. S. Burt. *Structural holes: The social structure of competition*. Harvard University Press, Cambridge, MA, 1992.

[6] T. Carnes, C. Nagarajan, S. M. Wild, and A. van Zuylen. Maximizing influence in a competitive social network: a follower's perspective. In *ICEC '07: Proceedings of the ninth international conference on Electronic commerce*, pages 351–360, New York, NY, USA, 2007. ACM.

[7] W. Chen, Y.Wang, and S. Yang. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM International Conference on Knowledge Discovery and Data Mining*, pages 199–208, 2009.

[8] A. Clauset. Finding local community structure in networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 72(2):026132, 2005.

[9] P. Domingos and M. Richardson. Mining the network value of customers. In *Proceedings of the 7th ACM International Conference on Knowledge Discovery and Data Mining*, pages 57–66, 2001.

[10] P. Dubey, R. Garg, and B. D. Meyer. Competing for customers in a social network. Cowles Foundation Discussion Papers 1591, Cowles Foundation, Yale University, Nov. 2006.

[11] R. Ghosh and K. Lerman. Community detection using a measure of global influence. In *Proceedings of the 2nd KDD Workshop on Social Network Analysis (SNAKDD'08)*, 2008.

[12] R. Ghosh and K. Lerman. Leaders and negotiators: An influence-based metric for rank. In *posted in Proceedings of 3rd International Conference on Weblogs and Social Media*, 2009.

[13] M. Gladwell. *The Tipping Point: How Little Things Can Make a Big Difference*. Back Bay Books, January 2002.

[14] 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*, 2001.

[15] M. Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 1978.

[16] M. S. Granovetter. The Strength of Weak Ties. *The American Journal of Sociology*, 78(6):1360–1380, 1973.

[17] S. Gregory. An algorithm to find overlapping community structure in networks. *Knowledge Discovery in Databases: PKDD 2007*, pages 91–102, 2007.

[18] 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, New York, NY, USA, 2004. ACM.

[19] A. Java, P. Kolari, T. Finin, and T. Oates. Modeling the spread of influence on the blogosphere. In *Proceedings of the 15th International World Wide Web Conference*. Citeseer, 2006.

[20] D. Kempe, J. M. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM International Conference on Knowledge Discovery and Data Mining*, pages 137–146, 2003.

[21] Klout raises 1.5 million dollars to measure influence and authority on twitter. `http://techcrunch.com/2010/04/28/klout-raises-1-5-million-to-measure-influence\-and-authority-on-twitter`.

[22] J. Kostka, Y. A. Oswald, and R. Wattenhofer. Word of Mouth: Rumor Dissemination in Social Networks. In *15th International Colloquium on Structural Information and Communication Complexity (SIROCCO), Villars-sur-Ollon, Switzerland*, June 2008.

[23] 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, New York, NY, USA, 2006. ACM.

[24] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM International Conference on Knowledge Discovery and Data Mining*, pages 420–429, 2007.

[25] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs: Patterns and a model. In *Society of Applied and Industrial Mathematics: Data Mining (SDM07)*, 2007.

[26] M. Mathioudakis and N. Koudas. Efficient identification of starters and followers in social media. In *EDBT '09: Proceedings of the 12th International Conference on Extending Database Technology*, pages 708–719, New York, NY, USA, 2009. ACM.

[27] Memetracker data. `http://www.memetracker.org/data.html`.

[28] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69, September 2003.

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

[30] A. Pothen, H. D. Simon, and K.-P. Liou. Partitioning sparse matrices with eigenvectors of graphs. *SIAM J. Matrix Anal. Appl.*, 11(3):430–452, 1990.

[31] M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. In *Proceedings of the 8th ACM International Conference on Knowledge Discovery and Data Mining*, pages 61–70, 2002.

[32] E. M. Rogers. *Diffusion of Innovations, Fourth Edition*. Free Press, 4 edition, February 1995.

[33] J. P. Scott. *Social Network Analysis: A Handbook*. SAGE Publications, January 2000.

[34] S. Wasserman and K. Faust. *Social Network Analysis*. Cambridge University Press, 1994.

[35] D. J. Watts. The accidental influentials. `http://faves.com/users/jlam/dot/76589374563`.

[36] D. J. Watts and P. S. Dodds. Influentials, networks, and public opinion formation. *Journal of Consumer Research*, 34(4):441–458, December 2007.

[37] S. Zhang, R. Wang, and X. Zhang. Identification of overlapping community structure in complex networks using fuzzy c-means clustering. *Physica A: Statistical Mechanics and its Applications*, 374(1):483–490, 2007.