Blog Cascade Affinity: Analysis and Prediction

Hui Li
herolee@pmail.ntu.edu.sg

Sourav S. Bhowmick
assourav@ntu.edu.sg

Aixin Sun
axsun@ntu.edu.sg

School of Computer Engineering
Nanyang Technological University, Singapore, 639798

ABSTRACT

Information propagation within the blogosphere is of much importance in implementing policies, marketing research, launching new products, and other applications. In this paper, we take a microscopic view of the information propagation pattern in blogosphere by investigating *blog cascade affinity*. A *blog cascade* is a group of posts linked together discussing about the same topic, and *cascade affinity* refers to the phenomenon of a blog's inclination to join a specific cascade. We identify and analyze an array of features that may affect a blogger's cascade joining behavior and utilize these features to predict cascade affinity of blogs. Evaluated on a real dataset consisting of 873,496 posts, our SVM-based prediction achieved accuracy of 0.723 measured by F_1. Our experiments also showed that among all features identified, the *number of friends* was the most important factor affecting bloggers' inclination to join cascades.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Search and Retrieval—*Information Filtering*; J.4 [**Computer Applications**]: Social and Behavior Sciences

General Terms

Algorithms, Experimentation

Keywords

Social networks, Network evolution, Blog cascade, Information flow

1. INTRODUCTION

The popularity of blogs has been increasing dramatically over the last few years. According to a recent report by Technorati[1] [19], a popular blog search engine, more than a half of the Internet users read blogs. Technorati have indexed more than 133 million blogs since 2002, and have tracked blogs in 81 languages by June, 2008. Blogs contain diverse varieties of information. General topics include personal diaries, experiences, opinions, information technology, and politics to name a few. Due to their accessible and timely nature, many bloggers surveyed have advertisement on their blogs. The mean annual revenue for blogs with advertisement is estimated to be $6,000 [19]. This figure jumps to $75,000 when we consider only those blogs having 100,000 or more unique visitors per month. Thus, blogosphere provides large amount of latest information on the Web and is of much importance in viral marketing and on-line advertising.

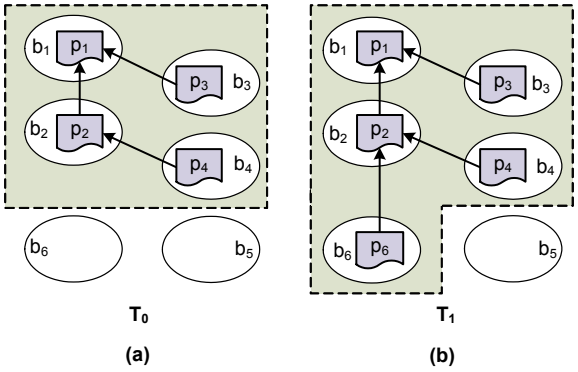

Figure 1: Blog cascade.

1.1 Motivation

A blog consists of several entries. Each entry within a blog, called a *post*, is time-stamped and the most recent entries always appear at the top. Bloggers can also create hyperlinks to other blogs or websites in their posts. The universe of all these blogs and their interconnections is often referred to as *blogosphere* [18, 19]. Blogosphere is an intuitive source for data involving the spread of information and influence within the network of bloggers [1, 7, 11, 18]. By analyzing the linking patterns from one blog post to another, we can infer the way information is propagated through the blog network over the Web. In particular, a piece of information flows from a post to another along the hyperlink between them. For example, consider the Figure 1(a). The ellipses represent different blogs (*e.g.,* b_1, b_2, b_3, b_4, b_5, and b_6), and each ellipse contains a set of posts. The edges in the figure indicate hyperlinks between posts. Assume that post p_1 in blog b_1 contains opinion about recent events related to the spread of H1N1 *virus*. Some time later, blog b_2 visited b_1 and wrote a post p_2 in response to

[1]http://technorati.com

CIKM'09, November 2–6, 2009, Hong Kong, China.

this topic of discussion and explicitly created a hyperlink to p_1. Subsequently, new posts will join this conversation by linking to existing posts. For instance, at time T_0, the structure of this conversation related to H1N1 virus containing a group of posts (p_1, p_2, p_3, and p_4) is depicted by the dashed rectangular component in Figure 1(a). Aggregating all the linked posts by backtracking the hyperlinks will result in a DAG (Directed Acyclic Graph), where each node is a post. Such a DAG is called a *cascade* [14, 20] (also known as conversation tree). All the posts in the same cascade typically discuss about a similar topic.

Observe that at time T_0 there were two blogs, b_5 and b_6, which did not join the conversation on H1N1 virus by writing a post and linking to the cascade. Now assume that at time $T_1 > T_0$ b_6 joined the cascade by writing a post p_6 and linking it explicitly to p_2. The modified structure of the cascade is now depicted in Figure 1(b). Notice that b_5 still did not join the conversation. Why did b_6 join the cascade but b_5 did not? Is it possible to predict the *cascade affinity* of b_5 and b_6 by analyzing the information embedded in the cascade at time T_0? In order to provide answers to these questions, *in this paper we propose an* SVM*-based technique that analyze an array of cascade features to predict which blogs are highly likely to join the cascade in the future.* We refer to the phenomenon of a blog's inclination to join a specific cascade as *cascade affinity*. In the sequel, we shall use blog and blogger interchangeably in the context of cascade affinity.

Although the notion of information cascade was formally introduced by Sushil Bikhchandani [4], it was first systematically studied in the context of blogosphere by Kumar et al. [11]. Majority of research on blog cascades [14] have focused their attention at the *macroscopic* level. In particular, these efforts investigated information flow in cascades, common shapes of cascades and their frequencies, and performed a series of topological analysis. In contrast, we take a *microscopic* view by analyzing cascade affinity behavior of individual bloggers. *To the best of our knowledge, this is the first approach that undertakes a systematic study to predict such behavior.*

The knowledge of a blogger's affinity to cascades is useful in several applications. It not only facilitates the design of advanced blogging system with more sophisticated personalized recommendations and filters, but also help us to set up intelligent strategies in on-line advertising. By predicting which blogs have stronger affinity to a cascade, we can make recommendations to those bloggers in case they have not yet read any post in the cascade. Consequently, we can influence the population faster by accelerating the information propagation process. In this way, new services or products can be disseminated and popularized in a shorter time. Further, we can also predict to what scale of population a cascade will finally expand so that when disseminating an advertisement along blog cascade we can understand the final effect of the advertisement ahead of time and adjust our advertisement strategy accordingly.

1.2 Overview

At first glance, it may seem that we can predict a blogger's affinity to a cascade by analyzing the textual content of existing posts in the cascade and estimating the overlap between the content of the blogger's previous posts and cascade content. However, such *content-aware* strategy is computationally expensive and may adversely affect the accuracy of prediction for several reasons. Firstly, the content of posts are often in conversational language containing flavors of abbreviated words and local lingo. Secondly, a blog cascade may consists of posts written in different languages. Thirdly, posts may only contain multimedia objects such as pictures or video clips. Consequently, these factors make content analysis significantly challenging. Hence, we take a *content-oblivious* strategy to address this issue.

We propose a group of content-oblivious features of a blog cascade that may influence a blog's affinity to the cascade. These features are as follows: *number of friends* that are already in the cascade, *popularity of participants* in the cascade, *number of participants* in the cascade, *time elapsed* since the genesis of the cascade, and *citing factor* of the blog. Note that all these features can be computed by analyzing only the link structure and topology of the cascade. For each of the proposed feature, we investigate how it influences a blog's affinity to the given cascade and performed a one-way analysis of variance (ANOVA) to test the significance of each feature's influence. Then we present an SVM classification-based approach that exploits these features to predict the probabilities of blogs' affinity to a cascade and rank them accordingly. Although we did not exploit the content of the posts, our experimental results demonstrated that our prediction strategy can generate high quality results ($F1$-measure of 72%). In summary, the main contributions in this paper are as follows.

- We propose an array of content-oblivious features that influence a blog's inclination to join a cascade. To the best of our knowledge, these features have not been studied together in the context of a blog network earlier. Further, we present different measures to calculate each feature's effect on the cascade affinity phenomenon.
- We formulate the task of predicting cascade affinity of blogs into a standard classification problem. We present an SVM classification-based technique to evaluate the probability of a blog's affinity to a particular cascade and rank blogs accordingly.
- We present an evaluation of our proposed prediction and ranking scheme demonstrating its practical significance using real data sets. In particular, our proposed technique performs the best when all features except *citing factor* is used.

The rest of this paper is organized as follows. Section 2 presents a brief review of related work. In Section 3, we introduce the data set as well as the cascade extraction process. We propose a group of features for modeling cascade affinity of each blogger in Section 4. In Section 5 we describe our proposed technique to measure and rank the probability of a blog to join a cascade. In Section 6, we conduct an empirical study to evaluate many aspects of our proposed approach and its effectiveness. The last section concludes the paper.

2. RELATED WORK

Much work have been done in the field of information flow modeling and word-of-mouth effect. We conduct a brief review of related work on these fields. Actually the work in this area can be traced back to the epidemic research in

Table 1: Definitions of symbols.

Symbol	Definition
b_j	blog j
c^i	cascade i
T^*	the timestamp of the last post in the data set
T^i	the timestamp when the first post appeared in c^i
$\phi^i(t)$	set of blogs that appeared in c^i before time t
ϕ^i	set of all the blogs that appeared in c^i, $\phi^i = \phi^i(T^*)$
$t^i(j)$	the timestamp when b_j joins c^i *if* $b_j \in \phi^i$; *otherwise*, $t^i(j) = T^*$
$post^i(t)$	the posts appeared in c^i before time t
$post_j(t)$	the posts appeared in blog j before time t
$\mathcal{K}$	friendship threshold

virus propagation problem [6]. Similar work have been done within large online social networks recently focusing on modeling the word-of-mouth effect in different social networks. Backstrom et al. [2] showed the probability of joining a social community depends on the number of acquaintances already in it. Leskovec et al. [12] showed that an individual's probability of buying a DVD increases with the number of recommendation he has received. There is a *saturation point* at the value of 10, which means after a person receives 10 recommendations on buying a particular DVD, the probability of buying does not increase anymore. Cha et al. [5] conducted a study on *Flickr* over the same problem. They showed the probability for a user to become a fan of a photo increases with the number of her friends who are already fans of the photo. These above work all focused on the *number of friends* feature. This feature is also used in our work to model the probability of a blog's affinity to a cascade. Additionally, as we shall see later, our work examines some other features that may affect this behavior.

Several recent papers have focused on modeling the information diffusion patterns within social networks, which is considered to play a significant role in political science and viral marketing [20]. In particular, several algorithms are proposed to find a set of nodes which have the most influence on the others so that by selecting those nodes as seeds we can make our piece of information spread over a large population [8, 10]. Gruhl et al. [7] modeled the information diffusion within blogosphere by defining a *read probability* and *copy probability* for each blogger, and iteratively computed the two and finally converged to the best solution. Agarwal et al. [1] proposed a ranking function for the blogs according to their influence based on the *influence* of posts appeared in each blog. The influence of a post is computed based on its length, comments, and a *propagation factor* which is the aggregated influence from the posts that linked to and from the current one. Another research by Ma et al. [15] focused on finding a set of k candidates as target for marketing strategy using heat diffusion models.

Our research differs from these studies in two ways. Firstly, these current approaches mainly focused on finding the most influential blogs in blogosphere [1], while ours is targeted on discovering the blogs that are most probably to be influenced by other blogs. Hence, our work is orthogonal to these efforts. A recent study showed that large-scale changes in public opinion are not driven by highly influential people who influence everyone else but by easily influenced people influencing other easily influenced people [21]. The authors investigated at a global scale the average size of cascades that are initiated by influential nodes and average nodes using different influence models. They showed that early *adopters* enrolled in a cascade is more important to affect the final cascade size than the initiators. Our work differs from it in that we study in detail under what situation a blogger will be influenced as well as to retrieve the most easily influenced individuals. Secondly, in our work we propose a group of features of blogs and cascades to model the probability of a blog to join a cascade.

In a recent work, Karagiannis et al. [9] studied the human behavior related to email responses. They showed that the *email replying probability* depends on a series of factors. By conditioning on each individual factor, they can achieve moderate prediction gains with respect to predicting replied emails. Putting together all the factors achieves a significant prediction gain. In contrast, our work analyzed the joining behavior of each individual blogger using a group of features. Additionally, we also proposed a ranking scheme that can compute the probability of a blog to join a cascade.

Table 2: Statistics of the data set.

Property	Value
Number of posts	873,469
Number of blogs	156,195
Number of blog-to-blog edges	340,124
Number of edges with weight ≥ 2	139,974
Number of cascades	7,269

3. DATA PREPARATION

In this section, we first introduce the real-world data set we have used for our study. Then, we present our approach of cascade extraction from the data set. In the sequel, we shall use the notations shown in Table 1 to represent different concepts. Generally, we shall use superscript to denote a cascade identifier and subscript to denote a blog identifier.

3.1 Data Set

We extracted our blog data set in September, 2008 using Technorati API[2]. The data set contains blog posts published from June, 2008 to September, 2008. We first selected the group of top 100 blogs indexed by Technorati as seeds. From these seeds, we retrieved the blogs that had linked to these seeds in their posts, and then we iteratively retrieve the posts that linked to the previous level till the sixth level which has been shown as the upper boundary size for most chain cascades [14]. From the XML collection of blogs, we can get the post-to-post relationships. Notice that a post of blog b_i linking to another post of blog b_j do not always indicate a friendship that author of b_i knows author of b_j or b_i regularly reads b_j's blog. So we additionally extracted blog-to-blog relationships with weighted edges where the *weight* of an edge from b_i to b_j indicates the number of times b_i has cited b_j's posts. Such a case, to some extent, indicates that b_i does not read b_j's blog by chance. We use this weighted graph as an indication of friends by filtering out the edges with weight less than a *friendship threshold* $\mathcal{K}$. The characteristics of the data set is shown in Table 2. For each blog, the posts that do

[2]http://technorati.com/developers/api

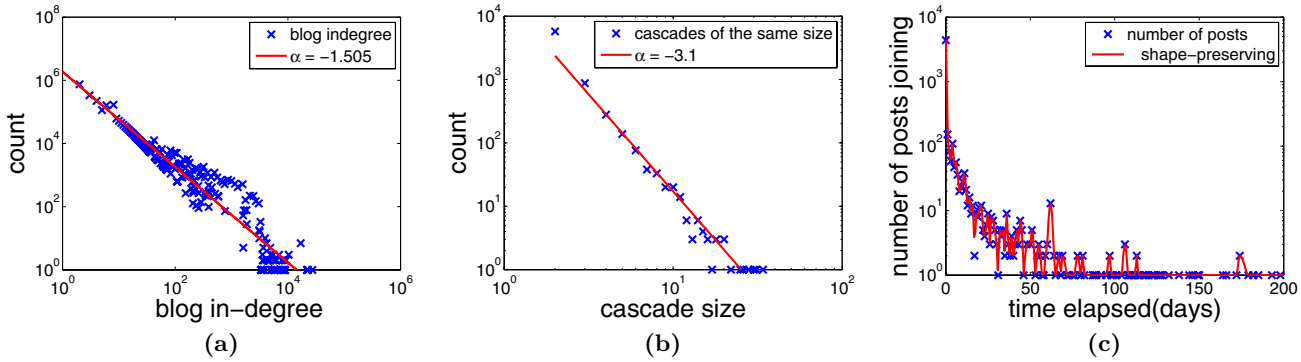

Figure 2: (a) Blog in-degree distribution (b) Cascade size distribution (c) Number of posts joining versus days elapsed.

Algorithm 1: Cascade extraction algorithm.

Input: A set of post-to-post relations $\mathcal{E} = \{e_1, e_2, \ldots, e_m\}$, each element is a pair of posts connected by a hyperlink

Output: A set of isolated cascades $\mathcal{C} = \{c^1, c^2, \ldots, c^s\}$ each of which comprised of connected posts

begin
  initialize each cascade as a single link $\mathcal{C} \longleftarrow \mathcal{E}$;
  while $\exists c^p, c^q$ *and* $c^p \cap c^q \neq \emptyset$ **do**
    forall $c^i, c^j \in \mathcal{C}$ **and** $c^i \neq c^j$ **do**
      if $\phi^i \cap \phi^j$ **then**
        add j to i: $c^i \longleftarrow c^i, c^j$;
        remove j: $\mathcal{C} \longleftarrow \mathcal{C} \setminus \{c^j\}$;
end

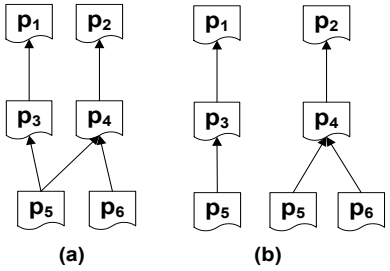

Figure 3: Different approaches for cascades extraction: (a) observed post-to-post relationship, it is also the cascade identified by our approach; (b) the cascades identified from (a) using the approach in [14].

not participate in any cascade are excluded from our data set. Figure 2(a) shows the in-degree distribution of blogs indexed by Technorati till September, 2008. This figure is plotted using the information extracted from our data set. It is shown to follow a power law distribution with exponent equal to -1.505, while in [14] this exponent is reported to be -1.7. Such a phenomenon indicates a few blogs are more connected than the rest. It is consistent with the result of "preferential attachment" model (rich get richer) [3].

3.2 Cascade Extraction

Recall that each blog participates in a cascade by writing a post which links to another post that is already in the cascade. We denote a set of cascades as $\mathcal{C} = \{c^1, c^2, \ldots, c^s\}$. The algorithm for extracting cascades from our data set is shown in Algorithm 1.

Note that the proposed cascades extraction procedure is slightly different from the one described in [14]. Let us elaborate on this further. Consider the scenario in Figure 3(a) which depicts posts and hyperlinks between them. Based on [14], each cascade should have only one initiator (topmost post). Hence, the scenario illustrated in Figure 3(a) have to be considered as two different cascades (have two initiators p_1 and p_2) as depicted in Figure 3(b). In contrast, we treat the scenario in Figure 3(a) as one cascade. The intuitive justification for this is as follows. Observe that the posts in Figure 3(a) are all linked together. That is, both p_1 and p_2 share some common posts in the conversation (*e.g.*, p_5). This may indicate that all these posts are discussing about a common topic. Hence, it makes sense to consider them as part of a single cascade instead of separating them into different ones.

The next step is to post-process the extracted cascades to eliminate the ones which have been there not more than a month till the time T^*. The set of "matured" cascades extracted after the post-processing is represented as: $\mathcal{C} = \{c^i | T^i \leq T^* - 30\}$. The number of cascades detected after filtering out the immature ones is shown in Table 2. The reason for post-processing the cascade set is as follows. We need to ensure that the extracted cascades can provide a robust and accurate framework for feature extraction and subsequent prediction. However, quantifying values of different features based on immature cascades (cascades which have not absorbed all potential participants) will distort the prediction accuracy of cascade affinity. Many participants may join these cascades after time T^* and consequently adversely affect the modeling of the ground truth based on the features set. Obviously, this may result in a deviation between our knowledge about the participants of these cascades and the ground truth. It is worth mentioning that it is not possible to justify the prediction performance without knowing the ground truth.

Figure 2(b) shows the distribution of cascade size extracted from our data set. It is defined as the number of blogs within a cascade. The X-axis is the different sizes of cascades and the Y-axis represents the number of cascades.

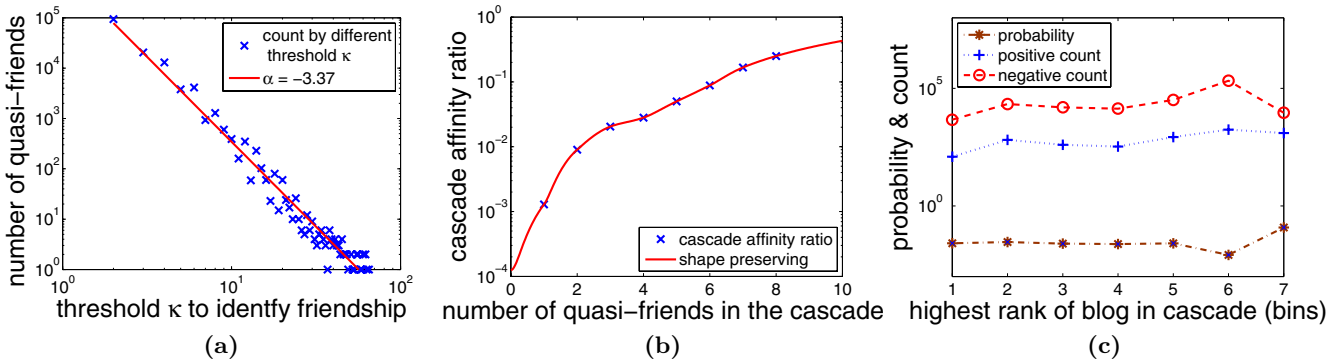

Figure 4: (a) Number of quasi-friends versus $\mathcal{K}$ (b) Cascade affinity ratio versus number of quasi-friends (c) Joining probability by cascade rank.

The minimum size of cascades is defined as 2 which is the trivial case, while the maximum size of a cascade is found to be 34 in our data set. The distribution of cascade size also follows a power law. The exponent found in our data set is -3.1, while this exponent is found to be -2 in the data set used by Leskovec et al. [14]. This deviation is due to the differences between the characteristics of the two data sets and different definition of a cascade in these two approaches.

4. ANALYSIS OF CASCADE FEATURES

In this section, we first present an array of content-oblivious cascade features that may influence a blog's affinity to a cascade. Then, we conduct a one-way variance analysis (ANOVA) on each of these features to quantify their significance related to cascade affinity.

4.1 Elapsed Time

First we present the role of the *elapsed time*. Informally, it refers to the difference between the time a blogger joins a cascade and the cascade creation time. Formally, it is defined as follows.

Definition 1. Let $t^i(j)$ be the time a blogger b_j joins a cascade c^i. Let T^i be the time of creation of c^i. Then, the elapsed time, denoted as $d^i(j)$, is defined as follows:

$$d^i(j) \quad = \quad t^i(j) - T^i$$

We use day as the unit of elapsed time as most bloggers write posts once per day. The distribution of this feature is shown in Figure 2(c). The X-axis represents the time elapsed in days, while the Y-axis represents the number of blogs that join cascades at a specific elapsed time. Observe that 91% bloggers join a cascade during the first week. After that affinity to cascades drops almost exponentially with elapsed time. Note that the above results deviate from other types of social networks, shown in [13], where the authors found that the average number of edges attached to each node did not change much over the lifetime of the node.

4.2 Number of Friends

We introduce the notion of *quasi-friend* to model friendship within blogosphere based on post citings. Formally, *quasi-friend* is defined as follows.

Definition 2. Given two blogs b_1 and b_2, b_1 is a quasi-friend of b_2 if and only if b_2 cites b_1's posts more than $\mathcal{K}$ times.

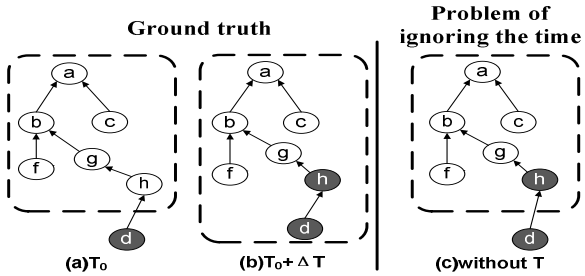

Figure 5: Effect of friendship creation time.

A quasi-friend indicates that b_2 probably often reads b_1's blog. This probability of frequent reading is controlled by the friendship threshold $\mathcal{K}$. Obviously, $\mathcal{K}$ will affect the number of quasi-friends discovered. As shown in Figure 4(a), $\mathcal{K}$ affects the number of quasi-friends exponentially with exponent $\alpha = -3.37$. Notice that if we set $\mathcal{K}$ to a large value then we may extract a very limited number of quasi-friends for a blog. Hence, we set $\mathcal{K}$ to 2 by default. We shall justify this value empirically in Section 6.2. Note that quasi-friendship is *directed*. That is, b_2 is not a quasi-friend of b_1 unless b_1 has cited b_2 more than $\mathcal{K}$ times. Given a value of $\mathcal{K}$, we denote the set of quasi-friends of a blog b_j as $F_j = \{f_1, f_2, \ldots, f_r\}$, where each element f_r is a blog.

Several recent papers have shown that personal behavior in a social network is highly affected by the person's neighbors [2, 5, 12]. Hence, the number of friends a blogger may have in a cascade is an important feature that may influence her decision to join the cascade. Naïvely, the number of friends a blogger has in a cascade can be computed at *any* time after she has joined the cascade. However, this may mislead us from the actual phenomenon as the number of friends is highly influenced by the temporal state of the cascade. Let us elaborate on this further. Consider the Figure 5(a). Each node is a blog and the dashed rectangle denotes a cascade at a particular time. Edges represent hyperlinks related to this cascade. Assume that a blog d joined it at time T_0. Note that at time T_0, d did not have any friend in that cascade. We refer to T_0 as *joining time*. Now assume that at time $T_0 + \Delta T$ node h became a friend of d as shown in Figure 5(b). We refer to this time when a friendship is created as *friendship creation time*. Observe that the number of friends d had during joining time and friendship creation time may be different. However, if we discard these two different phenomenons, then at any time

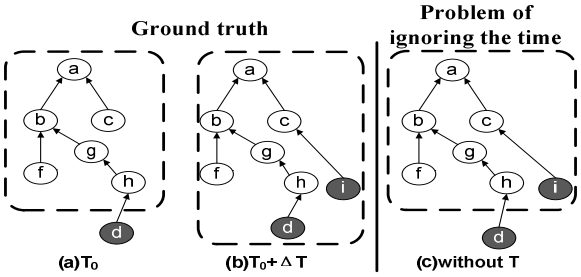

Figure 6: Effect of friendship creation time (contd.).

after $T_0 + \Delta T$ it may seem that d had a friend h in this community when she joined it (Figure 5(c)). Obviously, this is not an accurate reflection of the ground truth. Note that existing work ignore these two types of temporal features while modeling number of friends in a social network. There is another problem if we ignore the above temporal behavior. Consider the Figure 6(a), which represents the same scenario as depicted in Figure 5(a). Now assume that another blog i, who is a friend of d, joined this community at time $T_0 + \Delta T$ as shown in Figure 6(b). If we do not distinguish between times T_0 and $T_0 + \Delta T$, then it may seem that d had a friend i in the cascade when she joined it (Figure 6(c)). However, the truth is that when d joined this cascade at time T_0, she did not have any friend. Hence in our approach, we distinguish between the joining time and the friendship creation time to accurately reflect the ground truth. As we shall see in Section 6.2, this distinction improves the cascade affinity prediction performance significantly.

In our approach, we represent the set of blogs having α quasi-friends in a cascade c^i using $\Gamma^i(\alpha)$ taking into consideration the time $t^i(j)$. It is computed as follows.

$$\Gamma^i(\alpha) \quad = \quad \{b_j \big| |F_j(t^i(j)) \bigcap \phi^i(t^i(j))| = \alpha\}$$

$F_j(t^i(j))$ denotes the set of blogs that became a quasi-friend of j's before time $t^i(j)$, $\phi^i(t^i(j))$ is the set of blogs that appeared in c^i before time $t^i(j)$. Note that by incorporating $t^i(j)$ in our approach, we make a contribution to address the above issues (Fig 5(c) and Fig 6(c)). Based on $\Gamma^i(\alpha)$, we define the notion of *cascade affinity ratio* with respect to the number of quasi-friends.

Definition 3. Given the set of $\Gamma^i(\alpha)$, the cascade affinity ratio, denoted as P_α, is defined as:

$$P_\alpha = \frac{\sum\limits_i |\Gamma^i(\alpha) \bigcap \phi^i|}{\sum\limits_i |\Gamma^i(\alpha)|}$$

We computed P_α for the whole collection of cascades, and plotted the values in Figure 4(b). The X-axis is the α value (number of quasi-friends in a cascade). The Y-axis represents the values of P_α. From the figure, we observe a diminishing return phenomenon. That is, beyond a number each additional quasi-friends in the cascade will contribute less to the probability of joining that cascade. This number is around 7 in this figure. Note that the curve showed in the figure follows similar trend as found in other social networks in [2], and also in [12] where the author found a saturation point in the probability of buying a DVD by the number of recommendations received.

4.3 Popularity of Participants

Next we study the effect of *popularity* of cascade participants on the cascade affinity of a blogger. The idea is similar to the preferential attachment model which was first proposed in [3]. In this model, whenever a new vertex arrives in a network it attaches an edge to an existing vertex with a probability proportional to that of the old vertex's degree. Newman performed a series of analysis on the model in [17]. Leskovec et al. [13] also showed a similar pattern in some real-world data sets. Here we conduct an analysis based on this model. However, in our study when a blog joins a cascade we consider the model at the cascade-level whereas the above approaches consider it at the node level. Then, the *popularity* of a cascade c^i is the highest *rank* of the blogs in the cascade. Formally, it is defined as follows.

Definition 4. Let $D(b)$ be the rank *of a blog b. Then the popularity rank of a cascade c^i that b_j wants to join, denoted as $D_j(c^i)$, is defined as:*

$$D_j(c^i) = \min_{b \in \phi^i(t^i(j))} (D(b))$$

Note that the *rank* of each blog is based on its in-degree (indexed by Technorati). A blog having the largest in-degree has the highest rank as 1. Observe that the above definition can be intuitively explained from the social aspect. When a blogger b_j reads a post p_r she can also see other posts in the same cascade by tracing back the hyperlinks. If there is a popular blog which has a large in-degree in that cascade, then b_j will probably join this cascade. Interestingly, this effect is not so obvious in our result shown in Figure 4(c). The X-axis in the figure is the popularity rank of cascades. A cascade having lower rank means it contains a more popular blog. We plot the numbers of blogs that join a cascade ("positive count") and those who do not ("negative count") by varying the ranks. The curve labeled "probability" represents the ratio: $\frac{\text{positive count}}{\text{positive count+negative count}}$. As shown in the figure although the values along X-axis is in log-scale, the number of joined blogs in each bin do not vary much. This phenomenon indicates that a minority of cascades which have high popularity ranks influence a large number of bloggers to join.

4.4 Number of Participants

Intuitively, a blogger may have stronger affinity to a cascade which has absorbed a lot of participants. Hence, we now conduct an analysis using *number of participants* in a cascade as a feature. We compute the probability of joining a cascade as a function of the number of participants existing in the cascade. The *number of participants* is formally defined as follows.

Definition 5. Let $t^i(j)$ be the time when a blog b_j joins a cascade c^i. Then, the number of participants in c^i at time $t^i(j)$, denoted as $N_j(c^i)$, is defined as:

$$N_j(c^i) = |\phi^i(t^i(j))|$$

Figure 7 shows the probability of joining a cascade as a function of the number of participants in that cascade. The number of blogs inside a cascade ranges from 1 to 33. The probability of joining a cascade with β participants, referred

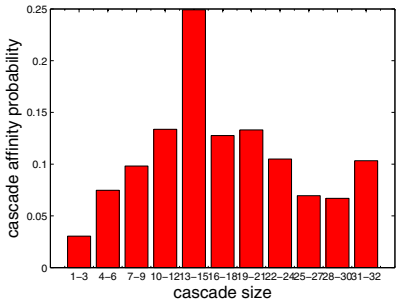

Figure 7: Cascade affinity probability versus cascade size.

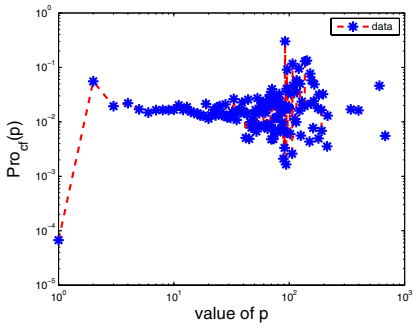

Figure 8: $Pro_{cf}(p)$ versus number of citations.

to as *cascade affinity probability* (denoted as $Pro(\beta)$), can be computed as follows.

$$Pro(\beta) = \frac{\sum_{c^i} |\{b_j | N_j(c^i) = \beta, b_j \in \phi^i\}|}{\sum_{c^i} |\{b_j | N_j(c^i) = \beta\}|}$$

We separate the cascades size range into 11 bins each with length 3. The height of each bar denotes the mean of the three cascade affinity probability values inside that bin. Notice that at the beginning, as the number of participants grows, the probability slightly grows, but after some point, the probability drops down. There is a peak at the point of cascades with the size $13-15$. It indicates that before a cascade absorbed $13-15$ participants, the probability for a blog to join this cascade increases. This represents the cascade initiation period where many new blogs keep on joining the cascade. However, after the number of participants in the cascade has reached a value between 13 and 15, the probability of a blog joining this cascade drops down to a stable value. This represents the stable period after a cascade has got enough attention.

4.5 Citing Factor

The features discussed above are all related to the cascade that a blog is inclined to join. Here we analyze a personal characteristics related to the joining behavior of each blogger. The reason for analyzing this feature is based on the hypothesis that a blogger b_j is more inclined to join a cascade if b_j likes to cite others' posts.

Definition 6. Let $out(\cdot)$ be the number of outlinks of $\cdot$. Then the citing factor of a blogger b_j, denoted as $H_j(c^i)$, is defined as:

$$H_j(c^i) = |out(post_j(t^i(j)))|$$

We can compute the probability for a blog b_j with p citations to join a cascade as follows.

$$Pro_{cf}(p) = \frac{\sum_{c^i} |\{b_j | H_j(c^i) = p, b_j \in \phi^i\}|}{\sum_{c^i} |\{b_j | H_j(c^i) = p\}|}$$

The result is shown in Figure 8. It is distributed almost uniformly with the change to the number of out-links. It is evident that this feature is not very informative as far as cascade affinity is concerned.

4.6 ANOVA Test

We now perform a series of variance analysis on each of these features. For each feature, we compare the values between blogs which finally joined a cascade and those did not using the one-way analysis of variance (ANOVA) to test whether the difference is really caused by the feature values or just by noise in the data. The F and p-values for each feature is shown in Table 3. The result shows that the p-value for citation factor is 0.968 while other features are all less than 0.05. It indicates that the different values of citation factor in both groups should only be considered as noise. The remaining four cascade features are all significant for predicting cascade affinity of a blogger.

Table 3: ANOVA test on cascade features.

Feature Name	*F*	*p-value*
Time elapsed	6.88	$\ll 0.001$
Number of friends	2.85	0.017
Popularity of participants	1.50	0.029
Number of participants	4.36	$\ll 0.001$
Citing factor	0.77	0.968

5. CASCADE AFFINITY PREDICTION

In this section, we describe how the features discussed in previous section can be exploited to predict bloggers who may join a cascade. The prediction involves two steps, namely *candidate blog extraction* and *cascade joining prediction*. We elaborate on these steps in turn.

5.1 Candidate Blog Extraction

For a given cascade, all blogs in the blogosphere are potential blogs that may join the cascade in the future. Nevertheless, many of these potential blogs have no interaction (*e.g.,* read the posts) with the blogs/posts already in the cascade and are unlikely to join the cascade. We therefore only consider a much smaller set of *candidate blogs* that are likely to read one or more posts in the cascade. The candidate blogs are those that have at least one quasi-friend in the given cascade. Formally, for a given cascade c^i, the candidate blogs $cand(c^i)$ that may join c^i is given by the following equation.

$$cand(c^i) = \{j \big| F_j \cap \phi^i \neq \emptyset\} \quad (1)$$

The algorithm for extracting candidate blogs is shown in Algorithm 2. Recall that quasi-friend is defined based on the number of times (*i.e.,* $\mathcal{K}$) a blog cites posts from another blog. Hence, the number of candidate blogs extracted for a given cascade naturally depends on the threshold $\mathcal{K}$. In our experiments, we set $\mathcal{K} = 2$ by default.

For all cascades in our data set, there are 312,414 candidate blogs extracted by Algorithm 2. On average, 43 candi-

Algorithm 2: Candidate blog extraction algorithm.

Input: cascade set $\mathcal{C} = \{c_1, c_2, \ldots, c_s\}$ extracted from the data set
Output: candidates Δ^i for each cascade c^i
begin
  foreach *cascade* $c^i \in \mathcal{C}$ **do**
    foreach *blog* $b_j \in \phi^i$ **do**
      $\Delta^i(j) = \{r | b_j \in F_r(t^i(j))\}$;
      $\Delta^i = \Delta^i \bigcup \Delta^i(j)$;
end

dates are extracted for each cascade. Naturally, the number of candidate blogs increases along the number of participants in a cascade. Particularly, for a cascade having fewer than 10 participants, there are 39 candidate blogs on average; for a cascade having $11 - 20$ participants, this value increases to 64 candidates on average; for a cascade having more than 20 participants, there are 81 candidates on average. From the numbers reported, candidate blog extraction greatly reduces the number of blogs to be considered in the prediction with respect to the total number of blogs in our data set. As an evaluation of candidate extraction, Table 4 shows 76.1% blogs that join a cascade have at least a quasi-friend in it when we set $\mathcal{K} = 2$.

5.2 Cascade Joining Prediction

As discussed in Section 1, in advertising we need to select a limited set of blogs that have the highest probability of joining a cascade. Ideally, we would like to compute a score for each candidate blog extracted with respect to a cascade indicating its likelihood of joining the cascade.

Based on the features identified in the preceding section, the prediction task can be naturally formulated as a binary classification task. Many existing classifiers (*e.g.,* Naïve Bayes, k-Nearest Neighbors, and Support Vector Machines) indeed return a category relevance score for each data instance to be classified indicating its likelihood of belonging to a pre-defined category.

In our experiments, we adopted Support Vector Machines (SVM) classifier [22] due to its promising results reported in many data mining/machine learning tasks. The training of SVM learns a hyperplane that separates the positive training examples from the negative ones with the largest margin. The hyperplane is defined by a vector $\mathbf{w}$ and a parameter b to be learned from the training data. The learned model computes a score for an unlabeled object $\mathbf{x}$ using its decision function $f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} - b$. In our setting, a larger $f(b_j)$ indicates more likelihood of b_j joining the target cascade. To learn an SVM classifier, those candidate blogs that eventually joined and did not join the target cascades were used as positive and negative examples, respectively.

6. EXPERIMENTS

6.1 Experimental Setting

To evaluate the effectiveness of the features in predicting cascade affinity of candidate blogs, we conducted experiments on our data set using 5-fold cross validation. That is, the data set was randomly partitioned into 5 parts and in each evaluation, 4 parts were used as training data and the remaining part was used as test data. The results reported are averaged over the 5 runs.

Table 4: Effect of different values of $\mathcal{K}$.

Value of $\mathcal{K}$	*Candidate size (max. recall)*	*Highest* F_1*-measure*
$\mathcal{K}$=1	946,329 (0.916)	0.702
$\mathcal{K}$=2	312,414 (0.761)	**0.723**
$\mathcal{K}$=3	80,482 (0.242)	0.227

The commonly used performance evaluation measures in classification tasks are *precision*, *recall* and F_1. Precision, denoted by Pr, is the percentage of blogs that eventually joined the target cascade among all blogs predicted to be joining. Recall, denoted by Re, is the percentage of the correct predictions among all blogs that eventually joined the target cascade. Note that, recall is computed with respect to all blogs that finally joined the target cascade regardless of whether the blogs are identified as candidate blogs or otherwise. $F_1 = \frac{2 \times Pr \times Re}{Pr + Re}$ is the harmonic mean of precision and recall. However, both precision and recall are threshold-dependent. A higher threshold leads to higher precision but lower recall. In our experiments, we are more interested in the effectiveness of the features in ranking the candidate blogs according to the likelihood of joining the target cascade. We therefore adopted the area under Precision-Recall curve (AUC-PR) as the evaluation metric.

6.2 Experimental Results

Justification of candidate set. Recall that the number of candidate blogs is affected by the parameter $\mathcal{K}$. As $\mathcal{K}$ increases, the number of quasi-friends identified decreases. Consequently, the candidate blog set shrinks. As a result, the maximum recall decreases, but the prediction performance may not. To determine the optimum value for $\mathcal{K}$, we conducted the prediction using different values of $\mathcal{K}$. Table 4 shows the sizes of candidate blog sets for different $\mathcal{K}$ as well as the highest F_1-measures achieved by selecting the best SVM thresholds. Observe that best F_1-measure is achieved for $\mathcal{K} = 2$. Hence, in the subsequent experiments we shall set $\mathcal{K} = 2$.

Comparison of feature sets. Recall that we have identified five features for cascade affinity prediction, namely *number of friends*, *popularity of participants*, *number of participants*, *citing factor*, and *elapsed time*. To evaluated the effectiveness of these features, we conducted 6 sets of experiments. The first set of experiments used all 5 features for prediction. This feature set is denoted by "ALL" in Table 5. In each of the following five experiments, one feature is removed. For instance, "A-NF" denotes that the feature *number of friends* is removed and the remaining four features were used for prediction. In Table 5, a '✓' indicates that the feature is used and '-' otherwise.

The prediction performances measured by AUC-PR are reported in the last row in Table 5. Using all the five features, the prediction achieved AUC-PR of 0.599. The following observations are made:

- Removal of *number of friends* resulted in significant drop in prediction performance to 0.044 indicating that *number of friends* is the most important factor that affects a blogger's cascade affinity. We validated this by performing another experiment using only the *number*

Table 5: Feature set notations and prediction performance in AUC-PR

Features/AUC-PR/Feature set	ALL	A-NF	A-PP	A-NP	A-CF	A-ET
Number of friends	✓	-	✓	✓	✓	✓
Popularity of participants	✓	✓	-	✓	✓	✓
Number of participants	✓	✓	✓	-	✓	✓
Citing factor	✓	✓	✓	✓	-	✓
Elapsed time	✓	✓	✓	✓	✓	-
AUC-PR	0.599	0.044	0.584	0.595	**0.603**	0.598

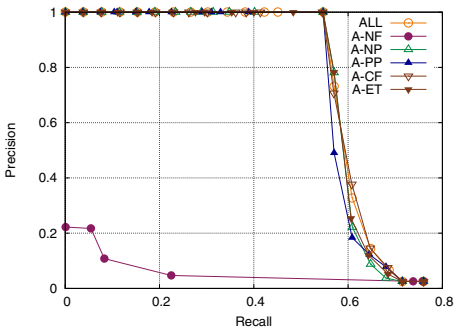

Figure 9: Precision-Recall Curves for different features.

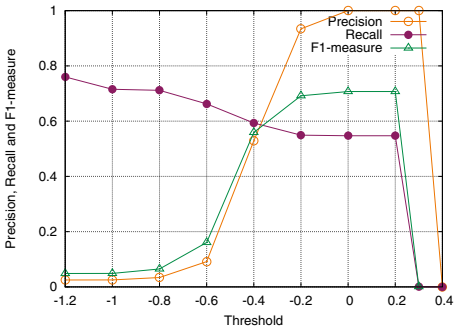

Figure 10: Precision, Recall and $F1$-measure for "A-CF".

of friends as feature. The result AUC-PR in this case is 0.445.

- Removal of either *popularity of participants*, *number of participates*, or *elapsed time* led to a very small performance degradation. These three features indeed contributed to the cascade affinity modeling.
- An interesting observation is that removal of the *citing factor* led to a better AUC-PR than using all the five features. This result clearly indicate that the *citing factor* introduced noise in the prediction, which is consistent with our ANOVA test results reported in Section 4.6. The remaining four features: *number of friends*, *popularity of participants*, *number of participates*, and *elapsed time* achieved the best performance.

For the completeness of the results, Figure 9 plots the Precision-Recall curves of using six different feature sets. All the five runs (except for "A-NF") achieved almost perfect precision before recall reached 0.57. Sharp drop of precision is then observed along with the increase of recall.

Figure 10 shows the precision, recall and F_1-measure by varying the threshold for the feature set "A-CF", which has the best prediction performance among all the approaches.

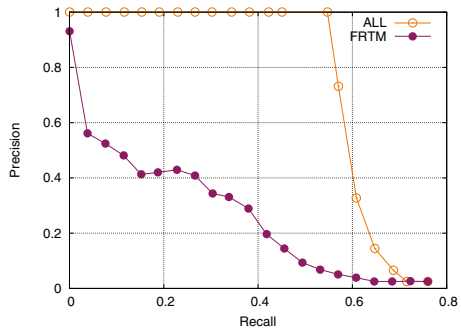

Figure 11: Significance of time for modeling number of friends.

Significance of time for modeling number of friends. Recall that in Section 4.2, we illustrated the significance of time in modeling the number of friends. To justify the goodness of our solution, we compared it with the approach that ignores time. Specifically, if we discard temporal issue in modeling quasi-friends then the definition of $\Gamma^i(\alpha)$ (the set of blogs having α friends in cascade c^i) is modified as follows.

$$\Gamma^i(\alpha) = \{b_j \big| |F_j(T^*) \bigcap \phi^i| = \alpha\}$$

We updated the *number of friends* feature in each candidate vector using the above formula. Using the updated feature vectors, we performed the prediction again. The performance of ignoring the time in *quasi-friend* identification shows a small AUC-PR 0.203 whereas our proposed solution achieves 0.599. The comparison between the Precision-Recall curve of this approach and our proposed solution is shown in Figure 11. Both of the curves use all the five features. "FRTM" represents the approach that discards the temporal aspects in *number of friends*. It is clear that time is an important factor in *quasi-friend* identification as it achieved significantly better prediction compared to the approach that ignores time.

Prediction of top-k bloggers. To study the prediction of accuracy of top-k blogs that are inclined to join a cascade, we computed the precision of our approach to retrieve top-k bloggers ranked based on the predicted scores. Specifically, for each cascade c^i having more than k positive samples, we generate the top-k predicted blogs and compute the *precision* as follows: $Pr^i(k) = \frac{\#\text{true positive}}{k}$. Then for a given k, we compute the *average precision*, denoted as $Pr_{avg}(k)$, using the following formula.

$$Pr_{avg}(k) = \frac{\sum_{|\phi^i| \geq k} Pr^i(k)}{\big|\{c^i \big| |\phi^i| \geq k\}\big|}$$

Table 6 shows average precision values for different k values highlighting the goodness of our approach. Note that

predicted score	label	target cascade ID	candidate blog	URL of the posts that joined the target cascade
0.8856	1	3442	http://redux.quinews.com	http://redux.quinews.com/2008/06/nba-finals-game-1-react/
0.8854	1	532	http://redux.quinews.com	http://redux.quinews.com/2008/06/cohens-on-race-and-politics/
0.8852	1	4530	http://redux.quinews.com	http://redux.quinews.com/2008/06/google-launching-gmail-labs-tonight/
0.8844	1	1032	http://genealogy.darlingranges.com	http://genealogy.darlingranges.com/genealogy-2008-05-06-181713/
0.8841	1	5411	http://politics.nuovoportale.com	http://politics.nuovoportale.com/huffpo-mccain-mooched-off-the-vietnamese-taxpayers
0.8841	1	4705	http://redux.quinews.com	http://redux.quinews.com/2008/05/spencer-tunick-section-2008-people-at-the/
0.8841	1	7230	http://www.dailynewscaster.com	http://www.dailynewscaster.com/2008/06/16/orbiting-the-blogoshpere-2/
0.884	1	2039	http://redux.quinews.com	http://redux.quinews.com/2008/05/haze-review-610-score-swedish-gamereactor/
0.8839	1	2822	http://www.francislarkin.com	http://www.francislarkin.com/2008/06/fivethirtyeightcom-electoral-projections-done-right
0.8836	1	5933	http://redux.quinews.com	http://redux.quinews.com/2008/05/does-chyler-leigh-sex-tape/

Figure 12: Top 10 candidates that are most probable to join a cascade.

Table 6: Average precision versus top-k.

k	1	2	3	4	5
$Pr_{avg}(k)$	**0.970**	0.783	0.707	0.734	0.750

$Pr_{avg}(k)$ may not monotonically decrease with increasing k as the number of cascades in the denominator depends on k. Figure 12 shows the top-10 candidate blogs over entire cascades collection. If the candidate is a positive sample, we also showed the corresponding URL of the post that joins the target cascade.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we analyzed a large publicly available collections of blog information, to investigate bloggers' behavior and interaction with blog cascades. We have identified in total five features, namely number of friends, popularity of participants, number of participants, time elapsed since the genesis of the cascade, and citing factor of the blog, that may play important role in predicting blog cascade affinity so as to identify most easily influenced bloggers. Such bloggers play important role in several real-world applications such as viral marketing. Note that our proposed features are derived from structural information of the cascades without any content analysis of posts/blogs. We performed ANOVA test on these features and showed that all of them, except citation factor, have significant impact on cascade affinity. The cascade affinity prediction is then formulated as a classification task and SVM classifier is employed in our experiments. Using the prediction scores from SVM, the candidate blogs can be ranked according to their probability of joining a cascade. We have evaluated different combinations of the features and our results on cascade affinity prediction is consistent with the ANOVA test. The four features that have significant impact on cascade affinity achieved the best prediction accuracy of 0.603 measured by AUC-PR. Our experimental results also showed that the number of friends plays a significant role in blog cascade affinity prediction.

As part of future work, we intend to investigate how to exploit cascade contents effectively along with structural features for predicting cascade affinity. In particular, recent results showed that *cascade types* may indicate the genre of the content in a cascade [16]. We wish to exploit them in the context of cascade affinity prediction.

8. REFERENCES

[1] N. Agarwal, H. Liu, L. Tang, and P. S. Yu. Identifying the influential bloggers in a community. In *ACM WSDM*, 2008.

[2] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: Membership, growth, and evolution. In *ACM KDD*, 2006.

[3] A. L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, Oct. 1999.

[4] S. Bikhchandani, D. Hirshleifer, and I. Welch. A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades. *Journal of Political Economy*, 100(5):992–1026, 1992.

[5] M. Cha, A. Mislove, and K. P. Grummadi. A measurement-driven analysis of information propagation in the flickr social network. In *ACM WWW*, 2009.

[6] P. S. Dodds and D. J. Watts. Universal behavior in a generalized model of contagion. *Physical Review Letters*, 92(21):218701+, May 2004.

[7] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *ACM WWW*, 2004.

[8] J. Hartline, V. Mirrokni, and M. Sundararajan. Optimal marketing strategies over social networks. In *ACM WWW*, 2008.

[9] T. Karagiannis and M. Vojnović. Behavior profiles for advanced email features. In *ACM WWW*, 2009.

[10] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *KDD*, 2003.

[11] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *ACM WWW*, 2003.

[12] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM Trans. Web*, 1(1):5, 2007.

[13] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *KDD*, 2008.

[14] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs: Patterns and a model. In *SDM*, 2007.

[15] H. Ma, H. Yang, M. R. Lyu, and I. King. Mining social networks using heat diffusion processes for marketing candidates selection. In *ACM CIKM*, 2008.

[16] M. McGlohon, J. Leskovec, C. Faloutsos, M. Hurst, and N. Glance. Finding patterns in blog shapes and blog evolution. In *ICWSM*, 2007.

[17] M. Newman. The structure and function of complex networks. *SIAM Rev*, 45:167–256, 2003.

[18] A. Stewart, L. Chen, R. Paiu, and W. Nejdl. Discovering information diffusion paths from blogosphere for online advertising. In *ACM ADKDD*, 2007.

[19] Technorati, http://www.technorati.com/blogging/state-of-the-blogosphere/. *State of the Blogosphere*, 2008.

[20] D. J. Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9):5766–5771, April 2002.

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

[22] C. C. Chang and C. J. Lin. *LIBSVM: a library for support vector machines*, 2001, software available at *http://www.csie.ntu.edu.tw/∼cjlin/libsvm*.