Data Min Knowl Disc (2014) 28:442–474
DOI 10.1007/s10618-013-0307-0

Affinity-driven blog cascade analysis and prediction

Hui Li · Sourav S Bhowmick · Aixin Sun · Jiangtao Cui

Received: 30 August 2012 / Accepted: 13 March 2013 / Published online: 26 March 2013

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 *macroscopic* and *microscopic* content-oblivious features that may affect a blogger's cascade joining behavior and utilize these features to predict cascade affinity of blogs. Based on these features, we present two non-probabilistic and probabilistic strategies, namely support vector machine (SVM) classification-based approach and *Bipartite Markov Random Field*-based (BiMRF) approach, respectively, to predict the probability of blogs' affinity to a cascade and rank them accordingly. Evaluated on a real dataset consisting of 873,496 posts, our experimental results demonstrate that our prediction strategy can generate high quality results ($F1$-measure of 72.5 % for SVM and 71.1 % for BiMRF) comparing with the approaches using traditional or singular features only such as elapsed time, number of participants which is around 11.2 and 8.9 %, respectively. Our experiments also showed that among all features identified, the *number of quasi-friends* is the most important factor affecting bloggers' inclination to join cascades.

Responsible editor: Bing Liu.

H. Li (✉)· J. Cui
School of Computer Science and Technology, Xidian University, Xi'an, China
e-mail: hli@xidian.edu.cn

S. S. Bhowmick · A. Sun
School of Computer Engineering, Nanyang Technological University, Singapore, Singapore

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 (2008)[1], 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 variety 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 (Technorati 2008). This figure jumps to $75,000 for those blogs with 100,000 or more unique visitors per month.

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* (Stewart et al. 2007; Technorati 2008). Blogosphere is an intuitive source for data involving the spread of information and influence within the network of bloggers (Agarwal et al. 2008; Gruhl et al. 2004; Kumar et al. 2003; Stewart et al. 2007). 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 Fig. 1a. 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 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 Fig. 1a. Aggregating all the linked posts by backtracking the hyperlinks will result in a directed acyclic graph (DAG), where each node is a post. Such a DAG is called a *cascade* (Leskovec et al. 2007b; Watts 2002) (also known as *conversation tree*). Cascade is the most common phenomenon of information propagation within blogosphere. All posts in the *same* cascade typically discuss about a similar topic.

Observe that at time T_0 there are 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

[1] http://technorati.com

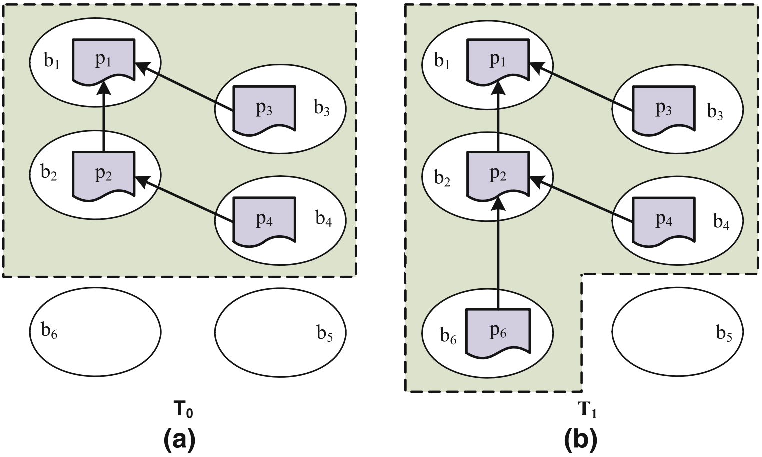


Fig. 1 Blog cascade

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 Fig. 1b. 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 question, *in this paper we propose probabilistic and non-probabilistic techniques based on Bipartite Markov Random Field (BiMRF)* (Shi et al. 2009) *and support vector machine (SVM), respectively, to analyze an array of macroscopic and microscopic cascade features for predicting which blogs are highly likely to join the cascade in the future*[2]. We refer to the phenomenon of a blog's inclination to join a specific cascade as *cascade affinity*.

Although the notion of information cascade was formally introduced by Sushil Bikhchandani (Bikhchandani et al. 1992), it was first systematically studied in the context of blogosphere by Kumar et al. (2003). Majority of research on blog cascades (Leskovec et al. 2007b) 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 hybrid view by analyzing cascade affinity behavior of individual bloggers from both *microscopic* and *macroscopic* level. *To the best of our knowledge, this is the first approach that undertakes a systematic study to predict such behavior.*

1.2 Applications

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

[2] A shorter version of this work has been published in (Li et al. 2009).

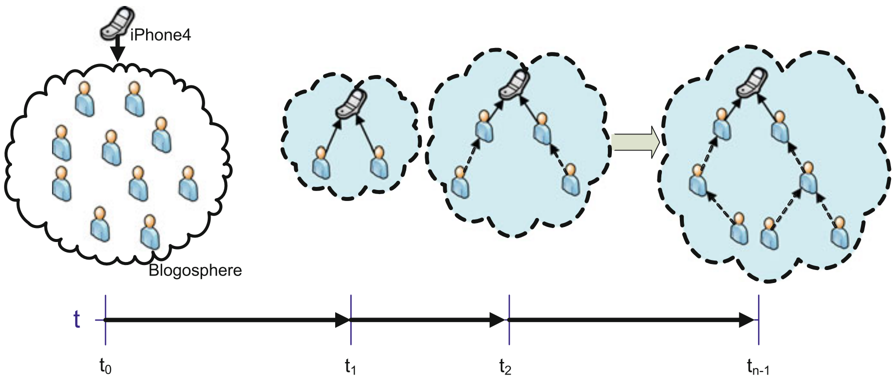


Fig. 2 An example application of blog cascade affinity prediction

strategies in online 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. Furthermore, we can 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 strategies accordingly. For example, assume that the release of "iPhone4" may trigger off a cascade within the blogosphere. At the very beginning, there is no post talking about *iPhone4* at time t_0 (see Fig. 2). Later at time t_1 there are two posts talking about this topic which initiates a cascade over "iPhone4". By studying the cascade affinity of many candidate blogs who are most probable to join this cascade, we can predict that two new bloggers have high chance to join this cascade at time t_2. In Fig. 2, a person with a dashed outgoing arrow represents a blogger who is predicted to join the cascade at current timestamp whereas a solid arrow denotes the blogger who has already joined it. Thus, we may estimate the final scope of the cascade at t_{n-1} by iteratively predicting the set of bloggers who may join it at the next timestamp. Our ability to forecast the final scope of the population involved in this cascade at an early stage paves way to more judicious adjustment of advertisement strategies and budget ahead of time.

1.3 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 *macroscopic* and *microscopic* features of a blog cascade that may influence a blog's affinity to the cascade. The *macroscopic* features are associated with the overall structure of a cascade such as *time elapsed* since the genesis of the cascade, *number of participants* in the cascade, and the shape of the cascade defined by the *star-likeness ratio*. On the other hand, the *microscopic* features (*number of quasi-friends* of a blogger in the cascade, *popularity of participants* in the cascade, *citing factor*, and *initiator-media link*) are related to the blogs or posts of a cascade. Note that all these features are computed by analyzing only the link structure and topology of the cascade. For each of the proposed features, 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 two non-probabilistic and probabilistic methods, namely SVM classification-based approach and *Bipartite Markov Random Field*-based (BiMRF) (Shi et al. 2009) approach, respectively, that exploit these features to predict the probability of blogs' affinity to a cascade and rank them accordingly. 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 72.5 % for SVM and 71.1 % for BiMRF). In summary, the main contributions in this paper are as follows.

- We propose an array of content-oblivious macroscopic and microscopic 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 take two different methods to evaluate the probability of a blog's affinity to a particular cascade and rank them accordingly, namely SVM-based and BiMRF-based classification strategies.
- We present an exhaustive evaluation of our proposed prediction and ranking methods demonstrating their effectiveness and practical significance using real-world datasets. In particular, our proposed techniques perform the best when all features except *citing factor* is used. Further, our results demonstrate that the *number of quasi-friends* feature is the most important factor affecting bloggers' inclination to join cascades.

The rest of this paper is organized as follows. Section 2 presents a brief review of related work. In Sect. 3, we introduce the dataset as well as the cascade extraction process. The macroscopic and microscopic cascade features and their analysis are described in Sects. 4 and 5, respectively. Section 6 describes our proposed models to measure and rank the probability of a blog to join a cascade. In Sect. 7, we conduct an exhaustive empirical study to evaluate many aspects of our proposed techniques and their effectiveness. The last section concludes the paper.

2 Related work

2.1 Community affinity

Much work have been done in the field of information flow modeling and word-of-mouth effect. The work in this area can be traced back to the epidemic research in virus propagation problem (Satorras and Vespignani 2001; Strang and Soule 1998; Newman 2002; Pastor-Satorras and Vespignani 2002; Wang et al. 2003; Dodds and Watts 2004; Clements et al. 2010). 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. (2006) showed that the probability of joining a social community depends on the number of acquaintances already in it. Leskovec et al. (2006; 2007a) reported 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. (2009) conducted a study on *Flickr* over the same problem. They showed that 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 quasi-friends* feature, which can be considered as a microscopic feature of a cascade. Hence, this feature is also used in our work to model the probability of a blog's affinity to a cascade. However, in contrast to the aforementioned work, we exhaustively examine several additional *microscopic* as well as *macroscopic* cascade features that may affect this behavior.

2.2 Information diffusion

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 (Watts 2002; Rogers 2003; Gruhl et al. 2004; Iribarren and Moro 2009; Chen et al. 2009b; Lerman and Hogg 2010). 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 (Kempe et al. 2003; Hartline et al. 2008; Wang et al. 2010; Chen et al. 2009a; Kimura et al. 2009; Adams et al. 2010; Lee et al. 2010). Gruhl et al. (2004) 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. (2008) 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. (2008) focused on finding a set of k candidates as target for marketing strategy using heat diffusion models. Recently, Bao et al. (2010) proposed *AdHeat*, which diffuses hint words of influential users to others and then matches advertisements for each user with aggregated hints.

Our research differs from the aforementioned studies in two key ways. Firstly, existing approaches mainly focused on finding the most influential blogs in blogosphere (Agarwal et al. 2008) whereas the goal of this research is to discover 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 (Watts and Dodds 2007). 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 retrieval of 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.

Karagiannis et al. (2009) studied the human behavior (Davidson et al. 2012) 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 proposed two different models for ranking the blogs according to their probabilities of joining a cascade.

2.3 Retweeting in microblogging

In a different media, Pal and Counts (2011) used the count of original tweets, conversational tweets, and re-tweets of a tweeter as features to rank the *authority* of each tweeter in the context of different topics. They employed a *Gaussian Mixture Model* to compute the authority score of each tweeter. Formally, the authority score for twitter i can be computed as the following.

$$R_G(x_i) = \prod_{f=1}^{d} [\int_{-\infty}^{x_i^f} N(x; \mu_f, \sigma_f)]^{w_f} \tag{1}$$

In the above equation, w_f is the weight that is put on feature f; x_i^f is the associated value of node i on feature f; $N(x; \mu_f, \sigma_f)$ is the univariate Gaussian distribution with model parameters as μ_f and σ_f. The authority score defined above helps in devising a total ordering under "$\leq$" over all the users. To validate their results, they conducted a survey to rate the authority of the tweeters and use it as the ground truth for authority ranking. Blog posts may contain much more information than microblogs. A blog post may consist of text in different languages, urls, tables, photos, video clips, etc. Content-based study in microblogging requires processing all these information with different formats, which can introduce plenty of unsolvable problems. One main contribution of this paper is to resolve the aforementioned problems that multimedia content and multi-language text introduced. We show that cascade behavior can be predicted using only link-based features which avoid these problems.

Recently, Goyal et al. (2012) proposed a *credit distribution* (CD) *model* that leverages on historical *action logs* of a network to learn how influence flows in the network and use this to estimate influence spread. An *action log* is a set of triples (u, a, t) which says user u performed action a at time t. The basic idea is that if user v takes action a and later on v's friend u does the same, then the authors assume that action a have propagated from v to u. Based on this assumption the CD model assigns "credits" to the possible influencers of a node u whenever u performs an action. The sophisticated variant of this model distinguishes between different influenceability of different users by incorporating a *user influenceability function*. It is defined as the fraction of actions that u performs under the influence of at least one of its neighbors (e.g., v) and is learnt from the historical log data. In contrast to our approach, this model suffers from two key limitations. Firstly, it depends on the availability of large amount of historical action logs to compute influence probability as well as user influenceability. Unfortunately, historical action logs may not be available to end-users in many real-world social networks.

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 Dataset

We extracted our blog dataset in September, 2008 using Technorati API[3]. 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 (Leskovec et al. 2007b). 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 does 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 dataset is shown in Table 2. For each blog, the

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

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
$s(g)$	Star-likeness ratio of graph G
$G(c^i)$	Shape of cascade c^i
$\mathcal{K}$	Friendship threshold
$ini(c^i)$	Initiator of the cascade c^i
$I(c^i)$	Initiator-media link of the cascade c^i

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
Cascade size = 2	5674
Cascade size = 3	883
Cascade size > 3	712

posts that do not participate in any cascade are excluded from our dataset. Figure 3a 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[4] with exponent equal to -1.505 whereas in (Leskovec et al. 2007b) 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 gets richer) (Barabasi and Albert 1999).

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 outlined in Algorithm 1.

[4] We adopted the method described in paper Chekuri et al. (2006) for fitting power-law distributions

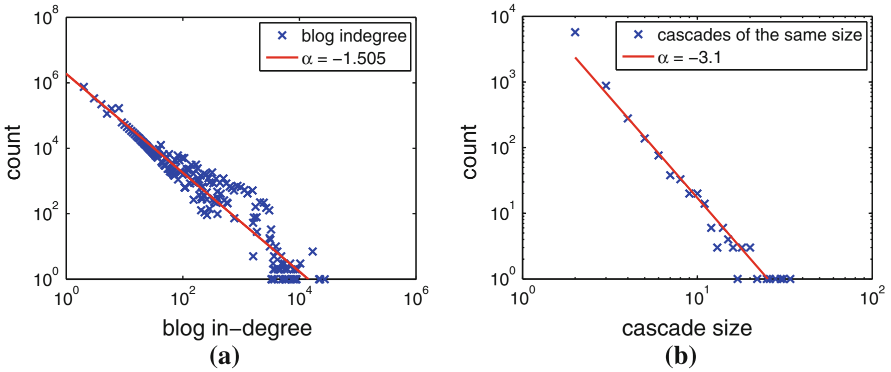


Fig. 3 **a** Blog in-degree distribution. **b** Cascade size distribution

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
```

Note that the proposed cascades extraction procedure is slightly different from the one described in (Leskovec et al. 2007b). Let us elaborate on this further. Consider the scenario in Fig. 4a, depicting blog posts and hyperlinks between them. Based on (Leskovec et al. 2007b), each cascade should have only one initiator (top-most post). Hence, the scenario illustrated in Fig. 4a have to be considered as two different cascades (have two initiators p_1 and p_2) as depicted in Fig. 4b. In contrast, we treat the scenario in Fig. 4a as one cascade. The intuitive justification for this is as follows. Observe that the posts in Fig. 4a 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: $\mathcal{C} = \{c^i \big| T^i \leq T^* - 30\}$. 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

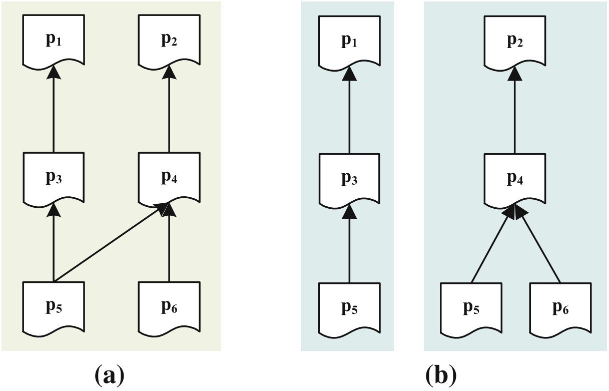


Fig. 4 Different approaches for cascades extraction: **a** observed post-to-post relationship, it is also the cascade identified by our approach; **b** the cascades identified from (**a**) using the approach in (Leskovec et al. 2007b)

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 3b shows the distribution of cascade size extracted from our dataset. It is defined as the number of blogs within a cascade. The X and Y-axes represent different sizes of cascades and the number of cascades, respectively. 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 dataset. The distribution of cascade size also follows a power law. The exponent found in our dataset is -3.1 whereas this exponent is found to be -2 in the dataset used by Leskovec et al. (2007b). This deviation is primarily due to the differences between the characteristics of the two datasets and different definition of a cascade in these two approaches.

4 Macroscopic features

We now present an array of content-oblivious cascade features that may influence a blog's affinity to a cascade. We classify these features into two types, namely *macroscopic* and *microscopic* features. The former refers to features that are associated with the entire cascade whereas the latter refers to features of the blogs or posts of a cascade. In this section, we begin with three macroscopic features, namely *elapsed time*, *number of participants*, and *star-likeness ratio*. In the next section, we shall elaborate on the microscopic features.

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. We use day

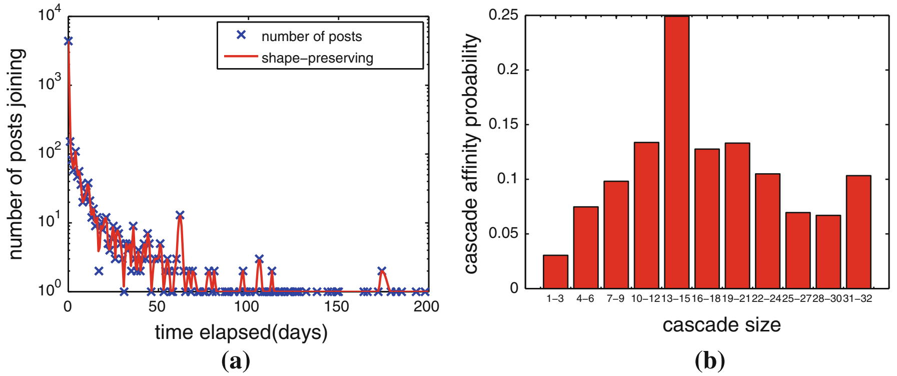


Fig. 5 **a** Number of posts joining versus days elapsed. **b** Cascade affinity probability versus cascade size

as the unit of elapsed time as most bloggers write posts once per day. The distribution of this feature is shown in Fig. 5a. 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 (Leskovec et al. 2008), 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 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. Interestingly, we observed that in only a very small fraction (0.2 %) of cascades the number of posts is more than the number blogs. It indicates that bloggers seldom re-posts in the same cascade such that the number of posts is always the same as the number of blogs in a cascade. Consequently, in the sequel we uniformly use number of blogs to represent the *size* of a cascade. The *number of participants* is formally defined as follows.

Definition 1 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 5b 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 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.

4.3 Star-likeness ratio

Recent results showed that the type of *cascade topology* may indicate the genre of the content in a cascade (McGlohon et al. 2007). We now investigate whether the topology of a cascade influences a blog's affinity to join it. Firstly, we extracted different cascade shapes in the dataset and classified the cascades into 176 different shapes. We observed that the most common shape contains only two posts. The top-13 frequent cascade shapes which appear at least 25 times in the dataset are depicted in Fig. 6. The shapes are listed according to descending order of their frequencies (G_1 is the most frequent cascade shape). Further, the shapes can be classified into two groups, namely *chain* and *star*. Informally, a chain has only one leaf node whereas a star is an n-order shape having $n - 1$ leaves. Notice that in this dataset chains appear more frequently than stars with respect to the same cascade size (i.e., G_2 is more frequent than G_3 and G_{11}). Besides, shapes containing multiple-initiators are less frequent than the single-initiator ones (i.e., G_3 is more frequent than G_{11}). Moreover, cascade frequency does not necessarily decrease with the increase in cascade size (i.e., G_6 is more frequent than G_7).

Secondly, we investigate the probability of blogs to join a cascade by varying the cascade shapes. Formally, the *cascade shape* of c^i is defined as the following.

Definition 2 Let $G_1, \ldots, G_n$ denote the set of cascade shapes extracted from the dataset. If cascade c^i_t (cascade c^i before time t) follows the shape G_s, then the shape of c^i_t, denoted as $g(c^i_t)$, is defined as: $g(c^i_t) = G_s$.

Figure 7a shows the probability of joining a cascade as a function of the cascade shapes. Notice that, the probability of joining a cascade is measured using the temporal shape of each cascade. The probability of joining a cascade of shape G_s can be

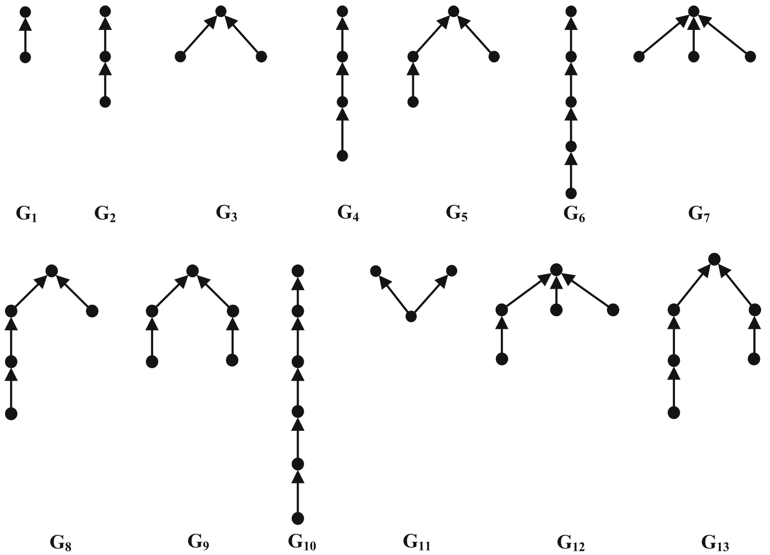


Fig. 6 Common cascade shapes by frequency

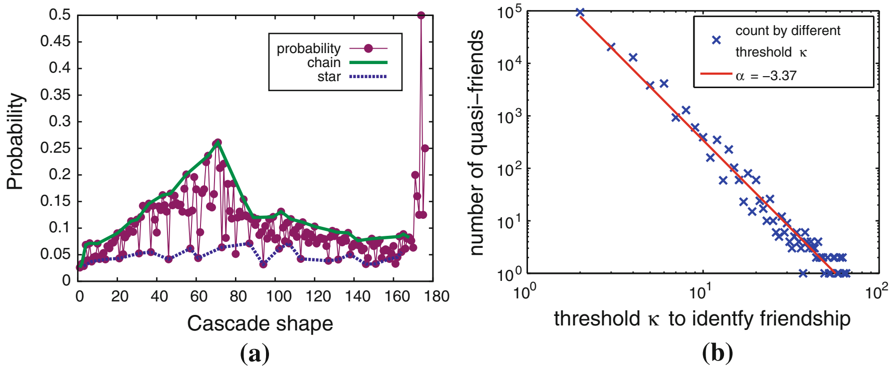


Fig. 7 **a** Joining probability by cascade shape. **b** Number of quasi-friends versus $\mathcal{K}$

computed as follows.

$$Pro_{top}(G_s) = \frac{\sum_{c_t^i} |\{b_j | g(c_t^i) = G_s, b_j \in \phi^i(t)\}|}{\sum_{c_t^i} |\{b_j | g(c_t^i) = G_s\}|}$$

In general, the curve in Fig. 7a is not informative enough to lead to any conclusion on the relationship between cascade shapes and the probability of joining a cascade. However, if we plot the probability curves for star (i.e., G_3, G_7) or chain (i.e., $G_1,\ G_2,\ G_4,\ G_6,\ G_{10}, \ldots$) cascades, then it is clear that chain-shaped cascades are more probable to to attract new blogs to join them compared to their star-shaped counterparts. This phenomenon may be due to the fact that new bloggers can easily see all the blogs within a chain cascade by tracking the hyperlinks one by one.

Table 3 Star-likeness ratio of frequent shapes

	$s(G)$	$Pro_{top}(G)$
G_1	1	0.026
G_2	0.5	0.037
G_3	1	0.029
G_4	0.333	0.069
G_5	0.667	0.040
G_6	0.25	0.072
G_7	1	0.033
G_8	0.5	0.045
G_9	0.5	0.047
G_{10}	0.2	0.071
G_{11}	0.25	0.042
G_{12}	0.75	0.048
G_{13}	0.4	0.053

In contrast, new bloggers may only see a small portion of a star-shaped cascade due to its topology. Also, observe that blogs are much more probable to join a shape with multiple roots (i.e., $Pro_{top}(G_{11}) = 0.042$) compared to other shapes (i.e., $Pro_{top}(G_2) = 0.033$, $Pro_{top}(G_3) = 0.029$) with the same size.

As discussed above, chain cascades are more probable to attract new blogs compared to their star-shaped counterparts. However, shapes other than chain or star are hard to classify and describe. Thus, we propose a new feature called *star-likeness ratio* to describe how much different a shape is from a star and utilize it in the future prediction of cascade affinity.

Definition 3 Let $root(G)$ and $leaf(G)$ denote the root nodes and leaf nodes of graph $G(V, E)$, respectively. That is, $root(G) = \{v | \nexists u \in V, \overrightarrow{vu} \in E\}$ and $leaf(G) = \{v | \nexists u \in V, \overrightarrow{uv} \in E\}$. Then the star-likeness ratio of graph G, denoted as $s(G)$, is defined as:

$$s(G) = \frac{|leaf(G)|/|root(G)|}{|V| - 1}.$$

The ratio falls within the range (0, 1]. A shape with ratio close to 1 indicates that it is closest to star topology. All star-shaped cascades exhibit the same ratio value of 1. For example, consider the frequent shapes in Fig. 6. The star-likeness ratio ($s(G)$) as well as the join probability $Pro_{top}(G)$ of these shapes are listed in Table 3. Observe that the three shapes whose star-likeness ratio are the most (i.e., $s(G) = 1$) exhibit the least joining probability. This suggests that star-likeness ratio can contribute to the analysis of the cascade joining behavior.

5 Microscopic features

In this section, we first investigate four microscopic features of blog cascades that may play important role in cascade affinity prediction, namely *number of quasi-friends*, *popularity of participants*, *citing factor*, and *initiator-media links*. We conclude this section by conducting a one-way variance analysis (ANOVA) on these macroscopic and microscopic features to quantify their significance related to cascade affinity.

5.1 Number of quasi-friends

We introduce the notion of *quasi-friend* to model friendship within blogosphere based on post citings. Unlike other social media platforms (i.e. Facebook, Youtube, Twitter) where users can set up friendship links or select users to listen to, there is no explicit definition on friendship in blogosphere. Thus, we need to find a way to define quasi-friendship in blogosphere. Similar way has been adopted by other researchers (Guice 1995). Formally, *quasi-friend* is defined as follows.

Definition 4 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.

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 Fig. 7b, $\mathcal{K}$ 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 Sect. 7.2. 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 (Backstrom et al. 2006; Cha et al. 2009; Leskovec et al. 2007a). Hence, the number of quasi-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 quasi-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 quasi-friends is highly influenced by the temporal state of the cascade. Let us elaborate on this further. Consider the Fig. 8a. 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 quasi-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 quasi-friend of d as shown in Fig. 8b. We refer to this time when a friendship is created as *friendship creation time*. Observe that the number of quasi-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 after $T_0 + \Delta T$ it may seem that d had a quasi-friend h in this community when she joined it (Fig. 8c). 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 quasi-friends in a social network.

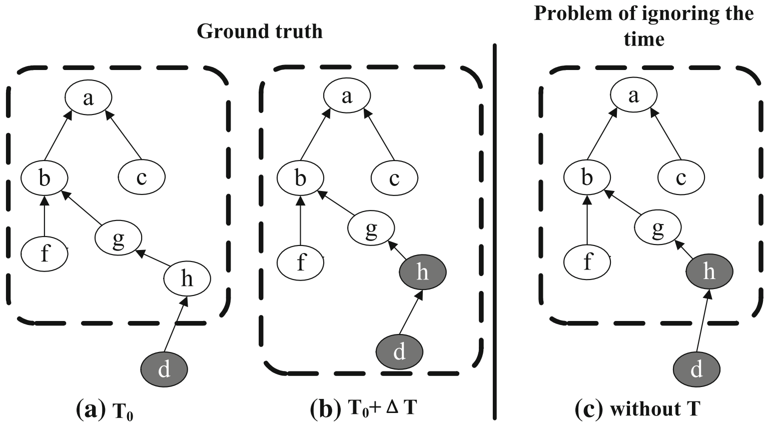


Fig. 8 Effect of friendship creation time

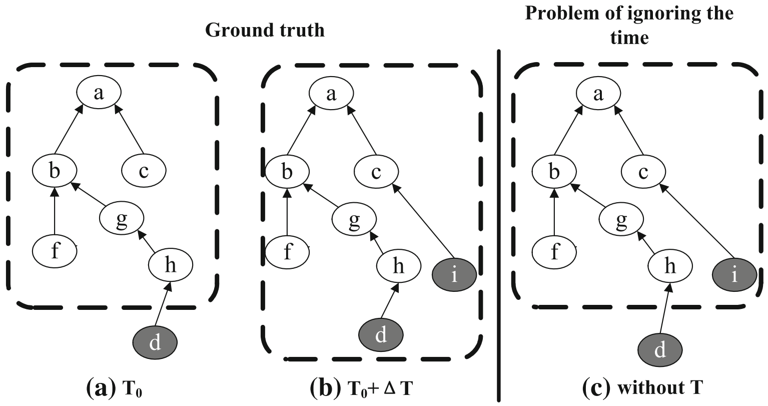


Fig. 9 Effect of friendship creation time (contd.)

There is another problem if we ignore the above temporal behavior. Consider the Fig. 9a, which represents the same scenario as depicted in Fig. 8a. Now assume that another blog i, who is a quasi-friend of d, joined this community at time $T_0 + \Delta T$ as shown in Fig. 9b. If we do not distinguish between times T_0 and $T_0 + \Delta T$, then it may seem that d had a quasi-friend i in the cascade when she joined it (Fig. 9c). However, the truth is that when d joined this cascade at time T_0, she did not have any quasi-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 Sect. 7.2, 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.

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

$F_j(t^i(j))$ 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

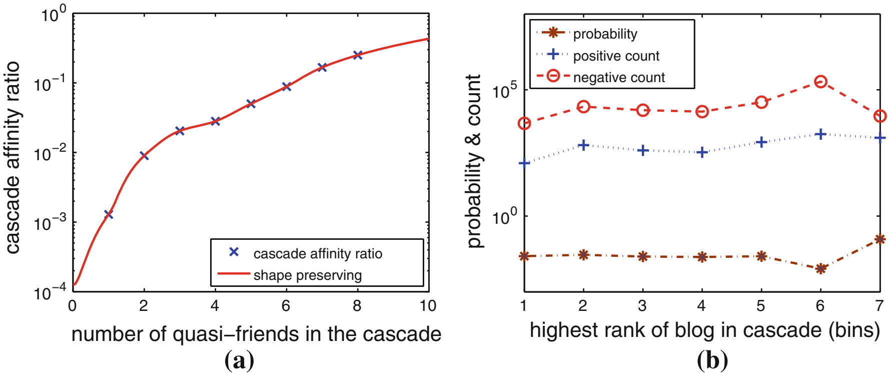


Fig. 10 **a** Cascade affinity ratio versus number of quasi-friends. **b** Joining probability by cascade rank

by incorporating $t^i(j)$ in our approach, we make a contribution to address the above issues (Figs. 8c, 9c). Based on $\Gamma^i(\alpha)$, we define the notion of *cascade affinity ratio* with respect to the number of quasi-friends.

Definition 5 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 Fig. 10a. 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 (Backstrom et al. 2006), and also in (Leskovec et al. 2007a) where the author found a saturation point in the probability of buying a DVD by the number of recommendations received.

5.2 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 (Barabasi and Albert 1999). 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 (Newman 2003). Leskovec et al. (2008) 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 6 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 Fig. 10b. 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.

5.3 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 7 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 Fig. 11a. 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.

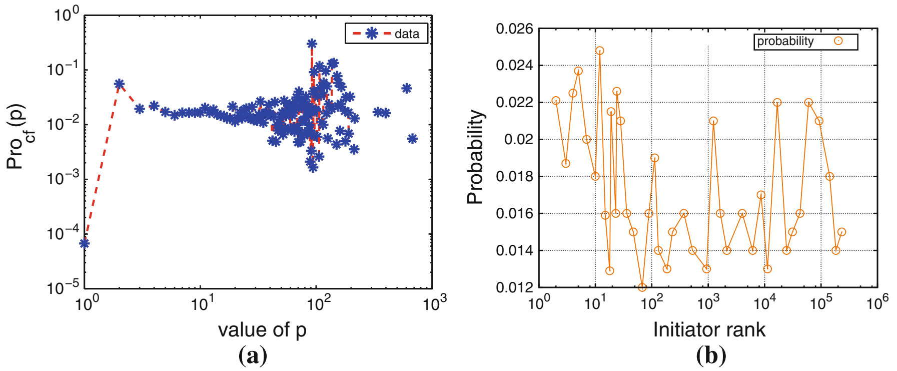


Fig. 11 **a** $Pro_{cf}(p)$ versus number of citations. **b** Joining probability by initiator rank

5.4 Initiator-media link

Lastly, we investigate a feature that is associated with the initiator(s) of a cascade. Recall that the initiator of a cascade is the first blogger who initiates discussion in the cascade. We conduct a series of analysis involving the cascade initiators in order to study whether there is any correlation between the cascade affinity and the cascade initiators. We first formally define *cascade initiator*.

Definition 8 Let $c^i(B, E)$ be a cascade containing blogs $B = \{b_1, b_2, \ldots, b_s\}$ with post–post links E. The initiator $ini(c^i)$ of the cascade c^i is defined as follows

$$ini(c^i) = \{b_j | \nexists b_i \in B, \overrightarrow{b_j b_i} \in E\}.$$

Notice that it is possible to have more than one initiator in a cascade. For example, consider the cascade in Fig. 4a. Both p_1 and p_2 are initiators according to the above definition.

We now investigate two interesting properties of initiators of blog cascades using the Technorati dataset. We extract all the initiators for each cascade. Firstly, we investigated the correlation between the probability of a blog to join the cascade c^i and its *initiators' popularity*. Similar to Sect. 5.2, the *popularity* of initiators of a cascade c^i is the highest *rank* of the initiators in that cascade: $\min_{b \in ini(c^i)}(D(b))$. We plot the probability of a blog to join a cascade by varying the *initiators' popularity* in Fig. 11b. Observe that there does not exist any correlation between these two factors.

Secondly, we examine a property of the initiators related to their out-links to other media resources. Initiators in a cascade are the posts that do not reference any other post in that cascade. However, they may reference non-blog media sources such as Flickr, Youtube, etc. We refer to these links as *initiator-media links*. Formally, it is defined as follows.

Definition 9 Let c^i denote the cascade. Then, the initiator-media link of c^i, denoted as $I(c^i)$, is defined as:

Table 4 Initiator data

Initiators	Count	Avg. size of cascades	Ratio of join
Initiators with initiator-media links	3,782	23.32	0.11 %
Initiators without out-links	4,537	12.51	0.06 %

Table 5 ANOVA test on cascade features

	Feature name	F	p-value
Macro features	Time elapsed	6.88	$\ll 0.001$
	Number of participants	4.36	$\ll 0.001$
	Sar-likeness ratio	1.66	0.024
Micro features	Number of quasi-friends	2.85	0.017
	Popularity of participants	1.50	0.029
	Citing factor	0.77	0.968
	Initiator-media link	1.38	0.034

$$I(c^i) = \begin{cases} 1 \text{ if } \exists b \in ini(c^i),\ \ b \text{ hyperlinks to non-blog URLs} \\ 0 \text{ otherwise.} \end{cases}$$

Table 4 reports the statistics of initiators that link to these media resources against those which do not. Observe that the average size of cascades containing initiator-media links is larger than those which do not have such link. It means that the initiators who reference other media resources are probable to generate larger cascades than the ones that do not. This phenomenon suggests that bloggers are more inclined to write post on a topic when they have found related resources from many different media.

Thus, it indicates that bloggers are more probable to join the cascades whose initiators referenced other media resources. Hence, we propose to use this property as another feature to predict the cascade affinity. For each cascade, we first identify the initiators within it. After that, we test whether the initiators have hyperlinks to web pages that belong to non-blog domains.

5.5 ANOVA Test

In this section, we conduct a one-way variance analysis (ANOVA) on each of the above macroscopic and microscopic features to quantify their significance related to cascade affinity. 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 of each feature is shown in Table 5. 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 six cascade features are all significant for predicting cascade affinity of a blogger.

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 $\Delta^i$ for each cascade $c^i$
begin
    foreach cascade $c^i \in \mathcal{C}$ do
        foreach blog $b_j \in \phi^i$ do
            $\Delta^i(j) = \{r \big| b_j \in F_r(t^i(j))\}$;
            $\Delta^i = \Delta^i \bigcup \Delta^i(j)$;
end
```

6 Cascade affinity prediction

In this section, we describe how the features discussed in previous sections 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.

6.1 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 c and (c^i) that may join c^i is given by the following equation.

$$cand(c^i) = \{j \big| F_j \cap \phi^i \neq \emptyset\} \tag{2}$$

The algorithm for extracting candidate blogs is outlined 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 dataset, there are 312, 414 candidate blogs extracted by Algorithm 2. On average, 43 candidates 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 6 shows 76.1 % blogs that join a cascade have at least a quasi-friend in it when we set $\mathcal{K} = 2$.

6.2 SVM-based cascade affinity prediction

We now present two techniques that exploits the macroscopic and microscopic features to predict blogs that may join a cascade. One takes a *non-probabilistic* approach whereas the other is *probabilistic* in nature. In this subsection, we present the former approach first. We discuss the probabilistic approach in the next subsection. In Sect. 7, we shall empirically compare these two strategies.

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. We adopted a non-probabilistic binary classifier SVMs (Chang and Lin 2001) due to its promising results reported in many data mining/machine learning tasks. SVM models all the samples including both positive and negative ones as points in high dimensional space. The training of SVM learns a hyperplane in the space. The hyperplane learned from SVM model should be able to separate the positive training examples from the negative ones with the largest margin.

Formally, the training data in this paper is represented as a set of points in a 7-dimensional space: $\mathcal{D} = \{(\mathbf{x}_i, y_i)|\mathbf{x}_i \in \mathbb{R}^7, y_i \in \{-1, 1\}\}$. y_i is either 1 or -1, indicating the class to which the point belongs. $\mathbf{x}_i$ is a 7-dimensional vector. Our target is to find a hyperplane that divides the points having $y_i = 1$ from those having $y_i = -1$ with the maximal margin. Actually, any hyperplane can be written as the set of points $\mathbf{x}$ that satisfies the following equation:

$$\mathbf{w} \cdot \mathbf{x} - b = 0$$

where $\mathbf{w}$ is normal vector and perpendicular to the hyperplane. The vector $\mathbf{w}$ and a parameter b to be learned from the training data by minimizing the function:

$$\mathbf{w}^\top \cdot \mathbf{w}$$

subject to

$$y_i(\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1.$$

In order 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. Moreover, all the candidates extracted using Algorithm 2 are formulated as vectors in order to fit in the model. For example, a candidate sample b_j is represented as 7-dimensional vector:

$$\overrightarrow{b_j} = [b_{j1}, b_{j2}, \ldots, b_{j7}]^\top$$

Within the vector, each entry is the value of one of the seven macroscopic and microscopic features discussed in Sects. 4 and 5.

Given the learned model, we compute a score for an unlabeled object b_j using its decision function:

$$f(b_j) = \mathbf{w} \cdot b_j - b.$$

In our setting, a larger $f(b_j)$ indicates more likelihood of b_j joining the target cascade.

6.3 Bipartite Markov Random Field-based (BiMRF) cascade affinity prediction

Other than SVM, we then adopted a probabilistic approach to predict the likelihood of joining a cascade which is based on *Bipartite Markov Random Field* (BiMRF). BiMRF (Shi et al. 2009) models the group joining behavior as a bipartite graph where the vertices at one side of the graph are associated with the variables $B = \{b_i\}_{i=1}^{N}$ which represent users, and the vertices at the other side of the graph are associated with variables $C = \{c^j\}_{j=1}^{M}$ which represent cascades. The advantage of using BiMRF model in this problem is that it can explicitly incorporate the relationship between bloggers and cascades. Moreover, it has been proved to be more effective than other approaches in modeling the group joining behavior (Shi et al. 2009). Similar to SVM-based approach, based on the observed value of each feature, we study the joining behavior at different time step using BiMRF model. In the model, each user is a 7-dimensional feature vector $b_i = [b_{i1}, b_{i2}, \ldots, b_{i7}]^\top$, the values of which may change over time. Let O denote all the observations, including users and their features, cascades and their features as well as the connections of users. Let $E = \{E_{ij}^t : 1 \leq i \leq N, 1 \leq j \leq M$ and $1 \leq t \leq T\}$ be a set of random variables where $e_{ij} = 1$ if the user b_i joins cascade c^j at time t; otherwise it is 0. Let $\{e\}$ denote an instance of E. Then given the observations, BiMRF defines a conditional distribution as follows:

$$p(\{e_t\}|O) = \frac{1}{Z(w)} \exp(\sum_{k=1}^{K} \omega_k f_k(\{e_t\}, O))$$

where f_k are feature functions and ω_k are their weights which will be learned. Given the observed features, $p(\{e\}|O) = \prod_{t=1}^{T} p(\{e\}|O)$.

Formally, the dataset is a pairing of observations and joining behaviors (i.e., $\mathcal{D} = \{\langle\{e\}, O\rangle\}$). The best model to fit the data is the one with the maximum conditional likelihood: $\mathcal{L} = \log p(\{e\}|O)$. We define feature functions to compute the feature values $f_k(\{e\}, O)$ in the above equation. For example, the feature function to compute the feature value of *number of participants* is as follows.

$$f_{nop}(e_{ij}^t = 1, c^i, b_j, t) = N_j(c^i).$$

Thus, the optimized ω_k is learned by maximizing the likelihood:

$$\mathcal{L} = \log p(\{e\}|O).$$

In line with (Shi et al. 2009), the optimization is achieved using L-BFGS (Limited Memory Broyden-Fletcher-Goldfarb-Shanno) algorithm (Liu and Nocedal 1989).

With the weights ω_k learned from the training data, the probability that user b_j joins cascade c^i at time t is given by computing the marginal probability

$$p(e_{ij}^t = 1|O).$$

7 Experiments

7.1 Experimental setting

For both prediction models, we conducted experiments on our data set using 5-fold cross validation to evaluate the effectiveness of the features in predicting cascade affinity of candidate blogs. 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.

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.

7.2 Experimental results

7.2.1 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 6 shows the sizes of candidate blog sets for different $\mathcal{K}$ as well as the highest F_1-measures achieved by selecting the best thresholds in both models. Observe that in both models the best F_1-measures are achieved at $\mathcal{K} = 2$. Hence, in the subsequent experiments we shall set $\mathcal{K} = 2$.

7.2.2 Comparison of feature sets

Firstly, we compare the prediction performance using either macro features or micro features in order to find which groups of features are more important in affinity prediction.

Table 6 Effect of different values of $\mathcal{K}$

Value of $\mathcal{K}$	Candidate size (max. recall)	Highest F_1-measure	
		SVM	BiMRF
$\mathcal{K}$=1	946,329 (0.916)	0.707	0.702
$\mathcal{K}$=2	312,414 (0.761)	0.725	0.711
$\mathcal{K}$=3	80,482 (0.242)	0.227	0.227

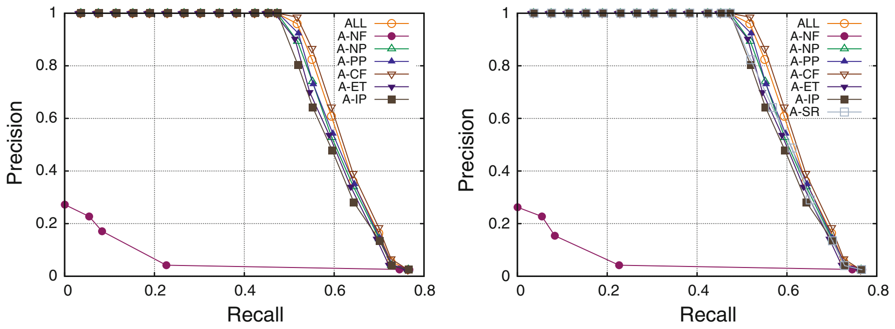
Table 7 Feature set notations and prediction performance in AUC-PR

Features/AUC-PR/feature set	ALL	A-NF	A-PP	A-NP	A-CF	A-ET	A-IP	A-SR
Number of quasi-friends	✓	-	✓	✓	✓	✓	✓	✓
Popularity of participants	✓	✓	-	✓	✓	✓	✓	✓
Number of participants	✓	✓	✓	-	✓	✓	✓	✓
Citing factor	✓	✓	✓	✓	-	✓	✓	✓
Elapsed time	✓	✓	✓	✓	✓	-	✓	✓
Initiator-media link	✓	✓	✓	✓	✓	✓	-	✓
Sar-likeness ratio	✓	✓	✓	✓	✓	✓	✓	-
AUC-PR (SVM)	0.615	0.066	0.588	0.604	0.625	0.604	0.592	0.595
AUC-PR (BiMRF)	0.610	0.055	0.588	0.599	0.618	0.587	0.587	0.589

Secondly, recall that we have identified seven features for cascade affinity prediction, namely *number of quasi-friends*, *popularity of participants*, *number of participants*, *citing factor*, *elapsed time*, *star-likeness ratio* and *initiator-media link*. To evaluate the effectiveness of these features, we conducted 8 sets of experiments. The first set of experiments used all 7 features for prediction. This feature set is denoted by “ALL” in Table 7. In each of the following seven experiments, one feature is removed. For instance, “A-NF” denotes that the feature *number of quasi-friends* is removed and the remaining four features were used for prediction. In Table 7, a ‘✓’ indicates that the feature is used and ‘-’ otherwise.

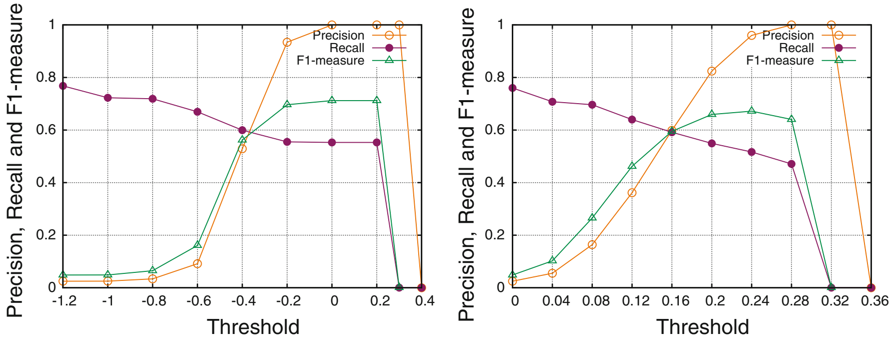
The prediction performances measured by AUC-PR are reported in the last two rows in Table 7. Using all the seven features, the prediction achieved AUC-PR of 0.615 for SVM and 0.610 for BiMRF. We can make the following observations:

- Removal of *number of quasi-friends* resulted in significant drop in prediction performance to 0.046 in both models indicating that *number of quasi-friends* is the most important factor that affects a blogger’s cascade affinity. We validated this by performing another experiment using only the *number of quasi-friends* as feature. The AUC-PR in this case is 0.445 for SVM and 0.434 for BiMRF.
- Removal of either *popularity of participants*, *number of participates*, *star-likeness ratio*, *elapsed time* or *initiator-media link* led to a small performance degradation. These five features indeed contributed to the cascade affinity modeling.
- An interesting observation is that removal of the *citing factor* in both models led to better AUC-PR than using all the seven features. This result clearly indicates that the *citing factor* introduced noise in the prediction, which is consistent with our ANOVA test results reported in Sect. 5.5. The remaining six features: *number*



(a) SVM **(b)** BiMRF

Fig. 12 Precision-Recall Curves for different features



(a) SVM **(b)** BiMRF

Fig. 13 Precision, Recall and $F1$-measure for “A-CF”

of quasi-friends, *popularity of participants*, *number of participates*, *elapsed time*, *star-likeness ratio* and *initiator-media link* achieved the best performance.

- We conducted additional two experiments using only macroscopic or microscopic features. Microscopic features achieved AUC-PR of 0.533 for SVM (0.519 for BiMRF) whereas macroscopic ones only exhibit AUC-PR of 0.064 for SVM (0.059 for BiMRF). This is because the *number of quasi-friends* is one of the microscopic features.

For the completeness of the results, Figure 12 plots the Precision-Recall curves of using eight different feature sets. Under SVM model, all the seven 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. In contrast, all the seven runs of BiMRF model (except for “A-NF”) achieved almost perfect precision before recall reached 0.48. As the recall increases, precision in BiMRF drops down more smoothly than that of SVM model. However, SVM and BiMRF show almost the same AUC-PR. Thus, both models can be applied in cascade affinity prediction.

Figure 13 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. Both precision and recall of BiMRF model change more smoothly than

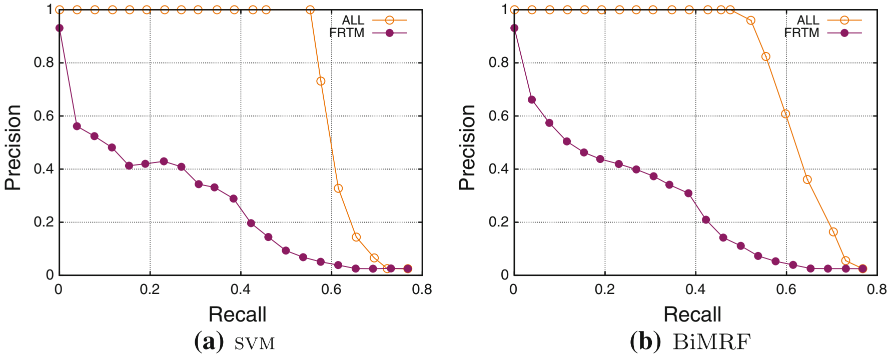


Fig. 14 Significance of time for modeling number of quasi-friends

those of SVM model as the thresholds increase, indicating that SVM model results in a clearer margin between the score of positive samples and negative ones.

7.2.3 Significance of time for modeling number of quasi-friends

Recall that in Sect. 5.1, we illustrated the significance of time in modeling the number of quasi-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 quasi-friends in cascade c^i) is modified as follows.

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

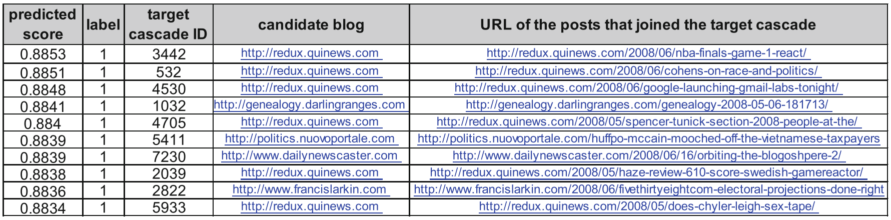
We updated the *number of quasi-friends* feature in each candidate vector using the above formula. Using the updated feature vectors, we performed the prediction again. The performance of ignoring the time in *quasi-friend* identification shows a small AUC-PR 0.211 (0.216 for BiMRF) whereas our proposed solution achieves 0.615 (0.610 for BiMRF). The comparison between the Precision-Recall curve of this approach and our proposed solution is shown in Fig. 14. Both of the curves use all the seven features. "FRTM" represents the approach that discards the temporal aspects in *number of quasi-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. In fact, whenever K is set to 1, 2, or 3, the experiment result indicates that distinguishing joining time and friend creation time is important, we select not to show the detailed statistics on K=1 and 3 to avoid making the paper to lengthy and tedious.

7.2.4 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

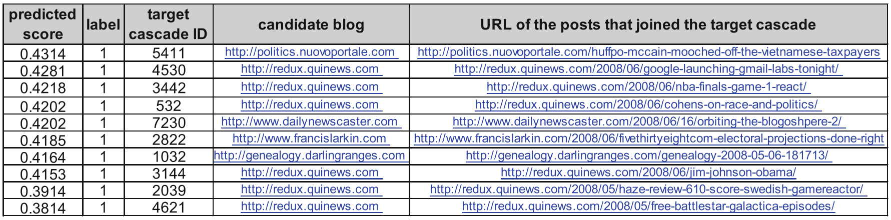
Table 8 Average precision versus top-k

k	1	2	3	4	5
$Pr_{avg}(k)$ (SVM)	0.970	0.783	0.722	0.734	0.763
$Pr_{avg}(k)$ (BiMRF)	0.970	0.762	0.716	0.722	0.744



predicted score	label	target cascade ID	candidate blog	URL of the posts that joined the target cascade
0.8853	1	3442	http://redux.quinews.com	http://redux.quinews.com/2008/06/nba-finals-game-1-react/
0.8851	1	532	http://redux.quinews.com	http://redux.quinews.com/2008/06/cohens-on-race-and-politics/
0.8848	1	4530	http://redux.quinews.com	http://redux.quinews.com/2008/06/google-launching-gmail-labs-tonight/
0.8841	1	1032	http://genealogy.darlingranges.com	http://genealogy.darlingranges.com/genealogy-2008-05-06-181713/
0.884	1	4705	http://redux.quinews.com	http://redux.quinews.com/2008/05/spencer-tunick-section-2008-people-at-the/
0.8839	1	5411	http://politics.nuovoportale.com	http://politics.nuovoportale.com/huffpo-mccain-mooched-off-the-vietnamese-taxpayers
0.8839	1	7230	http://www.dailynewscaster.com	http://www.dailynewscaster.com/2008/06/16/orbiting-the-blogoshpere-2/
0.8838	1	2039	http://redux.quinews.com	http://redux.quinews.com/2008/05/haze-review-610-score-swedish-gamereactor/
0.8836	1	2822	http://www.francislarkin.com	http://www.francislarkin.com/2008/06/fivethirtyeightcom-electoral-projections-done-right
0.8834	1	5933	http://redux.quinews.com	http://redux.quinews.com/2008/05/does-chyler-leigh-sex-tape/

Fig. 15 Top 10 candidates that are most probable to join a cascade (SVM)



predicted score	label	target cascade ID	candidate blog	URL of the posts that joined the target cascade
0.4314	1	5411	http://politics.nuovoportale.com	http://politics.nuovoportale.com/huffpo-mccain-mooched-off-the-vietnamese-taxpayers
0.4281	1	4530	http://redux.quinews.com	http://redux.quinews.com/2008/06/google-launching-gmail-labs-tonight/
0.4218	1	3442	http://redux.quinews.com	http://redux.quinews.com/2008/06/nba-finals-game-1-react/
0.4202	1	532	http://redux.quinews.com	http://redux.quinews.com/2008/06/cohens-on-race-and-politics/
0.4202	1	7230	http://www.dailynewscaster.com	http://www.dailynewscaster.com/2008/06/16/orbiting-the-blogoshpere-2/
0.4185	1	2822	http://www.francislarkin.com	http://www.francislarkin.com/2008/06/fivethirtyeightcom-electoral-projections-done-right
0.4164	1	1032	http://genealogy.darlingranges.com	http://genealogy.darlingranges.com/genealogy-2008-05-06-181713/
0.4153	1	3144	http://redux.quinews.com	http://redux.quinews.com/2008/06/jim-johnson-obama/
0.3914	1	2039	http://redux.quinews.com	http://redux.quinews.com/2008/05/haze-review-610-score-swedish-gamereactor/
0.3814	1	4621	http://redux.quinews.com	http://redux.quinews.com/2008/05/free-battlestar-galactica-episodes/

Fig. 16 Top 10 candidates that are most probable to join a cascade (BiMRF)

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.

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

Table 8 shows average precision values for different k values highlighting the goodness of our approach. Note that $Pr_{avg}(k)$ may not monotonically decrease with increasing k as the number of cascades in the denominator depends on k.

Figures 15 and 16 show 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.

7.2.5 *Using only the number of quasi-friends as feature*

As mentioned above, we found that *number of quasi-friends* is the most important factor that affects a blogger's cascade affinity. To further validate this, we conducted

Table 9 Average precision versus top-k using only *number of quasi-friends*

k	1	2	3	4	5
$Pr_{avg}(k)$ (SVM)	0.643	0.561	0.597	0.571	0.573
$Pr_{avg}(k)$ (BiMRF)	0.612	0.515	0.535	0.564	0.558

another experiment using only the *number of quasi-friends* as feature. The AUC-PR in this case is 0.445 for SVM and 0.434 for BiMRF.

In fact, just as we did in Figs. 15 and 16, if we list the top-10 candidate blogs over the entire collection for the approach of using only the *number of quasi-friends* in SVM or BiMRF, we can also get 10 positive samples which eventually join the target cascades. However, when we compute the average precision values for for different k values as what we did in Table 8, the approach of using only the *number of quasi-friends* shows poor result which is shown in Table 9. It indicates that using only the *number of quasi-friends* does not perform so well as the other approaches combining several informational features. It can only found very limited number of candidate blogs which are most possible to join target cascades over the entire collection but fails to give enough precision in order to predict the affinity in each individual cascade.

7.2.6 *Comparison of the performance between SVM and BiMRF model*

Referring to the prediction results shown in Table 7 and Fig. 13, we compare the performance between the SVM and BiMRF models. The observations are as follows.

- In both models, removal of the *citing factor* perform the best, then is the approach using all the features. However, the removal of either *elapsed time* or *initiator-media link* does not perform better than the removal of *popularity of participants* in BiMRF, which is not the case in SVM.
- In general, both SVM and BiMRF models show satisfactory prediction results with AUC-PR over 0.61 when using the best feature set “A-CF”. However, SVM performs slightly better than BiMRF model for all the other feature sets.
- The precision and recall curves by varying the threshold of both SVM and BiMRF model almost exhibit the same shape, except that the curves of BiMRF model is smoother than those of SVM. It suggests that SVM model tends to produce clearer margin between positive class and negative one.
- Both SVM and BiMRF models exhibit the best average precision with 0.970 when predicting the top-1 blogger according to Table 8. In addition, the average precision of top-k ($k = 2, \ldots, 5$) of both methods do not vary much. Moreover, 8 out of 10 records in Fig. 16 also appear in Fig. 15, indicating that the results of both models do not vary much.

8 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 seven macroscopic and microscopic features, namely number of quasi-friends, popularity of participants, number of participants, time elapsed since the genesis of the cascade, star-likeness ratio, initiator-media link 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 a non-probabilistic (SVM-based) and probabilistic (BiMRF classifier) methods are employed. Using the prediction scores from the SVM-based approach or the conditional probability from BiMRF, 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. In general, SVM model performs slightly better than BiMRF model in most of the feature sets. Moreover, SVM model tends to generate clearer margin between positive class and negative one. The six features that have significant impact on cascade affinity achieved the best prediction accuracy of 0.625 (0.618 for BiMRF) measured by AUC- PR. Our experimental results also showed that the number of quasi-friends plays a significant role in blog cascade affinity prediction. The features proposed in this paper may not be independent to each other (i.e., *number of participants* and *number of quasi-friends*). As part of future work, we intend to investigate the correlation between different features and how they influence the cascade affinity.

On the other hand, we intend to study micro-blogging behavior using the model and results proposed in this paper. However, the model we proposed in this paper may not be applied directly in micro-blogging scenario, as there exist several differences, which are shown as follows, between retweeting in micro-blog and blog cascade. Firstly, instead of web pages, all micro-blog posts are kept in the server by some operator (i.e., Twitter). An arbitrary blogger can view any blog posts that is published on the Web. However, an arbitrary micro-blogger can only see the content posted by those whom he is listening to. Secondly, micro-blog posts are platform dependent in that a twitter user can never re-tweet a post that is published at another platform such as Tumblr. Thirdly, there is only one initiator in microblogging which is different with blog cascade. Lastly but not the least, collecting data for blog posts study can be done by crawling the static web pages. However, we can only get tweets by querying the database of Twitter through the API provided by Twitter. Twitter API can only provide the tweets that appeared within 2 weeks' time. Thus, in order to apply out model towards micro-blog domain, definition of some features may be changed, such as quasi-friends, elapsed time, star-likeness ratio, popularity of participants. Moreover, some other features which are specific in micro-blog need to be taken into consider. In fact, this is what we intend to do in future work.

Acknowledgments Part of the work was done when the first author was pursuing PhD in School of Computer Engineering, Nanyang Technological University, Singapore. This work is partly supported by NSFC 61202179, 61173089

References

Adams B, Phung DQ, Venkatesh S (2010) Discovery of latent subcommunities in a blog's readership. TWEB 4(3):12:1–12:30

Agarwal N, Liu H, Tang L, Yu PS (2008) Identifying the influential bloggers in a community. In: WSDM '08: Proceedings of the 1st ACM international conference on web search and data mining, pp 207–218

Backstrom L, Huttenlocher DP, Kleinberg JM, Lan X (2006) Group formation in large social networks: membership, growth, and evolution. In: KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 44–54

Bao H, Chang EY (2010) Adheat: an influence-based diffusion model for propagating hints to match ads. In: WWW '10: Proceedings of the 19th international conference on, World wide web, pp 71–80

Barabasi AL, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512

Bikhchandani S, Hirshleifer D, Welch I (1992) A theory of fads, fashion, custom, and cultural change as informational cascades. J Political Econ 100(5):992–1026

Cha M, Mislove A, Gummadi PK (2009) A measurement-driven analysis of information propagation in the flickr social network. In: WWW '09: Proceedings of the 18th international conference on, World wide web, pp 721–730

Chang CC, Lin CJ (2001) LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm. Accessed 10 Feb 2013

Chekuri C, Even G, Kortasrz G (2006) A greedy approximation algorithm for the group steiner problem. Discret Appl Math 154(1):15–34

Chen H, Tiño P, Yao X (2009b) Predictive ensemble pruning by expectation propagation. IEEE Trans Knowl Data Eng 21(7):999–1013

Chen D, Tang J, Li J, Zhou L (2009a) Discovering the staring people from social networks. In: WWW '09: Proceedings of the 18th international conference on, World wide web, pp 1219–1220

Clements M, De Vries AP, Reinders MJT (2010) The task-dependent effect of tags and ratings on social media access. ACM Trans Inf Syst 28:21:1–21:42

Davidson I, Gilpin S, Walker PB (2012) Behavioral event data and their analysis. Data Min Knowl Discov 25(3):635–653

Dodds PS, Watts DJ (2004) Universal behavior in a generalized model of contagion. Phys Rev Lett 92(21):218, 701+

Goyal A, Bonchi F, Lakshmanan Laks VS (2012) A data-based approach to social influence maximization. PVLDB 5(1):73–84

Gruhl D, Guha RV, Liben-Nowell D, Tomkins A (2004) Information diffusion through blogspace. In: WWW '04: Proceedings of the 13th international conference on, World wide web, pp 491–501

Guice SL (1995) Creating Communities of Readers: A Study of Children's Information Networks as Multiple Contexts for Responding to Texts. Journal of Literacy Research 27(3):379–397

Hartline JD, Mirrokni VS, Sundararajan M (2008) Optimal marketing strategies over social networks. In: WWW '08: Proceedings of the 17th international conference on, World wide web, pp 189–198

Iribarren JL, Moro E (2009) Impact of human activity patterns on the dynamics of information diffusion. Phys Rev Lett 103(3):038, 702+

Karagiannis T, Vojnovic M (2009) Behavioral profiles for advanced email features. In: WWW '09: Proceedings of the 18th international conference on, World wide web, pp 711–720

Kempe D, Kleinberg JM, Tardos É (2003) Maximizing the spread of influence through a social network. In: KDD '03: Proceedings of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 137–146

Kimura M, Saito K, Motoda H (2009) Blocking links to minimize contamination spread in a social network. ACM Trans Knowl Discov Data 3:9:1–9:23

Kumar R, Novak J, Raghavan P, Tomkins A (2003) On the bursty evolution of blogspace. In: WWW '03: Proceedings of the 12th international conference on, World wide web, pp 568–576

Lee C, Kwak H, Park H, Moon SB (2010) Finding influentials based on the temporal order of information adoption in twitter. In: WWW '10: Proceedings of the 19th international conference on, World wide web, pp 1137–1138

Lerman K, Hogg T (2010) Using a model of social dynamics to predict popularity of news. In: WWW '10: Proceedings of the 19th international conference on World wide web, ACM, New York, NY, USA, WWW '10, pp 621–630

Leskovec J, Adamic LA, Huberman BA (2006) The dynamics of viral marketing. In: EC '06: Proceedings of the 7th ACM conference on Electronic commerce, ACM, New York, NY, USA, pp 228–237

Leskovec J, Adamic LA, Huberman BA (2007a) The dynamics of viral marketing. TWEB 1(1): Article 5. doi:10.1145/1232722.1232727

Leskovec J, Backstrom L, Kumar R, Tomkins A (2008) Microscopic evolution of social networks. In: KDD '08: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 462–470

Leskovec J, McGlohon M, Faloutsos C, Glance N, Hurst M (2007b) Cascading behavior in large blog graphs: Patterns and a model. In: SDM '07: Society of Applied and Industrial Mathematics: Data Mining

Li H, Bhowmick SS, Sun A (2009) Blog cascade affinity: analysis and prediction. In: CIKM' 09: Proceeding of the 18th ACM conference on Information and knowledge management, ACM, New York, NY, USA, CIKM '09, pp 1117–1126

Liu DC, Nocedal J (1989) On the limited memory BFGS method for large scale optimization. Math Program 45(3):503–528

Ma H, Yang H, Lyu MR, King I (2008) Mining social networks using heat diffusion processes for marketing candidates selection. In: CIKM '08: Proceeding of the 17th ACM conference on Information and, knowledge management, pp 233–242

McGlohon M, Leskovec J, Faloutsos C, Hurst M, Glance N (2007) Finding patterns in blog shapes and blog evolution. In: International Conference on Weblogs and Social Media, Boulder, Colo

Newman MEJ (2002) Spread of epidemic disease on networks. Phys Rev E 66(1):016, 128+

Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45:167–256

Pal A, Counts S (2011) Identifying topical authorities in microblogs. In: WSDM '11: Proceedings of the Forth International Conference on Web Search and Web Data Mining, ACM, New York, NY, USA, pp 45–54

Pastor-Satorras R, Vespignani A (2002) Epidemics and immunization in scale-free networks. ArXiv Condensed Matter e-prints/0205260

Rogers EM (2003) Diffusion of innovations, 5th edn. Free Press, New York

Satorras RP, Vespignani A (2001) Epidemic spreading in scale-free networks. Phys Rev Lett 86(14): 3200–3203

Shi X, Zhu J, Cai R, Zhang L (2009) User grouping behavior in online forums. In: KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, NY, USA, pp 777–786

Stewart A, Chen L, Paiu R, Nejdl W (2007) Discovering information diffusion paths from blogosphere for online advertising. In: ADKDD '07: Proceedings of the 1st international workshop on Data mining and audience intelligence for advertising, ACM, New York, NY, USA, pp 46–54

Strang D, Soule S (1998) Diffusion in organizations and social movements: from hybrid corn to poison pills. Annu Rev Sociol 24:265–290

Technorati (2008) State of the blogosphere. Tech Rep http://www.technorati.com/blogging/state-of-the-blogosphere/. Accessed 3 Mar 2010

Wang Y, Chakrabarti D, Wang C, Faloutsos C (2003) Epidemic spreading in real networks: An eigenvalue viewpoint. IEEE Symposium on Reliable Distributed Systems 0:25+

Wang Y, Cong G, Song G, Xie K (2010) Community-based greedy algorithm for mining top-K influential nodes in mobile social networks. In: KDD '10: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, NY, USA, pp 1039–1048

Watts D (2002) A simple model of global cascades on random networks. P Natl Acad Sci USA 99(9):5766–5771

Watts DJ, Dodds PS (2007) Influentials, networks, and public opinion formation. J Consumer Res 34: 441–458