Information Cascades in the Blogosphere: A Look Behind the Curtain

Manos Papagelis, Nilesh Bansal, Nick Koudas
University of Toronto
Canada
(papaggel, nilesh, koudas)@cs.toronto.edu

Abstract

With an increasing number of people that read, write and comment on blogs, the blogosphere has established itself as an essential medium of communication. A fundamental characteristic of the blogging activity is that bloggers often link to each other. The succession of linking behavior determines the way in which information propagates in the blogosphere, forming cascades. Analyzing cascades can be useful in various applications, such as providing insight of public opinion on various topics and developing better cascade models.

This paper presents the results of an excessive study on cascading behavior in the blogosphere. Our objective is to present trends on the degree of engagement and reaction of bloggers in stories that become available in blogs under various parameters and constraints. To this end, we analyze cascades that are attributed to different population groups constrained by factors of gender, age, and continent. We also analyze how cascades differentiate depending on their subject. Our analysis is performed on one of the largest available datasets, including 30M active blogs and 700M posts. The study reveals large variations in the properties of cascades.

Introduction

Blogging activity has been proliferating over the last years. An important characteristic of blogging is that blog posts may link to posts of other bloggers leading to a discussion. The succession of the linking process determines the way in which information propagates in the blogosphere forming *information cascades*. Analyzing such cascades can be useful in various domains, such as providing insight on public opinion on a variety of topics (Gruhl *et al.* 2005) or developing better cascade models (Leskovec *et al.* 2007).

This paper presents the results of an excessive study on cascading behavior in the blogosphere. The study collected and analyzed information of cascading behavior that is available in *BlogScope* (Bansal & Koudas 2007). BlogScope is an analysis and visualization tool for the blogosphere and is currently tracking over 30 million active blogs with almost 700 million posts making it one of the largest available datasets for our analysis. Each blog in BlogScope is associated with a *blog profile*, which usually includes information on the gender, age, and location of the author.

Previous work on cascading behavior in blogs has assumed spreading of information regardless of the context of the posts (e.g., its topic), and highlighted the need for further analysis that would be useful to the development of more accurate patterns of information propagation (Leskovec *et al.* 2007). Our approach is to analyze cascades that are attributed to different population groups constrained by factors of *gender*, *age*, and *continent*. Then, we analyze how cascades differentiate in five subjects: *technology*, *politics*, *financial*, *sports*, and *entertainment*. In each case, we report on the degree of engagement, the reaction times, and the structural properties of the cascades.

This study is the first attempt to gain insight of the cascading behavior in blogs under various parameters and constraints, and contributes to a better understanding of the linking activity in the blogosphere. Such insight would be useful to the development of better models of cascading behavior, better ranking measures, better predictive models of the spread of ideas and behaviors, as well as in applications that make use of the diffusion process.

Methodology

This section describes the methodology we follow to compute cascades. First, we introduce terminology and notation. Then, we present the datasets employed in the analysis and how they are collected. Finally, we present the observation measures on which cascades are compared.

Preliminaries

Let $U(B, P)$ represent the blogosphere, where B is the set of all blogs and P the set of all posts. Each blog $b \in B$ consists of a set of posts $P^b \subseteq P$ (Figure 1(a)). Also assume that each post $p \in P$ is associated with a unique timestamp t_p that corresponds to the time of its submission, allowing for a total temporal ordering of the posts in P. Further, let $\ell_{p_y \to p_x}$ represent a link from a post p_y in the future to a post p_x in the past. For each link $\ell_{p_y \to p_x}$ we also define $\Delta_{p_x}^{p_y}$ to be the difference in the submission times of post p_y and p_x. Note that $\Delta_{p_x}^{p_y} = t_{p_y} - t_{p_x} > 0$.

Now, let $G(P, L)$ be a graph where P is the set of all posts and L is the set of all links between posts. We call this graph the *post graph* (Figure 1(b)). Let $\hat{L}$ represent the set of all links in L but with reversed direction. Let also the graph $\hat{G}(P, \hat{L})$. A cascade $C(P^C, L^C)$ is an induced graph of the graph $\hat{G}$ where $P^C \subseteq P$ and $L^C \subseteq \hat{L}$ (Figure 1(c)).

A cascade C can be thought of as a directed graph, where nodes represent posts and edges represent information flow between posts. Note that the direction of an edge in C follows the information propagation (the actual permalink follows the opposite direction). We denote the in-degree of a node $\nu \in C$ as $deg^-(\nu)$ and its out-degree as $deg^+(\nu)$. A node $\nu \in C$ with $deg^-(\nu) = 0$ is called a *source* and a node $\nu \in C$ with $deg^+(\nu) = 0$ is called a *sink*. A cascade C has only one source, which represents the *initiator post* p_i. Any node $\nu \in C$ with $deg^-(\nu) > 1$ is called a *connector node* and represents a post that has permalinks to at least two posts in the past, and therefore connects branches of different cascades. Throughout the study we assume that a connector node participates (i.e., it is re-evaluated) in all the cascades that it connects. For each post p in the cascade C we define its reaction time R_p as the difference in the submission times of p and the initiator post p_i (i.e., $R_p = \Delta^p_{p_i}$). We define the following properties for a cascade C:

- *cascade size* (C_s): The number of nodes in C, excluding the initiator post p_i.
- *cascade height* (C_h): The height of the spanning tree obtained by traversing the cascade graph C using a *depth-first search* (DFS) algorithm. The algorithm starts at the source and at each step visits adjacent nodes giving priority to the node whose the post has smaller timestamp. The algorithm remembers previously visited nodes and will not revisit them (Figure 1(d)).
- *minimum reaction time* (R_{min}): The minimum reaction time of all posts in the cascade (excluding p_i).
- *mean reaction time* (R_{mean}): The mean reaction time of all posts in the cascade (excluding p_i).
- *maximum reaction time* (R_{max}): The maximum reaction time of all posts in the cascade (excluding p_i).

We consider a cascade C with $C_s = 0$ and $C_h = 0$ to be a *trivial cascade* (i.e., a single post). A *non-trivial cascade* C has $C_s \geq 1$, $C_h \geq 1$ and all links obey time order (i.e., $\Delta > 0$).

Dataset Description

In order to analyze the cascading behavior of posts under various parameters and constraints we abide by the following method. First, we formulate a sample dataset consisting of a set of posts satisfying the required specifications. These posts serve as the initiator posts for the analysis. Then, we retrieve the cascades triggered by these posts by monitoring the blogosphere (i.e., approx. 30M blogs with 400M posts) for a specific time frame. The monitoring refers to the process of searching and retrieving all posts that have back links to the initiator post in the specified time frame. For each of the retrieved posts, we continue monitoring the blogosphere looking again for backlinks. This describes a recursive process that eventually computes a cascade (trivial or not-trivial) for each of the initiator posts. The recursion stops when there are not any backlinks to any of the intermediately retrieved posts. Note that all posts are allocated the same time frame to evolve, therefore there is no truncation occurring for posts later in the cascade. For the scope of our analysis we employed two sample datasets; one for analysis of blog profiles and one for analysis of different subjects. Following we present details on each of the sample datasets.

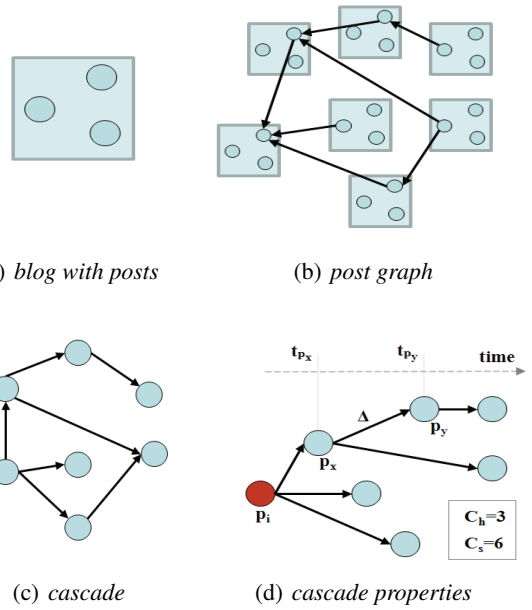


Figure 1: Example Cascade

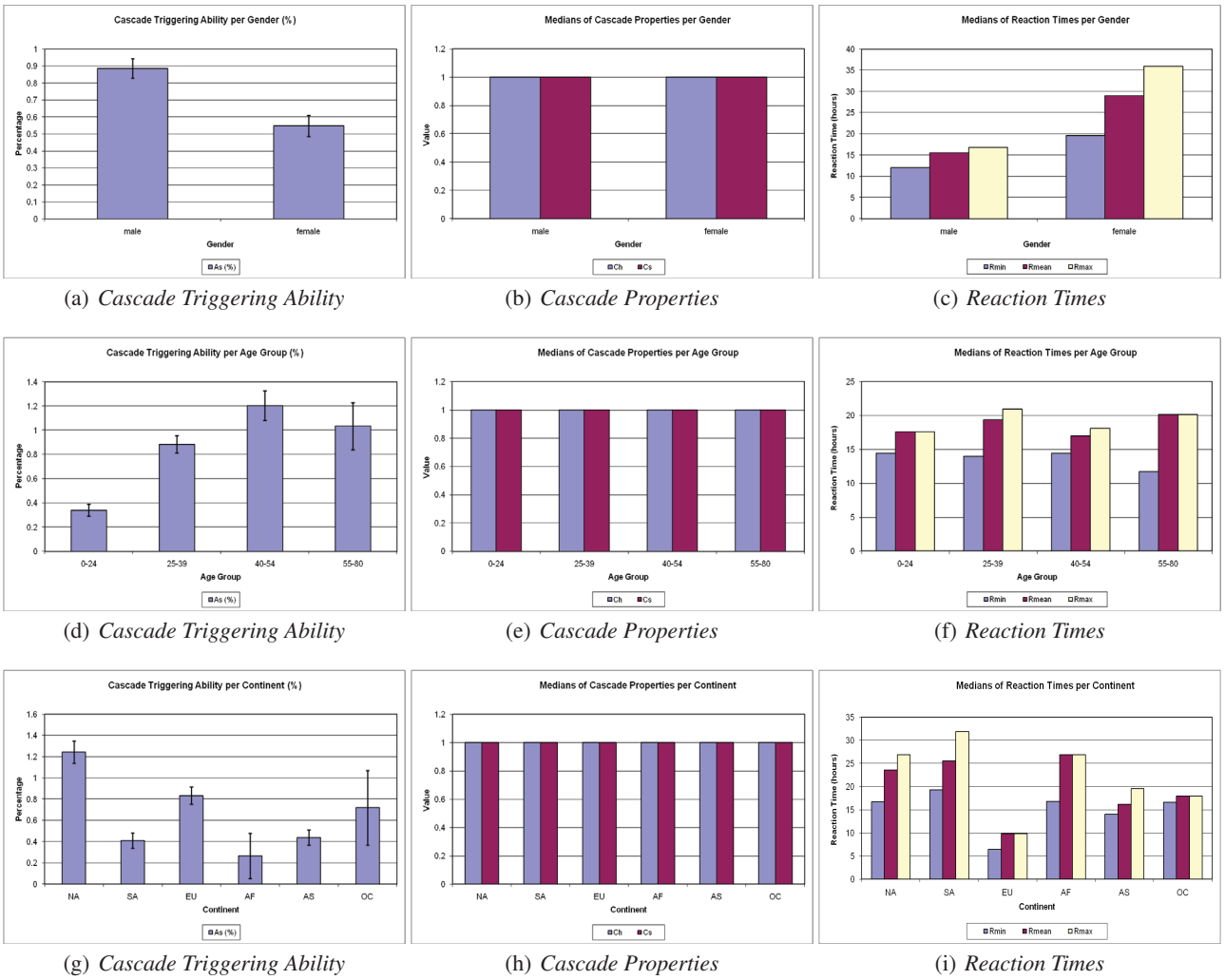
Sample of Posts With Complete Profile The study considers 160k posts submitted by 5000 blogs in the period starting on 01-Jun-2008 and ending on 10-Jun-2008. These blogs have *complete profile* information (i.e., gender, age, and country is defined). For each post we monitor the blogosphere looking for cascades in a 30-days time frame starting at the post's submission day. We had a bias in favor of blogs with large number of posts. The final set consisted of blogs that had at least 17 posts in these 10 days with the most active ones having hundreds of posts. The average number of posts per blog in the dataset was 32.

Sample of Posts in Different Subjects The study considers around 6000 posts distributed in five subjects: *technology*, *politics*, *financial*, *sports*, and *entertainment*. For each subject we manually collect a number of representative blogs. Then, for each blog we obtain a set of posts submitted between 01-Jun-2008 and 10-Jun-2008. For each post we monitor the blogosphere looking for cascades in a 30-days time frame starting at the post's submission day. We had a bias in favor of more authoritative blogs taking into account the number of inlinks to a blog in the last year. The higher the number, the more authoritative a blog is.

Observation Measures

We present measures that characterize the cascading behavior of a set of posts S. Let $S_C \subseteq S$ represent the set of posts in S that were able to trigger non-trivial cascades. Formally, we define the *cascade triggering ability* A_S of a set of posts S to be the ratio of $|S_C|$ over $|S|$: $A_S = \frac{|S_C|}{|S|}$.

Each post in S_C corresponds to an initiator post and forms a non-trivial cascade with properties of size, height, as well as minimum, mean and maximum reaction time (C_s, C_h, R_{min}, R_{mean}, R_{max} respectively). Since these properties typically have very skewed distributions, we report in our



(a) *Cascade Triggering Ability* (b) *Cascade Properties* (c) *Reaction Times*

(d) *Cascade Triggering Ability* (e) *Cascade Properties* (f) *Reaction Times*

(g) *Cascade Triggering Ability* (h) *Cascade Properties* (i) *Reaction Times*

Figure 2: Cascades per Gender (a, b, c), Age Group (d, e, f), and Continent (g, h, i)

analysis the *medians* of their distributions in order to characterize the behavior of a set of cascades. (i.e., the median of the cascade sizes, the median of the cascade heights, the median of the minimum, the mean, and the maximum reaction times.). Throughout our analysis, we analyse cascading behavior based on these observation measures.

Analysis

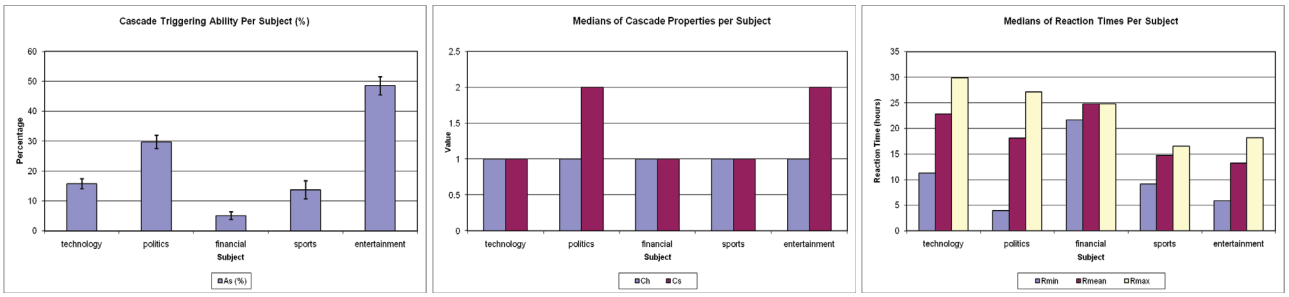
The raw data obtained from the cascades requires an exhaustive analysis, in order to reject potential wrong cascades and avoid the existence of possible deviations in results. Briefly, the analysis consists of the detection of unusual observations (outliers) from the initial cascades and elimination of particular types of cascades (i.e., self-links, invalid links, spam, duplicates). Once the wrong cascades are rejected, the evaluation measures are computed on the clean dataset. When we report on the cascading behavior ability A_s of a sample, the confidence intervals at the 95% confidence level are also stated (error bars on graphs). As aforementioned, medians are reported for the rest of the observation measures.

Analysis of Cascades by Population Group

The cascades obtained allow the comparison of cascading behavior among different population groups. The analysis first compares cascades of posts submitted by bloggers of different gender (male, female). It is then extended to compare the situation in age groups (0-24, 25-39, 40-54, 55-80) and continents (North America (NA), South America (SA), Europe (EU), Africa (AF), Asia (AS), Oceania (OC)).

Cascades per Gender Figure 2(a) highlights that there are notable variations on the cascade triggering ability (A_S) between male posts (0.9%) and female posts (0.5%). These cascades are typically small and short ($C_s = 1$, $C_h = 1$) in both male and female posts (Figure 2(b)). However, male posts exhibit shorter reaction times in the blogosphere, but their discussions appear to be more *ephemeral* (i.e., finish soon after they start) compared to the female posts that last more (Figure 2(c)).

Cascades per Age Group Figure 2(d) highlights that there are notable variations on the cascade triggering abil-



(a) *Cascade Triggering Ability* (b) *Cascade Properties* (c) *Reaction Times*

Figure 3: Cascades per Subject

ity (A_S) among people of different age groups. The blogosphere seems to mostly engage in cascades triggered by posts of people in the age group of 40-54 (1.2%) and 55-80 (1.1%) and less in posts of people in age groups of 0-24 (0.3%) and 25-39 (0.9%). However, cascades in all age groups are typically small and short ($C_s = 1$, $C_h = 1$) (Figure 2(e)). Finally, the reaction time of the blogosphere to senior posts is shorter than to posts from younger and discussions generally last more (Figure 2(f)). It is also evident that discussions on posts coming from people in the 0-24 age group are late to start and ephemeral.

Cascades per Continent Figure 2(g) highlights that there are notable variations on the cascade triggering ability (A_S) of posts among continents. People appear to engage more in posts coming from NA, followed by EU, OC, AS, SA, and AF with values 1.2%, 0.8%, 0.7%, 0.4%, 0.4%, and 0.3% respectively. However, cascades attributed to all continents are typically small and short ($C_s = 1$, $C_h = 1$) (Figure 2(h)). Finally, the reaction times vary slightly among continents with the exception of EU posts that exhibit reactions that are immediate, but very ephemeral (Figure 2(i)). On the other hand reaction times in NA, SA and AF are not that timely but discussions last longer.

Analysis of Cascades by Subject

The cascades obtained allow the comparison of cascading behavior of posts in relation to their subject that varies among technology, entertainment, sports, financial, and politics. Figure 3(a) highlights that there are notable variations on the cascade triggering ability (A_S) of posts depending on their subject. Entertainment posts are much more likely to trigger cascades (50%). Politics posts have also a high probability (30%). On the other hand, financial posts rarely trigger cascades (only 5%). Technology and sports posts fall somewhere in the middle with 15%, and 13% respectively. Note that this trend appears to be independent to the number of posts in each subject. For example, even if there are almost as many financial posts as entertainment posts in our sample, the latter are 10 times more likely to launch cascades. In all subjects cascades are typically small and short ($C_s = 1$, $C_h = 1$) (Figure 3(b)). However, there are variations in the reaction times of the cascades (Figure 3(c)). The blogosphere seems to be more reactive to politics, sports and entertainment posts as indicated by the corresponding R_{min} values. However, politics posts have much larger R_{max} values which indicates that they continue to occupy the blogosphere for a longer period. This is also true for the technology posts. On the other hand, sports, financial and entertainment posts appear to be more ephemeral.

Tests of Significance

We performed statistical significance tests to assess evidence for our claims. For each case a null hypothesis was tested, stating that samples are not significantly different. When the null hypothesis was rejected, we were performing post-hoc analysis to determine the set of different pairs in each case. The tests of significance supported our claims that a blogger's profile (each of gender, age, and continent) and the subject of a post can significantly differentiate the cascading behavior. We omit details on the tests of significance due to space limitations.

Conclusions

Our analysis revealed notable variations of the cascading behavior depending on (a) the blogger's profile and (b) the subject of a post. More specifically:

- **cascade triggering ability**: Posts are more likely to trigger cascades if they are coming from males than from females, if they are submitted by middle-agers or seniors than younger, if they are related to entertainment or politics as opposed to sports, technology, or finance.
- **cascade size and height**: Structural properties of cascades, such as size and height, follow power law distributions with only a few cascades being large and deep. A typical cascade is small and shallow (i.e., $C_h = 1$, $C_s = 1$). Practically, this is a sign of limited discussion taking place in the blogosphere.
- **reaction times**: Reaction times are shorter for posts submitted by males, where the author is middled-aged or senior, or the post relates to politics or entertainment.

References

Bansal, N., and Koudas, N. 2007. Searching the blogosphere. In *WebDB*.

Gruhl, D.; Guha, R.; Kumar, R.; Novak, J.; and Tomkins, A. 2005. The predictive power of online chatter. In *KDD*.

Leskovec, J.; McGlohon, M.; Faloutsos, C.; Glance, N.; and Hurst, M. 2007. Cascading behavior in large blog graphs. In *SDM*.