Supplementary Material

Please cite this article as:

Information Cascades in the Blogosphere: A Look Behind the Curtain. *Manos Papagelis, Nilesh Bansal, Nick Koudas. Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media (ICWSM 2009).*

Engagement and Reaction in the Blogosphere

Manos Papagelis, Nilesh Bansal, Nick Koudas

University of Toronto, Canada

(papaggel, nilesh, koudas)@cs.toronto.edu

(Supplementary material to complement the article that appeared in ICWSM 2009.)

Abstract

The changing trends in the use of web technology that aims to enhance interconnectivity, self-expression and information sharing on the web has led to the development of online, many-to-many communication services such as blogs. Information elements in blogs can be interlinked. The succession of linking behavior determines the way in which information propagates among blogs, forming cascades. Analyzing cascades can be useful in various domains, such as providing insight of public opinion and developing better cascade models. In this paper, we present trends on the degree of engagement and reaction of bloggers in stories that become available in blogs under various parameters and constraints. We analyze cascades that are attributed to different population groups constrained by factors of gender, age, and continent, and how cascades differentiate depending on their subject. Our analysis is performed on one of the largest available datasets and reveals large variations in the properties of cascades.

1 Introduction

The process of ideas and practices spreading through a population, contagiously, with the dynamics of an epidemic, has long been of interest in social sciences. Its systematic study developed in the middle of the 20th century, into an area of sociology known as the *diffusion of innovations*. The theory of diffusion of innovations examines the effect of word of mouth and explores the role of social networks in influencing the spread of new ideas and behaviors. At a particular point in time, some nodes in the network, become aware of new ideas, technologies, fads, rumors, or gossip and they have the potential to pass them on to their friends and colleagues, causing the resulting behavior to cascade through the network. The initial research on this topic was empirical, but in the 1970s economists and mathematical sociologists began formulating basic mathematical models for the mechanisms by which ideas and behaviors diffuse through a population [15]. Two fundamental models for the process by which nodes adopt new ideas have been considered, the *threshold model* [4] and the *independent cascade model* [3]. Models of diffusion of innovations in a social network have since been considered in many disciplines including mathematical epidemiology, viral marketing, game theory (see [8] and the references therein).

More recently, the changing trends in the use of web technology that aims to enhance interconnectivity, self-expression and information sharing on the web has led to the development of online, many-to-many communication services such as blogs. A *blog* is a website with regular entries of commentary called *blog posts* or simply *posts*. The collective community of all blogs and their interconnections is known as the *blogosphere*. Information diffusion in technological networks raises interesting connections to the theoretical models [9]. As news and stories become available in the real-world, they spread in the blogosphere forming information cascades. The capacity to collect and analyze information cascades can be useful in various domains, such as providing insight on public opinion on a variety of topics [5] or developing better predictive models of the spread of ideas and behaviors [13]. It can also be useful in applications that make use of the diffusion process, such as in the problem of maximizing the spread of influence in a social network [7, 14] and the problem of early detecting outbreaks in a network [12].

The cascade process may be modeled by inferring links through textual similarity, as in [6] where authors focus on the diffusion of topics in blogs, or in [11] where authors track the news cycle through short, distinctive phrases that travel intact through online text. Cascades may as well be modeled through links among blog posts, such as in [1], where linking patterns of political blogs are explored, and in [10], where community-level behavior is analyzed in an online social network inferred from blog-roll links (links to affiliated blogs) among blogs. Our work is mostly related to work in [13], where authors study the temporal and topological patterns of cascades in large blog graphs based on link analysis. In that work, authors assumed spreading of information regardless of the *context* of the posts, and highlighted the need for further analysis that would be useful to the development of more accurate patterns of information propagation.

In this paper, we present trends of 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*. Then, we analyze how cascades differentiate in five subjects: *technology*, *politics*, *financial*, *sports*, and *entertainment*. In each case, we report on the structural properties of cascades and try to identify and quantify any variability in the cascading behavior. In our study we collected and analyzed information of cascading behavior that is available in *Blogscope* [2]. BlogScope is an analysis and visualization tool for the blogosphere and is currently tracking almost 36 million blogs with over 800 million posts making it one of the largest available datasets for our analysis.

2 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.

2.1 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 $\ell_{p_y \rightarrow p_x}$ represent a link from a post p_y in the future to a post p_x in the past. For each link $\ell_{p_y \rightarrow p_x}$ 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)).

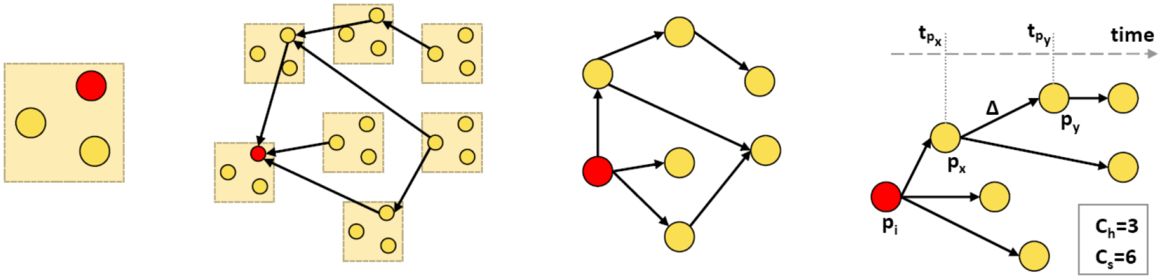
A cascade C can be thought of as a directed graph, where nodes represent posts and edges represent information flow between posts. Note that the direction of an edge in C follows the information propagation (the actual permalink follows the opposite direction). We denote the in-degree of a node $\nu \in C$ 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., $R_p = \Delta_{p_i}^{p}$). 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., $\Delta > 0$).

2.2 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. 36M blogs with 800M posts) for a



(a) *blog with posts* (b) *post graph* (c) *cascade* (d) *cascade properties*

Figure 1: Example Cascade

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 cascades attributed to variable blog profiles and one for analysis of cascades of different subjects. Following we present details on each of the sample datasets.

2.2.1 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, meaning that the gender, age, and country of the blogger is defined (blogs in Blogscope are associated with a *blog profile*). 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.

2.2.2 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.

2.3 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 $|S|$:

$$A_S = \frac{|S_C|}{|S|} \tag{1}$$

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 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.

3 Analysis

The raw data obtained from the cascades requires a quality check in order to reject potential wrong cascades and avoid the existence of possible deviations in results. In particular, the quality check takes the following into account:

- *Closed-world Assumption*: We make a closed-world assumption that out-links to posts are valid only if they are out-links to posts in the dataset.

- *Total-ordering of Posts Assumption*: We assume that there is a total-ordering of the posts in the database based on the timestamps of the posts. This assumption allows to ignore links from a post in the past to a post in the future. Such links are infrequent but resident in the dataset and are mostly due to post updates or differences in time-zones.

- *Self-links Removal*: We are interested in information cascades in a blog level. Links between posts of the same blog are considered self-links and are eliminated from the study.

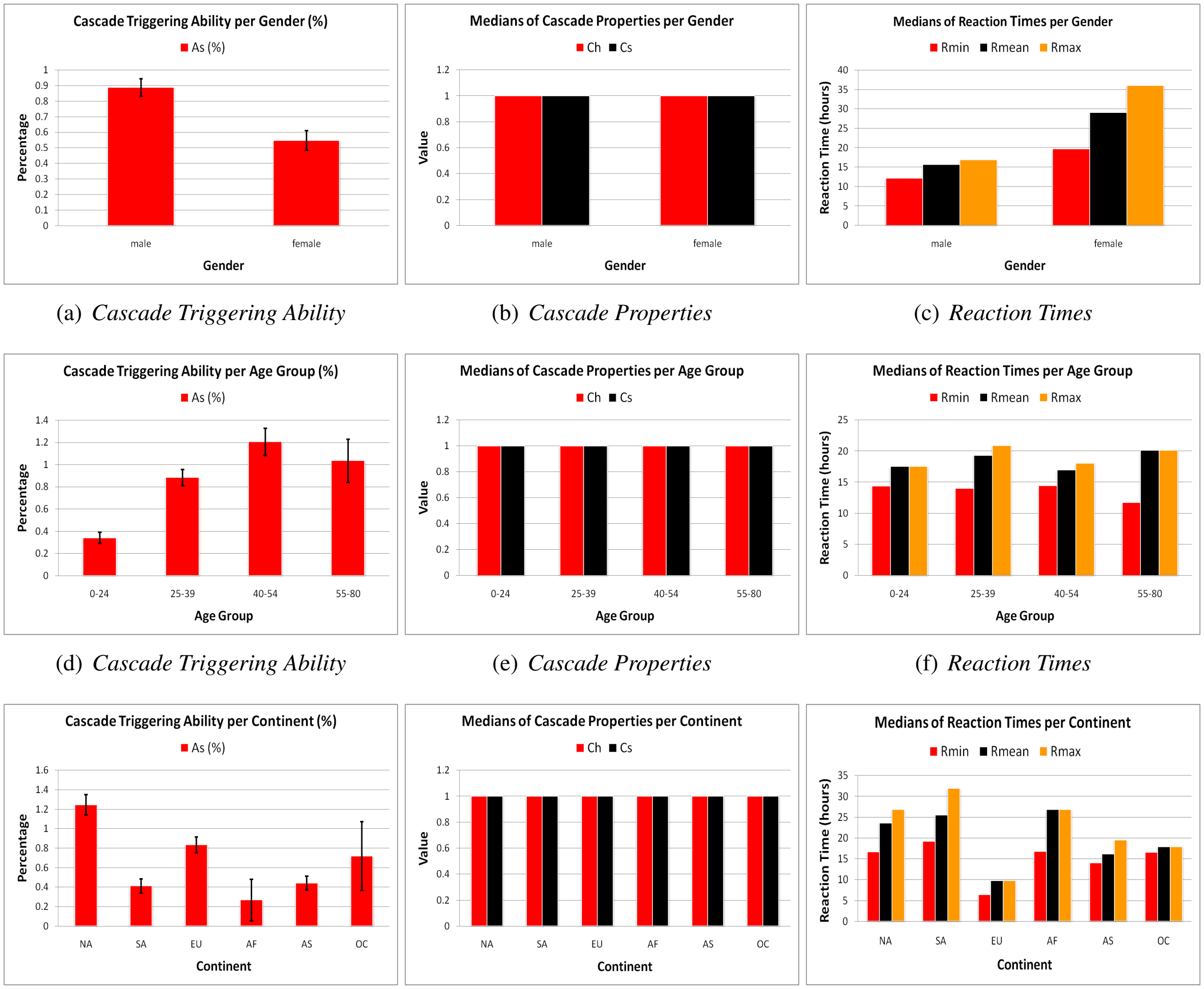
- *Spam and Duplicate Post Detection and Elimination*: We make use of the Blogscope spam analyzer to filter out posts that are likely to be spam. We also detect and eliminate any duplicate posts from the datasets.

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.

3.1 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)).



(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)

Cascades per Age Group: Figure 2(d) highlights that there are notable variations on the cascade triggering ability (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.

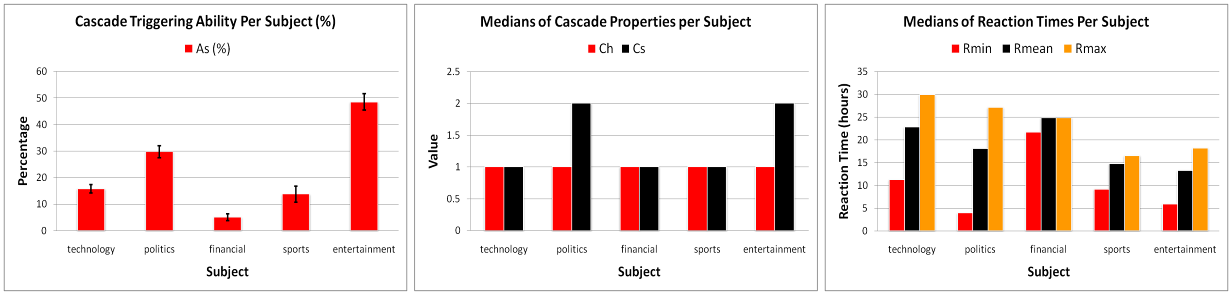
Cascades per Continent: Figure 2(g) highlights that there are notable variations on the cascade triggering ability (A_S) of posts among continents. People appear to engage more in posts coming from NA, followed by EU, OC, AS, SA, and AF with values 1.2%, 0.8%, 0.7%, 0.4%, 0.4%, and 0.3% respectively. 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.

3.2 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.

4 Validation

Thoughout our study we based our observations on sample populations. But, how dependent our observations on these samples are? Statistical inference allows to assess evidence in favor of or against some claim about the population from which the sample has been drawn. The methods of inference we use to support



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

Figure 3: Cascades per Subject

or reject our claims based on sample data are known as *tests of significance*.

4.1 Tests of Significance

The first step in testing for statistical significance is to postulate a *null hypothesis* H_0 and a *research hypothesis* H_a. For the case of k-samples, a null hypothesis usually states that samples are not significantly different (i.e., they come from the same population). To test whether the null hypothesis should be rejected or accepted, we perform, for a selected probability of error level (alpha level), a *statistical significance test* that compares the k-samples. If the null hypothesis is rejected, indicating that at least two samples are significantly different, we then need to perform a *post-hoc analysis* to determine which of the k-samples are different. Due to the nature of the quantities that we are reporting in our study we employ two types of significance tests.

4.1.1 Tests for k Proportions

The cascade triggering ability A_s is a proportion that measures, on a [0, 1] scale, how many of the posts in a sample set are able to trigger non-trivial cascades (see Equation 1). The statistical test we use to compare k proportions is known as a Chi-square test. Since the Chi-square test is an asymptotical test (i.e., vulnerable to errors with small values), we actually run the simulations based test (Monte Carlo test) with 5000 simulations. For the various cases in our study, we formally define the following hypotheses:

H_0^A: *The proportions are not significantly different*

H_a^A: *At least one proportion is significantly different* from another

In the case that the null hypothesis is rejected by the test, we perform post-hoc analysis using the Marascuilo procedure to identify which of the series differ. The Marascuilo procedure simultaneously tests the differences of all pairs of proportions for the several samples under investigation.

4.1.2 Non-parametric Tests for k Independent Samples

The observation measures C_s, C_h, R_{min}, R_{mean}, R_{max} are all real numbers that describe a cascade. The values of these properties typically have skewed distributions. Thus, to compare k-samples of one of these properties we employ a method that does not assume a normal population, such as the Kruskal-Wallis test. This test is a non-parametric method for testing equality of population medians among groups. Intuitively, it is identical to a one-way analysis of variance with the data replaced by their ranks. For the various cases in our study, we formally define the following hypotheses:

H_0^B: *The samples are not significantly different*
H_a^B: *The samples do not come from the same population*

In the case that the null hypothesis is rejected by the Kruskal-Wallis test, we perform post-hoc analysis to identify which of the series differ. We use the bonferroni adjustments post-hoc test to identify these pairs.

4.2 Results

4.2.1 Cascades by Population Group

We performed statistical significance tests for each of the various segmentations of the population in our study (i.e., Gender, Age Group, Continent). For each segmentation, the computed p-value of the Chi-square test was lower than the significance level $alpha = 0.05$, so one should reject the null hypothesis H_0^A and accept the alternative H_a^A. Therefore, we can conclude that a blogger's profile (each of gender, age, and continent) can significantly differentiate the cascade triggering ability A_s of a post. Moreover, for each population segmentation, the post-hoc analysis showed that all pairs are significantly different except for the pairs (25-39, 55-80) and (40-54, 55-80) regarding the age group segmentation and the pairs (NA-OC), (SA-AF), (SA-AS), (SA-OC), (EU-OC), (AF-AS), (AF-OC), (AS-OC) regarding the continent segmentation.

For the case of C_s, C_h, R_{min}, R_{mean}, R_{max}, the Kruskal-Wallis test showed that the samples do not come from the same population, thus rejecting the null hypothesis H_0^B. The only exception was in the reaction times of the age group segmentation where the null hypothesis was accepted, indicating no significant difference in the reaction times of posts coming from diverse age groups. The post-hoc analysis revealed the set of different pairs in each case. We omit details on pairs due to space limitations.

4.2.2 Cascades by Subject

We performed statistical significance tests for the cascades of different subjects in our study. The computed p-value of the Chi-square test was lower than the significance level $alpha = 0.05$, so one should reject the null hypothesis H_0^A and accept the alternative H_a^A. Therefore, we can conclude that the subject of a post can significantly differentiate its cascade triggering ability A_s. Moreover, the post-hoc analysis showed that only the pair (technology, sports) is not significantly different, while all other pairs are.

For the case of C_s, C_h, R_{min}, R_{mean}, R_{max}, the Kruskal-Wallis test showed that the samples do not come from the same population, thus rejecting the null hypothesis H_0^B. The post-hoc analysis revealed the set of different pairs in each case. We omit details on pairs due to space limitations.

5 Discussion

Analysis of the cascading behavior of blogs is a challenging research topic. We focused on the effect of incorporating heterogeneity into the analysis of information cascades in the blogosphere. 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.

The issue that different bloggers present different potential to trigger cascades provides useful insights for the development of better prediction models. In particular, microscopic analysis of blogger behaviors could serve as the basis for analyzing macroscopic properties of networks, such as prediction models of the spread of ideas and methods to measure influence among blogs. It might also be useful in the development of tools that make use of the diffusion process (e.g., in fields of epidemiology and viral marketing).

Regarding the subject of a post, the issue that different topics present different background requirements could explain some of the variability of the cascading behavior we observe. For example, one would expect that entertainment posts are more likely to trigger cascades than financial posts, since it is much easier to discuss entertainment than finance topics. It may also be relevant that different user groups are associated with different topics. The latent semantics of these values and how they correlate is intricate. However, it is important that our work not only clearly identifies this variability, but also tries to quantify it.

References

[1] L. A. Adamic and N. Glance. The political blogosphere and the 2004 u.s. election: divided they blog. In *LinkKDD*, 2005.

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

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

[4] M. Granovetter. Threshold models of collective behavior. *The American Journal of Sociology*, 83(6), 1978.

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

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

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

[8] J. Kleinberg. Cascading behavior in networks: Algorithmic and economic issues. *Algorithmic game theory*, 24:613–632, 2007.

[9] J. Kleinberg. The convergence of social and technological networks. *Commun. ACM*, 51(11), 2008.

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

[11] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *KDD*, 2009.

[12] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *KDD*, 2007.

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

[14] M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. In *KDD*, 2002.

[15] T. C. Schelling. *Micromotives and Macrobehavior*. W. W. Norton & Company, Inc., 1978.