A Study on Characteristics of Topic-Specific Information Cascade in Twitter

Geerajit RATTANARITNONT†, Masashi TOYODA†, and Masaru KITSUREGAWA†

† Institute of Industrial Science, The University of Tokyo 4-6-1 Komaba, Meguro-ku, Tokyo, 153–8505 Japan

E-mail: †{aomi,toyoda,kitsure}@tkl.iis.u-tokyo.ac.jp

Abstract In this paper, we study patterns of information diffusion and behaviors of participating users in Twitter. We investigate characteristics of hashtag cascade in various topics by exploiting distributions of user influence, which are cascade ratio, tweet ratio, and time interval. We show that topics of major hashtags can be characterized by these distributions. For example, people using political hashtags often influence many of their friends and continuously discuss on the topics. Our experiments also show that the hashtags can be roughly clustered into topics using only those measures, and miss-clustered hashtags have some special roles in their topics.

Key words Social network, Information diffusion, Web mining

1. Introduction

Nowadays people can keep in touch with each other on social networking sites such as Facebook, Twitter, and MySpace. People connecting to online social networks can share interests and activities with their friends, and even make new friends all over the world. The resulting networks grow rapidly and gained significant popularity on the Internet.

Many researchers have studied various aspects of online social networks such as network structure (e.g., community detection), user relationships (e.g. link prediction), and information flow between users (e.g. information diffusion). In this paper, we perform a research on Twitter's network structure to understand patterns of information diffusion and behaviors of participating users. This will help us to verify whether campaigns or advertisements for viral marketing are successful according to marketing strategies or not. For example, new products are spread among a large number of people or talked frequently among a specific group of people.

We analyze the spread of information according to topics of most frequently used hashtags and find the characteristics across them. To investigate a large amount of data, we consider three probability distributions of user influence, which are cascade ratio, tweet ratio, and time interval. The cascade ratio indicates an ability of a user to cascade information to his/her neighborhoods, the tweet ratio measures how much a user is involved in each topic, and the time interval determines how long time a hashtag is used since the first post. We show that topics of major hashtags can be characterized by these distributions. For example, people using political hashtags often influence many of their friends and continuously discuss on the topics. But people using earthquake hashtags seldom influence their friends and talk very few times about this topic during first period of time. Our experiments also show that the hashtags can be roughly clustered into topics using only those measures, and miss-clustered hashtags have some special roles in their topics. For instance, "jishin" and "genpatsu" which both are belong to the earthquake topic. However, since "jishin" is directly related to the Great East Japan Earthquake, people talked about it very much during that time and in turn rarely talked when the situation was back to normal. On the other hand, "genpatsu" is related to the nuclear power plant, people thus continued discussing about it even after the quake stopped because the nuclear problem has been unsolved yet.

The Twitter dataset used in this paper is crawled from March 11, 2011 to July 11, 2011. It consists of 260 thousand users and 783 million tweets. Because our dataset is captured during the Great East Japan Earthquake, it becomes one of the most frequently posted topics in Twitter. Therefore, we take earthquake including with politics, media, and entertainment as four major topics into account throughout the study.

The rest of this paper is organized as follows. Section 2 introduces related work on information diffusion in online blogging and social networking services. Section 3 explains the dataset. In Section 4, we describe three different distributions (cascade ratio, tweet ratio, and time interval) and investigate the characteristics of information diffusion over four categories (earthquake, politics, media, and entertainment). Then we conduct further analysis by using clustering algorithm in Section 5. Finally, we conclude this paper and future work in Section 6.

2. Related Work

Information diffusion in online community has been studied for a decade. Gruhl *et al.* [4] studied the dynamics of information propagation in weblogs. They investigated characteristics of long-running topics due to outside world events or within the community. Leskovec *et al.* [6] also studied information propagation in weblogs. They proposed a simple model that mimics the spread of information in blogspace and is similar to propagation found in real life.

Instead of blogsphere, Liben-Nowell *et al.* [7] traced the spread of information at individual level and found that information reach people in a narrow deep pattern, continuing for several hundred steps. Similarly, Sun *et al.* [10] conducted an analysis on information diffusion in Facebook and discovered that large cascade begins with a substantial number of users who initiate short chains.

In most recent years, as Twitter becomes one of the most popular micro-blogging services and allows us to obtain its data via Twitter API, it gains much interest from many researchers [2], [3], [5], [8], [9], [11]〜[13]. Romero *et al.* [9] studied information spread in Twitter and shown that controversial political topics are particularly persistent with repeated exposures comparing to other topics. Moreover, rather than understanding how information itself is spread, Bakshy et al. [1] exploited information cascade to identify influencers in Twitter. Scellato et al. [11] also extracted geographic information from information dissemination process and utilized it to improve caching of mulitimedia files in a Content Delivery Network. Unlike others, this work study characteristics of information diffusion over different topics in Twitter in term of cascade ratio and tweet ratio. Because we will understand how people interact with each other and how information is cascaded; in addition to identifying influencers as in [1], [3], [5], [12], we can utilize this work to verify success of viral marketing strategy.

3. Twitter Dataset

We crawled the Twitter dataset from Twitter API from March 11, 2011 when the Great East Japan Earthquake took place to July 11, 2011. We obtained 260 thousand users and 783 million tweets. Our data collection consists of user profiles, timestamp and tweet contents including with retweet source and mention destination. Then we define our interested users, network and hashtags as below.

3.1 Users

In this paper, we consider users who have at least one tweet during the period of dataset. Therefore, we have 260 thousand users as active users.

Table 1 An example of hashtags in each category

Earthquake	Politics	Media	Entertainment
jishin	bahrain	nicovideo	madoka_magica
genpatsu	iranelection	nhk	akb48
prayforjapan	wiunion	news	atakowa
save_fukushima	teaparty	fujitv	tigerbunny
save_miyagi	gaddafi	ntv	anohana

3.2 Network

Instead of friend-follower graph, we regard directed links among users when user A has at least one retweet from or mention to user B and call this relationship as outgoing neighborhood. Including with the idea of active users, we gained 31 million links.

3.3 Hashtags

In order to study information cascade according to different topics, we treat a hashtag as a representative of the topic users talk about. Although URL is another choice, we choose hashtag over URL because it provides the sense of topic more comprehensive than URL. In other words, URL is too specific. One topic can be indicated by a large number of URLs.

We select top 100 frequently used hashtags from the dataset and categorize them according to topics. We have four main categories which are earthquake, politics, media, and entertainment. There are 21, 32, 13, and 11 hashtags in each category respectively. Table1 shows an example of hashtags in each category. Whereas earthquake category is about the Great East Japan Earthquake, politics category is related to political issues and events all over the world especially the uprising events in the Middle East, e.g., "bahrain" hashtag. Media category is represented by communication channels, such as, television networks and video sharing websites; and most of them are Japanese channels. Finally, entertainment category refers to television programs, movies and artists; and the majority are again Japanese animations.

4. User Influence Distributions

To study the characteristics of information diffusion over massive dataset, we define three distributions related to user influence which are cascade ratio, tweet ratio, and time interval in this section respectively.

4.1 Cascade Ratio

Cascade ratio determines the proportion of how much a user can influence his/her neighborhoods to spread a hashtag comparing to all users who used the same hashtag. We captured the cascade by tracing the time each user firstly used a given hashtag. Thus, cascade score of a user is a number

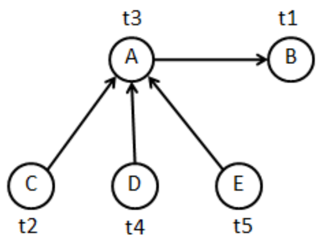


Fig. 1 Information Cascade

of his/her immediate incoming neighborhoods that reposted the hashtag after him/her as shown in Fig.1. A node and a directed edge in the graph represents a user and a link of our network respectively, while t indicates the first timestamp each user posted a given hashtag. According to the figure, user C,D, and E have ever retweeted from or mentioned to user A at least once during our dataset period. However, only user D and E used the same hashtag after user A, the cascade score of user A thus equals to two which refers to user D and E. The cascade ratio cr of a user u posting a hashtag h is then defined as below:

$$cr(u,h) = \frac{C(u,h)}{U(h)} . \quad (1)$$

where $C(u,h)$ is the cascade score of the user u posting the hashtag h and $U(h)$ is a set of all users using h.

Fig.2a illustrates the probability distribution of cascade ratio of all hashtags according to four categories which are earthquake, politics, media, and entertainment respectively. x is cascade ratio and y is the number of occurrences of cascade ratios normalized by total number of users using a given hashtag. The plot is in log-log coordinate and calculated as a cumulative distribution function, where y or $P(x)$ is the probability at a value greater than or equal to x.

Each line remains horizontally at the beginning and then starts to fall down at each cascade ratio assigned to a user. Between any two points, the higher slope is, the more users have those corresponding cascade ratio values. According to Fig.2a, the earthquake and the media categories are steep at small cascade ratio. That means a number of people in these two categories have comparatively low cascade ratio. We can imply that people used hashtags independently not because of seeing from their friends' tweets. In turn, the hashtags themselves are hot topics or general words so that people know them already. For example, "jishin" which means earthquake in Japanese language and "nhk" which is Japan's national public broadcasting organization. In contrast, the politics and the entertainment categories go down at higher cascade ratio comparing to the previous two categories. We can say that a number of people have relatively high cascade ratio. When a user post a hashtag in these categories, many of his/her friends will also post it after him. This is because the hashtags are discussion topics or special words known only in a specific group of people. For instance, "sgp" which stands for Smart Girl Politics, a non-profit organization for conservative women activists; and "akb48" which is a popular Japanese female idol group.

For easy to see the difference among four categories, Fig.3a shows the average cascade ratio distribution of all hashtags in each category. We see that 90% of all users who use hashtags in the earthquake and the media categories have cascade ratio less than 0.005 which means they directly influence less than 0.5% of all users, while 2.7% and 0.8% for the politics and the entertainment categories respectively. Although an overall view of the entertainment category in Fig.2a has similar trend as the politics category, its average line in Fig.3a is in the middle between the politics category and the earthquake-media categories. We can imply that a group of people who participate in entertainment topic is less dense than a group of people who talk about political topic.

4.2 Tweet Ratio

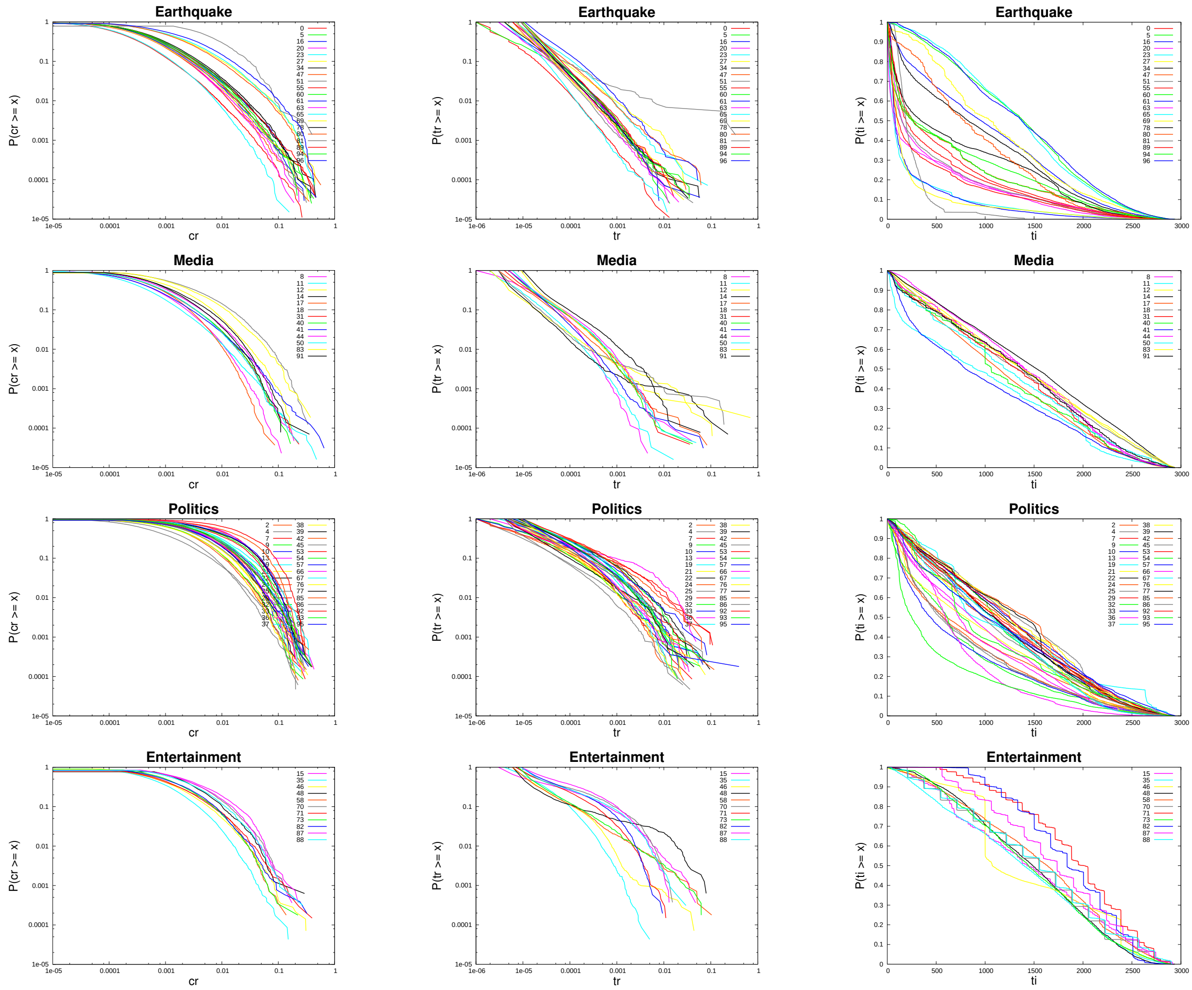
Another measure is tweet ratio which determines how much users engage in each topic. The tweet ratio tr of a user u posting a hashtag h is then simply defined as below:

$$tr(u,h) = \frac{T(u,h)}{W(h)} . \quad (2)$$

where $T(u,h)$ is the number of tweets containing the hashtag h posted by the user u and $W(h)$ is a number of all tweets containing h.

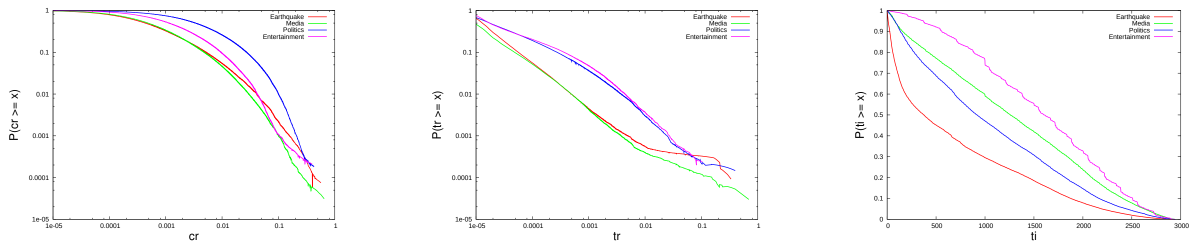
Fig.2b shows the probability distribution of tweet ratio of all hashtags according to four categories. x is tweet ratio and y is the number of occurrences of tweet ratios normalized by total number of users using a given hashtag. Each line is plotted in log-log coordinate and calculated as a cumulative distribution function, where y or $P(x)$ is the probability at a value greater than or equal to x.

We see that earthquake and media categories are quite straight comparing to other two categories, which means a number of people in the earthquake and media categories have comparably low tweet ratio. In other words, they posted tweets containing hashtags but repeated to use those hashtags very few times. This is because the hashtags are related to announcements and people then just passed information to others without discussion. In this case, although there is one line explicitly separating from other lines in the earthquake category, it still has the same interpretation with different numbers of involving users. It refers to "pf_anpi" hashtag, which is safety information for finding missing people during the Great East Japan Earthquake. Alternatively, politics and entertainment categories are more curved. More people repetitively used same hashtags many times than people in the first two categories. The hashtags themselves are



(a) Cascade ratio (b) Tweet ratio (c) Time interval

Fig. 2 Probability distributions of all hashtags in each category



(a) Cascade ratio (b) Tweet ratio (c) Time interval

Fig. 3 Average distributions of all hashtags in each category

discussion topics among groups of interested people.

Additionally, Fig.3b demonstrates the average tweet ratio distribution due to different four categories. We see that they are separate clearly into two groups. In more details, 90% of all users who use hashtags in the earthquake and the media categories have tweet ratio less than 0.00005 which means they only posted 0.005% of overall tweets containing a given hashtag. On the other hand, the politics and the entertainment categories have 0.03% which is six times higher than the first two categories.

4.3 Time Interval

The last measure used to investigate the characteristics of information diffusion is time interval which indicates how long time a hashtag is used since the first post. Therefore, the time interval ti of a tweet tw containing a hashtag h is straightforwardly defined as the difference in time between tw and the first tweet of h.

Fig.2c demonstrates the probability distribution of time interval of all hashtags according to four categories. x is time interval in hour(s) and y is the number of occurrences of time intervals normalized by total number of tweets comprising a given hashtag. Each line is plotted as a cumulative distribution function, where y or $P(x)$ is the probability at a value greater than or equal to x.

The majority of earthquake category fall down abruptly at first period. That means a large number of tweets were posted soon after the topics were raised to Twitter. And a number of tweets gradually decreased when time passed. People talked very much about the Great East Japan Earthquake during that time and they in turn rarely said about it when the situation was back to normal. Contrarily, the majority of media and politics categories lay in a diagonal. We can imply that the number of tweets did not change according to time. People continually talked about these topics. For example, "newsjp" which is about daily news in Japan and "syria" which is related to events in Syria especially the 2011 Syrain uprising from March 15, 2011 until now. Lastly, the majority of entertainment category lie above the diagonal at the beginning and thereafter fall under the diagonal. We can say that a number of tweets increased over time. This is possibly because people talked a little about television programs before they are on air. However, after on air for a while, they were already known and people thus said much about them. Besides, some hashtags in this category are sawtooth. We see that there are approximately three peaks in each 500 hours or one peak per week. It is likely that people talked much on the on air day. For example, "tigerbunny" and "anohana" which are both Japanese animations on air once a week on television channels.

To be easy to compare the difference among four categories, Fig.3c illustrates the average time interval distribution of all hashtags in each category. Specifically, there are 55%, 23%, 32%, and 8% of all tweets containing hashtags in the earthquake, the media, the politics, and the entertainment categories during the first 500 hours after their first post respectively. After that, during the first 500 - 1000 hours, there are 15%, 18%, 22%, 17% of all tweets in each category accordingly.

5. Information Cascade Clustering

In last section, we analyzed the characteristics of information diffusion across topics by manually dividing hashtags into four categories; earthquake, media, politics, and entertainment. In this section, we further investigate the characteristics of information diffusion by using clustering algorithm.

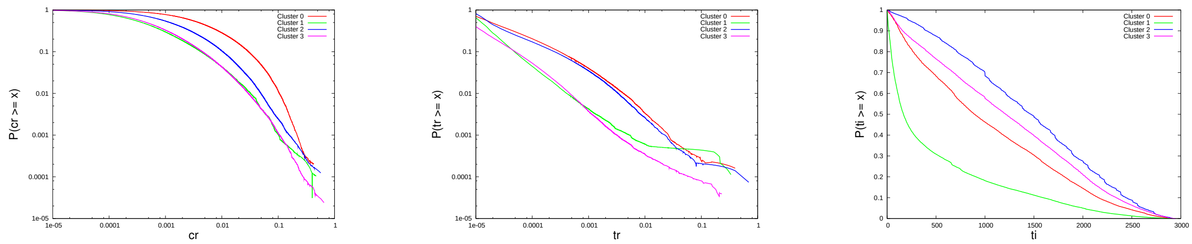
We performed k-means clustering based on the distributions of cascade ratio, tweet ratio, and time interval. Then we compared the result with categories from the earlier section. Table2 illustrates proportion of four categories assigned to each cluster when we choose the number of clusters as $k = 4$. Interestingly, we see that the majority of each category are put into different clusters even we do not consider tweet contents. That means each category has different characteristics over cascade ratio, tweet ratio, and time interval.

Table 2 Proportion of each category in each cluster

No. of hashtags	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Total
Earthquake	0	14	3	4	21
Media	0	0	2	11	13
Politics	28	1	2	1	32
Entertainment	0	0	10	1	11

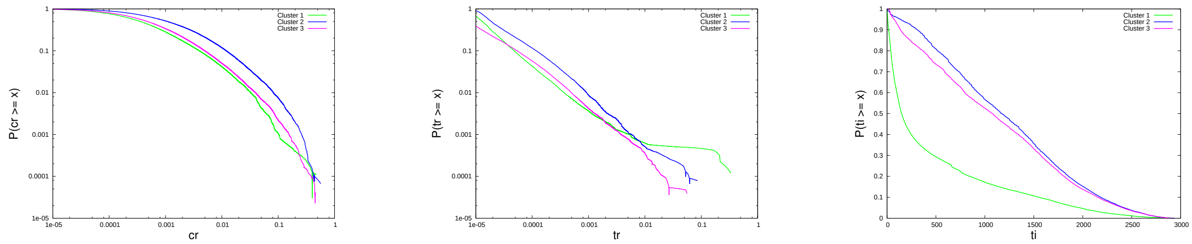
Fig.4 shows the average distributions of all hashtags according to four clusters. Cluster 1 and 3 to which the majority of the earthquake and the media catogories are assigned respectively have low cascade ratio and low tweet ratio. People tweeted about these topics a few times on their own selves not because of influenced by their friends. However, a number of tweets in cluster 1 decline when time passed. People talked very much at the beginning of events and then almost stopped saying at the end of events. In another way, a number of tweets in cluster 3 almost remains the same as well as in cluster 0 which are from the politics category. That is, people continuously talked over time. Anyhow, cluster 0 instead has high cascade ratio and high tweet ratio. People talked about the same topics many times and many of their friends also started talking after seeing their tweets. That means the political topic is specific to groups of interested people, hence they discussed much between each other. In case of cluster 2 which are almost from the entertainment category, although it is has high tweet ratio as same as cluster 0, it stays in the middle among cluster 0, 1, and 3 in term of cascade ratio. We can explain that the entertainment topic is much discussed among specific groups of people than the earthquake and the media topics but not as much as the political topic. Moreover, a number of tweets in cluster 2 tended to increase according to time. People talked about the entertainment topic much more after it became known.

According to Table2, whereas most of hashtags in different categories are divided into different clusters, some of them are assigned to other clusters belonging to other categories. Here we focus on the earthquake category. The majority are assigned to cluster 1, while the minority are assigned to cluster 2 and 3. Fig.5 demonstrates the average distributions of all hashtags in the earthquake category according to three clusters. Even though they seem very similar in term of cascade ratio and tweet ratio, they are obviously different in term of time interval. The hashtags in cluster 1 are directly related to the Great East Japan Earthquake such as "jishin" and "save_miyagi". Therefore, people talked very much during that time and in turn almost did not talk when the situation got back to normal. On the other hand, the hashtag in cluster 2 and 3 are related to the nuclear power plant such as "nuclearjp" and "save_fukushima". People thus con-



(a) Cascade ratio (b) Tweet ratio (c) Time interval

Fig. 4 Average distributions of all hashtags in each cluster



(a) Cascade ratio (b) Tweet ratio (c) Time interval

Fig. 5 Average distributions of earthquake category in each cluster

tinued discussing about the nuclear problem even after the quake stopped because it has not been recovered yet.

6. Conclusion

In this paper, we studied the characteristics of information diffusion according to topics of most frequently used hashtags in Twitter. Here we focused on four different topics which are about earthquake, politics, media, and entertainment. We analyzed the probability distributions of three measures, cascade ratio, tweet ratio, and time interval. We have found that the earthquake and the media categories are hot topics or general words so that most of people know them already and use hashtags by themselves not because of being influenced from their friends. On the other hand, the politics and the entertainment categories are discussion topics or special words known among groups of interested people. Moreover, people talked much about the earthquake topic and almost stopped talking when the situation was back to normal, while they continuously talked about the media and political topics over time and talked much more about the entertainment topic after it was already known. Finally, as future work, we need to explore the characteristics of more categories in larger dataset, such as technology and games category.

References

[1] E. Bakshy, J.M. Hofman, W.A. Mason, D.J. Watts, Everyone's an Influencer: Quantifying Influence on Twitter. In 4th International Conference on Web Search and Data Mining, pp. 65–74. ACM (2011)

[2] C. Castillo, M. Mendoza, B. Poblete, Information Credibility on Twitter. In 20th International Conference on World Wide Web, pp. 675–684. ACM (2011)

[3] M. Cha, H. Haddadi, F. Benevenuto, K.P. Gummadi, Measuring User Influence in Twitter: The Million Follower Fallacy. In 4th International Conference on Weblogs and Social Media, pp. 10–17. AAAI (2010)

[4] D. Gruhl, R. Guha, D. Liben-Nowell, A. Tomkins, Information Diffusion Through Blogspace. In 13th International Conference on World Wide Web, pp. 491–501. ACM (2004)

[5] H. Kwak, C. Lee, H. Park, S. Moon, What is Twitter, a Social Network or a News Media?. In 19th International Conference on World Wide Web, pp. 591–600. ACM (2010)

[6] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, M. Hurst, Patterns of Cascading Behavior in Large Blog Graphs. In 7th SIAM International Conference on Data Mining, pp. 551–556. SIAM (2007)

[7] D. Liben-Nowell, J. Kleinberg, Tracing Information Flow on a Global Scale Using Internet Chain-Letter Data. In The National Academy of Sciences, pp. 4633–4638. PNAS (2008)

[8] B. Meeder, B. Karrer, A. Sayedi, R. Ravi, C. Borgs, J. Chayes, We Know Who You Followed Last Summer: Inferring Social Link Creation Times in Twitter. In 20th International Conference on World Wide Web, pp. 517–526. ACM (2011)

[9] D.M. Romero, B. Meeder, J. Kleinberg, Differences in the Mechanics of Information Diffusion Across Topics: Idioms, Political Hashtags, and Complex Contagion on Twitter. In 20th International Conference on World Wide Web, pp. 695–704. ACM (2011)

[10] E. Sun, I. Rosenn, C. Marlow, T. Lento, Gesundheit! Modeling Contagion through Facebook News Feed. In 3rd International Conference on Weblogs and Social Media, pp. 146–153. AAAI (2009)

[11] S. Scellato, C. Mascolo, M. Musolesi, J. Crowcroft, Track Globally, Deliver Locally: Improving Content Delivery Networks by Tracking Geographic Social Cascades. In 20th International Conference on World Wide Web, pp. 457–466. ACM (2011)

[12] J. Weng, E.-P. Lim, J. Jiang, Q. He, TwitterRank: Finding Topic-Sensitive Influential Twitterers. In 3rd International Conference on Web Search and Data Mining, pp. 261–170. ACM (2010)

[13] S. Wu, J.M. Hofman, W.A. Mason, D.J. Watts, Who Says What to Whom on Twitter. In 20th International Conference on World Wide Web, pp. 705–714. ACM (2011)