A Study on Relationships between Information Cascades and Popular Topics 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 perform a research on Twitter's user network to understand patterns of information cascade and behaviors of participating users in various topics by exploiting three measures, which are cascade ratio, tweet ratio, and time of tweet. We show that hashtags in different topics have different cascade patterns in term of these measures. However, some hashtags even in the same topic have different cascade patterns. For instance, the earthquake hashtags can be divided into the hashtags directly related to the Great East Japan Earthquake, the media-related hashtags, and the political-related hashtags or the hashtags about the nuclear power plant. We surprisingly discover that such kind of hidden relationship between topics can be revealed by using only three measures rather than considering tweet contents.

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. Information is then said to be cascaded over the Internet. For example, people in Japan spread "Operation Yashima" on Twitter to conserve electricity due to the Great East Japan Earthquake. This kind of situation is an emergency and needs to be reached a large number of people within short time. Unlike other activities, for instance, Fukushima Daiichi Nuclear Power Plant faced failures according to the Great East Japan Earthquake. Because this is a serious problem and cannot be solved immediately, much of discussion and concerns are continually talked by people including experts.

Since different activities tend to have different ways information spread on the network, studying patterns of information cascade would help organizations to examine behaviors of public relation campaigns. Therefore, in this paper, we perform a research on Twitter's user network to understand patterns of information cascade and behaviors of participating users in various topics such as earthquake and political topics. We verify whether different topics really have different cascade patterns or not by exploring three measures, which are cascade ratio, tweet ratio, and time of tweet. The cascade ratio determines how much people can influence their friends, the tweet ratio determines how much people talk in each topic, and lastly the time of tweet determines how long a topic is still popular in the network. We consider Twitter hashtags as representatives of topics and conduct experiments on a real Twitter dataset.

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. We select top 500 frequently used hashtags from the dataset and categorize them according to topics. We found that the majority fall into six topics which are earthquake, media, politics, entertainment, sports, and idiom. We firstly study the pattern of hashtag cascades in each topic by using statistical approach. We then further investigate the relationship between cascade patterns and topics by using clustering algorithm. Our results show that hashtags in different topics have different cascade patterns in term of cascade ratio, tweet ratio, and time of tweet. For example, the earthquake topic has low cascade ratio, low tweet ratio, and short lifespan, while the political topic has high cascade ratio. However, some hashtags even in the same topic have different cascade patterns. For instance, the earthquake hashtags can be divided into the hashtags directly related to the Great East Japan Earthquake, the media-related hashtags, and the political-related hashtags or the hashtags about the nuclear power plant. We discover that such kind of hidden relationship between topics can be surprisingly revealed by using only three measures rather than considering tweet contents.

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 measures of users' influence and posting behaviors, and investigate the characteristics of information diffusion over six major topics. 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 blogging services has been studied for a decade [1], [6], [10], [11]. Gruhl *et al.* [6] studied the dynamics of information propagation in weblogs. They investigated characteristics of long-running topics due to outside world events or within the community. Adar *et al.* [1] developed a tool to visualize the flow of individual URLs over a blog network. Leskovec *et al.* [11] 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, researchers are also interested in information diffusion on other networks especially upcoming social networks [3], [7], [9], [12], [14], [16], [18]. Liben-Nowell *et al.* [12] 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.* [16] 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], [4], [5], [8], [13], [15], [17], [19], [20]. Romero *et al.* [15] studied information spread in Twitter and showed 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. [2] exploited information cascade to identify influencers in Twitter. Scellato et al. [17] also extracted geographic information from information dissemination process and utilized it to improve caching of mulitimedia files in a Content Delivery Network.

Although various measures are studied to explain the patterns of information cascade, there are possibly more standard measures to distinguish them in different topics, for instance, earthquake and political topics. Besides, it is still unclear which measure are the most effective. We thus explore three measures, which are cascade ratio, tweet ratio, and time of tweet, to express the cascade patterns in various topics.

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. Our data collection consists of user profiles, timestamp and tweet contents including retweets. We started crawling from famous Japanese users, Japanese users who have many followers. We firstly got timelines of these users, then repeatedly expanded the set of users by tracing retweets and mentions in their timelines. We then obtained 260 million users as active users and 783 million tweets. 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. This is because retweet-mention relationship is stronger than friend-follower relationship. We extracted 31 million links by considering only active users.

To study information cascade according to different topics, we treat a hashtag as a representative of the topic users talk about. We select top 500 frequently used hashtags from the dataset and categorize them according to topics. We have six major categories which are earthquake, politics, media, entertainment, sports, and idiom. Table1 shows examples of hashtags in each category. First, earthquake category is mainly about the Great East Japan Earthquake. Second, politics category is related to political issues and events all over the world. Many of them refer to the uprising events in the Middle East. Third, media category is represented by communication channels, such as, television networks, news channels, and video sharing websites. Forth, entertainment category refers to television programs, movies and artists especially Japanese animations. Fifth, sports category corresponds to sports teams and tournaments. Most of them are Japanese baseball teams. Finally, idiom topic is a popular phrase used as Twitter culture. Although it is still unclear that the idiom topic should be really treated as the topic or not, we include this in our work because it was studied by Romero *et al.* [15].

4. Measures of Users' Influence and Posting Behaviors

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 the number of his/her immediate incoming neighborhoods that reposted the hashtag after him/her as shown in Fig.1. A

Table 1 Examples of hashtags in each topic

Topic	Total	Examples
Earthquake	54	jishin, genpatsu, prayforjapan, save_fukushima, save_miyagi
Media	49	nicovideo, nhk, news, fujitv, cnn
Politics	102	bahrain, iranelection, wiunion, teaparty, gaddafi
Entertainment	85	madoka_magica, akb48, atakowa, tigerbunny, anohana
Sports	20	hanshin, f1jp, dragons, sbhawks, cwc2011
Idiom	41	nowplaying, shoutout, followme, justsaying, pickone

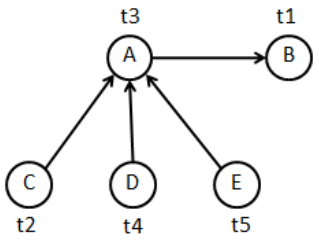


Fig. 1 An example of hashtag cascade

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 are seeing user A's posts. When user A start to use a hashtag, only user D and E use the same hashtag after him. 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:

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

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 shows point-wise average cascade ratio distributions. 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. The red line is the point-wise average distribution of a particular topic, the blue line is the point-wise average distribution of all hashtags, and the green line is 90% confidence interval. In addition to the point-wise average distributions, we calculate the 90% bootstrap confidence intervals to test a null hypothesis. Using 95% confidence interval does not change resulting patterns. Our null hypothesis is that the particular topic has no difference in cascade ratio from a set of all hashtags. If 90% confidence interval do not contain average distribution of a topic, we can reject the null hypothesis and conclude by 90% confidence level that the topic has statistically significant difference in cascade ratio from the population. Otherwise, we cannot conclude by 90% confidence level that the topic has no difference in cascade ratio from the population.

According to Fig.2a, The earthquake, media, sports, and idiom topics have relatively low cascade ratio. People participating in these topics used hashtags independently not because of seeing from their friends' tweets. On the contrary, the political topic has relatively high cascade ratio. When people posted political hashtags, many of their friends started to post the same hashtags after them.

4.2 Tweet Ratio

The second measure is tweet ratio, the proportion of how many times a user uses a hashtag comparing to all tweets of the same hashtag. The tweet ratio tr of a user u posting a hashtag h is then simply defined as below:

$$tr(u,h) = \frac{T(u,h)}{\sum_u T(u,h)} \tag{2}$$

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

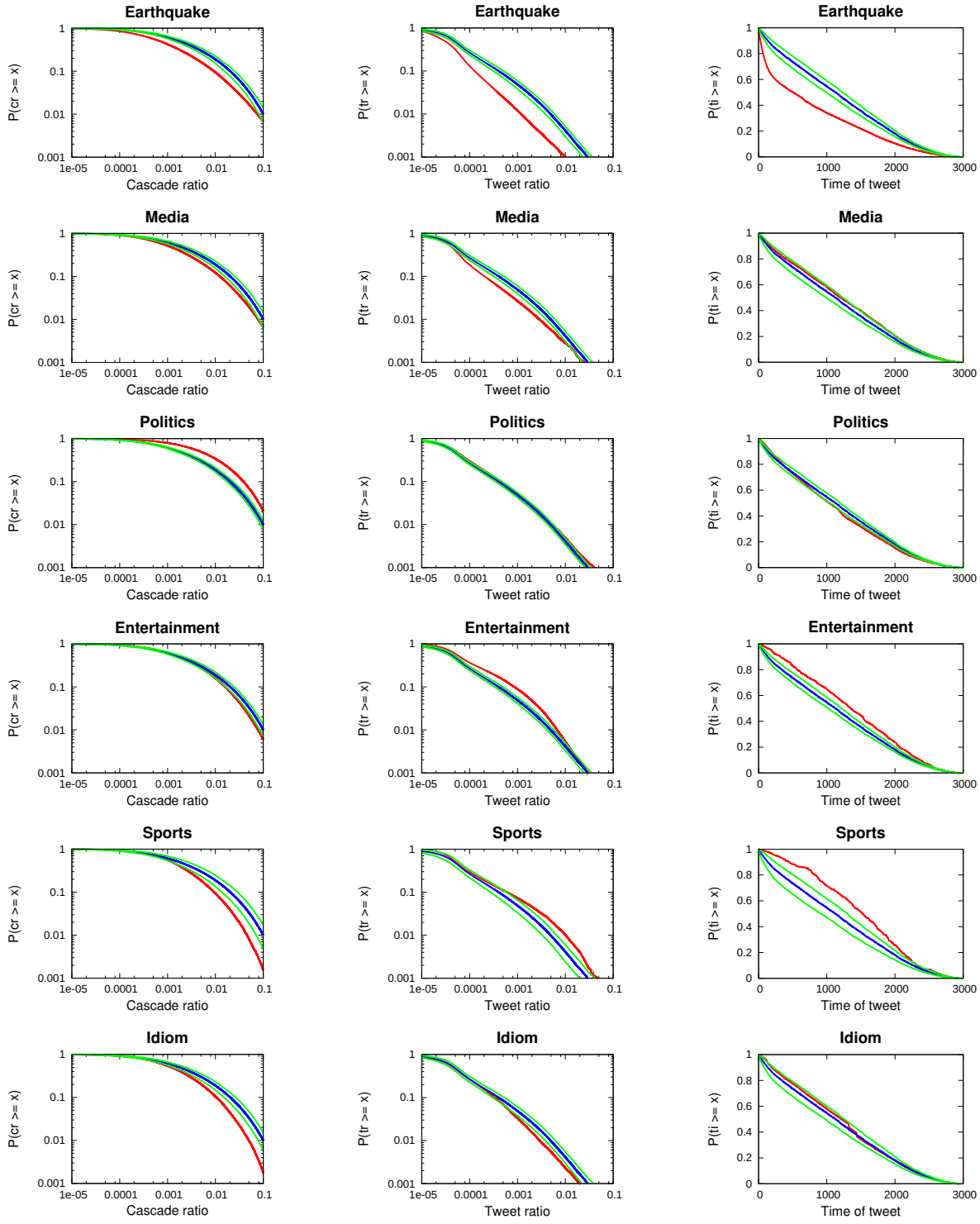
Fig.2b illustrates point-wise average tweet ratio distributions. x is tweet ratio and y is the number of occurrences of tweet ratios normalized by total number of users using a given hashtag. Each line is plotted in log-log coordinate and calculated as a cumulative distribution function, where y or $P(x)$ is the probability at a value greater than or equal to x. The red line is the point-wise average distribution of a particular topic, the blue line is the point-wise average distribution of all hashtags, and the green line is the 90% confidence interval.

The earthquake, media, and idiom topics have relatively low tweet ratio. People in these topics repeated to use same hashtags very few times. On the other hand, the political topic has relatively high tweet ratio. People repetitively posted same hashtags about the political topic many times.

4.3 Time of Tweet

The third measure is time of tweet which is time of each usage of a hashtag from its first appearance. The time ti of a tweet tw containing a hashtag h is then straightforwardly defined as the difference in time between tw and the first tweet of h.

Fig.2c shows point-wise average time distributions. x is time of tweet in hour(s) and y is the number of occurrences of time normalized by total number of tweets comprising a given hashtag. Each line is plotted as a cumulative distribu-



(a) Cascade ratio (b) Tweet ratio (c) Time of tweet

Fig. 2 Point-wise average distributions of each topic

tion function, where y or $P(x)$ is the probability at a value greater than or equal to x. The red line is the point-wise average distribution of a particular topic, the blue line is the point-wise average distribution of all hashtags, and the green line is the 90% confidence interval.

The earthquake topic falls down at first period. A large number of tweets were posted soon after the topics were raised to Twitter and gradually decreased when time passed. We can imply that people talked very much about the Great East Japan Earthquake during that time and in turn rarely said about it when the situation was back to normal. Conversely, the entertainment and sports topics lay in a diagonal. The number of tweets did not change according to time. People continually talked about these topics during the period of time.

Table 2 Patterns of hashtag cascades in each topic

Topic	Cascade ratio	Tweet ratio	Time of tweet
Earthquake	L	L	L
Media	L	L	-
Politics	H	-	-
Entertainment	-	H	H
Sports	L	H	H
Idiom	L	L	-

Table 3 Clustering result when $k = 6$

No. of hashtags	c0	c1	c2	c3	c4	c5
Earthquake	25	9	1	6	7	0
Media	1	20	1	13	9	2
Politics	0	4	41	3	31	15
Entertainment	0	11	5	37	6	6
Sports	0	2	0	17	0	1
Idiom	1	15	1	8	10	0

4.4 Patterns of Topic-Sensitive Hashtag Cascades

By using cascade ratio, tweet ratio, and time of tweet, we summarize patterns of hashtag cascades according to six major topics as in Table 2. "H" means high, "L" means low, and - means No statistically significant difference from the population.

The earthquake topic has low cascade ratio, low tweet ratio, and short lifespan. The media and idiom topics have same patterns, which are low cascade ratio and low tweet ratio. The political topic has high cascade ratio. The entertainment and sports topics have similar patterns, which are high tweet ratio and long lifespan, and the sports topic additionally has low cascade ratio.

5. Relationships between Cascade Patterns and Topics

In this section, we further investigate the relationship between cascade patterns and popular topics in Twitter and examine the effectiveness of each measure we described in earlier section. We perform k-means clustering based on the distributions of cascade ratio, tweet ratio, and time of tweet. Each hashtag is represented as a vector of values captured from n points in each distribution. For each hashtag, we select n=93 points proportional to the x-axis. For example, in case of time of tweet, our interested range of the x-axis is 0-3000. Then we divide this range into 93 bins and capture the y-axis value at each bin.

We use Euclidean distance as a distance measure and randomly assign each hashtag to a cluster at initialization. Considering six major topics in our study, we vary the number of clusters as $k = 6, 7, 8$. Since k-means algorithm provides different results depending on the initialization, we perform five trials for each k and evaluate clustering results by using normalized mutual information (NMI). Instead of other evaluation measures such as purity and F measure, it can be used to compare clustering quality with different numbers of clusters. For each trial, we compute NMI to evaluate clustering results. We then pick up the trial that provides the highest NMI at each k. Since those results when $k = 6, 7, 8$ have the same trend, we then choose the result of $k = 6$ to consider throughout this study.

Table 3 illustrates clustering result when $k = 6$. We can conclude that hashtags from the same topic or the topics having similar patterns of cascade are assigned into the same cluster. For example, the majority of the earthquake topic are assigned into cluster 0. In the same way, because the media and sports topics have same cascade patterns, the majority of these two topics are put together into cluster 1.

However, some of them even from the same topic have different behaviors and thus put into other clusters. For example, the hashtags in the earthquake topic are mainly divided into cluster 0, 1, and 4. The hashtags in cluster 0 are directly related to the Great East Japan Earthquake such as "jishin", "save_miyagi", and "84ma" (Operation Yashima). On the other hand, the earthquake hashtags in cluster 1, which the majority of the media topic are assigned to, are hashtags such as "iwakamiyasumi" (a journalist who spread information about nuclear power plant after the accident at Fukushima Daiichi Nuclear Power Plant) and "nicojishin". We can see that they are somehow related to the media topic. Likewise, the earthquake hashtags in cluster 4, which its major members are the political topic, are hashtags such as "fukunp" and "cnic" (Citizen's Nuclear Information Center). Because they are about the nuclear power plant which needs the Japanese government to concern and take actions on, they are said to be political-related.

In the same way as the media hashtags, they are primarily split into cluster 1, 3, and 4. The hashtags in cluster 1 are Japanese television media such as "fujitv", "nhk", and "tvasahi", while the media hashtags in cluster 3 are Japanese Internet media such as "r_blog" (Rakuten blog), "ameblo" (Ameba blog), and "2chmatome". Furthermore, the media hashtags in cluster 4, which its major members are again the political topic, are hashtags such as "aljazeera", "wikileaks", and "alarabiya". Since these kind of media mainly serve political news, they are thus said to be political-related too.

Lastly, the entertainment and sports hashtags are largely assigned into the same cluster, cluster 3. The entertainment hashtags here are Japanese animations and artists such as "tigerbunny" and "akb48" respectively, while the sports

hashtags are Japanese baseball teams such as "hanshin" and "dragons". It is probably that both of them are hobbies, gain much interest from their fans and thus share common behaviors.

Due to the above analysis, it is interesting that we can discover hidden relationship between topics by using only three measures rather than seeing tweet content

6. Conclusion

We studied the patterns of information cascade in six popular topics in Twitter, which are earthquake, media, politics, entertainment, sports, and idiom. We found that different topics mostly have different patterns of hashtag cascades in term of cascade ratio, tweet ratio, and time of tweet. For example, the earthquake topic has low cascade ratio, low tweet ratio, and short lifespan, while the political topic has high cascade ratio. However, some hashtags even in the same topic have different cascade patterns. For instance, the earthquake hashtags can be divided into the hashtags directly related to the Great East Japan Earthquake, the media-related hashtags, and the political-related hashtags or the hashtags about the nuclear power plant. We discover that such kind of hidden relationship between topics can be surprisingly revealed by using only three measures rather than considering tweet contents.

Finally, as future work, we need to explore other useful characteristics such as expert level of individual users and verify which measures are the most appropriate to explain patterns of hashtag cascades in different topics. Moreover, we need to investigate other clustering algorithms and other similarities whether they still provide the same results or not.

References

[1] E. Adar, L.A. Adamic, Tracking Information Epidemics in Blogspace. In 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI2005), pp. 207–214, 2005.

[2] E. Bakshy, J.M. Hofman, W.A. Mason, D.J. Watts, Everyone's an Influencer: Quantifying Influence on Twitter. In 4th ACM International Conference on Web Search and Data Mining (WSDM2011), pp. 65–74, 2011.

[3] E. Bakshy, B. Karrer, L.A. Adamic, Social Influence and the Diffusion of User-Created Content. In 10th ACM Conference on Electronic Commerce (EC2009), pp. 325–334, 2009.

[4] C. Castillo, M. Mendoza, B. Poblete, Information Credibility on Twitter. In 20th International Conference on World Wide Web (WWW2011), pp. 675–684, 2011.

[5] M. Cha, H. Haddadi, F. Benevenuto, K.P. Gummadi, Measuring User Influence in Twitter: The Million Follower Fallacy. In 4th International AAAI Conference on Weblogs and Social Media (ICWSM2010), pp. 10–17, 2010.

[6] D. Gruhl, R. Guha, D. Liben-Nowell, A. Tomkins, Information Diffusion Through Blogspace. In 13th International Conference on World Wide Web (WWW2004), pp. 491–501, 2004.

[7] D. Kempe, J. Kleinberg, E. Tardos, Maximizing the Spread of Influence through a Social Network. In 9th ACM SIGKDD Knowledge Discovery and Data Mining (KDD2003), pp. 137–146, 2003.

[8] 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 (WWW2010), pp. 591–600, 2010.

[9] J. Leskovec, L.A. Adamic, B.A. Huberman, The Dynamics of Viral Marketing, ACM Transaction on the Web, 1(1):5, May 2007.

[10] J. Leskovec, L. Backstrom, J. Kleinberg, Meme-tracking and the Dynamics of the news Cycle. In 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2009), pp. 497–506, 2009.

[11] 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 (SDM2007), pp. 551–556, 2007.

[12] D. Liben-Nowell, J. Kleinberg, Tracing Information Flow on a Global Scale Using Internet Chain-Letter Data. In the National Academy of Sciences, 105(12):4633–4638, March 25, 2008.

[13] 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 (WWW2011), pp. 517–526, 2011.

[14] M.E.J. Newman, S. Forrest, J. Balthrop, Email Networks and the Spread of Computer Viruses. Physical Review E, 66(035101), 2002.

[15] 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 (WWW2011), pp. 695–704, 2011.

[16] E. Sun, I. Rosenn, C. Marlow, T. Lento, Gesundheit! Modeling Contagion through Facebook News Feed. In 3rd International AAAI Conference on Weblogs and Social Media (ICWSM2009), pp. 146–153, 2009.

[17] 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 (WWW2011), pp. 457–466, 2011.

[18] D.J. Watts, A Simple Model of Global Cascades on Random Networks. In the National Academy of Sciences, 99(9):5766–5771, April 30, 2002.

[19] J. Weng, E.-P. Lim, J. Jiang, Q. He, TwitterRank: Finding Topic-Sensitive Influential Twitterers. In 3rd ACM International Conference on Web Search and Data Mining (WSDM2010), pp. 261–170, 2010.

[20] 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 (WWW2011), pp. 705–714, 2011.