8-2014

On macro and micro exploration of hashtag diffusion in Twitter

Yazhe WANG
Singapore Management University, yazhe.wang.2008@phdis.smu.edu.sg

Baihua ZHENG
Singapore Management University, bhzheng@smu.edu.sg

DOI: https://doi.org/10.1109/ASONAM.2014.6921598

Published in Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining: ASONAM 2014 : Beijing, China, August 17-20, 2014, Pages 285-288.
http://doi.org/10.1109/ASONAM.2014.6921598

On Macro and Micro Exploration of Hashtag Diffusion in Twitter

Yazhe Wang
School of Information Systems
Singapore Management University
Singapore
yazhe.wang.2008@phdis.smu.edu.sg

Baihua Zheng
School of Information Systems
Singapore Management University
Singapore
bhzheng@smu.edu.sg

***Abstract*—This exploratory work studies hashtag diffusion in Twitter. The analysis is conducted from two aspects. From the macro perspective, we study general properties of hashtag diffusion, and classify hashtags into three main classes based on their temporal dynamics referred as "single spike", "multi-spikes", and "fluctuation", and find that each of these classes has some unique characteristics. From the micro perspective, we investigate individual diffusion. We adopt Edelman's "topology of influence" theory to identify four type of users with different influence levels in diffusion based on their dynamic retweet behaviors. The results of our study are useful for gaining more insights of information diffusion in Twitter.**

***Keywords*-information diffusion; Twitter; hashtag;**

I. Introduction

Twitter is a fast growing online social media, which is featured by fast information diffusion. In this work, we study hashtag diffusion in Twitter. Hashtags are the user specified topic keywords in tweets prefixed by "#". An information diffusion process of a hashtag involves all the tweets containing the hashtag. Each tweet has a timestamp of when it was published and the user information of who published this tweet, and possiblly the id of the original tweet which it retweeted.

We first study general properties of hashtag diffusions in Twitter. We classify hashtags into three main classes based on their temporal dynamics referred as "single spike", "multi-spikes", and "fluctuation". We find that temporal dynamics of hashtags are closely related to their semantics, and each of these classes has some unique characteristics. Then, we investigate each individual diffusion. We adopt Edelman's "topology of influence" (TOI) theory to identify four categories of key players in a diffusion namely idea starters, amplifiers, adapters, and commentators. We propose a quantitative metric for each category, and study the properties of the users in different categories. We find the role that a user plays in a diffusion is related to his structural properties in the network. Our findings are useful for gaining insights of the mechanism of information diffusion in Twitter.

II. Related Works

There are many works that study the properties of information diffusion in social networks. For example, Kwak, et al. study retweet diffusions in Twitter [1]. They find that most of the diffusions do not go beyond one hop in the network and have durations no more than one day. Different from their work, we study hashtag diffusion, and we focus more on the temporal dynamics of diffusion. Gruhl, et al. categorize topic diffusions in blogspace by their daily frequency patterns, and discuss two patterns: sustained chatter and sharp rise spikes [2]. These patterns are similar with the fluctuation pattern and spiky pattern of hashtag diffusion that we find in this work. Lehmann, et al. also study the temporal patterns of diffusion, but focus on the patterns with single peak [3]. Our study is not limited to diffusions with single spike as we also study diffusions with multi-spikes and fluctuation patterns. Budak, et al. study topics that are diffused through the news media in Twitter, and address structural properties of the topics (i.e., whether the topics are diffused among clustered or distributed users) [4]. While we examine more features that are both structural and non-structural.

In this work, we also study users' role in each individual diffusion. Most of the existing works that study users' role in information diffusion focus on only one category: the information spreaders/influential users, who trigger large cascade of diffusion [5]–[8]. However, we consider the large-scale diffusion in Twitter as the result of the collaboration of users with different roles rather than the effect of some influential users only. We study four categories of user roles in a diffusion process based on Edelman's TOI theory. Tinati, et al. examine the user categories also based on Edelman's TOI theory [9]. However, the metrics for these categories defined in our work are different from theirs. Moreover, they provide detailed discussion only on the idea starters, but we analyze all of the four categories.

III. Dataset Description

We use a dataset that contains 12 million tweets posted by 46,560 Singapore Twitter users from May 1, 2012 to May 30, 2012. We extract hashtags from this dataset. As we are only interested in relatively large-scale diffusions, we filter out all the hashtags that appear in less than 100 tweets. Then, we exclude potential spam hashtags that are used by only a few users and non-English hashtags.Finally, we obtain a total of 153 hashtags. We also construct a user network based

Table I: Basic properties of the hashtag diffusion studied.

Property	Description
Hashtag length	The number of characters of the hashtag.
Tweet/User size	The number of tweets/users that contain/post the hashtag.
Average user tweets	The average number of tweets containing the hashtag generated by a user.
Duration	The number of days between the first appearance and last appearance of the hashtag in the dataset.
Tweet spreading speed	The average number of tweets posted containing a hashtag per day.
User spreading speed	The average number of users who post about a hashtag per day.
Retweet ratio	The proportion of retweets among all the tweets (both original tweets and retweets) containing the hashtag.
User link density	The density of links (in the user mention network) among the users who post tweets containing the hashtag.
Temporal pattern	The time series of the daily frequency of tweets containing the hashtag.

Table II: Hashtags in the three classes.

Class	Num. of hashtags	Examples
Single spike	59 (38.5%)	HappySunnyDay, SS4EncoreDay2, 4yearswithSHINee, NanHuaProblems NBASG12, HappyAnniveSHINee4th
Multi-spikes	20 (13.1%)	ChannelUJump, F1, FF, Hougangby-election, LFC, LionsXII
Fluctuation	74 (48.4%)	Nowplaying, TWFanmiIy, Travel, beauty, business, fashion

Table III: Properties of the hashtags in three classes.

Property (avg.)	Single spike	Multi-spikes	Fluctuation
Hashtag length	11.6	8.0	7.0
Tweet size	681.9	1052.5	3304.2
User size	152.5	229.5	525.0
Average user tweets	2.7	3.4	7.0
Duration	13.5	23.5	27.3
Tweet spreading speed	119.4	56.5	120.4
User spreading speed	30.5	9.67	19.1
Retweet ratio	40%	49%	32%
User link density	0.035	0.029	0.012

on mention interactions between crawled Twitter users. We form a directed link from a user A to a user B if A mentioned B in his tweets, and we assign the total number of times that A mentioned B as the weight of the link.

IV. Analysis of Results

A. General Properties of Diffusion

A diffusion is defined by a hashtag with the tweets containing the hashtag and the users who generate these tweets. We study various interesting properties of hashtag diffusion as summarized in Table I.

First, we classify the hashtags based on their temporal patterns (i.e., daily frequency of tweets containing a hashtag) manually, and identify three main classes of patterns:

- *Single spike pattern*: The appearances of the hashtags of this pattern during the studied time period change from infrequent to very frequent suddenly, and then drop to infrequent drastically.
- *Multi-spikes pattern*: The occurrences of the hashtags of this pattern show multiple spikes during the studied time period.
- *Fluctuation pattern*: Hashtags with this pattern appear continuously with moderate frequencies through a long period of time.

Table II lists the number of hashtags and some random examples in each class. We find that in our dataset, single spike and fluctuation patterns consist of the main stream of hashtag diffusion in Twitter (i.e., around 87% of the hashtags belong to these two classes). Therefore, these two classes are of the most interest to our study.

We further characterize the three classes of hashtags with other properties. Table III lists the average value of the studied properties. We observe that the hashtags with the single spike pattern have longer length than those of the other two patterns. It is because these hashtags are mostly related to external events or specific topics, and long phrases are usually used for description (e.g., HappyAnniveSHINee4th and NanHuaProblems). Whereas, the fluctuation pattern is usually related to general topics described by simple words or short phrases (e.g., beauty and travel). Therefore, temporal patterns are closely related to the semantic of hashtags. We then find that the hashtags of single spike pattern have smaller tweet size and user size, and the hashtags of fluctuation pattern have the largest tweet size and user size. Moreover, the single-spike hashtags have significantly shorter lifespan, and the fluctuation hashtags last for the longest time. Intuitively, this could be explained by the nature of different ways that people engage in event-specific topics and general-interest topics. An event-specific topic is usually interesting to a specific (relatively small) group of people, and the discussion is only hot within a short period around the time the event happened. However, a general-interest topic is widely acceptable by many people, and its attractiveness is long lasting. In addition, by examining the average user tweets, we find that Twitter users tend not to generate many tweets about event-specific hashtags with spiky patterns. However, they like to repeatedly contribute to general-interest hashtags with the fluctuation pattern. Next, we calculate the spreading speeds of the hashtags in terms of the tweets and the users. We find that the single-spike hashtags are almost as efficient as fluctuation hashtags on engaging large quantities of tweets because they have almost the same tweet spreading speed. However, the single-spike hashtags are more efficient on engaging users according to the user spreading speed. In addition, we find that the

single-spike hashtags have the largest retweet ratio (i.e., 40%), although the retweet ratio of the fluctuation pattern is slightly lower (i.e., 32%). These indicate that users like to re-share other users' ideas when discussing event-specific topics, while they are more willing to generate their own ideas when talking about general-interest topics. Finally, we calculate the user link density based on the user mention network. We find that the users involved in the single-spike hashtag diffusions are more densely connected than the users involved in the fluctuation hashtag diffusions. This implies event-specific topics tend to distribute among the users that are more closely related, while general-interest topics may distribute through the users in different local communities.

B. Users' Role in Diffusion

In this section, we apply Edelman's *topology of influence* (TOI) theory [10] to understand different user roles in each individual hashtag diffusion. The TOI theory profiles users by the following five categories based on how their social behaviors fit into *online* communication channels: *Idea starters* (IS) like to start new ideas during a conversation. *Amplifiers* (Amp.) share opinions of others rather than generate their own, and enjoy being the first one to do so. *Adapters* (Ad.) read memos from a broad context outside of their traditional sphere of knowledge, and tailor them to their niche groups. *Commentators* (Com.) do not usually initiate new ideas but like to add comments. *Viewers* do not contribute to the conversation, but only consume the information.

The TOI theory only provides conceptual descriptions of the five user categories. In our study, we redefine these concepts to fit the Twitter hashtag diffusion context, and propose a quantitative metric for each of them.

We define the idea starters as the users whose tweets are frequently retweeted by other users in a conversation. We use the term "conversation" and "hashtag diffusion" interchangeably. We define a score function for idea starters as the average number of retweets that a user gets for each of his tweets.

$$S_{IS}(u) = \frac{\sum_{t \in T_u} |RT^t|}{|T_u|}, \tag{1}$$

where u is a user, t represents a tweet, T_u is the set of original tweets that u published in a conversation, and RT^t is the set of retweets of t.

Then, we define an amplifier based on the number of times that he is ranked as the first few who retweet a tweet in a conversation. Equation 2 provides the score function.

$$S_{Amp.}(u) = \sum_{t \in RT_u} \frac{|RT^{t.orig.}|}{rank(t, RT^{t.orig.})} * |RT_u^{first}|, \tag{2}$$

where RT_u is the set of retweets that are generated by u, $t.orig.$ is the original tweet that t reweets given t is a retweet, $rank(t, RT^{t.orig.})$ evaluates the rank of t in the set $RT^{t.orig.}$ of retweets that re-post $t.orig.$ sorted based on ascending order of the published time, and RT_u^{first} is the set of retweets that are generated by u and are the first ones in the retweet chains.

Table IV: Proportions of users in different role categories.

Hashtag class	IS	Amp.	Ad.	Com.
All Hashtags	5.6%	10.9%	2.8%	8.9%
Single spike	6.0%	9.6%	2.7%	6.7%
Multi-spikes	5.9%	13.7%	3.1%	8.0%
Fluctuation	3.6%	8.2%	2.5%	10.4%

We define an adapters as a user who retweet from many other different users. Equation 3 defines the score function.

$$S_{Ad.}(u) = |RTU_u|, \tag{3}$$

where RTU_u is the set of users from whom u has retweeted.

We define a commentator as a user who actively tweet many times in a conversation, but is not retweeted by many other users. Equation 4 defines the score function.

$$S_{Com.}(u) = \frac{|T_u|}{\sum_{t \in T_u} |RT^t| + 1} \tag{4}$$

Unfortunately, it is difficult to capture the footprints of viewers who only read tweets on Twitter. Thus, this user category is not examined.

We locate the users of different categories by calculating the scores of the four role categories of each user in each hashtag diffusion. Then we assign a user to a role category if his score of that category is above the average of the none-zero scores of all the users in the diffusion. We first examine the proportions of the users of each category (see Table IV). We find that interestingly for the hashtags with the fluctuation pattern, the proportion of the idea starters (i.e., 3.6%) is noticeably lower than that of the other two classes, while the proportion of the commentators is higher. It shows less influence of the idea starters in the discussions of general-interest topics, and the ordinary users have more genuine ideas to contribute. However, the discussions of event-specific topics are more influenced by the idea starters.

Next, we study properties of the users within different role categories. We study 5 network structural properties based on the user mention network, including in-degree (d_{in}), out-degree (d_{out}), total degree (d_{total}), betweenness (btw), in-out degree ratio ($\frac{d_{in}}{d_{out}}$), boundary spanner ($\frac{btw}{d_{total}}$), and 1 activity (act) property evaluates the average number of tweets generated by a user daily.

Table V displays the average values of the studied properties. By observing the degree properties, we find that the idea starters have significantly higher in-degree than the users of other categories, they are the most popular users who are mentioned by many other users, and the adapters have the highest out-degree, they are the users who mention many other users. The commentators have the lowest in-degree, so they are the less popular users who are less mentioned by others. Moreover, we find that the idea starers have the

Table V: Characterizing the users of different roles based on their properties.

Properties	IS	Amp.	Ad.	Com.	All users
d_{in}	89.1	24.4	24.1	14.7	22.7
d_{out}	25.7	23.1	30.9	21.1	20.0
d_{total}	114.8	47.5	55.0	35.8	42.8
btw	1.9E6	8.4E5	9.0E5	6.0E5	7.3E5
$\frac{d_{in}}{d_{out}}$	8.9	0.6	0.6	0.4	1.2
$\frac{btw}{d_{total}}$	1.2E4	9.9E3	1.3E4	9.7E3	9.3E3
act	32.8	26.2	38.4	29.9	24.7

most uneven in-out degree ratio, they have much more incoming links than the out-going links. While the users in the other three categories tend to have less in-links than out-links. Next, we find that the idea starters have the highest betweenness value followed by the adapters. We believe that the high betweenness of the idea starters is caused by the significantly high degree value, and the betweenness value of the adapters is also high because they are the users who tend to lie on the boundaries of different local communities to re-share ideas. To further verify these, we calculate the boundary spanner score. We find that the adapters do have the highest boundary spanner score. Moreover, we find that the commentators have the lowest betweenness values. Therefore, they are the relatively marginalized users in the network. It explains why their tweets do not get much attention from other users. Finally, an interesting finding is that among all these four categories, it is the adapters but not the idea starters who are the most active. It indicates that the idea starters are the users who generate ideas with high quality rather than large quantity. However, the adapters, who like to combine and re-share ideas from many others users, are the most busy and active ones.

V. Conclusions

In this work, we analysis hashtag diffusion in Twitter. From macro aspect, we examine general properties of diffusion, and find two typical classes of hashtags based on temporal dynamics, namely single spike and fluctuation. The hashtags of these two categories are evidently different in their semantics as well as many other properties. The findings provide incentives for designing different models to simulate the hashtag diffusion of different classes. Then, from the micro aspect, we analysis users' role in individual hashtag diffusion based on Edelman's TOI theory. We define quantitative metrics to locate four types of users, and find the role that a user plays is related to his network structural.

Our results in this work are promising, however are limited by the scope of the dataset. It is an interesting open problem to investigate generality of our findings.

Acknowledgment

This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office, Media Development Authority (MDA). This research is also funded through a research grant (12-C220-SMU-007) from MOE's AcRF Tier 1 funding support through Singapore Management University.

References

[1] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 591–600.

[2] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information diffusion through blogspace," in *Proceedings of the 13th International Conference on World Wide Web*, ser. WWW '04. New York, NY, USA: ACM, 2004, pp. 491–501.

[3] J. Lehmann, B. Gonçalves, J. J. Ramasco, and C. Cattuto, "Dynamical classes of collective attention in twitter," in *Proceedings of the 21st International Conference on World Wide Web*, ser. WWW '12. New York, NY, USA: ACM, 2012, pp. 251–260.

[4] C. Budak, D. Agrawal, and A. El Abbadi, "Structural trend analysis for online social networks," *VLDB Endowsment*, vol. 4, no. 10, pp. 646–656, Jul. 2011.

[5] M. Kitsak, L. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. Stanley, and H. Makse, "Identification of influential spreaders in complex networks ," *Nature Physics*, vol. 6, no. 11, pp. 888–893, Aug 2010.

[6] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '03. New York, NY, USA: ACM, 2003, pp. 137–146.

[7] C. Lee, H. Kwak, H. Park, and S. Moon, "Finding influentials based on the temporal order of information adoption in twitter," in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 1137–1138.

[8] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, "Everyone's an influencer: quantifying influence on twitter," in *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, ser. WSDM '11. New York, NY, USA: ACM, 2011, pp. 65–74.

[9] R. Tinati, L. Carr, W. Hall, and J. Bentwood, "Identifying communicator roles in twitter," in *Proceedings of the 21st International Conference on World Wide Web*, ser. WWW '12 Companion. New York, NY, USA: ACM, 2012, pp. 1161–1168.

[10] H. Jonathan, "The fire hose, ideas, and topology of influence@ONLINE," Jun 2010. [Online]. Available: http://www.edelmandigital.com/2010/06/24/the-fire-hose-ideas-and-topology-of-influence/