23rd Workshop on Information Technology and Systems (WITS2013),
Milan, Italy, December 2013

User Roles in the Time of Crisis: A Social Media Analysis

Nargis Pervin, Remy Cazabet, Hideaki Takeda, Fujio Toriumi and Anindya Datta

Abstract

Dynamics of information propagation in Twitter has been studied in the context of retweet practices. In order to build the network of information flow, either the activity network or the follower network has been considered. However, because of unstandardized retweet practices the bias in diffusion analysis is not avoidable in prior approaches. By combining both activity network and follower network, we introduce the concept of "Information Diffusion Impact" (IDI) on network. With a Twitter dataset during the Great Eastern Japan Earthquake we characterize important user roles in information propagation in the time of crisis and discuss evolution of user roles over time.

1 Introduction

Microblogging services such as Twitter turns out to be the most popular service by far which can disseminate up-to-the-minute information very quickly. This massive information resource has proven to be useful for event detection [10], political uprising [1], and organizing people in the time of crisis [6, 9]. However, little is known about how information is propagated and people behave during natural calamity.

The principal factor of information diffusion in Twitter is the so-called act of retweeting, which allows users to share information published by others. Diffusion of information in Twitter has already been studied by several researchers in the context of trending topic detection [10], modeling retweetability of a tweet [2, 11], and adoption cascades [5]. But the different roles played by users in this diffusion of information needs further investigation.

Chu et al.[3] have proposed a system to classify Twitter accounts into human, bots, and cyborgs (human assisted bots or bot assisted humans) computing entropy of inter-tweet delays. Tinati et al. [12] have proposed a model to classify several types of users, that they call Idea Starters, Amplifiers, Curators, Commentators, and Viewers. However, as we will explain in Section 3, their retweet network may suffer from a strong bias. Furthermore, the metrics used to identify the roles of the users are based on a specific model, called 'Topology Of Influence'. On the contrary, in this paper, we propose definitions of user roles directly grounded on the impact the users have on the network. Our goal is in two folds: first we analyze the retweet dataset along with the follower network to characterize different user roles in the context of information diffusion. Secondly, we study the change of user roles in the time of crisis.

Herein, we present a realistic approach of finding the retweet chain from tweet content. Upon analysing the retweet chain we classify user roles in three categories, namely, "idea starter", "amplifier", and "transmitter" in the context of information diffusion in the network. Later, we analyze the change of user roles in the time of crisis using a Twitter dataset of the Great East Japan Earthquake in 2011.

This paper is organized as follows. Section 2 presents the dataset description used in this study. Section 3 describes the user role classification followed by experimental settings and results in Section 4 and Section 5 respectively. Finally, Section 6 concludes our findings.

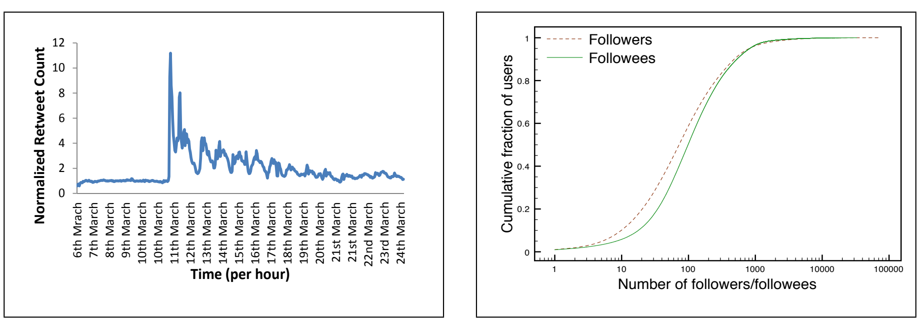
2 Dataset Description and Preparation

2.1 Dataset Description

Tweet Data: We used a Twitter dataset collected during the earthquake in 2011 described thoroughly in [13]. The dataset covers a period of 20 days (from 5^{th} March, 2011 to 24^{th} March, 2011), and consists of 362,435,649 tweets posted by 2,711,473 users in Japan. This dataset is remarkable by its completeness: the authors have checked that 80% to 90% of all published tweets by these users were present in this dataset.

Fig. 1(a) shows the retweet count for a period of 20 days normalized to cut off daily variations. The first two major peaks represent the two big earthquakes on 11^{th} and 12^{th} March as reported in [14]. After the disaster, retweet count progressively returns to its normal average values.

Follower Network Data: In Twitter, follower network depicts the social relationship among the users. Follower information has been collected by crawling Twitter API in May, 2013 for the active users who have been mentioned more than 20 times in 20 days. Follower network dataset consists of 300,104 users and 73,446,260 relationships information. Degree distribution has been shown in Fig. 1(b) by plotting cumulative fraction of users against the number of followers/followees of user.



(a) Tweet Distribution over days (Normalized)

(b) Cumulative fraction of users by degree

Figure 1: Retweet distribution and degree distribution

2.2 Dataset Preparation

In this work, we have investigated the activity network and the static follower network of the Twitter users simultaneously. In Twitter, the retweet functionality allows users to share information with their friends and followers, generating a network of retweeters. Here, this retweet network is considered as activity network. In our case, to analyze the information diffusion for each tweet, we were interested in the retweet sequence of each tweet, which we name as retweet chain.

How to find Retweet Chain: In most recent work, particularly the previous work proposing a classification of user's roles [12], the diffusion of information is directly extracted from the content of the tweets. For instance, if a tweet published by user u_1 is composed of the following pattern:

RT @u_0 *tweet* then one considers that the information diffused directly from u_0 to u_1. Retweet chains are identified by tweets containing several references, i.e *RT* @u_1 *RT* @u_0 *tweet*, or consecutive citations, such as u_1 posting the retweet : *RT* @u_0 *tweet* followed by u_2 posting *RT* @u_1 *tweet*.

However, in reality, after one step of citation, this has two important biases:

- users tend to keep only the original author of the tweet, and not intermediates, in particular to meet the 140 character limit of Twitter. Even when using the official Retweet function of Twitter, only the initial poster is kept, which will strongly increase the number of direct retweet and in turn the apparent role of the original poster in the diffusion of information.
- users frequently retweet after seeing a tweet several times, as it has been shown in [8]. As a result the user cited as source might not be fully representative of the information diffusion.

In this work, to characterise the diffusion of information, we therefore adopt a combination of both the follower network and the retweeter information from tweet. A retweet chain is simply defined as the sequence of all published tweets containing the original content, ordered by their publication time. To study the information flow, we combine this information considering that, each time a user publish a tweet, all his followers can see the information. We can therefore know by whom the user might have been informed, independently of the user who appears as source in the tweet itself.

3 User Role Classification

We classify user roles in the light of information propagation through retweeting. By analyzing retweet chain the users are classified into three categories, "idea starter", "amplifier", and "transmitter".

- Idea starters are the users who are able to launch new ideas which will spread broadly in the network. They are the users whose ideas will reach many.
- Amplifiers are users who do not publish interesting content by themselves, but have the potential to diffuse information published by others to many new people.
- Transmitters are users who act as bridges between several communities in the network. If an idea starter publish an interesting tweet in a given community of the network, amplifiers will spread this tweet in the same community, but transmitters are necessary to reach other communities, which in turn will result in transmission of the information more broadly.

We base our user role definitions on the concept of *Information Diffusion Impact (IDI)*, namely for a user, the number of times he made one user aware of one piece of information. Therefore, making 10 people aware of one information and making one person aware of 10 different pieces of information result in the same IDI value. This notion is very important, as it allows to compare

the impact of different roles. For each user, we can compute a value of *IDI* for each behavior (idea starter, amplifier, and transmitter), which represents the impact of this user on the diffusion of information: how many people were impacted by his publication of popular tweets? How many people became aware of a tweet by his action of retweeting? And how many people could access the information because the user transmitted it to another community? These values are therefore comparable.
Notations used in this paper has been listed in Table 1.

Notation	**Meaning**
N^t	Number of new people aware of tweet t
N_u	Number of new people made aware by user u
C_i	Community i
$follower_{C_i}(u)$	Follower set of user u in community C_i
$RC(t)$	Retweet Chain of tweet t
$order(t,u)$	Position of user u in the retweet chain of tweet t

Table 1: Notation Table

Idea Starter: Idea starter can be conceptualized as the one who creates the original information. Idea starters are important as their tweets are retweeted broadly in the network depending on the importance of the content. Using u's follower information, for each user u in the retweet chain of tweet t (*RC(t)*), we compute the number of new people (N_u) u makes aware of. The total number of people (N^t) in the network aware of the tweet t is given by
$N^t = \sum_{u \in RC(t)} N_u$.
Here, N^t is the number of different users made aware of the tweet, which is the impact of idea starter u for tweet t. Hence, the overall impact of u as an idea starter is defined by,

$$IdeaStarter(u) = \sum_t \sigma N^t, \sigma = \begin{cases} 1, & \text{if } order(t,u) = 1. \\ 0, & \text{otherwise.} \end{cases}$$

It is worthy to note that, to become a good idea starter it is not necessary to be followed by many people.
Amplifier: Amplifiers are considered as the individuals who share other's information and make many people aware of it. They are important as they are followed by many users and as a result, amplifiers make a large fraction of users aware of the information. Unlike idea starter, impact of amplifiers depend only on their immediate neighbors. For each tweet t, u participated but was not the idea starter, we compute the number of new people u makes aware of denoted by N^t.

$$Amplifier(u) = \sum_t \sigma N^t, \sigma = \begin{cases} 1, & \text{if } order(t,u) > 1. \\ 0, & \text{otherwise.} \end{cases}$$

We should note that this value is usually less than the number of followers of u, as some of his followers have already been made aware of the tweet. Therefore, users who appear early in the retweet chain, or who tend to inform users following few people will naturally have higher amplifier scores.
Transmitter: It is now accepted that most social networks have a strong community structure [4]. The Twitter follower network is no exception and its analysis reveals clearly defined modules. In

this paper, we used the Fast OSLOM algorithm [7] to detect communities in our follower network because it is fast and can find overlapping communities. The algorithm found 8 communities in our follower network with an average size of 44668 nodes per community. By manual investigation, we found obvious meaning for some communities, such as a community of foreigners and a community of users related to night life (disc jokeys, hip-hop celebrities, etc.).

We observed that communities play an important role in the diffusion of information. Most tweets are diffused only in a fraction of the communities. For instance, let us consider that a tweet is diffused in a community if it is retweeted by its users more than half as much as expected by picking users randomly. Then only 12% of tweets are diffused in the foreigner community of 43,375 users, while 64% of tweets are diffused in another community of comparable size. This indicates that tweets are not easily transmitted to the foreigner community. In contrary, more than 95% of tweets are considered transmitted to the largest community of size 150,235. We observed that only 38% of tweets that diffuse in the foreigners' community also diffuse in the largest community. This suggests that transmitters are needed to spread the information into or outside of this community.

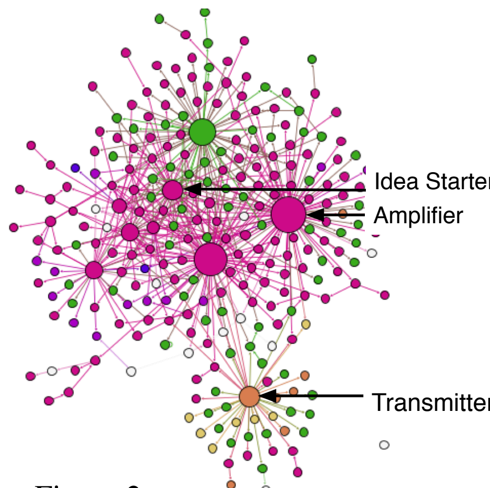


Figure 2: Retweet network for a popular tweet

Therefore, we identify transmitters as users who spread tweets initially stuck in a community A to another community B. We consider that a tweet is stuck in a community if the first 20 retweets are in the same community. Therefore, the first user from a different community B to retweet is considered as transmitter to B if he again gets retweeted by other people from B.

The impact of a transmitter for one tweet is simply the number of people who gain access to the information by his retweet. More formally, the effective number of users informed about the transmission of tweet t in community C_j is the summation of the number of followers of retweeters in community C_j.

$$Transmitter^t_{C_j}(u) = |follower_{C_j}(u)| + \sum_{u_k \in C_j} \sigma |follower_{C_j}(u_k)|, \sigma = \begin{cases} 1, & \text{if } order(t, u_k) > order(t, u). \\ 0, & \text{otherwise.} \end{cases}$$

where $order(t, u)$ is the position of user u in $RC(t)$ and $follower_{C_j}(u)$ represents the number of followers of user u in community C_j.

If a user belongs to several communities, he can be transmitter to different communities for a single tweet. If he does not transmit t to a community C_j, then $Transmitter^t_{C_j}(u) = 0$. $Transmitter^t(u) = \sum_j Transmitter^t_{C_j}(u)$

Hence, overall transmitter score of u for all tweet he is transmitter can be given by $Transmitter(u) = \sum_t Tranmitter^t(u)$

Figure 2 shows the retweet network corresponding to a popular tweet, where each node represents a user and an edge $v \rightarrow u$ exists if v follows u and $order(t, v) > order(t, u)$ in $RC(t)$. Node color represents the community it belongs and size of node is indicator of number of follower of the node in the network. Using our metric we identify the idea-starter, amplifiers, and transmitters in the retweet chain. One can note that the idea-starter is not followed by many people, as the size of the node is comparatively small. Amplifier is the one with many followers and well-connected in the network. On the other hand transmitter is the node from a different community where he diffuses the information.

4 Experimental Settings

Using Twitter dataset described in Section 2.1, we find the top 10,000 popular tweets, where popularity is measured by number of times it has been retweeted. For all these 10,000 tweets, retweet chain has been formed, which is basically the chain of users in chronological order of their retweet of the original post. Our tweet dataset (5^{th} March-24^{th} March) have been divided into three time windows, pre-crisis (5^{th} March-10^{th} March), during-crisis(11^{th} March-17^{th} March), and post-crisis(18^{th} March-24^{th} March). Using the follower information idea starter, amplifier, and transmitter IDI value of each user in all three time windows have been computed.

5 Experimental Results

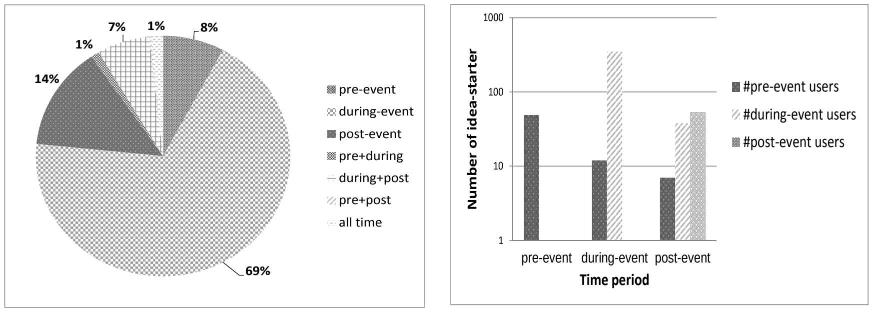
5.1 Evolution of User Roles over Time

In different tweets, one user might have different roles, idea starter/amplifier/transmitter. We have measured the individual impact for each role. In Figure 3 and Figure 4 we have computed the percentage of users who retained and disappeared as idea-starter and amplifier respectively in three time windows. Figure 3(a) shows the overall distribution of the idea starter in the three time-windows. 69% of the total idea starters were emerged only during the earthquake and 7% of the popular idea starters remained popular after the event also. To understand the transition of the idea-starter from one time-window to another, we analysed the proportion of the idea-starters in pre-event time-window who retained in other time-windows. From Figure 3(b), we see that out of 49 users in pre-event time-window, only 12 retained during the event. After the event it was only 7. Also, a large number (349) of idea-starters emerged during the event and 38 of them retained and 53 new users emerged after the event.

Similar analysis has been carried for amplifiers in Figure 4(a) and Figure 4(b). Overall distribution of the amplifier in Figure 4(a) shows that the number of new amplifiers only during the event is very high (95%) and only 4% of the amplifiers who emerged during the event continued to contribute after the event also. Figure 4(b) shows that popular amplifiers in pre-event time-window (13) tends to be popular during the event (11) and after the event (9), though a high number of amplifiers appeared only during the event (7400). Comparing Figure 3 and Figure 4, a large number of idea-starters and amplifiers in the post-event time-window were from during-event time-window.

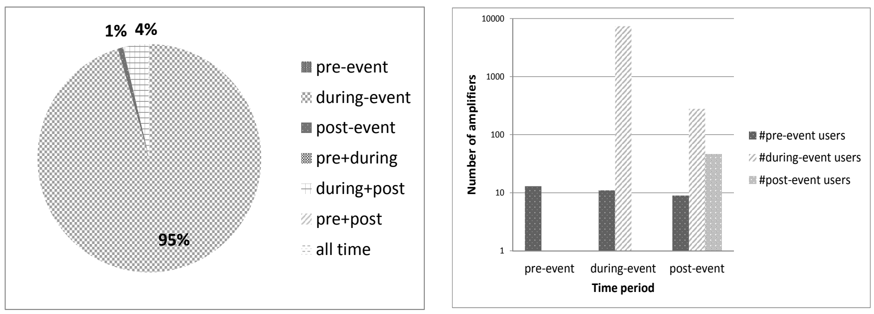
5.2 Association of User Roles

In Figure 5, we plotted (in log scale) Idea Starter Impact against Amplifier Impact for each user for three time-windows. We have divided the region into four quadrants as shown in Figure 5(d)). We have plotted only the users for whom sum of idea-starter impact and amplifier impact is at least 100,000, which means that overall the user has impacted 100,000 users in the network. In pre-event time-window (Figure 5(a)), number of high-impact idea starters (in quadrant 4) is comparatively larger than high-impact amplifiers (quadrant 2). Number of super-users in pre-event time-window



(a) In all time windows (b) Pre-and during event window

Figure 3: Distribution of role retention as idea starter in pre-, during- and post-event time window

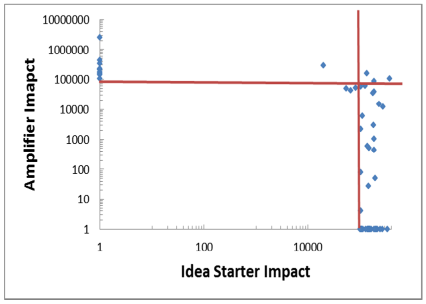


(a) In all time windows (b) Pre-and during event window

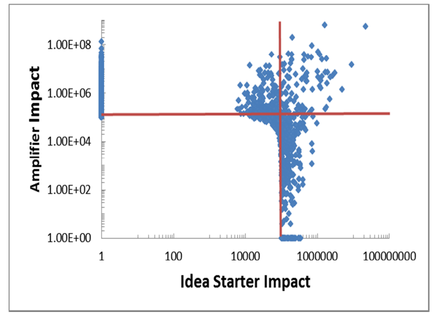
Figure 4: Distribution of role retention as amplifier in pre-, during- and post-event time window

is comparatively lower than other time-windows. In during-event time window (Figure 5(b)), there is a strong increase in the number of users in all quadrants and number of super-users are maximum during the event, contributing a lot in launching important ideas and spreading to others in the network. Also we have observed that many idea-starters started behaving as amplifier during the disaster. For instance, the user ‘earthquake_jp’ is a bot which posts information on earthquake and seismic intensity. In the pre-event time window it acted only as good idea starter. However, during the event it started retweeting others’ tweets and became potential amplifier, as commonly referred as cyborg in the literature [3]. Interestingly, after the event (Figure 5(c)) it became again a bot. Unlike ‘earthquake_jp’, user ‘nhk_pr’ was idea-starter as well as amplifier in all time. Particularly during the disaster, he became very popular both as idea starter and amplifier and also remains as potential idea starter and amplifier after the event. Interestingly, in the post-event time-window, many users were observed with high impact in dual roles and some new users emerged as potential idea starter and amplifier after the event and the number of super-users increases compared to pre-event time-window.

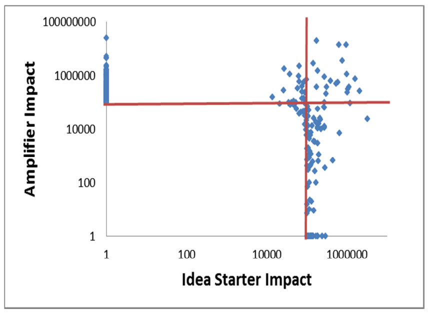
We do not present similar experiments for transmitter as getting high transmitter value is rarer than other two metrics. However, a comparison of top 100 idea-starters, amplifiers, and transmitters is carried out, which reveals that 21 users were listed in both top-idea starter and top-amplifier, 7 were listed in both amplifiers and transmitters and 1 was in top-idea starters and transmitters. The popular celebrity with Twitter id ‘ayu_19980408’ was there in all three top-lists which indicates her potential influence in the network.



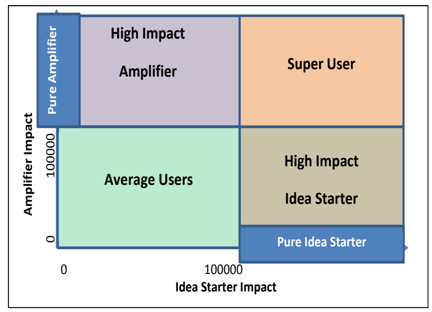
(a) Pre-event Time window



(b) During-event window



(c) Post-event window



(d) Division of users in four quadrants

Figure 5: Idea-starter vs. Amplifier impact in pre-, during- and post-event time window

5.3 Transmitter's Topology

According to raw *IDI* values, transmitters do not have an impact as high as the two other roles. However, many transmitters had a strong impact on the diffusion of information with 15 users having a overall transmitter score above 100,000 and 538 users with a score above 10,000.

We can identify two categories of transmitters. The first category corresponds to users who are frequent transmitters to small communities (Table 3). They were transmitters for several dozens of tweets, but the community to which they transmitted is not very large, resulting in relatively low scores. On the contrary, some of the users (Table 2) with the highest transmitter scores are transmitters in less than 5 tweets. But they are transmitting an information from small communities to the largest ones. Therefore, a tweet which could have reached only a fraction of all users without transmission, reach most of the network after transmission. For example, one tweet was launched by someone from the night life community, asking for help for refugees in a town affected by the tsunami. After being retweeted by some people from the same community, it was transmitted to the other communities and became widely retweeted. The impact of the transmission is therefore very high in this case.

Name	**Transmission IDI**	**#transmission**	**Informed community size**
polyjam	177332	2	150235
lythiriana	156351	2	117193
rt_report	150909	2	150235

Table 2: Top Transmitters by IDI

Name	**Transmission IDI**	**#transmission**	**Informed community size**
satos_cafe_bar	57945	102	11462
ayu_19980408	50153	102	12177
bigkottakromac	106834	58	20824

Table 3: Top Transmitters by number of transmission

5.4 IDI of User Role and Number of Followers

We investigated the Pearson correlation among overall idea-starter, amplifier, and transmitter impact of each user with that of their number of followers. We find that number of followers is not correlated with idea-starter impact (correlation =0.2827), amplifier impact (correlation =0.4352), and transmitter impact (correlation =0.0273).

Figure 6 shows the contrast of 100 top-followed users with idea starter, amplifier, and transmitter impact. One can note that amongst the top-followed users, the roles are very different and they have very different *IDI* impact. Hence, metrics like number of followers cannot determine the user-roles or even the global IDI of the user in the network.

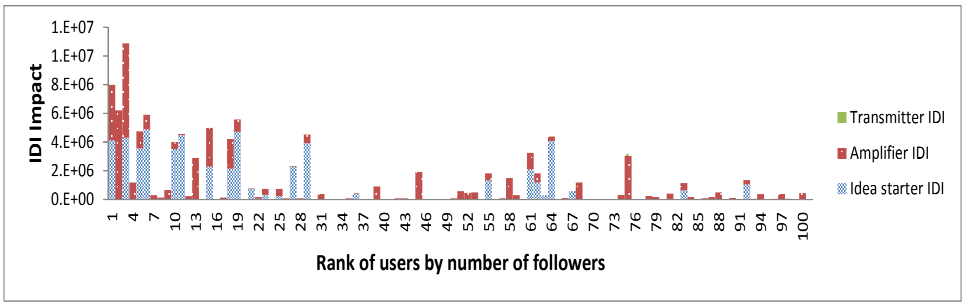


Figure 6: Comparison of number of followers with IDI impact of three roles

6 Conclusion

Information diffusion in Twitter happens mainly by the act of retweeting. In most recent works only tweet content is analyzed to build the diffusion network. However, common unstandardized retweet practices introduce bias in the analysis, which can be mitigated by embedding follower network information along with the tweet content. In this study we systematically constructed and analyzed retweet chains to define user's role based on the concept of *Information Diffusion Impact (IDI)*. In the context of information diffusion we identify three important user roles, namely "idea starter", "amplifier", and "transmitter". While idea-starters are good at launching interesting ideas, amplifiers are needed to diffuse the information broadly in the network, and transmitters are important to reach different community. Analyzing Twitter dataset of the Great East Japan earthquake data, we observe the change of these user roles in three different time-window, before, during, and after the disaster. Our findings show that popular idea starters and amplifiers tend to be popular during the disaster, but a large number of new users also emerge as good idea-starters and amplifiers. We also identified super-users with high potential to both amplify and launch new ideas.

Acknowledgement

We thank Genta Kaneyama (Cookpad Inc.) for assistance in collecting data from Twitter.

References

[1] F. Abel, Q. Gao, G.-J. Houben, and K. Tao. Analyzing user modeling on twitter for personalized news recommendations. In *Proceedings of the 19th international conference on User modeling, adaption, and personalization*, UMAP'11, pages 1–12. Springer-Verlag, 2011.

[2] D. Boyd, S. Golder, and G. Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *Proceedings of the 2010 43rd Hawaii International Conference on System Sciences*, HICSS '10, pages 1–10. IEEE Computer Society, 2010.

[3] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia. Who is tweeting on twitter: human, bot, or cyborg? In *Proceedings of the 26th Annual Computer Security Applications Conference*, ACSAC '10, pages 21–30. ACM, 2010.

[4] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, June 2002.

[5] 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*, WWW '04, pages 491–501. ACM, 2004.

[6] A. L. Hughes and L. Palen. Twitter Adoption and Use in Mass Convergence and Emergency Events. In *ISCRAM Conference*, May 2009.

[7] A. Lancichinetti, F. Radicchi, J. J. Ramasco, and S. Fortunato. Finding Statistically Significant Communities in Networks. *PLoS ONE*, 6(5), 2011.

[8] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*, 1(1):5, 2007.

[9] M. Mendoza, B. Poblete, and C. Castillo. Twitter under crisis: can we trust what we rt? In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 71–79. ACM, 2010.

[10] N. Pervin, F. Fang, A. Datta, K. Dutta, and D. E. VanderMeer. Fast, scalable, and context-sensitive detection of trending topics in microblog post streams. *ACM Trans. Management Inf. Syst.*, 3(4):19, 2013.

[11] B. Suh, L. Hong, P. Pirolli, and E. H. Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Proceedings of the 2010 IEEE Second International Conference on Social Computing*, SOCIALCOM '10, pages 177–184. IEEE Computer Society, 2010.

[12] R. Tinati, L. Carr, W. Hall, and J. Bentwood. Identifying communicator roles in twitter. In *Mining Social Networks Dynamics,(MSND workshop)*, WWW '12 Companion, pages 1161–1168, 2012.

[13] F. Toriumi, T. Sakaki, K. Shinoda, K. Kazama, S. Kurihara, and I. Noda. Information sharing on twitter during the 2011 catastrophic earthquake. In *2nd International Workshop on Social Web for Disaster Management (swdm2013)*, WWW '13 Companion, pages 1025–1028, 2013.

[14] Wikipedia. List of foreshocks and aftershocks of the 2011 thoku earthquake. http://en.wikipedia.org/wiki/List_of_foreshocks_and_aftershocks_of_the_2011_T%C5%8Dhoku_earthquake.