Using network properties to analyze users' role in Twitter in time of crisis

Rémy Cazabet[*1] Nargis Pervin[*2] Fujio Toriumi[*3] Hideaki Takeda[*1]

[*1] National Institute of Informatics [*2] National University of Singapore

[*3] The University of Tokyo

Twitter and online social networks in general are known to be useful alternative sources of information in time of crisis. Using a large dataset of Tweets published in Japan during the period of the earthquake and tsunami of March 2011, we used network analysis to identify the roles of users in the diffusion of information. In a first part, by analyzing retweet chains, we identified 3 categories of users: "idea starters", "amplifiers" and "transmitters", all being necessary to the efficient diffusion of information in the network. In a second part, we studied in details how the degree of users in the network affects their capacity to diffuse efficiently information.

1. Introduction

Online Social Networks have attracted a lot of attention recently. One interesting aspect of these platforms is that they can be used by users to share and diffuse information, in particular in situations where traditional sources of information are not reliable of efficient. Theses cases encompass political crisis and natural disasters, such as earthquakes and hurricane. In this paper, we will investigate the effects of a specific natural disaster, namely the earthquake and Tsunami in Japan of 2011, and its effect on users behaviors on Twitter. After briefly introducing our dataset, we will present the results of a first work, looking in details at the different roles the users can have in the network. We will then present the results of a second work, this time on the impact of the degree of nodes on the capacity to publish highly retweeted tweets. In both works, we will first present general results, without taking the crisis period into account, and then discuss the change in behavior due to it.

Dataset description

1.1 Tweet Data

We used a Twitter dataset covering the great Tohoku earthquake in Japan in March 2011 and described thoroughly in [Toriumi 2013]. The dataset covers a period of 20 days (from March 5, 2011 to March 24, 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 evaluated that 80% to 90% of all published tweets by these users were present in the dataset.

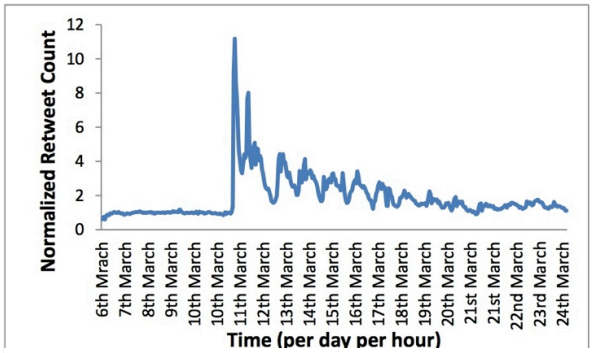

Fig. 1 Normalized retweet count

Fig. 1 shows the normalized retweet count. The first two major peaks correspond to the two main earthquakes on March 11 and 12. After the disaster, retweet count progressively returns to its normal average values.

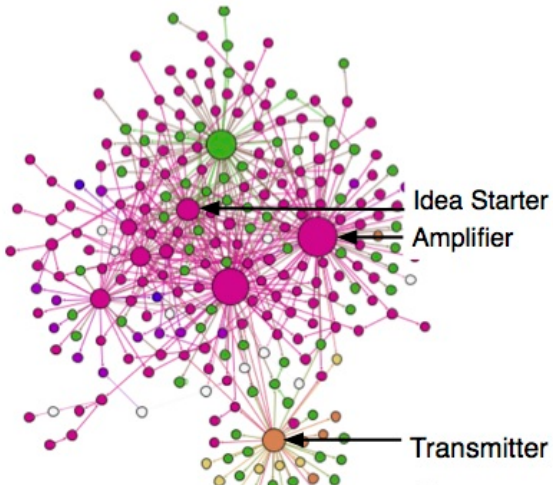

Fig. 2 Example of retweet chain with roles

1.2 Follower network data

To complete the original dataset, we collected the follower relationships between the most active users, namely those mentioned more than 20 times in our dataset. This follower network consists of 300,104 users and 73,446,260 relationships.

2. Studying the complementary roles of users

In a first approach, we studied what kind of different roles the users were able to play in the diffusion of information. Building upon the paper [Tinati 2012], we defined 3 categories, illustrated in Fig. :

- **Idea starter (ID)**. These users are the ones who create the original information, making it available to their direct followers. These original tweets can, consequently, be diffused much further in the network, through retweets.

Contact: Hideaki Takeda, National Institute of Informatics, 2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, takeda@nii.ac.jp

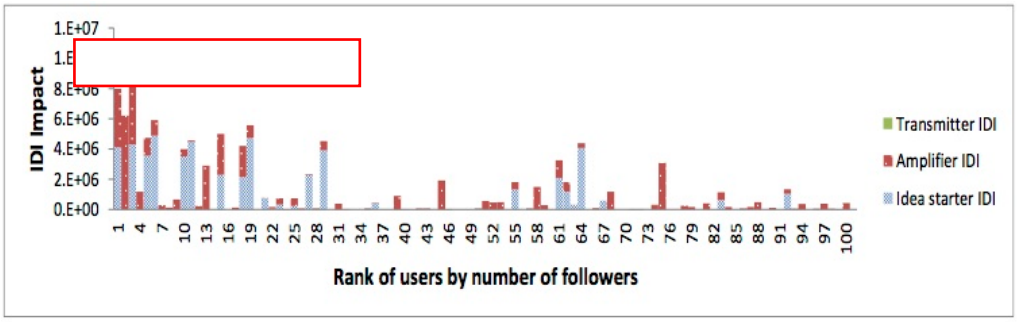

Fig. 3 Roles for the top 100 users

- **Amplifiers (AMP)**. These users make many new people aware of tweets they have not published themselves. Typically, they are users with many followers.
- **Transmitters (TR)**. Users in social networks are known to form communities, i.e. group of users relatively weakly connected to each other. These users have the particularity to relay information from one community to another.

We didn't want to assume that roles were or not exclusive, so, we decided to define a metric for each role, which can be computed for each user. We can therefore attribute to each user a score of ID, a score of AMP and a TR score. We proposed metrics which have the advantage of being comparable, that is, their value represents how the user impact the diffusion of information in the networks through his behavior in each role. More precisely, we define a unit that we call IDI, for Information Diffusion Impact. This metric corresponds to the number of people affected by the actions of the user. An IDI of 1 correspond to the fact that one user can see one piece of information he never saw before. An IDI of 10 can means that 1 user saw 10 new information items, or 10 users saw one piece of information.

To describe formally the computation of the metrics, we must

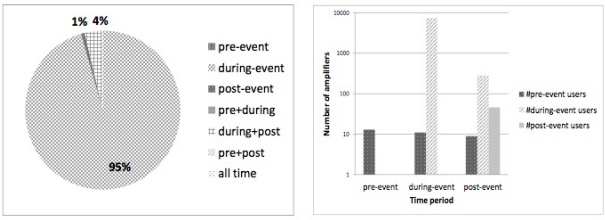

Fig. 4 Evolution of top amplifiers

first describe the concept of cascade of information. For each tweet *t*, we compute its cascade of information *c*, which is the ordered list of all users who had the possibility to see this tweet. It is constructed according to the following algorithm, starting from the ordered list of all the retweeters:

c<-[ ]
For u in RTL(t)
  $c <- c \cup [\, fol(u) \setminus c \,]$
End for

With
-RTL(*t*) the ordered list of all retweeters of tweet *t*
-Fol(*u*) the list of all followers of user *u*

We additionally define the following functions :
-Order(*u,c*) which returns the rank of *u* in the cascade *c*
- $\inf(u,c) = \{u1 | u1 \epsilon fol(u) \wedge order(u1,c) > order(u,c)\}$ which corresponds to the list of users who could acces the information for the first time because of *u* retweet.

We can now define formally our metrics:
$ID(u) = \sum_{c|order(u,c)=1} |c|$
$AMP(u) = \sum_c |\inf(u,c)|$
$TR(u) = \sum_c \sum_{u1 \in \inf(u,c) \wedge newCom(u1)} |TR(u1)| + 1$
Where newCom(u1) represent the fact that *u1* belongs to a community different than the ones of the nodes before. To identify communities in the network, we used the OLSOM community detection algorithm [Lancichinetti 2011], considered as one of the most efficient, and fast enough to handle our large dataset.
For each user in our dataset, we can compute its value of ID, AMP and TR, which can be compared as they are in the same unit, IDI, representing the impact on the network.

2.1 Experimental results

Complete Period

In a first step, we studied the roles on the whole dataset. We observed that users with similar number of followers could have very different role scores (Fig. 3). Some users, even though they have many followers, do not play a role of transmitters. Notable users in these categories are news media and bots (accounts whose tweets are published in an unsupervised manner by computer programs. For example, a popular bot publish a tweet with some details for each earthquake occuring in Japan). On the contrary, some users which are very good amplifiers do not use this potential to transmit their own information. Finally, some users have high scores in both roles. Transmitters scores are comparatively smaller for the top users of the network, but, for some particular users with fewer followers, transmitter scores can also be high.

Effect of the crisis

In a second step, we studied the impact of the earthquake on the roles adopted by users. We partitioned our dataset in 3 sections of 1 weak each, corresponding to the period before the earthquake, directly following the earthquake, and after the crisis. For amplifiers (Fig. 4), we can see that users who were already amplifiers before the crisis mostly kept this role. A very large number of new users became amplifiers during the crisis, and some of them kept this new role even after the core of the crisis.

For idea starters, we observed different behaviors. Most of those who were influential before the crisis did not keep this role

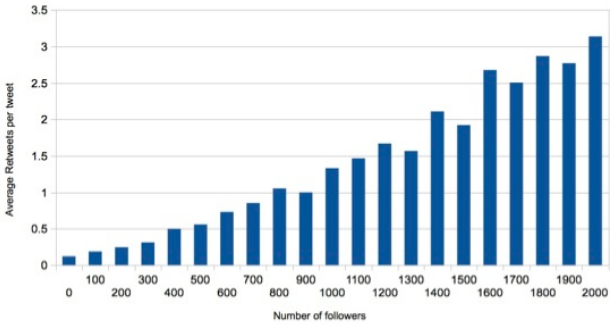

Fig. 5 Correlation degree/average reweets

after the earthquake. As for amplifiers, a large amount of new idea starters were revealed by the crisis. It's also interesting to stress that the ratio of the role between amplifiers and idea starter was changed during the crisis. In the first period, we could find more influential idea starters than influential amplifiers (IDI above 100.000). But as soon as the earthquake happened, we could observe more than 10 times more influential amplifiers than idea starters. It seems that during a crisis, the role of people transmitting information published by other users become even more crucial.

3. Diffusion capacity and nodes degree

In this second work, we were interested in the relation between the degree of a node in the follower network (the number of followers this user has) and his probability for being retweeted. More precisely, a strong correlation had been already observed [Kwak 2010] between the degree of a user and the average number of time his tweets are retweeted. We have observed the same phenomenon on our dataset (Fig. 5). However, it is also known that the distribution of the retweet chain lengths follows a power law. As the average value of power law distribution is not representative of this distribution, we proposed to directly study the parameters of the distribution, in order to better understand it. The interest of such a work is to better understand why the average number of tweets increases for popular users; is it just because, having a lot of followers, it's more likely for them to be retweeted a few times, or is it because they are also more likely to diffuse seminal tweets, which can reach most people in the network ?

3.1 Computing the parameters

A power law distribution function can be expressed by the following formula:

$$P(x) = Cx^{-\alpha}$$

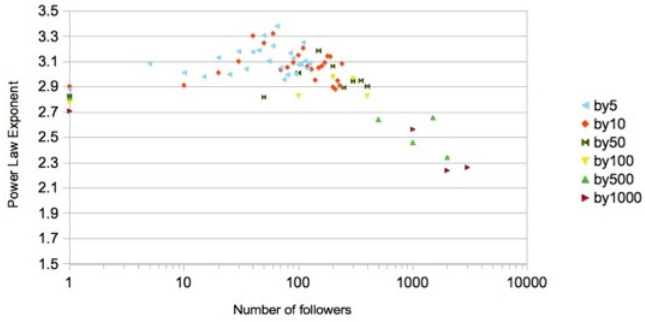

Fig. 6 Relation between Degree and Exponent

where C is a normalization constant. The distribution is also bound by an *xmin* value. The two parameters *xmin* and α define the distribution. α defines the slope of the distribution (said otherwise, the relative frequency of rare events), while *xmin* represents the minimal value for which the power law is respected. Therefore, we searched how these parameters were changing according to the degrees of nodes. To do so, we created groups of users of comparable degrees and computed the parameters for each of these groups. Due to the power law distribution of degrees of nodes, we used several granularity levels for the similarity of nodes, and kept only results for groups of users with enough users to be reliable. To compute the parameters of the distributions, we followed the recommendation

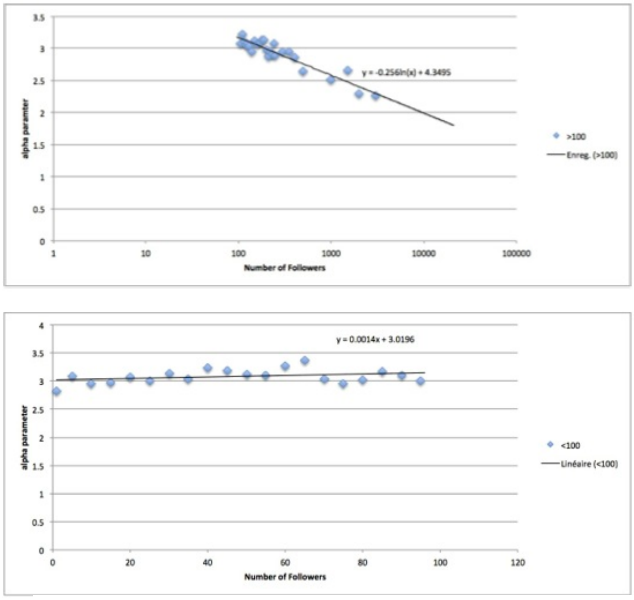

Fig. 7 Details of exponent/degree relations

discover power law in real data. In our paper, we do not try to claim that the power law is unarguably the unique possible fit for our data, but we checked that the power law was a convincing fit, reliable enough to represent the tendencies of the relation which interest us.

3.2 Results

We found that *xmin* was varying mostly for nodes with few followers. On the other hand, we could find a clear relation between α and the degree of nodes.

This relation is presented in fig.6. We can see that from 0 until approximately 100 followers, α do not vary much and slightly increases (Fig. 7). For users with more than 100 followers on the contrary, α decreases logarithmically with the degree. Said differently, for these users, the more followers they have, the more it is likely to have « rare events », tweets retweeted widely in the network.

This observation nuance some recent results and a common belief that « everyone can be popular » on web 2.0 platforms. In order to check this observation, we computed in the dataset the relative representation of nodes of a given degree in chains of a

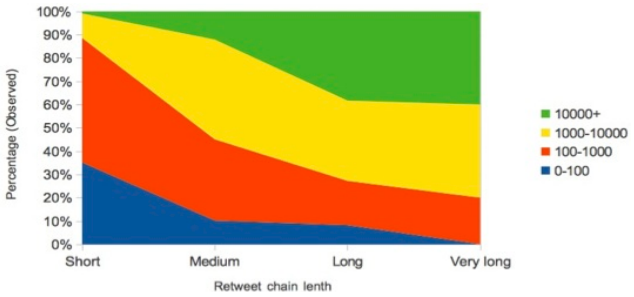

Fig. 8 Relation degree/long chains

given length. In order to have significant quantities of data, we separated retweet chains in 4 categories:

- Short (1-10 retweets)
- Medium (10-50 retweets)
- Long and (50-500 retweets)
- Very long (500+ retweets)

We similarly created 4 groups of users:

- Few followers (0-100 followers)
- Medium number of followers (100-1000)
- Many followers (1000 – 10000)
- Super hubs (10000+)

In figure 8, we show the correlation between these two categories. We can observe that, while users with less than 1000 followers are responsible for most of the tweets retweeted only a few times, they represent only a small fraction of very long chains. On se contrary, the super hubs, while representing only a tiny fraction of the users of the network (less than 1%), account for more than 30% of the long and very long retweet chains. Even though it is possible for not famous users to generate widely diffused tweets, this is an exception, and the rule is that very popular users are far more likely to publish seminal tweets.

3.3 Effects of the crisis

Our last experiment was to look at the effect of the crisis on the distribution of retweet chain lengths. To do so, we computed the global α parameter on each day, instead of on the whole network as before. It has already been shown that in time of crisis, more tweets were published on tweeter. But many factors can be responsible of this observation. For example, it might be due to the implication of users who de not post tweets usually, or to the publication of more original tweets, or more retweets, among many hypothesis. Using the same technique as before, we computed α for each day. What we observe, as seen on (fig. 9), is a sudden fall in this parameter, which slowly rise back to its original value. We can interpret this as a much higher probability for tweets to become seminal. A tweet published by users of a given degree is fare more likely to be diffused to a large portion of the network during and right after the crisis than before.

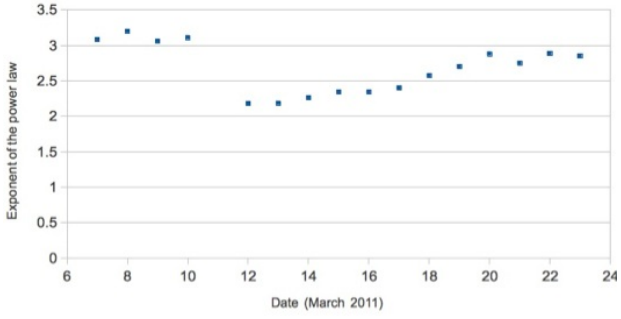

Fig. 9 Effect of the crisis on the exponent

4. Conclusion

In this paper, we investigated the roles of users in the diffusion of information in Twitter. In a first part, we identified 3 roles, and proposed metrics to quantify the involving of users in each of these roles. One key point of these metrics is that they are comparable between themselves, as they represent the global impact on the network through the IDI unit. We found that some users could have very strong values in one metric but very low in another, which reflects a real specialization. In particular, users such as bots and news media tend to be very influential idea starters, but often do not use this influence to be amplifiers. Some more results can be found in [Pervin 2013].

The findings of the second part of this paper are that the capacity of users with low degrees to publish tweets widely diffused in the network might be overestimated. In fact, we show that the more followers a user has, the more likely he is to publish this kind of seminal tweets. It is only for users with less than 100 followers that this probability is stable with the degree. Some complementary results can be found in [Cazabet 2013].

Finally, in both cases, we studied the impact of the crisis. We found strong impact on the behaviors of the users. The two results are complementary, since the first one shows that the amplifier role is the most affected by the crisis, with many people becoming strong amplifiers, while the second one shows that the probability of being widely retweeted increases strongly because of this event.

In future works, it would be interesting to search for other roles for users. It would also be enlightening to compare the results obtained on tweeter with the ones we could have on different Online Social Networks.

References

[Cazabet 2013] R. Cazabet, N. Pervin, F. Toriumi, H. Takeda, "Information Diffusion on Twitter: everyone has its chance, but all chances are not equal." *Signal-Image Technology & Internet-Based Systems (SITIS)*.

[Clauset 2009] A. Clauset, C. R. Shalizi, and M. E. Newman, "Power-law distributions in empirical data," *SIAM review*, vol. 51, no. 4, pp. 661–703.

[Kwak 2010] 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*.

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

[Pervin 2013]N. Pervin, R. Cazabet, F. Toriumi, H. Takeda, "User roles in the time of crisis : A social media analysis" *Poster in WITS 2013*

[Tinati 2012] R. Tinati, L. Carr, W. Hall, and J. Bentwood. Identifying communicator roles in twitter. In *Mining Social Networks Dynamics,(MSND workshop)*.

[Toriumi 2013] F. Toriumi, T. Sakaki, K. Shinoda, K. Kazama, S. Kurihara, and I. Noda, "Information sharing on twitter during the 2011 catastrophic earth- quake," in *Proceedings of the 22nd international conference on World Wide Web companion*.