Proceedings of the 4th International Conference on Computing and Data Science (CONF-CDS 2022)
DOI: 10.54254/2755-2721/2/20220621

Why do rumors spread so fast on Twitter?

Weijie Xie

South China Normal University, 510898 China

382455438@qq.com

Abstract. This research aims to find out why rumors spread so fast on Twitter. So based on the concept of social networks and Katz Centrality, I decide to use "Crawlers" to collect data from Twitter and generate "Real News Spreading Network", "Rumors Spreading Network" as well as graphs that illustrate their variation with post time delay. And I find that although the original size of rumors may not be as large as the original size of real news, more and more people spread rumors as time goes by, while the amount of people who read real news only grows a little. What's more, it is users who are not very important(judging from their Katz Centrality value's rank in certain news spreading networks) in the social network that play key roles in spreading the news(both rumors and real news). So I deduce that it might be ordinary users (they are not Internet celebrities, and don't have many followers) who should be aware of being informed of many rumors because they might be the key factors to spreading or stopping rumors.

Keywords: Social network, Rumors, True News, Twitter, Katz Centrality.

1. Introduction

As is known to all, nowadays technology is developing at a fierce speed. As a result, the internet is developing rapidly too, and a variety of social apps are being used by people frequently: facebook, Twitter, WeChat…Therefore, some people prefer to gain information and read news on social apps instead of reading the newspaper or watching TV. But unlike the news on newspapers or official television, many news on social apps are fake." A lie can travel halfway around the world while the truth is still putting on its shoes", it is a famous saying created by Mark Twain, and this saying describes the exact situation of rumor spreading on social apps nowadays. The quickness of rumor spreading cause huge harm to society, because many works are based on information in the present day, and rumors could also damage one's fame as well. So to find out an efficient way to curb rumor spreading, we need to know why rumors travel so fast.

Many experts and talented people have already researched this question, and most of their results are accurate and sophisticated. However, I am trying to figure out a method that could allow readers to understand the reason easier. So I use a different perspective to study why rumors spread so fast, based on a concept of Katz centrality and some quantitative measures of networks.

Most methods are so sophisticated and depend on many factors. In some scenarios, it is hard to collect enough data for analysis. Our proposed method focuses on three common variables which are the parent

DOI: 10.54254/2755-2721/2/20220621

nodes, child nodes and post time of rumors spreading. As a result, our method can make an efficient judge and it is applicable to different scenarios.

Few methods conduct both qualitative analysis and quantitative analysis on rumor spreading. Therefore, we propose a comprehensive method that combines qualitative analysis and quantitative analysis.

In terms of qualitative analysis, we visualize three different diagrams for rumors and real news respectively, including the Spreading Network, the Histogram of Katz Centrality and Frequency for Top-20 Influential Nodes, and the Spreading Network Varying with Post Time Delay.

In terms of quantitative analysis, we carefully select three kinds of quantitative measures which are Density, Diameter, and Clustering.

Compared to the existing methods of rumor spreading, our proposed approach could make the conclusion more intuitive and comprehensive.

2. Twitter and Rumors spreading

Twitter is a kind of social app that is extremely popular and it’s based on the concept of the social network, which is formed by people, and they are linked to each other by their various relationships. Twitter is a manifestation of the social network, it consists of many users, and there are relationships between users——some users may follow others (like subscription on Twitter), so at the same time, some users are followed. Some news that is spread between users is not confirmed or verified, they are generally defined as rumors.

The public fears will be exaggerated while rumors travel through various channels on a social network. There is a noted model called SIS, introduced by Daley and Kendall, which could be used to describe this phenomenon. It’s a model that drew inspiration from the SIR model, which is used in epidemiology. Evenly blended populations are divided into three kinds in the SIS model. The first “S” refers to spreaders (In rumors spreading, it represents people who actively spread rumors), “I” refers to ignorants (i.e. people who are unaware of rumors), and the second “S” refers to those people who have received rumors but stop spreading.

However, it oversimplifies network topology. Network entails so many factors (e.g., social links and, webpage links), which make it way more complicated than homogeneously uniform networks. Rumors don’t equal the epidemic, people could decide what to do with them (believing or spreading). The description and the criticism of the SIS model are mentioned in the paper "How Rumors Spread and Stopover Social Media: a Multi-Layered Communication Model and Empirical Analysis"[1], which I think accurately pointed out the incompleteness of representing the law of social networks.

Though as diverse as networks are, they share some common traits. Any two people are connected through a principle called “six degrees of separation”. Also, the diameter between two nodes in the network which is constructed by web pages and links is only 19 [2]. These traits are also mentioned in the paper “Why rumors spread so quickly in social networks”, which is very inspiring.

In addition, social networks could be described as a directed graph. A node’s Katz centrality represents their centrality in a network in graph theory. This concept was first introduced in 1953 by Leo Katz, which is used to measure an actor’s(or node’s) influence within a social network. Katz centrality is more comprehensive than typical centrality in measuring importance, for the latter only take the shortest path between a pair of actors into account but the former consider the whole number of walks between a pair of actors.

To find out the key factors of spreading rumors, top-K influential nodes are needed to be found. About this subject, a considerable amount of research has been conducted: Centrality theory, diffusion models, heat diffusion theory, evidence theory, etc[3]. The techniques I mentioned above are frequently used for figuring out top-K influential nodes in a network.

Li et al. (year) introduced a method based on evidence theory to identify influential nodes in a network of networks (NON). It divides the complex network into sub-networks such as groups of centrality networks. Distance matrix (D) is computed for each of these networks, and D represents the centrality among the nodes. This matrix D is further used to calculate centrality networks which help to figure out

DOI: 10.54254/2755-2721/2/20220621

basic probability assignment (BPA). Nodes with high centrality values are considered to be influential in NON.

We assume that nodes choose their communication partners randomly from their neighbors but exclude the node they contacted just before. In the directed graph, we consider the spread of a single piece of information initially manifest at a single node. It’ s also assumed that the process of rumor spreading is synchronized, which means that in each time step, each node communicates with its neighbor to exchange information at discrete points of time. This assumption could get rid of a lot of complexity because, in the real world, rumors don’ t travel at an identical speed. But making this simplification doesn’ t greatly affect the accuracy, for the networks are large enough.

Twitter users communicate in different ways, text, post on their pages, share others’ posts, etc. But in the graph, we only assumed the link between them is their sharing between two nodes or the following relationship because more complicated mechanisms will be needed if forms of communication are all modeled.

3. Background Knowledge

3.1. Katz centrality

The Katz centrality of a node vi is computed as:

$$C_{Katz}(v_i) = \alpha \sum_{j=1}^{n} A_{j,i} C_{Katz}(v_j) + \beta \tag{1}$$

Of which, α is a constant called the damping factor, which is generally considered to be less than the maximum eigenvalue, λ namely α< 1/ λ， β is an offset constant, also known as an exogenous vector, used to avoid Zero Center values. Centrality tends to diverge with α ⩾ λ.

3.2. Density

Density could be used to describe how densely each node in a network is connected. In the online social network, it is also used to measure the density of social relationships and their varying trend. In addition, large-sized networks have a lower density than those networks whose size is relatively small, and its value can be computed by this formula:

$$d(G) = \frac{2L}{N(N-1)} \tag{2}$$

where n is the number of nodes and L is the number of edges in G (a network) [4]. The formula of density is put forward in the paper "Graphs over time: densification laws, shrinking diameters and possible explanations.".

3.3. Diameter

The diameter is the maximum eccentricity(The eccentricity of a node v is the maximum distance from v to all other nodes in G.)[5].

3.4. Clustering

The local clustering coefficient Ci of node i is the ratio of the actual number of edges Ei between its adjacent ki nodes to the total number of possible edges, i.e:

$$C_i = \frac{2E_i}{[k_i(k_i-1)]} \tag{3}$$

DOI: 10.54254/2755-2721/2/20220621

Clustering coefficient C is defined as the arithmetic mean of all nodes' clustering (Ci) coefficient, its computing formula is:

$$\boldsymbol{C} = \sum_{i=1}^{n} C\, i \tag{4}$$

Where n is the number of total nodes and i is the order of a node in a certain network[6]. The formula for Clustering coefficient is put forward in the paper "Comparison of Different Generalizations of Clustering Coefficient and Local Efficiency for Weighted Undirected Graphs".
For a network, the clustering coefficient is used to describe the closeness between nodes and their neighbors[7]. "Small world network characteristics of biological systems" put forward the use of clustering coefficient in 2007.

4. Methodology

The code was written in Python using VScode. We use python data structures to write the algorithms. Because Python is relatively easier for students to learn, besides it contains numerous functions. Its code structures are more comprehensive than other programming languages, and it could fulfill a function with a few lines of code.

In terms of data used for generating directed graphs, a dataset will be needed. The aim is to collect a certain number of Twitter users who read and share rumors and real news, to build directed networks which we named "Rumors Network" and "Real News Network", used to describe the communication between users (Their interaction of spreading rumors or truth). In addition, I also intend to generate graphs that show the variation of the spreading network with post-time delay.

The program I am about to use for collecting data is "Crawlers". It's a script/program which collects information following certain rules. Compare with other means of searching data, it has the following advantages:1.it saves the researcher time: the searching process is automatically done by the program;2.it collects a larger amount of data than humans: Because it's done by computers, and computers have more powerful relative functions than humans;3.it is money-saving: I don't have to recruit more staff or a complicated system to collect data.

But, likely, the Twitter dataset is not available to me, so I plan to use crawlers to collect data from Twitter. Crawlers firstly find the URL of the data, then send a request to this address. After that, it collects the data that is sent by the URL server, next it uses python to get the aiming data from the source code. Next, I remove data that are irrelevant and save that are useful. Eventually, I get data and could use them to generate the network model.

Then I use the data we collect to calculate Katz centrality. As I mentioned above, "Rumors Network" and "Real News Network" are directed graphics in this experiment. Users are vertices and the edges are their communication (like sharing real news or spreading rumors). I defined the in-degree as how many rumors (real news)the users read, and the out-degree represents the number of rumors(real news)they spread to others.

Also, I use those data to generate graphs, which could help us to see how much those networks vary with a post-time delay so that we could see the growing trend of real news or rumors.

5. Result

To figure out whether certain news is rumor or real news, we use an online rumor debunking service provided by www.snopes.com[8]. From March to December 2015, we collected a total of 778 reported incidents, 64% of which were rumors. For each of them, we extract keywords from the last part of the Snopes URL, for example., http://www.snopes.com/pentagon-spends-powerball admission ticket. We refine keywords by adding, deleting or replacing words, and repeat these keywords until the composite query can get fairly accurate twitter search results. We use scripts to download the "Live" search results

DOI: 10.54254/2755-2721/2/20220621

from Twitter. In addition, we need to balance the two classes, so we further added some non-rumor events from two public datasets [9] [10].

From the short review above, key findings emerge. We use data of Twitter users who share rumors or real news and those who retweet after reading them to generate 20 each spreading network of rumors and real news, and 20 each histogram to show the relationship between frequency and Katz centrality value of Top-20 influential nodes. Next, we decided to choose some that are representative to post in this paper (Figure. 1—Figure. 4). It's clear that the nodes of rumor networks are fewer than the Real News network. And although for each histogram, their ranges of Kat centrality value are different, the top-20 influential nodes of them have the highest frequency in the far left of the histogram.

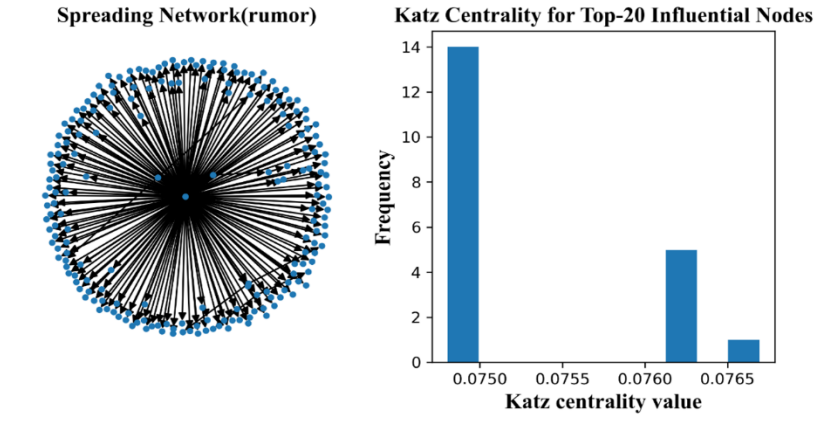

Figure. 1. Spreading Network of certain rumors and the histogram of Katz Centrality and Frequency for Top-20 Influential Nodes.

The graph on the left of Fig.1 describes the trend and shape of a certain rumor spreading on the social network. The histogram shows the relationship between frequency and Katz centrality value of Top-20 Influential Nodes in this spreading network.

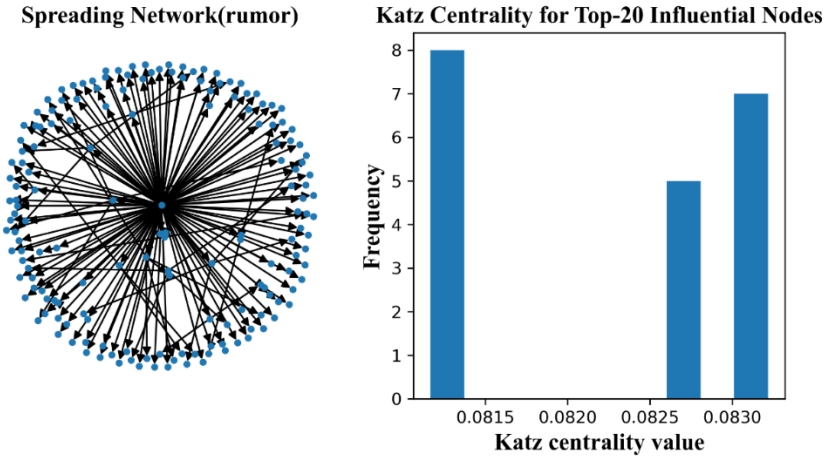

Figure. 2. Spreading Network of certain rumors and the histogram of Katz Centrality and Frequency for Top-20 Influential Nodes.

The graph on the left of Fig.2 describes the trend and shape of a certain rumor spreading on the social network. The histogram shows the relationship between frequency and Katz centrality value of Top-20 Influential Nodes in this spreading network.

DOI: 10.54254/2755-2721/2/20220621

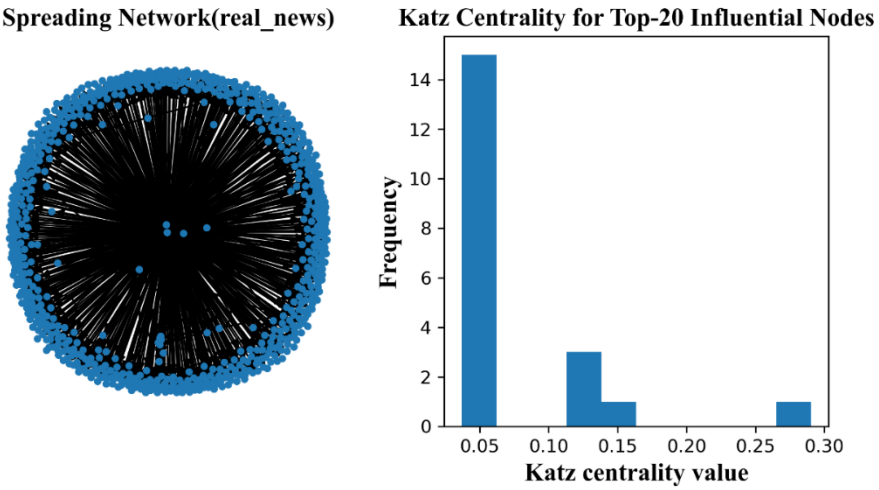

Figure. 3. Spreading Network of certain real news and the histogram of Katz Centrality and Frequency for Top-20 Influential Nodes.

The graph on the left of Fig.3 describes the trend and shape of certain real news spreading on the social network. The histogram shows the relationship between frequency and Katz centrality value of Top-20 Influential Nodes in this spreading network.

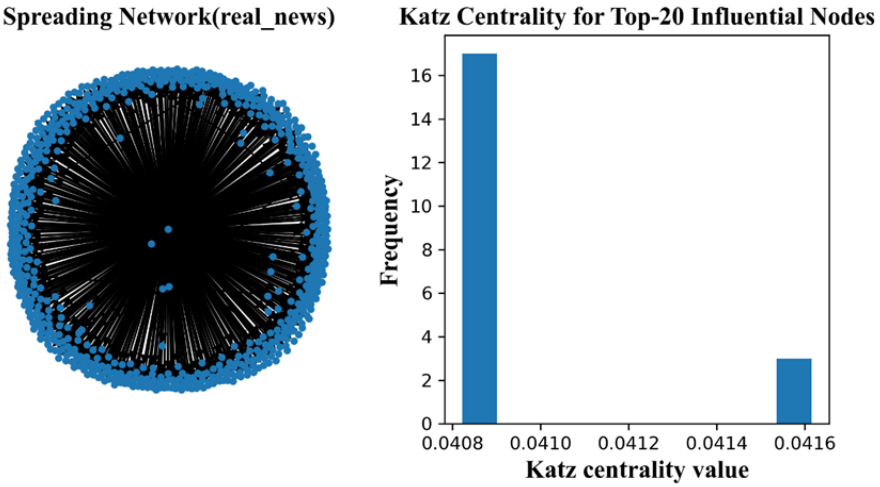

Figure. 4. Spreading Network of certain real news and the histogram of Katz Centrality and Frequency for Top-20 Influential Nodes.

The graph on the left of Fig.4 describes the trend and shape of certain real news spreading on the social network. The histogram shows the relationship between frequency and Katz centrality value of Top-20 Influential Nodes in this spreading network.

Fig.5 and Fig.6 are about two networks' changes with time-varying. It could be found that nodes in this rumor network are relatively few compared with nodes in the true news network in the beginning. But it is noticeable that Real News Spreading Network doesn't vary a great deal, while the Rumor Spreading becomes much denser as time goes by. In the final stage, nodes in this rumor network are more than nodes in the chosen real news network.

DOI: 10.54254/2755-2721/2/20220621

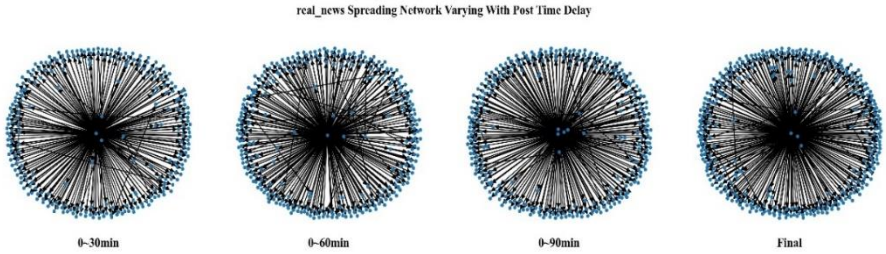

Figure. 5. Real news Spreading Network varying with post time delay.

These four graphs in Fig.5 show the changing trend of this real news network within a given time.

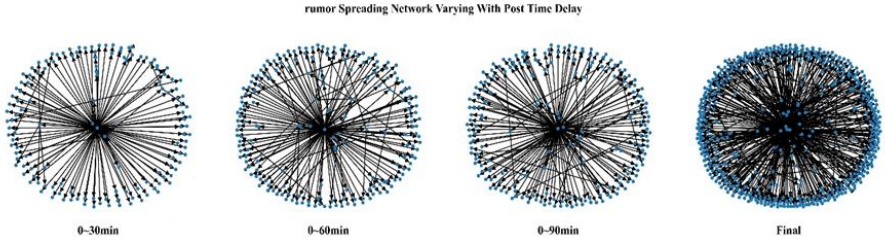

Figure. 6. Rumor Spreading Network varies with post time delay.

Table 1. Some quantitative measures of rumor(real news) network.

Metrics	Rumor spreading network	Real News spreading network
average density	0.009699	0.004505
average diameter	7.342995	6.852217
average clustering	0.004888	0.005251

Table 1 shows some quantitative measures of these two kinds of networks. It lists the average density, average diameter, and average clustering of all the rumor-spreading networks and real news-spreading networks we collected respectively. We only post the average data of them because the data we collected is much. It is clear that the value of average density and average diameter of rumor networks are higher than the ones of the real news network, while real news networks have a higher average clustering value.

6. Discussion

From Fig.1—Fig.4 we could notice that in common cases the nodes in Rumors Spreading Network are much fewer than the ones in Real News Spreading Network. In other words, the size of rumor spreading maybe not be so large in the beginning. According to our commonsense in daily surfing on the Internet, it is not difficult to conclude that most rumors are spread by unofficial users(e.g.users that may only have a few followers, or those users that are not so influential in the whole social network).In addition, users are of less importance and play key roles in spreading the news(both real news and rumors).

However, as time goes by, the size of rumors usually gets larger while the size of real news only increases a little. In other words, even if only a few people come into contact with a certain rumor, a few days later the rumor may be much more well-known. From my opinion, that's probably because the content of rumors is usually more exaggerating(maybe the intent of the original spreader is to attract more people to view his post so that he could earn more popularity). The information flow here shows that false tweets are first published by low impact users, and then some popular users join in to promote the dissemination. However, tweets that tell the truth are first published by popular users and directly disseminated by many ordinary users.

DOI: 10.54254/2755-2721/2/20220621

As for the result of quantitative measures of rumor networks and real news networks, we could also make some conjecture. According to the concept of density, we find that size of most rumor networks is smaller than that of real news networks. As for diameter, rumors likely spread wider than real news, for the average maximum distance of nodes in rumor networks is longer. In terms of the average clustering coefficient, we infer that users have a closer relationship in real news networks than in rumors network, which means that rumors are more likely to be spread by users, for not every single spreader of news are friend, conversely, they don't have any relationship with each other.

But the results have some limitations. Because we don't have any access to the database of Twitter, so we could not obtain detailed information about the spreaders(e.g.how many followers do they have respectively). So the conclusion above maybe not be so accurate. And due to the lack of available data, the following network could not be generated due to these limitations(e.g.not being able to access the Twitter database).

7. Conclusion

In conclusion, this research aims to identify why rumors spread so fast on Twitter. Based on the concept of the social network, quantitative measures of social network and Katz Centrality, as well as some data about news spread on Twitter that are found on Github, it could be concluded that the key factors to the quickness of rumor spreading are ordinary Twitter users. The result indicates that though the original size of rumors may not be large, they grow fiercely as time goes by, and most spreaders do not play important roles in the whole social network(judging from their Katz Centrality value). Although the research clearly illustrates the trend of how fast rumor grows as well as Twitter users who are not very important in social networks are the main spreaders. After all, we could not define groups without detailed information about them. But we could conclude from the quantitative measures that rumors are more likely to be spread by different people.

So future studies should firstly collect data about spreaders: how many followers do they have; Whether they are internet celebrities.

Despite these limitations, this research tells us that we should not blame Internet celebrities or some influential people for the quickness of rumor spreading. Instead, we should pay more attention to ourselves, considering most readers (including myself) are not of great significance in social networks, so we may probably be the main reason why rumors travel so fast. Therefore, we should be able to distinguish right from wrong and not trust any news easily until they are confirmed to be real.

This research mainly contributes to figuring out users that are don't play significant roles in social networks are the key factors to the quickness of rumors spreading, and helping the government to make decision appropriately on curbing the spread of rumors. With these findings, both government and citizens will be aware that everyone should be responsible for the rumors spreading, not some significant celebrities.

References

[1] Qin, Zhiwei; Cai, Jian; and Wangchen, H.Z. (2015) "How Rumors Spread and Stopover Social Media: a Multi-Layered Communication Model and Empirical Analysis," Communications of the Association for Information Systems: Vol. 36, Article 20.

[2] Doerr B, Fouz M, Friedrich T. Why rumors spread so quickly in social networks[J]. Communications of the ACM, 2012, 55(6): 70-75.

[3] Zhan J, Gurung S, Parsa S P K. Identification of top-k nodes in large networks using Katz centrality[J]. Journal of Big Data, 2017, 4(1): 1-19.

[4] Leskovec, J. , J. M. Kleinberg , and C. Faloutsos . "Graphs over time: densification laws, shrinking diameters and possible explanations." Proceeding of the eleventh ACM SIGKDD international conference ACM, 2005.

[5] Python Document. https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.distance_measures.diameter.html last accessed 2022/6/17.

DOI: 10.54254/2755-2721/2/20220621

[6] Wang, Yu; Ghumare, Eshwar; Vandenberghe, Rik; Dupont, Patrick (2017). "Comparison of Different Generalizations of Clustering Coefficient and Local Efficiency for Weighted Undirected Graphs". Neural Computation.
[7] Xiebin, liushenquan, liyanfeng, etc. Small world network characteristics of biological systems [c]// National Conference on nonlinear dynamics and motion stability, 2007.
[8] MA, Jing; GAO, Wei; MITRA, Prasenjit; KWON, Sejeong; JANSEN, Bernard J.; WONG, Kam-Fai; and CHA, Meeyoung. Detecting rumors from microblogs with recurrent neural networks. (2016). Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI 2016). 3818-3824. Research Collection School Of Computing and Information Systems.
[9] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In Proceedings of WWW.
[10] Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. 2013. Prominent features of rumor propagation in online social media. In Proceedings of ICDM.