International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 24 Issue 3 – FEBRUARY 2017 (SPECIAL ISSUE).

Detecting Influence in Twitter Network

K.SATHIYAKALA[#1]

[#] *Assistant Professor Department of Software Computer Science, Shanmuga Industries Arts & Science College,*
Manalurpettai Road, Tiruvannamalai, India

***Abstract*- An online social network is a platform with many advantages- it brings people together and keeps the communication active; it helps to exchange information online as well as offline; it is a marketing tool to engage consumers; it is also an election campaign tool for politicians. In such social systems, we also tend to observe different nature of social bonding. This correlation which may arise due to factors like homophily, social influence or external factors, emphasizes the behavioral patterns observed in individuals.**

Social Influence plays a major role in any social media marketing. The aim is to identify the presence of social influence in the network and then target the influencers. This when done strategically can lead to the success of a product launch.

In this paper, we confirm the presence of social influence in Twitter network by carrying out two simple tests- randomization test and reversal of edges test. We first do our analysis on three models of social correlation depicting- homophily, influence and no-correlation followed by a simulation on real Twitter data. Our simulation results show a strong social correlation attributed by social influence in Twitter network.

***Keywords*- Correlation, Homophily, Social Influence, Hashtag**

I. INTRODUCTION

Over the last few years, Twitter has defined a brand new way of online communication. People discuss myriad topics ranging from business, politics, celebrity and sports, breaking news and new age activism. Twitter has also revolutionized the expression of thoughts by the use of hashtags. Likewise, various other social networking sites like Facebook have brought people closer and provided a platform to exchange world views.

In all such online systems at times we see a very strong cascading effect in the nature of acceptance or rejection of an idea or a fact. This effect could mainly be due to the presence of social influence in the network. It has now become important to find the most influential node in any social network. In Twitter network e.g., even when a user has maximum number of followers he may not have the most influential standing in the network. Thus, twitter runs its own algorithm to score the influencers.

A study of social influence in many ways is very important for researchers to identify new trends in human behavior and social ties, and for businesses it is important as they identify new dimensions.

The nature of social ties in an online setting also defines one’s social behavior. When two friends join a Twitter community, it is defined as an act of homophily, as people with like mindedness performed an action together. When a user joins a Twitter community because his friends suggested him to, it is defined as an act of influence, as he was motivated by others to perform this action. Thirdly, when a user joins a Twitter community after learning about it from a newspaper article, it is defined as an act of confounding factor, as a newspaper article is an external factor contributing to the action.

The research carried out on distinguishing influence from correlation in [1] rules out influence as a source of correlation in the tagging behavior observed in Flickr network. Although related, our work focuses on proving that influence is indeed prevalent in a social setting like Twitter and we perform our study by observing its hashtag behavior.

II. SOCIAL CORRELATION

We study a social network G, where G is a directed graph generated from an unknown probability distribution. Here, we focus on each node performing a certain *task* for the first time. This *task* in our Twitter study refers to the use of a certain hashtag by an individual for the first time. We tend to observe that the same hashtag is being used by other nodes in the network at a different point in time. Now this nature of social affiliation could have three reasons – homophily, social influence or external factors.

In a network model, where the source of correlation is purely homophily, the set of nodes, W which has used a particular hashtag is selected according to some distribution and then the graph G is selected from these W nodes.

In a network model, where the source of correlation is due to confounding factors, we observe a correlation in the network G and the set of nodes W, due to some external variable X.

In the third model, which is the model of social influence is the most probable reason why we observe correlation in a social setting. In this model, the graph G is picked according to a certain distribution, after which the system is observed for a time period {1…T}. Then it is checked whether the propagation of the same behavior happens in the adjacent nodes in every time step between 1…T. Each adjacent node, of a node that has already performed the action may choose to adopt the same behavior or may not. As choosing to adopt the same behavior has binary outcome, we perform a logistic regression.

III. Methodology

In this section we describe the statistical equations used to perform the analysis and also the tests based on which we prove that Twitter network indeed has a very strong presence of influence. We briefly explain the randomization test and the reversal of edges test which determine the source of correlation in the Twitter network.

A. *Measuring social correlation*

In any social network the decision to become active at a time t is determined by how many active friends a person has in the network at that point in time. The estimation of this probability of choosing to be active is measured by the following logistic function:

$$p(a) = \frac{e^{\alpha \ln(a+1)+\beta}}{1+ e^{\alpha \ln(a+1)+\beta}} \qquad (1)$$

Where a is the number of friends who are already active and α, β are correlation coefficients, estimated later in the study. The measure of correlation is the given by the coefficient α.

$$\prod_a P(a)^{Y_a}(1-p(a)^{N_a}) \qquad (2)$$

Equation (2) is a maximum likelihood expression which is used to estimate the values of the coefficients α, β. Here, $Y_{a,t}$ is the number of active users, who started using a particular hashtag at a time t and $N_{a,t}$ is the number of inactive users, *p(a)* is obtained from equation 1 above. These values of $Y_{a,t}$ and $N_{a,t}$ are then summed up, $Y_a = \Sigma\, Y_{a,t}$ and $N_a = \Sigma\, N_{a,t}$.

B. *Randomization Test*

The randomization test or the time shuffle test is performed to identify the presence of social influence in a network. Anagnostopoulos *et al.* in [1] say that, if there is influence in the social network then the timing of an action by a node should depend on the timing of activation of her friend.

We perform the test at first with our randomly generated graph G and the set of W active nodes during a time period of [0, T]. Then Y_a and N_a values are computed, for $a \le R$, where R is a constant. We calculate the social correlation coefficient α using the maximum likelihood function.

The next step is to randomly shuffle the timestamps of each activated node in the set W, let us call this set W' which forms a part of the graph G. We then measure the correlation coefficient, α. If both the measures of α and α(W') are very different, then we can infer that influence is indeed present in the network.

C. *Reversal of Edges Test*

This test is similar in idea to that carried out in the study of obesity by *Christakis et al.* in [2]. The edge running from a→b means, "a is friends with b" and the only possibility of influence is from a to b. In the reversal of edges test, we change the direction of this friendship from b to a, b→a. We then measure the correlation coefficient in this case. Intuitively, the estimates of α in both cases should be very different as the direction of influence is reversed. The calculation of Y_a, N_a values and running the logistic regression takes place here too.

IV. Probabilistic models of social CORRELATION

We create three probabilistic models as defined by *Anagnostopoulos et al.* in [1] namely, no-correlation model, influence model and correlation model.

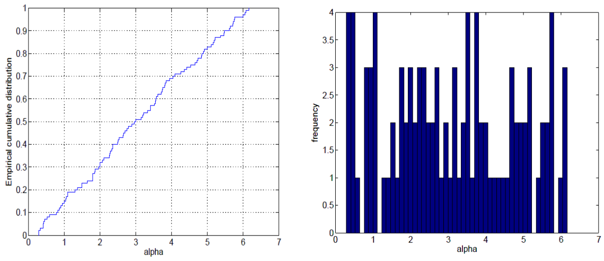
In the no-correlation model, the action is generated randomly so that we follow the nature of a real network. We take a note of the number of active nodes at each timestamp and how many new nodes get activated in the following timestamps between [0, T]. Equations (1) and (2) are then used to calculate the rest of the values to complete the model. In the influence model, we use a variety of α, β values. Here, at each timestamp a node whose friend is already active decides to become active or not and thus is a binary event. We use equation (1) to decide whether or not the node becomes active. In the correlation model, we pick a number of random centers. Then we perform a random permutation to select the set of nodes, W around these centers. We then repeat the same process followed in the no-correlation model for the activation of nodes.

A. Experiment on Randomly Generated Graph

We perform our experiments on the artificially created network. Figure 1 and Figure 2 are obtained after applying logistic regression to the influence model and correlation model respectively. The values of α in both the cases are positive.

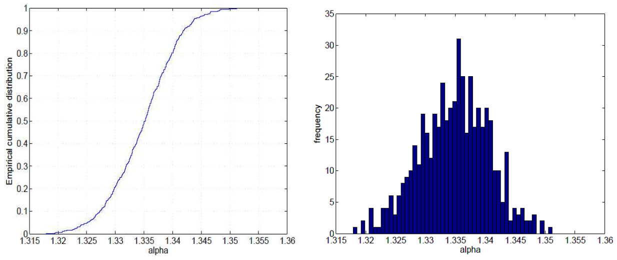
We now present the results of shuffle test and edge reversal test on influence model and correlation model. In Figure 3, we can notice that there is a shift to the left in the cumulative density function (CDF) which is indicative of the fact that on reversing the edges there is a decrease in the value of α. In Figure 4 we see that even after the shuffle test there is not much difference in the values of α. This is in line with our analysis and the values of α are very close with and without the shuffling of timestamps.

In Figure 5 and Figure 6 we present the results of edge reversal test performed on influence model and correlation model. In Figure 5 we see that there is a similarity in the result obtained for the influence model in the shuffle test and the edge reversal test. We see that there is a vast difference in the



(a) Empirical CDF (b) Histogram

Figure 1: Distribution of α for the influence model.



(a) Empirical CDF (b) Histogram

Figure 2: Distribution of α for the co-relation model.

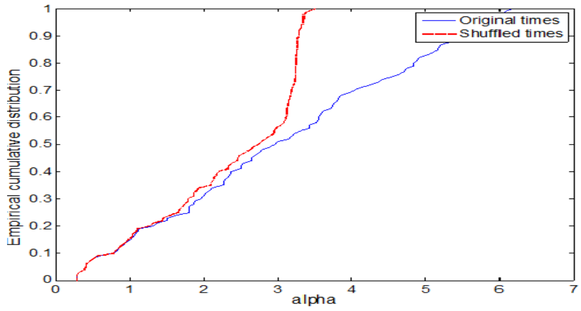


Figure 3: Time-Shuffle test for the Influence model.

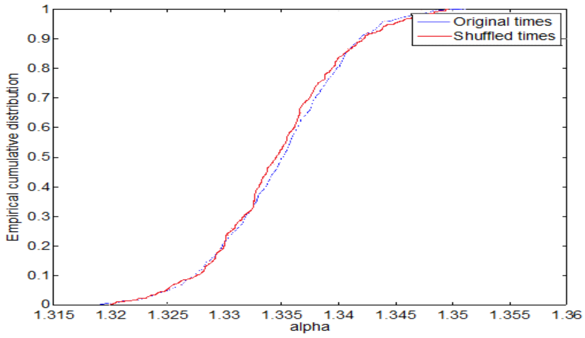


Figure 4: Time-Shuffle test for the Correlation model.

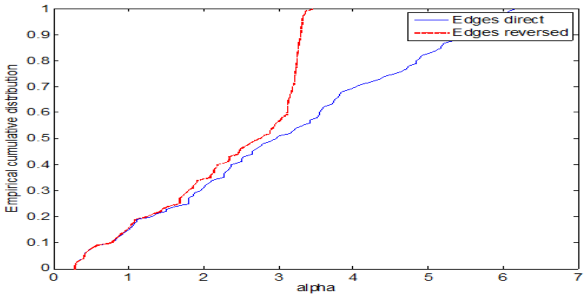


Figure 5: Edge Reversal Test for the influence model.

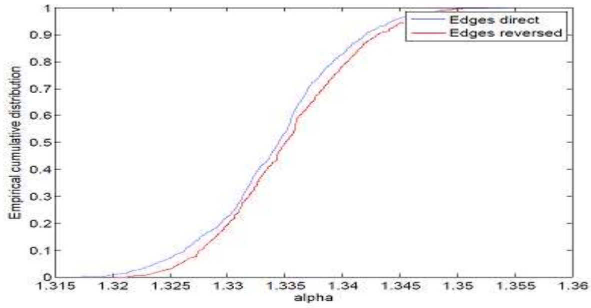


Figure 6: Edge Reversal Test for the correlation model.

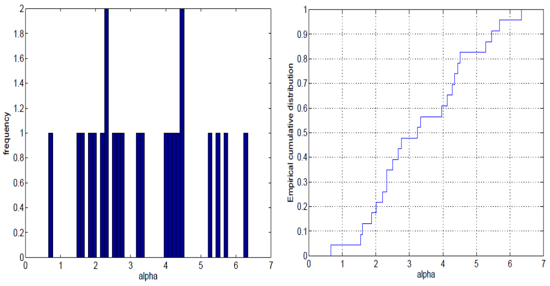
V. EXPERIMENTS ON TWITTER DATA

We perform our analysis on Twitter dataset. Twitter is a popular microblogging social platform where every message posted by a user is termed as a tweet. A tweet when posted can be viewed by an audience of the user's *followers*. On Twitter user-user communication can happen through *retweets* and *mentions*. A retweet happens when the content posted by a particular user is posted again by another user. Thus, the popularity of a tweet goes up in the forum. A *mention* is when a user is mentioned by another user in his tweet. A *hashtag* which is a popular feature of Twitter allows users to categorize tweets. The symbol # is called a hashtag. When a word is preceded by this symbol it becomes a keyword and when this hashtag becomes very popular, it becomes a topic that is trending. For e.g. "Freaky Friday" is expressed as #FF.

The dataset we use for our analysis is the publicly available dataset by *Conover et al*. [3] which was collected between September 14th and November 1st, 2010, during the run-up to the November 4th U.S. congressional midterm elections. The dataset has three versions, each specifying a network representing different types of tweets.

Retweet- This network contains only directed retweet edges between users. If X retweets content posted by Y, then an edge runs from X to Y. This indicates that information has diffused from X to Y.

Mention- This network contains only directed reply edges between users. If X mentions Y in a tweet, then an edge runs from X to Y. This indicates that information may have propagated from X to Y.



(a) Histogram (b) Empirical CDF

Figure 7: Distribution of α for the *retweet* network.

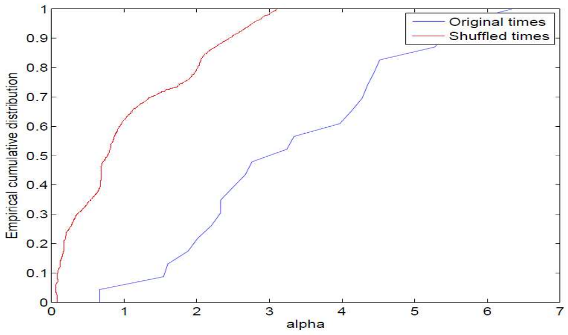


Figure 8: Time-Shuffle test for the *retweet* network.

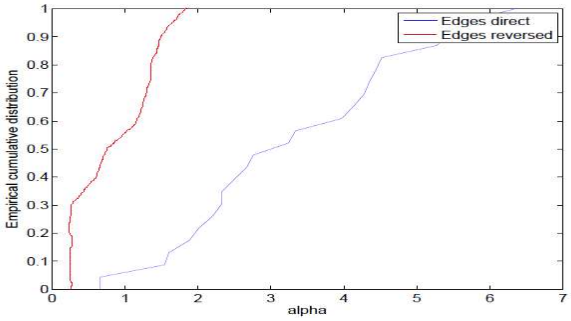
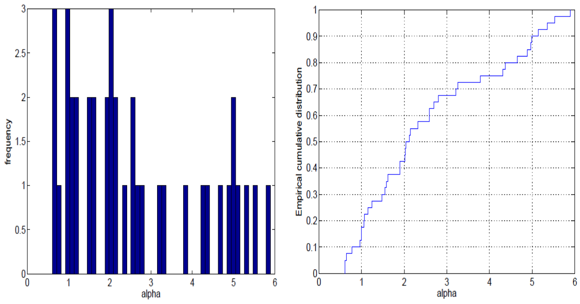


Figure 9: Edge-reversal test for the *retweet* network.



(a) Histogram (b) Empirical CDF

Figure 10: Distribution of α for the *mention* network.

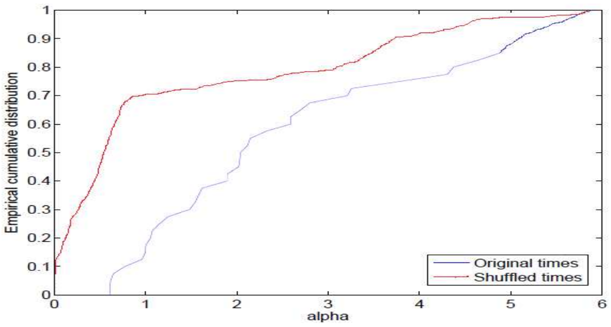


Figure 11: Time-Shuffle test for *mention* network.

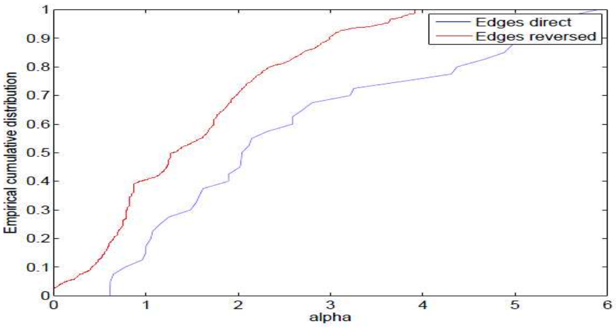


Figure 12: Edge-reversal test for *mention* network.

All- This network contains both the types of edges mentioned above.

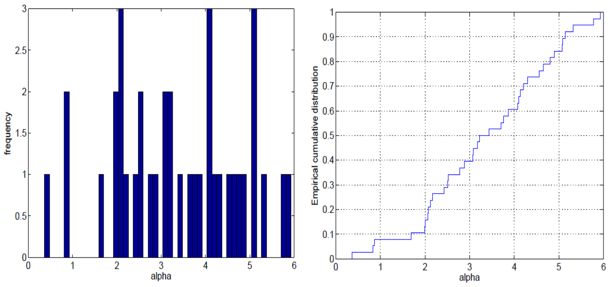
The *retweet* network consists of 18,470 nodes and the *mention* network consists of 7,175 nodes. These nodes are studied for our analysis. The networks are not symmetric and hence an edge running from x→y only means that the flow of information is from x to y and reverse is not true.

The hashtags are political in nature and some of them are #p2 meaning "Progressives 2.0", #tcot meaning "Top Conservatives on Twitter", etc. We pick one hashtag at a time and study the network and then do the tests.

In Figure 7, we see the distribution of α for all the hashtags in the *retweet* network. We can see that there is influence in the use of hashtags in this network. To validate this, we perform the two tests and in Figure 8 and Figure 9 we see that the values of α are shifted to the left.

In Figure 10, we see the distribution of α for all the hashtags in the *mention* network. In this network we see the presence of influence as well. To validate this, we perform our tests on this network and from Figure 11 and Figure 12 we observe that the tests verify the correlation to be based on influence too.

We now do a similar analysis for the *all* network. Like in the previous cases we see that this network is also a network of influence. A further test of detecting influence only confirms its presence in the hashtag usage of Twitter network.



(a) Histogram (b) Empirical CDF

Figure 13: Distribution of α for the *all* network.

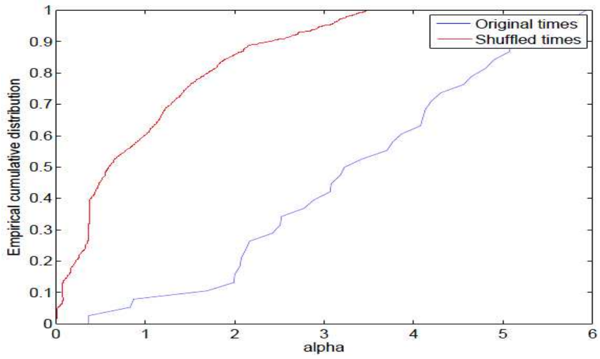


Figure 14: Time-Shuffle test for the *all* network.

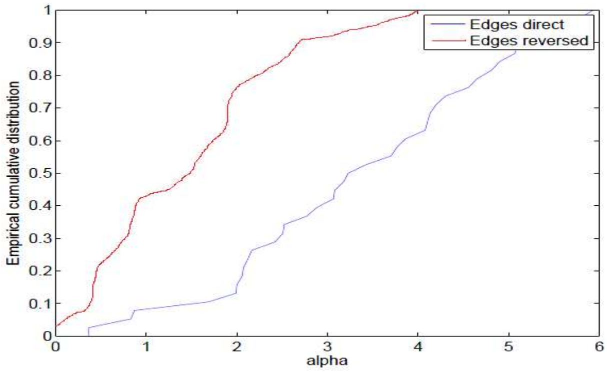


Figure 15: Edge-reversal test for *all* network.

In the research performed on tagging behavior in the Flickr dataset, *Anagnostopoulos et al*. [1] prove that even though there is correlation in the tagging behavior this is not due to influence. We perform our analysis on the Twitter data and verify the significance of the tests in correctly identifying influence in the Twitter network.

VI. References

[1] ArisAnagnostopoulos, RaviKumar, MohammadMahdian: Influence and correlation in social networks. KDD’08: 7-15.

[2] N. A. Christakis and J. H. Fowler. The spread of obesity in a large social network over 32 years. The New England Journal of Medicine, 357(4):370{379, 2007.

[3] Conover, M.; Ratkiewicz, J.; M. Francisco; Goncalves, B.; Flammini, A.; and Menczer, F. 2011. Political Polarization on Twitter. Proc of the 5th Intl. Conf on Weblogs and Social Media (ICWSM).