Predicting Information Diffusion Patterns in Twitter

Eleanna Kafeza[1], Andreas Kanavos[2], Christos Makris[2] and Pantelis Vikatos[2]

1. Athens University of Economics and Business, Greece, kafeza@aueb.gr
2. Computer Engineering and Informatics Department, University of Patras, Greece
{kanavos, makri, vikatos}@ceid.upatras.gr

Abstract. The prediction of social media information propagation is a problem that has attracted a lot of interest over the recent years, especially because of the application of such predictions for effective marketing campaigns. Existing approaches have shown that the information cascades in social media are small and have a large width. We validate these results for Tree-Shaped Tweet Cascades created by the ReTweet action. The main contribution of our work is a methodology for predicting the information diffusion that will occur given a user's tweet. We base our prediction on the linguistic features of the tweet as well as the user profile that created the initial tweet. Our results show that we can predict the Tweet-Pattern with good accuracy. Moreover, we show that influential networks within the Twitter graph tend to use different Tweet-Patterns.

Keywords: Information Diffusion, Machine Learning, Social Media Analytics

1 Introduction

Social contagion, or as it is often called word-of-mouth, is mainly based on viral marketing strategies. There is a plethora of tools in the area of social media analytics that are available to help marketers to develop a real time viral marketing strategy. A central question when developing such strategies is how information is diffused within the social network. Recently an augmenting body of research has focused on examining the nature and structure of information diffusion from the point of view of an information cascade. For example, in the case of Twitter, an information cascade is the followers of a user that retweeted the initial user message. We look into the problem of defining typical cascades and predict when they will appear in the social graph. We argue that the content of the message and the type of network diffusion are the two main dimensions that lead to robust cascade prediction.

The problem of how information spreads within social media has been examined from different perspectives in the literature. In several cases, information flow has been associated to virus contamination and a diverse set of models have been developed trying to count the spread of information [1], [6], [19]. Yet,

there are fundamental differences in the case of information spread as opposed to virus spread mainly because the spread is based on a network structure and is influenced by the message. For instance, it has been observed in studies [19] that social media users who are several steps afar tend not to propagate information of each other. Although probability decay models have been proposed to capture that element, it is not certain that such decay occurs in a homogenous manner. Moreover, existing studies show that large cascades are rare and usually, larger cascades are utilized in terms of width and not depth. These results are in accordance to our findings. Hence, since most cascades appear, we need a more pragmatic approach in order to predict one.

In the problem of Twitter message diffusion, we take a different approach and similarly to [2] we investigate the actual users diffusion. Our objective is to predict the information diffusion of users for messages related to a specific topic. Our contribution is that we present a solid method for predicting tweet cascades within the Twitter graph. We argue that not all users and their connections behave in the exact same way. We assume that there are two aspects that drive the information spread, the message itself and the network structure. So, for different message subjects different diffusion occurs. For a given subject we examine the frequencies of all possible diffusions and thus we extract some simple and basic patterns of diffusion. We then, associate the content of tweets with the propagation patterns, using linguistic analysis and machine learning techniques. Given a tweet of a user, we are then able to predict its pattern path. We do so, with very good accuracy as our results indicate. Hence given a user's tweet on a subject we predict the approximate path that the tweet will take into the Twitter graph. Moreover, we show that influential communities tend to have a diverse set of propagation patterns.

The rest of the paper is organized as follows: Section 2 is the related work. In Section 3 we present the methodology of our approach. The implementation and our results are presented in Section 4. We conclude in Section 5 where we also present our future work.

2 Related Work

In [17], a nonstandard form of Bayesian shrinkage implemented in a Poisson regression is proposed. There, the approach identifies the specific users who most influence others' activity and does so considerably better than simpler alternatives. The authors find that for the social networking site data, approximately one-fifth of a user's friends actually influence their activity level on the site.

Also, in [15], the authors present a study towards finding influential authors in Twitter brand-page communities (many enterprises have set-up their official webpages) where an implicit network based on user interactions is created and analyzed. More specifically, author profile and user interaction features are combined in a decision tree classification framework (DT framework) and thus, a novel objective evaluation criterion is used for evaluating these features.

Another work was [10], where a novel method to find influentials by considering both the link structure and the temporal order of information adoption in Twitter is adopted. Recently, in [9], the authors consider the issue of choosing influential sets of individuals as a problem in discrete optimization. The optimal solution is NP-hard for most models that have been studied, including the model of [5]. The framework proposed in [16], on the other hand, is based on a simple linear model where the solution to the optimization problem can be obtained by solving a system of linear equations. The generality of the mentioned NP-hard models lies between that of the polynomial-time solvable model of [16] and the very general model of [5], where the optimization problem cannot even be approximated to within a non-trivial factor.

In [18], the authors compare a wide assortment of node-level network measures (degree centrality, clustering coefficient, network constraint, and eigenvector centrality), testing their robustness to different forms of measurement error. They also investigated network-structural properties (average clustering, degree distributions) as explanations for the varying effects of measurement error. In addition, Leskovec et al. [11] deal with information cascades; they are phenomena in which an action or idea becomes widely adopted due to influence by others. They develop a scalable algorithm and set of techniques to illustrate the existence of cascades as well as to measure their frequencies. From their experiments, they found that most cascades are small, cascade sizes approximately follow a heavy-tailed distribution, the frequency of different cascade subgraphs depends on the product type and last that these frequencies reflect more subtle features of the domain in which the recommendations are operating.

In article [4], a business process classification framework to put the research topics in a business context is employed while providing a critical perspective on business applications of social network analysis and mining. Furthermore, in [12], authors provide some assumptions concerning epidemic models. They find that the probability of purchasing a product increases with the number of recommendations received, but quickly saturates to a constant and relatively low probability. Also, they present a simple stochastic model that allows for the presence of relatively large cascades for a few products, but reflects well the general tendency of recommendation chains to terminate after just a short number of steps; they observe that the most popular categories of items recommended within communities in the largest component reflect differing interests between these communities.

Another empirical study of user-to-user content transfer occurring in the context of a time-evolving social network in Second Life (e.g. a massively multiplayer virtual world) is presented in [2]. There, adoption rates quicken as the number of friends adopting increases and this effect varies with the connectivity of a particular user. Also, in [3], authors find out that the largest cascades tend to be generated by users who have been influential in the past and who have a large number of followers. The authors in [13] focused on the problem of link prediction in Microblogs and proposed the notion of social distance based on the interaction patterns. The main idea is that social networks exhibit homophily

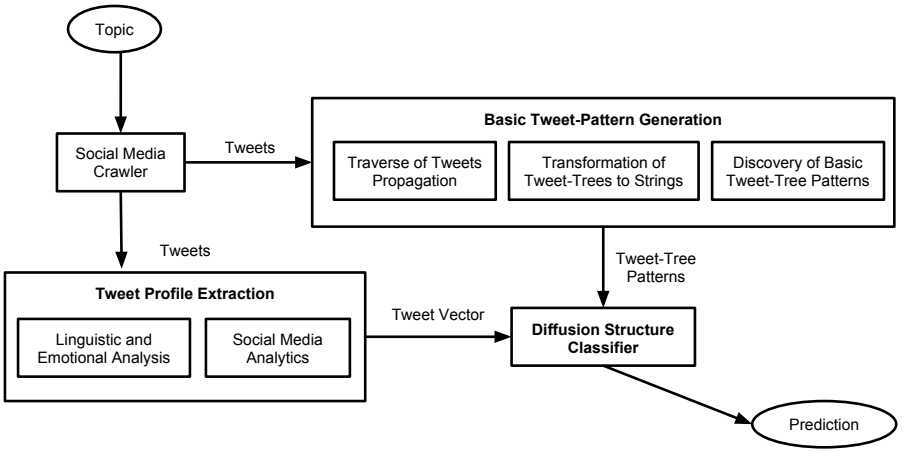

Fig. 1. System Architecture of the Tweet Prediction Pattern System

and that the agents prefer to create ties with other agents who are close to them. In [6], methods for identifying influential spreaders in online social networks are presented and in following, a taxonomy of various approaches employed to address diffusion modeling techniques is proposed.

Finally, Peng et al. [14] introduced the basic characteristics of the diffusion process of multi-source information via real datasets collected from Digg (e.g. a social news aggregation site); in following, a mathematical model to predict the information diffusion process of such multisource news is used.

3 Methodology

In our model we examine information diffusion in Twitter as Tree-Patterns. Our methodology for predicting the tweet cascade is based on the linguistic and emotional content of the tweets as well as the user communication behavior. According to Twitter: "A Retweet is a re-posting of someone else's Tweet. Twitter's Retweet feature helps you and others quickly share that Tweet with all of your followers". Although ReTweets might have other forms as well (for example the use of RT in the beginning of the message denotes that it is a ReTweet), still most of the ReTweets are done according to the formal Twitter definition using the @username convention. This is the conversion we use to retrieve ReTweets. Figure 1 depicts the architecture of our system.

We represent the information cascade as a Tree-Pattern i.e. the set of nodes that retweeted the initial user tweet. Note that all of these nodes are related with the "follow" relationship since only a user's followers can retweet their message.

The Social Media Crawler takes a topic as input and retrieves the information from Twitter. It collects all the Tweets and their corresponding ReTweets on the specific topic. The Tweet Profile Extraction is the module where the Tweets are inserted and they are abstracted as vectors that contain linguistic and emotional information about the Tweet as well information regarding the user behavior as depicted through the Twitter analytics i.e. number of followers etc. The Basic Tweet-Pattern Generation is a module that creates the frequent basic Tweet Propagation Patterns. The Tweet Vector and the Patterns are the input to the Diffusion Structure Classifier that predicts the Tweet-Pattern that will be followed by a given user tweet.

3.1 The Social Media Crawler

In order to retrieve information from the Twitter graph we use a crawler topic-based sampling approach where Tweets are collected via a keyword search query. Our data source is Tweets retrieved from the Twitter through the Twitter API (e.g. Twitter4J[1]) for a specific period of time (e.g. a couple of hours). We retrieved all the Tweets that were relevant to the subject *#MH370* concerning Malaysia Airlines Flight 370 disappearance. Our dataset contains 13000 Tweets that have been done by 11130 users.

The crawler is responsible for sampling and traversing the Twitter media; it also collects information regarding the users' activity. More specifically, we extract 6 basic user features which better describe user communication behavior in Twitter. The set of 6 features contain the number of Followers, the number of Direct Tweets, the number of ReTweets, the number of Conversational Tweets (e.g. if a user replies to a post), the Frequency of user's Tweets (e.g. how "often" an author posts Tweets) and last the number of Hashtag Keywords (e.g. words starting with the symbol # and can specify the thematic category of a specific Tweet) as in [7], [8].

3.2 Generating the Tweet-Pattern

To measure how tweets propagate through the Twitter network we represent tweet propagation as a tree. The root of the tree is the user that posts the initial tweet. Every child node is a follower of the initial user that has retweeted the initial tweet. This continues recursively until there are no nodes that retweeted the initial tweet. Hence each Tweet-Tree represents all the followers that retweeted the same message as the root user.

In Figure 2 we present some Tweet-Trees returned by our crawler. As previous studies have shown and as our results have shown as well, there are no large cascades of information, and Tweet-Trees tend to be small. So, our first objective is to find a set of representative Tweet-Patterns.

We observe that isomorphic Tweet-Trees exhibit the same diffusion. Isomorphic paths have the level number of a node as invariants, the number of paths

[1] Twitter4J library: http://twitter4j.org/en/index.html

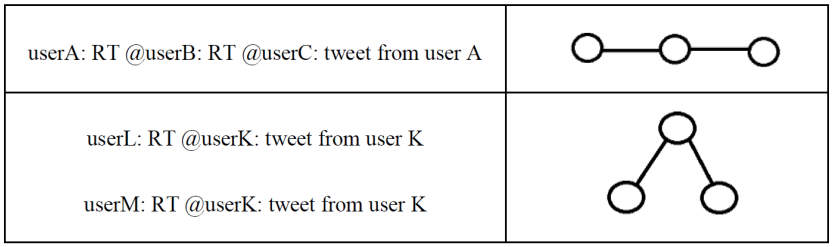

Fig. 2. Examples of Tweet-Trees

from the root to the leaves, the number of levels in the tree, the number of leaf descendants of a node and the level number of a node. In information diffusion we are interested in the spread of information. Additionally, we are interested in the number of the leaf descends and not the specific position of the subtree that these descendants come from. Therefore we represent in the same way all isomorphic Tweet-Paths, taking as their representative the left-most variant.

For example in Figure 3, we see three different Tweet-Trees. All of them exhibit the same information diffusion thus we model them choosing one representative; the one depicted in (c).

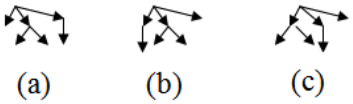

Fig. 3. Isomorphic Tweet-Trees that represent the same information Diffusion

3.2.1 Representing Tweet-Trees as Strings We encode the diffusion of each tweet as a string and use top-down approach to traverse the tree. The root is the tweet that has been produced by a certain user. Let k be the maximum number of different users that have posted a ReTweet from a user from the whole dataset. Hence given a root tweet, there are at most k different possibilities of diffusion for each node of the tree. The order of these possibilities depends on the depth and width of diffusion. Starting from the root we use symbols to mark the diffusion. The placement of symbols begins from root to the next level. If there is more than one node in the current level we order the created symbols in reverse lexicographic way and we put the symbols respectively.

For each node there are k different possibilities of diffusion to the next level. In total there are k^m different strings where k are different possibilities of diffusion to the next level and m the max depth of the trees. We have to note that based on the above construction, isomorphic trees have the same string representation.

3.2.2 Identifying the Basic Tree-Patterns We use edit distance as a metric to identify the basic tree patterns. In particular the edit distance between two strings s_1, s_2 of size d_1, d_2 respectively is defined as he number of symbol operations (insertion, deletion and substitution) that must be performed in order to transform one string into the other. The edit distance is equal to the cost of the alignment of the two strings that is defined as the number of character mismatches when putting the one string under the other embedding spaces in both of them in order to produce strings of equal length. The edit distance can be computed with a simple dynamic programming algorithm as it is nicely described in [17] in time $d_1 * d_2$. More specifically a two dimensional table D of size $d_1 * d_2$ is defined where $D[i, j]$ is equal to the edit distance between $s_1[1 \dots i]$ and $s_2[1 \dots j]$ and can be expressed as the minimum of $D[i-1, j] + 1, D[i, j-1] + 1, D[i-1, j-1] + (i, j)$, where (i, j) is an indicator variable equal to 1 if $s_1[i]$ and $s_2[j]$ are different, otherwise it is 0. The needed solution is given by $D[d_1, d_2]$ and it can be produced by a simple traversal of the array from top to bottom and from left to right.

Each produced string is compared to others using a table of dynamic programming via edit distance and the frequent strings are gathered in bins. As a result of the above process, a set of basic Tweet Tree-Patterns are extracted. These frequent patterns constitute the pre-assigned labels of each tweet for the training of the classification models.

3.3 Using Linguistics to Represent the Tweet as a Vector

After having specified the basic Tweet-Patterns, we argue that given a tweet on a specific topic the diffusion basic pattern that this tweet will follow is based on the tweet content and on the user profile. Hence we represent each tweet as a vector in which we extract linguistic and emotional information as well as information regarding the user that posted the tweet.

The following are the two main characteristics that we include in the Tweet-Vector:

- Linguistic and Emotional characteristics, which are produced by the Linguistic Inquiry and Word Count (LIWC) software which produce 80 features that include linguistic and psychological use of language as well as personal concerns.
- Social Media Analytics, which can be used to monitor and capture user's behavior. The Followers of a user, the number of Tweets, ReTweets, Conversations and Hashtag Keyword as well the Frequency of Tweets, are some aspects that differentiate user behavior.

Thus every tweet is represented as a vector that contains the 80 features that were extracted when the tweet was processed with LIWC for linguist extractions and the 6 social analytics metrics that represent information about the user that created the initial tweet. We predict the Tweet-Pattern based on that vector using a variety of classification algorithms. The performance is evaluated by the F-Measure metric.

4 Implementation and Results

As already mentioned our dataset contains 13000 Tweets that have been done by 11130 users. In these Tweets, the maximum depth of the Tweet-Tree was 4 and the maximum width was 92. Thus we verify existing results which claim that the depth of the Tweet-Tree is small and the width large [10]. We encode the tweet information diffusion using 92^4 different representations as previously explained. After applying the methodology presented in Section 3.2, we end-up with the following basic Tweet-Patterns:

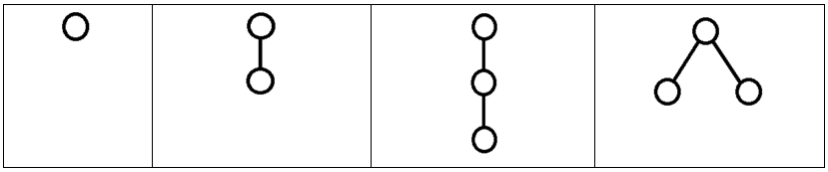

Fig. 4. The Basic Tweet Patterns

As explained in Section 3.2, the next step in our methodology is to represent the tweet as a vector based on its linguistic characteristics and its user profile as described in the analytics. We used LIWC as a tool to extract linguistic characteristics and we used Twitter analytics to extract the user profile. As a result, we created a vector of 86 characteristics, as presented in Table 1.

Table 1. The Characteristics of the Tweet Vector

Features	**#**	**Description**
LIWC	80	4 general descriptor categories (total word count, words per sentence, percentage of words captured by the dictionary, and percent of words longer than six letters), 22 standard linguistic dimensions (e.g., percentage of words in the text that are pronouns, articles, auxiliary verbs, etc.), 32 word categories tapping psychological constructs (e.g., affect, cognition, biological processes), 7 personal concern categories (e.g., work, home, leisure activities), 3 paralinguistic dimensions (assents, fillers, nonfluencies), and 12 punctuation categories (periods, commas, etc)
Twitter Metrics	6	Followers, Tweets, ReTweets, Conversations, Frequency, Hashtag Keywords

We separated dataset to training and test set, using two approaches: a) K-Fold Cross-Validation (K=10 Fold) and b) Leave-One-Out Cross-Validation. The concept of using both techniques is that splitting with 10-Fold Cross-Validation, important information can be removed from the training set. However, the

Table 2. F-Measure for each Classifier

Classifiers	F-Measure
AdaBoost	0.71
IBK	0.71
J48	**0.88**
JRip	**0.89**
Multilayer Perceptron	0.81
Naive Bayes Classifier	0.61
PART	**0.88**
RotationForest	**0.88**
SMO	0.77

Leave-One-Out Cross-Validation technique evaluates the classification performance based on one sample.

We used the WEKA library[2] for the classifiers. Table 2 shows the results of F-Measure for each classifier. The classifiers that give us the best results are depicted in bold. We can observe that JRIP achieves the highest F-Measure value. In addition, the next best classifiers are J48, PART as well as RotationForest. Moreover, we see that almost all classifiers achieve an F-Measure above 60%; only one classifier achieves 61%, while the rest are above 70%. Hence the accuracy of our prediction is high which validates our hypothesis that based on the linguistic aspects of the tweet and the user analytics profile, we can predict the Tweet-Tree.

As an application of our approach, we examine the different patterns that occur in highly influential communities. Based on our previous work [8], we extract the existing communities of the graph taking into consideration the personality of the users. We rank these communities based on the number of Tweets. The first in the rank are the most influential communities i.e. the communities where most Tweets occur. We are interested in finding out the type of diffusion that occurs in these communities.

We expect that the top influential communities are those that diffuse more the information. The first Tweet-Pattern which is a node by itself does not exhibit any diffusion since the tweet was not retweeted. Hence we expect that for the influential communities the diffusion patterns should mainly be the other three. Figure 5 verifies that expectation. It shows that when ranking the influential communities in ascending order based on the number of tweets that occurred in them, then the pattern of singe tweet (diffusion class 1) occurs less to the most influential community (com1). While the rest patterns (diffusion classes 2, 3 and 4) occur more in com1 and com2 than in com3. Therefore we conclude that communities where the number of tweets is large (influential communities) have more diffusion where ReTweets occur either as in sequence or in width.

[2] Weka toolkit: http://www.cs.waikato.ac.nz/ml/weka/

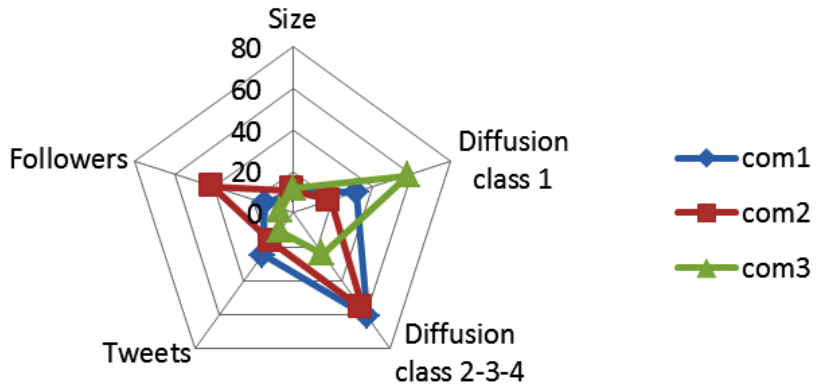

Fig. 5. Information Diffusion in Top Communities

5 Conclusions and Future Work

Recently there has been a growing interest in the literature for examining how information is diffused in social networks especially in the case where information cascades are short and usually with small depth and large width. In our work we use existing evidence supported by our results, to show that these Tweet-Trees have specific patterns. Hence we find the representative Tweet-Tree patterns that information follows when propagated in Twitter. We argue that this Tree-Pattern information diffusion can be predicted given a user's tweet and we use linguistic and user profiling information in order to do so based on machine learning techniques. Our results show that we can predict the basic Tree-Pattern with good accuracy. Our contribution is important especially in cases where marketers are interested in identifying influential users and networks and estimate the propagation of their messages. We show that in influential networks there is a set of different Tree-Patterns that occur.

In our future work, we are interested in examining the overall propagation of information within a whole Twitter network based on our prediction methodology. Also, as a next step we will consider time as a factor that influences the information diffusion.

References

1. S. Aral and D. Walker. (2010). Creating Social Contagion through Viral Product Design: A Randomized Trial of Peer Influence in Networks. ICIS.
2. E. Bakshy, B. Karrer and L. A. Adamic. (2009). Social Influence and the Diffusion of User-Created Content. EC, pp. 325-334.
3. E. Bakshy, J. M. Hofman, W. A. Mason and D. J. Watts. (2011). Everyone's an Influencer: Quantifying Influence on Twitter. WSDM, pp. 65-74.

4. F. Bonchi, C. Castillo, A. Gionis and A. Jaimes. (2011). Social Network Analysis and Mining for Business Applications. ACM Transactions on Intelligent Systems and Technology, Volume 2, Number 3.
5. P. Domingos and M. Richardson. (2001). Mining the Network Value of Customers. KDD, pp. 57-66.
6. A. Guille, H. Hacid, C. Favre and D. A. Zighed. (2013). Information Diffusion in Online Social Networks: A Survey. SIGMOD Record, Volume 42, Number 2, pp. 17-28.
7. E. Kafeza, A. Kanavos, C. Makris and D. Chiu. (2013). Identifying Personality-based Communities in Social Networks. LSAWM (ER).
8. E. Kafeza, A. Kanavos, C. Makris and P. Vikatos. (2014). T-PICE: Twitter Personality based Influential Communities Extraction System. IEEE International Congress on Big Data.
9. D. Kempe, J. M. Kleinberg and E. Tardos. (2003). Maximizing the Spread of Influence through a Social Network. KDD, pp. 137-146.
10. C. Lee, H. Kwak, H. Park and S. B. Moon. (2010). Finding Influentials Based on the Temporal Order of Information Adoption in Twitter. WWW, pp. 1137-1138.
11. J. Leskovec, A. Singh and J. M. Kleinberg. (2006). Patterns of Influence in a Recommendation Network. PAKDD, pp. 380-389.
12. J. Leskovec, L. A. Adamic and B. A. Huberman. (2007). The Dynamics of Viral Marketing. ACM Transactions on the Web (TWEB), Volume 1, Number 1.
13. D. Liu, Y. Wang, Y. Jia and J. Li. (2014). From Strangers to Neighbors: Link Prediction in Microblogs using Social Distance Game. Diffusion Networks and Cascade Analytics (WSDM).
14. C. Peng, K. Xu, F. Wang and H. Wang. (2013). Predicting Information Diffusion Initiated from Multiple Sources in Online Social Networks. ISCID, Volume 2, pp. 96-99.
15. H. Purohit, J. Ajmera, S. Joshi, A. Verma and A. P. Sheth. (2012). Finding Influential Authors in Brand-Page Communities. ICWSM.
16. M. Richardson and P. Domingos. (2002). Mining Knowledge-Sharing Sites for Viral Marketing. KDD, pp. 61-70.
17. M. Trusov, A. V. Bodapati and R. E. Bucklin. (2010). Determining Influential Users in Internet Social Networks. Journal of Marketing Research (JMR).
18. D. J. Wang, X. Shi, D. A. McFarland and J. Leskovec. (2012). Measurement error in network data: A re-classification. Social Networks (SOCNET), Volume 34, Number 4, pp. 396-409.
19. F. Wu, B. A. Huberman, L. A. Adamic and J. R. Tyler. (2004). Information Flow in Social Groups. Physica A, pp. 327-335.