Observing Behaviors of Information Diffusion
Models for Diverse Topics of Posts on VK

Aisulu Khairullina*, JooYoung Lee†
Department of Computer Science
Innopolis University
Innopolis, Russia 420500
*ais.khairullina@gmail.com, †j.lee@innopolis.ru

Gwan Jang‡, Sung-Hyon Myaeng§
Department of Computer Science
Korea Advanced Institute of Science and Technology
Deajeon, Republic of Korea
‡gjang@kaist.ac.kr, §myaeng@kaist.ac.kr

Abstract—The way information spreads through society has changed significantly over the past decade with the advent of online social networking. It is also observed that users have distinct behaviors, i.e., the topics of conversations shared among users, based on which social media platforms they use. However, many previous approaches for predicting information spreading in social networks do not consider this versatility. In this paper, we examine Independent Cascade (IC) information diffusion model which assumes that each node independently influences its neighboring nodes. We show the results of applying IC model to the biggest Russian social network Vkontakte (VK). We first apply the model to synthetic networks and compare the results with the real networks extracted for different topics. The results support our hypothesis that the behavior of information diffusion in social media is different based on the topics shared. Our results also show that IC model does not properly describe the diffusion processes in VK.

I. INTRODUCTION

Communication is essential to all individuals and their communities. People are entitled to participate in communications and making decisions within societies. The main purpose of communication is sharing and spreading information among participants of the groups. The importance of understanding how information is spread within a community has drawn interests of researchers as tools of social network analysis have made it easier to investigate information diffusion in social networks.

Information diffusion is a process by which a piece of information is spread and reaches individuals through interactions. The process involves two types of actors: sender(s) and receiver(s). Another key point in information diffusion process is a medium through which the diffusion takes place. For instance, the medium for a rumor spreading is the personal communication between individuals. Social network platforms, or online services those focused on facilitating the building of social networks or social relations among people who share interests, activities, backgrounds, or real-life connections. A social network service consists of a representation of each user (often a profile), his/her social links, and a variety of additional services. Most social network services are web-based and provide means for users to interact over the Internet, such as e-mail and instant messaging. Online community services are sometimes considered as a social network service though, in a broader sense, social network service indicates an individual-centered service whereas online community services are group-centered. Social networking sites allow users to share ideas, activities, events, and interests within their individual networks. There exist many types of these services and each of them has a distinct way of interacting, publishing and sharing for its users. Therefore, the characteristics of social network media should be considered when studying information diffusions as each network is unique and it affects the behaviors of users in it.

Lately, online social services have become not only platforms for communications among individual users but also effective tools for company-customers communications. Commercial companies are eager to establish reputations in online social communities to advertise their product or services and increase profits. However, to create an effective business, it is necessary to consider preferences of potential customers and build trust and loyalty. Thus, companies are interested in what people talk and feel about their products or services, to make customers share the experiences or opinions on their products as much as possible, i.e., to make the product information diffuse. Online social networks provide them with a great opportunity to gather opinions and draw attention of a large group of people. In addition, there is no need to make extra effort for companies to diffuse the information in social networks since users send messages, publish posts on the profile wall, share, and re-share voluntarily.

On the other hand, it is extremely hard to define initial time and set of users who can initiate the most efficient diffusion. This set may have various features: users with the high status online (friends number, comments amount under the wall posts, etc.), or in real life, similar interests, etc. People are highly influenced by each other, what is popular for others becomes popular for them as well. However, as far as the diffusion processes are still remaining under research, some part of users interaction keeps being covered and using it without preliminary testing might lead to unexpected consequences.

II. RELATED WORKS

Communication activities may seem to have specific features such as uncertainty or haphazardness, but many of research papers prove the opposite. Even more, they admit the important role of online social networks in human interactions and information diffusions [1]. The relationship between users or tie strength has a strong influence on information spreading. As strong tie, they consider frequent online messaging,
appearing on the same photo, commenting each other posts or the same one. Subjects who were exposed to a link shared by a friend from whom the subject received three comments are 2.83 times more likely to share than subjects exposed to a link shared by a friend from whom they received no comments. For those who were not exposed, the same comparison shows that subjects are 3.84 times more likely to share a link that was previously shared by the stronger tie. However, authors highlight that although strong ties are individually more influential, the effects of strong ties is not large enough to match the sheer abundance of weak ties.

Existing methods of estimating information diffusion in social networks study diffusion patterns arising from seven online domains (Yahoo!Kindness, Yahoo!Voice, Zync, the Secretary Game, Twitter News Stories and Videos, Friend Sense). Each of them has a various type of contents, modes of sharing, network density and time of information spreading. Researchers classified social platforms by cost, time, edges sharing, network density and time of information spreading. Each of them has a various type of contents, modes of spread for several weeks, Twitter on Twitter policy, but edges do not involve active communication because of privacy secretary Game, Twitter News Stories and Videos, Friend Sence). Yahoo!Kindness, Yahoo!Voice, Zync, the Secretary Game, Twitter News Stories and Videos, Friend Sense. The vast majority of cascades are small and are described by a simple tree structure that terminates within one degree of an initial adopting seed. As for large ones, diffusion takes place within one degree of a few dominant individuals [2].

Many scientist look for ways of using patterns of diffusion in a real life, especially, for viral marketings. In [3], authors address the problem of identifying small number individuals through whom the information can be diffused to the network as soon as possible to solve the diffusion minimization problem in mobile social networks from different aspects. Authors proposed a community-based algorithm which leverages the community structure and a distributed set-cover algorithm which collects information by probing messages in a distributed way. Admittedly there is at least one active node in the community and if the condition is ruled out then communities need to be merged in order to make diffusing nodes number equal or less than the number of communities. Also, there exist studies in which how information is exchanged in a common. The vast majority of cascades are small and are described by a simple tree structure that terminates within one degree of an initial adopting seed. As for large ones, diffusion takes place within one degree of a few dominant individuals [2].

For both models, nodes may have two states. The active state is that when the node has already influenced by information in diffusion. The inactive state is inherent the node which is unaware of the information or not influenced. In the Independent Cascade model, information flows over the network through the cascade. The model describes the process where active nodes influence inactive neighbor nodes with given diffusion probability. Note that the diffusion probability for IC model needs to be defined in advance. However, in practice, it is impossible to set the correct values in advance and we need to estimate the values using the real data.

Those models show the information propagation in directed network $G = (V, E)$ with no self-loops, where $V$ is set of the nodes and $E$ is the links between them. Nodes may change states only in one direction: from inactive to active. The initial set of active nodes is given in advance [16], whereas we use reposts graph in which it is obvious that our set $S$ includes only one node that is the original author id of the post.

III. INFORMATION DIFFUSION MODELS

Communications with friends, colleagues, neighbors or community members have influence on our decision making processes in everyday life. Also, our perceptions of information rely on the relationships with the source of the information. Therefore, it is important to reason about how information travels through a network to understand diffusion difference among various types of information. As some viruses are more infectious than others, some information spreads faster than others. For example, urgent news tends to travel fast regardless of media including mouth to mouth. On the other hand, some types of information do not attract people’s attention and spread slow. For instance, private information such as having a baby and getting married usually travels only to a closed group of acquaintances. Hence, studying the behaviors of information diffusion on social networks could benefit many applications which need to attract wide audiences. Traditionally, marketers have been trying to analyze information diffusion to find influential users who can maximize the diffusion so that they can be used for target marketing [10], [11]. To better benefit from this analysis, it is important to consider what types of information needs to be diffused.

Since influence plays a leading role in information spreading process in our real life, we may assume that it has the same part in online information propagation as well. In order to solve the influence maximization problem in online and networks, the various information diffusion models had been created. The most used models in these studies are Independent Cascade (IC) [12], [1] and Linear Threshold (LT) [13], [14], [15] models.

For both models, nodes may have two states. The active state is that when the node has already influenced by information in diffusion. The inactive state is inherent the node which is unaware of the information or not influenced. In the Independent Cascade model, information flows over the network through the cascade. The model describes the process where active nodes influence inactive neighbor nodes with given diffusion probability. Note that the diffusion probability for IC model needs to be defined in advance. However, in practice, it is impossible to set the correct values in advance and we need to estimate the values using the real data.

Based on those aspects, several general types of informa-
A. Problem Statements

Many papers describe information diffusion using worldwide online social networks such as Twitter, Facebook, Google+, etc. Those networks show an extremely high rate of activeness of users and provide a huge amount of data. However, diffusion patterns which are effective for global network may not be applicable for local ones since diffusion process depends on many factors such as country, mentality, culture, economical and political conditions and technological progress.

We target to study users of online social networks who share similar geographical locations and historical background and culture. We selected Vkontakte (VK) online social platform which is the largest Russian social network in Europe. As of November 2014, VK had about 280 million accounts. The biggest share in country distribution belongs to Russia (70.26%), Ukraine (18.4%) and Belarus (4.6%) are in the second and third place, respectively. Kazakhstan (1.63%), Moldova (0.54%), USA (0.54%) and Germany (0.49%) come next. VK allows users to create profiles, communities, public pages, to message each other privately, to post and share messages, photos, audio, video on their profile walls. Each post may contain up to 10 attachments: media files, maps, and documents, etc. News feeds can be switched between all news (default) and most interesting modes.

The goal is to define patterns of spreadings of VK user’s wall posts, ways of diffusion optimization in terms of duration and density as well as to determine influence of the post structure, semantic features on the diffusion process. We applied Independent Cascade model to the posts of various categories and compare results with real network post diffusion.

IV. DATA COLLECTION AND DATASET CHARACTERISTICS

Python (v.2.7) and Open Vkontakte Application Programming Interface (API) were used to create a program to access and collect the data from Vkontakte. Open VK API is a system for developers which enables easy authorization of VK users on the detached site, access to information about users friends, photos, audios, videos and other VK data. For that purpose, VK API methods are provided. Note that the API methods have a limitation in the number of processing information units. Some methods, for example, wall.getReposts and groups.getMembers, are able to return a list of not more than 1000 users.

We define the different categories of posts and we selected the most interesting cases, i.e., the posts which have relatively large numbers of reposts and depth, for each of them. This allowed us to study the relationships between post categories and their propagation patterns.

1) Advertisement: In this category, we define posts with advertisement or marketing campaign in social networks where a user is strongly motivated to make a repost for an opportunity to win a prize. Usually, this type of posts are published by communities rather than individual users.

2) Social appeal: In this category, we include posts that are calling for help such as in finding a missing child or in finding shelter for animals. This type of posts are published by a user in the aim to attract attentions of friends and friends of friends. It is closely related to the strength of the tie between individuals.

3) Entertainment posts: In this category, we include informative and humor posts of users. This type attracts larger audience than the second type. It can be published by a user and by a community as well. We also observe entertainment posts attract more likes than sharings.

V. EXPERIMENTS AND RESULTS

For the experiment we extracted the data of the wall post diffusion and then create post owner friendship network. We assume that the post can be reposted inside the network only, by a friend of the user with given probability (see Figure 4). Although in some cases, this is not true for the real information diffusion process: if post owner doesn’t deny access to his (her) wall, any user of Vkontakte social network may repost the post.
Thus, we set the post owner as a seed (the initial active node) and apply Independent Cascade model to imitate propagation with the given probability. As probability which has to be set in advance we define $\kappa = 1/\bar{d}$, where $\bar{d}$ stands for the average degree of edges in the network, in our case we can assume that $\bar{d}$ as average friends number for each user of the network. It is necessary to note that the given node can be influenced by the only one parent node but activate any of its child nodes. As a result, we receive the set of nodes which are activated with given probability.

We assume that depending on topic category each post follows different diffusion patterns, thus we attempt to model information propagation for each post category assignment individually. The code was written using Python (v.2.7). After implementing the Independent Cascade model we compare the synthetic dataset results with the real one. The results of the experiment are shown in Table I.

### VI. Discussions

We now have ways to compare the diffusion process with respect to Independent Cascade model for the Vkontakte observed dataset. Being able to apply the model to the users friends network enables us to analyze the diffusion process.

The results presented in Section V show that Independent Cascade model could not illustrate the information diffusion process in Vkontakte online social platform, no matter which category post belongs to. We applied Independent Cascade model to the post owner’s friends network to investigate how many active nodes appear at the end. The results received from IC model and the real diffusion results have a large gap. However, the gap is less in the case of entertainment post, but it is difficult to compare with other categories since the entertainment posts in the most cases have less reposts (in average 10 times less than advertisement or social appeal posts).

However, if the owner had published several posts then all of them should have different propagation results in real life, while Independent Cascade model gives the same results for each post. Thus we can conclude that in a real network, the topic of the post is as important as the number of friends and relation strength, therefore obtaining such background knowledge is necessary to better understand the information diffusion process.

According to the results, we may assume that the number of friends (size of the friends network) is not a propagation leverage. For advertisement category, the number of activated nodes in real diffusion is much higher than the estimated ones. Since the IC model estimates based on the probability given to each edge, which is $1/\bar{d}$, the estimated number of active nodes are proportional to the probability. However, as shown in Figure 1, in real scenario almost all of the neighbors of the source node got activated. In this specific example, the influence of the advertisement was much higher than expected and the diffusion behavior was different from what we expected.

In the case of social appeal category, the number of active nodes IC model estimated is much higher than the true number of active nodes. Since we observed that the posts in this category tends to diffuse deeper than other categories, the difference in the result is due to the fact that this types of posts diffuse deeper but not wider. If the diffusion process goes deep but not wide, the total number of active nodes becomes small. However, since the model doesn’t consider types of the posts, it estimates based on only the edge probability. If the average degree of nodes in the network is high, the estimated number of active nodes becomes high even when the diffusion happens vertically, not horizontally.

The third case is entertainment posts which gives the closest estimation results with IC model. Since entertainment posts are the most popular types of posts in social networks, the behavior of this type of posts can best describe average
diffusion process and it is observed that these posts do not have biased behavior patterns such as social appeal posts.

VII. CONCLUSION AND FUTURE WORK

We considered the problem of analyzing information diffusion processes in a social network using Independent Cascade model. We extracted Vkontakte social network dataset for post propagations, community members, and user friends network. We assumed that information propagation is closely related to the topics and post types, thus we defined several categories for communities and posts types. We found that there is a relation between network size and depth and also between posts and community types. We further applied the IC model to the post owner’s friends network and investigated how the results differ from a real information diffusion. We set the same initial active nodes as in the real post diffusion and define the probability of diffusion. We observed that Independent Cascade model does not properly predict information diffusion in Vkontakte and we suspect that the relations (tie strength) played a crucial role in the information propagation. Therefore, we plan to apply Linear Threshold model to our data to examine if the diffusion in Vkontakte follows the model and users are dependent from other users opinion. In addition, we plan to consider geolocations of the posts to understand how mobility of users can affect the information diffusion process. This approach can also reveal the physical extent of information diffusion considering real locations of active users in a network.

REFERENCES


