Diffusion Patterns of Social Network Posts

Alexander Gubanov
Yuliya Mundrievskaya

Center of Applied Big Data Analysis
Tomsk State University
Email: derzhiarbuz@yandex.ru
muo@data.tsu.ru

Ida M. Pu

Department of Computing
Goldsmiths, University of London, UK
Email: idapuone@gmail.com

***Abstract*—Social network posts as an efficient means of communication directly reflect users' interests and engagements. Despite challenges there are strong interests in understanding how social network posts efficiently spread information. In this article some diffusion patterns of social network posts are explored. The information spreading via post chains based on partial data of popular social network is studied to gain insights of the problem of information diffusion. Mathematical models for information cascades are proposed and future research directions are discussed.**

Keywords–information diffusion; posts; reposts; walls; social networks; social media; probability; mathematical modelling

I. INTRODUCTION

Social networks are real world systems where people (referred to as *actors*) interact with each other. Actors are represented as nodes of the network, and their interactions are represented as ties between them. Social networks appear in forms of social networking websites (*online platforms*) maintained by serving institutions (*services*), such as Facebook, Twitter, Telegram, etc. In this paper, we focus on diffusion issues of posts of messages based on the data from the on-line social network "Vkontakte".

The "Vkontakte" [1] is one of the largest social networks on the previous-soviet space with 80-100 millions visitors daily. Nodes of the network represent users (individuals or groups), and ties can be a mutual (undirected) link, such as friendship or/and directed link, such as subscribing. Users exchange information by means of private *messages* or/and public *posts*. Posts are publications in texts or/and multimedia (images, sound and videos) on webpages. Recurrent posts are referred to as *reposts* of the original post, and the dedicated display areas for posts are referred to as *walls*. Figure 1 shows a screenshot as an example of the wall and users' posts on the Vkontakte site.

Posts on the wall are also queued in a *newsfeed* (*poster*) and become visible on devices of subscribed users (users for short hereafter). Any post can be reposted by other users, and appears on their walls and in newsfeeds of subscribers (as friends or/and followers). With such iterative processes, multiple chains of posts are formed and information is spread like epidemic.

As a social media website, users can also interact with each other via the platform. They can, for example, leave comments on posts, vote for the favourite, exchange dialogues, etc. At the backend of the website, the notions of walls, posts and newsfeeds are evolved as self-sufficient software agents

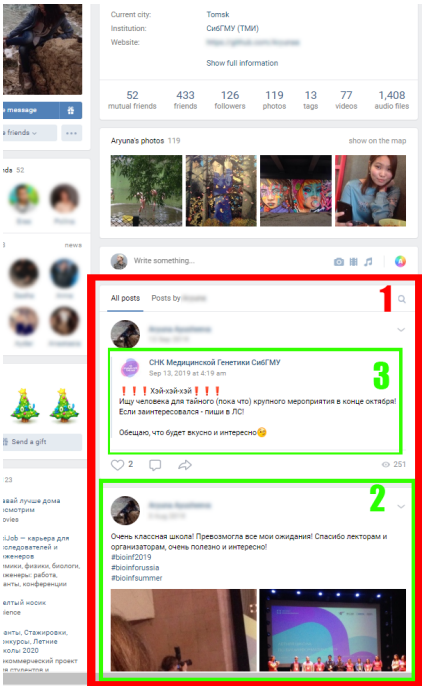

Figure 1. An example of the wall and posts (1:wall, 2:post, and 3:repost)

empowered by technologies allowing communications across different platforms.

Posts, as an efficient means of communication, reflect users' interests and engagements directly. Few businesses do not want to understand them. Despite challenges there are strong interests in understanding how posts spread information. In this article, we model the spread of the posts based on the data of this social network in order to help eventually answer interesting questions, such as the followings:

1) Why do people spread a piece of information? Is it because they like a given piece of news or is it because they just intend to be similar to their social surroundings?
2) How can we classify a piece of information by its spreading behaviour? Can we decouple properties of information from the properties of network?
3) How can we classify people by their contribution in information spreading? For example, how can we find someone whose role in information spreading is significantly different from the average (e.g., opinion leaders, non-conformists, information brokers, etc.)?

 ISBN: 978-1-61208-783-2

To answer these questions, different approaches are considered. The use of network information can vary in existing work. Some approaches discard information about network structure ([2], [3]), some assume that links are unobserved or irrelevant and should be reconstructed ([4], [5]), or is assumed to be correct and fully observable ([6], [7]). Influence of social conformism (or social pressure), is firstly covered by threshold models [8] then accounted in several models [7]. The modelling techniques are also vary: there are machine learning approaches, such as [3], statistical [9], probabilistic ([7], [4]), and game theoretic models [10], etc. Also the roles of different nodes in information spreading (especially opinion leaders) are often studied by analysis of network structure and topic content analysis ([11]–[12]).

Despite efforts, we could not find an approach that can be applied directly to solve our research problems which tend to be more localised in nature, practical and richer in social contexts. Most of the known approaches are useful to predict wide information epidemics on explicit, simple and large networks, such as Twitter, but not readily for Vkontakte. In case of local information (for example, the significant events in a city of average size) the social network can be small (about a hundred of reposts for one post), the number of data can be insufficient and the observed ties in the network can be incomplete or not perfectly relevant.

Known statistical and mean-field approximation approaches, commonly used for prediction of epidemic outbreaks, are not useful here, because the specific of local information can be subjective and exclusive, e.g. interesting only a certain group of people. The assumption that N (number of nodes) is infinite is not applicable for a local cluster. The common assumption of an average infection rate for everyone in the network and the impact of the node depending only on its network characteristics (degree, centrality, etc.) is also wrong in practical settings.

Social contexts are over simplified in models. For example, high degree nodes do not necessarily mean richer information sources. Nodes (users) can be faked by bots and ties (friendships) can be commercialised, biased, or true friendships can be hidden from newsfeeds. Results handled well in some visual approaches may not necessarily be easily obtained in other probabilistic models.

Hence, it would be inappropriate to use a model assuming that the network is known and relevant (IC, SIR and so on). On the other hand, we cannot discard the network information since the amount of data we possess is relatively small. For small datasets, machine learning approaches are not very useful.

Finally, the real local cascade usually has a group of completely isolated nodes (i.e., that are disconnected from any other nodes) in the cascade, and we should assume that they got the information through unobservable ties. Of course, we cannot discard them, neither. As we can see, after all, our small objects are not necessarily simple.

Our goal is to develop a sensitive tool for working with such type of information, that could give us insights to answer the interesting questions above.

The rest of the paper is arranged as follows. Section II briefly describes a model of information cascades. Section III explores the patterns of information spreading using visualisation techniques. Section IV proposes a model of information diffusion. Section V provides the summary and directions of further development.

II. INFORMATION CASCADE

Information cascade is a phenomenon in which a number of actors make same decisions in a sequential fashion. It is a two-step process in which a Yes-No decision is required by each actor whose decision can be influenced by others. Information cascades occur when the external information from previous participants overrides one's own private judgement.

In our model we assume the followings:

1) Each actor decides what to repost and whether or not to repost.
2) The limited action space is (repost, notrepost).
3) The actions of reposts are sequential in chronological order.
4) Each actor observes the reposts so far.
5) An actor cannot directly observe what other actors are in favour of but (s)he can infer that they like the posts enough if they repost it.

Our modelling is based on two parts: the underlying network of relations (the network) and information cascade (cascade) on this network.

A. Network

The nodes in our network model are identifiers of users and/or groups of users. The links/edges between nodes represent connections/relationships in social network, undirected for "friendship" and directed for "subscribing", corresponding to the model of undirected graph and digraph respectively.

It is impossible to model the entire social network in the real world because of its huge size and data availability. Hence, for a specific post we restrict the network by nodes that have made a repost or liked the post and their friends. Such a network is referred to as the *underlying network* (*network* for short) for specific information cascade.

To build the network we use the Vkontakte public Application Program Interface (API). Due to the privacy policy of Vkontakte that allows users to hide their information, about 30 percents of nodes are actually "hidden". It means that we do not know whether or not they have made a post. Also we do not know all incident links of them (but some links can be reconstructed from other nodes).

The network is modelled as a graph $G = \{U, E\}$, where U is a set of user node identifications (u_i) and E is a set of links (u_i, u_j) between a pair of nodes, where $i = 1, \cdots, N$ and $j = 1, \cdots, N$, and N is the number of nodes in the network. To distinguish a *user* and *group*, u_i can be positive or negative. The positive u_i is used to represent an user, and the negative u_i represents a group of users.

B. Cascade

The information cascade is a sequence of moments when post appeared on user's walls. It is defined as a sequence of the pairs $\{(u_i, t_i)\}_{1 \le i \le M}$, where u_i denotes a node from the network and t_i is a timestamp of the moment when repost (or original post) on the wall of this node is appeared. M is the

total number of nodes that are "infected" by the information, i.e., that having the post on their wall. This data is also collected using the public Vkontakte API.

It is Vkontakte's policy, however, to forbid general public viewers from accessing any repost chain. Hence, we have no explicit information about information spreading path.

C. Spreading issues

Processes of information spreading are widely studied often using a Twitter social network as the source of experimental data. For every retweet (analog of repost in Vkontakte) it is known exactly, who retweeted whose tweets, so it is possible to build an explicit graph of information propagation. This graph is always a tree. Analysis results of such graphs are used to identify spreading parameters (for example, the intensity of "infection") and important nodes (opinion leaders). But this "exactness" also conceals one significant trait of how people spread the information. Sometimes people choose to manifest something due to its respectable source. Their decision of a repost (or retweet) may not necessarily be purely driven by their inner motives, but also by the apparent positive responses, such as the big number of surroundings who translate the same piece of information. It is a common behaviour involving opinions, social norms, and trusts, etc. [13].

We assume that the fact of repost (retweet) from some node does not mean that it is a merit of only this node. It can be the cumulative contribution of all nodes made the post which was seen by the reposting user before. It can not be substituted by defining individual infection rates or probabilities as ([14], [4]). However, some models [7] takes it into account. We do it also.

III. Pattern of Spreading

The first step in analysing the spreading of posts is visualisation of the pattern of spreading. This visual technique is useful for practice, also the patterns can be analysed using structural network analysis (measuring centrality, modularity, etc) which is widely covered by appropriate software (for example Gephi). As long as the pattern of spreading is directed aperiodic graph, some bibliometric techniques can also be used, for example main path analysis [Batagelj2014].

The pattern is a way to display the combination of information cascade and underlying network, so the researcher could see key properties of both to be able to make conclusions. It is acyclic oriented graph $G_p = \{U_p, E_p, D_p\}$ with weighted nodes where $U_p \subset U$ is a set of infected nodes (i.e., having a post or repost on their walls). There are M nodes in U_p. $E_p = \{(u_i, u_j) : (u_j, u_i) \in E : u_i \in U_p, u_j \in U_p, t_i < t_j\}$ is a set of ordered pairs of nodes from U_p, representing directed edges such that there is a link between u_j and u_i (in general it means u_j is a friend of or subscribed on u_i) in underlying network and u_i made the repost earlier than u_j. So the edge represents a potential act of information propagation, because u_j is able to see u_i's post in the newsfeed before it decides to make reposts. There is also $D_p = \{d_{u_i}\}_{1 \le i \le M}$, a set of weights for nodes, where d_{u_i} is the degree of u_i in underlying network. The example can be seen in Figure 2 and Figure 3.

Figure 4 shows the pattern of spreading of real posts. The size of the node is defined by its degree in underlying network (i.e., the number of neighbours the node could influence). The colour becomes darker as the number of outgoing edges on patterns increases (the darker it is in red, the more neighbours are actually influenced). Note that in this case we use only mutual links for users (Vkontakte is more friendship-style than subscriber-style network and the most of links between users are mutual/undirected). Hence, there is no need to define different in- and out- degrees in the underlying network.

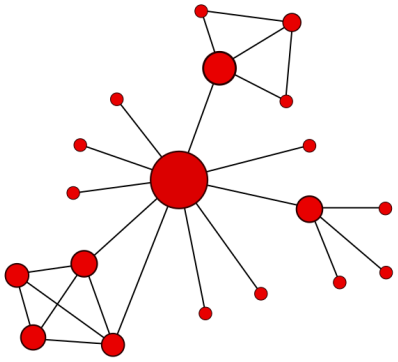

Figure 2. Underlying network

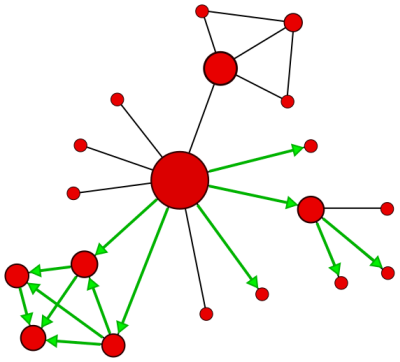

Figure 3. Green is a pattern of spreading

In the figure, we can see the separate cluster of users (1), the nodes that seems to be information brokers or influencer in their cluster (like 2, which has not very high degree, but probably affected a lot of neighbours), the nodes that have no influence at all (3, it has high degree, but infects nobody). Also there is a group of isolated nodes (4) that means that they got the information through inobservable path (private messages or hidden users).

Although the patterns of spreading help researchers make hypotheses and explore intuitive solutions, mathematical models are necessary for rigorous analysis and prediction.

IV. Models of the Information Diffusion

On-line social networks are set up mainly for information diffusion. A number of widely used dynamic models are known as *infection models* (SI, SIR, SIS) [15] of social contagion. They are considered as good for describing diffusion of certain types of information, such as hot news, memos, rumours and in other situations when people become infected regardless his will.

Another type of information spreading model, known as the *threshold model*, describing a situation when each node

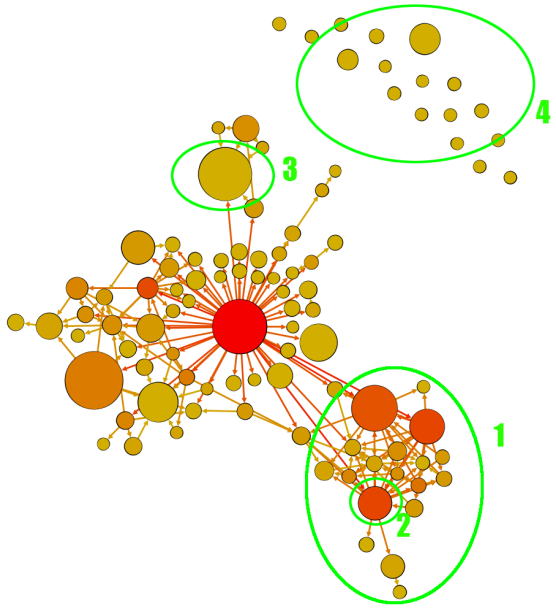

Figure 4. Pattern of spreading

makes a decision to spread the information based on social reinforcement (e.g. social norms or opinions) and susceptibility to information strongly depends on how many neighbours are infected. ([13], [8]).

We build our model as a development of probabilistic SI model for networks. For each edge between an infected and susceptible node, there is a probability of passing the infection during each small time interval Δt that is proportional to $\Delta t + o(\Delta t)$. Our goal is to define this probability.

The SI model is continuous and the time delay in the edge matters. It is interesting that there is a discussion about how important it is to take this delay into account. Some authors [3] insist that in real world social networks this delay is too noisy and it is better to discard it, while other authors [16] think that it is a very important measure and should be taken into consideration.

In our model we took into account the minimum amount of properties of the process that are crucial for its description. On one hand, we consider every data available and try to use the information about known connections between users also. On the other, we respect the fact that the information is not necessarily completely relevant. Some connections are hidden or inobservable.

A. Model parameters

The model has four parameters: namely, observed infection rate θ, *conformism* level (threshold) κ, decay (or obsolence, a kind of recovery rate analogue) δ and unobserved infection rate ρ.

The first is the infection rate θ. In classical network SI model the probability of susceptible $j'th$ node to be infected during small time interval Δt is equal to

$$n_j(t)\theta\Delta t + o(\Delta t) \tag{1}$$

where $n_j(t)$ is a number of infected neighbours at the moment t.

The second is *conformism* κ defining a threshold: a fraction of node's neighbours that should be infected to significantly increase the chance of infection of the node. It is a modifier for infection rate depending on the number of neighbours infected and a total number of neighbours. Thus, the probability of infection during small time interval is

$$n_j(t)\tau_\kappa(j,t)\theta\Delta t + o(\Delta t) \tag{2}$$

where $\tau_\kappa(j,t)$ is threshold function, for example

$$\tau_\kappa(j,t) = \begin{cases} 1, \frac{n_j(t)}{N_j} \geqslant \kappa \\ \epsilon, \frac{n_j(t)}{N_j} < \kappa, \end{cases} \tag{3}$$

where N_j is total number of neighbours that node j has and ϵ is small.

Note $n_j(t)$ and $\tau_\kappa(j,t)$ depend on t and will not write this dependence below. Here can also be used different function which ascends or descends being regulated by κ.

The third is a *background* parameter ρ, which defines the intensity of contagion through unobservable channels. To handle this infection process we take a classical non-network ST/SIR model's assumptions that each node connects to all other nodes. The assumption is that the number of all information sources (observed and unobserved) is proportional to the number of observed sources. The probability of infection should then be

$$\left(\rho N_I(t) + n_j\tau_\kappa(j)\theta\right)\Delta t + o(\Delta t), \tag{4}$$

where $N_I{}'(t)$ is an overall intensity of contagion ($N_I(t)$ is a number of nodes infected at time t).

The last parameter is a *decay* δ. As the older posts have less chance to be seen in one's newsfeed, the infection rate for each infected node should decrease over time. There can be different types of decay, for the exponential one the probability of contagion during Δt is

$$\left(\sum_{i\in A(t)} \rho e^{-\delta(t-t_i)} + \tau_\kappa(j)\sum_{i\in A_j(t)} \theta e^{-\delta(t-t_i)}\right)\Delta t + o(\Delta t), \tag{5}$$

where $A(t)$ is a set of all infected nodes at time t, $A_j(t)$ is a set of infected neighbours of node j at time t and t_i is the moment at which node i was infected.

Note, that using step decay function (i.e., that equal to 1 until some moment, and 0 after) gives a classic SIR model recover behaviour. In this case setting θ to 0 gives an infection equation for non-network SIR model, and setting ρ and κ to 0 gives an infection equation for network SIR model. However, if step recover function is good for describing biological infections, it is unsuitable for information, so gradual exponential decay is more preferable. Note that if theta, rho and kappa characterise the post, delta is a property of the social network. Thus, we can assume that it is the same for every publication. This assumption is helpful when we start to

 ISBN: 978-1-61208-783-2

estimate parameters not just for one information cascade, but for the set of cascades.

As long as our model is a kind of SI, not SIR (as there is no recovery, the post stays on the wall forever, or if it was removed, we have no information about it) the equation (5) defines the model behaviour. The initial condition is a set $A(0)$ of nodes that was infected at the time $t = 0$.

B. Detecting significant nodes

In general every node has its own parameters θ (influence) and κ (conformism). It would be very useful to estimate all of them. Indeed, we cannot identify all individual θ's and κ's for each node because of small amount of data.

The common practical question is "What the most influential node is?". Or "What the less conform node is?". Thus, we should find one or several nodes whose parameters are differ the most from the average. We propose to use greedy algorithms. The algorithms search for the node for which increased θ makes the model better (higher likelihood). Then the second one, etc. The same is for decreased θ's and κ's.

Using such kind of approach should help us to find most "distinctive" nodes without overestimating the model.

V. Conclusion

In this paper we defined an object of study, obtained the data and formulated research questions. We considered several existing approaches and found that they cannot be applied directly to our cases. Patterns of spreading visualisation gave us intuitions on what is the object of research looks like and allows to make some decisions and hypothesis for practical use. The four parameter model of information spreading provides opportunities to answer several questions about an essence of information. The different roles of the nodes, however, still need to be considered. The next step of our studies is to estimate model parameters and then to develop a procedure to detect nodes whose behaviours significantly differs from average.

Acknowledgment

We would like to thank Jacqueline W. Daykin for helpful discussion, the anonymous reviewers for valuable comments, and financial supports for the research, The first author would also like to thank the overseas scholarship that allows his research visit to Goldsmiths, University of London.

References

[1] On-line social network "Vkontakte". [Online]. Available: https://vk.com

[2] A. Najar, L. Denoyer, and P. Gallinari, "Predicting information diffusion on social networks with partial knowledge," in Proceedings of the 21st International Conference on World Wide Web, ser. WWW'12 Companion. New York, NY, USA: Association for Computing Machinery, 2012, pp. 1197–1204. [Online]. Available: https://doi.org/10.1145/2187980.2188261

[3] S. Bourigault, S. Lamprier, and P. Gallinari, "Representation learning for information diffusion through social networks: An embedded cascade model," in Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, ser. WSDM'16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 573–582. [Online]. Available: https://doi.org/10.1145/2835776.2835817

[4] M. Gomez-Rodriguez, L. Song, H. Daneshmand, and B. Schölkopf, "Estimating diffusion networks: Recovery conditions, sample complexity & soft-thresholding algorithm," The Journal of Machine Learning Research, vol. 17, no. 90, 1 2016, pp. 1–29.

[5] M. Rodriguez, D. Balduzzi, and B. Schölkopf, "Uncovering the temporal dynamics of diffusion networks," vol. abs/1105.0697, 05 2011.

[6] M. Kimura and K. Saito, "Tractable models for information diffusion in social networks," in LNCS, 09 2006, pp. 259–271.

[7] C. Lagnier, L. Denoyer, E. Gaussier, and P. Gallinari, "Predicting information diffusion in social networks using content and user's profiles," in Advances in Information Retrieval, P. Serdyukov, P. Braslavski, S. O. Kuznetsov, J. Kamps, S. Rüger, E. Agichtein, I. Segalovich, and E. Yilmaz, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 74–85.

[8] M. Keuschnigg, Granovetter (1978): Threshold Models of Collective Behavior, 01 2019, pp. 239–242.

[9] V. Krishnamurthy, S. Bhatt, and T. Pedersen, "Tracking infection diffusion in social networks: Filtering algorithms and threshold bounds," IEEE Transactions on Signal and Information Processing over Networks, vol. 3, no. 2, June 2017, pp. 298–315.

[10] C. Jiang, Y. Chen, and K. J. R. Liu, "Evolutionary dynamics of information diffusion over social networks," IEEE Transactions on Signal Processing, vol. 62, no. 17, Sep. 2014, pp. 4573–4586.

[11] A. Farzindar and W. Khreich, "A survey of techniques for event detection in twitter," Computational Intelligence, vol. 31, 09 2013.

[12] Q. Li, A. Nourbakhsh, S. Shah, and X. Liu, "Real-time novel event detection from social media," 04 2017, pp. 1129–1139.

[13] C. Kadushin, Understanding social networks: Theories, concepts, and findings. Oxford: Oxford University Press, 2012.

[14] K. Saito, M. Kimura, K. Ohara, and H. Motoda, "Learning continuous-time information diffusion model for social behavioral data analysis," in Advances in Machine Learning, Z.-H. Zhou and T. Washio, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 322–337.

[15] M. E. J. Newman, Networks. Oxford: Oxford University Press, 2010.

[16] A. Guille and H. Hacid, "A predictive model for the temporal dynamics of information diffusion in online social networks," in Proceedings of the 21st International Conference on World Wide Web, ser. WWW'12 Companion. New York, NY, USA: Association for Computing Machinery, 2012, pp. 1145–1152. [Online]. Available: https://doi.org/10.1145/2187980.2188254

 ISBN: 978-1-61208-783-2