Information diffusion in social networks

Vaganov Danila[a], Kutuzova Tatiana[b], Abubakirov Azat[c], Pankov Vasiliy[d]

[a] ITMO University, M4117

[b] ITMO University, M4117

[c] ITMO University, M4117

[d] ITMO University, M4117

R E P O R T I N F O

Key words:
Information diffusion
Stochastic model
Social networks
Random recursive tree
Barabasi-Albert
Complex networks
Scale-free networks.

A B S T R A C T

This paper is based on two main approaches: modeling of the online social network structure, that implemented by using modified Barabasi-Albert model, and modeling of information diffusion, which based on Random Recursive Tree. For developing and calibrating these models the dataset was crawled from a social media community about charity. Current results contain of processed and analyzed dataset, framework for calibrating and modeling cascades of information spreading and complete model for growing networks. The future work involves the extension of the RRT model and integration of these models for combining the network growth and the dissemination of information in it.

1. INTRODUCTION

Online social networks allow hundreds of millions of Internet users worldwide to produce and consume content. They provide access to a very vast source of information on an unprecedented scale. Online social networks play a major role in the diffusion of information by increasing the spreading of new information [1].

Scale-free networks with a degree distribution following a power law have been the focus of a great deal of attention in the literature. This type of network characterizes the degree distributions of many man-made and naturally-occurring networks.

Understanding the features and dynamics of information diffusion in social networks is crucial for businesses to promote products, but also for governments to predict and even regulate public opinion.

Barabasi and Albert [2] have given the first explanation of the scale-free distribution by reformulating Simon's model [3] in the context of growing networks.

For modeling information diffusion in online social networks there are a lot of approaches: Machine learning techniques have been applied to, for example, predict content popularity based on the previous popularity and the features related to the contents and users that shared the content [4], [5].

Stochastic models, such as cellular automata [6], Threshold models [7]–[9], Susceptible Infected Recovered (SIR) [10]–[12], and Linear Influence [13] have been studied to understand how the dynamics of information diffusion such as the spreading rate and the social network topology could influence a key feature of the diffusion process such as the popularity.

Based on information of considered community about charity in social network Vkontakte, we suppose, that there are a lot of active spreading information, that propagates deeply in relations of the users. Thus, for modeling information diffusion the Random Recursive Tree model will be considered [14]. For the modeling growing of the network, the model [15] will be implemented and modified.

The main goal of this paper is implement, modify and calibrate models, which mentioned above, according to the crawled data.

The rest of this paper is organized as follows: firstly, the data analysis, processing and network analysis will be described, then there is an introducing and evaluating of the considered models, and, finally,

the conclusion, future work and evaluating the contributions of all team members will be presented.

2. MODELS

This section contains the dataset and model descriptions.

2.1 Data description

The dataset used for the research contained data about charity-based social media community, that helps people with rare diseases gather money.

A typical post in this community contains a fundraising request and a link to the community dedicated to the person money is being raised for.

Vkontake has a very rigorous process of verification for charity (in order to become verified, the community must be approved by Vkontakte, the administering of the community is judged by VK representatives, and administrators also must have all kinds of official documents for their projects that describe their charity funds and what money is spent on; verified status for charity communities lasts for a year, after that they must undergo re-verification), many communities don't go through it, and the aforementioned community was created to mitigate this issue - administrators of the community check charity-related documents for fraud before posting a link to the fundraiser.

During data collection, first all posts from core community were gathered (each post is characterized with a chain of reposts, as shown in Fig. 1); then reactions on the posts were collected.

Community has a chain of posts with likes and comments, each with a chain of reposts; list of administrators and users; discussions (each discussion is a sequence of messages, probably with likes; some of messages can be addressed to a particular user).

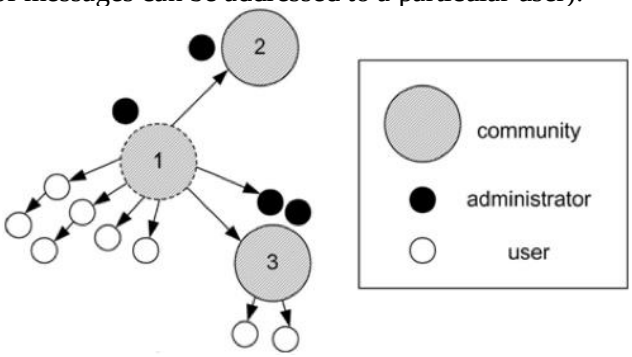
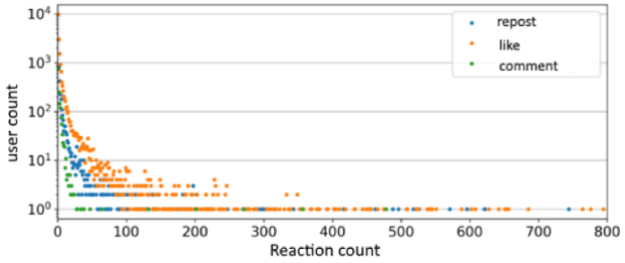


Figure 1 - A scheme of a post in core community 1

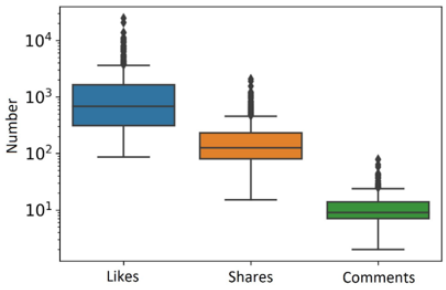
During data crawling, the information about posts, reposts and comments in the community during the interval of 508 days (2016-06-27 - 2017-11-13) had been gathered. 805 posts had been posted in the community during that interval.

During the aforementioned interval, the users have reposted 57 86, liked 564 971 and commented 3079 times. On average a post in the community has 71,9 reposts, 701,8 likes and 3,8 comments.

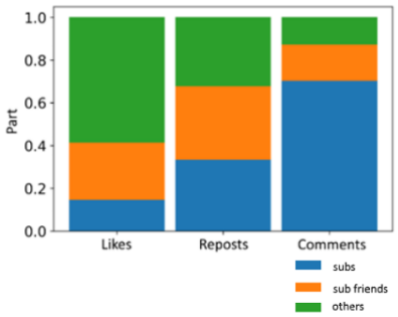
The distributions of these reactions are presented on Fig. 2.



a)



b)



c)

Figure 2 - User reactions

As we can see from these figures, most of the users don't react more than once and most of activity comes from non-subscribed users.

One of the main obstacles to our research were that the chain of information sharing in the dataset lacked depth (there were 3 levels at most) and that there was no information about the followers of non-subscribers. To address this, the scope of the research had been narrowed and additional data had been collected.

2.2 Network description

The target community consists of 294 345 followers and 33 478 368 friends of the followers.

The network of subscribers and their friends contains 80 million edges with 33 million nodes. Figure 3 represents the visualization of relationship network, which contains only strongly connected nodes (about 10% of all) inside the community. Colors corresponds to the classes of modularity, which divides the community into clusters according to mutual friends of the users. The size of the nodes

relates to the betweenness centrality (BC), i.e. than larger a value of BC, than larger the considered node. The degree distributions of the nodes on the layer of subscribers and on the 2nd layer will be considered at the p. 2.3.

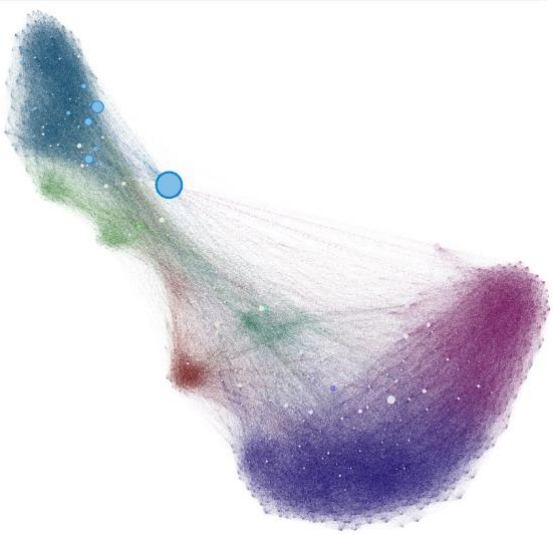


Figure 3 - Visualization of considered community

2.3 Network models

The distinguishing feature of scale-free network is power law degree distribution. The degree of a node in a network is the number of connections it has to other nodes. Degree distribution is the probability distribution of these degrees over the whole network [16] and can be presented as

$$f(d) = d^{-\gamma} \quad (1)$$

where d is a degree and γ is power law exponent which is described by the following formula

$$\gamma = 1 + N\left(\sum_i ln\frac{d_i}{d_{min}}\right)^{-1} \quad (2)$$

where N is number of nodes in network, d_i - degree of node i, d_{min} is a minimal value of degree in network.

2.3.1 Barabási–Albert model

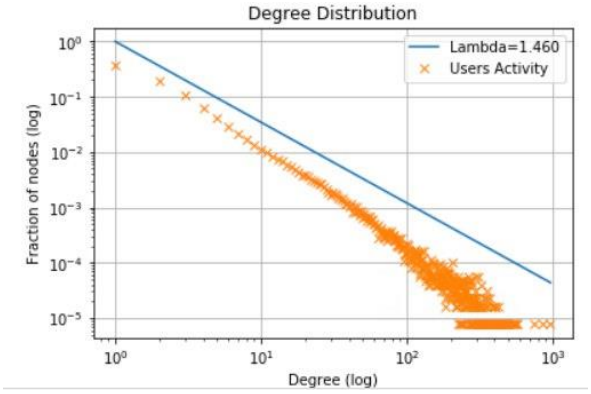
Scale-free network can be generated by Barabási–Albert model. This model produces a scale-free network with power law degree distribution [17].

The network develops following two steps:

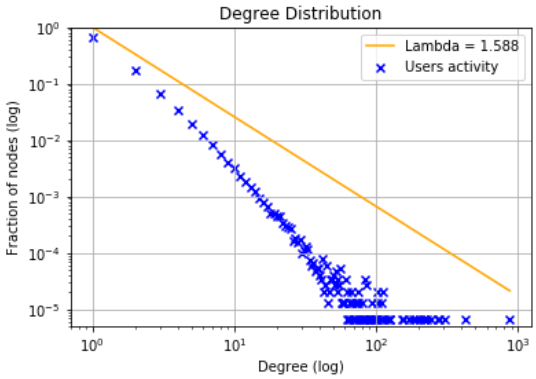
1.at each timestep we add a new node with m links that connect the new node to m nodes already in the network;

2.the probability $P(d)$ that a link of the new node connects to node i depends on the degree d_i as

$$P(d_i) = \frac{d_i}{\sum_j d_j} \quad (3)$$



a)



b)

Figure 4 – Degree distribution of network of community's subscribers: a) real data b) modelling result

Figure 4 presented that this model fits to the real network inside the community, because the degree distribution of real data has power law distribution. But if generating of extended (2nd level) network is necessary, advanced approach need to be used.

2.3.2 Advanced Model

More complex model based on BA model, but has some improvements. Model from paper[15] consists of three probabilities of growing network, which displayed in Figure 5(a, b, c), but there are not considered decaying network behavior of social networks. Thus, a new probability of removing node with lowest degree was introduced (Figure 5 – d).

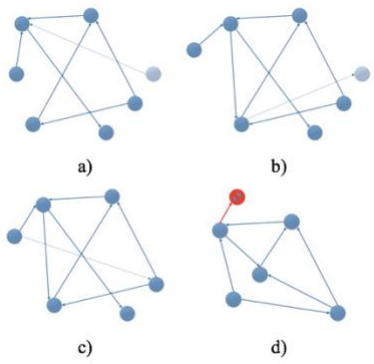


Figure 5 – Developing of network process: a) creation of node attaching an old node, b) creation of

node attached by a new link, c) creation of link between old nodes, d) removing of node and its links

At each timestep:

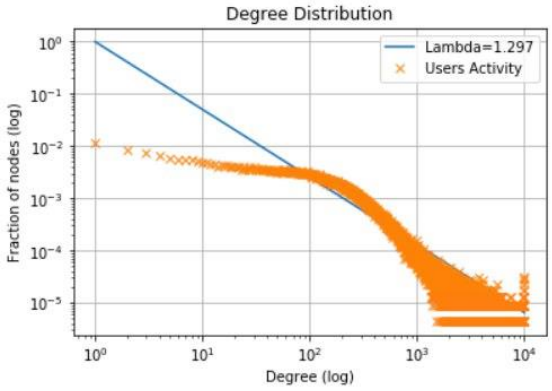
· with probability p, a new node is created attaching to a directed node;

· with probability q, a new node is created attached by a direct link;

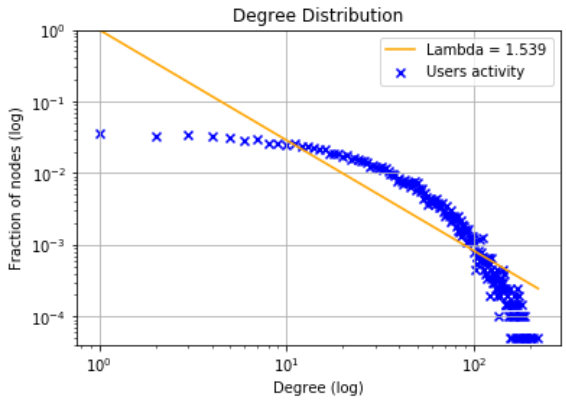
· with probability r, directed link is created between the old nodes;

· with probability d, an old node is removed.

The sum of first three probabilities must be equal 1. Probabilities p and q are user-defined, hence $r = 1 - r - q$. Proceeding from the fact that extracting of user occur much less often than add users, probability of removing of node is far less than values of other probabilities.



a)



b)

Figure 6 –Degree distribution of network of community's subscribers and their friends: a) real data b) modelling result

From 6 evidently that enhancement model is more useful than Barabási–Albert. This is due to the fact that this model is examine more fields of users behavioral in social media.

2.4 RRT model

The spreading information structure in social networks can be represented as a cascade tree. A cascade root is an author of new post (information). Edges between cascade nodes are shares of the post between community of the social network. The scheme of sharing activity is shown in the figure 7.

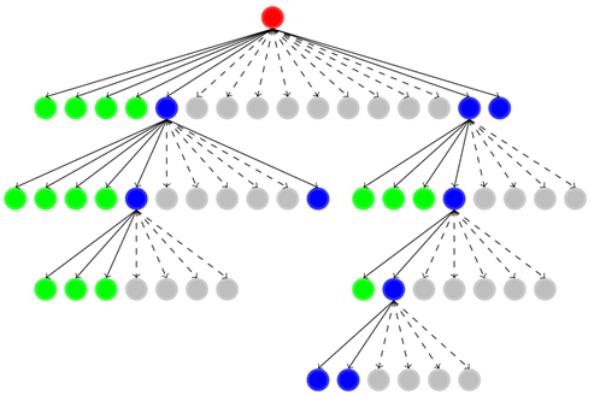


Figure 7 — Scheme of the information spreading in the social network. Red node represents an author of the information. Green nodes represent users that set "like" to the post. Blue nodes represent users that shared this node with their subscribers. Grey nodes represent users who just ignored the information.

A cascade tree has two fundamental properties that characterize the topology of the information spreading: average path length between nodes (5) and degree variance (6).

$$E[H] = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} H_{ij} \quad (5)$$

H_{ij} is the shortest path between node i and node *j*, *N* is a total number of nodes

$$Var[D] = \frac{\sum_{i=1}^{N} (d_i - E[D])^2}{N} \quad (6)$$

d_i is a degree of node i.

These two characteristics of the cascade were considered as metrics for evaluating the quality of the model, i.e. the model should capture these two key features. To model the spreading of the information Random Recursive Tree model [14] was considered and implemented. The idea of the model is pretty simple. The tree growing starts (t = 0) with the root that represents the author of the sharing content. In the step, new node connects to an existing node that is selected by the following rule: each existing node has an opportunity to be selected with the probability $\frac{d_i^{\theta}(t)}{\sum_{i=1}^{t} d_i^{\theta}(t)}$, $d_i(t)$ is a node *i* degree at time step *t*. Thus, the probability of the connection new node to node *i* is proportional to the degree of node *i* of the power of $\theta \in [0, +\infty]$ at step time *t*. Generation of the cascade tree continues until the tree size exceeds N. The parameter θ shows how the node degree influences on the choice of the node.

The considering dataset contains 659 cascade trees that have sizes between 10 and 50 nodes. During the experiments 100 cascade trees per size N were generated for different values of θ.

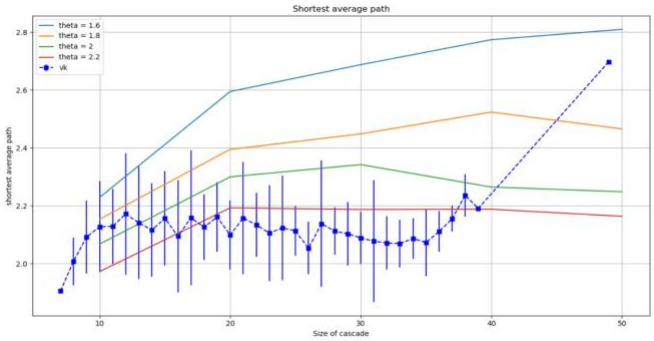


Figure 8 - The average path length in the cascade trees in VK dataset and the RRT models

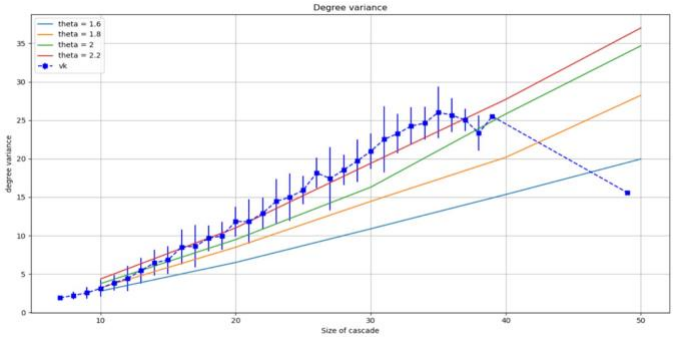


Figure 9 - Degree variance of the cascade trees in VK dataset and the RRT models

The obtained results can be interpreted as the following: Random Recursive Tree model successfully builds the cascades in terms of average path length and degree variance if the cascade size does not exceed 40 nodes. Modelling the cascades that have a big number of nodes is difficult because of a low quality of the dataset. It can be assumed that if the dataset has more information about the big size cascades, the approach based on RRT can be applied to model cascades with any sizes.

3. CONCLUSION

3.1 Results

The modified model for growing scale-free network, which successfully fitted to real distribution of relations between subscribers of considered community and their friends, was implemented. The tendencies of modelled and real-existing distributions are quite comparable.

The current state of Random Recursive Tree model is can be used for modeling of the cascades with the small depth. As result of this paper there are completely implemented framework for calibrating and modeling of information diffusion in online social networks. According to the average path length, which represents the size of the cascade tree increases as the size of the cascade tree increases, but for big cascades the data is not representative. As same as average path length, the degree variance grows according to size of cascade, except that when the size is large and the degree variance tends to decrease, because of lack of collected data. If the dataset contains more cascades with the more sharing depth and count of sharing, RRT will be probably applicable to generation the different cases.

3.2 Future work

The future work must be focused on the crawling the high-level quality dataset, combining of the both of models and extension of the information diffusion model. The new dataset gives the opportunity to calibrate the model more accurately and more deeply. Despite the importance of increasing the quality of data, the integration of the network-growing model and RRT model are more important. Thus, the improving of the RRT model by introducing the new parameter the same as theta in p. 2.4, but for other category of users, which subscribed on the personal communities of the people, who was aggregated in considered community, in theory, must gain the more representative result.

3.3 Contributions

Danila (dataset collecting, project management) — 26%

Tatiana (network model, slides for presentations) — 26%

Azat (random recursive tree model, refactoring) — 24%

Vasiliy (dataset analysis, data preprocessing) — 24%

All details of this project you can find here: https://github.com/tatiasha/ICS-Project

REFERENCES

[1] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic, “The Role of Social Networks in Information Diffusion,” *Proc. 21st Int. Conf. World Wide Web SE - WWW ’12*, pp. 519–528, 2012.

[2] A. L. Barabási and R. Albert, “Emergence of scaling in random networks,” *Science (80-.).*, 1999.

[3] H. A. SIMON, “ON A CLASS OF SKEW DISTRIBUTION FUNCTIONS,” *Biometrika*, 1955.

[4] C. Richier, E. Altman, R. Elazouzi, T. Jimenez, G. Linares, and Y. Portilla, “Bio-inspired models for characterizing YouTube viewcout,” in *ASONAM 2014 - Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2014.

[5] J. Cheng, L. A. Adamic, J. Kleinberg, and J. Leskovec, “Do Cascades Recur?,” 2016.

[6] J. Goldenberg, B. Libai, and E. Muller, “Talk

of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth," *Mark. Lett.*, 2001.
[7] M. S. Granovetter, "Threshold Models of Collective Behavior," *Am. J. Sociol.*, 1978.
[8] Q. Li, L. A. Braunstein, H. Wang, J. Shao, H. E. Stanley, and S. Havlin, "Non-consensus Opinion Models on Complex Networks," *J. Stat. Phys.*, 2013.
[9] B. Qu, Q. Li, S. Havlin, H. E. Stanley, and H. Wang, "Nonconsensus opinion model on directed networks," *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.*, 2014.
[10] L. Feng, Y. Hu, B. Li, H. E. Stanley, S. Havlin, and L. A. Braunstein, "Competing for attention in social media under information overload conditions," *PLoS One*, 2015.
[11] R. Pastor-Satorras and A. Vespignani, "Epidemic spreading in scale-free networks.," *Phys Rev Lett*, 2001.
[12] M. E. J. Newman, "Spread of epidemic disease on networks," *Phys. Rev. E - Stat. Physics, Plasmas, Fluids, Relat. Interdiscip. Top.*, 2002.
[13] J. Yang and J. Leskovec, "Modeling information diffusion in implicit networks," in *Proceedings - IEEE International Conference on Data Mining, ICDM*, 2010.
[14] L. Liu, B. Qu, B. Chen, A. Hanjalic, and H. Wang, "Modeling of Information Diffusion on Social Networks with Applications to WeChat," pp. 1–17, 2017.
[15] S. Aparicio, J. Villazón-Terrazas, and G. Álvarez, "A model for scale-free networks: Application to twitter," *Entropy*, vol. 17, no. 8, pp. 5848–5867, 2015.
[16] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Stat. Mech. complex networks*, 2002.
[17] M. Pósfai *et al.*, "Chapter 5: The Barabási-Albert Model," *Netw. Sci.*, 2015.