# Information Propagation Strategies in Online Social Networks

Rodion Laptsuev, Marina Ananyeva, Dmitry Meinster, Ilia Karpov,
Ilya Makarov, and Leonid E. Zhukov

National Research University Higher School of Economics, Moscow, Russia

`rodion123683@gmail.com`, `ananyeva.me@gmail.com`, `dlmeynster@edu.hse.ru`,
`karpovilia@gmail.com`, `iamakarov@hse.ru`, `lzhukov@hse.ru`

**Abstract.** Online social networks play a major role in the spread of information on a very large scale. One of the major problems is to predict information propagation using social network interactions. The main purpose of this paper is to construct heuristic model of a weighted graph based on empirical data that can outperform the existing models. We suggest a new approach of constructing the model of information based on matching specific weights to a given network.

**Keywords:** Social Network Analysis, Information Propagation, Social Networks

## 1 Introduction

Online social networks is one of the most effective and fast tools used for the spread of information. These technologies enable individuals to share information simultaneously with an audience of any size on different topics of interest. For instance, in online social networks, this process can be implemented via reposts that are posts copying information from another post while preserving a link to the source. Profound knowledge of core principles of information propagation and the ways of its spread provide us with a lot of opportunities. For example, one might influence the target groups of minimal sufficient number of active actors depending on the purpose: propagation of rumors, political propaganda, placement of a new product on the market, viral advertisement, and many others.

A lot of efforts have already been made in understanding the principles of information propagation and their formalisation through mathematical models. The range of currently applied models mostly consists of linear threshold models, cascade models and mixed models [3, 4, 8, 14]. However, there are still many challenges on the way to improving existing models. In this study, we propose a weighted graph model with tuned parameters. The weights are modified in order to fit the empirical data.

The article is organised as follows. In the beginning (Section 2), we give a brief overview of the most frequently used strategies of modelling information propagation. Section 3 contains the main definitions and notions on the topic. In Section 4, we introduce the dataset used for our experiments. In Section

5, we describe two models for weighted graph and baseline model for further verification. In Section 6, we conclude the paper and discuss possible directions for future work.

## 2 Related Work

The problems of information propagation, social influence maximization, and its applications to online social networks were widely studied in the research papers [14,15] . One of the main prerequisites for spread of information via social networks is that user receives new information and takes one of two possible strategies: to extend the information further or not. The propagation can occur either through direct messages or through reposts. Since the direct conversations are treated as confidential information, we focused our attention on the reposts. We consider only direct reposts leaving indirect reposting to future research.

There are two probability models used by researchers: Independent Cascade (IC) model [8, 9] and Linear Threshold model (LT) model [11]. Cascade models were borrowed from particle physics [17]. They are used for simulating processes similar to activation of any node as a result of one-off independent attempts by already activate neighbouring nodes. Aside from physics, the models were also inspired by medicine. Systematic study on the adoption of medical diffusion models in human networks started with the research by Coleman et al. [5]. In [19], Morris described the contagion theory of spread of behaviours. The basic premise behind the theory of social contagion is that a node is driven to adopt a behaviour based on the behaviours of its neighbours, more precisely the fraction of the neighbours who have adopted the behaviour. The idea was further generalised through linear threshold models [13] incorporating different weights for neighbours and different individual thresholds.

Regarding IC model, Kimura and Saito [17] proposed several shortest-path based influence cascade models and provided efficient heuristic algorithms for computations of influence spread under their models. In [3], researchers modified the algorithm by adding degree discount heuristics for the uniform Independent Cascade model with the same probabilities for all edges of graphs. One more feature proposed by the authors is called maximum influence arborescence, which is a tree in a directed graph where all edges are either pointing toward the root or pointing away form the root [9]. All the mentioned above models use specific features of the IC model, therefore, they can not be applied directly to LT models.

The cascade and linear threshold classes of models show the good results of influence spread and active actors maximization in comparison to other suggested models [13, 14]. Nevertheless, one of the most important disadvantages of models described above are slow time of running and not scalable properties. Other algorithms do not provide consistently good performance on influence diffusion [17]. A generalized cascade model was shown to be equivalent to the generalized threshold model [13] while providing theoretical performance guarantees for several hill-climbing strategies for many general instances of the NP-hard models for optimal solutions.

The first researchers who considered the influence maximization within probabilistic approach as an algorithmic problem were Domingos and Richardson [7]. They used methods based on Markov random fields in order to model the final state of the node activation in the network directly. Driven by its application in viral marketing, a lot of recent efforts in diffusion processes have focused on finding the set of nodes that would maximize the spread of information in a network, also called a target set selection problem. Despite the problem of scalability of their greedy algorithms, Kempe et al. [13] studied the influence maximization as a discrete optimization problem. In [18], the lazy-forward optimization model was presented, which selects new seeds in order to reduce the number of influence propagation evaluations. However, it is also not scalable to graphs with thousands nodes and edges or larger size and computational time is not efficient enough.

Consequently, several extensions of the original cascade and threshold models have been developed. In [2], the authors collected available traces of photos spread in social network Flickr and tried to reveal the role of friendship in the diffusion process and the length of photo's spread. The results showed that exchanged information between friends was likely to account for over 50% with significant delays at each hop and the obstacles to spread each photo widely among users. In [21], it was found that the long ties in social networks prohibited the complex cascade. The authors of [20] found a coupling between interaction strengths and the networks local structure in a mobile communication network. It was shown that if the weak ties are gradually removed, then a phase transition takes place in the network. In addition, it is important to mention several studies with designed machine learning algorithms in order to extract parameters of influence cascade models from empirical data sets [10, 17]. We have also used this principle to generate graphs in our paper.

## 3 Definitions

In our research, the term 'information propagation' implies the the diffusion of information via reposts among the members of given community. We denote the online social network as a graph $G(V, E)$, where $V$ stands for the set of vertices, $E$ is the set of directed weighted edges. The node may be defined as a user or a group of user, but taking into account our empirical dataset, we define one vertex as a member (subscriber) of community. The spread of reposts on the graph we call a 'wave' in further.

We define Information Cascade or Information Diffusion as a behaviour of information adoption by people in a social network resulting from the fact that people ignore their own information signals and make decisions from inferences based on peoples previous actions.

# 4 Dataset

We used Vkontakte as the source for uploading the dataset. We chose one political community ('United Russia') as an example for examining the way the information is spread via reposts among the community subscribers.

The dataset contains the user discussions, obtained during Parliament elections in Russia in 2016. It consists of approximately 6000000 messages and more than 36000 actors who re-posted a message more than 10 times.

One should say that there are still few problems with the quality of the database.

1. There is a difficulty modeling users and groups simultaneously, because two different types of objects are presented on the same graph.
2. Low value of betweenness centrality metric (lack of connections).
3. Groups do not make direct references to other groups, which are the sources of information.

The following pre-processing steps were conducted. First, the BIRCH (balanced iterative reducing and clustering using hierarchies) data mining unsupervised algorithm was used in order to perform hierarchical clustering over particularly large datasets. We also used Locality-sensitive hashing (LSH) to reduce the dimension of data set.

1. All given documents were sorted by length.
2. The analysis included a context window of 10000 documents.
3. The size of search window was chosen of 2 weeks.

# 5 Model Description

In our project we have used two different models. The first model constructs a weighted graph where edges reflect apriori extent to which users influence each other. The second model provides an advanced model of wave propagation on the baseline graph in order to adjust weights for better wave propagation modelling. This simulation is used to measure similarity of our model compared to information spread in real networks.

## 5.1 Architecture

**Weight generating model** We begin with two types of data that was taken from the real network. The first type is a list of communities with their users. With this data, a baseline graph is built with nodes corresponding to communities and edges connecting communities with common users. The weights correspond to Jaccard index for users of connected nodes in baseline graph.

Second type of data is "wave", which is a message that is posted in communities. Each wave is a list of users with time of their activation. For each wave, we consequently modify baseline weights and adjust the model based on new information on information propagation. The algorithm is correct with respect to different order of waives during graph processing. In what follows, we describe processing one wave.

In the beginning, we make all the nodes belonging to the wave active simultaneously. Then, we make look on the neighbours of active nodes. If neighbour is not active then we decrease the weight of edge from active node to its neighbour by applying a function of specific type depending on the current weight. Otherwise, we choose the node, which was activated earlier and increase weight of the edge connecting the first activated node to its neighbour activated later.

To change the weight we find the argument of chosen monotonic piecewise continuous function taking values in $[0,1]$, then change the argument by adding/subtracting constant parameter (computed via model fitting), and finally calculate the value of new weight as the function value on a new argument. We choose Sigmoid function among many of so-called "activation functions" [12] often used in emulating non-linear processes.

Let us demonstrate processing a simple graph with a short wave consisting of three vertices (see Fig. 1). In this wave, vertex 0 was activated in time $t_1$ and vertex 1 was activated in

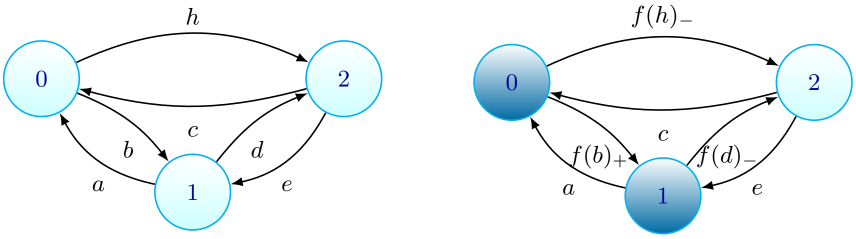

**Fig. 1.** Example of re-weighing on waves with 3 vertices.

Dark blue colour indicates activated status of a vertex. Left graph presents wave with activated vertex 0 at time $t_1$ and vertex 1 at time $t_2 < t_1$. Right graph shows updates of weights: the functions of weights on edges $f(d)$ and $f(h)$ were reduced due to inactivity of node 2, while $f(b)$ was increased for activation of node 1.

### 5.2 Model Evaluation

After adjusting weighted graph based on "wave" data, we aim to preserve the following property of modelling the information waves: we want to simulate the existing waves for the obtained graph and get almost the same number of activated vertices as in the original wave. The question of comparing the differences between sets of activated vertices is left for the future work.

We take a vertex with the earliest time of activation from $wave_i$ and make it active on our graph. Next, we make its neighbour active with probability

$p = W_{out}$, similar to IC model. In order to reflect the assumption that users who have seen information from a fairly reliable source and refused to get activated will never be activated with this wave, we remove all the nodes that were not activated if the weight edge connecting them is greater than parameter $Q$ standing for the threshold of the information spread reliability. Choosing proper parameter of reliability we guarantee convergence of our simulation to non-trivial solution close to the original wave. In what follows, we continue this process for all the activated neighbours for the graph without removed vertices. Hereby, we demonstrate the process of verification on the small network shown on Fig. 2 with simple case $Q = 0$ of complete reliability. Wave starts from vertex 0. Next, vertex 2 was activated and vertex 1 was not, so it was removed from the future analysis. Then, we analyse only neighbours of 2 that were not removed. Suppose that 3 was removed and now our algorithm stops and return the final amount of activated nodes by this wave. These two steps are represented on Fig.2. One can observe graphical representation of this process in Fig. 3.

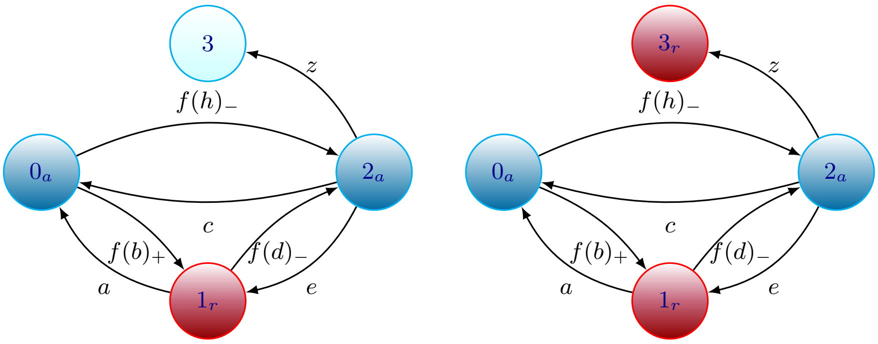

**Fig. 2.** Verification graphs

### 5.3 Implementation

First, we write an algorithm for assigning and evaluating the weights (see Algorithm 1). Next, we provide the step-by-step algorithm of implementing the simulations (see Algorithm 2).

### 5.4 Discussion

The main result of our project is a new model that could be used for modelling information propagation in social networks. In order to check applicability of our model we have built two metrics that will check if our algorithm works better than cascade model.

**Algorithm 1:** . Weight generating model

```
Data: $G(V, E, W), Waves = (S \subseteq V, T : V \to \mathbb{N})$
Result: $G(V, E, W_{new})$
for $wave$ in $Waves$ do
    for $U$ in $wave$ do
        for $V$: $U \to V$ do
            if $v \in wave$ then
                if $T_u < T_v$ then
                    $W_{uv} := \sigma(\sigma^{-1}(W_{uv} + \delta)$
                end
            else
                $W_{uv} := \sigma(\sigma^{-1}(W_{uv} - \delta)$
            end
        end
    end
end
```

**Algorithm 2:** . Simulation

```
Data: $G(V, E, W), s \in V$
Result: $\{F_i\}_{i=1}^{t}$
$F_0 := \{s\}$
$Excluded := \varnothing$
$t := 0$
while $F_t \neq \varnothing$ : do
    for $u \in F$ do
        for $v : u \to v$ do
            if $v \notin Excluded$ then
                $F_{t+1}.add(v)$ with probability $W_{uv}$
            end
        end
    end
    $Excluded.add(\{v | u \to v, v \notin F_{t+1}\})$
    $t := t + 1$
end
```

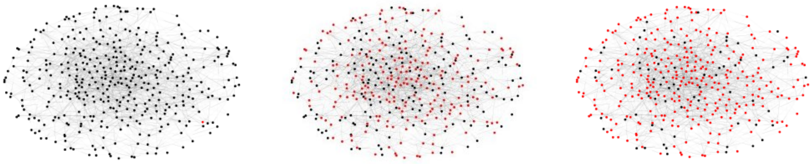

**Fig. 3.** Example of the nodes activation in information spread simulation. Three graphs, from left to right, demonstrate the spread of information in three moments of time: start, intermediate step and finish of simulation.

The first metric is an interval for activated node. In order to obtain this metric, we take some random node $V_0$ from wave. Next, we apply it to our model and look on the amount of nodes that were finally activated during 500 iterations. Then, we build the intervals of 95% values and repeat the same procedure for cascade model with different thresholds, observing which of them is narrower and less biased.

Next, we build two graphs of dependence between number of nodes and period of activation. The aim is to find there final activation value of each particular wave. Finally, we merge these three graphs in order to compare the results.

## 6 Conclusion

We proposed a weighted graph model with tuning specific weights based on empirical data that might outperform the existing models. There is still a lot of directions for improvements in this research area. Considering the last step performed in this paper, it would be a great improvement to suggest the model of optimisation for threshold parameter, which is required for not considering the inactive nodes (in other words, removing them from the final subset of activated nodes). It also makes sense to compare the results obtained via proposed model with those we obtained using the independent cascade model or any other state-of-the-art models (e.g., another type of cascade model — linear threshold model). Depending on the results we could conclude whether proposed model is more effective or better by any other criterion.

**Acknowledgements.** The article was supported within the framework of a subsidy by the Russian Academic Excellence Project '5-100' and RFBR grant 16-29-09583 "Methodology, techniques and tools of recognition and counteraction to organized information campaigns on the Internet". The data set used in this research paper was provided by the NRU HSE International Laboratory for Applied Network Research.

## References

1. Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012, April). The role of social networks in information diffusion. In Proceedings of the 21st international conference on World Wide Web. (pp. 519 – 528). ACM.
2. Cha, M., Mislove, A., Gummadi, K. (2009). A measurement-driven analysis of information propagation in the flickr social network. In: Proceedings of the 18th international conference on world wide web, WWW 09, pp. 721730.
3. Chen, W., Wang, Y., & Yang, S. (2009, June). Efficient influence maximization in social networks. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 199-208). ACM.
4. Chen, W., Wang, C., & Wang, Y. (2010, July). Scalable influence maximization for prevalent viral marketing in large-scale social networks. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1029-1038). ACM.

5. Coleman, J., Katz, E., & Menzel, H. (1957). The diffusion of an innovation among physicians. Sociometry, 20(4), 253-270.
6. Dodds, P. Watts, D. (2005). A generalized model of social and biological contagion. Journal of Theoretical Biology 232, 4, 587 - 604.
7. Domingos, P., & Richardson, M. (2001, August). Mining the network value of customers. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 57 – 66). ACM.
8. Goldenberg, J., Libai, B., Muller, E. (2001) Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth. Marketing Letters. 12:3, 211–223.
9. Goldenberg, J., Libai, B., Muller, E. (2001) Using Complex Systems Analysis to Advance Marketing Theory Development. Academy of Marketing Science Review.
10. Goyal, A., Bonchi, F., & Lakshmanan, L. V. (2011). A data-based approach to social influence maximization. Proceedings of the VLDB Endowment, 5(1), 73-84.
11. Granovetter, M. (1978). Threshold models of collective behavior. Am J Sociol 83(6):14201443.
12. Karlik, B., & Olgac, A. V. (2011). Performance analysis of various activation functions in generalized MLP architectures of neural networks. International Journal of Artificial Intelligence and Expert Systems, 1(4), 111-122.
13. Kempe, D., Kleinber, J., & Tardos, E. (2003). Maximizing the spread of influence in a social network. In Proceeding 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 137-146).
14. Kempe D., Kleinberg J., Tardos E. (2005) Influential nodes in a diffusion model for social networks. In: Proceedings of the 32nd international colloquium on automata, languages and programming (ICALP). Springer Berlin Heidelberg, Lisbon.
15. Kimura, M., & Saito, K. (2006, September). Tractable models for information diffusion in social networks. In European Conference on Principles of Data Mining and Knowledge Discovery (pp. 259-271). Springer Berlin Heidelberg.
16. Kostka, J., Oswald, Y. A., & Wattenhofer, R. (2008, June). Word of mouth: Rumor dissemination in social networks. In International Colloquium on Structural Information and Communication Complexity (pp. 185-196). Springer Berlin Heidelberg.
17. Liggett, T. (1985) Interacting Particle Systems. Springer.
18. Leskovec, J. (2010). Epinions social network.
19. Morris S. (2000). Contagion. Review of Economic Studies 67. pp. 57 – 78.
20. Onnela, J., Saramaki, J., Hyvonen, J., et al. (2007). Structure and tie strengths in mobile communication networks. PNAS 104(18):73327336.
21. Centola, D., Eguluz, V., Macy, M. (2007). Cascade dynamics of complex propagation. Physica A Stat Mech Appl 374(1):449456.
22. Subbian, K., Aggarwal, C., Srivastava, J. (2016) Mining influencers using information flows in social streams. ACM Trans Knowl Disc Data 10(3):26.
23. Zhao, J., Wu, J., Feng, X., Xiong, H., & Xu, K. (2012). Information propagation in online social networks: a tie-strength perspective. Knowledge and Information Systems, 32(3), 589-608.