Scientific Visualization, 2022, volume 14, number 2, pages 98 - 107, DOI: 10.26583/sv.14.2.09

Critical Paths of Information Dissemination in Networks

M.S. Ulizko[1,A,B], A.A. Artamonov[2,B], R.R. Tukumbetova[3,A,B], E.V. Antonov[4,B], M.I. Vasilev[5,B]

[A] Plekhanov Russian University of Economics,
Stremyannyy Pereulok, 36, 115093 Moscow, Russia
[B] National Research Nuclear University MEPhI (Moscow Engineering Physics Institute),
Kashira Hwy, 31, 115409 Moscow, Russia

[1] ORCID: 0000-0003-2608-8330, mulizko@kaf65.ru
[2] ORCID: 0000-0002-9140-5526, aartamonov@kaf65.ru
[3] ORCID: 0000-0002-1976-1390, rrtukumbetova@kaf65.ru
[4] ORCID: 0000-0003-1498-9131, eantonov@kaf65.ru
[5] ORCID: 0000-0002-1821-3763, vasilev.michael.98@gmail.com

Abstract

The development of information and communications technology in the 21st century has led to a shift in the distribution of information from traditional print media to electronic media. New means of informing the public have emerged, such as social networks and messengers, which made it possible to share information almost instantly and over long distances. In this regard, it has become a daunting task to analyze the patterns of information spreading, particularly, by studying information sources, revealing hidden interests and bias, identifying opinion leaders. The article presents methods for solving the problems of network analysis of identifying information dissemination patterns, determining «critical paths» of information dissemination by analyzing messages and news in the Telegram messenger.

Keywords: Network analysis, informational signal, edit distance, Telegram, Gephi, graph, critical path.

1. Introduction

Network information resources are the main channel for obtaining information about what is happening in the country and the world. The transition of traditional media from paper to electronic form is taking place, some publishers have completely abandoned the printed versions [1]. Due to the news abundance, many publishers are moving to instant messaging platforms - such as Telegram, WhatsApp.

Instant messaging platforms allow creating specialized channels both to inform the population, and to receive feedback (using the function "comment" and graphic response to the message) [2]. This method of conveying information is characterized by zero cost of entering the platform, as an organization does not bear any financial costs to maintain and support the resource.

Currently, plenty of news agencies, individuals (bloggers), commercial and governmental organizations, executive authorities have their own news channels in the Telegram messenger. The number of users following such channels varies from a few dozens to several million people.

The paper investigates the information dissemination models on the Telegram instant messaging platform. The object of study is an individual information message, which can be distributed within closed environment infinitely. Besides Telegram can be considered as one of the subtypes of the social network.

A significant number of articles and books have been devoted to the study of the phenomenon of social networks and methods of information dissemination [3, 4]. Separately, we can figure out [5] a considered social network as an object.

Researchers distinguish several classes of "game-theoretic" models of social networks:

- Models of mutual awareness.
- Models of coordinated collective action.
- Communication models.
- Stability models.
- Models of informational influence and management.
- Models of information confrontation.

The article considers models of informational influence and management that have the unique property of having “opinion leaders” [3]. We have developed a tool for collecting textual information and metadata related to reposts and reactions to information messages and methods of their graph analysis.

2. Methodology

Considering Telegram as one of the subtypes of a social network, we highlight that the nodes are entities such as personalities, channels and groups, and the edges are informational messages, membership of participants in groups, etc. [6, 7].

The first objective is to collect information from Telegram channels. This task can be solved using two resources - Telegram and TGStat website, which provides statistics about Telegram.

The joint use of the resources allows ensuring the completeness of the collected messages and metadata necessary for the development of the model of information distribution in the network. The collection process is shown below (Figure 1).

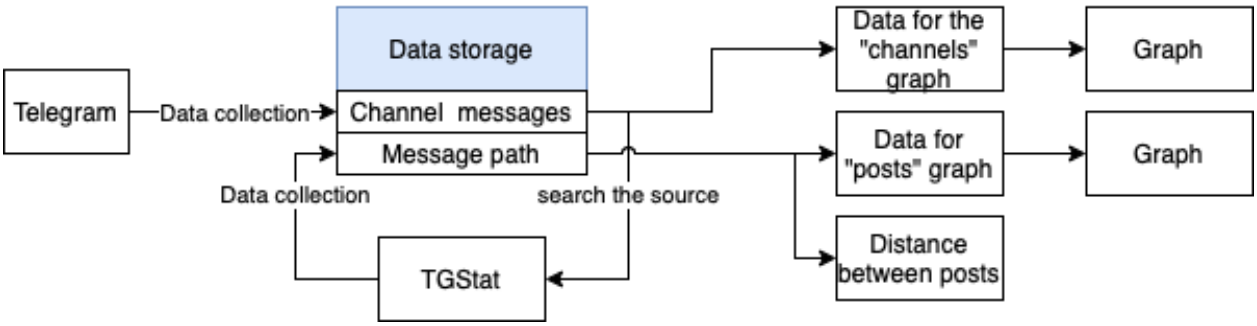

Figure 1. Data collection process

There are three major ways to interact with Telegram: using apps (mobile and desktop), using a browser, or using API. The last method is used, as it provides the most complete and the fastest data collection. According to Telegram API each message is characterized by 29 attributes, but the most significant in the research are:

- Id – message identifier in the channel,
- peer_id – channel information,
- date – publication date and time,
- message – publication text,
- fwd_from – information about forwarded messages.

The “fwd_from” attribute consists of 10 fields, the most significant are:

- date – publication date and time,
- from_id – information about the primary source,
- channel_post – message identifier in the primary source.

Telegram has more than 300,000 channels, for the purposes of the experiment, the analysis was conducted on the sample of 30 channels, with a total audience of more than 3.5 mil-

lion users. Channels were selected based on their audience and were not linked by their thematic focus.

After implementing the data collection, a data model was developed to build the directed graph. The nodes are the channels, the arcs (directed edges) indicate message forwarding from one channel to another. The node of the graph is defined by the following fields:

- Id – node identifier,
- label – channel ID according to Telegram API,
- name – channel name.

An arc is defined by the following fields:

- Source – identifier of the start node,
- Target – identifier of the end node,
- Type – edge type (directed),
- Weight – weight of the node,
- Date – date of message forwarding.

The dissemination of information message in Telegram can be considered from two perspectives: in terms of causes and predictions. Adrien Guille et al. [3] describe these models for distribution of messages between users, but this approach can also be applied to Telegram if we consider individual channels as users. We hypothesize that some channels are more interconnected than others. It is possible to identify this correlation, as well as to obtain a predictive model, by comparing distribution paths. So, we should consider their structure.

In terms of graph theory, a propagation path can be considered as an oriented tree, where each parent can have several children. The tree, as a graph, can be both numbered by vertices and unnumbered.

Considering the propagation path as a graph, we can use the following metric to compare two paths with each other [8]:

$$d(G_1, G_2) = 1 - \frac{|mcs(G_1, G_2)|}{\max(|G_1|, |G_2|)}, \quad (1)$$

where G_i – graph i, $\text{msg}(G_i, G_j)$ – maximum common subgraph, $|G_i|$ – the order of graph i

The disadvantages of this metric are that the algorithms to find the maximum common subgraph are labor-intensive, and that it does not take into account the graph structure. There are other metrics [9, 10], but they also do not take into account the type of a graph.

To determine the similarity of objects we will use edit distance [11]. Edit distance is a metric defined as the minimum number of one-character operations (insertion, deletion, replacement) needed to turn one sequence into another. In case of node-numbered trees, the node number will be the Telegram channel Id, so edit distance between two dissemination paths for different topics will be large. Therefore, in order to compare paths that propagate through different channels, unnumbered channels are used – in this case more general conclusions can be drawn.

Edit distance for oriented unlabeled trees can be calculated based on their string representation [12].

In this paper we consider the approach of visual graph analysis. To perform this we will map each object to a graph according to the following rules:

1. Each node will have a spread depth.
2. Each arc will have a weight inversely proportional to the forwarding time.

3. Results and discussion

3.1. Spread of messages between channels (statistical analysis)

395 795 informational messages were collected in the Telegram messenger over the 3 months period, then they were converted for graph representation, visualization was per-

formed using the tool Gephi [13]. That tool was chosen due to the user-friendly interface and the possibility of adding a dynamic component.

First, the graph for 30 channels and their output to external sources was plotted (Figure 2). In this figure channels for which messages were collected are highlighted in green, and those channels from which redirection took place are highlighted in red. Node size is directly proportional to the number of messages that were forwarded from this channel.

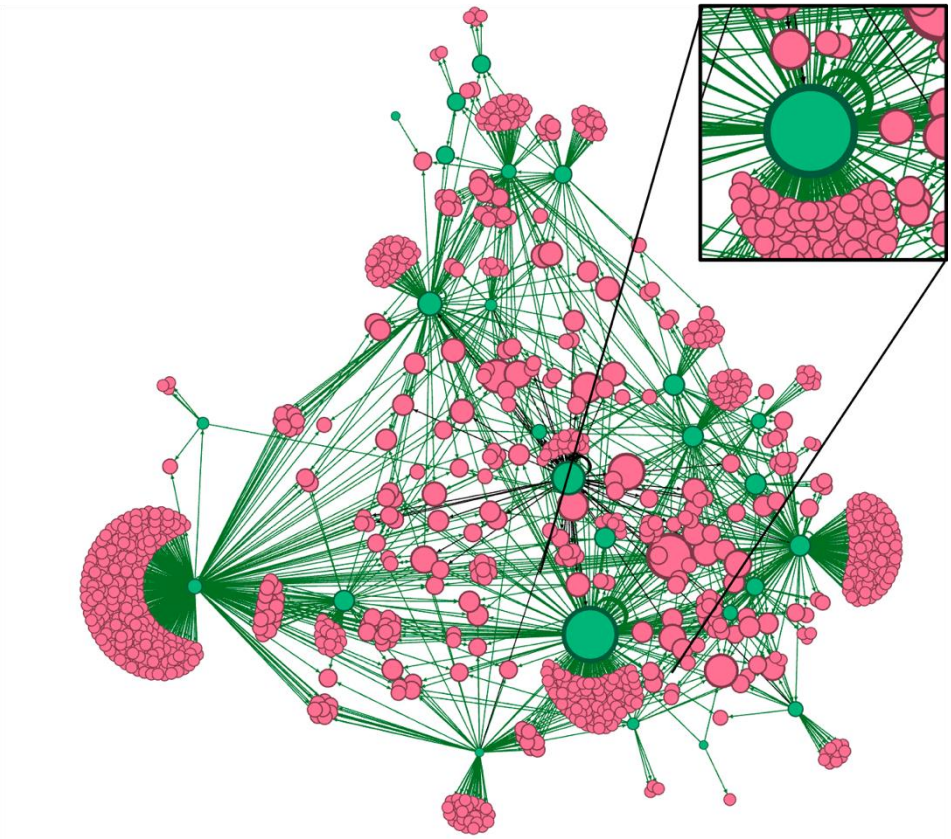

Figure 2. Interconnection of information channels

The graph shows a certain homogeneity of the information field in the center, which characterizes strong interrelations among the participants of the information field. Wide range of "buds" on the periphery indicates the number of channels being monitored. It is also possible to make preliminary conclusions about the amount of unique content generated by the channel. The more "dead-end" edges indicate the less unique content produced. However, in order to confirm this statement, it is necessary to conduct additional research to assess the ratio of content within the information channel, which is one of the extensions of the authors' research.

Note the presence of reflexive arcs (as in the highlighted fragment in Figure 2), that is, some channels refer to their own messages. This can confirm the authority of the channels, but negatively affects the overall perception of the information field.

The second stage was the expansion of the set of channels for message collecting to 35 pieces. Channels with the largest number of links were added to the sample, based on the analysis of the graph (Figure 2), and then new graphs were plotted. For them the reflexive connection was removed and additional color design was added to improve perception.

First, a graph is constructed to obtain overall statistics by channel (Figure 3). There are several particularly interesting aspects for researchers.

First, there are apparent opinion leaders (“8”, “3”, “23”), which are predominantly referred to by other channels, stand out. On the other hand, it can be traced that the major source of information for the channel “4” is the channel “5”, while the rest of the messages are taken from channels to which no other authoritative channel is affiliated.

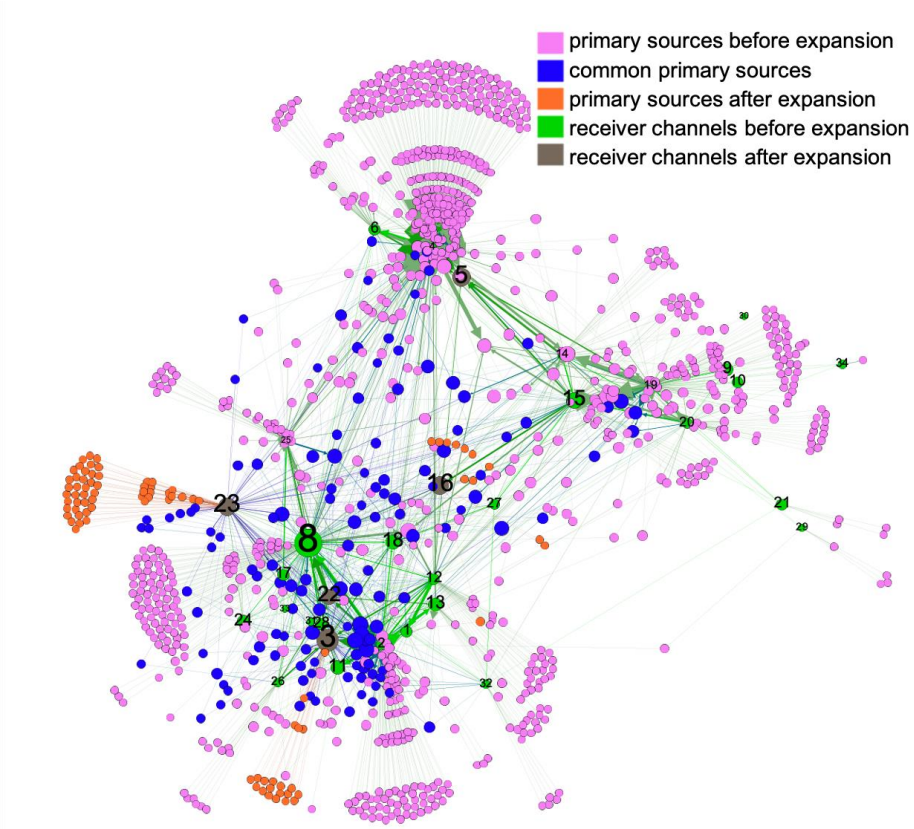

Figure 3. Interconnection of information channels at the second stage

The second graph was constructed with the dynamic component by the date of forwarding to the target channel (*Figure 4*).

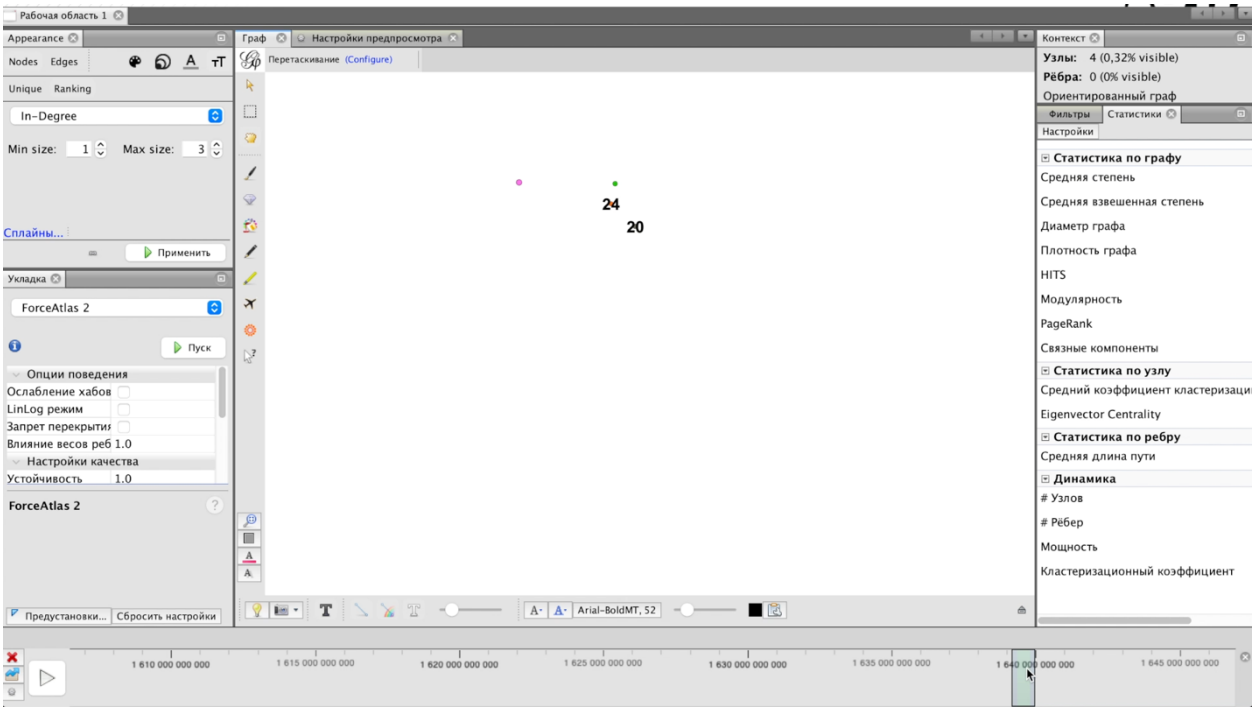

Figure 4. Dynamic graph of information channels interconnection

This approach allows estimating the flow of forwarded messages for different time intervals. The data sample presents the 3 months period of 2022, which is shown in the figure (by converting timestamp to date), we can see the burst of publications in the middle of the sam-

ple (about 20%), which indicates a significant newsworthy event that occurred in this time interval. In addition, we can note the timing of the interaction between the isolated channels (for example, “16”, “33”, “34”) and more central channels.

The limitation of this method is that the size of nodes (which is proportional to the number of links per channel) does not change its size in different time intervals.

3.2. Spreading messages between channels (dynamic analysis)

The distribution of an information message can be described using a probabilistic model. In this case, one of the possible ways for the analysis is the cascade model of infection [4]. Since the work is aimed at obtaining a predictive model and detection of anomalous phenomena in the graph, and the propagation time can take from several minutes to several days, this model is insufficient for the analysis. On the other hand, propagation paths can be compared by edit distance and visually. In this case, the graphical representation, theoretically, allows quick highlighting of anomalies on a single example.

The information signal propagation in Telegram is shown using Gephi. The label of edges is calculated as the time difference (in minutes) between the publication of a message in channel i and its forwarding in channel j, and the weight of an edge is determined by the following formula:

$$w_{i,j} = \frac{k}{d_{i,j} + 1}, \tag{2}$$

where k – proportionality coefficient (take k=10), $d_{i,j}$ – the time difference (in minutes) between the publication of a message in channel I and its forwarding in channel j.

170 objects were selected for analysis. Out of 170 objects, 48 pairs (0.4%) were found (on average 12 vertices in one graph) in which the editorial distance satisfies the condition:

$$r_{i,j} < \frac{\max(|G_i|, |G_j|)}{4} \tag{3}$$

Let's consider two graphs (Figure 5, Figure 6). On each of the above graphs a color scheme is used to separate nodes by the depth of propagation, and the numerical designation is the edge label.

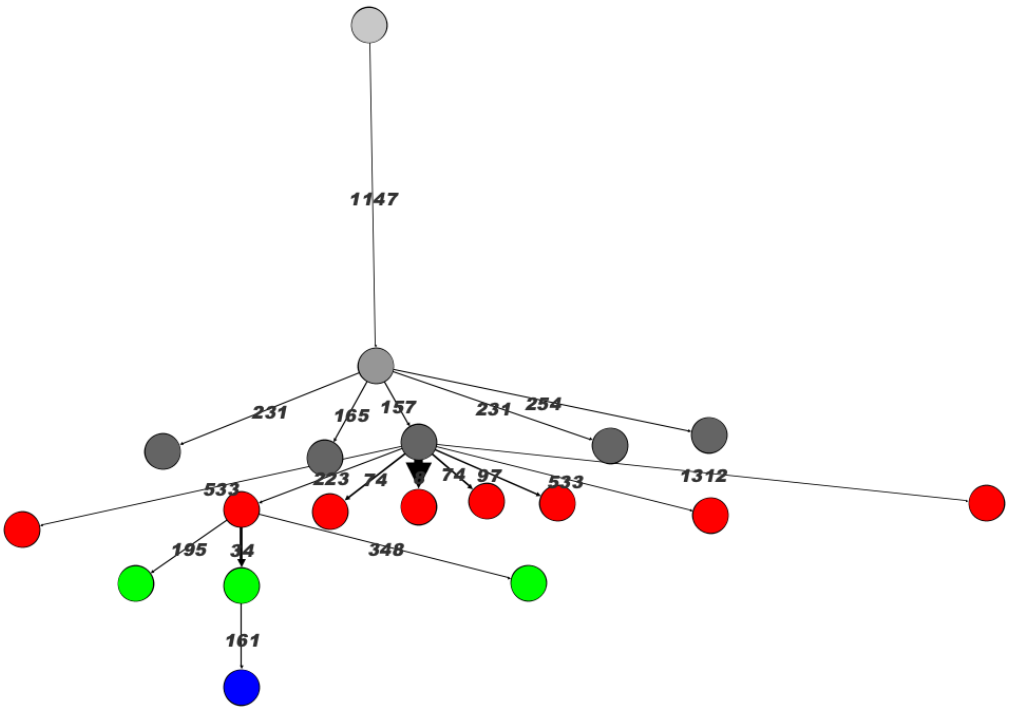

Figure 5. Hierarchical representation of the distribution graph

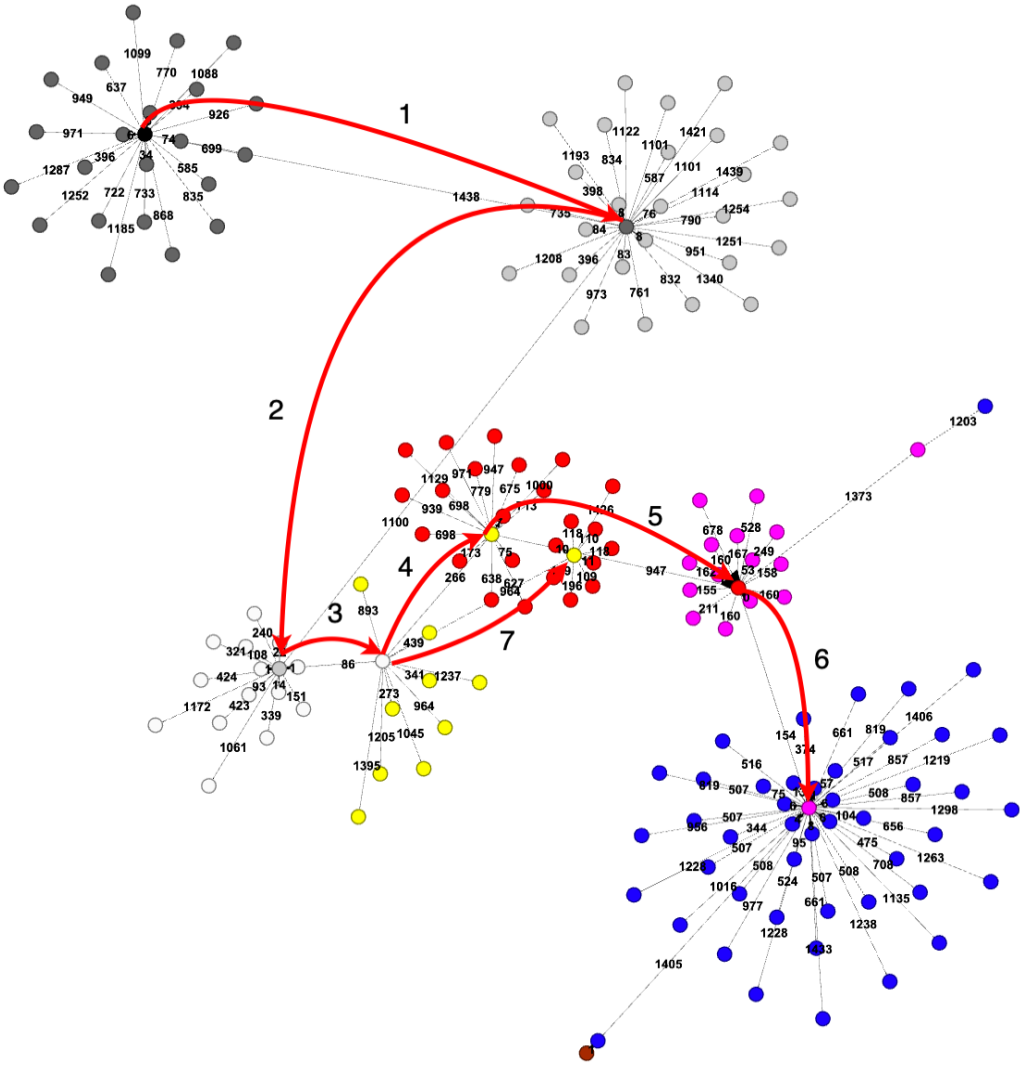

Figure 6. Propagation graph build by ForceAtlas2 algorithm

Two features can be emphasized on the first graph. First, some posts are forwarded almost immediately, while others are forwarded much later. On the other hand, some posts are forwarded to the exact minute, although they take quite a long time. For example, the first graph highlights edges with forwarding times of 231 minutes at level 2; with forwarding times of 74 and 533 minutes at level 3. Since there are 3 pairs of completely identical edges out of 18 edges (by time and absence of child elements) there is an assumption about the connection of channels-receivers in pairs with each other or about intentional forwarding of information signals.

The second graph represents one of the largest propagation networks, containing 165 nodes and a maximum depth of 7. The ForceAtlas 2 visualization algorithm has adequately placed the nodes by propagation depth. This example highlights the chain of nodes that corresponds to the main propagation. In the majority of cases, messages are forwarded from these nodes, indicating that these channels are authoritative. The identified chain is associatively similar to the critical path of the Gantt chart [14,15] used in project management theory.

Let's introduce the concept of critical path of signal propagation as a sequence of nodes participating in the propagation of information signal, providing the greatest information coverage.

In addition, let us introduce the characteristic of time. For this purpose, a graph is built, where each node has a label “time since the beginning of propagation” (in hours) (Figure 7).

This graph contains two critical paths which contain 8 nodes each, the time before forwarding to the end nodes was 858 hours (35 days) and 778 hours (32 days).

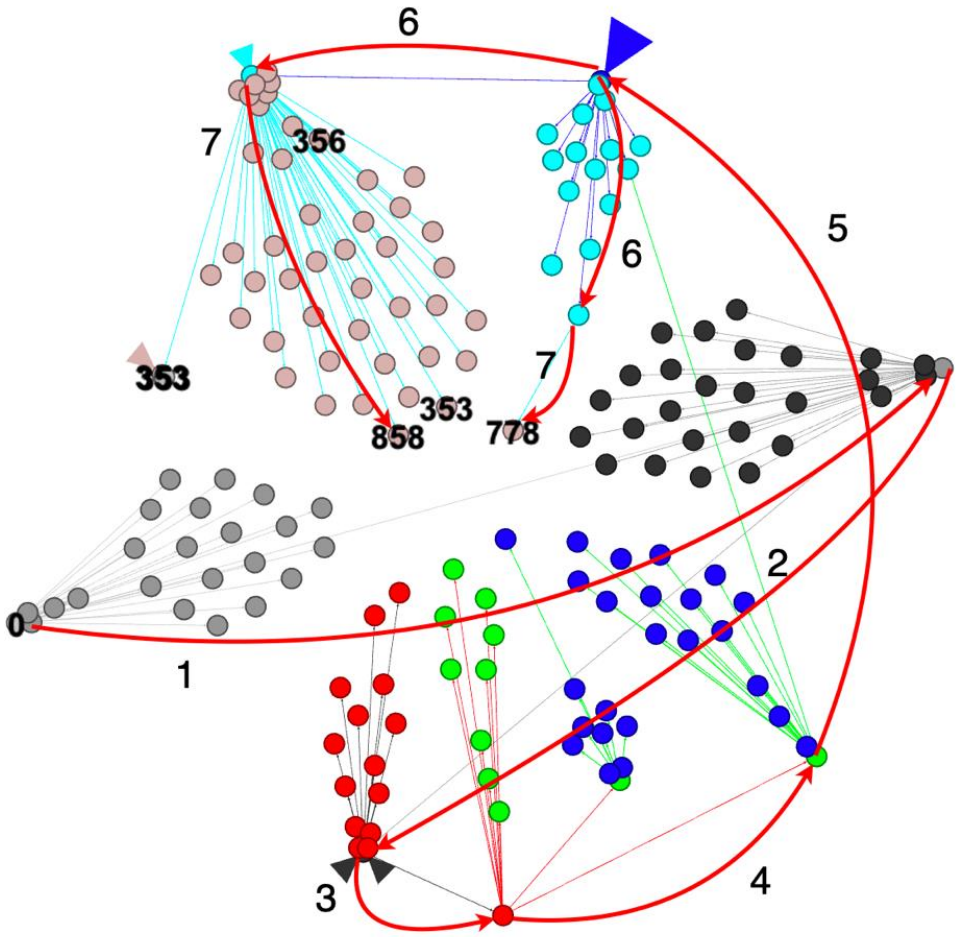

Figure 7. Critical paths of information signal propagation

4. Conclusion

Social networks and instant messaging platforms represent a field for research on information signals. The tasks of such research are to determine the nature of propagation, to reveal the relationships between individual subjects/objects, and to build the predictive model of propagation.

Two tasks were solved in the work: the analysis of information dissemination in Telegram for the most popular channels (30 channels in the initial sample) and the consideration of individual information signals. The analysis of channel interaction potentially allows identifying "opinion leaders", "aggregators" of information and bursts of forwarding activity.

The distribution of individual information messages can be described by graph theory, but comparing the constructed “trees” causes difficulties. Graphs can be compared by calculating the edit distance, but the impossibility of using temporal characteristics (forwarding time) imposes a limitation on the use of this metric. On the other hand, the visual representation of such objects allows highlighting critical places, such as identical ways of message propagation (the same propagation time). A separate class of tasks is the study of “critical paths” of signal propagation, which will allow solving a wide range of tasks related to management and confrontation in information networks, including identification of primary sources, ways of information dissemination, “opinion leaders”.

References

1. Huffpost, https://www.huffpost.com/entry/arthur-sulzberger-we-will_n_710251 – (Last accessed: 2022/04/15)..

2. Kulik, S. (2016). Factographic information retrieval for competences forming // 2016 3rd International Conference on Digital Information Processing, Data Mining, and Wireless Communications, DIPDMWC 2016, 2016, pp. 245–250, 7529397.

3. Guille, A., Hacid, H., Favre, C., & Zighed, D. A. (2013). Information diffusion in online social networks. In ACM SIGMOD Record (Vol. 42, Issue 2, pp. 17–28). Association for Computing Machinery (ACM). https://doi.org/10.1145/2503792.2503797.

4. M. Gomez Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In KDD ’10, pages 1019–1028, 2010.

5. Gubanov D.A., Novikov, D.A., Chkhartishvili A.G. Modeli vliyaniya v social'nyh setyah (obzor) // Upravlenie bol'shimi sistemami, 2009. №27. S. 205-281.

6. Ulizko, M. S., Antonov, E. V., Artamonov, A. A., Tukumbetova, R. R. (2020). Visualization of Graph-based representations for analyzing related multidimensional objects. In Scientific Visualization (Vol. 12, Issue 4). National Research Nuclear University MEPhI (Moscow Engineering Physics Institute). https://doi.org/10.26583/sv.12.4.12.

7. Cherkasskiy, A., Artamonov, A., Cherkasskaya, M., and Leonova, N. (2021). Methods for identifying an information object in social networks // Procedia Computer Science, 2021, 190, pp. 137–141. https://doi.org/10.1016/j.procs.2021.06.017.

8. Bunke H. and Shearer K. A graph distance metric based on the maximal common subgraph // Pattern Recognit. Lett., 1998, vol. 19, no. 3–4, pp. 255–259.

9. Wallis W., Shoubridge P., Kraetz M., and Ray D. Graph distances using graph union // Pattern Recognit. Lett. 2001. V. 22. P. 701–704.

10. Moskin, N. D. (2021). Metric for comparing graphs with ordered vertices based on the maximum common subgraph. In Prikladnaya Diskretnaya Matematika (Issue 52, pp. 105–113). Tomsk State University. https://doi.org/10.17223/20710410/52/7.

11. S. Y. Lu, A tree-to-tree distance and its application to cluster analysis, IEEE Trans. Pattern Anal. Mach. Intelligence, (1979), pp. 219-224.

12. Nettleton, D. F., & Salas, J. (2016). Approximate Matching of Neighborhood Subgraphs — An Ordered String Graph Levenshtein Method. In International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems (Vol. 24, Issue 03, pp. 411–431). World Scientific Pub Co Pte Lt. https://doi.org/10.1142/s0218488516500215.

13. Gephi, https://gephi.org – (Last accessed: 2022/04/01).

14. Kelley J E and Walker M R 1959 Critical-path planning and scheduling Proc. Eastern Joint Computer Conf. (IRE-AIEE-ACM) 160-173.

15. GanttPro / Kriticheskij put' v MS Project, https://blog.ganttpro.com/ru/kriticheskiy-put-critical-path-ms-microsoft-project/- (Last accessed: 2022/04/20).