The technique of structuring social network graphs for visual analysis of user groups to counter inappropriate, dubious and harmful information

M Kalameyets[1], A Chechulin[1] and I Kotenko[1,2]

[1]Laboratory of computer security problems, St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, Saint-Petersburg, 194100, Russia.
[2]International Laboratory of Information Security of Cyber-physical Systems, ITMO University, Saint-Petersburg, 197101, Russia.

Abstract. The paper proposes a data visualizing technique for analyzing social networks in order to identify and counteract inappropriate, dubious and harmful information. The proposed technique is based on the force-layout technique of drawing graphs in which the parameters of vertices and edges are calculated depending on the number of links. The paper provides an example of using the proposed data visualization technique for social networks analysis on the basis of visual analytics of several groups in the "VKontakte" social network in four display modes: without using the technique, in 2D, in 3D and in augmented reality. The example also contains a description of various drawing techniques and estimates them.

1. Introduction

Nowadays, the importance of social networks is growing, because they influence greatly on the individuality and behavior of people. At the same time, the reputation institute has spread widely in the Internet, therefore users have a growing trust in social networks as a relevant information source in comparison with traditional communication channels - television, radio, magazines, etc. But in addition to news, social networks have become a platform for the dissemination of inappropriate, dubious and harmful information, for example, the ideas of homeopathy, denial of the Holocaust, human immunodeficiency virus (HIV) dissidence, etc. It is extremely difficult to assess the scale of such information dissemination and to influence on its distribution [1]. A possible solution for discovering and analyzing information flows in social networks is to visualize data which allows one to represent complex processes in a simple graphic manner [2-4].

A distinctive feature of visual analytics is that all decisions are made by a person. Thus, it can be used to detect trends, identify and make individual decisions in cases where automated techniques (for example, machine learning methods) give a high percentage of false positives. Nevertheless, in view of the large amount of information analyzed, visualization of social networks often becomes unreadable and difficult to perceive in order to carry out effective analysis.

In this paper, in order to simplify the visual analysis of social networks, it is proposed to use the force-layout technique of graph drawing. The novelty of this work lies in the approach for the force-layout parameters calculation on the basis of statistical data, which allows to spatially cluster the vertices of the graph, thereby allocating clusters of users. The contribution of this work is the technique for selection of parameters for force-layout drawing of social networks in the interests of identifying and countering inappropriate, dubious and harmful information.

The paper is organized as follows: the second section reviews the techniques used to visualize social networks: the third section describes the proposed approach for calculating the parameters of the force-layout graph of the social network and describes the experiments conducted; the fourth section is an evaluation of the proposed approach and explanation of results; the fifth section conclude the paper results and plans for future work.

The paper uses the terminology which are used in the social network "VKontakte" and can differ from the terminology of other social networks:

1) User in social network represents user's webpage with a microblog in which he or she can post content like video, images, texts and audio.

2) Group in social network is a webpage which is administered by several users together. Users can join group to read its content. Groups can be created either by companies or news agencies, or by simple users.

3) Post is a content like video, images, texts and audio that can be posted by user or group.

4) Repost is a post which a user or a group copied from the page of another user or group. Repost contains information from whom it was copied.

5) Repost tree is a chain of reposts that occur when users repost posts.

2. Related work

Currently, there are many visualization models that allow one to display various processes occurring in social networks [5]. The most studied models include the models based on graphs, TreeMaps and matrices. So, with the help of graphs, one can effectively analyze the connectivity of users. Interval graphs models allow one to analyze the temporal relationship between different events.

In the form of tree, the repost tree can be represented, where the source of the post is the root.

TreeMaps models allow one to efficiently analyze metrics (such as the number of "likes", user coverage, post discussion) of reposts.

Using matrices one can analyze small graphs with a large number of links.

It is also possible to use more specific models, for example: combined models [6], models based on Voronoi maps [7] in the case of planar graphs and Voronoi TreeMaps [8] as an alternative to the TreesMaps, Chord diagrams and geo-map models. The list of possible models can be continued [5].

It also important to note, that the same dataset can be visualized using different visualization models [9-10]. For example, in Figure 1 you can see how single dataset with two types of links (links that form a hierarchical structure and links that form a planar structure) can be visualized using different visualization models [9].

At the same time, effective visual analytics should correspond to the principle of Ben Schneiderman [11]: to be able to view the entire data, be scalable and filterable.

The following principles can also be considered as the key requirements for visualization:

1) Ability to use different drawing techniques [12]. For graphs, this can be radial, force-layout, or cluster-based rendering (Figure 2).

2) The naturalness of the techniques of human-machine interaction. For example, a natural way to visualize 3D graphs is augmented reality, and in 2D mode, when using filters, one can use the capabilities of touch screens.

It should be noted that complex and multilevel visualization models (such as multi-level graphs, Chord diagrams, Voronoi TreeMaps, etc.), despite their clear advantage in the ability to display a large number of metrics, is more difficult to manage and percept. For this reason, the most common way to visualize social networks are graphs.

Graphs are the most universal visualization model, which can be relatively easily structured, and which supports data visualization capabilities even with hundreds of thousands of nodes and no fewer number of links.

The main disadvantage of the graph representation is that it becomes difficult to perceive by a large number of edges and vertices. In the field of visualization, there are many works devoted to clustering graph elements, ways of structuring, ways of drawing, building trees of different types, reducing the size of data, etc. Researchers who deal with this problem agree that in order to simplify the visual analysis, it is necessary to develop techniques for structuring the elements of the graph.
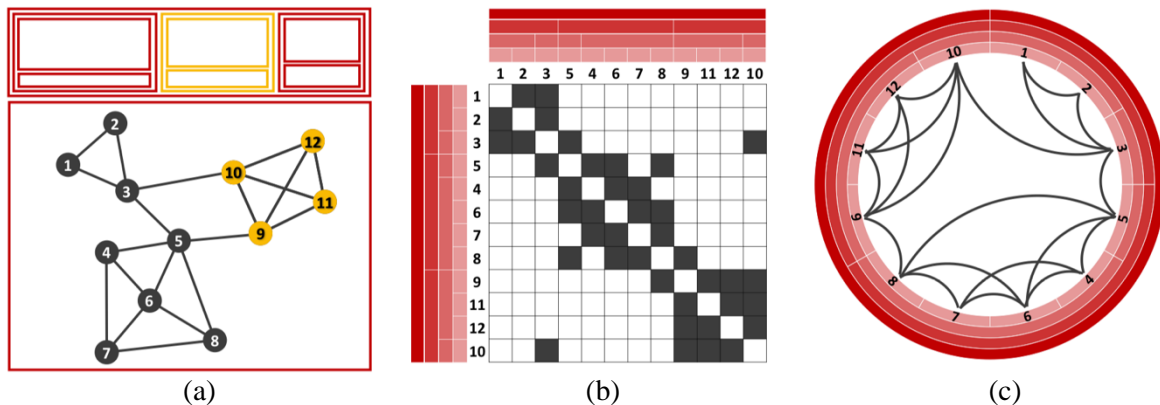
Figure 1. Dataset can be visualized using different visualization models: (a) Graph-TreeMap, (b) Tree-Matrix, (c) Chord diagram [9].
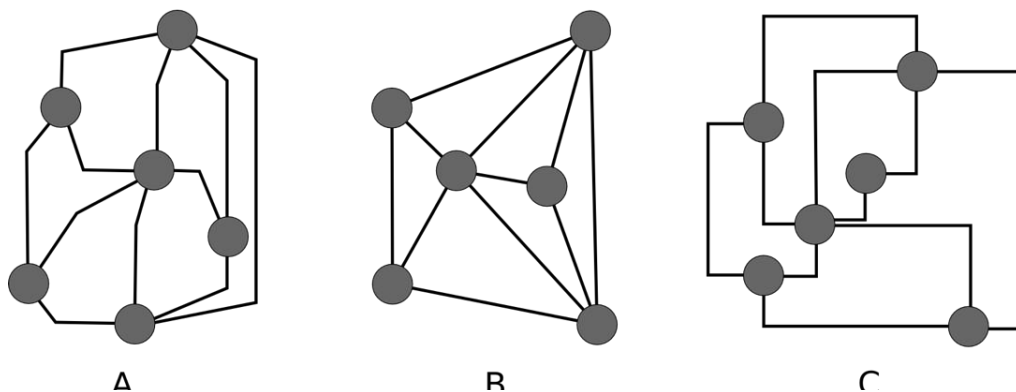

Figure 2. One graph can be depicted using different layouts [12]: (A) poly-line layout, (B) straight-line layout, (C) orthogonal layout.

For visual analysis of social networks, we propose a technique that is based on the force-layout drawing of the graph in which the parameters of vertices and edges are calculated depending on the number of links.

3. Technique of calculating the parameters of the force-layout graph of the social network

The force-layout technique of graph drawing is used to simplify perception by structuring the arrangement of vertices and the edges [13, 14]. The force-layout technique is based on physical simulation, which consists of three main components:

1) Each vertex is given the value of the charge, after which the vertices, similar to electrons with the same charge, are "repelled" from each other. The magnitude of this charge determines with what force this vertex repels. Thus, the vertices with a smaller charge will be closer to each other, while the vertices with a bigger charge will be farther away.

2) The edges of the graph connect the vertices similar to springs. Each edge is given a value of the length, which is similar to the length of the spring in the stretched form. Thus, the distance between the connected vertices is bounded by the length of the edge.

3) Each edge is given a value of strength, which is similar to the spring's elasticity. Strength determines the possibility of "stretching" and "squeezing" the rib. Thus, for edges with the strength equal to 1, the length of the edge is initially equal to the given, and in the other case, the edge will expand and contract depending on the values of strength and the charge of the vertices connected to them.

Since the vertices repel, and their edges restrain, the graph "comes into motion" until it comes to an equilibrium state. This state is the result of the algorithm. Getting a readable graph depends on the correct selection of these parameters for a certain graph structure.

For social networks we suggest to use the following technique:

1) For vertices that have only one edge and their edges:
   a. to set small charge value for vertex;
   b. to set maximum strength value of 1 for edge (the edge cannot be stretched);
   c. to set a small edge length.

2) For vertices that have the largest number of edges (statistically correspond to "outliers") and for edges between these vertices:
   a. to set a big charge value for vertex;
   b. to set maximum strength value of 1 (the edge cannot be stretched) for edges between these vertices (having the largest number of edges).
   c. to set a big edge length.

3) For all other vertices and edges,
   a. to set an average charge value for vertex;
   b. to set value of 0.1 for edge (the edge can be strongly stretched);
   c. to set an average edge length.

Let us consider examples of using this technique on the basis of visual analytics of several groups in "VKontakte" social network.

The initial data is a JSON file that has the following information: an array of objects that consists of users and groups; an array of links that consists of relations "repost" between users and users, users and groups, groups and groups.

Each vertex is a subject – user or group who reposted the post from the group or another group participant.

The edge indicates the presence of a repost between them.

Let us consider examples of implementation.

The simplest way to represent is a 2D graph. The graph of repost between users and groups is depicted in Figure 3 and Figure 4 in 2D display mode using the force-layout graph drawing technique, where the vertex is the user or group and the edge is the repost.
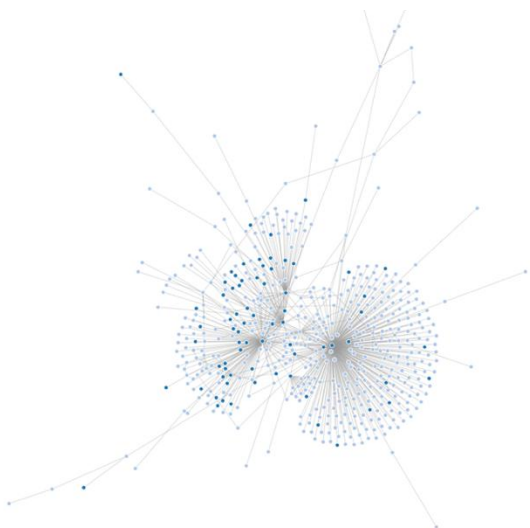

Figure 3. Force-layout technique of graph drawing
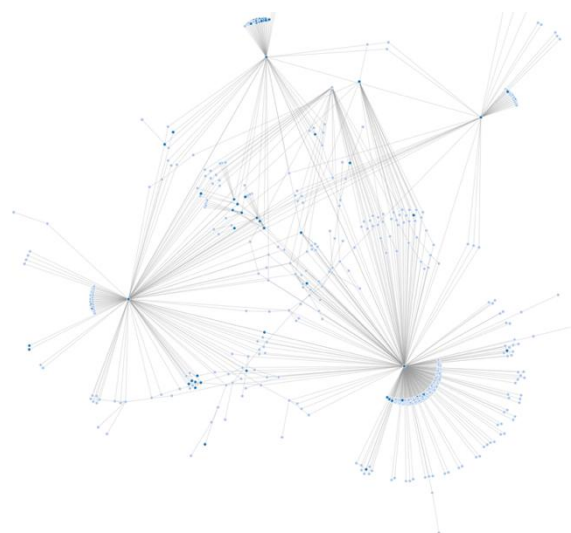in which standard parameters are used.

Figure 4. The force-layout technique of graph drawing in which for vertices and edges individual parameters are used depending on the number of edges at the vertex.

In Figure 3, the standard parameters for the vertices and edges of the graph are used: the charge values, edge lengths, and strengths are the same for all vertices and edges.

Obviously, the force-layout drawing technique simplified the perception of the graph, and spatially clustered some vertices.

Nevertheless, in view of the great connectivity, such a graph is still difficult-readable. For comparison, in Figure 4, the vertices and edges parameters, proposed in the technique, are used.

It is noticeable that the readability of the resulting image is ensured by the following:

1) The vertices, which have the greatest number of edges, tend to be located at a great distance from each other, forming a kind of skeleton of the graph, around which the remaining vertices will be located. Let us call such vertices as "skeleton vertex".

2) Vertexes with a single edge are grouped by semi-rings around the skeleton vertex associated with them, forming the "Hair Ball".

3) Other vertices are located between skeleton vertices, the posts of which they reposted. It can be seen that some of these vertices also form small groups if they reposted each other's posts.

An obvious advantage of such force-layout drawing parameters is a more visible spatial clustering of vertices.

Vertexes that have one edge and correspond to users or groups who repost posts from only one group are clustered by an arc near the skeleton vertex (usually, it is a group).

Vertices-users are clustered between common skeleton vertices. And skeleton vertices are located far enough from each other to distinguish clusters of common users between them.

We discuss the practical importance of clustering in the next section.

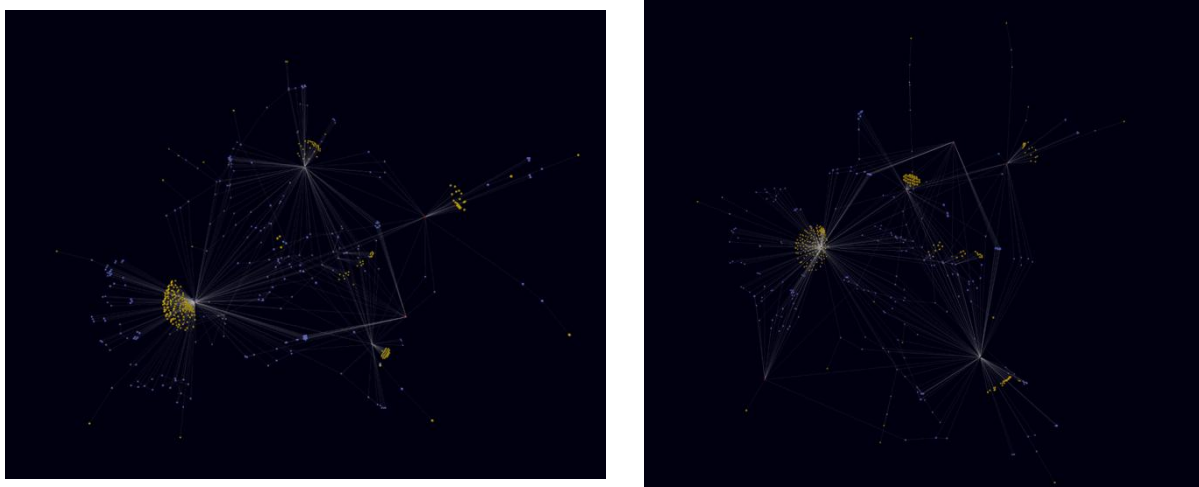The effectiveness of this technique of drawing is especially clear when the graph is depicted in 3D (Figure 5).

Figure 5. The force-layout technique of graph rendering in 3D in two angles;
the initial data used to construct the graph does not differ from the data
used for the graph in Figures 3 and 4.

Graph display in 3D allows to avoid overlapping edges on each other and even more spatially clustering vertices by using the third dimension.

It should be noted that the advantage of 3D rendering leveled by the inefficiency of human-machine interaction. For example, in 2D drawings (Figure 3 and Figure 4), scaling can be done by simply zooming of image, and in 3D drawing (Figure 5), one need to simulate a camera in 3D space that has 3 motor and 3 rotational degrees of freedom.

Thus, when zooming and navigating, the user is forced to use the mouse and keyboard, while often losing part of the image from view in view of unnatural camera control.

In order not to navigate in 3D space with the keyboard and mouse, one can use the graph visualization in augmented reality (Figure 6).
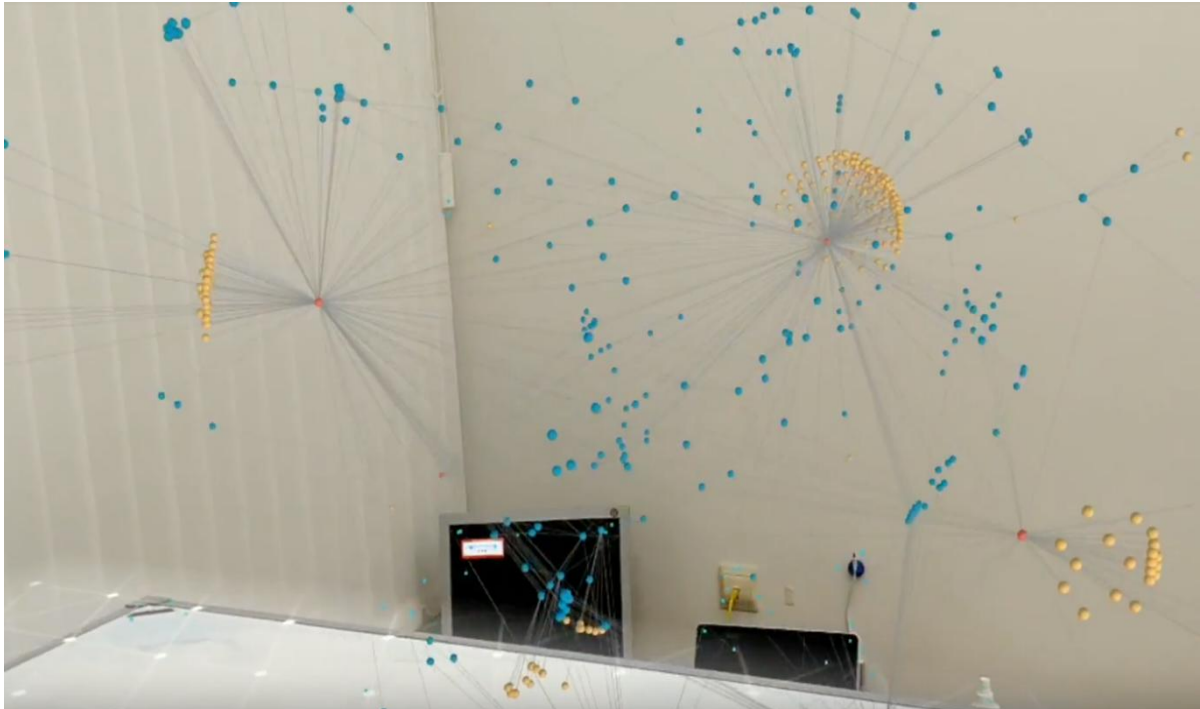
Figure 6. The force-layput technique of 3D graph drawing in augmented reality;
the initial data used to construct the graph does not differ from the data
used for the graph in Figures 5, 4 and 3.

Augmented reality makes the process of zooming and navigation more natural: a person physically approaches, moves away and turns his head, perceiving the graph as a real physical object.

Thus, several models are formed at once, which can be used for various tasks. For example, to navigate in large-sized graphs, one can use graphs in augmented reality, and to analyze local user groups, one can use 2D or 3D mode, depending on the degree of vertex connectivity.

4. Discussion

An obvious advantage of the proposed technique is the spatial clustering of the graph vertices into 4 categories (see Figure 7):

1) *The vertices of a graph with a large number of edges, forming a skeleton*. Some such vertices are highlighted by the red region in Figure 7. Usually, these vertices are representing big groups and they are the main sources of information dissemination. Smaller groups and users are clustered around this vertices.

2) *Clusters of vertices with one edge*. Usually, they are clustered by semi-rings around one skeleton vertex-group with which they are connected by this single edge. Some such clusters of vertices are identified by the blue region in Figure 7. These vertices are not the disseminators of information. Usually, these are endpoint users who reposted a single or several posts that were not further distributed. Nevertheless, such clusters of users create a coverage of the views of the post, and the size of the resulting arc can be used to judge the potential of this coverage.

3) *Groups of graph vertices that have edges or small sequences of edges with two or more skeleton vertices*. Usually, they are clustered between the skeleton vertices. Some such clusters of vertices are identified by the yellow region in Figure 7. These clusters typically represent users that are reposting posts from several public posts. This indicates their greater involvement in the analyzed topics.

4) *Small local user clusters*. Typically, this is a small number of users who are part of one of the previous clusters and at the same time repost posts from each other. Some such clusters of

vertices are distinguished by a green region in Figure 7. The fact that such users repost posts from the same groups and from each other may indicate that these users are familiar with each other.
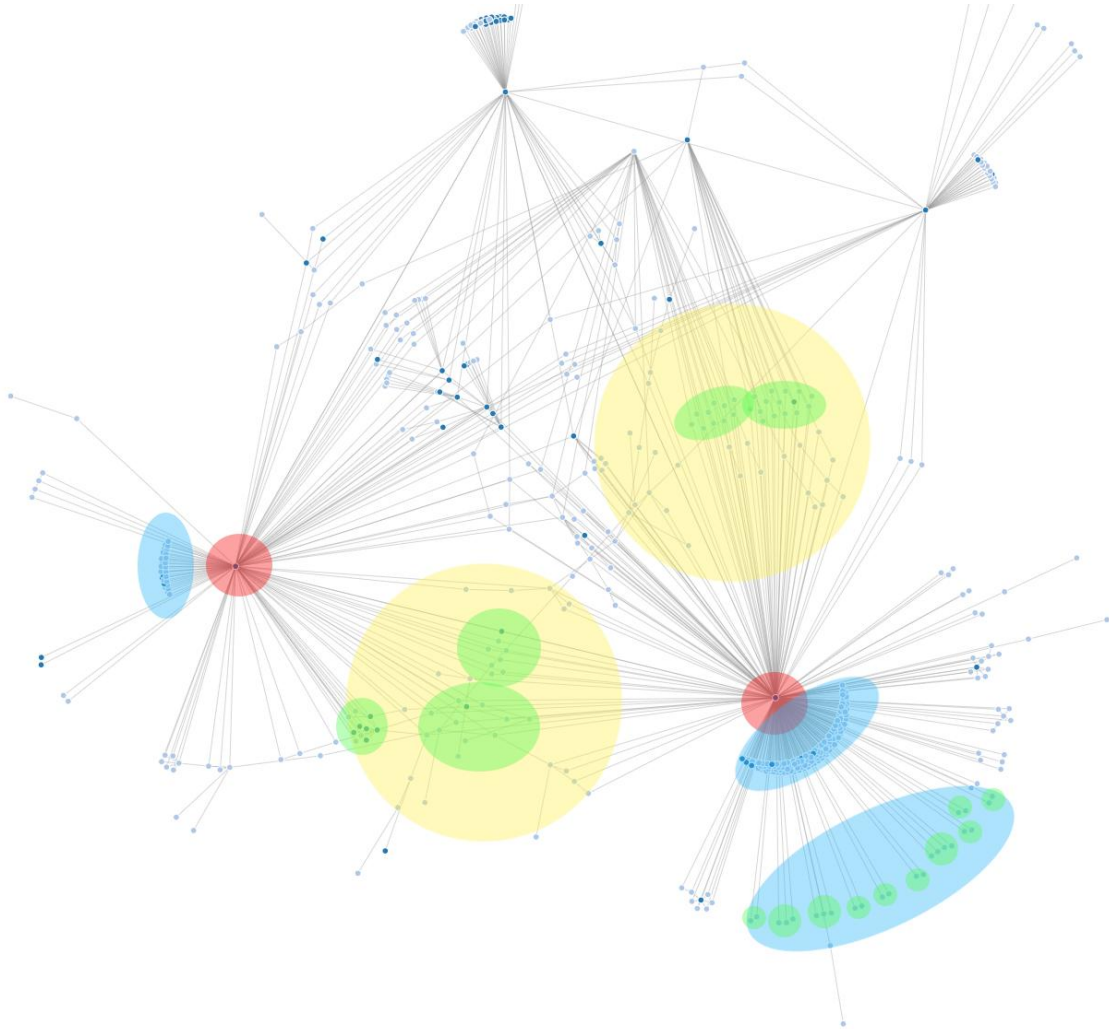

Figure 7. The force-layout technique forms four categories of clusters:
big groups that are main sources of information dissemination (red),
the endpoint users that do not distribute the information (blue),
involved users (yellow) and the potentially familiar users (green).

Thus, the vertices belonging to the first category are the primary sources or the most significant information distributors. The second category - the end users from whom the information is not further disseminated. The third category is users involved in the topic of group relations. The fourth category is local groups, probably familiar to each other.

At the same time, it is obvious that visual analysis of users of the third category is complicated. Through them, a large number of edges pass, and they are not so easy to distinguish in comparison with other categories.

Nevertheless, this disadvantage can be leveled by 3D mode and convenient computer-human interaction techniques that allow to locally change the layout of the graph and provide convenient scaling tools.

Thus, the proposed technique provides a simple way of visual analytics to identify user clusters and assess the extent of information dissemination, its potential coverage and the degree of social connectivity of users.

Examples of implementing the described visualization modes can be found on the following links:
1) for a 2D graph without using the proposed technique [15];
2) for a 2D graph that used the proposed technique [16];
3) for the 3D graph that used the proposed technique [17];
4) for the 3D graph that used the proposed technique in augmented reality (you need a mobile phone based on Android 8+ and Google AR Core support) [18].

5. Conclusion

The paper suggested the technique of visualizing data to analyzing social networks in order to identify and counteract inappropriate, dubious and harmful information.

The proposed technique is based on the force-layout technique of graph drawing in which the parameters of vertices and edges are calculated depending on the number of links.

As a result, the vertices are spatially clustered into 4 categories: big groups that are main sources of information dissemination, the endpoint users that do not distribute the information, involved users and the potentially familiar users.

The paper provides the example of using the proposed data visualization technique for social networks based on visual analytics of several groups in "VKontakte" in four display modes: without using a technique, in 2D, in 3D and in augmented reality.

The paper provides the evaluation that explains the meaning of the clusters on the picture which was obtained as a result of the experiment.

In future works, it is planned to explore the possibility of using other technique for structuring social network graphs, and also create computer-human interaction techniques that are necessary for visual analysis, which will effectively apply various layout techniques for different parts of one graph.

6. References

[1] Kotenko I, Chechulin A, Shorov A and Komashinsky A 2014 Analysis and Evaluation of Web Pages Classification Techniques for Inappropriate Content Blocking *Lecture Notes in Computer Science* **8557** pp 39-54.
[2] Zhao J, Cao N, Wen Z, Song Y, Lin Y and Collins C 2014 # FluxFlow: Visual analysis of anomalous information spreading on social media *IEEE transactions on visualization and computer graphics* **20** pp 1773-1782.
[3] Smith M, Shneiderman B, Milic-Frayling N, Mendes E, Barash V, Dunne C, Capone T, Perer A and Gleave E 2009 Analyzing (social media) networks with NodeXL *Proc. fourth international conference on Communities and technologies* pp 255-264.
[4] Ribarsky W, Wang D and Dou W 2014 Social media analytics for competitive advantage *Computers & Graphics* **38** pp 328-331.
[5] Chen S, Lin L, and Yuan X 2017 Social media visual analytics *Computer Graphics Forum* **36** pp 563-587.
[6] Kolomeec M, Gonzalez-Granadillo G, Doynikova E, Chechulin A, Kotenko I and Debar H 2017 Choosing Models for Security Metrics Visualization *Lecture Notes in Computer Science* **10446** pp 75-87.
[7] Kolomeets M, Chechulin A and Kotenko I 2016 Visualization Model for Monitoring of Computer Networks Security Based on the Analogue of Voronoi Diagrams *Lecture Notes in Computer Science* **9817** pp 141–157
[8] Balzer M, Deussen O and Lewerentz C 2005 Voronoi treemaps for the visualization of software metrics *Proc. 2005 ACM symposium on Software visualization* 165-72.
[9] Vehlow C, Beck F and Weiskopf D 2015 The state of the art in visualizing group structures in graphs *Eurographics Conference on Visualization (EuroVis)-STARs* **2**.
[10] Ghoniem M, Fekete J and Castagliola P 2005 On the readability of graphs using node-link and matrix-based representations: a controlled experiment and statistical analysis *Information Visualization* **4** pp 114-135.

[11] Craft B and Cairns P 2005 Beyond guidelines: what can we learn from the visual information seeking mantra? *Proc. Ninth International Conference on Information Visualisation* pp 110-118.

[12] Jusufi I 2013 Multivariate networks: visualization and interaction techniques.

[13] Novikova E, Kotenko I, Fedotov E 2015 Interactive Multi-view Visualization for Fraud Detection in Mobile Money Transfer Services *International Journal of Mobile Computing and Multimedia Commmunications* **6** 4 pp 72-97.

[14] Novikova E and Kotenko I 2018 Visualization-Driven Approach to Fraud Detection in the Mobile Money Transfer Services *Algorithms, Methods and Applications in Mobile Computing and Communications* IGI Global pp 205-236.

[15] Laboratory of Computer Security website. Implementation example: 2D graph without using the proposed technique. URL: http://comsec.spb.ru/files/forceWithout.html. Access date (02.09.2018).

[16] Laboratory of Computer Security at SPIIRAS website. Implementation example: 2D graph with using the proposed technique. URL: http://comsec.spb.ru/files/forceWith.html. Access date (02.09.2018).

[17] Laboratory of Computer Security at SPIIRAS website. Implementation example: 3D graph that used the proposed technique. URL: http://comsec.spb.ru/files/force3D.html. Access date (02.09.2018).

[18] Laboratory of Computer Security at SPIIRAS website. Implementation example: 3D graph that used the proposed technique in augmented reality (you need a mobile phone based on Android 8+ and Google AR Core support). URL: http://comsec.spb.ru/files/forceAPK.html. Access date (02.09.2018).

Acknowledgments