Report for the 2nd stage of the implementation of the project of the Russian Science Foundation No. 18-71-10094 "Monitoring and counteraction to malicious influence in the information space of social networks"

The development of information technologies, the global Internet, social networks and mass communication leads the scientific community to radically revise their ideas about information attacks and how to defend against them. Not only in scientific area, but also in the media, the aspect related to national security is being discussed, namely: information and psychological security, information wars and information intervention. So, as never before, the question is urgent of developing methods to combat the deliberate dissemination of theories of conspiracy, terrorism, HIV dissent, and many other destructive ideas and their propaganda.
The greatest acuteness of this problem manifests itself in attempts to recognize targeted informational influences (attacks) on subjects (individual or collective, for example, on a person, family, social group of people or organization) and to counter such attacks.

This project "Monitoring and counteraction to harmful influence in the information space of social networks" can be divided into 2 main interconnected parts - monitoring and counteraction.

As part of the monitoring, the collection, pre-processing and semantic analysis of information objects is carried out, as well as the identification of sources, distribution channels and target audience of malicious information impact. In order to successfully counteract such an influence, it is necessary to take into account possible countermeasures, ways to identify the countermeasures goal, evaluate its effectiveness and support decision making when choosing the most promising countermeasures available.

The first step in analyzing a social network is collecting data. While working on the project, a software prototype with a graphical user interface was developed to collect, preprocess, and save heterogeneous social network objects.

To analyze the data collected, several prototypes were developed: a prototype for semantic analysis (analyzes the text using machine learning methods), a prototype for attack analysis (analysis of the source of the attack and who it is aimed at) and a prototype for analysis of distribution channels (analysis of how the attack is distributed).

To analyze the text content of social networks, a prototype component of the semantic analysis of malicious information objects in social networks was developed using data mining methods. The component classifies texts that are contained in potentially harmful posts on social networks using training methods based on positive and unlabeled data (eng. Positive-unlabeled learning) and multiclass classification. The developed prototype was experimentally studied on a dataset that contains posts on six potentially malicious topics and text information from social network’s posts on a random safe topic. Comparing the results of the experimental study allows us to draw conclusions about the sufficiently high quality and stable operation of the proposed approach to detecting malicious information objects.

Appropriate techniques and a software prototype have also been developed for attack analysis. An information attack always contains its source and the target audience as its main components. The difference between these components is the time of occurrence of malicious information objects – objects placed by the source fall within the time interval set by the expert [t0, t0+e], while other objects that received access to malicious objects or participated in their

distribution after t0+e belong to the target audience. The algorithm for analyzing the source that has a spontaneous nature segments its audience by the degree of activity, thus allowing us to get closer to the accounts that are the instigators of information impact. The algorithm for analyzing the target audience gives an idea of the gender and age of the target audience and its location. The proposed algorithms are implemented in a software prototype, and the result of their work in the form of graphs and pie charts is available in the graphical user interface.

For the analysis of information distribution channels, a prototype of their analysis was developed. The distribution channel is the audience, which, although it is not the initiator of the attack, but contributes to its distribution. This audience can be represented in the form of a network where users and groups that disseminate information are connected to each other. Two modules were developed for analysis: parallel processing and visualization. The parallel processing module using the GPU analyzes the network structure and allows one to calculate: key network nodes (opinion leaders), channel width, channel framework, social circles of the channel and estimate the attenuation of information in the channel. The visualization module allows one to visualize the channel in the form of a 3D image, so the operator can visually evaluate it. These modules were implemented as software prototypes and were also tested using the analysis of 5 different distribution channels as an example.

Data collection, text analysis, analysis of attacks and distribution channels are links of one chain, called monitoring the information space of social networks. A prototype was developed for it, combining the above prototypes into a single system. The monitoring stage consists of a tracking phase, where social connections are monitored and potentially malicious information objects are discovered, and a dangerous influence detection phase, where a graph of information distribution channels is constructed and measures of centrality are calculated, the source of the attack is separated from the target audience, segmentation and statistical analysis are performed. The developed software prototype implements all stages of the monitoring component and allows the operator to control the entire process from detecting an information attack at the data collection stage to visualizing and downloading the results of their analysis.

In order to counter attacks on social networks, it was necessary to determine:
1) how can one counteract an attack?
2) how to determine the objects of counteraction?
3) how to determine for what object what countermeasure should be applied?
4) how to evaluate the effectiveness of counteraction?

For this, appropriate methods, models and techniques have been developed.
Countermeasures are diverse and can be classified according to various criteria. Developed classification allows us to divide countermeasures based on their object, type, method, breadth and executor of impact as well as the group to which this countermeasure belongs. The grouping of countermeasures is based on the expected result, namely: switching of attention, blurring of attention, discrediting, reducing the intensity or blocking of the object of impact. Based on the proposed classification, we developed a countermeasure model. The generated list of countermeasures is a single table, where rows are showing examples of countermeasures for each of the selected groups and columns are showing their place in the proposed classification. It is assumed that this list will allow to determine countermeasures targets as well as offer to the system operator the most effective of them.

In order to choose the goals of counteraction, an appropriate method was developed. The project explored the structure of social networks and interactions between elements, methods of disseminating information and measures aimed at counteracting them. Based on the data studied, a model of target selection for the implementation of measures to counter malicious information in social networks is constructed. A step-by-step methodology for assessing complexity for the implementation of countermeasures is proposed, which allows ranking measures according to the degree of complexity of their implementation. At the same time, the system configuration is quite flexible, the operator can be both the executive authority and the parent protecting the child from the effects of malicious information. Depending on who the system operator is, he adjusts it initially using the methodology for selecting complexity factors based on expert estimates. Also, in the project for optimal application of the selected measures, a methodology for assessing the priorities of objects that will be affected in the process of combating malicious information is proposed. The set of disparate methods is combined into a single method that determines the sequence of methods and the formation of the final overall assessment of the counteraction.

After forming a list of countermeasures and a list of goals, the system operator needs to decide which countermeasures and to whom they should be applied. For this, a method was developed that allows one to provide countermeasures and impact targets in the form of an image - a matrix. The technique includes three methods:
1) The method of creating a visualization model - the formation of a data model necessary for visualization.
2) The method of rendering the visualization model in the form of a matrix. In this matrix, rows are objects, columns are measures, and cells are efficiency indicators.
3) Methods of visual selection of countermeasures by the operator. Using this method, the operator filters and sorts the elements of the matrix, and then chooses to whom what measures should be applied.
Thus, a method was developed that forms an image, with which one can choose exactly which countermeasures should be applied.

In order to evaluate the effectiveness of countermeasures, methods, models and techniques have been developed to evaluate the effectiveness and efficiency of countermeasures against harmful effects in social networks.
By efficiency, in the context of this study, we will understand the relationship between the achieved result and the resources used.
The first developed method of estimating effectiveness of countermeasures based on the use of the different characteristics of information objects and users, for example, the increase in the number of views, the increase in the number of likes, the increase in the number of comments, the increase in the number of reposts, etc.
The second developed method for evaluating the effectiveness of counteraction measures is based on evaluating the user's state at some point in time, since the ultimate goal of countering malicious influence in social networks is to minimize the user's interest in malicious content.
The proposed metrics and methods for evaluating the effectiveness of measures to counteract malicious influence in social networks are proposed to be used as part of an integrated method for evaluating the effectiveness of measures to counteract malicious influence in social networks. At the first step, the effectiveness of counteraction measures for each class of goals (a separate information object, user, or group of users) can be used to evaluate the expected effect of the applied countermeasure. To assess the user's state in relation to malicious content and form final conclusions about the effectiveness of the applied measure to counteract

malicious influence, it is proposed to use the second method and the social network user's state indicator.

All developed methods must be combined into a single method, which will allow for information objects to determine countermeasures. For this, a comprehensive counteraction method was developed, which combines the previously obtained methods and includes:
1) Initialization of initial values (system setup).
2) The choice of countermeasures in 5 steps: obtaining possible goals based on monitoring data; identification of all possible goals, countermeasures and assessments; calculation of efficiency before countermeasures are applied; selection of a list of combinations of objects of and countermeasures from all possible; application of countermeasures selected by the operator.
3) Methodology for evaluating the effectiveness of countermeasures - assesses how effective the measures taken were and generates a report.
An interaction model was also developed, which is a set of modules that are executed in turn at each step.

Also, as part of the project, there were presented the results at 5 conferences in Russia and abroad. 4 publications are included in the citation indexes of Scopus and / or Web of Science and 8 publications are included in the RSCI system. 4 computer programs registered. For the next stage of the project (2020-2021), two defenses of dissertations are planned for the degree of candidate of technical sciences using the results of this project.

At St. Petersburg State University of Telecommunications named after prof. M.A. Bonch-Bruevich conducted a practice course for masters of the 1st semester 2019/2020, which used the results obtained in the framework of the project.