Social Networks: Analysis, Algorithms and Their Implementation

Svitlana Popereshnyak, Iryna Yurchuk

Taras Shevchenko National University of Kyiv, Bohdan Hawrylyshyn str. 24, Kyiv, UA-04116, Ukraine

Abstract

In this research the main types of social networks and methods of their analysis are considered. The types of ties and analysis of the problems of ties in the social network are analysed. The relationship between graph theory and social network analysis was studied and established. Typical database was developed. The concepts of graph theory "centrality" and "intermediate centrality" were used to build a prototype based on the proposed method of analyzing ties in cyberspace.

Keywords [1]

Social network analysis, connections in social networks, social graph, graph theory, centrality.

1. Introduction

The digital nature of information facilitates the ability to obtain data about networks. The notion of communication systems and relationships was formed long before the invention of the Internet. However, the collection of personal data is a time-consuming and difficult procedure; people sometimes do not understand who is in their personal network (or how strong this or that connection), and the researcher needs to collect accurate data on interactions. These problems can be minimized through online research because in them the information is digital and encoded through the act of sending a message or adding one using the functionality of the website [1]. Also, in digital social networks, it is easy to copy messages for further analysis.

Methods of mathematical statistics, graph theory, differential equations are used to study social networks. Social media analysis is closely related to graph theory and big data analysis.

The analysis of social networks is based on the mathematical theory of graphs, as well as empirical research in the field of social psychology and anthropology. While the first group of scientists discovered various laws for constructing abstract nodes and lines, the latter found that nodes and lines are the most convenient to denote the relationship between people. Since both groups of researchers worked at the same time (in the second half of the twentieth century), they agreed to develop a series of metrics and methods to identify the basic structures of complex empirical phenomena.

Currently, the analysis of social networks is one of the most intensively developing areas not only in sociology, but also in other humanitarian and technical disciplines. The interest in them is dictated by the fact that the fact of the special position of this object of research is becoming quite obvious, entailing a new set of explanatory models and analytical tools that are outside the framework of conventional research methods - both quantitative and qualitative.

A social network as a way of organizing social knowledge requires a special methodological approach that differs from traditional methods of analyzing sociological information.

Purpose of work - build a prototype of software that finds the power of connections between social network actors.

The aim of research – prototype of the method of intersubjective connections analysis in social networks.

COLINS-2021: 5th International Conference on Computational Linguistics and Intelligent Systems, April 22–23, 2021, Kharkiv, Ukraine

EMAIL: spopereshnyak@gmail.com (S. Popereshnyak); i.a.yurchuk@gmail.com (I. Yurchuk)
ORCID: 0000-0002-0531-9809 (S. Popereshnyak); 0000-0001-8206-3395 (I. Yurchuk)

2. Background

In this section main concepts and mathematical terms are considered.

2.1. Networks and its analysis

A network is a set of nodes (such as people, organizations, websites, or government entities) and the relationships between them (oriented or not). For example, for the postal network is the principle to order the nodes (senders and recipients). A social network organized by the software is not usually an oriented network of friends (users).

Network analysis makes good results when all network nodes belong to the same object class. To explore more than one type of object (such as bloggers and commentators), "two-level analysis," which involves your own set of distinctions have to be used.

There are the following types of networks:

- Solid networks (e-mail, mailing lists, and social networks of the Internet space): links are clearly defined, it is easy to form a group structure and identify specific users. They are built using lists, which allows you to reach any user;
- Partial networks: they are a compromise between the desire to cover a whole network and the fact that some whole networks are just too massive to cover them completely. The researcher can start with a single web page or several pages (so-called "sowing"), then he will look for pages related to that sowing, and then pages related to these pages. The sampling process ends when a sufficient number of pages have been coll ected; when all possible pages are collected; or when the sample meets certain criteria (for example, when all pages with more than 400 words are collected);
- Ego networks: they are represented by a spontaneous sample of users, the researcher can collect data or a star-shaped network (ego-node and its connections with other nodes), or a complete ego-network (which also includes connections of other nodes with each other). Data collection on ego networks can be based on the already available results of various research techniques and interviews;
- Social networks: people is together in pre-established interpersonal relationships such as kinship, friendship, classmates, colleagues, business partners, etc. A connection is built one at a time. The main reason for people to join a social network is to maintain old relationships and create new ones to expand their network. There are social networks for communication (Facebook), for sharing media content (Instagram), for collective bargaining (Quora), for authoring (Twitter), social bookmarking services (Pinterest), for interests (Goodreads).

In Figure 1, three types of networks (solid, partial and ego networks) are illustrated.

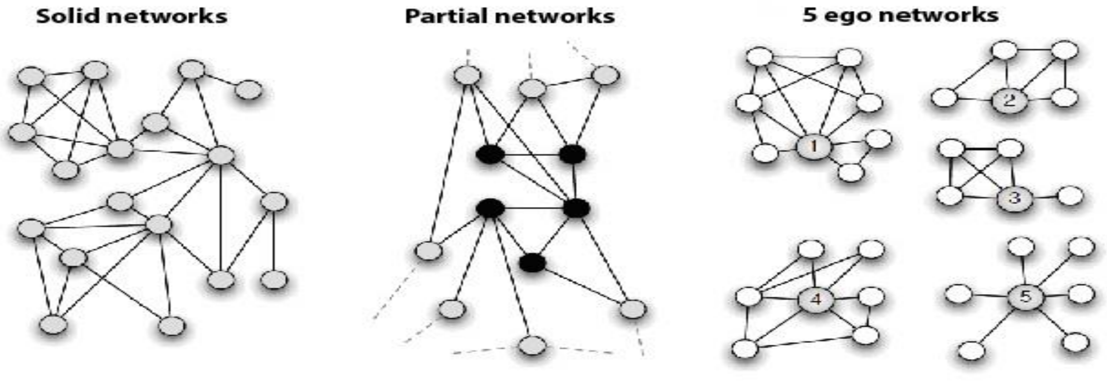

Figure 1: The examples of solid, partial and ego networks are shown

Social networks analysis (SNA) is effectively used to combat money laundering, identity theft, network fraud, cyberattacks, and others. In particular, SNA techniques were used in the investigation of illegal securities transactions conducted by the Australian Securities and Investment Commission.

The author of some books and an expert in security and data analysis, Jesus Mena, called SNA as "a technique of data mining that represents their structure in the form of interconnected objects" [2]. This technique is based on mathematical disciplines such as graph theory and matrix algebra, and provides analysts with tools that can model and study the structure of relationships between different objects.

According to the basic theories, there are two main tools of SNA. They are the matrix and the relationship diagram. Accordingly, the algorithm of the analyst's work with the collected data includes the following stages: the creation of a matrix (or matrices) of connections, construction of a diagram of connections (graph), analysis of connections.

Many researchers have paid attention to the issue of studying social networks [3-8].

For example, social network analysis approach [5] is a way to quantify the social and intelligent behaviors as displayed by animals in social structures. The study analyzes some of such network measures like modularity-based algorithms for community formation, intergroup and intragroup dynamics.

2.2. Basic concepts of graph theory

A vertex degree is a number of edges of a graph G that are incidental to a vertex v. During calculating, a loop is counting twice [9]. A vertex degree is denoted by $\deg(v)$. The maximum and minimum degrees of the vertices of the graph G are denoted by Δ(G) and δ(G), respectively.

The sequence of degrees of the vertices of a non-orientative graph is a non-increasing sequence [8] The sequence of degrees of vertices is an invariant of the graph, so for isomorphic graphs, it is the same. However, the sequence of degrees of vertices is not a unique characteristic of the graph: in some cases, nonisomorphic graphs also have the same sequence [1].

The problem of degree sequences is to find some or all graphs with a given non-increasing sequence consisting of natural numbers (zero degrees can be ignored because their number is changed by adding or removing isolated vertices). A sequence that is a sequence of powers of any graph is called a graphical sequence. It follows from the formula for the sum of powers that any sequence with an odd sum (such as 3, 3, 1) cannot be a sequence of powers of a graph. The opposite is also true: if a sequence has an even sum, it is a sequence of degrees of the multigraph. The construction of such a graph is quite simple: it is necessary to combine the vertices of odd degrees in pairs, to the left unfilled vertices must be added loops

It is more difficult to implement a simple graph with a given sequence. Erdos-Gallai theorem states that the non-increasing sequence d_i (for i = 1,…,n) can be a sequence of a simple graph under certain conditions [10].

According to Havel-Hakimi algorithm [10] if non-increasing sequence (d_1, d_2, …, d_n) is a degrees sequence of a simple graph, then ($d_2 - 1$, $d_3 - 1$, …, dd1+1 − 1, dd1+2, dd1+3, …, d_n) is some sequence of degrees of a simple graph. This fact allows building a polynomial algorithm for finding a simple graph with a given sequence.

Intermediate centrality measures the number of times a node acts as a bridge on the shortest path between two other nodes [11].

Power centrality of vertex v for a graph $G \coloneqq (V,\ E)$, where $|V|$ is a set of vertices and $|E|$ is a set of edges, can be denoted as

$$C_D(v) = deg(v).$$

The calculation of the degree of centrality for all nodes in the graph takes $\Theta(V^2)$ at a dense adjacency matrix of the graph representation and $\Theta(E)$ at a sparse matrix. The definition of centrality at the node level can be extended to the whole graph, in this case, we are talking about the centralization of the graph. Let $v*$ be a vertex with the most power of centrality. Let *X*:=(*Y*, *Z*), where $|Y|$ is a connected graph which maximized the following number (with y* vertex, which is the most power of centrality at *X*):

$$H = \sum_{j=1}^{|Y|}[C_D(y*) - C_D(y_i)].$$

Then the degree of centrality of a graph *G* equals to

$$G_D(G) = \frac{\sum_{j=1}^{|Y|}[C_D(v*) - C_D(v_i)]}{H}.$$

A value H is a maximum if a graph X contains a unique central vertex such that all other vertices are connected with. In this case *H*=(*n*-1) (*n*-2).

3. Program design

3.1. Finding ties

To find a connection between different people in this work uses a social network in which a person is registered and he has a list of friends. To connect all human friends, the API is used to parse all friends (A) of a person, and all friends of friends (B). After that, the program removes from set B all the friends who are not in set A. Then the program looks for how a friend is connected with other friends and gives information in the form of a list (see Fig. 2).

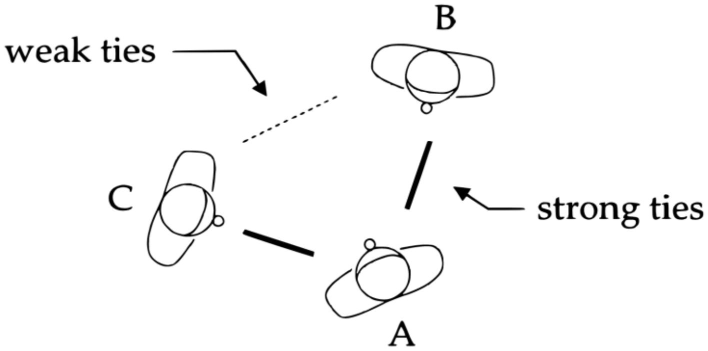

Figure 2: The connections ties

The "weak ties hypothesis" states with help of the probability theory and mathematics it can be used to predict that if A is related to B and C, then there is a strong possibility that B and C are also related [12].

3.2. Determination of weak and strong ties

The powerful factor for the tie construction is a relationship between friends and friends of friends. The connections between them are the most important thing to create a graphical representation of a graph of friends.

To detect the strength of communication (weak or strong), the program will make a comparison between all friends on the following parameters:

- School - the system will compare data of friends and find friends:
  a. who study in the same schools,
  b. go to school, which is located in the same region in a big city,
  c. or all schools from a small town (where 2-5 schools) and shops,
- Work - the program compares data of friends and finds friends:
  a. in profiles whose job data match,
  b. or companies, where they work, are located in the same city and work in the same field,
- Communities - the program will compare groups of friends and search for shared groups that both have. It is also a criterion for determining the strength of the connection between a user.
- Music (selective) - this criterion for those people who want to find among their friends people who have common favors in music.
- Photo - the program will analyze the number of tags in the photo of certain person, in the photos which are in the profiles of friends. The weight of these connections is calculated.
- Information that meets the above criteria is broken down into a database. And then the system has a complete database with information about all friends from the list of friends in a social network. Then the collected data about each friend will be compared with others. There is a gradation of bond strength: weak ties, medium and strong ties.

Weak ties according to the system of criteria are those ties that will correspond to 0 or 1 criterion.
Medium ties according to the system of criteria are those ties that will correspond to 2 criteria.
Strong ties according to a system of criteria are those that will correspond to 3 or more criteria.

This is not a definitive system of criteria. There are many other variations that can be used to strengthen and define relationships.

3.3. Database design

There are the following basic elements: Person, Playlist, Song, ListCommunities, Community, Photo, Album, Tag (see Figure 3).

The Person table is used to store all users who can be parsed using the API. All users' data will be compared with others.

The Playlist is the connecting table between the Person table and the Songs table. For example, each person has music that is in a playlist.

The Song table describes the song in the playlist for future comparison of the two playlists. Usually, everyone has albums that have photos, but often there can be many such albums.

The album table gives a foreign key to person and has only one user field. Each photo in this Photo table is described by three fields: idphoto, name, date.

The tag table has information about each tag of our user in the photos of other users.

The ListCommunities table is the same as the Album or Playlist. Everyone can visit some communities. There are many communities, but the list of communities that a particular person attends is unique, even if there are two people with the same communities on their list. The information in this table is also used to find the strength of the links.

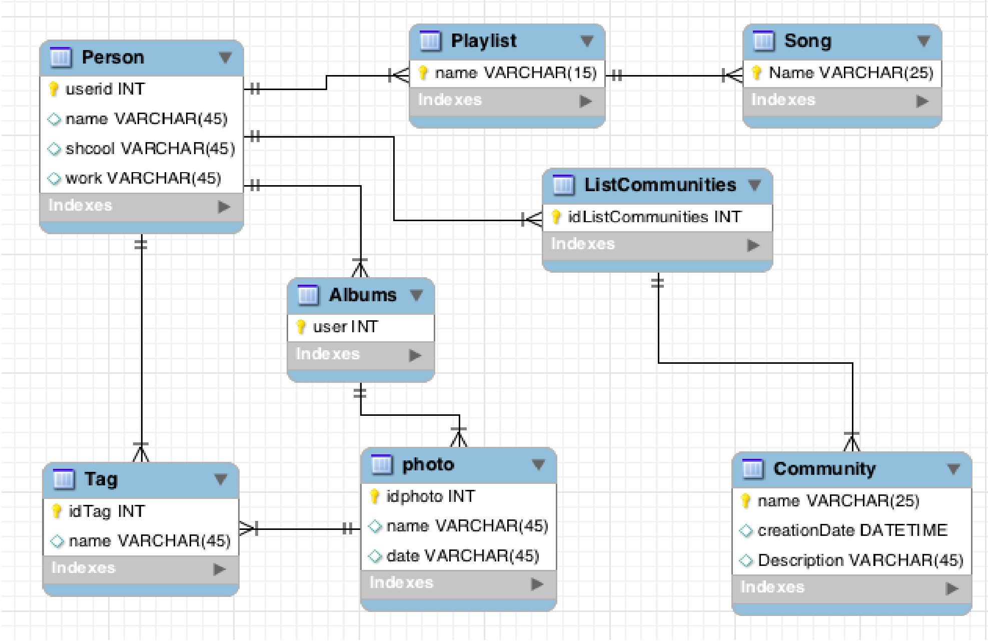

Figure 3: The database schema

The Community table is a table that describes the community and has three fields: name, creationDate, Description.

Basic tools of development

RabbitMQ is a platform that implements a messaging system between software system components based on the AMQP (Advanced Message Queuing Protocol) standard. Uses the standard AMQP (Advanced Message Queuing Protocol), supports horizontal scaling to build a cluster architecture, supports data storage on disk, support for HTTP, XMPP and STOMP, is the implementation of clients to access RabbitMQ for some programming languages: Java, .NET, Perl , Python, Ruby, PHP, etc., there are various plug-ins (such as a plug-in for monitoring and management via HTTP or web interface or a plug-in "Shovel" for transferring messages between brokers).

NetworkX Python library for studying graphs and networks. NetworkX is free software that is distributed under a BSD license. Gives you the ability to convert graphics from multiple formats, build random graphs or build them gradually, find subgraphs, clicks, K-cores, explore contiguity, degree, diameter, radius, center, intermediate, etc., draw networks in 2D and 3D.

Matplotlib is a Python programming library for data visualization with 2D graphics (3D graphics are also supported). The resulting images can be used as illustrations in publications.

Matplotlib is a flexible, easily configurable package that, along with NumPy, SciPy, and IPython, provides features similar to MATLAB. The package currently works with several graphics libraries, including wxWindows and PyGTK.

4. Model of finding leaders in the social network

The task of finding leaders is one of the main in the study of social networks. It is closely related to the task of finding clusters in a graph, namely, some communities that are similar in common. The main steps are to recognize leaders in accessible social graphs, the introduction of metrics to combine different characteristics and take into account the importance of each metric for the problem.

4.1. Functional model of the system

Let consider the functional model of the system, which can be represented graphically using the context diagram IDEF0 for "Search for social network leaders"(Fig. 4).

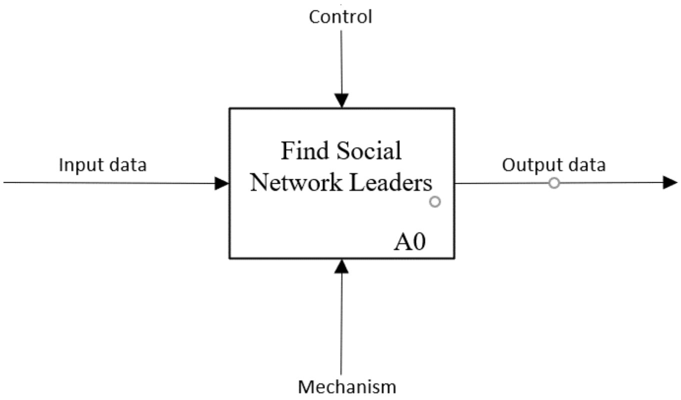

Figure 4: The context diagram

First, we will make a general description of the system, and then perform a functional decomposition. In Fig. 5 the functional model of the system with external links is shown.

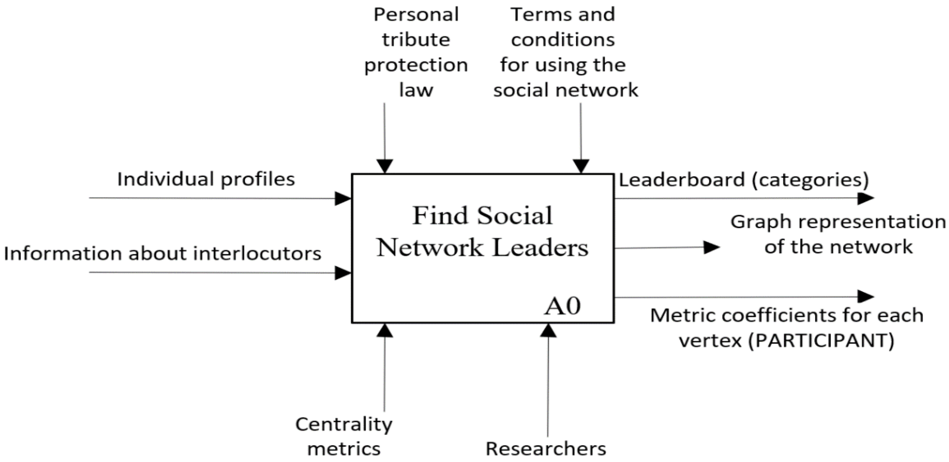
Personal tribute protection law
Terms and conditions for using the social network
Individual profiles
Information about interlocutors
Find Social Network Leaders
A0
Leaderboard (categories)
Graph representation of the network
Metric coefficients for each vertex (PARTICIPANT)
Centrality metrics
Researchers

Figure 5: Search for social network leaders

In Fig. 6 a decomposition of previous diagram on three blocks with consecutive direct is presented

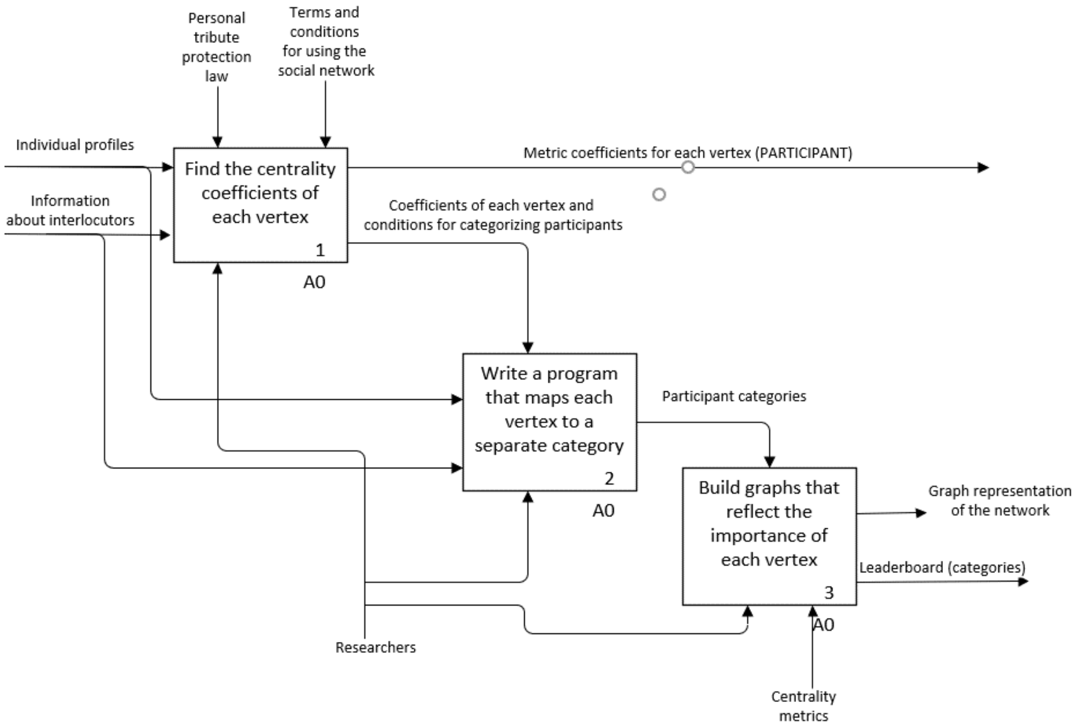
Personal tribute protection law
Terms and conditions for using the social network
Individual profiles
Information about interlocutors
Find the centrality coefficients of each vertex
1
A0
Metric coefficients for each vertex (PARTICIPANT)
Coefficients of each vertex and conditions for categorizing participants
Write a program that maps each vertex to a separate category
2
A0
Participant categories
Build graphs that reflect the importance of each vertex
3
A0
Graph representation of the network
Leaderboard (categories)
Researchers
Centrality metrics

Figure 6: Decomposition of the first level of the functional block "Finding leaders of a social network"

As the result of applying IDEF0 to the current system, we obtain a model of this system, consisting of a hierarchically ordered set of diagrams.

4.2. Description of the program

The program is made by the Jupyter Notebook environment in a full-featured Python programming language.

The main elements of the program consist of two files.

The first file is Neural_network_graph.ipynb. This module loads the graph of the oriented neural network of neurons and synapses of the C. elegans worm using the built-in functions of the NetworkX library, determines the centrality coefficients for all metrics considered and sets the necessary parameters for visualization using Gephi. The exponential centrality metric is considered first, followed by the proximity metric, the intermediate metric, the eigenvector metric, the PageRank metric, and the load metric. After calculating all the places for each vertex, the values of the vertices of the compatible metric are determined. To visualize, you must specify the size and color of each vertex. The greater the coefficient of centrality, the brighter the color of the top and has a larger size. After analyzing the results obtained for each metric, the maximum and minimum significance coefficients of the vertices were determined.

The second file, called Gnutella_graph.ipynb, looks at an oriented neural network graph with 8,717 vertices and 31,525 edges. The difference from the previous file is that the vertices were not assigned a color or size, but only the necessary calculations were performed to determine the coefficients of the compatible metric.

Preservation of the result: since the analysis is for research purposes, a pickle will suffice.

Before saving dictionaries, it would be logical to delete the keys whose values are None. These are blocked or deleted accounts. Simply put, these id will be present in the column, because they have someone in the friends, well, the program will save on the number of keys in the dictionary. The saved result must be unloaded so as not to collect again ID.

For drawing a graph, Figure API and Draw Networkx are used. In Fig.7 a graph of mutual friends, a total of 306 vertices and 2,096 edges is shown.

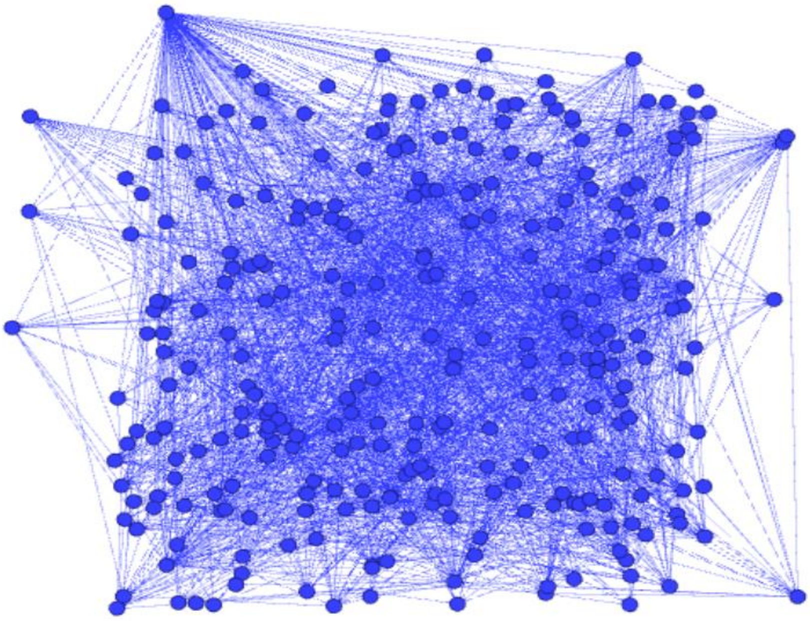

Figure 7: Visualization of friends as a graph

4.3. The results of the experiment

To analyze the selected network, an indicator metric was the first selected that calculates the number of edges incident to each vertex, and then normalizes the value obtained for each vertex. To identify leaders in the network, the nodes with the highest coefficient of significance in terms of metrics have a larger size and brighter color. Yes, the first group (with the highest odds - leaders) is red, the second - yellow, the third - green, the fourth - blue and the fifth - without color (with a label).

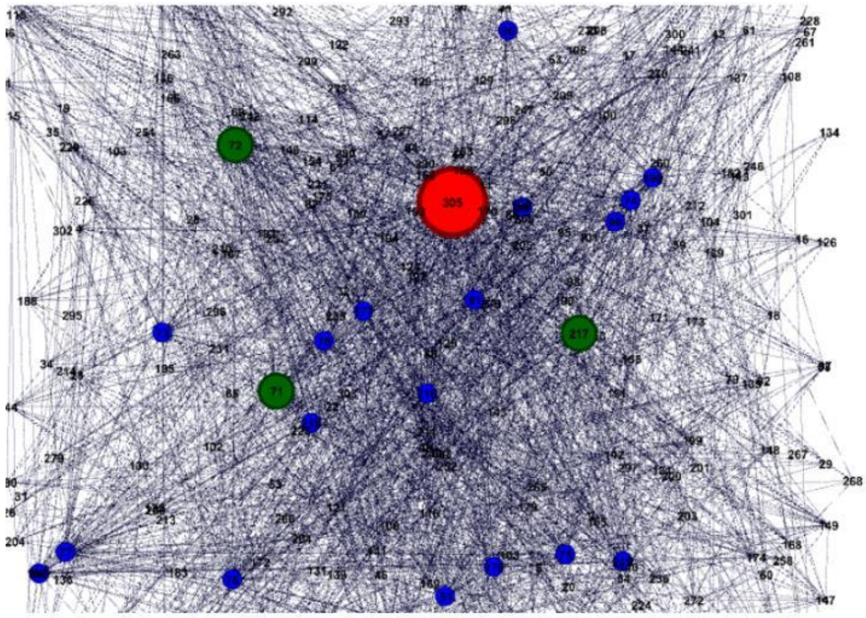

Figure 8: Leaders of the social network by indicator centrality

The first group includes vertices with coefficients from 0.4 to 0.5, the second - from 0.3 to 0.4, the third - from 0.2 to 0.3, the fourth - from 0.1 to 0, 2 and to the fifth all other vertices, the coefficient of which is less than 0.1. As can be seen from the figure, the leader with a coefficient of significance equal to 0.45085 was the vertex with the number 305. None of the vertices fell into the second group, but the third group received three at once - 71 with a coefficient of 0.28136 and 72nd with a coefficient of 0.27119, as well as the 217th with a coefficient of 0.20339. Therefore, it can be concluded that these above nodes have the most connections with other vertices. But these links do not reflect their quality and importance to the graph, but only their quantity.

So, next for consideration is the proximity metric, in which all the vertices were also divided into groups, the colors for the groups were chosen in the same way as for the indicator centrality. The results are shown on Fig. 9.

The only difference is that there are no groups with coefficients between 0.3 and 0.5, these groups did not get any of the vertices. The maximum coefficient was again at the top of 305 and was equal to 0.57578. Blue peaks have a coefficient from 0.1 to 0.2, and from 0.2 to 0.3 - green. As you can see, the vertex 305 not only has many connections to others, but it is also well suited for data transmission.

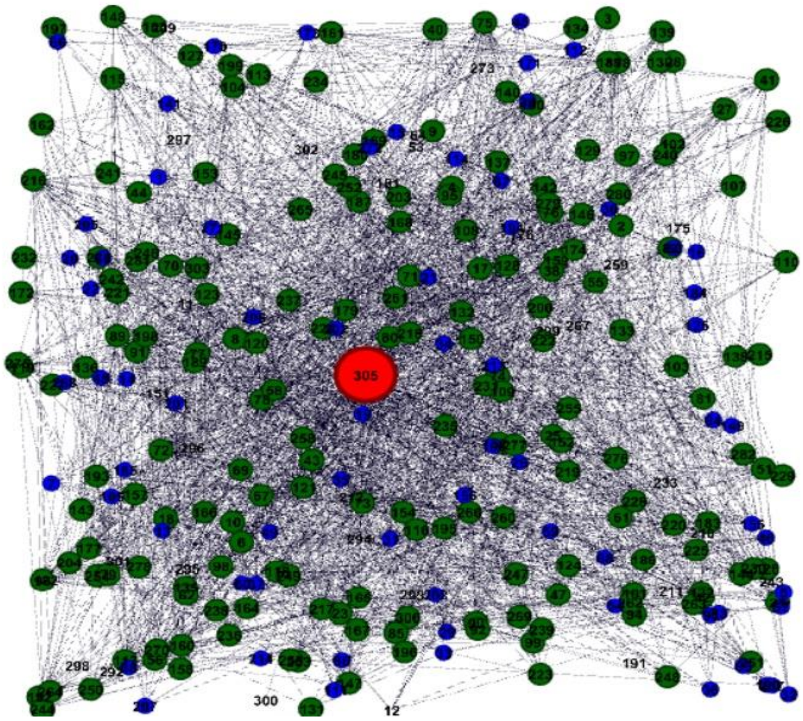

Figure 9: Social network leaders on proximity metrics

Continuing the study, we obtain that the peak with the maximum coefficient was again the 305th (significance factor - 0.12464). The high significance factor for this metric indicates that the vertex has a high impact and popularity on the network, because it is indicated by many other "quality" vertices. However, for the load metric and the intermediate vertex 305 had extremely low coefficients, so it did not fall into the "top five" for the compatible metric.

5. Networks as applicable to public health

It is not easy to use apply networks for public health, since there is a compromise between people and their relationships. There are many basis to construct ties as edges, for example, sexual, emotional, transactional, and so on.

If we use social networks for modeling infectious disease transmission, some requires not only the parameters of what causes disease to spread (and not spread) but also a representative network of individuals are needed. For creating this representative network, data are required. We will assume true network census data are not available and data are thus egocentric.

6. Conclusions

Social networks are an important component of communication in modern life. Their analysis is necessary both to understand the current situation and to improve some aspects of life. As a result of the study, a prototype of the method of analysis of intersubjective connections of the social network was built. During the work, the analysis of search and establishment of the type of connections and research of social network problems was carried out. The relationship between graph theory and social network analysis was studied and established. As part of this work, the problem of analysis of ties in social networks was considered and the main methods of its solution were determined with the help of graph theory. Using the concepts of graph theory "centrality" and "intermediate centrality", a method of analyzing connections in cyberspace was proposed. This method was used to build a prototype for social network analysis. The RabbitMQ, NetworkX, and Matplotlib ties algorithm include depth setting, data sending, response retrieval, and optimization.

Further research involves the neurosophic approach as a generalization of fuzzy concepts in the graph, which will allow a deeper interpretation of the relationships and their quality.

7. References

[1] S Popereshnyak, O Suprun, Tools and methods for intersubjective relationships in cyberspace forecasting, in: Proceedings of the 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT ’2017, Lviv, Ukraine, 2017, pp. 244-247. doi:10.1109/STC-CSIT.2017.8098779.

[2] J. Mena, Investigative Data Mining for Security and Criminal Detection. Butterworth Heinmann, 2003.

[3] Q. Han, M. Gu, L. You, F. Miao, Rumor Spreading with Cross Propagation in Multilayer Social Networks, in: Proceedings of the Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking, ISPA/BDCloud/SocialCom/SustainCom ’2019, Xiamen, China, 2019, pp. 1641-1645. doi: 10.1109/ISPA-BDCloud-SustainCom-SocialCom48970.2019.00242.

[4] M. Pruthvi, S. Karthika, N. Bhalaji, "SMART COLLEGE”- Study of Social Network and IoT Convergence, in: Proceedings of the 2nd International Conference on IoT in Social, Mobile, Analytics and Cloud, I-SMAC ’2018, Palladam, India, 2018, pp. 100-103. doi: 10.1109/I-SMAC.2018.8653787.

[5] S. Garg, T. Gandhi, B. Panigrahi, Social Network measures association with social and intelligent behaviors in Dolphin network, in: Proceedings of the 11th International Conference on Cloud Computing, Data Science & Engineering, Confluence ’2021, Noida, India, 2021, pp. 655-659. doi: 10.1109/Confluence51648.2021.9377088.

[6] M. Morzy, P. Kazienko, T. Kajdanowicz, Priority rank model for social network generation, in: Proceedings of the ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM ’2016, San Francisco, CA, USA, 2016, pp. 315-318. doi: 10.1109/ASONAM.2016.7752251.

[7] F. Long, N. Ning, C. Song, B. Wu, Strengthening Social Networks Analysis by Networks Fusion, in: Proceedings of the ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM ’2019, Vancouver, BC, Canada, 2019, pp. 460-463. doi: 10.1145/3341161.3342939.

[8] S. A. Bitaghsir, S. Kashipazha, A. Dadlani, A. Khonsari, Social-aware Mobile Road Side Unit for Content Distribution in Vehicular Social Networks, in: Proceedings of the 2019 IEEE Symposium on Computers and Communications, ISCC ’2019, Barcelona, Spain, 2019, pp. 1-6, doi: 10.1109/ISCC47284.2019.8969669.

[9] U. Brandes, A Faster, Algorithm for Betweenness Centrality, The Journal of Mathematical Sociology, volume 25 (2004). doi:10.1080/0022250X.2001.9990249.

[10] N. Salia, C. Tompkins, O. Zamora, An Erdős-Gallai type theorem for vertex colored graphs. Graphs and Combinatorics, volume 35, 2019, pp. 689–694. doi:10.1007/s00373-019-02026-1

[11] P. Erdős, I. Miklós, Z. Toroczkai, A Simple Havel–Hakimi, Type Algorithm to Realize Graphical Degree Sequences of Directed Graphs, The Electronic Journal of Combinatorics, volume 17 (2009). doi: 10.37236/338.

[12] A. Rapoport, Contributions to the Theory of Random and Biased Nets, Bulletin of Mathematical Biophysics, volume 19 (4), 1957, pp. 257–277. doi:10.1007/BF02478417.