VISUALIZATION AND CLUSTER ANALYSIS OF SOCIAL NETWORKS

M.I. Khotilin, A.V. Blagov

Samara National Research University, Samara, Russia

Abstract. The article is devoted to the analysis of social networks, which are presented by graphs. The paper presents approaches to the modeling of distributions of social networks as well as the algorithms used for finding communities, as well as accounts that have the greater impact on the community.

Keywords: big data, graphs, data visualization, data analysis, clustering, modularity, SCAN.

Citation: Khotilin MI, Blagov AV. Visualization and Cluster Analysis of Social Networks. CEUR Workshop Proceedings, 2016; 1638: 843-850. DOI: 10.18287/1613-0073-2016-1638-843-850

Introduction

In the modern world it is continuously generated a huge amount of data, whether the data received from the satellite, or sensors in the aircraft, bank transactions, patient diagnostic data, etc. A special place is occupied by social networks. The significance of social networks is due to the fact that, on the one hand they are the subject of socialization of people, and on the other - the most powerful and affordable political, ideological and economic instrument [1]. A number of papers are dedicated to researches of social networks as systems, which contain large volumes of data [2, 3, 4].

Large amounts of data as well as the relationships (connections) between them must be present in comfortable and readable form. The data from social networks can be presented in various forms: a tag cloud, charts, historical flows [5], but graphs are more often used for this purpose.

1 Representation of the network as a graph

Generally, when it talks about objects representing the network, such as social, data visualization concept is closely related to the concept graphs. An important task is to present links in social networks to identify different kinds of dependencies.

The graph is a collection of non-empty set of vertices and the set of edges: (- set of vertices, - set of edges). The vertices in a graph, which describes the social network, are user accounts, and edges - the connections between them. For example, a subscription in the network such as Twitter, and the attitude of the "friendship" in social networks like The Facebook (figure 1).

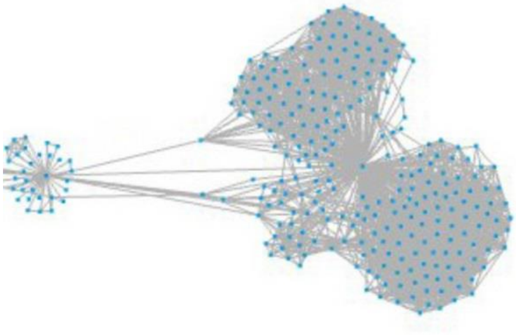


Fig. 1. A fragment of the graph of social network "VKontakte"

One of the main related characteristics that should be considered is the metric. The metric of the graph is based on the notion of distance.

For example, we call the distance $d\left(x_i,\ x_j\right)=d_{ij}$ between vertices x_i and x_j of the graph $G\left(X,U\right)$ the length of the shortest path, which connecting these vertices. By chain length the number of its edges is meant. Then, the function $d\left(x_i,\ x_j\right)$, defined on the set of edges U of graph G, is called graph metric.

The degree of a vertex $x_i \in X$ of graph is the number of edges incident to this vertex - $d(x_i)$.

Empirically, it has been proved that the degree distribution of the various segments of the vast majority of social networks has the following form (figure 2). f_k - is the proportion of vertices of $G\left(X,U\right)$, which has the degree $d(x_i)=k$.

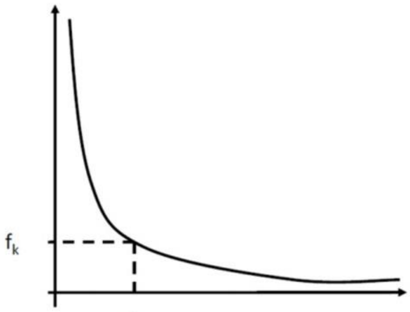


Fig. 2. The distribution of the degrees of vertices

$$p(k) = Ck^{-\alpha},$$

$$p(k) = \frac{z^k}{k!} e^{-z}.$$

Coefficients and are found for a particular segment of the social network.

2 Clustering and communities finding algorithms based on the modularity

To simplify the graph, and also for finding the so-called "communities" in a social network, which is described by graph, the clustering is applied.

There are a number of algorithms and approaches for clustering, one of which is the modularity [8-9].

This functionality was proposed by Newman and Girvan in the process of developing clustering algorithm of graph vertices [7]. Under the modularity means a scalar value from the interval [-1, 1], expressed by the following formula:

$$\mathrm{Q} = \frac{1}{2\mathrm{m}} \sum_{i,j} \left(A_{ij} - \frac{d_i d_j}{2m} \right) \delta(C_i, C_j),$$

where A - adjacency matrix, A_{ij} - *(i,j)* element of matrix A, d_i - the degree of *i* graph vertex, C_i – the label of the vertex (the number of the community, to which the vertex is belongs), *m* –the total number of edges in the graph, $\delta(C_i, C_j)$ - delta-function (one, if $C_i = C_j$, zero otherwise).

The task of finding isolation of communities in the graph is reduced to search such C_i, which will maximize the value of modularity.

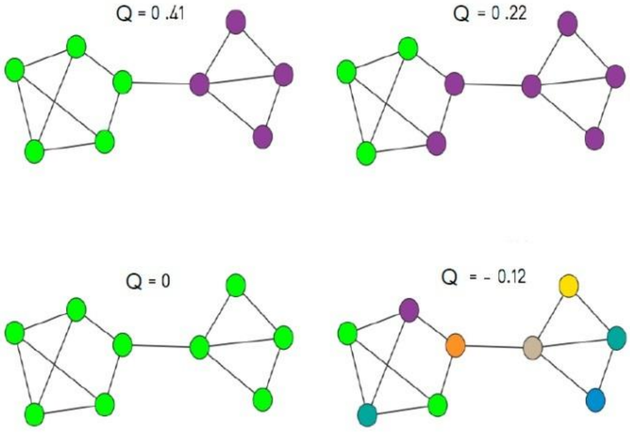


Fig. 3. The value of the modularity factor in the allocation of different clusters

The advantages of modularity may include the following:

- modularity simply interpreted. Its value is equal to the difference between the proportion of edges within the community and expected share of links, if the edges were placed randomly;
- it is possible to count modularity effectively with small changes in the clusters.

However, there are also some disadvantages:

- the functionality is not continuous, and the task of its optimization is discrete. The approximation schemes are used to find the global optimum. Some of them are really optimize the functional value, while others choose the value of the modularity by the best solution found, without warranty of local optimality of solutions;
- there is a resolution problem (the functionality does not "see" small communities). This problem is solved by using the modified functionality, which retains all the advantages and adds scale parameter [7].

As modularity describes the quality of separation of the graph in the groups, the problem of finding the optimal partition of the graph can be approached by solving the problem of maximizing of modularity. However, using a simple brute force method to solve this problem is almost impossible, since the number of options for separating n nodes into k groups grows exponentially with n To solve this problem the "greedy algorithm" of optimization of modular functions was proposed. This algorithm has its base in an association of two groups giving the highest increase modularity step by step.

Let us consider some decomposition of the N nodes into k groups (N – a set of nodes with the number of elements n) [8]. The modularity function will be equal:

$$Q_1 = \frac{1}{m}\sum_{l=1,l\neq i,l\neq j}^{k}\left(m_l - \frac{(d(N_l))^2}{4m}\right) + \frac{1}{m}\left(m_i + m_j - \frac{(d(N_i))^2 + (d(N_j))^2}{4m}\right).$$

Now join the group i and j in one, which is denoted as an$N_{i\cup j} = N_i \cup N_j$. The modularity function for the new graph will be:

$$Q_2 = \frac{1}{m}\sum_{l=1,l\neq i,l\neq j}^{k}\left(m_l - \frac{(d(N_l))^2}{4m}\right) + \frac{1}{m}\left(m_{i\cup j} - \frac{(d(N_{i\cup j}))^2}{4m}\right).$$

The number of edges within the group $N_{i\cup j}$ is equal to the sum edges within groups $N_i$and N_j plus the number of edges between them. In other words:

$$m_{i\cup j} = m_i + m_j + m_{i,j}$$

The degree of the united group $N_{i\cup j}$is equal to the sum of groups of degrees $N_i$and N_j:

$$d(N_{i\cup j}) = d(N_i) + d(N_j),$$

Consequently:

$$(d(N_{i\cup j}))^2 = (d(N_i))^2 + (d(N_j))^2 + 2d(N_i)d(N_j).$$

Taking this into account, we obtain:

$$\Delta Q = Q_2 - Q_1 = \frac{1}{m}\left(m_{i,j} - \frac{2d(N_i)d(N_j)}{4m}\right) = \frac{1}{m}\left(m_{i,j} - \frac{d(N_i)d(N_j)}{2m}\right).$$

This implies that the greatest growth of modularity occurs by combining groups $N_i$and N_j, for which the value:

$$\Delta(N_i, N_j) = m_{i,j} - \frac{d(N_i)d(N_j)}{2m}$$

is maximal.

It is also seen that the combination of groups, between which there are no edges ($m_{i,j}$=0), can not give increasing of the modularity.

3 Classification of vertices, finding the most significant, SCAN algorithm

Apart from finding a community in a social network which is represented by the graph, the classification of vertices in the graph has a considerable interest.

In the article [7] the following classification of the vertices was introduced (see figure 4):

- core – is the vertex which contains in ε-neighborhood at least μ vertices
- hub – is an isolated vertex which has neighbors belonging to two or more different clusters;
- outlier – an isolated all its neighbors of which either belong to only one cluster or do not belong to any cluster.

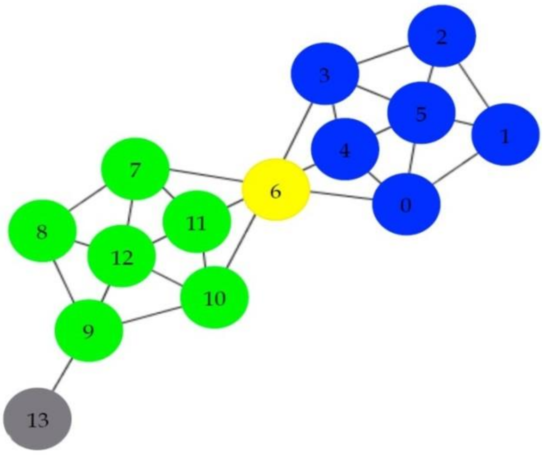


Fig. 4. An example of a graph with different types of vertices

To perform such classification a SCAN algorithm is proposed [7]. The principle of its work is described as follows.

The search begins by first visiting each vertex once to find structure-connected clusters, and then visiting the isolated vertices to identify them as either a hub or an outlier (hub or outlier).

SCAN performs one pass of a network and finds all structure-connected clusters for a given parameter setting. At the beginning all vertices are labeled as unclassified. SCAN classifies each vertex either a member of a cluster or not a member. For each vertex that is not yet classified, SCAN checks whether this vertex is a core. If the vertex is a core, a new cluster is expanded from this vertex. Otherwise, the vertex is labeled as not a member of the cluster.

To find a new cluster, SCAN starts with an arbitrary core V and search for all vertices that are structure reachable from V. This is sufficient to find the complete cluster containing vertex V. new cluster ID is generated which will be assigned to all found vertices.

SCAN begins by inserting all vertices in ε- neighborhood of vertex V into a queue. For each vertex in the queue it computes all directly reachable vertices and inserts those vertices into the queue which are still unclassified. This is repeated until the queue is empty.

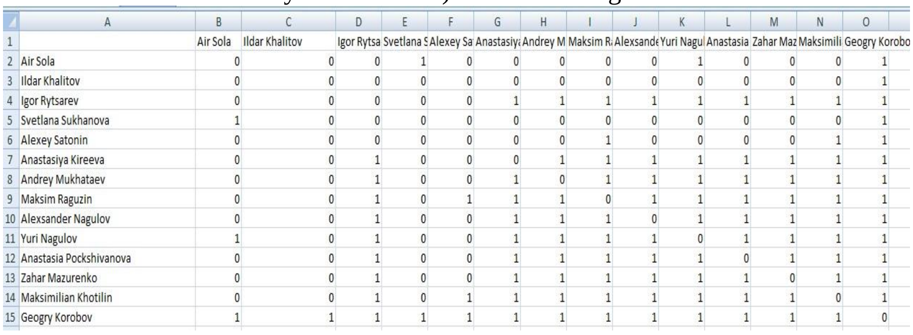
The non-member vertices can be further classified as hubs or outliers If an isolated vertex has edges to two or more clusters, it is may be classified as a hub. Otherwise, it is an outlier [7].

A distinctive feature is the presence of parameters μ and ε , which can be set by the user or an expert. At the same time the finding of the optimal values of these parameters can be carried out using a machine learning system, using a certain network segments.

4 Experiment and results

In order to test the algorithms mentioned above, the analysis of the community of Samara airsoft group "BPF" if social network “Vkontakte” was produced in order to identify possible communities by breaking into clusters.

The adjacency matrix for the graph (in this case, the state of the friendship between members of the community to each other) is shown in Figure 5



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1		Air Sola	Ildar Khalitov	Igor Rytsa	Svetlana S	Alexey Sa	Anastasiya	Andrey M	Maksim R	Alexsande	Yuri Nagul	Anastasia	Zahar Maz	Maksimili	Geogry Korobo
2	Air Sola	0	0	0	1	0	0	0	0	0	1	0	0	0	1
3	Ildar Khalitov	0	0	0	0	0	0	0	0	0	0	0	0	0	1
4	Igor Rytsarev	0	0	0	0	0	1	1	1	1	1	1	1	1	1
5	Svetlana Sukhanova	1	0	0	0	0	0	0	0	0	0	0	0	0	1
6	Alexey Satonin	0	0	0	0	0	0	0	1	0	0	0	0	1	1
7	Anastasiya Kireeva	0	0	1	0	0	0	1	1	1	1	1	1	1	1
8	Andrey Mukhataev	0	0	1	0	0	1	0	1	1	1	1	1	1	1
9	Maksim Raguzin	0	0	1	0	1	1	1	0	1	1	1	1	1	1
10	Alexsander Nagulov	0	0	1	0	0	1	1	1	0	1	1	1	1	1
11	Yuri Nagulov	1	0	1	0	0	1	1	1	1	0	1	1	1	1
12	Anastasia Pockshivanova	0	0	1	0	0	1	1	1	1	1	0	1	1	1
13	Zahar Mazurenko	0	0	1	0	0	1	1	1	1	1	1	0	1	1
14	Maksimilian Khotilin	0	0	1	0	1	1	1	1	1	1	1	1	0	1
15	Geogry Korobov	1	1	1	1	1	1	1	1	1	1	1	1	1	0

Fig. 5. The adjacency matrix

The graph with edges that describe the state of dependency of "friendship" and vertices - members of the group is shown in Figure 6.

For this graph the modularity had been calculated. With the resolution of 0.5 the value of the modularity parameter was 0.021, while the graph has been divided into five communities. By the way in the communities data includes accounts of people who are:

- students together in one university;
- graduated from the same school;

- those who have been invited by the participant Raguzin;
- those who have been invited by the participant Korobov;
- those who have been invited by the participant YuriNagulov.

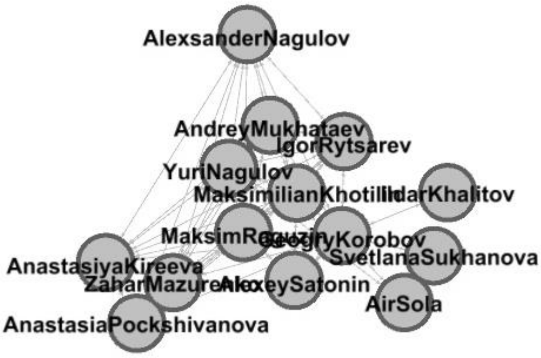


Fig. 6. The graph of the community

This division into clusters can be seen in the figure below:

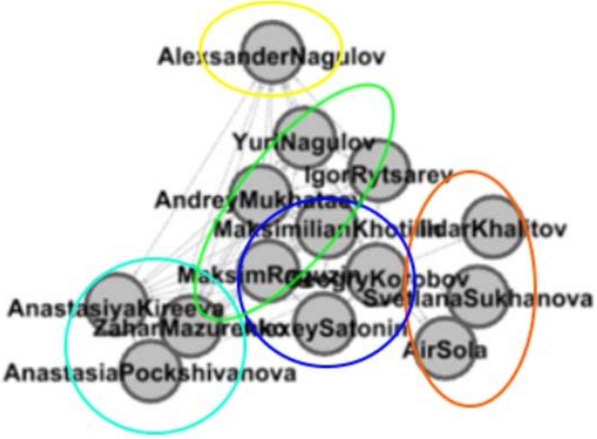


Fig. 7. The division of the graph into clusters

It should be noted quite satisfactory performance of the algorithm for finding the communities on a graph, which represents a relatively small segment of the real social network.

Conclusion

Presentation of social networks in the form of a graph and its further analysis, including clustering and finding dependencies, is an urgent task in Big Data. Using the methods described in this article and approaches permits to produce the classification of the social network segments and find the elements of greatest interest, for example, users, affecting several separate communities (in the graph representation - such as the edge of "hub"). When finding the power distribution of the nodes of the graph, which described the social network, it is possible to carry out modeling of social networks with a given distribution.

The algorithms presented in this article are planned for completion and use in the study of social segments of Samara region. The authors have developed the necessary tools to graph visualization of necessary social networks segments and distributed methods of processing of high-dimensional graphs.

References

1. Tan W, Blake MW, Saleh I, Dustdar S. Social-network-sourced big data analytics. IEEE Internet Computing, 2013; 5: 62-69.
2. Semertzidis K, Pitoura E, Tsaparas P. How people describe themselves on Twitter. Proceedings of the ACM SIGMOD Workshop on Databases and Social Networks, 2013; 25-30.
3. Blagov A, Rytcarev I, Strelkov K, Khotilin M. Big Data Instruments for Social Media Analysis. Proceedings of the 5th International Workshop on Computer Science and Engineering, 2015; 179-184.
4. Protcenko VI, Kazanskiy NL, Serafimovich PG. Real-time analysis of parameters of multiple object detection systems. Computer Optics, 2015; 39(4): 582-591 [In Russian]. DOI: 10.18287/0134-2452-2015-39-4-582-591.
5. Ivanov PD, Lopukhovsky AG. Big Data technologies and different methods of presenting large data. Science and innovations, 2014; 9: 1-10 [In Russian].
6. Gastner M, Michael T, Newman ME. Optimal design of spatial distribution networks. Physical Review, 2006; 74: 016117-016126.
7. Newman ME, Girvan M. Finding and evaluating community structure in networks. Physical Review 2004; 69: 026113-026115.
8. Xu X. Scan: a structural clustering algorithm for networks. Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, 2007; 824-833.
9. Newman ME. Fast algorithm for detecting community structure in networks. Physical Review, 2004; 69(6): 066133-066135.