# Centrality Measures in Large and Sparse Networks

Fuad Aleskerov
HSE
ICS RAS
Moscow, Russia

Sergey Shvydun
HSE
ICS RAS
Moscow, Russia

Natalia Meshcheryakova
HSE
ICS RAS
Moscow, Russia

Vyacheslav Yakuba
ICS RAS
HSE
Moscow, Russia

***Abstract*—The problem of quick detection of central nodes in large networks is studied. There are many measures that allow to evaluate a topological importance of nodes of the network. Unfortunately, most of them cannot be applied to large networks due to their high computational complexity. However, if we narrow the initial network and apply these centrality measures to the sparse network, it is possible that the obtained set of central nodes will be similar to the set of central nodes in large networks. If these sets are similar, the centrality measures with a high computational complexity can be used for central nodes detection in large networks. To check the idea, several random networks were generated and different techniques of network reduction were considered. We also adapted some rules from social choice theory for the key nodes detection. As a result, we show how the initial network should be narrowed in order to apply centrality measures with a high computational complexity and maintain the set of key nodes of a large network.**

***Index Terms - network analysis; centality measures; sparse networks; social choice rules; computational complexity***

## I. Introduction

Over the past several decades, the detection of key nodes in the network has been one of the main subjects in network theory. There has been developed many indices that measure the centrality level of each node of the network. Some of them are based on the number of links to the other nodes with or without respect to the importance of adjacent nodes (degree measures, eigenvector centrality, PageRank, etc.). Other techniques consider how close each node is located to the other nodes of the network in terms of the distance or how many times it is on the shortest paths connecting any given node-pairs, etc. The more features are taken into account, the more difficult in terms of computational complexity these centrality measures are.

Unfortunately, due to enormous growth of information most real-life networks are large and complex, which leads to the fact that some centrality measures are difficult to calculate due to their high computational complexity. There are many ways how to deal with this situation. In some cases, the centrality measures with a low computational complexity are used instead of more complex ones; in other cases, centrality measures for which fast parallel algorithms exist are applied.

In this paper we focus on another approach of central nodes detection in large networks. The main idea is the following: instead of calculating centrality measures with a high computational complexity on a large network, we can somehow narrow it and calculate these measures on a sparse network. If the set of key nodes of the sparse network is identical to the set of key nodes of the large network, then there is no need to calculate centrality measures on the large network. To check the idea, we randomly generated several large networks, applied several techniques of network reduction and then compared how a set of central nodes was changed. We also made an attempt to adapt some rules from the social choice theory and compared the results for large and sparse networks.

The paper is organized as follows. First, different measures of central nodes detection in networks are given. Second, we describe how we generate random networks and narrowed them. Finally, we compare how different are the sets of key nodes in large and sparse networks.

## II. Centrality Measures

We consider the following centrality measures:

- *In-degree centrality* – the number of in-going links of the node;
- *Out-degree centrality* – the number of out-going links of the node;
- *Degree centrality* – the total number of links of the node;
- *Closeness centrality* - the inverse from the average distance from a given node to all other nodes of the network;

- *Betweenness centrality* – the number of shortest paths going through a specific node;
- *PageRank centrality* – the rank of a node which depends on the link structure of a given network;
- *Eigenvector centrality* - the values of the first eigenvector of the network adjacency matrix.

The detailed description of centrality measures is provided in [1-3].

It is necessary to consider the computational complexity of these centrality measures. In-degree, out-degree, degree and closeness centrality are measures with a low computational complexity. In order to calculate the betweenness centrality, the shortest paths between all node-pairs should be considered. Thus, the betweenness centrality is difficult to calculate for large networks. As for the PageRank and eigenvector centrality, these measures have a high computational complexity but there are many approximate algorithms with a low computational complexity integrated in standard software packages that implement the main idea of these centrality measures.

We also consider some rules from social choice theory based on the pairwise comparison of elements:

- Copeland rules 1-3 – depends on the cardinality of the upper and lower counter sets of each element [4];
- Strong top cycle (STC) – a maximal set of undominated elements [5, 12-13];
- Minimal externally stable set (MES) – a set of externally stable elements [6, 14];
- Untrapped set (UT) –a set of untrapped elements [7].
- Uncovered sets ($UC^{I}$, $UC^{II}$) – a set of uncovered elements [7-11];

The detailed description of these rules is provided in [4-14]. The matrix-vector representation of rules based on majority relation was designed in [15].

Unfortunately, since these rules are based on a pairwise comparison, their computational complexity is extremely high so it is difficult to apply them to the network analysis. However, our motivation of using these rules is the following. First, these rules are the results of different ideas how the optimal social choice should look like. Thus, these rules have an easy and clear interpretation. Second, if the results of these rules for sparse networks are similar to the results of these rules for large networks, then these rules can be computed on sparse networks. Third, it is important to compare the relation of these rules with existing centrality measures. Finally, although an application of these rules in empirical studies was bounded by the limits of computation, the interest in these solutions is revived with the development of computer technology.

## III. Network Generation and Reduction

To compare centrality measures in large and sparse networks 10 networks were randomly generated. The degree distribution of each generated network is exponential since most real large and complex networks have an exponential degree distribution. Each network contains 10000 nodes while the number of links varies from 250000 to 325000. More information on random networks is given in Table I.

TABLE I. Networks Statistic

| Network Features | Value |
|---|---|
| Number of nodes | 10000 |
| Number of links | 250000-325000 |
| Average degree | 25.5-32.5 |
| Network density | 0.003 |
| Network diameter | 7-9 |
| Average path length | 3.3-3.5 |
| Average clustering coefficient | 0.006-0.007 |
| Weakly connected components | 1 |
| Strongly connected components | 1 |

As it was mentioned before, we compared centrality measures in large and sparse networks. There are many ways how to obtain a sparse network from a large one. However, all techniques should satisfy the following requirement: the procedure of network reduction should be simple and quick.

In our paper we consider the edges elimination procedure which is based on the idea of composition of super-threshold procedures. Let us remind that the super-threshold procedure consists in the choice of those elements for which the estimate criterion is more (less) than some pre-defined threshold [16]. The composition of super-threshold functions is the sequential application of super-threshold procedures so the result of the first super-threshold procedure is the input for the second super-threshold procedure. In our research we use the Copeland values as the estimate criteria for network reduction while the threshold is assigned differently.

Thus, the edges elimination procedure is done as follows. First, the Copeland values are calculated for the initial large network. Then, the nodes for which the Copeland value is more (less) than the pre-defined threshold are eliminated. Next, for a sparse network obtained from the previous step the Copeland values of each node are re-calculated and the super-threshold procedure is performed again. Thus, the elimination procedure is applied sequentially until no more nodes can be eliminated from the network by the super-threshold procedure.

Since the elimination procedure depends on the pre-defined threshold value, different thresholds are considered. The super-threshold procedure was applied on both sides (more than threshold and less than threshold) since the links of the initial network can be interpreted differently – as a network of domination (node X is better than node Y) and as a network of citation (node X cites node Y). Thus, the following thresholds are considered

- For the domination network (node X is better than node Y):
  - Copeland 1: mean value, worst 20%, 30%, 40%, 50%, 60% intervals;
  - Copeland 2: mean value, worst 20%, 30%, 40%, 50%, 60% intervals;
  - Copeland 3: mean value, worst 20%, 30%, 40%, 50%, 60%, 70% intervals;

- For the citation network (node X cites node Y):
  - Copeland 1: mean value, worst 5%, 10%, 15%, 20%, 25%, 30% intervals;
  - Copeland 2: mean value, worst 5%, 10%, 15%, 20%, 25%, 30% intervals;
  - Copeland 3: mean value, worst 5%, 10%, 15%, 20% intervals.

Thus, 38 ways of network reduction are considered.

The motivation of using the Copeland values is rather simple. First, these rules are easy to compute since they take into account only information about in-going or out-going links of each node. Second, these rules have a clear interpretation. Third, in some cases these methods are equal to the in-degree or out-degrees centrality measures which can also be used as criteria for the edges elimination. However, we should emphasize that the use of these rules does not limit us to use some other techniques of network reduction in further research.

Thus, the elimination procedure was applied for 10 randomly generated networks. For each network about 700 sparse networks were obtained. Fig. 1-4 indicates how many steps were performed to narrow one of the initial networks (axis Y is a network size, axis X is an iteration number).

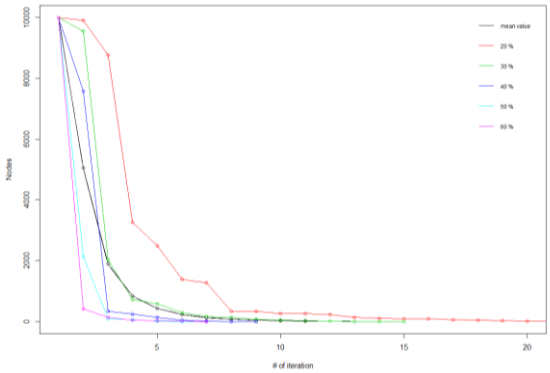


Fig. 1. Edges and nodes elimination by the Copeland rules 1,2 (elimination of nodes with Copeland values less than the pre-defined threshold).

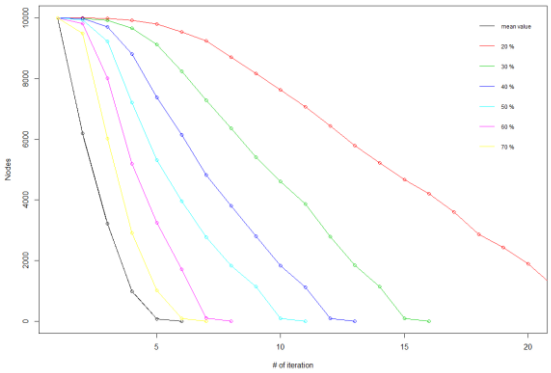


Fig. 2. Edges and nodes elimination by the Copeland rule 3 (elimination of nodes with Copeland values more than the pre-defined threshold).

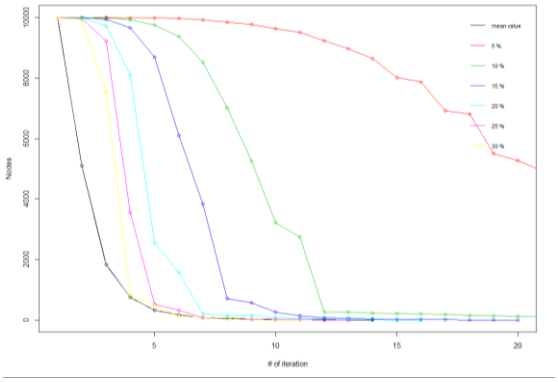


Fig. 3. Edges and nodes elimination by the inverse Copeland rules 1,2 (elimination of nodes with Copeland values more than the pre-defined threshold).

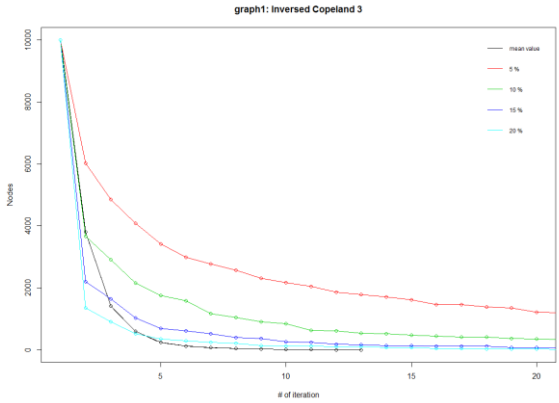


Fig. 4. Edges and nodes elimination by the inverse Copeland rule 3 (elimination of nodes with Copeland values less than the pre-defined threshold).

Fig. 1-4 show that few number of steps (less than 5) are needed to eliminate almost all nodes of the initial network by the Copeland rules 1,2 (see Fig.1) and the inverse Copeland rule 3 (see Fig.4). On the contrary, the Copeland rule 3 and the inverse Copeland rules 1,2 requires more steps to eliminate the initial network of size 10000.

The elimination procedure was performed in R 3.2.2 software package.

## IV. Results

To compare the centrality measures in large and sparse networks we firstly calculated these measures for all initial networks of size 10000 and their subnetworks (sparse networks) obtained by different elimination procedures. The computation of existing centrality measures was done in R 3.2.2 software package while the computation of rules based on majority relation was performed with the use of C#.

For each network a set of central nodes was chosen. The procedure was done as follows: a node is included in the set if its centrality measure is more or equal than some pre-defined threshold. The threshold can be assigned differently; in our work nodes which are included into x% interval are considered as central.

Thus, for each centrality measure we can define how many central nodes of the network of size 10000 remained in the set of central elements in sparse networks. The results are provided in Fig. 5-11 where each line represents some way of edges and nodes elimination.

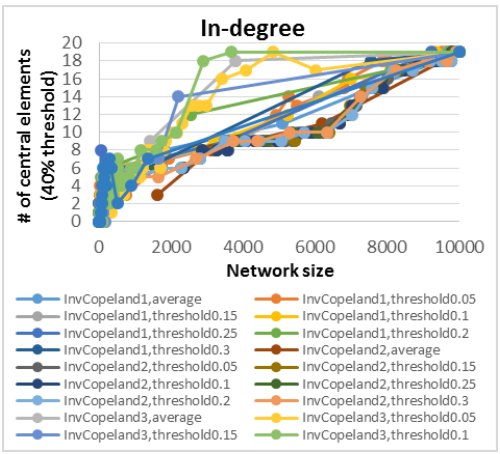


Fig. 5. Total number of nodes considered as central in the initial network by the in-degree centrality measure remained in the set of central nodes in sparse networks.

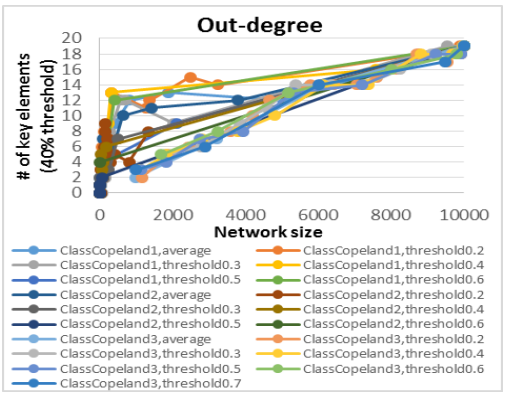


Fig. 6. Total number of nodes considered as central in the initial network by the out-degree centrality measure remained in the set of central nodes in sparse networks.

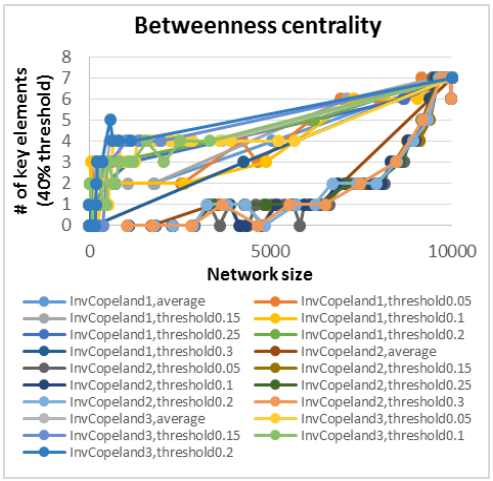


Fig. 7. Total number of nodes considered as central in the initial network by the degree centrality measure remained in the set of central nodes in sparse networks.

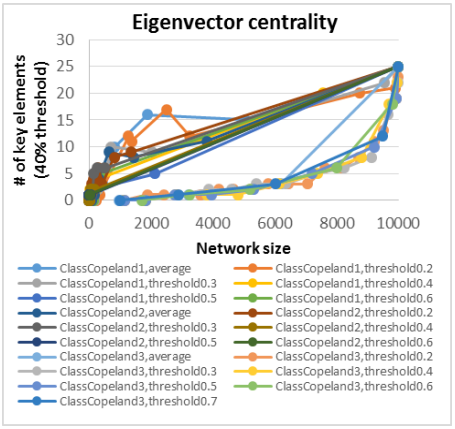


Fig. 8. Total number of nodes considered as central in the initial network by the eigenvector centrality measure remained in the set of central nodes in sparse networks (the links of the network interpreted in terms of domination).

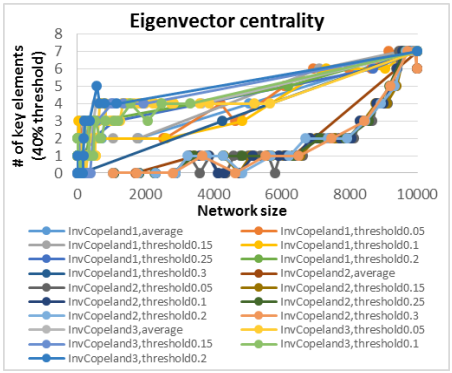


Fig. 9. Total number of nodes considered as central in the initial network by the eigenvector centrality measure remained in the set of central nodes in sparse networks (the links of the network interpreted in terms of citation).

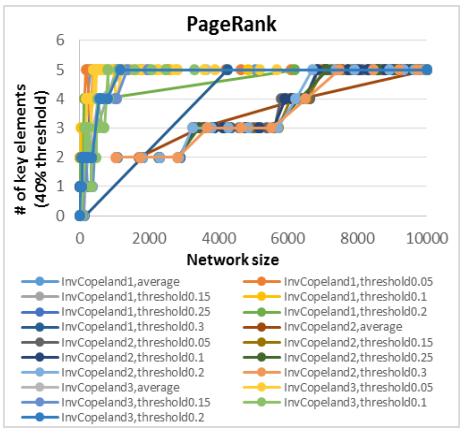


Fig. 10. Total number of nodes considered as central in the initial network by the PageRank centrality measure remained in the set of central nodes in sparse networks.

The results are the following. Best nodes by the in-degree centrality measure in the network of size 10000 remain central in networks of size 4000 if the elimination procedure is applied by the inverse Copeland rule 3 (see Fig. 5). Nodes which are considered as central by the out-degree centrality measure in the network of size 10000 remain central in networks of size 1000-2000 if the elimination procedure is applied with the use of the Copeland rule 1 (see Fig. 6). Such results came from the

fact that degree centrality measures in some cases coincide with the Copeland rules.

As for more sophisticated centrality measures, the results are the following. Nodes, which were chosen by the betweeness centrality in the network of size 10000, remain central in the network of size 1000-2000 if the elimination procedure is applied by the inverse Copeland rule 3 (Fig. 7). Nodes, which were chosen by the eigenvector centrality in the network of size 10000, remain central in the network of size 1000-3000 if the elimination procedure is applied by the Copeland rule 1 (see Fig. 8) or by the inverse Copeland rule 3 (see Fig. 9). Nodes, which were chosen by the PageRank centrality, remain central in the network of size 500-1500 if the elimination procedure is applied by the inverse Copeland rule 3 (see Fig. 10).

We also applied rules based on the majority relation to our random networks. First, we studied how many nodes are chosen by different rules in the initial network. Such information can be used to evaluate the significance of the rule. The results are provided in Table II.

TABLE II. NUMBER OF NODES CHOSEN BY DIFFERENT RULES IN RANDOM NETWORKS.

| Network number | Number of nodes chosen from the initial network | | | | |
|---|---|---|---|---|---|
| | $UC^{I}$ | $UC^{II}$ | STC | MES | UT |
| 1 | 10 | 9230 | 2 | 19 | 337 |
| 2 | 6 | 9205 | 0 | 21 | 299 |
| 3 | 11 | 9205 | 0 | 13 | 372 |
| 4 | 13 | 9225 | 2 | 11 | 425 |
| 5 | 15 | 9242 | 1 | 0 | 446 |
| 6 | 16 | 9215 | 0 | 12 | 413 |
| 7 | 8 | 9217 | 0 | 21 | 219 |
| 8 | 14 | 9173 | 0 | 12 | 388 |
| 9 | 7 | 9258 | 1 | 10 | 165 |
| 10 | 11 | 9161 | 1 | 16 | 123 |

As it is shown above, the uncovered set II ($UC^{II}$) choose almost all nodes of the initial network while the strong top cycle (STC) is almost always empty. Hence, if we also take into account the computational complexity of these rules, we can conclude that the output of these rules does not provide any valuable information. As for other rules, the untrapped set (UT) contains rather large amount of elements (several hundreds) while the uncovered set I ($UC^{I}$) and the minimal externally stable set (MES) are quite small.

We also considered how the size of each set varies with the network size. The uncovered set I ($UC^{I}$) contains no more than 100 elements for all subnetworks of size less than 10000. The total number of elements included in the uncovered set II ($UC^{II}$) is close to the size of the network. The strong top cycle (STC) contains no more than 800 elements for all networks of size less than 10000, however, in 30% of cases STC is empty. The cardinality of the minimal externally stable set (MES) does not exceed 1600 elements and the total number decreases with the size of the network. The untrapped set (UT) choses relatively small number of elements (less than 10 in 45% of cases); in 25% of cases the set is empty. However, for sparse networks of size close to 10000, the cardinality of the untrapped set increases dramatically and in some cases it is equal to 500-700.

Thus, we focused on the rules that choose relatively small number of elements. Among considered rules based on majority relation, only the uncovered set I ($UC^{I}$) and the minimal externally stable set (MES) satisfy the condition. Thus, we examined only for these rules how many central nodes of the initial network will be chosen in sparse networks. The results are provided in Fig. 11-12.

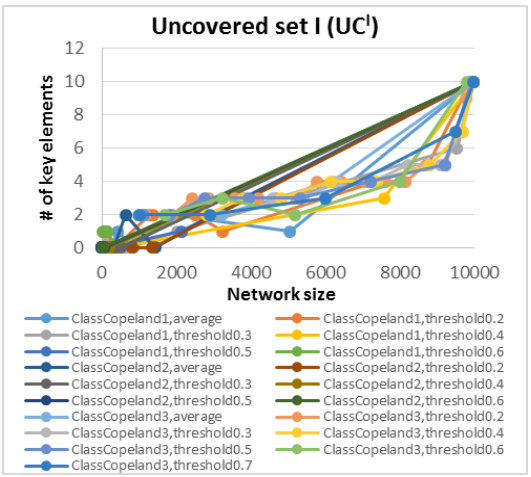


Fig. 11. Total number of nodes included in the set $UC^{I}$ in the initial network remained in the set $UC^{I}$ in sparse networks.

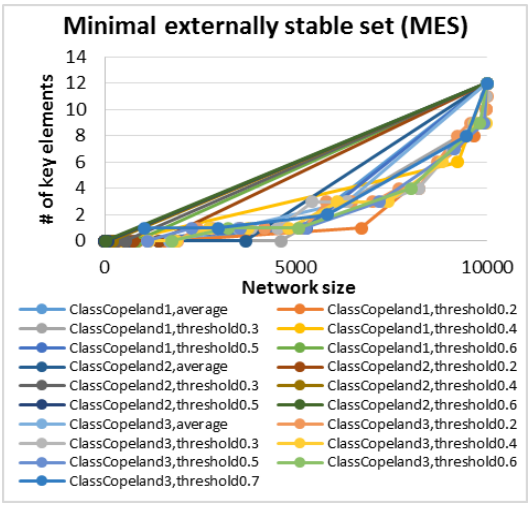


Fig. 12. Total number of nodes included in the set MES in the initial network remained in the set MES in sparse networks.

The results show that central nodes of the initial network are not included in the set MES in sparse networks of size less than 9000 (see Fig. 12). In other words, the set MES of the initial network does not coincide with the set MES of sparse networks. As for the uncovered set I ($UC^{I}$), only small number of nodes included in the set $UC^{I}$ in the initial network remain in the set $UC^{I}$ in sparse networks of size 3000+.

## V. CONCLUSION

We studied how the centrality measures differs in large and sparse networks. The focus of this paper is on their computational complexity. Since most centrality measures

cannot be applied to large networks, we proposed to compute them on sparse networks. To test our hypothesis, 10 networks with exponential degree distribution were randomly generated and different network reduction techniques were considered. As result, we showed which elimination procedure should be applied in order to maintain the set of key nodes chosen in large networks in sparse networks.

We also studied different rules from social choice theory in order to define which of them can be applied for key nodes detection. The results showed that for most rules the set of central nodes of large networks does not coincide with the set of central nodes of sparse networks.

The results of this work are not final; however, it demonstrates the main idea of our research. In further research, some other centrality measures as well as elimination procedures will be considered. It is also necessary to mention that we applied our idea to randomly generated networks; however, our approach can also be applied to the real networks.

## ACKNOWLEDGMENT

The work was partially financed by the International Laboratory of Decision Choice and Analysis (DeCAn Lab) of the National Research University Higher School of Economics and by the Laboratory of V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences. The research was carried out in the framework of the Russian Foundation for Basic Research (RFBR) (grant №15-01-02847 "Influence measures in network structures").

## REFERENCES

[1] Bonacich P. Technique for Analyzing Overlapping Memberships// Sociological Methodology, Vol.4, 1972, pp.176-185.

[2] Brin S., Page, L. The anatomy of a large-scale hypertextual Web search engine// Comput. Netw., 30, 1998, pp.107-117.

[3] Newman M.E.J. Networks: An Introduction. – Oxford, UK: Oxford University Press, 2010.

[4] Aleskerov F.T., Khabina E.L., Shvarts D.A. (2006) Binary relations, graphs and collective decisions, Moscow: HSE Publishing House (in Russian).

[5] Ward B.: Majority Rule and Allocation. *J. Confl. Resolut*. 5. 379-389 (1961)

[6] Aleskerov F., Kurbanov E.: A Degree of Manipulability of Known Social Choice Procedures. In: Alkan A., Aliprantis Ch., Yannelis N. (eds.) Current Trends in Economics: Theory and Applications, pp. 13-27. Springer, Berlin/Heidelberg/New York (1999)

[7] Duggan J.: A systematic approach to the construction of non-empty choice sets. *Soc. Choice &Welf*. 28. 491-506 (2007)

[8] Duggan J.: Uncovered sets. Mimeo (2006)

[9] Miller N.: A new solution set for tournaments and majority voting: Further graph-theoretical approaches to the theory of voting. *Amer. J. Pol. Sci*. 24. 68-96 (1980)

[10] Fishburn P.: Condorcet social choice functions. *SIAM J. Appl. Math*. 33. 469-489 (1977)

[11] McKelvey R.: Covering, dominance and institution-free properties of social choice. *Amer. J. Pol. Sci*. 30. 283-314 (1986)

[12] Schwartz T.: On the Possibility of Rational Policy Evaluation. *Theory & Decis.* 1. 89-106 (1970)

[13] Schwartz T.: Rationality and the Myth of the Maximum. Noûs. 6. 97-117 (1972)

[14] Subochev A.: Dominating, Weakly Stable, Uncovered Sets: Properties and Extensions. Avtomatika i Telemekhanika (Automation & Remote Control). 1. 130-143 (2010)

[15] Aleskerov F. T., Subochev A. Modeling optimal social choice: matrix-vector representation of various solution concepts based on majority rule // Journal of Global Optimization. 2013. Vol. 56. No. 2. P. 737-756.

[16] Aizerman M., Aleskerov F. Theory of Choice. Elsevier, North-Holland, 1995.