Graphs decomposition using modified spectral clustering method

D.T. Yakupov[1], V.V. Mokshin[1]

[1]Kazan National Research Technical University named after A.N. Tupolev – KAI, K. Marks Str. 10, Kazan, Russia, 420111

Abstract. Among a large number of tasks on graphs, studies related to the placement of objects with the aim of increasing the information content of complex multi-parameter systems find wide practical application (for example, in transport and computer networks, piping systems, in image processing). Despite years of research, accurate and efficient algorithms cannot be found for placement problems. It is proposed to consider the solution of the allocation problem in the context of decomposition of the initial network into *k* regions, in each of which a vertex with some centrality property is searched. This article provides an analysis of sources for solving the problem of placement in graphs, as well as methods of decomposition of graph structures. Following the main provisions of the theory of spectral clustering, the disadvantages of the splitting applied criteria *Rcut* and *Ncut* are indicated. It is shown that the application of the distance minimization criterion *Dcut* proposed in this paper allows to obtain high results in the decomposition of the graph. The obtained results are based on the examples of searching for sensor placement vertices in the known ZJ and D-Town networks of the EPANET hydraulic modeling system.

1. Introduction

Graph models provide an opportunity to study an object based on its topology, without delving into the physical nature of the processes occurring in the system under consideration, which, in turn, greatly simplifies the calculations [1-3]. Among the many problems on graphs, studies related to splitting the original graph into a predetermined number of connected disjoint components have found wide practical application [4-8]. Methods of decomposition of graph structures make a significant contribution to the speed of search algorithms, which is especially important in conditions of restrictions on computational and time resources. However, widespread algorithms of spectral clustering based on minimization of *Rcut* and *Ncut* sections do not always allow to solve the problem of object placement in the best way. The reason for this is that the decomposition by these criteria takes into account the number of cut edges and the size of the resulting subgraphs, but does not take into account the distances between the vertices and the nature of their location within these subdomains. This paper provides an example showing that decomposing the original graph of 12 vertices and 12 edges into two parts, there are two splitting options that meet the *Rcut* and *Ncut* criteria, but when considering these options in terms of the distances between the vertices within these subdomains, the second option is preferred. The use of this criterion, designated in the paper as *Dcut*, in the decomposition of graphs allowed us to solve the problem of placing objects in the network with a high quality result, comparable and even better than spectral methods based on *Rcut* or *Ncut* criteria, which confirms the applicability of the *Dcut* criterion in spectral methods of clustering graphs.

2. The analysis of the sources

The solution of the problem of finding the optimal placement of objects in k nodes for a network with the number of vertices $|C|$ by the full search method requires $\frac{|C|!}{k!(|C|-k)!}$ iterations. For example, for a small network with $|C|$=100 and k=5 it is required $10^{7,88}$ iterations, which makes practical application of this method impossible. Trial-and-error, greedy, and stochastic algorithms are the most widely used approaches for solving the placement problem.

The trial-and-error algorithm is based on iterations. This approach provides a fairly close to optimal solution, but requires a lot of time.

The principle of the greedy algorithm is to make locally optimal decisions at each stage, assuming that the final solution will also be optimal. The solution is fast, but not accurate.

Solving by evolutionary computation, simulated annealing algorithms is based on choosing combinations of nodes placing objects on the basis of the probabilistic approach does not provide warranty of solution time and solution as a whole.

In [8, 9] the solution of the placement problem is proposed to be considered in the context of the decomposition of the initial network into k regions, in each of which a vertex with some property of centrality is searched. When decomposing a graph, it is necessary to minimize the number of edges connecting the vertices of different subdomains. A prerequisite for this is the connectivity of the selected subgraphs.

Many methods are used to solve the graph decomposition problem [10-15].

Since the 1990s, spectral graph theory has been used in many fields [16]. The main advantage of spectral graph theory is simplicity, so any system represented as a graph can be analyzed only by the spectrum of the associated matrix.

Fiedler in [17] showed that the eigenvector that corresponds to the second smallest eigenvalue of the Laplacian matrix can be used to solve the problem of bipartite graph decomposition. Hagen and Kahng [18] introduce the criterion of rational sections (*Rcut*) to assess the quality of decomposition. Shi and Malik in [19] use Fiedler's conclusions for iterative bipartite partitioning and introduce the normalized sections criterion (*Ncut*). The development and application of the theory of spectral clustering are also considered in [20-25].

3. Theory of spectral clustering of graphs

3.1. Fundamentals

The class of spectral decomposition methods [26-29] combines elements of graph theory and linear algebra. They are based on the application of the properties of eigenvalues and vectors of the Laplacian matrix of the graph.

The eigenvectors contain information about the topology of the graph. Based on the problems, the spectral graph theory uses: the main eigenvector [30], Fiedler eigenvector [17], a group of the first k eigenvectors [19]. A review of spectral clustering methods is presented in [31, 32].

Spectral clustering algorithms consist of three main steps:

1) For the original graph G, the adjacency matrix W, the matrix of degrees of vertices D, the Laplacian matrix L are forming. In addition to the non-normalized Laplacian matrix, its normalized equivalents are also used, for example, the Laplacian matrix normalized by the random walk method [19] or the symmetric normalized Laplacian matrix [20].

2) Determination of eigenvalues and eigenvectors of the non-normalized or normalized Laplacian matrix, which are used in the formation of the matrix of eigenvectors U.

3) Division of the set of vertices into k clusters by classical clustering methods, for example, the k-means method in relation to the matrix U.

3.2. Graph cut point of view

Methods of spectral clustering are aimed at obtaining such subgraphs that the difference between the constituent elements of the subdomain is minimal with the maximum difference between the

subgraphs. In this case, the subgraphs must be connected and balanced in size. To implement these conditions, the criteria proposed in [18, 19]:

$$Rcut(C_1, \dots, C_k) = \sum_{i=1}^{k} \frac{cut(C_i, G)}{|C_i|} \rightarrow min \quad (1)$$

$$Ncut(C_1, \dots, C_k) = \sum_{i=1}^{k} \frac{cut(C_i, G)}{vol(C_i)} \rightarrow min \quad (2)$$

where G - is the initial graph, C_i – i-th subgraph, k is the number of sub-areas to divide the original graph, $cut(C_i, G)$ - the sum of the weights of the cut edges, $|C_i|$ – is the quantity of vertices in the subgraph i, $vol(C_i)$ – the sum of the weights of edges in subgraph i.

It should be noted that the values of both criteria tend to a minimum with decreasing edge sections and with balancing subgraphs ($|C_1| = |C_2| = \cdots = |C_k|$ or $vol(C_1) = vol(C_2) = \cdots = vol(C_k)$). According to [32], the *Rcut* criterion is preferred for decomposition by non-normalized matrices, and *Ncut* is preferred for decomposition by normalized matrices.

However, the criteria under consideration do not always clearly indicate a solution. Figure 1 shows a graph with 12 vertices and 12 edges. The weight of each edge, according to the figure, is 10.

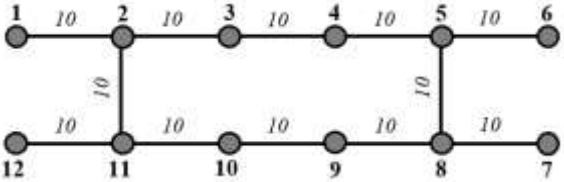


Figure 1. the Original graph with 12 vertices and 12 edges.

When searching for the optimal decomposition on k=2 subgraphs, we get two variants (figure 2).

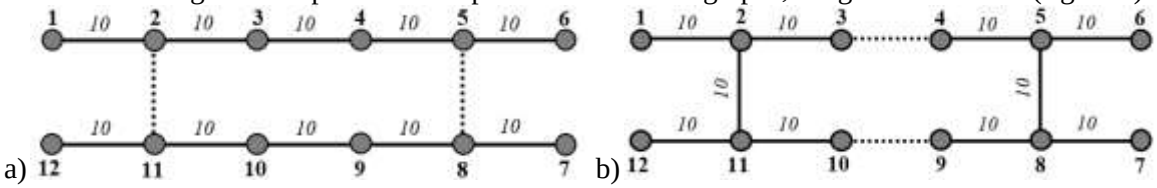


Figure 2. Decomposition of the original graph into 2 parts: a) with a partition on edges 2-11 and 5-8; b) with a partition on edges 3-4 and 9-10.

Next, according to (1) and (2), we define the values of the *Rcut* and *Ncut* criteria for the first (figure 2.a) and second (figure 2.b) variants of the partition:

$$Rcut_1 = \frac{cut_{2,11} + cut_{5,8}}{|C_1|} + \frac{cut_{2,11} + cut_{5,8}}{|C_2|} = \frac{10+10}{6} + \frac{10+10}{6} = 6{,}67 \quad (3)$$

$$Ncut_1 = \frac{cut_{2,11} + cut_{5,8}}{vol(C_1)} + \frac{cut_{2,11} + cut_{5,8}}{vol(C_2)} = \frac{10+10}{50} + \frac{10+10}{50} = 0{,}8 \quad (4)$$

$$Rcut_2 = \frac{cut_{3,4} + cut_{9,10}}{|C_1'|} + \frac{cut_{3,4} + cut_{9,10}}{|C_2'|} = \frac{10+10}{6} + \frac{10+10}{6} = 6{,}67 \quad (5)$$

$$Ncut_2 = \frac{cut_{3,4} + cut_{9,10}}{vol(C_1')} + \frac{cut_{3,4} + cut_{9,10}}{vol(C_2')} = \frac{10+10}{50} + \frac{10+10}{50} = 0{,}8 \quad (6)$$

From calculations it is clear that from the point of view of *Rcut* and *Ncut* both variants of splitting give the same result. However, when solving placement problems, it is important to consider the distances between all vertices in subgraphs. Tables 1 and 2 show the distances between vertices in subgraphs when decomposing by the first (figure 2.a) and second variants of the partition (figure 2.b).

It can be seen from the tables that the second variant of splitting the graph into 2 parts provides a smaller distance length in subgraphs, which indicates a greater degree of grouping of vertices in subgraphs in the second variant of decomposition.

Table 1. Distances between vertices in subgraphs (decomposing by the first variant of the partition).

Subgraph C_1							***Subgraph C_2***						
Nodes	***1***	***2***	***3***	***4***	***5***	***6***	***Nodes***	***7***	***8***	***9***	***10***	***11***	***12***
1	0	10	20	30	40	50	***7***	0	10	20	30	40	50
2	10	0	10	20	30	40	***8***	10	0	10	20	30	40
3	20	10	0	10	20	30	***9***	20	10	0	10	20	30
4	30	20	10	0	10	20	***10***	30	20	10	0	10	20
5	40	30	20	10	0	10	***11***	40	30	20	10	0	10
6	50	40	30	20	10	0	***12***	50	40	30	20	10	0
Sum: 700							**Sum: 700**						

Table 2. Distances between vertices in subgraphs (decomposing by the second variant of the partition).

Subgraph C_1'							***Subgraph C_2'***						
Nodes	***1***	***2***	***3***	***10***	***11***	***12***	***Nodes***	***4***	***5***	***6***	***7***	***8***	***9***
1	0	10	20	30	20	30	***4***	0	10	20	30	20	30
2	10	0	10	20	10	20	***5***	10	0	10	20	10	20
3	20	10	0	30	20	30	***6***	20	10	0	30	20	30
10	30	20	30	0	10	20	***7***	30	20	30	0	10	20
11	20	10	20	10	0	10	***8***	20	10	20	10	0	10
12	30	20	30	20	10	0	***9***	30	20	30	20	10	0
Sum: 580							**Sum: 580**						

3.3. Distance minimization criteria

Let's consider the solution of the problem of placing k =2 objects on the basis of the preliminary decomposition of the graph according to options 1 and 2.

The problem under consideration can be formulated as follows. There is a graph G whose nodes are characterized by certain parameter estimates P_i, and whose edges are characterized by weights W_j. After setting the next object in the vertex, the estimates are recalculated according to the formulas:

$$P_S = 1 \tag{7}$$

$$P_i = max(P_y \cdot 1/W_{i,y}) \tag{8}$$

where P_S - evaluation of the deterministic value of the parameter in the node setup of the object, P_y - evaluation of the deterministic value of the parameter of node neighbor, $W_{i,y}$ - is the weight of edge connecting two adjacent vertices i and y.

It is necessary to find such an arrangement of objects in the nodes that provides a minimum of the average value of the uncertainty estimation of the target parameter:

$$F = 1 - mean(P) \rightarrow min \tag{9}$$

The values of the estimates P_i of each vertex for the decomposition variants 1 and 2 are shown in figure 3:

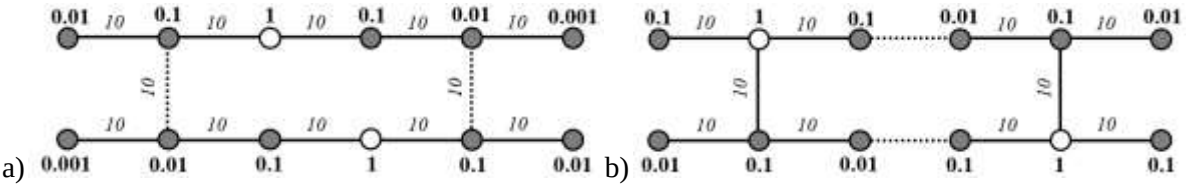


Figure 3. Values of P_i scores: a) splitting by edges 2-11 and 5-8; b) splitting by edges 3-4 and 9-10.

White color indicates the vertices where the objects are placed, next to the vertices are given the values of the determinism estimates P_i. As a result, we get:

$$F_1 = 1 - mean(P) = 0{,}796 \tag{10}$$

$$F_2 = 1 - mean(P') = 0{,}780 \tag{11}$$

The value of F_1 is greater than F_2, which means that the placement based on the decomposition of the second option gives better results. This result is mainly due to the fact that the vertices in the subgraphs of the second variant of the decomposition are grouped more tightly.

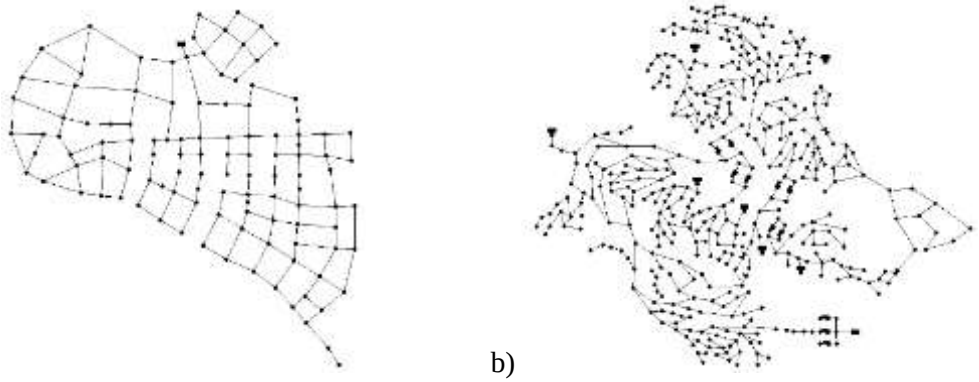
Thus, the use of a criterion that takes into account the length of distances in subgraphs is justified in solving the problems of placing objects on graphs. In this paper, we propose the following criterion for minimizing distances (12):

$$Dcut(C_1, \dots, C_k) = \sum_{i=1}^{k} \left(\frac{cut(C_i, G)}{vol(C_i)} * \frac{dist_i}{|C_i| * (|C_i| - 1)} * \frac{|C|}{|C_i|} \right) \rightarrow min \quad (12)$$

where $dist_i$ - is the sum of the distances between all vertices in subgraph i, $|C|$ - is the number of all nodes in the original graph.

4. Case study

The proposed solution, based on the application of a criterion that takes into account the distance lengths in subgraphs, was tested on the example of solving the problem of placing pressure sensors in water supply networks ZJ and D-Town of EPANET hydraulic modeling system (figure 4).



a) b)

Figure 4. Water supply networks: a. ZJ, b. D-Town.

ZJ is a network with 114 nodes and 164 pipes, D-Town has 407 nodes and 459 pipes. Nodes (consumers) of the considered networks are characterized by pressure determinism estimates, and edges (pipelines) are characterized by lengths L_j. After installing the next sensor in the network, the determinism estimates are recalculated by the formulas:

$$P_S = 1 \quad (13)$$

$$P_i = \max(P_y \cdot \alpha_1 \cdot \alpha_2^2 \cdot f(L_{i,y})) \quad (14)$$

where P_S - assessment of determination of pressure values in the node setup of the sensor, P_y - assessment of determination of pressure values of the node-neighbor, α_1 – the estimated error of determination of specific resistance of the pipeline, α_2 - estimating the error of determining the values of water consumption, $f(L_{i,y})$ - is a function of the length of the pipeline section to the next node.

The task is to arrange these sensors in the nodes in such a way that provides a minimum of the average value of the estimation of the uncertainty of pressure in the network (9).

For the ZJ network, options for installing sensors in the number from 1 to 10 are considered, for the D-Town network - from 1 to 20. Nodes with the highest centrality in the group are selected as sensor placement vertices.

Solutions are considered: trial and error (*TE*), greedy algorithm (*Gr*), algorithms based on spectral clustering (*SC*). Algorithms based on spectral clustering (*SC*) are considered in the context of using various criteria: *SCr* - spectral clustering by *Rcut* criterion, *SCn* - spectral clustering by *Ncut* criterion, *SCd* - spectral clustering by *Dcut* criterion. The criteria to assess the effectiveness of the algorithms: 1) average uncertainty estimates (*F*), 2) number of iterations ($\sum Iter$), 3) elapsed time (*T*), 4) accuracy rate ($1 - \bar{\delta}$), where δ is the relative error between the results of the considered algorithm and the algorithm of trial-and-error, 5) the highest relative error (*max(δ)*).

Tables 3 and 4 show the results of calculating the performance indicators of the algorithms. The best accuracy scores are shown in bold.

Table 3. Performance indicators of algorithms (ZJ network).

Indicator	*TE*	*Gr*	*SCr*	*SCn*	*SCd*
F	0,571	0,590	0,572	**0,570**	**0,570**
$\sum Iter$	1140	55	**10**	**10**	**10**
T, мин	18,9	**0,6**	1,7	1,8	1,8
$\mathbf{1-\bar{\delta}}$, %	100,0	96,8	100,1	**100,2**	**100,2**
max(δ),%	0,0	7,6	1,1	1,0	1,0

Table 4. Performance indicators of algorithms (D-Town network).

Indicator	*TE*	*Gr*	*SCr*	*SCn*	*SCd*
F	0,535	0,591	0,561	0,542	**0,539**
$\sum Iter$	8140	210	**20**	**20**	**20**
T, мин	188,5	**4,6**	6,5	6,6	6,7
$\mathbf{1-\bar{\delta}}$, %	100,0	89,5	95,2	98,7	**99,3**
max(δ),%	0,0	17,4	9,8	7,3	3,2

The solution obtained by the trial-and-error algorithm was chosen as the reference solution. The main problem of this algorithm is the necessary time and computational resources (1140 and 8140 iterations take 18.9 and 188.5 minutes, respectively). The fastest (0.6 and 4.6 minutes) and at the same time less accurate is greedy algorithm (96.8% and 89.5%). Algorithms based on spectral clustering showed close to the reference result at low computational cost (about 1.7-1.8 and 6.5-6.7 minutes). The application of the *Rcut* criterion provides a solution with an accuracy of 100.1% and 95.2%, the *Ncut* criterion - 100.2% and 98.7%. The best results among the methods of spectral clustering for the *Dcut* criterion are 100.2% and 99.3% in relation to the results of the trial-and-error algorithm.

Thus, the application of the proposed *Dcut* distance minimization criterion for graph decomposition allowed us to solve the problem of placing objects in the network with a high quality result, comparable and even better than spectral methods based on *Rcut* or *Ncut* criteria, which confirms the applicability of the *Dcut* criterion in spectral methods of graph clustering.

5. Conclusion

This article offers a look at the problems of solving the problems of placing objects in the network in the context of finding a solution in pre-defined subdomains obtained using the tools of the theory of spectral clustering of graphs by the criterion of minimizing distances in the desired subgraphs. The analysis of sources for solving the problem of placement in graphs, as well as methods of decomposition of graph structures are given. It is shown that many combinatorial problems on graphs can be solved with acceptable accuracy and in a short time, performing a search not in the entire set of graph elements, but on local sets grouped by a certain criterion. The proven theory of spectral clustering of graphs is proposed as a decomposition tool. Following the main provisions of this theory, the disadvantages of the applied criteria for splitting *Rcut* and *Ncut* are indicated. It is shown that the application of the *Dcut* distance minimization criterion proposed in this paper allows us to obtain good results when decomposing a graph. The obtained results are based on the examples of searching for sensor placement vertices in the known ZJ and D-Town networks of the EPANET hydraulic modeling system.

Further work will be aimed at studying the possibilities of applying the *Dcut* criterion in solving various combinatorial problems on graphs using spectral clustering methods.

6. References

[1] Mokshin, V.V. Recursive-regression self-organization of complex systems analysis and control models / V.V. Mokshin, I.M. Yakimov, R.M. Yulmetyev, A.V. Mokshin // Nelineiniy mir. – 2009. – Vol. 7(1). – P. 66-76.

[2] Mokshin, V.V. Parallel genetic algorithm for selection of significant factors influencing the evolution of a complex system // Vestnik Kazanskogo gosudarstvennogo tekhnicheskogo universiteta im. A.N. Tupoleva. – 2009. – Vol. 3. – P. 89-93.
[3] Yakimov, I.M. Modeling of complex systems in GPSS W simulation environment with advanced redactor / I.M. Yakimov, A.P. Kirpichnikov, V.V. Mokshin // Vestnik Kazanskogo tekhnologicheskogo universiteta. – 2014. – Vol. 17(4). – P. 298-303.
[4] Saifudinov, I.R. Optimization model of graph representation for allocation of significant structures on the example of visual data preprocessing / I.R. Saifudinov, V.V. Mokshin, P.I. Tutubalin, A.P. Kirpichnikov // Vestnik tekhnologicheskogo universiteta. – 2018. – Vol. 21(5). – P, 121-129.
[5] Galimova, G.I. Statistical analysis of the urban water consumption for administrative-business sector / G.I. Galimova, D.T. Yakupov // Amazonia Investiga. – 2018. – Vol. 7(17). – P. 414-425.
[6] Saifudinov, I.R. Method of allocation of significant areas in graphic information for decision support systems / I.R. Saifudinov, V.V. Mokshin, A.P. Kirpichnikov, P.I. Tutubalin, L.M. Sharnin // Vestnik tekhnologicheskogo universiteta. – 2018. – Vol. 21(6). – P. 146-152.
[7] Mokshin, V.V. Method of formation of complex system analysis model / V.V. Mokshin, I.M. Yakimov // Informatsionnie tehnologii. – 2011. – Vol. 5. – P. 46-51.
[8] Galimova, G.I. Application of Spectral Clustering Methods in Pipeline Systems Graph Models / G.I. Galimova, I.D. Galimyanov, D.T. Yakupov, V.V. Mokshin // Helix. – 2019. – Vol. 9(5). – P. 5607- 5614.
[9] Di Nardo, A. Sensor placement in water distribution networks based on spectral algorithms / A. Di Nardo, C. Giudicianni, R.Greco // EPiC Series in Engineering. – 2018. – Vol. 3. – P. 593-600.
[10] George, A. Computer Solution of Large Sparse Positive Definite Systems / A. George, J.W. Liu – Prentice-Hall, Englewood Cliffs, NJ, 1981.
[11] Schloegel, K. Graph partitioning for high performance scientific simulations / K. Schloegel, G. Karypis, V. Kumar // CRPC Parallel Computing Handbook, Morgan Kaufmann, 2001.
[12] Kernighan, B.W. An efficient heuristic procedure for partitioning graphs / B.W. Kernighan, S. Lin // The Bell System Technical Journal. – 1970. – Vol. 49(2). – P. 291-307.
[13] Fiduccia, C.M. A linear time heuristic for improving network partitions / C.M. Fiduccia, R.M. Mattheyses // Proc. 19th IEEE Design Automation Conference, 1982. – P. 175-181.
[14] Zheleznyakova, A.L. Efficient decomposition methods of unstructured adaptive grids for high-performance calculations in solving computational aerodynamics problems // Fiziko-himicheskaya kinetika v gazovoy dinamike. – 2017. – Vol. 18(1).
[15] Van Dongen, S. Graph Clustering by Flow Simulation // University of Utrecht, The Netherlands, 2000.
[16] Cvetkovic, D. Graph spectra in computer science / D. Cvetkovic, S. Simic // Linear Algebra Appl. – 2011. – Vol. 434. – P. 1545-1562.
[17] Fiedler, M. Algebraic connectivity of graphs // Czechoslov. Math. J. – 1973. – Vol. 23. – P. 298-305.
[18] Hagen, L. New spectral methods for ratio cut partitioning and clustering / L. Hagen, A. Kahng // IEEE Trans. Comput. - Aided Des. – 1992. – Vol. 11(9). – P. 1074-1085.
[19] Shi, J. Normalized cuts and image segmentation / J. Shi, J. Malik // IEEE Trans. Pattern Anal. Mach. Intell. – 2000. – Vol. 22. – P. 888-905.
[20] Ng, A.Y. On spectral clustering: Analysis and an algorithm / A.Y. Ng, M.I. Jordan, Y. Weiss // Advances in Neural Information Processing Systems (NIPS). – 2002. – Vol. 14. – P. 849-856.
[21] Ary-Catro, E. Spectral clustering based on local linear approximations / E. Ary-Catro, G. Chen, G. Lerman // Electronic Journal of Statistics. – 2011. – Vol. 5. – P. 1537-1587.
[22] Chen, G. Kernel spectral curvature clustering (KSCC) / G. Chen, S. Atev, G. Lerman // Dynamical Vision Workshop, IEEE 12th International Conference on Computer Vision – Kyoto, Japan, 2009. – P. 765-772.

[23] Chen, G. Spectral curvature clustering (SCC) / G. Chen, G. Lerman // J. Comput. Vision. – 2009. – Vol. 81. – P. 317-330.
[24] Guo, Q. Manifold clustering via energy minimization / Q. Guo, H. Li, W. Chen, I.-F. Shen, J. Parkkinen // ICMLA: Proceedings of the Sixth International Conference on Machine Learning and Applications – Washington, DC, USA, 2007. – P. 375-380.
[25] Luxburg U. Consistency of spectral clustering / U. Luxburg, M. Belkin // Ann. Statist. – 2008. – Vol. 36. – P. 555-586.
[26] Simon, H.D. Partitioning of unstructured problems for parallel processing // Comput. Syst. Eng. – 1991. – Vol. 2. – P. 135-148.
[27] Williams, R.D. Performance of dynamic load balancing algorithms for unstructured mesh calculations – Technical Report, California Institute of Technology, 1990.
[28] Chan, T.F. A framework for the analysis and construction of domain decomposition preconditioners / T.F. Chan, D.C. Resasco // Technical Report CAM-87-09 – University of California, Los Angeles, 1987.
[29] Spielman, D.A. Spectral partitioning works: Planar graphs and finite element meshes / D.A. Spielman, S.-H. Teng // Linear Algebra and its Applications. – 2007. – Vol. 421. – P. 284-305.
[30] Wei, T.-H. Algebraic Foundations of Ranking Theory – University of Cambridge, Cambridge, UK, 1952.
[31] Mohar, B. The Laplacian spectrum of graphs // Graph Theory, Combinatorics, and Applications – Wiley: Hoboken, NJ, USA, 1991. – P. 871-898.
[32] Luxburg, U. A tutorial on spectral clustering // Statistics and Computing. – 2007. – Vol. 17. – P. 395-416.