Date of publication DEC-31, 2023, date of current version DEC-20, 2023.
www.computingonline.net / computing@computingonline.net

Print ISSN 1727-6209
Online ISSN 2312-5381
DOI 10.47839/ijc.22.4.3360

Cluster Analysis of Information in Complex Networks

OKSANA KYRYCHENKO, SERHII OSTAPOV, IHOR MALYK
Yuriy Fedkovych Chernivtsi National University, Chernivtsi, 58012, Ukraine

Corresponding author: Oksana Kyrychenko (e-mail: o.kyrychenko@chnu.edu.ua).

ABSTRACT The research is devoted to the study of information in complex networks, namely: calculation of statistical characteristics and cluster analysis of data. Special software (crawler) was developed for direct data collection from the web space. In addition, a new structure of information technology has been developed for the collection, processing, and storage of large volumes of data collected from the web space. With the help of this structure, statistical characteristics of different segments of the web space (Ukrainian – edu.ua, Polish – edu.pl and Israeli – ac.il) are studied and their cluster structure is studied. The study of the cluster structure of web space zones was carried out using the spectral clustering algorithm of PIC (Rower iteration clustering). The results of the search for the optimal number of clusters using the "elbow" method and the k-Core decomposition method are presented, graphs illustrating the cluster structure of the investigated subnets are drawn. The paper also proposes a new approach to solving the problem of clustering and finding the optimal number of clusters when clustering objects are given by unstructured data (graphs) based on the spectral analysis of the stochastic matrix of the given graph. On this basis, a new method developed by the authors for determining the optimal number of clusters is proposed. Model examples are given and testing of the new method based on Monte Carlo simulation is performed. The optimal number of clusters was found by four methods: the "elbow" method, the k-core decomposition method, the silhouette method, and a new method developed by the authors. A conclusion is made concerning the accuracy of the developed new method, its advantages and disadvantages.

KEYWORDS complex networks; information system; random matrix; graph clustering; Monte Carlo method.

I. INTRODUCTION

TODAY, many scientific publications are devoted to the statistical cluster analysis of large volumes of data. A significant number of these publications study the structure and characteristics of the web space [1-5]. Most often, the structure of such data is presented using a graph. This approach turned out to be very successful for studying the characteristics of complex networks. Considering the web space as a directed graph, when static pages were taken as nodes, and the links between these pages were taken as arcs, [6] proved that the www space is a scale-free network. Similar studies were carried out in [3-11], which proved the high level of the worldwide network development. This was confirmed by the authors of [12], who studied the statistical characteristics of the national web domains of Brazil, Chile, Greece, Korea, and Spain. As for the Ukrainian segment of the web space, insufficient attention is paid to these issues [1, 13]. Nevertheless, based on the ensemble of the results of the above listed studies, it was possible to depict the structure of the web space [11-15].

However, collection and analysis of statistical characteristics of the huge data sets that characterize the modern web is becoming increasingly difficult: it is not easy to process millions of nodes with their connections and internal links. Therefore, in this study the clustering process to process such data will be used. One of the tasks of clustering consists in the reduction of the data dimensionality according to the characteristics by which objects are classified. The idea of all clustering methods is that the objects located in a specific cluster should be as similar as possible in terms of semantics. A large number of data partitioning approaches have been based on the clustering process [16-18].

Considering all the above, it is clear that clustering of large volumes of data greatly simplifies their analysis for various tasks: identification of the structure of a set of objects; simplification of further analysis of data arrays in order to make the necessary decisions; reduction of the amount of data storage in the case of an excessively large sampling; selection of atypical objects, which by their characteristics do not belong to any of the clusters, etc. Thus, the task set can be considered as a problem of clustering in Big Data, using limit theorems based on the theory of random matrices [19, 20]. These theoretical calculations can be used to divide Smart Grid systems into parts.

In all such cases, an adjacency matrix is built based on the initial data, which is then examined using various clustering methods. For example, in works [21-25] such studies were carried out with the involvement of spectral clustering methods.

However, for their work the most popular clustering methods require to enter the number of clusters, into which the original data set should be divided. In some problems this is not an issue, but very often there are cases where this quantity is unknown. There are also many methods for determining the "optimal" number of clusters, but there may arise certain issues in comparing the results of their work, prospects, ease of use and other characteristics.

Based on the above, the main aim of this study is to compare the effectiveness of different methods for determining the optimal number of clusters in the data set: the "elbow" method and k-core decomposition.

We also proposed a new method for finding the optimal number of clusters. The new method for determining the optimal number of clusters is based on the spectral decomposition of the transition matrix for the Markov process corresponding to the graph.

II. METHODS OF RESEARCH

To achieve the set goal, the following tasks were performed.

1. *Collecting and summarizing information from web pages.* To perform this task, we developed special software (the Java programming language was used), a detailed description of which can be found in [26]. The basis of this system is the crawler, the architecture of which is presented in Fig. 1.

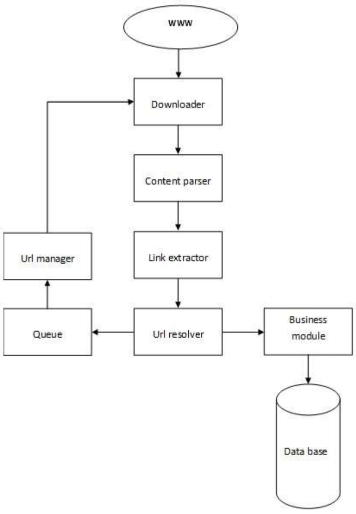


Figure 1. Generalized architecture of the developed crawler for collecting information from web pages

The developed software is completely controlled by the settings before starting work – this is its main advantage. The user can specify a list of web page addresses – entry points, and if necessary, the indexing depth, etc., which allows to fully control the process of searching and indexing pages, as well as calculating the main statistical parameters of the network under study. A brief description of functionality of the modules presented in Fig. 1 is as follows:

- *Downloader* – downloads the content of the page by the specified URL, processes the header of the request by the page.
- *Content parser* – parses the content of the downloaded web page, highlighting the structure of html-tags and the information being inside the tags.
- *Link extractor (link receiver)* – finds and extracts links from the content of the downloaded page, taking into account external and internal links.
- *Url resolver (link processor)* – transforms received links according to the rules of link normalization. Such transformations prevent the same resources from being downloaded more than once.
- *Business module* – ensures the preservation of the appropriate link structure in the database.
- *Url manager (link manager)* – generates a list of links for downloading.
- *Database* – used to save the web graph, i.e. pages, their link structure, various statistics. MySQL DBMS was used as a database management system.

Moving through the given web pages, the program collects the structure of links to other pages, analyzing the links that are on the web pages and the connections between these pages. The results of the investigation are written to the database, and the crawler moves on to the next entry point. If this "wandering" results in coming across a page that is already in the database, then re-examination is not carried out, only a new connection is added.

2. *Presentation of the information collected by the crawler in the form of a graph.* This task is performed using the GrafX framework on the Spark open source cluster computing system [27]. For the studies subnet, the degree of each node is established, the clustering coefficient is determined, and the probability distributions of nodes are constructed based on input and output connections. By combining the calculated dependencies for the input and output subnets, we can obtain the statistical characteristics of the undirected graphs of the web pages of the studied zones of the web space.

3. *Determination of the optimal number of clusters and their centers by the chosen method.* After constructing the graph, we use the selected methods (k-Core decomposition [28] or the "elbow" method [29, 30]) to determine the optimal number of clusters and cluster centers for the studied segment of the web space.

To implement the "elbow" method, we used MLib (Machine Learning Library) from the Spark framework. It involves the following steps [30]:

- *k-means* clustering is performed, while the value of the penalty function is calculated and recorded.
- A graph of the dependence of the penalty function on the given number of clusters is constructed.
- We select the number of clusters at which the greatest bend in the graph occurs. This number will be the optimal number of clusters.

Another way to determine the optimal number of clusters is the *k*-Core decomposition method.

GraphX on Spark was also used for this. Such a decomposition enables to quickly find the core of the graph, that is, the maximally connected subgraph, in which each vertex is connected to at least *k* vertices in the subgraph. The *k*-core schedule is often used in large-scale network analysis. This is an *O(m)* algorithm, where *m* is the number of threads in non-parallel calculations [28]. Its main goal is to find a strong subgroup, the members of which play the role of communicators on the graph. Each node in the subgraph should have at least *k* degree.

k-Core decomposition has the following properties:

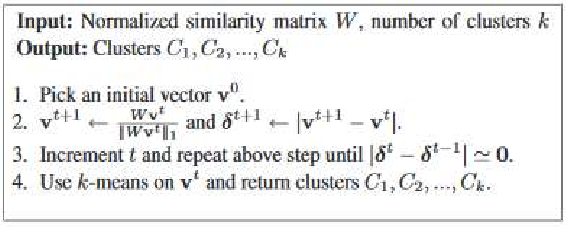
$$\forall\, u \in V : k-core(u)=k \leftrightarrow \begin{cases} \text{There is such maximal subgraph } V_k \text{ that} \\ \forall v \in V_k : \deg(v) \ge k, \\ \text{and} \\ \text{There is no such subgraph } V_{k+1} \text{ that} \\ \forall v \in V_{k+1} : \deg(v) \ge k+1. \end{cases} \tag{1}$$

The pseudocode of the method looks like this:

```
Procedure k-core decomposition
    Input
    Graph: data in vertex is (degree, bool)
    Output
     Graph: data in vertex is K
    Pseudo Code
    While
                Initial the Pregel send initial MSGs to all node
            Graph.Vertex update (intitial MSGs)
              MSGs = message merge (all message sent)
             While messages.count > 0
                 Graph.Vertex update (messages)
                 MSGs = message merge (all message sent)
         End while
             K += 1
    End while
    Vertex update stage(messages)
                If (message.bool ) messages.degree =
      max(origin, new degree)
                 Message.bool = origin && new bool
    Message send stage
    If ( ! v1.bool || ! v2.bool)
                 empty
    Else if(v1.degree == k && v2.degree > k)
                 Send to v2 (1,true)
                 Send to v1 (0,false)
    Else
                 Iterator.empty
    Message merge stage
    (Sum, a.bool && b.bool)
```

4. Carrying out clustering. This stage was performed by us using the PIC-clustering method (Power Iteration Clustering).

In the PIC algorithm, the search for centroids of clusters is performed on the basis of the Peron-Frobenius eigenvector of the normalized matrix W. Moreover, the search for this eigenvector is based on iterative methods [21]. The PIC algorithm pseudocode is given below:



```
Input: Normalized similarity matrix $W$, number of clusters $k$
Output: Clusters $C_1, C_2, ..., C_k$

1. Pick an initial vector $\mathbf{v}^0$.
2. $\mathbf{v}^{t+1} \leftarrow \frac{W\mathbf{v}^t}{\|W\mathbf{v}^t\|_1}$ and $\delta^{t+1} \leftarrow |\mathbf{v}^{t+1} - \mathbf{v}^t|$.
3. Increment $t$ and repeat above step until $|\delta^t - \delta^{t-1}| \simeq \mathbf{0}$.
4. Use $k$-means on $\mathbf{v}^t$ and return clusters $C_1, C_2, ..., C_k$.
```

As a result of performing such tasks, a partition of the collected data set into clusters is obtained, and it is possible to compare the performance of different clustering algorithms.

Graphically, the sequence of performing the specified tasks with the involvement of the selected methods is presented in Fig. 2.

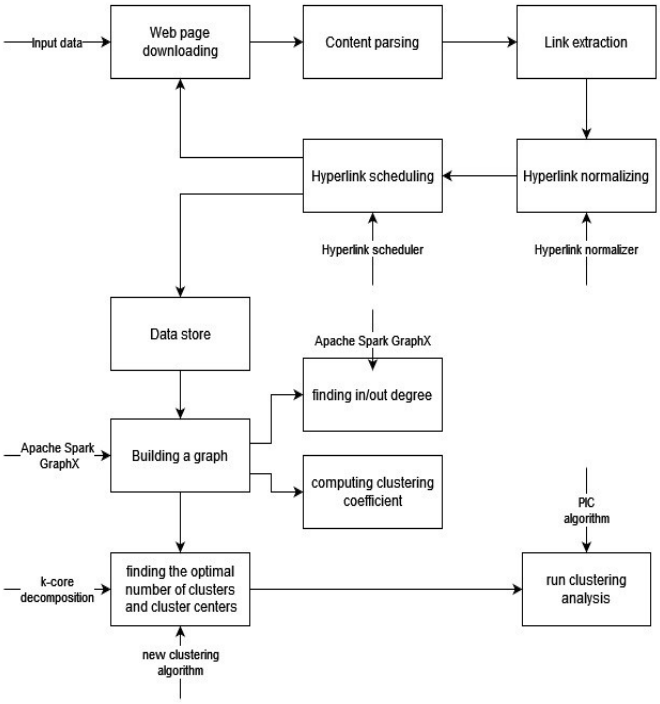


Figure 2. Generalized scheme of performing the statistical cluster analysis in complex networks

Performing all the specified steps results in obtaining the cluster structure of the data sets received and possibility to analyze their cluster characteristics. A more detailed description of the information technology obtained is given in [31].

III. THE RESULTS OF THE CONDUCTED RESEARCH AND THEIR DISCUSSION

Using the developed methodology, based on statistical methods, a cluster study of the next web space areas was performed: academic segments of Ukraine (*edu.ua*), Poland (*edu.pl*) and Israel (*ac.il*). For each segment, the probability distribution of nodes by input and output connections was constructed, network clustering coefficients were calculated, average values of the node degree for undirected graphs were constructed and determined [32]. The results of the statistical analysis of the obtained results are presented in Fig. 3-6.

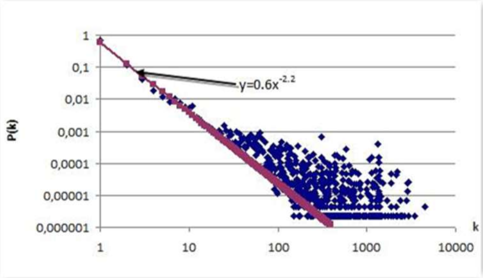


Figure 3-1. Probability distribution of nodes by degrees on input links for the *edu.ua* zone

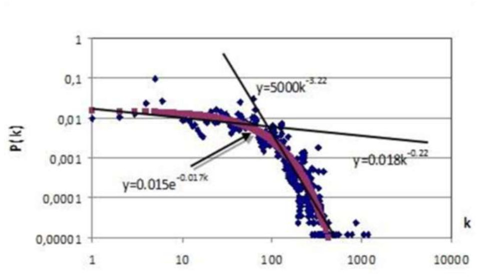


Figure 3-2. Probability distribution of nodes by degrees on output links for the *edu.ua* zone

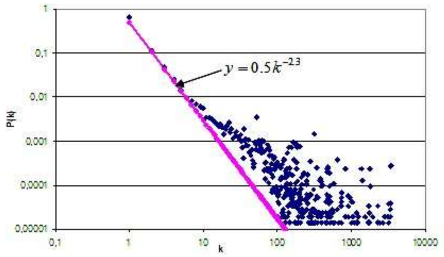


Figure 4-1. Probability distribution of nodes by degrees on input links for the *edu.pl* zone.

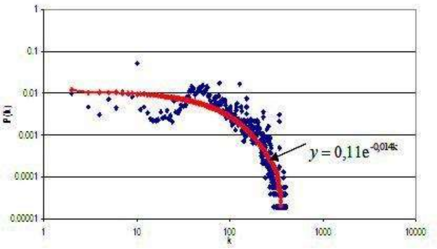


Figure 4-2. Probability distribution of nodes by degrees on output links for the *edu.pl* zone

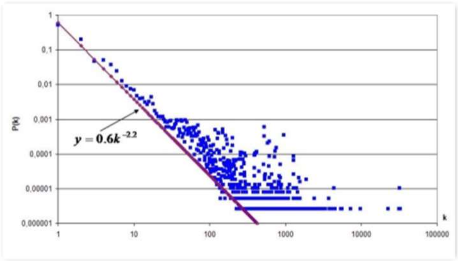


Figure 5-1. Probability distribution of nodes by degree on the input links for the *ac.il* zone

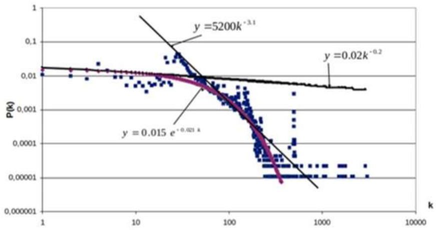


Figure 5-2. Probability distribution of nodes by degree on the output links for the *ac.il* zone

Figures 3-5 show that all the academic areas of the web space of the three countries have very similar statistical characteristics. According to input connections, all zones studied correspond to the "small world" concept: the probability distribution shows a power dependence with indicators (-2.2)÷(-2.3). In addition, if the Israeli and Polish academic segments can be considered fully developed networks, then the same can be said about the Ukrainian academic segment of the web space, because its statistical characteristics are quite similar to the first two.

A similar situation is observed in terms of output connections: all three studied academic zones show similar statistical characteristics and are best approximated by exponential dependencies with fairly close exponents: from -0.014 (*edu.pl*), -0.017 (*edu.ua*) to -0.021 (*ac.il*). Therefore, here too it can concluded that the Ukrainian academic zone is also in full correspondence to the development trends of the web space. Nevertheless, unlike previous researchers, we cannot claim that the statistics of the original connections also corresponds to the power law, as it was asserted in [1, 12-13].

Figures 3-5 prove that the investigated networks consist mainly of nodes that have a significant number of outgoing connections, i.e. the average degree of the node is shifting towards large values. The estimation of average degrees for these networks gives results in the range ⟨k⟩=55÷120, confirming our conclusion. These values correlate with the area of change of the approximating straight lines, which occurs at k≈100. If we formally calculate the average value of the node ⟨k⟩ degree over the input links, then, as expected, it is shifted towards small values, which are several times smaller than those of the output links for the corresponding networks.

It should be noted that the specified statistical characteristics are not a feature of the academic zone of *edu.ua*, but are characteristic of other zones of the Ukrainian segment of the web space.

As a result of the conducted research, the clustering coefficients of the studied segments were also obtained (Table 1). The data indicate a large number of nearest neighbor cross-references [32].

Table 1. Clustering coefficients for subnetworks

Name of zone	edu.ua	ac.il	edu.pl
Clustering coefficient	0,11	0,104	0,088

It should be mentioned that for all networks, the clustering coefficients fluctuate around the value of 0.1, which also indicates similar statistical characteristics of all studied segments.

The study of the cluster structure of the obtained data sets started with determining the optimal number of clusters using the two methods described in the previous section: the "elbow" method and *k*-core decomposition.

Fig. 6-8 present the obtained graphs of the penalty function for all investigated zones of the web space.

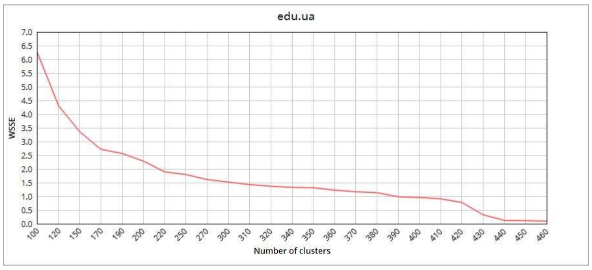


Figure 6. Dependence of the penalty function on the number of clusters for the *edu.ua* zone

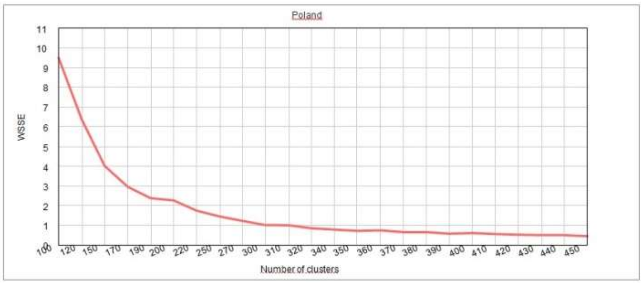


Figure 7. Dependence of the penalty function on the number of clusters for the *edu.pl* zone

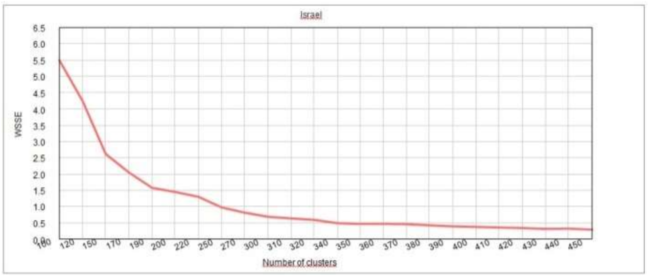


Figure 8. Dependence of the penalty function on the number of clusters for the *ac.il* zone

One can see that all the presented dependencies are almost of the same nature. After determining the point of the greatest inflection of the graphs from the above dependencies, the optimal number of clusters is obtained: 220 – for *edu.ua*; 210 – for *edu.pl*; 190 – for *ac.il*. Again, as it can be observed, the number of clusters calculated by the "elbow" method turned out to be approximately the same for all the investigated segments of the web space.

The results of determining the optimal number of clusters by the k-core decomposition method and comparison with those obtained using the "elbow" method are presented in Table. 2.

Table 2. Optimal number of clusters using "elbow" and k-Core decomposition methods

Name of web space segment	**'Elbow" method**	***k*-core decomposition method**
Ukrainian web space segment (*edu.ua*)	220	229
Polish web space segment (*edu.pl*)	210	213
Israeli web space segment (*ac.il*)	190	196

It can be seen from Table 2 that we obtained a good agreement of the optimal number of clusters for all investigated areas of the web space, performed by different methods.

Therefore, only clustering is to be performed, for which the PIC method is used. The results of it are presented in Fig. 9-11.

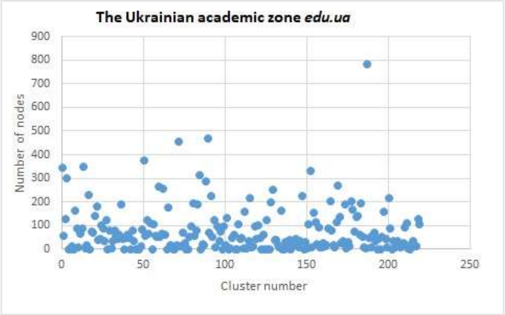


Figure. 9. Cluster structure of the Ukrainian academic zone *edu.ua* (220 clusters)

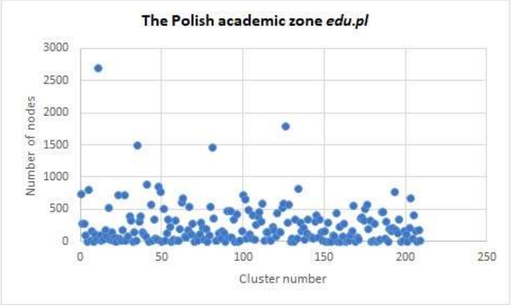


Figure 10. Cluster structure of the Polish academic zone *edu.pl* (210 clusters)

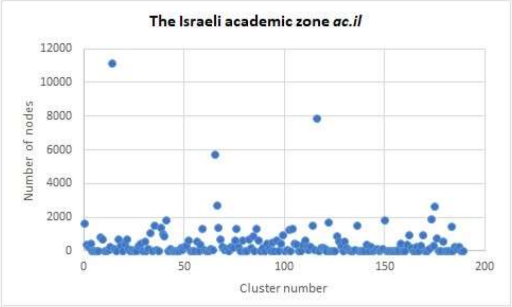


Figure 11. Cluster structure of the Israeli academic zone *ac.il* (190 clusters)

The analysis of Figures 9-11 shows that the Israeli academic segment has the most homogeneous (developed) cluster structure: the predominant number of nodes in its clusters is up to 2,000, although there are clusters with 6,000-11,000 nodes. The Polish *edu.pl* zone consists mainly of clusters containing up to 1000 nodes, but there are also clusters of 1500-2700 nodes. As Fig.9 shows, the Ukrainian segment of *edu.ua* is the least developed structure: the main number of clusters contains up to 300 nodes, and the largest cluster contains approximately 800 nodes, which is significantly less than the Polish one, and moreover, than the Israeli zone.

Obviously, such results indicate that although the Ukrainian segment of the web space shows statistical characteristics similar to those of the networks of other countries, it should be developed in the direction of increasing the Internet representation of educational institutions, and especially, in the formation of connections between websites of various educational institutions forming the space of *edu.ua*.

IV. A NEW METHOD FOR FINDING THE OPTIMAL NUMBER OF CLUSTERS

In [33], the authors present a new method for determining the optimal number of clusters in the network, which is based on the spectral properties of the transition matrix of the corresponding Markov chain. It should be pointed out that the main idea of the method construction is based on the asymptotic properties of the spectrum of random matrices, which in turn is one of the mathematical models of Big Data systems [19, 20]. The basis of the method is the fact that for a Markov process, the multiplicity of the eigenvalue λ=1 for the transition probability matrix coincides with the number of connecting classes. In addition, for homogeneous clusters of large dimensions based on Wigner's circular law, almost all eigenvalues will be concentrated in a circle of radius $R = O\left(\frac{1}{\sqrt{N}}\right)$, where N is the number of vertices in the cluster.

Thus, the research will be based on the spectral properties of the transition probability matrix

$$P_{ij} = \frac{A_{ij}}{\sum_{j=1}^{N} A_{ij}},$$

where matrix A is the adjacency matrix for the system, in which A_{ij} is equal to the number of transitions from node i to node j. To determine the optimal number of clusters, the following ratio was proposed in [33].

$$\mathrm{k_{opt}} = \#\left\{\lambda_\mathrm{i}: |\lambda_\mathrm{i}(\mathrm{P}) - 1| \le \frac{1}{\alpha\sqrt{\mathrm{N}}}\right\},$$

where the parameter α depends on the distribution of elements $\mathrm{A_{ij}}$. However, in practice, determining the value of α is time-consuming, so here the ratio will be used

$$\mathrm{k_{opt}} = \#\left\{\lambda_\mathrm{i}: \mathrm{Re}(\lambda_\mathrm{i}) > \max_{\mathrm{i}=1,\dots,\mathrm{N}} |\mathrm{Im}(\lambda_\mathrm{i})|\right\}.$$

It should be noted that for sufficiently large clusters based on Wigner's circular law [34-35],

$$\max_{\mathrm{i}=1,\dots,\mathrm{N}} |\mathrm{Im}(\lambda_\mathrm{i})| = \mathrm{O}\left(\frac{1}{\sqrt{\mathrm{N}}}\right).$$

Let us test the considered new method based on Monte Carlo simulation with $M = 10^4$ realizations. In the simulation, we will use $N = 10^2, \dots, 10^5$ vertices with $S_1, \dots, S_k$ clusters. We will use Several basic parameters will also be used for the analysis:

1. λ_1 – the Poisson distribution parameter, based on which the number of ties for the vertices of the graph located in different clusters is modeled,
$$A_{ij} \sim Pois(\lambda_1), i \in S_m, j \in S_l, m \neq l.$$
2. λ_2 – the Poisson distribution parameter, based on which the number of ties for the vertices of the graph located in the same cluster is modeled, i.e.
$$A_{ij} \sim Pois(\lambda_1 + \lambda_2), i, j \in S_m.$$
3. (L, H) – the minimal and maximal numbers of objects in the cluster.
4. k – the number of clusters used in the simulation.

To search for the optimal number of clusters using four methods will be used: the "elbow" method, the *k*-core decomposition method, the silhouette method [36, 37] and a new method developed by the authors.

The results of the simulation, i.e. examples of finding the optimal number of clusters, are presented in Table 3.

Table 3. Simulation results of the average number of clusters and the root mean square deviation from the average number $\mu(\sigma)$

Modeling hyperparameters $(\lambda_1, \lambda_2, L, H, k)$	"Elbow" method	k-core decom-position	Silhouette method	New method
$(10, 1, 200, 300, \mathbf{20})$	17.4 (4.25)	16.1 (7.3)	16.4 (5.3)	10 (8)
$(10, 2, 200, 300, \mathbf{20})$	18.5 (2.31)	18.2 (3.74)	17.4 (5)	19 (5)
$(10, 5, 200, 300, \mathbf{20})$	19.1 (2.21)	19.3 (2.35)	18.3 (4.5)	20 (0.5)
$(10, 10, 200, 300, \mathbf{20})$	19.4 (1.85)	19.5 (1.35)	19 (2.83)	20 (0.23)
$(10, 20, 200, 300, \mathbf{20})$	19.5 (1.11)	19.8 (0.98)	19.4 (1.95)	20 (0.16)
$(10, 40, 200, 300, \mathbf{20})$	19.6 (0.65)	19.86 (0.43)	19.71 (1.03)	20 (0.09)
$(10, 100, 200, 300, \mathbf{20})$	19.98 (0.115)	19.99 (0.09)	19.87 (0.26)	20 (0.02)
$(10, 1, 200, 300, \mathbf{150})$	173 (25.7)	164 (12.4)	171 (25.4)	78 (53)
$(10, 5, 200, 300, \mathbf{150})$	161 (15.3)	157 (9.21)	163 (18.4)	144 (7.9)
$(10, 20, 200, 300, \mathbf{150})$	156 (7.89)	153.4 (7.31)	158.3 (13.23)	148 (4.28)
$(10, 40, 200, 300, \mathbf{150})$	153 (4.26)	152.1 (4.16)	154.5 (9.31)	149.3 (2.11)
$(10, 40, 20, 300, \mathbf{150})$	156 (5.73)	154.3 (10.23)	145.5 (15.18)	144.3 (13.7)

As these calculations show, for small values of the $\frac{\lambda_2}{\lambda_1}$ ratio, the estimations of the optimal number of clusters based on the "elbow" method, k-core decomposition, and the silhouette method are more accurate than that of the new method. However, at $\frac{\lambda_2}{\lambda_1} > 30\%$, the proposed method shows a smaller error in determining the exact number of clusters, as well as a smaller standard deviation compared to the other methods. One can see the same results in the next two figures. Fig. 12 shows that the width of the reliability interval significantly decreases for the new method as the ratio $\frac{\lambda_2}{\lambda_1}$ increases. For the other three methods, the narrowing of the width of the reliability interval occurs more smoothly, i.e. the error for these methods is higher.

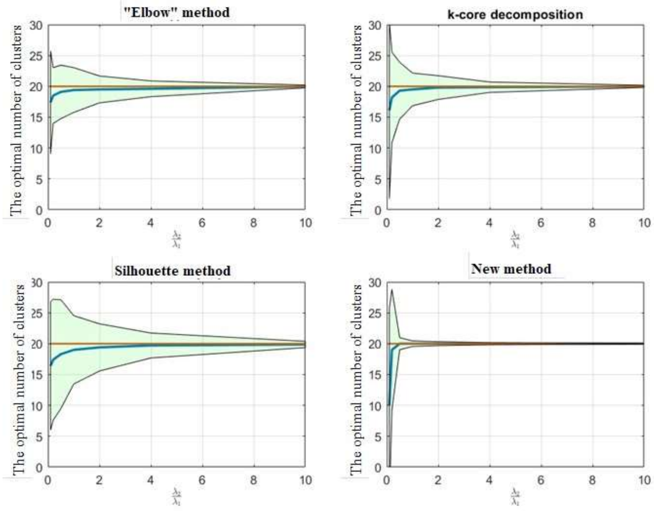


Figure 12. The optimal number of clusters determined by four methods together with a 95% confidence interval

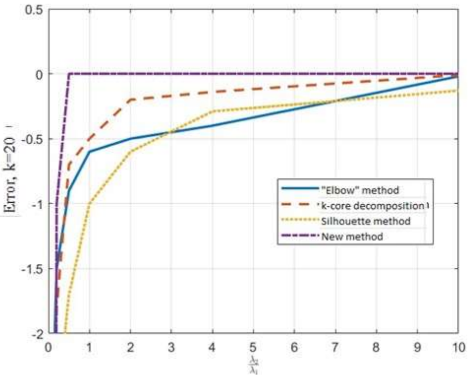


Figure 13. Clustering error for four clustering methods at k=20

As Fig. 13 shows, the convergence of the optimal number of clusters obtained by the new method occurs much faster with the growth of the $\frac{\lambda_2}{\lambda_1}$ ratio, and when $\frac{\lambda_2}{\lambda_1} > 1$, all four methods will give the correct result on average. According to our assumptions, this means that if the average number of transitions in clusters is twice as large as that between the clusters, all four methods will give the same result. However, in the range $\frac{\lambda_2}{\lambda_1} \in (0.3; 1)$, the proposed new method demonstrates better results.

The disadvantages of the proposed method include the fact that a significant variation in the size of the clusters can lead to a decrease in the average value of $\mathrm{k_{opt}}$. This result follows from the last line of Table 3. Therefore, it can be concluded that the proposed new method is sensitive to the presence of clusters of small dimensions.

V. CONCLUSIONS

The research performed led to the following conclusions:

1. The methodology of statistical cluster analysis of information in complex networks has been developed. It consists of the stages of collection, statistical data processing and carrying out the clustering process.

2. A crawler has been developed, which has a modular architecture, is fully configured by the user and allows him to control the process of data collection in the web space.

3. The statistical characteristics and cluster structure of the Ukrainian educational segment *edu.ua*, the Polish subnet *edu.pl* and the Israeli academic zone *ac.il* were studied using the developed methodology. A comparative analysis was carried out. It is shown that the Ukrainian subnet *edu.ua* demonstrates statistical characteristics similar to those of the others, but there lack of the nodes of educational institutions is observed, as demonstrated by the cluster analysis. The number of nodes in the clusters is the smallest of all other studied segments.

4. Two methods of determining the optimal number of clusters in the data set were compared: the "elbow" method and k-core decomposition. Both methods showed almost the same results. Nevertheless, we consider the k-core decomposition method to be the most promising for cluster analysis, as there already exist the algorithms enabling to change the k-core quickly at the dynamic change of the corresponding segment of the web space with no need to traverse the entire graph, unlike the "elbow" method.

5. For determining the optimal number of clusters, the authors developed a new method based on the spectral distribution of the transition matrix for the Markov process corresponding to the graph. The new method of finding the optimal number of clusters showed greater (higher) accuracy compared to other known methods ("elbow" method, silhouette method and k-core decomposition). It should be noted that for the range $\frac{\mu_W}{\mu_B} \in (1.2, 1.4)$ the proposed method has a higher accuracy than other methods, where μ_W is the average number of connections in a cluster, μ_B is the average number of connections between clusters in the case of the Poisson distribution of connections.

6. The disadvantages of the proposed method include several important factors. Firstly, in the theoretical explanations it is assumed that the clusters are of the same type, i.e. the distribution in the clusters is the same or at least has the same average values, i.e.

$$EA_{ij} \approx c; \;\; i,j \in S_m, m = 1, \dots k.$$

Secondly, the developed method is sensitive to the presence of small clusters that may be outliers. This drawback can be eliminated by considering the removal of outliers at the stage of data preprocessing. Thirdly, the method works best when the number of clusters is proportional to that of nodes as $O\left(\frac{N}{\log(N)}\right)$, that is, the number of clusters grows much more slowly than the number of new nodes.

In further research, it is planned to use a new clustering method for real Big Data systems, including Smart Grid systems.

References

[1] Yu. Holovatch, O. Olemskoi, C. von Ferber, T. Holovatch, O. Mryglod, I. Olemskoi, V. Palchykov, "Complex networks," *Journal of Physical Studies,* vol. 10, no. 4, pp. 247–289, 2006. https://doi.org/10.30970/jps.10.247

[2] M. E. J. Newman, "The structure of scientific collaboration networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 2, pp. 404–409, 2001. https://doi.org/10.1073/pnas.98.2.404.

[3] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, issue 2, pp. 167–256, 2003. https://doi.org/10.1137/S003614450342480.

[4] S. H. Strogatz, "Exploring complex networks," *Nature*, vol. 410, pp. 268-276, 2001. https://doi.org/10.1038/35065725.

[5] M. E. J. Newman, "Models of the small world," *J. Stat. Phys,* vol. 101, pp. 819–841, 2000. https://doi.org/10.1023/A:1026485807148.

[6] A. Broder, R. Kumar, F. Maghoul et al. "Graph structure in the web," *Proceedings of the 9th World Wide Web Conference on Computer networks*, 2000, vol. 33 (1), pp. 309-320.

[7] A. Barrat, M. Weight, "On the properties of small-world networks models," *The European Physical Journal,* vol. 13, pp. 547–560, 2000. https://doi.org/10.1007/s100510050067.

[8] D. J. Watts, S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature,* vol. 393, pp. 440–442, 1998. https://doi.org/10.1038/30918.

[9] L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley, "Classes of small-world networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 21, pp. 11149–11152, 2000. https://doi.org/10.1073/pnas.200327197.

[10] D. J. Watts, *Small Worlds: The Dynamics of Networks Between Order and Randomness,* Princeton University Press, 1999, 262 p. https://doi.org/10.1515/9780691188331.

[11] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, issue 2, pp. 167-256, 2003. https://doi.org/10.1137/S003614450342480

[12] R. Baeza-Yates, C. Castillo, E. N. Efthimiadis, "Characterization of national web domains," *Journal ACM Transactions on Internet Technology*, vol. 7, no. 2, pp. 9–33, 2007. https://doi.org/10.1145/1239971.1239973.

[13] V. V. Pasichnyk, N. M. Ivanushchak, "Research and modelling of complex networks," *Eastern-European Journal of Enterprise Technologies*, vol. 2, no. 3(44), pp. 43–48, 2010.

[14] J. M. Kleinberg, "Navigation in a small world," *Nature,* vol. 406, p. 845, 2000. https://doi.org/10.1038/35022643

[15] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, "Random graphs with arbitrary degree distribution and their applications," *Physical Review* E 64, 026118, 2001. https://doi.org/10.1103/PhysRevE.64.026118.

[16] N. Mishra, R. Schreiber, I. Stanton, and R. Tarjan, "Clustering social networks," In Anthony Bonato and Fan R. K. Chung, editors, *Algorithms and Models for the Web-Graph, Vol. 4863 of Lecture Notes in Computer Science*, chapter 5, pp. 56–67. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[17] P. Domingos and M. Richardson, "Mining the network value of customers," *Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'01*, New York, pp. 57–66, 2001. https://doi.org/10.1145/502512.502525.

[18] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, issues 3-5, pp. 75–174, 2010. https://doi.org/10.1016/j.physrep.2009.11.002

[19] R. C. Qiu, P. Antonik, *Smart Grid using Big Data Analytics: A Random Matrix Theory Approach,* Wiley, 2017, 632 p.

[20] T Tao, *Topics in Random Matrix Theory*, American Mathematical Society, 2023. 282 p.

[21] F. Lin and W. W. Cohen, "Power iteration clustering," *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, Haifa, Israel, June 21-24, 2010, pp. 1-10.

[22] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, pp. 395–416, 2007. https://doi.org/10.1007/s11222-007-9033-z.

[23] T. Xiang and S. Gong, "Spectral clustering with eigenvector selection," *Pattern Recognition*, vol. 41, issue 3, pp. 1012–1029, 2008. https://doi.org/10.1016/j.patcog.2007.07.023

[24] A. Y. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," *Advances in Neural Information Processing Systems 14 (NIPS 2001*), 2001, pp. 1-8.

[25] T. C. Ramos, J. Mourão-Miranda and A. Fujita, "Spectral density-based clustering algorithms for complex networks," *Front. Neurosci*., vol. 17, article 926321, 2023. https://doi.org/10.3389/fnins.2023.926321.

[26] O. Kyrychenko, "Features of software architecture for collecting and analyzing statistical information on the global network," *Information Technology: Computer Science, Software Engineering and Cyber Security*, no. 2, pp. 107–112, 2023. https://doi.org/10.32782/it/2023-2-13.

[27] R. Xin, J. Gonzalez, M. Franklin, and I. Stoica "GraphX: A resilient distributed graph system on spark," *AMPLab, EECS*, UC Berkeley 2013. https://doi.org/10.1145/2484425.2484427

[28] S.-T. Cheng, Y.-C. Chen, and M.-S. Tsai, "Using k-Core decomposition to find cluster centers for k-Means algorithm in GraphX on Spark," *Proceedings of the 2017 Eighth International Conference on Cloud Computing, GRIDs, and Virtualization Cloud Computing'2017*, 2017, pp. 93-98.

[29] A. Kuraria, N. Jharbade, & M. Soni, Manish, "Centroid selection process using WCSS and elbow method for K-Mean clustering algorithm in data mining," *International Journal of Scientific Research in Science, Engineering and Technology*, vol. 4, issue 11, pp. 190-195, 2018. https://doi.org/10.32628/IJSRSET21841122

[30] Determining the number of clusters in a data set. [Online]. Available at: https://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set

[31] O. Kyrychenko, "Information technology for statistical cluster analysis of information in complex networks," *Computer Systems and Information Technologies*, vol. 4, pp. 47–51, 2022. https://doi.org/10.31891/csit-2022-4-7

[32] O. L. Kyrychenko, S. Ostapov, I. Kanovsky, "Investigation of the certain internet domain statistical characteristics," *Eastern-European Journal of Enterprise Technologies*, vol. 6, no. 12(66), pp. 91–96, 2014. https://doi.org/10.15587/1729-4061.2013.19698

[33] O. L. Kyrychenko, I. V. Malyk, & S. E. Ostapov, "Stochastic models in artificial intelligence development," *Bulletin of Taras Shevchenko National University of Kyiv. Physics and Mathematics*, no. 2, pp. 53–57, 2021. https://doi.org/10.17721/1812-5409.2021/2.7

[34] E. P. Wigner, "On the distribution of the roots of certain symmetric matrices," *The Annals of Mathematics*, Second Series, vol. 67, no. 2, pp. 325-327, 1958. https://doi.org/10.2307/1970008

[35] E. P. Wigner, "Characteristic vectors of bordered matrices with infinite dimensions," *The Annals of Mathematics*, second series, vol. 62, no. 3, pp. 548-564, 1955. https://doi.org/10.2307/1970079

[36] L. Lenssen, E. Schubert, Erich, "Clustering by direct optimization of the medoid silhouette," *Proceedings of the International Conference on Similarity Search and Applications (SISAP'2022*), 2022, pp. 190–204. https://doi.org/10.1007/978-3-031-17849-8_15

[37] F. Batool, C. Hennig, "Clustering with the average silhouette width," *Computational Statistics & Data Analysis,* vol. 158, 2021. https://doi.org/10.1016/j.csda.2021.107190

OKSANA KYRYCHENKO, Assistant Professor, Department of Mathematical Problems of Control and Cybernetics, Yuriy Fedkovych Chernivtsi National University, Chernivtsi, Ukraine.

Field of scientific interests: research of complex networks, social networks, complex networks and their statistical characteristics, engaged in cluster analysis, machine learning.

https://orcid.org/0000-0003-0282-9958
e-mail: o.kyrychenko@chnu.edu.ua

PROF. SERHII OSTAPOV, Doctor of Physical and Mathematical Sciences. Head of the Computer Systems Software Department of Yuriy Fedkovych Chernivtsi National University.

Field of scientific interests: information security, analysis of complex networks, software for scientific research, modeling based on cellular automata.

https://orcid.org/0000-0002-4139-4152
email: s.ostapov@chnu.edu.ua.

PROF. IHOR MALYK,

Doctor of Physical and Mathematical Sciences, Professor, Department of Mathematical Problems of Control and Cybernetics, Yuriy Fedkovych Chernivtsi National University, Chernivtsi, Ukraine.

Field of scientific interests: stochastic analysis, financial mathematics, machine learning, simulation of random processes.

https://orcid.org/0000-0002-1291-9167
email: i.malyk@chnu.edu.ua