Formation of Network of Scientists in Cybersecurity Field

Dmytro Lande[1,2], Andrii Snarskii[1,2], Oleh Dmytrenko[2,1], Chen Li[3,4], Xianyi Li[3,4], and Jianping Guo[4]

[1]National Technical University «Igor Sikorsky Kyiv Polytechnic Institute», 37, Prosp. Peremohy, Kyiv, 03056, Ukraine
[2]Institute for Information Recording of National Academy of Sciences of Ukraine, 2, Mykoly Shpaka Street, Kyiv, 03113, Ukraine
[3]Qilu University of Technology (Shandong Academy of Sciences), Daxue Road 3501, Changqing District, Jinan, 250353, Shandong Province, China
[4]Information Research Institute (Shandong Academy of Sciences), Jinan, 250353, Shandong Province, China

Abstract

This work considers the networks of scientists, which take into account not only the relationship of co-authorship, but also the thematic proximity of their scientific interests. The unique feature of the presented approach is its use of a typical scientometric service and consideration of tags or descriptors of topics attributed by scientists to themselves and other authors of articles indexed by this scientometric service. During the implementation of this approach, a special algorithm is used to scan the resources of the scientometric service and obtain a representative set of authors or co- authors as network nodes. The weight of the connection between scientists in the considered network is determined by the meaningful correlation of their scientific fields, which is measured by the number of matching descriptors. Clustering algorithms enable the identification of groups of highly connected nodes that correspond to scientific schools and teams of scientists capable of collaborating on joint projects. The software implementation of the proposed approaches and methods uses the Perl and Python programming languages, publicly available information scanning utilities, and Gephi graph analysis and visualization software.

Keywords

Network of scientists, co-authorship network, scientometric service, information network scanning, topic descriptors, cluster analysis, cybersecurity.

1. Introduction

As a result of the development of scientific information systems, new opportunities have appeared, allowing us to assess the level of scientists, scientific schools and to study the patterns of scientific interaction [1].

At this time, the task of selecting expert groups, forecasting the joint work of scientists in various fields [2], in particular, in the field of cyber security, is relevant. Considering the relationship of common scientific interests of different scientists and/or co-authorship, it is possible to form networks that can be used to solve this problem. Networks of co-authors are already a traditional tool for studying the regularities of scientific cooperation, with the help of which it is possible to obtain not only scientometric assessments but also to identify experts for solving complex tasks. The largest scientific information services allow scientists to create their own profiles containing relevant scientometric information. Numerous works are dedicated to the study of networks of co-authors, as

XXII International Scientific and Practical Conference "Information Technologies and Security (ITS-2022)", November 16, 2022, Kyiv, Ukraine
EMAIL: dwlande@gmail.com (D. Lande); asnarskii@gmail.com (A.Snarskii); dmytrenko.o@gmail.com (O. Dmytrenko); 413886457@qq.com (C. Li); sdly_lxy@163.com(X. Li); jianpingdou@126.com (J. Gou)
ORCID: 0000-0003-3945-1178 (D. Lande); 0000-0002-4468-4542 (A. Snarskii); 0000-0001-8501-5313 (O. Dmytrenko); 0000-0001-6994-1253 (C. Li);0000-0001-9253-3370 (X. Li); 0000-0001-6012-1195 (J. Gou)

well as the Google Scholar service [3]. This fact confirms the relevance of the performed research. The task of building and researching co-authorship networks, as well as citation networks, is one of the first tasks of scientometrics, which is still relevant at this time. Modern scientometric services are based on the methods for forming networks of co-authors, determining significant nodes, network structure, citation research, as well as relevant corpora, etc. In particular, work [4] provides a method for assessing the importance of nodes in this network, which is based on the improved PageRank algorithm. The work also offers a scheme for assessing the contribution of each author to the work. Work [5] analyses the co-authorship network in order to find interdisciplinary scientific communities, and work [6] examines the Topic Flow Network (TFN), which is built using information about each author and the abstract of the article.

This work aims to present a novel approach for constructing a network of connections between scientists by deliberately exploring available scientometric services, forming and researching a network of scientists, and considering the relationships between co-authorship and meaningful correlations of their research directions.

Network scanning means the selecting a small amount of the most important content from large networks that for technological reasons cannot be fully scanned [7]. In many modern studies of networks, the mechanisms of their scanning are used, after which conclusions about the topology of such networks are made. The work [8] shows that this approach is wrong. The reflections of initial networks obtained using various scanning algorithms can significantly differ and only partially reflect the properties of the initial networks. This is because the properties of these reflections significantly depend on the algorithms used for scanning.

The co-authorship network can become quite large if it is not restricted to a specific topic, such as the topic of the first author who is the starting point of the process of forming the network.

This effect significantly complicates the perception of the network and leads to "theme drift" effect. There is also the same spelling of the names and initials of various scientists. To overcome these effects, thematic filtering is applied, i.e. descriptors are used, attributed to the authors of the scientometric network, which determine their thematic focus. Adherence to these descriptors determines the size of the co- authorship networks formed and their growth dynamics. Identifying clusters in such networks can also serve as a basis for recognizing scientific schools, expert groups, and more.

It is advisable to use models tested on peering networks (peer- to-peer, P2P) when forming co-authorship networks. Peer-to-peer networks consist of nodes, each of which interacts with only a subset of other nodes, which is exactly the same as a co-authors network. When a node receives a request, its local index is searched. And, if the request is successful, the result is returned. Otherwise, the request is forwarded over the network. In our case (scanning the network of co-authors), it is advisable to forward the request over the network in all cases, if some restriction conditions are not satisfied. The network is scanned using the Breadth-first search (BFS) [9] method, where the request from a starting node is directed to all neighbors (the closest according to certain criteria), and scanning is limited only by the parameter of author citations.

Scientists with fewer citations than a designated threshold are excluded. Consequently, a complete scan of the nodes determined by this parameter and the descriptors is performed, and the resulting network is considered.

Let us consider the conditions of the problem formally, namely, let's assume, A is the set of authors, A_i is an author with an index *i*. P_i is a profile of the authorA_i. Let's denote the set of all existing descriptors as D. We are interested in the descriptors included in the author's profile. Simplistically, we will assume that a profile is a set of descriptorsP_i and $d_j \in D$ is a descriptor with an index *j*. Let's denote $\hat{d}_j^i$ as a sign of the presence of a descriptor with index *j* of an author with index *i:*

$$\hat{d}_j^i = \begin{cases} 1, d_j \in P_i, \\ 0, d_j \notin P_i. \end{cases} \quad (1)$$

The author with the index *i* is matched with a vector$\bar{A}_i = \left(\hat{d}_1^i, \hat{d}_2^i, \ldots, \hat{d}_{|D|}^i\right)$.

We will consider the scalar product of the corresponding vectors as the thematic proximity of the interests of the authors with indices *i* and *k*:

$$Sim(A_i, A_k) = (\bar{A}_i, \bar{A}_k). \quad (2)$$

Let's denote the relation of co-authorship between authors that have indices i and k as$Co(A_i, A_k) \in \{0,1\}$.

Accordingly, in these notations, the connection in the network of scientists between authors with indices i and k is equal

$$Link(A_i, A_k) = Sim(A_i, A_k) + C \cdot Co(A_i, A_k), \quad (3)$$

where$C$is constant, which is chosen by an expert.

The set of all possible$Co(A_i, A_k)$ values forms a co-authorship matrix. Thus, the matrix corresponding to the network of scientists is a combination of a network of thematic interests and a co-authorship network. As a result, the matrix of the network of scientists is denser.

2. Algorithm

The algorithm for scanning the scientometric network of the scientometric information service and the further formation of the network of scientists was adapted to the real network of co-authors of the service (Google Scholar is considered as such a service) as follows (**Помилка! Джерело посилання не знайдено.**):

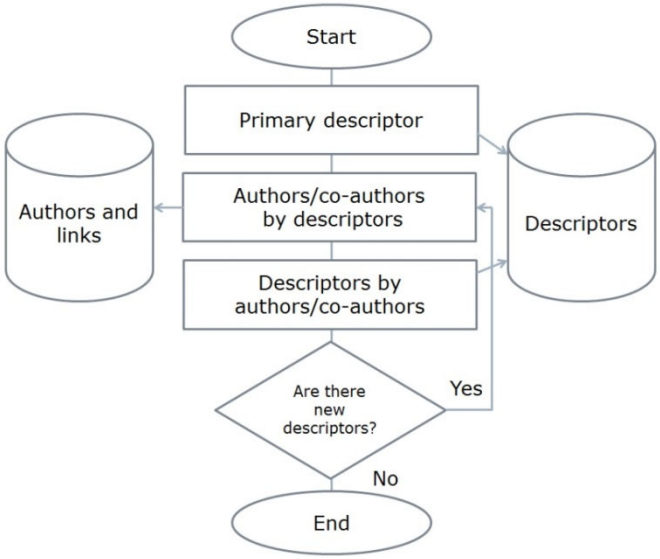

Figure 1: Advanced Google Scholar Citations service scanning algorithm

1. A descriptor is defined as the basic one for scanning (initially, one node is selected, in our case, it is obvious - "Cyber Security", Figure 2) is selected.

For the selected descriptor/descriptors, all scientists who have assigned themselves these descriptors (written in their profiles) are chosen using the scientometric service. As a result of this selection, the authors are placed in a sorted order - the authors with the most citations are shown at the beginning. To form a network using scanning, authors with a citation value equal to or greater than τ predetermined threshold value (for example, τ = 10,000) are considered.

2. The list of descriptors assigned to authors and defined at step 2 is considered. From among these, descriptors that correspond to the primary topic are selected. This process can be carried out by a specialist, expert, or automated method, such as by using specific keywords like "security", "access", "intrusion", '"deception", etc. (chosen by the knowledge engineer).In this particular case, the authors' pages for the first descriptor contain descriptors related to the primary topic, such as "CyberSecurity," "Cybersecurity," "Access Control Models Architectures," "Secure Cloud and IoT Computing," "Wireless Security," "Intrusion Detection," "Deception Detection," "Cloud Forensics Access Control," and more.

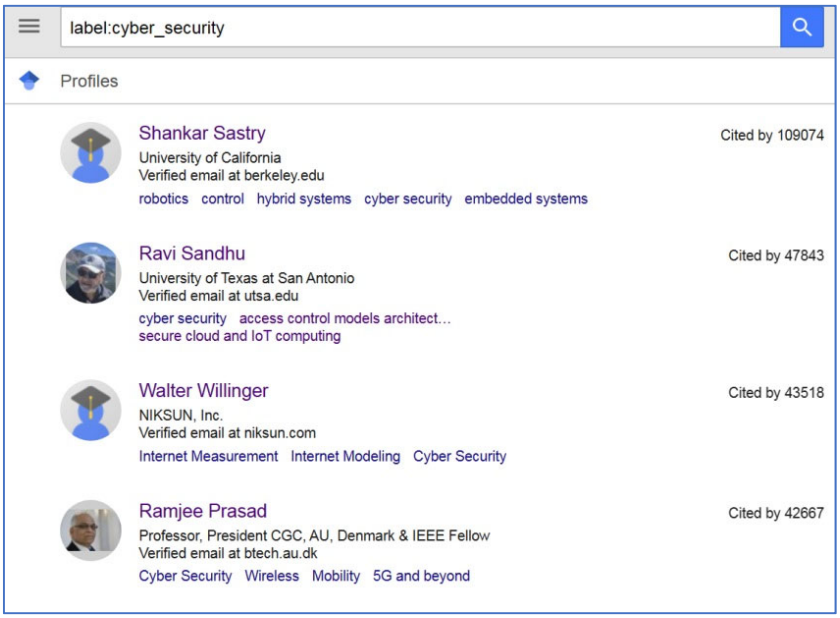

Figure 2: A fragment of the search results for the descriptor "Cyber Security"

3. For each of the authors selected at step 2, their co-authors with a citation value not less than the specified threshold are also considered. Among these co-authors, only those scientists whose descriptors are close to the primary topic of "Cyber Security" are considered as nodes of the network. These authors are also included as nodes in the future network of scientists. Descriptors that correspond to them are also taken into account, such as "Network Security," "Computer Security," "Data Breach Analysis," "Cybercrime Investigation," "IT Security," "Security and Privacy," among others.
4. For all selected descriptors, authors who have assigned themselves these descriptors are selected. If the list of authors with a citation value greater than 10,000 is exhausted for all selected descriptors, the process ends.

The given algorithm converges due to the limited number of scientists covered by the scientometric service. The weight of connections between nodes in the network is determined by the number of shared descriptors corresponding to the authors. Additionally, if there is a co-authorship relationship between the authors, a constant value is added to the weight of the corresponding connections.

Cluster analysis methods enable the identification of closely- related groups of scientists, co-authors, scientific schools, and expert groups. In this context, a scientific school refers to an informal team of researchers from different generations who are united by a shared program and research style, and are led by a recognized leader.

Figure 3a shows a fragment of the network of scientists in the cybersecurity field, which is formed according to the given algorithm with a citation threshold τ equal to τ=10000.As we can see, the network of scientists contains 1486 nodes and has one connectivity component, and explicit clusters, which were determined by modularity algorithm using the Gephi program environment [10, 11].

To calculate the characteristics of the network as a whole, parameters such as the number of nodes, edges, the average distance between nodes, diameter of the network (the largest geodetic distance), and the network density (the ratio of the number of edges to the maximum possible number of edges) are used. Determining cliques (subgroups or clusters in which nodes are more strongly interconnected than other members), selecting components (internal parts of the network not interconnected with other parts), and finding jumpers (nodes whose removal can lead to network collapse) are some of the topical problems in the study of complex networks.

The division of the network into groups is estimated by the clustering coefficient, which reflects the ratio of the number of connections between neighborsto the total possible number of such connections. The overall graph clustering coefficient is calculated as:

$$C = \frac{1}{N}\sum_{i=1}^{N}\frac{E_i}{k_i(k_i-1)},$$

where $N$is the number of nodes, k_i is the number of connections of the i-th node, E_i is the number of nodes adjacent to the i-th node, connected directly. The closer the value of the coefficient is to 1, the greater the probability of a cluster structure.

The modularity of elements and the graph as a whole is an essential characteristic of a graph. The modularity of a node is a value that evaluates the degree to which chains and clusters of components are connected, in proportion to the links of different components. In the cryptic vision of modulation, there can be distinctions like:

$$Q = \sum_{i=1}^{N} (e_{ii} - a_i)$$

where e_{ij} is an element of the adjacency matrix of the graph, equal to the ratio of the number of edges connecting two societies $i$and$j$, to the total number of edges in the network, $a_i = \sum_{j=1}^{N} e_{ij}$is the ratio of the number of edges connecting vertices in the society$i$ to the total number of edges. The high modularity of the network indicates a strong connection in the societies - clusters and a weak connection of the network itself.

The parameters of the formed network (Figure 3a) are as follows:

- number of nodes: 1486
- number of connections: 56937
- graph density: 0.052
- average node degree: 76.63
- graph diameter: 6
- average clustering coefficient: 0.62
- average path length: 2.55
- modularity: 0.484
- number of clusters according to the criterion of modularity with a distributive resolution of 1: 10.

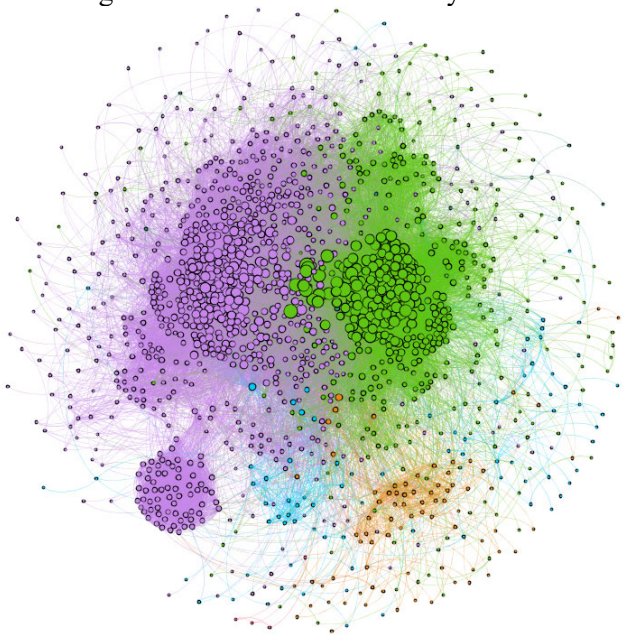

Figure 3a: Contours of the network of scientists

The network is divided into subnetworks using the modularity centrality algorithm. The average modularity of the network is 0.484, indicating active interaction among highly interconnected scientific groups. The algorithm identified 10 clusters in the network.

Figure 3b displays the central fragment of the network of scientists working in the field of cybersecurity, built using the specified algorithm and citation threshold.

One of the most important network parameters is the degree distribution of its nodes. In the case of a network of scientists, the node list ranked by degree is shown in Figure 4. The horizontal axis represents the rank of the network node, and the vertical axis shows the degree of the node. The high

degree of accuracy in approximating the values on the graph to a logarithmic curve (R^2=0.95) suggests an exponential distribution of node degrees.

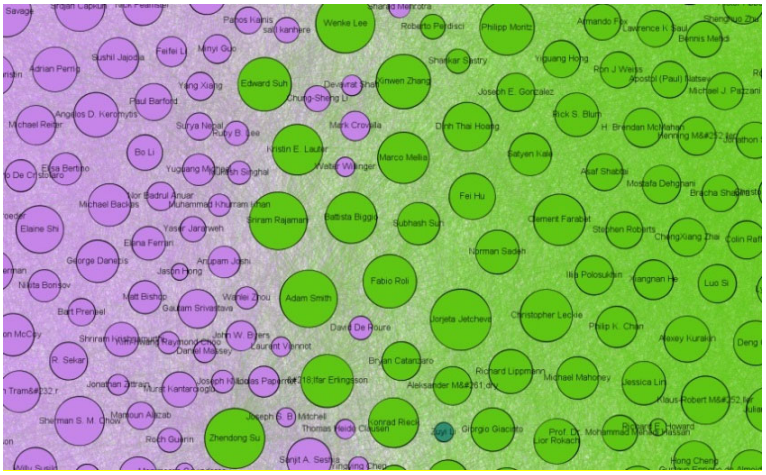

Figure 3b: A fragment of the network of scientists

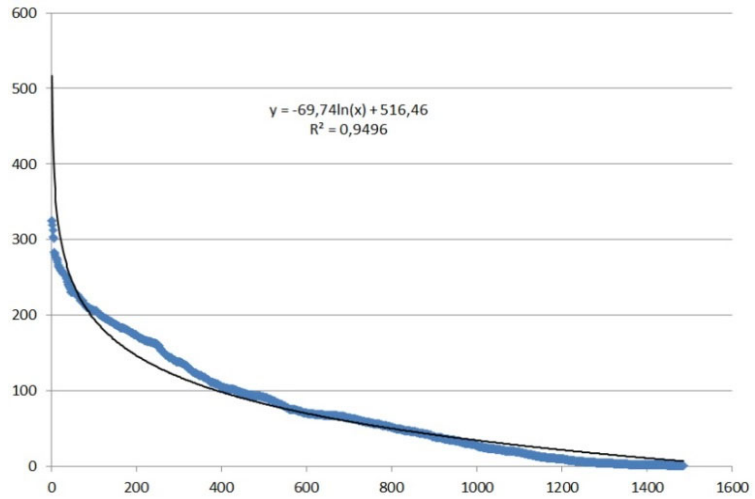

Figure 4: Rank distribution of node degree of the network of scientists

For comparison, Figure 5a shows the contours of the network of collaboration of scientists in the field of cybersecurity, built according to the part of the above algorithm with a citation threshold equal to τ=10000. The basis for building such a network is the co-authorship matrix - the set of all possible values $Co(A_i, A_k)$. We can see that the co-authorship network (those same 1486 nodes) has low connectivity and fuzzy clusters, which were also determined by modularity classes.

The parameters of the formed network are as follows:

- number of nodes: 1486
- number of connections: 2921
- graph density: 0.003
- average node degree: 3.93
- graph diameter: 19
- average clustering coefficient: 0.326
- average path length: 6.94
- modularity: 0.766
- number of clusters according to the criterion of modularity with a distribution resolution of 1: 40.

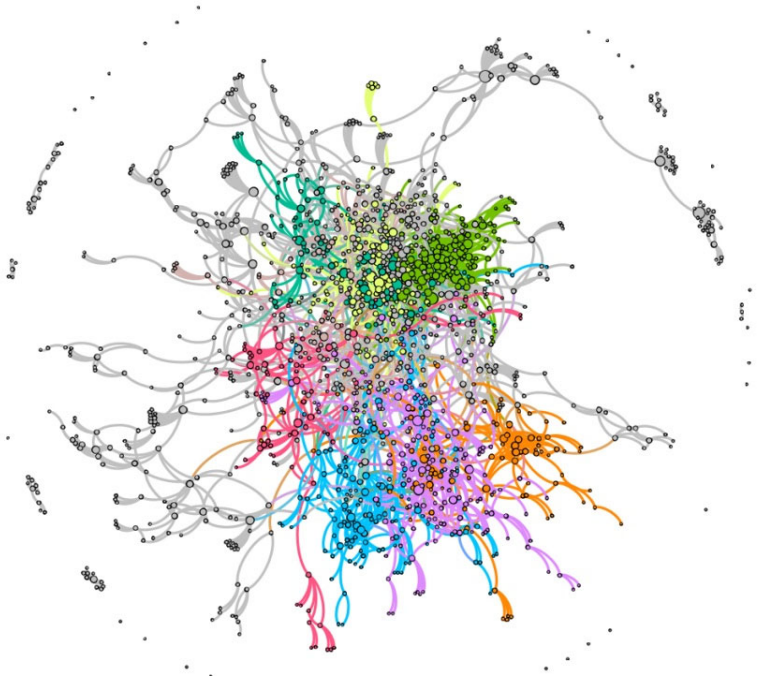

Figure 5a: Contours of the co-authorship network

The network is divided into loosely connected subnets using the modularity of each node (groups of nodes). The network's modularity is 0.766, indicating the active interaction of small scientific groups, relative to the size of the entire network. Thealgorithm identified 40 clusters in the network.

Table 1displays the top 20 cybersecurity scientists whose nodes have the highest degrees.

Table 1

List of the 20 most important nodes of the network of scientists

Person rank	Person	Node degree
1	Andreas Terzis	325
2	Jorjeta Jetcheva	325
3	Zhendong Su	319
4	Wenke Lee	313
5	Sriram Rajamani	304
6	Adam Smith	302
7	Philipp Moritz	284
8	Xinwen Zhang	282
9	T. V. Lakshman	280
10	Edward Suh	277
11	Fabio Roli	277
12	Dacheng Tao	276
13	Úlfar Erlingsson	273
14	Guo-Jun Qi	273
15	Christopher Leckie	270
16	Michael I. Jordan	267
17	Clement Farabet	264
18	Battista Biggio	263
19	Ghulam Muhammad	263
20	Marco Mellia	263

Figure 5b shows the central fragment of the cybersecurity academic co-authorship network.

Table 2 lists the 20 cybersecurity scientists whose nodes have the highest degrees in the co-authorship network of scientists.

For the co-authorship network, Figure 6 shows the node list ranked by degree. The high degree of accuracy in approximating the values on the graph to a logarithmic curve (R^2=0.99) suggests an exponential distribution of node degrees.

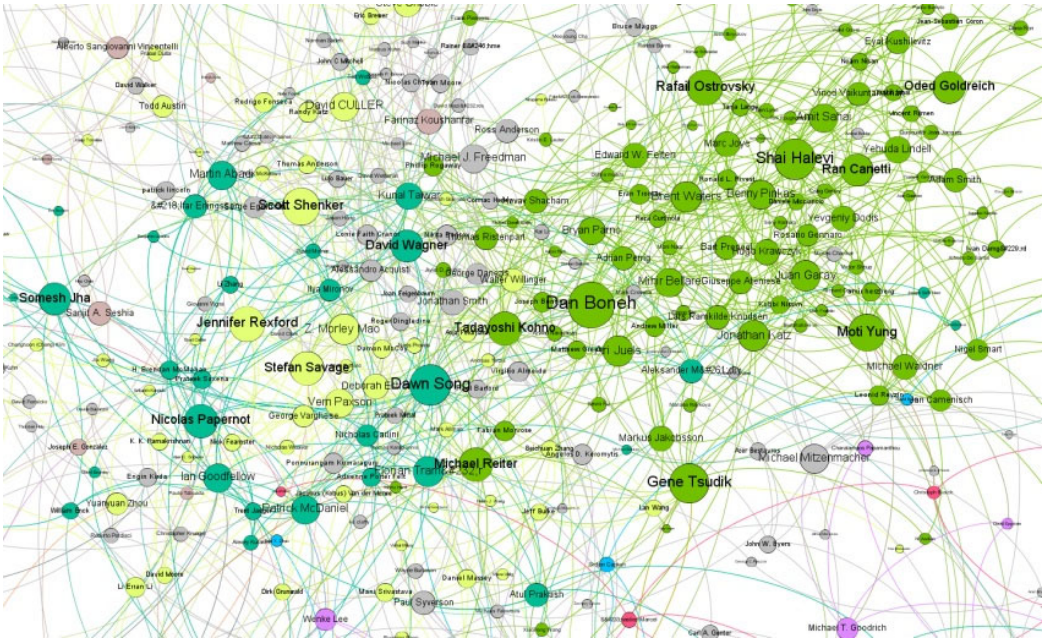

Figure 5b: A fragment of the co-authorship network

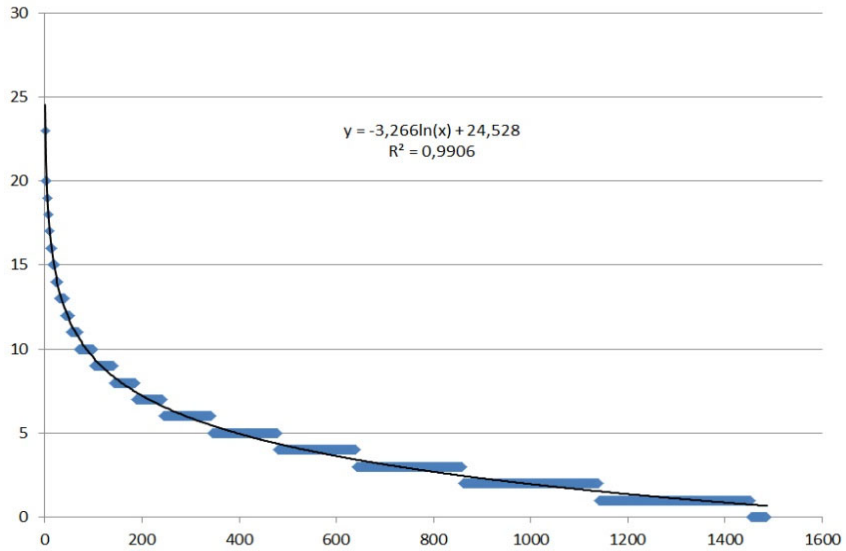

Figure 6: Rank distribution of node degree of the co-authorship network

Table 2
List of the 20 most important nodes of the co-authorship network

Person rank	Person	Node degree
1	Dan Boneh	23
2	Shai Halevi	20
3	Dawn Song	20
4	Federico Calzolari	20
5	Alessandro Gabrielli	19
6	Gene Tsudik	19
7	Scott Shenker	18
8	Moti Yung	18
9	Jennifer Rexford	17
10	Rafail Ostrovsky	17
11	Nicolas Papernot	16
12	Jiawei Han	16
13	Stefan Savage	16
14	Tadayoshi Kohno	16
15	Michael Reiter	16
16	Thomas S. Huang	15
17	David Wagner	15
18	Ran Canetti	15
19	Oded Goldreich	15
20	Somesh Jha	15

It is worth noting that the lists of scientists corresponding to the largest nodes in the two networks differ. Moreover, the indices of scientists corresponding to the largest nodes in the traditional co-authorship network, on average, exceed those parameters in the proposed network of scientists.

However, the proposed network has a number of important advantages for analysis:

- first of all, a clear clustering by topic, a limited number of clusters of scientists that clearly correspond to descriptors or in other words – topics;
- small graph diameter and average path length, which in practice can lead to the formation of expert groups of scientists who are not co-authors;
- and ultimately, considering not only the criterion of co-authorship, which increases the variability of solutions, allows for adjusting the relationship between clustering and thematic proximity.

3. Conclusions

We proposed and implemented an approach for forming a network of scientists within the subject area of cybersecurity. The algorithm for forming the network is limited by knowledge markers (descriptors) that are set in advance by scientists in their scientometric profiles.

It should be noted that there is a fundamental difference between the proposed model for the automatic formation of networks of scientists and existing models, which rely on direct participation of human experts in author selection. The proposed algorithm for forming the network of scientists uses both co-authorship relations and meaningful correlation of descriptors assigned to authors. Thus, the network scanning program uses the knowledge provided by the authors. This approach significantly expands the pool of experts.

In addition to the network under consideration, an adjacency network can also be considered. In this network the nodes are descriptors, and the connections are determined by the number of authors to whom corresponding pairs of descriptors are assigned. Such a network can be considered as a model of the domain defined by the primary descriptor.

The research results allow for scientific substantiation, automation, and acceleration of the procedure for selecting competent experts to solve various tasks in the field of cybersecurity.

The model was applied to the cybersecurity field within the Google Scholar service, but the proposed approach can be used for other scientific fields, or for other scientometric services.

4. References

[1] J. L. Ortega, How is an academic social site populated? A demographic study of Google Scholar Citations population, Scientometrics 104 (2015) 1-18.doi: 10.1007/s11192-015-1593-7.

[2] D.Lande, M.Fu, W.Guo, I.Balagura, I. Gorbov, H. Yang, Link prediction of scientific collaboration networks based on information retrieval, World Wide Web: Internet and Web Information Systems23(2020) 2239-2257.doi: 10.1007/s11280-019-00768-9.

[3] Google Scholar.URL: http://scholar.google.com/citations.

[4] J.Liu, Y.Li, Z.Ruan, G.Fu, X.Chen, R.Sadiq, Y. Deng, A new method to construct co-author networks,Phisica A419 (2015) 29-39.doi: 10.1016/j.physa.2014.10.006.

[5] M. Ullah, A. Shahid, I. Din, M. Roman, M. Assam, M.Fayaz, Y.Ghadi, H.Aljuaid, Analyzing Interdisciplinary Research Using Co-Authorship Networks, Complexity, 2022.2524491 (2022)13. doi: 10.1155/2022/2524491.

[6] B.Schäfermeier, J. Hirth, T. Hanika, Research topic flows in co-authorship networks,Scientometrics (2022) 1-28. doi: 10.1007/s11192-022-04529-w.

[7] Y.Chen, C.Ding, J.Hu, R.Chen, P.Hui, X.Fu,Building and analyzing a global co-authorship network using google scholar data, in: Proceedings of the 26th international conference on World Wide Web Companion (2017) 1219-1224. doi: 10.1145/3041021.3053056.

[8] D.Lande, O.Dmytrenko, Research of Topological Properties of Network Reflections Obtained Using Different Algorithms for Scanning Initial Networks,in: Shkarlet S. et al. (eds) Mathematical Modeling and Simulation of Systems. MODS 2021, volume 344 of Lecture Notes in Networks and Systems, Springer, Cham, 2022. doi: 10.1007/978-3-030-89902-8_26

[9] L.Paulino, C.Hannum, A.S.Varde, C.J.Conti, Search Methods in Motion Planning for Mobile Robots, in: Arai, K. (eds) Intelligent Systems and Applications. IntelliSys 2021, volume 269 of Lecture Notes in Networks and Systems, Springer, Cham, 2022.doi: 10.1007/978-3-030-82199-9_54.

[10] K.Cherven, Mastering Gephi Network Visualization, Packt Publishing, 2015.

[11] Gephi. URL: https://gephi.org/.