An Analysis of the Social and Conceptual Networks of CECIIS 2005 - 2010

Markus Schatten, Juraj Rasonja, Petar Halusek, Frane Jakelić
University of Zagreb
Faculty of Organization and Informatics
Pavlinska 2, 42000 Varaždin, Croatia
{markus.schatten, juraj.rasonja, petar.halusek, frane.jakelic}@foi.hr

Abstract. *Two network analyses and one keyword cloud analysis of the CECIIS 2005 - 2010 Proceedings bibliography are presented and discussed. First the social network of CECIIS authors is visualized using the k-core decomposition algorithm and properly analyzed. Afterwards a keyword cloud based conceptual analysis of keywords grouped by conference year is provided and most important topics are identified. In the end four conceptual networks of keywords are presented (keywords connected through authors, sessions, chairmans and papers) an the differences are discussed.*

Keywords. CECIIS; IIS; bibliography; social network analysis; conceptual network analysis

1 Introduction

Network theory or the new science of networks as Barabasi calls it [11] examines the properties of social, biological, transport, technological, physical, semantic and other types of networks. The study of social networks has a long tradition, but with the development of the Internet and contemporary information and communication technology it gained a great boost.

A network can be defined as a mathematical abstraction which consists of two parts: (1) nodes (which can represent people, organizations, countries, but also computers, life forms, molecules and concepts), and (2) links (which can represent any perceivable connection between nodes, e.g. friendship between people, trade between organizations, geographic neighborhood between countries, a wireless connection between computers, food chains in some ecosystem, linkages between molecules or semantic relationships between concepts in some language). If links are directed (e.g. a communication message, spreading of some contagious virus, social power etc.) than we say a network to be directed.

Formally, networks are characterized using graphs which are defined as the pair $G = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N} = \{n_1, n_2, \ldots, n_k\}$ is the set of nodes or vertices, and $\mathcal{E} = \{(n_i, n_j) | n_i, n_j \in \mathcal{N}\}$ is the set of edges or arcs. If the pairs of nodes in $\mathcal{E}$ are ordered, the graph is a directed graph or digraph. If the intensity of each edge is measurable, the graph is valued meaning that edges are annotated with their corresponding value.

In the following we will use construct a simple, undirected social network of authors (e.g. a scientific collaboration network [13, 12, 14]), in which two authors are connected if they have co-authored a paper. The network is, however, valued, since authors can co-author more than one paper. Thus, the value on each edge will be the number of papers the corresponding authors have written together.

Networks are often represented as the so called adjacency matrix $\mathbb{A} = [a_{ij}], a_{ij} \in \{0, 1\}$ for sake of simplicity. The matrix $\mathbb{A}$ is of size $k \times k$ where k is the number of nodes in a network. The elements of $\mathbb{A}$ are equal to $a_{ij} = 1$ if there exists and edge between the corresponding nodes (n_i and n_j). Otherwise $a_{ij} = 0$. If the network is undirected the matrix is symmetric. If the network is valued, the values of the edges are the elements of $\mathbb{A}$ instead of 1's.

The notion of bipartite, tripartite and n-partite graphs is of special importance to our study. A bipartite graph $G = (\mathcal{N}, \mathcal{E})$ is a graph for which set of nodes there exists a partition $\mathcal{N} = \{X, Y\}$, such that every edge has one node in X, and the other in Y, or more specifically the set of graph nodes is decomposed into two disjoint sets such that no two graph nodes within the same set are adjacent. For tripartite graphs there exists a partition into three disjoint subsets. The general case is that of n-partite graphs for which there exists a partition of n disjoint subsets of $\mathcal{N}$ with the stated properties.

A bipartite graph can for example be the network of authors (A) and papers (P) in which nodes are authors and articles (there exists a partition into two disjoint subsets A and P), and edges are the essential connections between authors and papers they have written. As we can see, there will never be a connection between two authors (e.g. authors do not write other authors) nor a connection between two two papers (e.g. papers aren't written by other papers). An example of a tripartite graph can be the networks of authors (A), papers (P) and keywords (K) which are used on a particular article. In the following we will analyze the 5-partite graph of authors (A), papers (P), keywords (K), sessions the papers were presented of (S) and chairmans of the particular sessions (C).

Here we must state that every n-partite graph can be represented using n $(n-1)$-partite graphs. For example the 4-partite graph $APKS$ can be represented using the following four tripartite graphs: APK, APS, AKS, PKS. Further, every tripartite graph can be represented through 3 bipartite graphs. For example the graph APK can without information loss be represented as AP, AK, PK. Graphs of lower partitioning are constructed by dropping all nodes of the excluded partition set as well as all edges these nodes participate in.

The representation through bipartite graphs is much less cumbersome than with tripartite and 4-partite graphs since the adjacency matrix of bipartite graphs can be rewritten as a $|X| \times |Y|$ matrix where X and Y are the partition sets, and $|W|$ is the cardinal number of set W. For example the graph AK can be represented as a matrix $\mathbb{AK} = [ak_{ij}]$ in which we put authors (the elements of A) as rows, and keywords (the elements of K) as columns. The values in this matrix ak_{ij} will equal 1 iff author i has used keyword j on some article.

The graph folding procedure is the mapping of one graph into another that always maps nodes from one graph into nodes of the other, and edges of one graph into edges of the other [8]. For the purpose of this study we will introduce one such graph folding procedure which allows us to obtain unipartite from bipartite graphs by using matrix multiplication with a transposed matrix. Let $\mathbb{AK}$ be the (bipartite) matrix of authors and keywords, then the folding of this graph by using the operation $\mathbb{AK} \cdot \mathbb{AK}^T$ allows us to construct a matrix that represent a social network of authors in which two authors are connected iff they have used the same keywords on any of their articles. The dual matrix $\mathbb{AK}^T \cdot \mathbb{AK}$ is a conceptual network of keywords in which two keywords are connected iff they have been used by the same author.

A similar procedure for tripartite networks has been used by Mika for his Actor-Concept-Instance folksonomy model [10]. He argued that a tripartite ontology model applies for social tagging systems like Delicious.[1] We will use this procedure to construct four conceptual networks:

- $\mathbb{AK}^T \cdot \mathbb{AK}$ - conceptual network of keywords connected through authors (two concepts are connected if they have been used by the same author);
- $\mathbb{SK}^T \cdot \mathbb{SK}$ - conceptual network of keywords connected through sessions (two concepts are connected if they have been used in the same session).
- $\mathbb{CK}^T \cdot \mathbb{CK}$ - conceptual network of keywords connected through chairmans (two concepts are connected if they have been hosted by the same chairman).
- $\mathbb{PK}^T \cdot \mathbb{PK}$ - conceptual network of keywords connected through chairmans (two concepts are connected if they have been hosted by the same chairman).

Since we are dealing with complex social and conceptual networks, we need an adequate visualization algorithm. Herein we will use the k-core decomposition algorithm described in [1]. The definition of this algorithm goes beyond this study which is why we will satisfy our selves with the simplified description that this algorithm attempts to find cores which represent mutually well connected nodes. These cores are arranged in a circular fashion such that the inner cores are comprised of nodes with higher degree (greater number of edges).

The rest of the paper is organized as follows: in section 2 we describe how the data was collected and how we implemented our analysis; in section 3 we give provide and analyze the social network of CECIIS authors; in section 4 we analyze particular keywords grouped by conference year by using keyword clouds; in section 5 the four conceptual networks of CECIIS (author, session, chairman and paper) are visualized and analyzed; in the end in section 6 we draw our conclusions.

2 Data Gathering and Implementation

In order to gather the data about the CECIIS bibliography and construct our 5-partite network we had to use various sources. Bibliographic data about IIS (years 2005, 2006 and 2007) and CECIIS (2008, 2009, and 2010) has only partially been available online. In particular the conference programme (including data about session names, chairmans, authors, and titles) was available for all considered years. Since 2008 additional metadata for each article has been available on a digital archive.[2] Unfortunately the year 2008 archive didn't include keywords.

Thus, in order to automate data collection we implemented a spider program using Python[3] and specifically the Scrapy[4] module, which considerably eased implementation. Since we had to deal with semistructured data, we used the XPath language to extract the relevant bibliography data including:

- Chairman/Chairwoman names and surnames;
- Session titles;
- Years of publication;
- Author names and surnames;

[1] http://www.delicious.com

[2] http://search.foi.hr

[3] http://www.python.org

[4] http://scrapy.org

- Paper titles;

- Keywords.

Data that couldn't be gathered automatically was collected manually from the appropriate proceedings [2, 3, 4, 5, 6, 7]. In this way data about 1380 distinct keywords, used on 434 distinct papers, written by 534 distinct authors, presented on 104 distinct sessions, held by 22 distinct chairman/chairwoman was collected.

All data was stored in a PostgreSQL[5] database and analyzed using the Python module NetworkX.[6] In order to visualize the constructed networks the tool LaNet-vi[7] was used. Keyword data was visualized into keyword clouds using Wordle.[8]

3 Social Network Visualization

The first constructed network from the collected data is the social network of co-authors depicted on figure 1.[9] The authors' names have been removed due to privacy concerns. The size of each node depends on the nodes degree, while the color depicts the number of edges to nodes inside the same core.

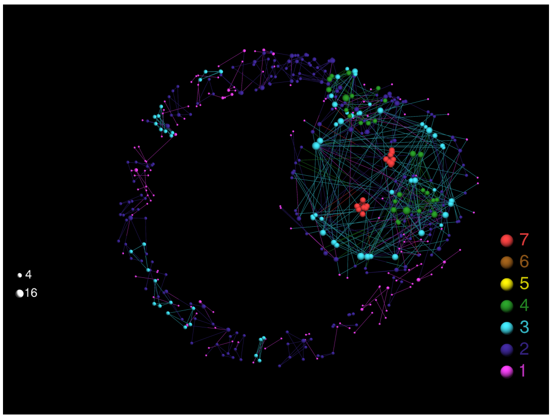

Figure 1: Social network of co-authors

As one can see from the figure the network is composed of two main cores, and a number of smaller cores which form cliques. The greatest (outer) core is comprised of authors who attended the conference sporadically (one or two times), while the second main (inner) core together with the cliques constitutes the most productive authors of the conference. There are also satellite nodes around these most productive authors and are their collaborators which (if publish further on the conference) will most likely join the core.

4 Keyword analysis

In the following we will analyze the most important conference keywords year by year. Each keyword-cloud has been constructed depending on keyword frequency on all conference papers for a given year. The greater the frequency of a keyword the greater the font. Figure 2 depicts the generated keyword-cloud for the year 2005.

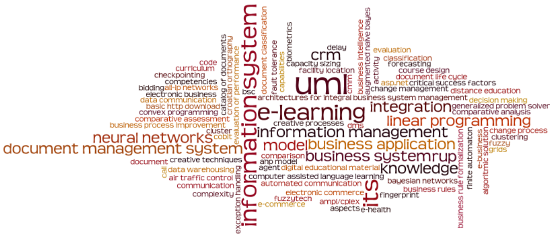

Figure 2: Keyword-cloud IIS 2005

As one can see from the figure, most important concepts as perceived by the authors in the year 2005 were related to UML, e-learning, information systems, ITS, neural networks, document management systems, CRM, integration, linear programming, information management, models, business applications, business systems, RUP as well as knowledge.

Figure 3 shows the keyword-cloud for the year 2006.

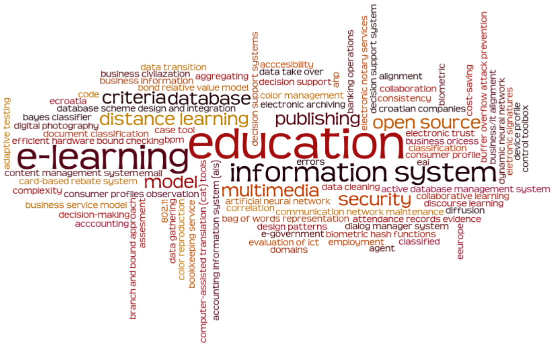

Figure 3: Keyword-cloud IIS 2006

As one can see from the image most important concepts in the 2006 conference were related to education, e-learning, information systems, criteriae, databases, distance learning, publishing, open source, models, multimedia as well as security.

On figure 4 the keyword cloud for the 2007 conference is shown.

As one can see from the figure, most important concepts that year were related to e-learning, data mining, security, UML, public services, information and communication technology, software development, multimedia, Java, graphic reproduction, SOA as well as ICT.

Figure 5 depicts the keyword-cloud for the 2008 conference.

In 2008 the most important concepts as perceived by the authors were e-learning, ICT, information

[5] `http://www.postgresql.org`

[6] `http://networkx.lanl.gov`

[7] `http://xavier.informatics.indiana.edu/lanet-vi`

[8] `http://www.wordle.net`

[9] Due to medium constraints these visualizations are of lower resolution. Please refer to `http://arka.foi.hr/~mschatten/ceciis2011` for high-resolution images.

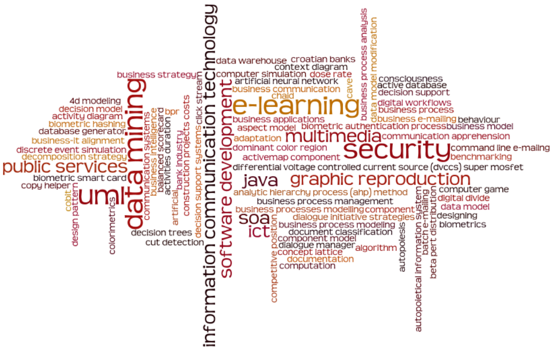

Figure 4: Keyword-cloud IIS 2007

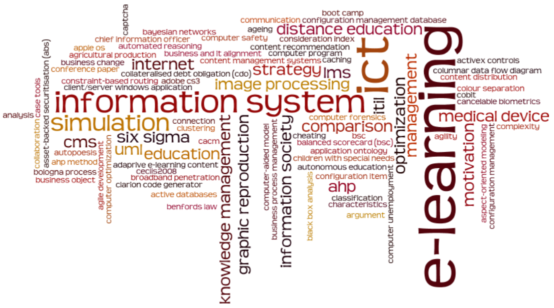

Figure 5: Keyword-cloud CECIIS 2008

systems, simulation, distance education, Internet, strategy, LMS, image processing, CMS, Six Sigma, UML, education, knowledge management, graphic reproduction, information society, ITIL, comparison, optimization, management, medical device and motivation.

The keyword cloud for the 2009 conference is shown on figure 6.

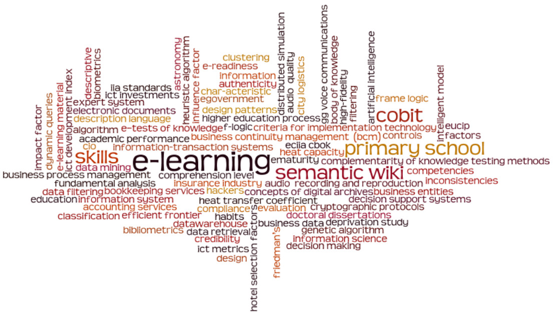

Figure 6: Keyword-cloud CECIIS 2009

This time most important concepts are related to e-learning, COBIT, primary school, semantic wiki and skills.

Figure 7 shows the keyword cloud for the 2010 conference.

In 2010 most important concepts as perceived by the authors were related to information systems, ICT, computer forensics, Internet, digital competence, e-learning, Coq, education, information, programming, open source as well as Croatia.

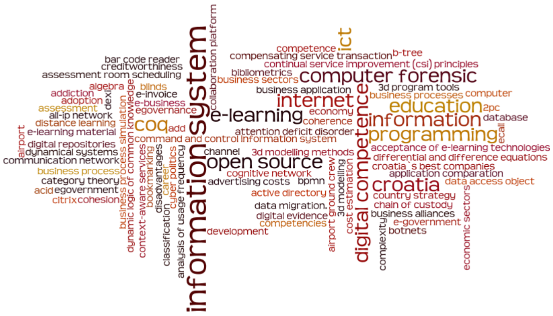

Figure 7: Keyword-cloud CECIIS 2010

5 Conceptual Network Analysis

In the following we present four visualizations of conceptual networks. On each visualization the node size reflects the nodes degree (the number of edges adjecent to the node) whereby the scale is given on the left side of the image. The nodes' color (scale given on the right side of the image) is of particular importance. Two nodes will have the same color if they have the same number of connections to all other nodes in the same core.

The first (depicted on figure 8) is the conceptual network where two keywords are connected if they are used by the same author. As one can see, the most inner core (red color) which is comprised of most mutually well-connected nodes includes the concepts: characteristic, systematization, evaluation method, agent, ontology, UML, parameter and biometrics. This network highlights keywords of the most important authors in the network (authors with highest degree).

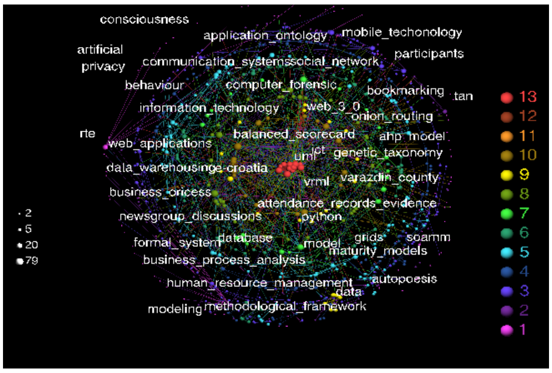

Figure 8: $\mathbb{AK}^T \cdot \mathbb{AK}$ keyword matrix (authors' graph)

The following (figure 9) shows the conceptual network in which two keywords are connected if they have been used on the same session. In this case the inner core of the network includes the keywords: bibliometrics, higher education, distance education, student, digital competence, motivation, multimedia, education, e-learning, and ICT. This network highlights keywords from the most prominent sessions.

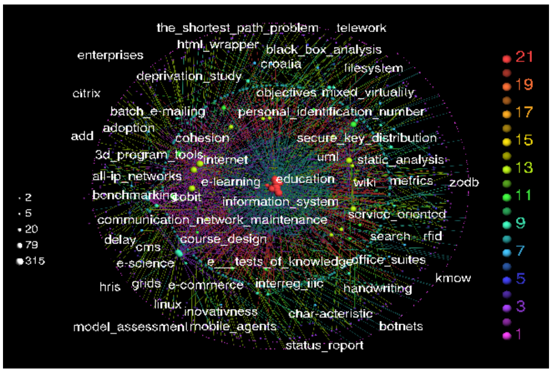

Figure 9: $\mathbb{SK}^T \cdot \mathbb{SK}$ keyword matrix (sessions' graph)

Figure 10 shows the conceptual network in which two keywords are connected if they were hosted by the same chairman. The inner core of this network contains the keywords: motivation, teaching approaches, information science, programming, ICT, e-learning, information society, distance education, methodology, UML, model, information system, Coq, TaOPis, datawarehouse, frame logic, education, semantic wiki. This network highlights keywords from the most important chairmans.

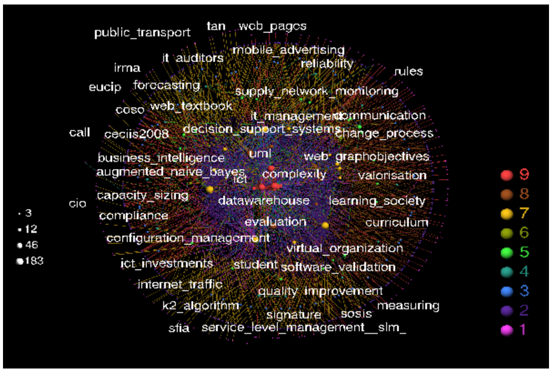

Figure 10: $\mathbb{CK}^T \cdot \mathbb{CK}$ keyword matrix (chairmans' graph)

Figure 11 depicts the conceptual network in which two keywords are connected if they were used on the same article. In this network no particular cores could be identified, but a number of cliques dealing with software (keywords: model, code, software, ICT), IT management (keywords: COBIT, maturity model, IT management, IT governance), information systems (keywords: management, BSC, business system, business application, information system, e-learning (keywords: education, multimedia, e-learning, distance learning, LMS, distance education, ICT), and multi-agent systems (ICT, multiagent system, agent). A number of dyads are also identifiable. The keyword ICT is a hub that connects the software, e-learning and multiagent system cliques.

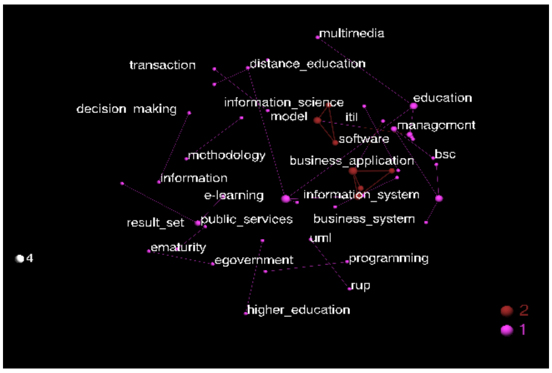

Figure 11: $\mathbb{PK}^T \cdot \mathbb{PK}$ keyword matrix (papers' graph)

6 Conclusion

In this paper a we presented a social and conceptual network analysis of the CECIIS proceedings in the last 5 years. The (short) social network analysis showed that the CECIIS conference has a rather small core of steady participants, while the majority of authors published only once or twice on it. This should be a concern to the future conference organizers: they should take actions which will motivate prospective authors to come back more often.

The keyword cloud analysis showed the most important keywords of the conference year by year. A constant through all analyzed years is the keyword e-learning, which seems to have been the most important topic for the last 5 years authors.

The analysis of the conceptual networks showed particularly important keywords depending on various aspects in which concepts were connected (authors, sessions, chairmans, and papers). Since these networks were very different, we can conclude an important thing about these (and probably similar) networks: context of association is crucial for conceptual network analysis.

On the other hand if we compare these networks to the results of Mika [10], another important difference appears. The conceptual networks constructed by Mika (using the network data of Delicious) made good sense: connected concepts were connected meaningfully. In our case, this wasn't the case, except in few cases for the paper based graph. This makes us ask the question, what is the difference between Delicious as a social system, and the CECIIS conference? There are two big differences: (1) the number of nodes in the network (Delicious has a huge number of actors, concepts and individuals in contrast to CECIIS), (2) the type of actors (CECIIS authors are scientists, while Delicious has a more variegated user base).

As Luhman states it, social systems are meaning processing systems [9, 15]. Our results make us wonder, when does a social system start to generate meaning? If the number of actors is the significant difference, how big should a conceptual network be in order to be meaningful. On the other hand, is it possible that educated actors are the inferior predictors? This and similar questions are subject to our future research.

References

[1] Alvarez-Hamelin, J. I., Dall'Asta, L., Barrat, A., and Vespignani, A. *Large scale networks fingerprinting and visualization using the k-core decomposition*, vol. 18 of *Advances in Neural Information Processing Systems.* MIT Press, Cambridge, MA, 2006, pp. 41–50.

[2] Aurer, B., and Bača, M., Eds. *Conference proceedings / 16th International Conference on Information and Intelligent Systems.* Faculty of Organization and Informatics, Varaždin, Croatia, 2005.

[3] Aurer, B., and Bača, M., Eds. *Conference proceedings / 17th International Conference on Information and Intelligent Systems.* Faculty of Organization and Informatics, Varaždin, Croatia, 2006.

[4] Aurer, B., and Bača, M., Eds. *Conference proceedings / 18th International Conference on Information and Intelligent Systems.* Faculty of Organization and Informatics, Varaždin, Croatia, 2007.

[5] Aurer, B., and Bača, M., Eds. *Conference proceedings / 19th Central European Conference on Information and Intelligent Systems.* Faculty of Organization and Informatics, Varaždin, Croatia, 2008.

[6] Aurer, B., Bača, M., and Rabuzin, K., Eds. *Conference proceedings / 20th Central European Conference on Information and Intelligent Systems.* Faculty of Organization and Informatics, Varaždin, Croatia, 2009.

[7] Aurer, B., Bača, M., and Schatten, M., Eds. *Conference proceedings / 21st Central European Conference on Information and Intelligent Systems.* Faculty of Organization and Informatics, Varaždin, Croatia, 2010.

[8] El-Kholy, E., and El-Esawy, E. Graph folding of some special graphs. *Journal of Mathematics and Statistics 1*, 1 (2005), 66–70.

[9] Luhmann, N. *Soziale Systeme: Grundriß einer allgemeinen Theorie.* Suhrkamp, Frankfurt, Germany, 1984.

[10] Mika, P. Ontologies are us: A unified model of social networks and semantics. *Web Semantics: Science, Services and Agents on the World Wide Web 5*, 1 (Mar. 2007), 5–15.

[11] Newman, M., Barabasi, A.-L., and Watts, D. J., Eds. *The Structure and Dynamics of Networks.* Princeton Studies in Complexity. Princeton University Press, Princeton and Oxford, 2006.

[12] Newman, M. E. J. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Phys. Rev. E 64*, 016132 (Jun 2001), 1–7.

[13] Newman, M. E. J. Scientic collaboration networks. i. network construction and fundamental results. *Phys. Rev. E 64*, 016131 (2001), 1–8.

[14] Newman, M. E. J. Coauthorship networks and patterns of scientific collaboration. In *Proceedings of the National Academy of Sciences* (2004), pp. 5200–5205.

[15] Schatten, M., and Bača, M. A critical review of autopoietic theory and its applications to living, social, organizational and information systems. *Društvena Istraživanja 108-109*, 4-5 (2010), 837–852.