Analyzing the network structure and gender differences among the members of the Networked Knowledge Organization Systems (NKOS) community

Fariba Karimi, Philipp Mayr and Fakhri Momeni

GESIS – Leibniz Institute for the Social Sciences,
Unter Sachsenhausen 6-8
50667 Cologne, Germany
firstname.lastname@gesis.org

Abstract. In this paper we analyze a major part of the research output of the Networked Knowledge Organization Systems (NKOS) community in the period 2000 to 2016 from a network analytical perspective. We focus on the papers presented at the European and U.S. NKOS workshops and in addition four special issues on NKOS in the last 16 years. For this purpose we have generated an open dataset, the "NKOS bibliography" which covers the bibliographic information of the research output. To analyze the co-authorship network of this community, we restrict our analysis to papers which have been authored by a minimum of two authors. This results in 123 papers with a sum of 256 distinct authors. We use standard network analytic measures such as degree, betweenness and closeness centrality to describe the co-authorship network of NKOS dataset. First, we investigate global properties of the network over time. Second, we analyze the centrality of the authors in the NKOS network. Lastly, we investigate gender differences in collaboration behavior in this community. The NKOS network is a typical co-authorship network with one large connected component, some smaller components and many isolated co-authorships or triples. The complete NKOS network consists of 97 (38%) women and 157 (62%) number of men and 2 unidentified names. The largest component represents 107 authors (41% of all authors). In the largest connected component we find 46 women and 59 men.

Keywords: NKOS workshops, Output analysis, Network analysis, Co-Authorship analysis, Central authors, Collaboration

1 Introduction

The Networked Knowledge Organization Systems (NKOS)[1] community in Europe and in the United States of America has held a long-running series of annual workshops at the European Conference on Digital Libraries (ECDL),

[1] For an introduction of KOS and NKOS and recent applications see [8, 14].

latterly renamed as the International Conference on Theory and Practice of Digital Libraries (TPDL), the Joint Conference on Digital Libraries (JCDL) and some other scattered events. The NKOS workshops in the US have started in 1997/1998 organized by Linda Hill, Gail Hodge, Ron Davies and others. Slightly later, the first NKOS workshop was organized in Europe at ECDL 2000 in Lisbon (Portugal) by Martin Doerr, Traugott Koch, Douglas Tudhope and Repke de Vries.

Typically, recent advances in Knowledge Organization Systems (KOS) have been reported at the annual NKOS workshops, e.g. including the Simple Knowledge Organization System (SKOS) W3C standard, the ISO 25964 thesauri standard, the CIDOC Conceptual Reference Model (CRM), Linked Data applications, KOS-based recommender systems, KOS mapping techniques, KOS registries and metadata, social tagging, user-centered issues, and many other topics[2]. Special issues on Networked Knowledge Organization Systems have been published in Journal of Digital Information in 2001 [8] and 2004 [24], in New Review of Hypermedia and Multimedia in 2006 [25] and recently in the International Journal of Digital Libraries in 2016 [14]. Recently, the NKOS workshop activities have accelerated again e.g. with two European NKOS in 2016 at the TPDL and Dublin Core conference and a revival of the US NKOS activities in 2017. In addition, the last two NKOS workshops at TPDL have resulted in formal conference proceedings published as CEUR Workshop Proceedings [15, 16].

The motivation of this paper is to analyze and visualize the past research output and collaborations of the NKOS community. We are focusing here on the informal part of this output, the paper presentations given at the past NKOS workshops. The specialty of this research output is that these research papers typically are not published in journals or conference proceedings. These papers appear just as oral presentations at the workshop and are documented on the corresponding websites. To cover this informal research output, we have collected presentation information from the workshop agendas. It is important to note how these workshop agendas are established. The practices at the NKOS workshops in the United States and Europe are different. In the United States, NKOS workshops were previously not based on an open call for papers contribution type, but rather via inviting speakers. This practice explains the relatively low ratio of co-authorship in the U.S. workshop series. From the beginning, in Europe, the NKOS workshops were based on accepting academic papers and resulted in open call for papers and subsequent peer review of submitted paper abstracts.

In the following, we report about network structures and gender differences among the members of the NKOS community as we could recall from the past European and U.S. workshop agendas and published special issues.

This paper is a largely extended version which based on the paper ”Analyzing the research output presented at European Networked Knowledge Organization

[2] Comprehensive review articles on KOS and NKOS topics have been published in [26, 9].

Systems workshops (2000-2015)" [18] presented at the 15th NKOS workshop at TPDL 2016. In [18], we focused on the European workshops and special issues. Meanwhile, we have extended the dataset and included the U.S. NKOS workshops and some other scattered NKOS events. So, this paper is able to give a more comprehensive overview of the international NKOS research community.

To the best of our knowledge, this paper is the first attempt to analyze the co-authorship network of NKOS in great details.

In the following sections we describe the underlying dataset (section 2), we perform network analysis (section 3), highlight some results of our analysis (section 4) and conclude our paper (section 5).

2 NKOS workshop bibliography dataset

For our analysis, we have compiled an open dataset derived from the "NKOS bibliography"[3]. The NKOS bibliography has been started in 2016 [18] and covers bibliographic information of all research papers presented at past NKOS workshops. Editing, organizing activities (incl. the introductions) at the workshops have not been covered in our dataset. Journal papers published in four special issues on NKOS [8, 24, 25, 14] which have been edited by members of the NKOS community in the same period have been added. These journal papers are the only formal publications in our analysis. In the end, we manually disambiguate author names of all papers. The bibliography is stored in single bibtex files (one bibtex file for each venue).

To this date, the NKOS bibliography covers:

- sixteen European NKOS workshops from 2000 to 2016. In total 16 workshop agendas: ECDL 2000, 2003-2010, TPDL 2011-2016, Dublin Core 2016,
- eight US NKOS workshop agendas: JCDL 2000-2003, 2005 and NKOS-CENDI 2008-2009, 2012,
- four special issues on NKOS and
- two scattered NKOS workshops at ISKO-UK 2011 and ICADL 2015.

For the analysis in this paper we have compiled all research presentations at NKOS workshops and papers published in special issues. For the following we restrict our analysis to papers which have been authored by a minimum of two authors. This restriction reduces the content of the dataset, e.g. the ECDL NKOS workshop from 2000 is missing in Table 1 because all papers were single author papers.

In total, this results in a dataset of 123 papers with a sum of 256 distinct authors (see Table 1)[4].

[3] The NKOS workshop bibliography is maintained in the following github repository: https://github.com/PhilippMayr/NKOS-bibliography.

[4] The data for this subset is available under https://github.com/PhilippMayr/NKOS-bibliography/tree/master/publications/ijdl17

year	nr. papers	nr. authors	nr. links	avg. clustering
2001	4	9	6	0.37
2002	3	10	13	0.8
2003	5	12	9	0.4
2004	13	39	47	0.65
2005	7	22	26	0.81
2006	11	33	39	0.73
2007	4	15	24	1.0
2008	7	15	9	0.2
2009	10	34	60	0.68
2010	8	21	19	0.61
2011	8	32	59	0.80
2012	6	26	56	0.92
2013	5	18	31	0.86
2014	6	16	13	0.85
2015	9	24	23	0.58
2016	17	60	114	0.75

Table 1: Overview of all NKOS papers sorted by years. In general, community shows high average clustering in many years indicating that there are many triangles in the network.

3 Network analysis of the NKOS community

In order to analyze the collaboration of the NKOS community we build a network of all authors at the workshops and special issues and compute various centrality measures for each author. A link in this network represents two authors who wrote a paper together. Therefore, if we have n_p number of papers and a paper i has m_i authors, the total number of pairs (links) E are

$$E = \sum_{i=1}^{n_p} \frac{m_i(m_i - 1)}{2} \qquad\qquad if \quad m_i \geq 1 \tag{1}$$

If two authors have published more than one paper together, we give weights to the link equivalent to the number of times they have collaborated in different papers. Thus, the resulting network is a weighted undirected graph.

In this paper, first, we investigate global properties of the network over time. Second, we analyze the centrality of the authors in this network. Lastly, we investigate gender differences in collaboration behavior in this community.

4 Results

Figure 1 demonstrates the overall NKOS co-authorship network. In this view each author has at least one co-author. The node color represents gender; purple for men and orange for women. This network contains 44 components. From the network illustrated in this figure we selected the largest component that is represented in Figure 3. 107 authors (41% of all authors) are connected in this component. The NKOS co-authorship network in the "NKOS bibliography" is a typical co-authorship network with one relatively large component, some smaller components and many isolated co-authorships or triples.

Figure 2 shows the degree distribution for this network. Despite being a rather small network, the degree distribution follows a similar trend as a power-law degree distribution that has been observed in other co-authorship networks [1, 11].

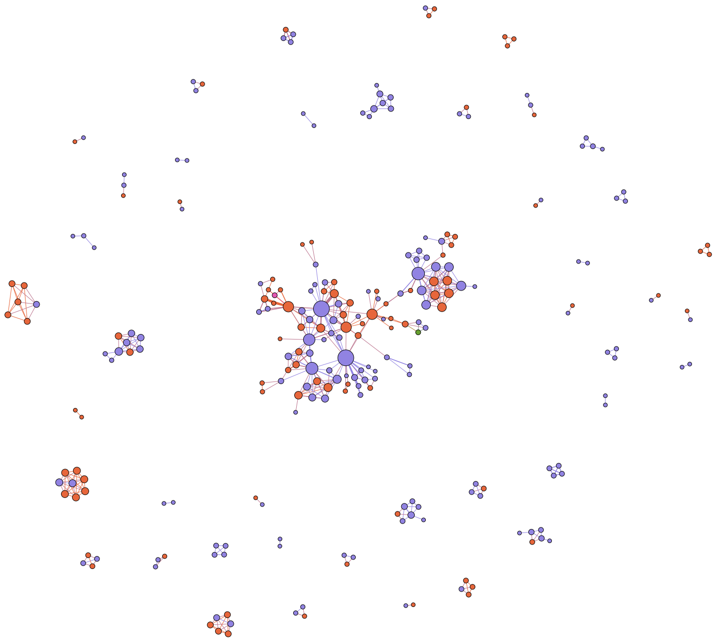


Fig. 1: Co-authorship network of the NKOS community. In general, the network is sparse and contain 44 isolated components. The largest connected component (the cluster in the middle) contains 107 number of nodes. Nodes are colored based on their gender. Purple nodes are men and orange nodes are women.

In Figure 3, the largest connected component, we can easily see that scientists tend to forge intra-institutional collaborations [6]. Good examples are the clusters from Johannes Keizer (FAO), Antoine Isaac (Vrije Universiteit Amsterdam/Europeana) and Philipp Mayr (GESIS). A large fraction of their co-authors are affiliated with the same institution. Also a tendency to select those co-authors who are in geographic proximity is visible in Figure 3. E.g. Douglas Tudhope (University of South Wales, UK) has a larger fraction of UK-affiliated co-authors.

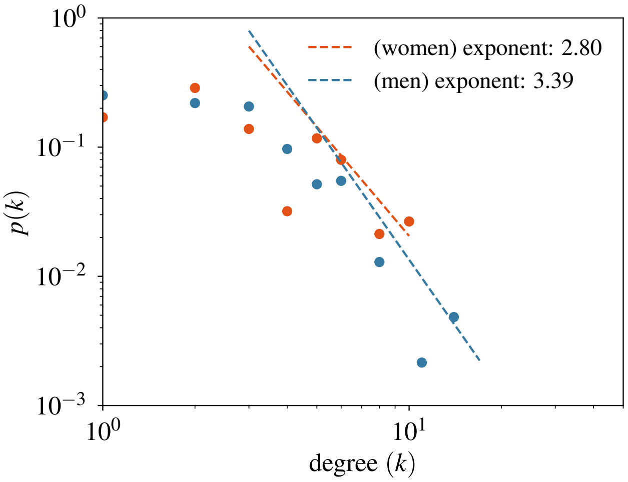


Fig. 2: Degree distribution of the NKOS network. Blue and orange colors indicate the distribution for men and women respectively. Although the network is small, it exhibits power-law degree distribution.

4.1 Node centralities

To detect the influence of authors on information exchange we calculated various measures of centrality namely degree centrality, betweenness centrality and closeness centrality of the authors. Here we only focus on the largest connected component (LCC) in order to have robust comparison.

Degree centrality is the most straightforward measure of centrality that depicts the importance of nodes in terms of total number of unique links. The authors with high degree centrality have established a wide collaboration with many different scholars.

Betweenness centrality indicates fraction of shortest paths between all pairs of nodes that pass through a node. The betweenness of a node indicates the node's ability to funnel the flow in the network [20]. In this network the author with a high betweenness has a large influence in transferring the information from one part of the network to another.

Closeness centrality indicates sum of shortest paths between a node to all other nodes [7]. If a shortest path between node u to v is $d(u, v)$ and the total number of nodes in the graph is denoted by N, closeness centrality of the node u is defined as follows:

$$c(u) = \frac{N-1}{\sum_{v}^{N-1} d(u, v)} \tag{2}$$

where $N-1$ in the nominator normalizes the measure so that it becomes size independent. Scholars with high closeness centrality are on average closer to other nodes in the network.

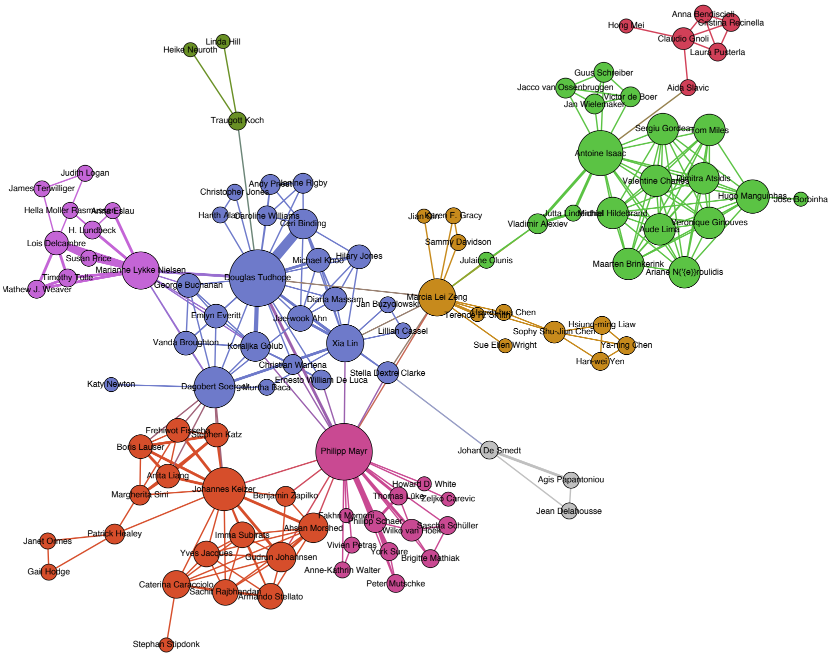
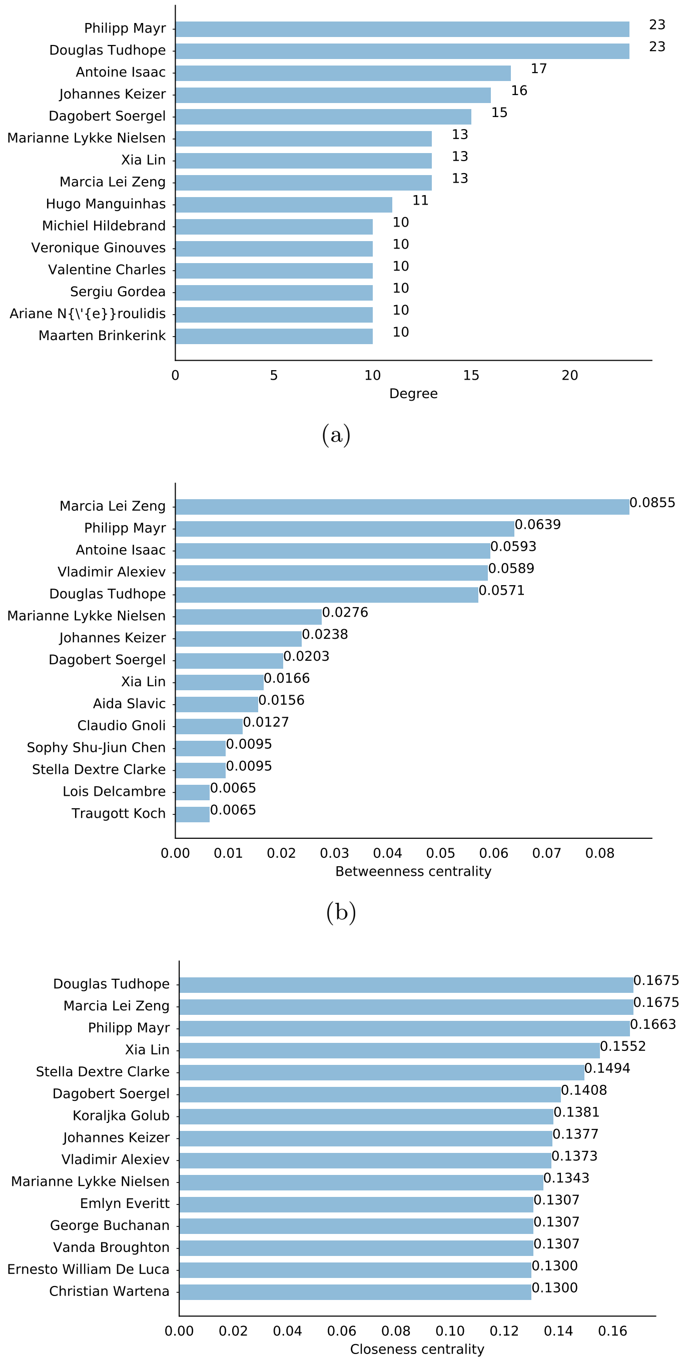


Fig. 3: Largest component in the NKOS co-authorship network. The network is clustered into 9 clusters using Louvain clustering method [2]. Nodes are colored based on their cluster and size of the node represents node's degree. Clusters are shaped based on the location of the groups and collaboration among members. Majority of the scholars in the largest component are based in Europe.



(a)

(b)

(c)

Fig. 4: Top 15 authors with the highest (a) Degree centrality, (b) betweenness centrality and (c) closeness centrality.

Figure 4 shows the comparison of centrality measures for top 15 authors in the largest connected component. It is interesting to note that authors centrality ranks may vary depending on the type of the centrality measures. For example, even though H. Manguinhas has relatively high degree centrality, this author does not appear in the top closeness or betweenness rank. A closer look at the author's location in the graph 3 shows that the author is embedded in the light green cluster with high clustering and few connectivity with other clusters.

Comparing closeness centrality and betweenness centrality also shows interesting results. Although some authors have a high closeness to other scholars, they may not have high betweenness centrality. For example, K. Golub has a relatively high closeness centrality due to special location of the author in connection with many other authors from different clusters. However, this author does not have a relatively high betweenness centrality because her network position does not allow to connect other further distanced clusters. In contrast, author A. Slavic does not have a high degree or a high closeness centrality, but this author has a high betweenness centrality due to connecting an almost isolated red cluster to the rest of the network. The same is true for T. Koch. It is important to note that while scholars with higher closeness centrality are on average closer to other scholars and thus can access novel ideas more frequently, authors with high betweenness centrality play a crucial role in transferring the knowledge in the community [10].

4.2 Structural holes and bridges

Weak ties play a crucial role in networks by connecting disconnected clusters and act as bridges in networks. Structural hole idea first coined by sociologist Ronald Burt, suggests that nodes can act as a mediator between two or more closely connected clusters. This is in particular important since novel ideas or information need to pass from these gatekeepers to transfer to other parts of the network. Here, we measure the effective size of a node based on the concept of redundancy. A persons ego network has redundancy to the extent to which her neighbors are connected to each other as well. In a simple graph, the effective size of a node u, $e(u)$, can be expressed as:

$$e(u) = n - \frac{2t}{n} \tag{3}$$

Where t is the number of the total ties in the egocentric network (excluding those ties to the ego) and n is the number of total nodes in the egocentric network (excluding the ego). The effective size can vary from 1 to the total number of links in the ego [3].

Figure 5 displays the top 15 ranked authors with respect to their effective size. The ranking suggests that in this community, nodes with high degree (hubs) also act as bridges between the clusters, thus, they can transfer novel ideas among their peers.

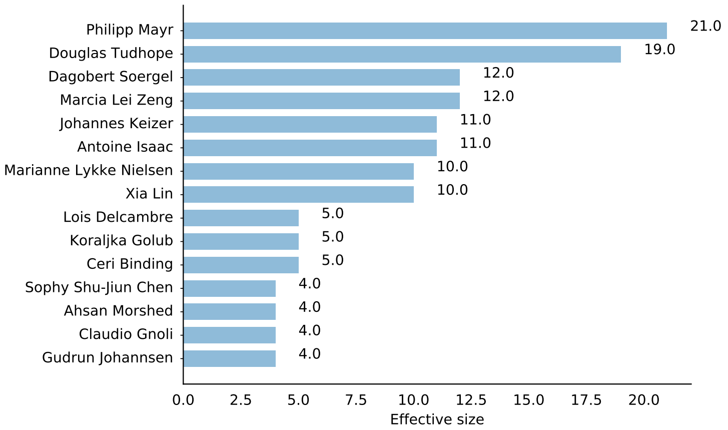


Fig. 5: Top 15 scholars with the highest effective size. The effective size indicates the ability of a node to connect otherwise disconnected nodes and therefore the node can act as a weak tie or bridge.

4.3 Gender differences in the co-authorship network

To infer the gender of the scholars, we use the state-of-the-art approach by combining the results of the first names and Google images of the scholars with their full names [13]. For the remaining unidentified names or names with initials, we manually check the author's online profile based on the title of their papers. Our complete network consists of 97 (38%) women and 157 (62%) number of men and 2 unidentified names. Compared to other scientific communities and in particular in science and engineering fields, this community shows a higher percentage of active women [11]. The share of women and men in the largest connected component also shows an interesting effect. We find 46 women and 59 men in the LCC which means women occupy 43% of the nodes in this component.

Homophily. In the first step, we measure homophily in this network. There are various ways to define homophily. Here, we use two well-defined measures. First measure of homophily is proposed by Newman that computes the Pearson correlation between attributes when corrected by what we would expect from node's degree [19]. The homophily varies between -1 (disassortativity) to +1 (complete assortativity). We find that gender assortativity in this community is 0.1. This means that there is a positive tendency among scholars in this community to collaborate with similar gender. One can observe the gender homophily from figure 1.

Although the assortativity measure captures the overall homophily in the network, it does not provide additional insights whether or not the nature of homophily is symmetric or asymmetric. Indeed, we have shown previously that asymmetric homophily can impact the degree centrality of the nodes and in

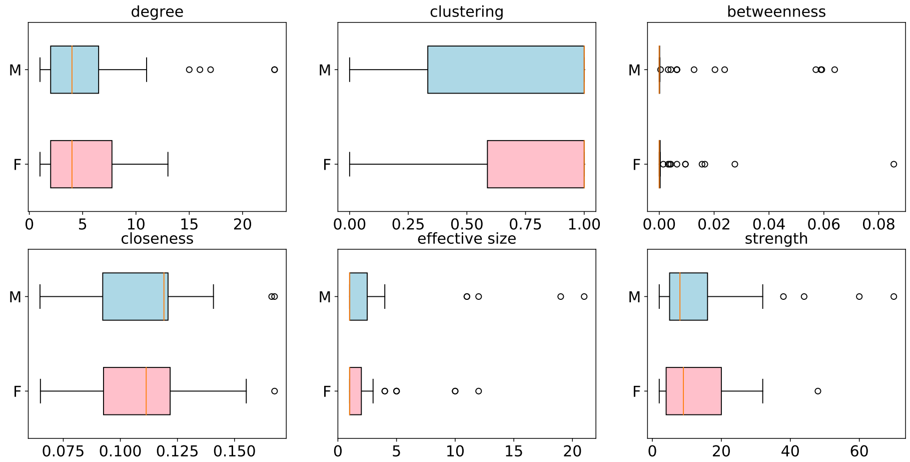


Fig. 6: Box plots indicating median and quartiles of network properties for male and female scholars in the largest connected component. Median is similar for majority of the node characteristics except for closeness centrality that is higher for men. With regards to degree centrality there are more outliers among men with high degree. For clustering, women have higher clustering on average than men. Men also show outliers with higher effective size and strength compared to women.

particular minority group in networks [12]. To capture the asymmetric nature of the homophily, we take a simple approach first proposed by Coleman (1958). In this case we measure the probability of links that exist between two scholars of the same gender. Let us denote the probability of links that exist among women as p_{ww} and among men as p_{mm}. To compare groups of different sizes, the probabilities are compared with group sizes and normalized by the maximum values. If the fraction of women is denoted by f_w and men by f_m, the Coleman index for women is:

$$C_w = \frac{p_{ww} - f_w}{1 - f_w} \tag{4}$$

Similar definition will apply for men. The maximum value for Coleman homophily index is 1. When applying this index to our network we get $C_w = -0.12$ and $C_m = -0.42$. These results suggest that the homophily among women is higher than the homophily among men in this network. Similar findings were also found in other co-authorship networks [11].

Network characteristics and gender differences. Next, we measure the network characteristics among men and women in the largest connected component. We use six measures of networks similar to the previous section. We also include strength of the node as the sum of all wighted links.

Figure 6 shows box plots comparing network measures for men and women. Overall, the median and quartiles for degree and betweenness are the same for men and women. Women show higher tendency for higher clustering compared to men. Men show higher median for closeness centrality compared to women. In addition, there are higher number of outliers among men in terms of the degree, effective size and strength compared to women.

5 Conclusion

In this paper, we have analyzed the collaborative research of authors and their connectivity for the special case of NKOS workshop activities including four special issues on NKOS. The results highlight the most active and central scholars in this community. We found differences among centrality measures of the scholars which indicates that scholars play a different role in their collaboration network. We also found the most influential scholars who act as bridges among the clusters. We found 9 clusters in the largest component that show scholars have higher tendency to collaborate with those in the same institution or the same geographic proximity [6]. Our analyses show that NKOS community is rather successful in bringing researchers from different domains together in recent years.

NKOS co-authorship network consists of 38% women in total, and the share of women in the largest connected component is 43%. The network shows positive gender homophily and the homophily among women is higher compared to men. We found on average men have higher closeness centrality compared to women. In addition, women have slightly higher clustering compared to men. Apart from these differences, we do not find any significant differences between men and women with respect to their centrality.

This study has some limitations. First of all, we have included just research paper presentations. Editing and organizing activities at the workshops, which have an enormous impact on the visibility and connectivity of researchers, have not been covered in our dataset. This leads to artifacts, e.g. Traugott Koch,[5] a long-term organizer of the NKOS workshops and editor of the early JoDI special issues on NKOS, is not covered very well in our dataset and the network.

Second, many influential papers (e.g. [9, 26]) and standardization activities (e.g. the W3C Recommendation for SKOS [17]), presented and discussed at NKOS events and published after the NKOS workshops are missing. This fact is of course reducing the expressiveness and completeness of the network.

Third, we have not included bibliometric data to complete our analysis. This is because most of the NKOS workshop activities (presentations) are not formally cited or even mentioned in scientific papers. In difference to the workshop output, the few journal papers in the special issues on NKOS are cited. Some works (e.g. [4, 5, 23, 21, 22]) are cited well in the literature. So adding citation data would be a next reasonable step to complete the dataset.

[5] Traugott Koch was an central protagonist and networker of the US and European NKOS community. He retired and left the NKOS community in 2012.

6 Future work

We are planning to extend the analysis of the NKOS network. In this way we first plan to complement the dataset with other NKOS research output. We also plan to analyze the development of topics in the titles and abstracts of the presentations and papers. Combining network analytic measures with bibliometric analysis (e.g. co-citations, bibliographic coupling) would complement our preliminary observations and advance our understanding of the role of gender and other attributes in scientific collaboration. We invite people to contribute to our open dataset.

7 Acknowledgment

We thank our colleague Marcia Lei Zeng (Kent State University) who provided us with internal information about the U.S. NKOS workshops. This work was partly funded by DFG, grant no. SU 647/19-1; the "Opening Scholarly Communication in the Social Sciences" (OSCOSS) project at GESIS.

References

1. Barabási, A.L.: Scale-free networks: a decade and beyond. science 325(5939), 412–413 (2009)
2. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment 2008(10), P10008 (2008)
3. Burt, R.S.: Structural holes and good ideas. American Journal of Sociology 110(2), 349–399 (2004)
4. Cranefield, S.: Networked knowledge representation and exchange using uml and rdf. Journal of Digital Information (2001), https://journals.tdl.org/jodi/index.php/jodi/article/view/30
5. Doerr, M.: Semantic problems of thesaurus mapping. Journal of Digital Information (2001), https://journals.tdl.org/jodi/index.php/jodi/article/view/31
6. Evans, T.S., Lambiotte, R., Panzarasa, P.: Community structure and patterns of scientific collaboration in Business and Management. Scientometrics 89(1), 381–396 (Oct 2011), http://link.springer.com/10.1007/s11192-011-0439-1
7. Freeman, L.C.: Centrality in social networks conceptual clarification. Social Networks 1(3), 215–239 (1978)
8. Hill, L., Koch, T.: Networked Knowledge Organization Systems: introduction to a special issue. Journal of Digital Information 1(8) (2001), https://journals.tdl.org/jodi/index.php/jodi/article/view/32/33
9. Hodge, G.: Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files (2000), https://www.clir.org/pubs/reports/pub91/pub91.pdf
10. Iyer, S., Killingback, T., Sundaram, B., Wang, Z.: Attack robustness and centrality of complex networks. PloS one 8(4), e59613 (2013)

11. Jadidi, M., Karimi, F., Wagner, C.: Gender disparities in science? dropout, productivity, collaborations and success of male and female computer scientists. arXiv preprint arXiv:1704.05801 (2017)
12. Karimi, F., Génois, M., Wagner, C., Singer, P., Strohmaier, M.: Visibility of minorities in social networks. arXiv preprint arXiv:1702.00150 (2017)
13. Karimi, F., Wagner, C., Lemmerich, F., Jadidi, M., Strohmaier, M.: Inferring gender from names on the web: A comparative evaluation of gender detection methods. In: Proceedings of the 25th International Conference Companion on World Wide Web. pp. 53–54. International World Wide Web Conferences Steering Committee (2016)
14. Mayr, P., Tudhope, D., Clarke, S.D., Zeng, M.L., Lin, X.: Recent applications of Knowledge Organization Systems: introduction to a special issue. International Journal on Digital Libraries 17(1), 1–4 (2016), http://link.springer.com/10.1007/s00799-015-0167-x
15. Mayr, P., Tudhope, D., Golub, K., Wartena, C., De Luca, E.W.: Proceedings of the 15th European Networked Knowledge Organization Systems (NKOS) Workshop. CEUR-WS.org (2016), http://ceur-ws.org/Vol-1676/
16. Mayr, P., Tudhope, D., Golub, K., Wartena, C., De Luca, E.W.: Proceedings of the 17th European Networked Knowledge Organization Systems (NKOS) Workshop. CEUR-WS.org (2017), http://ceur-ws.org/Vol-1937/
17. Miles, A., Bechhofer, S.: SKOS Simple Knowledge Organization System Reference (2009), https://www.w3.org/TR/skos-reference/
18. Momeni, F., Mayr, P.: Analyzing the research output presented at European Networked Knowledge Organization Systems workshops (2000-2015). In: Proc. of the 15th European Networked Knowledge Organization Systems Workshop (NKOS 2016). pp. 7–14. CEUR-WS.org, Hannover, Germany (2016), http://ceur-ws.org/Vol-1676/paper1.pdf
19. Newman, M.E.: Assortative mixing in networks. Physical review letters 89(20), 208701 (2002)
20. Opsahl, T., Agneessens, F., Skvoretz, J.: Node centrality in weighted networks: Generalizing degree and shortest paths. Social Networks 32(3), 245–251 (2010), http://dx.doi.org/10.1016/j.socnet.2010.03.006
21. Soergel, D., Lauser, B., Liang, A., Fisseha, F., Keizer, J., Katz, S.: Reengineering thesauri for new applications: the agrovoc example. Journal of Digital Information (2004), https://journals.tdl.org/jodi/index.php/jodi/article/view/112
22. Trant, J., with the participants in the steve.museum project: Exploring the potential for social tagging and folksonomy in art museums: Proof of concept. New Review of Hypermedia and Multimedia (2006), http://www.tandfonline.com/doi/abs/10.1080/13614560600802940
23. Tudhope, D., Alani, H., Jones, C.: Augmenting thesaurus relationships: possibilities for retrieval. Journal of Digital Information (2001), https://journals.tdl.org/jodi/index.php/jodi/article/view/181/160
24. Tudhope, D., Koch, T.: New Applications of Knowledge Organization Systems: introduction to a special issue. Journal of Digital Information 4(4) (2004), https://journals.tdl.org/jodi/index.php/jodi/article/view/109/108
25. Tudhope, D., Lykke Nielsen, M.: Introduction to Knowledge Organization Systems and Services. New Review of Hypermedia and Multimedia 12(1), 3–9 (2006)
26. Zeng, M.L., Chan, L.M.: Trends and Issues in Establishing Interoperability Among Knowledge Organization Systems. Journal of the American Society for Information Science and Technology 55(3), 377–395 (2004)