Social Analysis of the SEKE Co-Author Network

Rehab El Kharboutly
Software Engineering
Quinnipiac University
Hamden, CT 06518
Ruby.elkharboutly@quinnipiac.edu

Swapna S. Gokhale
Computer Science & Engg.
Univ. of Connecticut
Storrs, CT 06269
ssg@engr.uconn.edu

Abstract

We extract the co-author network over the entire history of the SEKE conference from 1988 through 2014. In this network, authors represent nodes and a pair of authors is connected by an edge if they have co-authored at least one article over the entire duration. We analyze this network using socio-centric and ego-centric network methods to study the extent to which the authors are involved in the SEKE community, and the patterns of collaboration between them. Socio-centric analysis reveals that most authors publish a very small number of articles, and collaborate within tightly knit circles. In fact, only a tiny fraction of the authors consistently return to SEKE to disseminate their research. Ego-centric measures of centrality confirm these findings by revealing that only a small percentage of the authors are structurally dominant, and influence the flow of communication among others. Based on these findings, we believe that strategically SEKE could benefit from cultivating a wider base of influential authors, promoting broader collaborations, and encouraging one-time authors to return.

Keywords

Co-author network, Clustering Coefficient, Centrality

1 Introduction and Motivation

The International Conference on Software Engineering and Knowledge Engineering (SEKE), now in its 27^{th} year, is a premier conference that aims at bringing together experts in either Software Engineering (SE) or Knowledge Engineering (KE) or both. Specifically, the conference seeks to emphasize the transference of methods between both domains [11]. Since its inception in 1988, SEKE has consistently expanded; both in terms of the number of papers and number of authors by welcoming contributions from traditional SE and KE topics as well as emerging areas.

Participation of researchers and authors is vital to the long-term survival of any conference. A conference is sustained by those authors who consistently return to the venue. However, for healthy growth, a conference should also seek to attract new contributors into its fold, and simultaneously foster collaborations between existing and new authors. Collaborative work offers many advantages, individually for the authors, collectively for the entire scientific community, and finally for the conference itself. It leads to cross-fertilization of ideas, sharing of resources, skills and workload, efficiency in the use of time, and avoidance of competition – all in the pursuit of a mutually shared, common goal. Collaborative research also encourages knowledge sharing, which is essential for knowledge creation, because a person's limited cognitive capability, and bounded rationality [25] imposes a natural limit on what an individual working alone can achieve. Thus, working together can increase the research productivity and impact of an individual [13]. Finally, fostering collaboration can strategically improve the quality, stature, and reputation of a conference, because when a conference spurs the collaboration, it is very likely that the new team chooses the same venue to publish their new joint work.

Co-authorship may be regarded as a strong evidence or an explicit product of collaborative work [10]. In fact, a significant proportion of scientific collaboration leads to co-authored articles. Collectively, such joint authorship of research articles leads to a network, where nodes are authors and links between two authors (nodes) represents at least one joint article between them. Such a co-author network can be considered to be the first-order approximation of complete scientific collaboration network [19]. Therefore, a study of the co-author network can offer insights into whether the social structure of a conference community is conducive to collaborative research. It can also offer suggestions on how such synergistic effort can be promoted.

In this paper, we study the co-author network of the SEKE conference, extracted from the DBLP, KSI and elec-

(DOI Reference Number: 10.18293/SEKE2015-092)

tronic proceedings spanning years 1988 through 2014. We studied this network using socio-centric and ego-centric analysis methods to understand the degree to which authors are embedded in the SEKE community, and the structural patterns of interactions among them. Socio-centric analysis reveals that most authors have published opportunistically (one or two) papers in SEKE, and only very few return consistently. Moreover, most authors collaborate within their small, tightly knit circles of $2-3$ collaborators. Approximate power-law spreads of ego-centric measures of centrality, namely, degree, closeness, and betweenness confirm these socio-centric observations by revealing that very few authors are structurally dominant, and control and influence the flow of information and communication among others. Based on these observations, we believe that the SEKE conference could derive long-term benefits by strategically: (i) cultivating a wider base of influential authors structurally embedded in the community; (ii) promoting broader collaborations; and (iii) encouraging one-time authors to return.

The rest of the paper is organized as follows: Section 2 describes data collection and pre-processing. Section 3 and 4 discuss socio-centric and ego-centric analyses respectively. Section 5 surveys related work. Section 6 concludes the paper and offers directions for future work.

2 Data Collection and Pre-Processing

We extract the co-author data from 26 editions of SEKE conferences from its inception in 1988 to its most recent in 2014. Of these years, proceedings for 2013 and 2014 were obtained electronically, data for 1991, 1997, and 1998 from the SEKE website, and the rest from DBLP. Since most data came from DBLP, and was retrieved in XML format, the textual citations obtained from the web and electronic proceedings were parsed and formatted into XML as well. To the best of our knowledge, DBLP does not provide a way to download data to a txt file, so we manually transferred the XML entries to a file. Figure 1 shows an example entry in the DBLP proceedings in the XML format.

We pre-processed this data to replace all the special characters with acceptable XML characters, especially in European names. Authors wrote their names in multiple formats, including first and last name, first and middle initials and last name but the most common representation was first initial, last name. Thus, we translated all the names into this common format. We noticed that many authors shared a last name, but it was very rare (only $4-5$ instances) for authors to share the combination of first initial and last name. We manually disambiguated between such authors by consulting their affiliations and emails; assuming that authors who share a name but not affiliation and/or email represent different individuals. We added unique tags to identify identical combinations that represent different individuals.

Table 1: Socio-centric Metrics

Metric	Value
Duration	$1989-2014$
Total Number of Authors	2990
Total Number of Papers	1738
Average Papers Per Author	1.49
Average Collaborators Per Author	2.46
Density	0.0059
Average Component Size	4.9
Largest Connected Component	1126
Average Path Length	7.329
Diameter	22
Average Clustering Coefficient	0.852
Number of triangles	4268

After pre-processing, we implemented a parser to extract pairs of collaborators from the XML entries. The pairwise list of authors created by the example XML entry is in Figure 2. This list was checked each time a newly created pair matches an existing entry in the list; if there is a match, the number of contributions for that pair is incremented, otherwise a new pair is added with a collaboration count of 1. We also maintain the number of papers and collaborators for each author. Finally, this list of collaborator pairs is used to create the adjacency matrix. We note that although we keep track of the collaboration count for each pair, for this analysis the adjacency matrix represents an unweighted graph. That is, if authors A and B have co-authored at least one paper, the corresponding element in the matrix is set to 1, otherwise it is set to 0. Altogether we processed 1738 articles written by 2990 authors to build the SEKE co-author network.

```
B. Cheng,R. Bourdeau
R. Bourdeau,G. Gannod
C. Cheng,G. Gannod
```

Figure 2: List of Author Pairs – DBLP Entry in Figure 1

3 Socio-centric Analysis

In this section, we discuss socio-centric metrics that are computed over all the nodes in the network. We compare these metrics, shown in Table 1, with those of other scientific communities within and beyond computer science.

3.1 Individual or Local Metrics

Individual metrics are computed locally by considering the immediate connections of each author to understand the

```
<inproceedings key="conf/seke/ChengBG94" mdate="2007-02-23">
<author>Betty H. C. Cheng</author> <author>Robert H. Bourdeau</author>
<author>Gerald C. Gannod</author> <title> The object-oriented
 development of a distributed multimedia environmental information system.
 </title>
<pages>70-77</pages>
<year>1994</year>
<crossref>conf/seke/1994</crossref>
<booktitle>SEKE</booktitle>
<url>db/conf/seke/seke1994.html#ChengBG94</url>
</inproceedings>
```

Figure 1: Example DBLP Entry in XML Format

author's involvement in the SEKE community. We found that on an average a SEKE author writes 1.49 articles, collaborates with 2.46 others, and on an average a SEKE article has 1.5 authors. These values are low compared to biology (6.4, 3.75 and 18.1) and physics (5.1, 2.53 and 9.7) co-author networks [22]. This difference may arise because biologists and physicists may need to collaborate more frequently and widely due of the experimental nature of their work. However, these metrics are consistent with the values for the co-author network in library and communication sciences (2.40, 1.80, and 2.24) [15], a field that may be closer to SEKE in terms of culture, traditions, practices, and norms. Figure 3, which shows the distribution of the number of authors per article further confirms that a very large percentage of SEKE articles has four or fewer authors. Articles with five or more authors are very rare, with seven being the maximum.

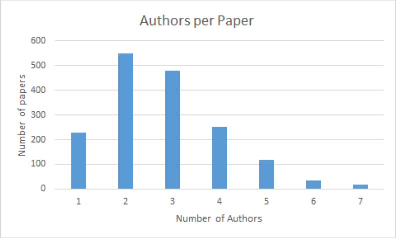

Figure 3: Distribution of Authors per Article

Figure 4, which shows the distribution of number of articles per author indicates that a large percentage of the authors publish only one article, and a very small percentage publishes three or more articles. This approximate power-law spread suggests that while the conference enjoys a very small loyal base, most authors opportunistically choose SEKE. Figure 5, which shows the distribution of the number of collaborators shows that a significant proportion of authors have between 1 and 6 collaborators. A group with two to three collaborators could represent a graduate advisor with his or her doctoral students. A very small percentage with no collaborators could represent solo authors. Finally, groups with 5 or more members could represent collaborations across institutions.

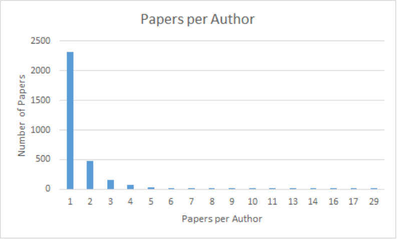

Figure 4: Distribution of Articles Per Author

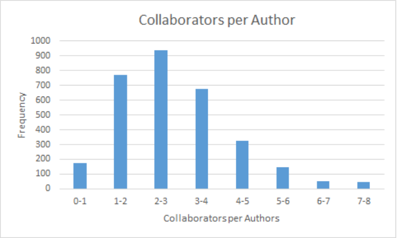

Figure 5: Distribution of Collaborators Per Author

3.2 Aggregate or Global Metrics

Aggregate social metrics, computed by considering the global network, will reveal the degree of closeness of the SEKE community structure.

- **Network Density:** The network density (D) is defined as the number of edges T to the number of possible edges and is given by Equation (1). The density of the SEKE co-author network is 0.0059, indicating an overall very sparsely connected network.

$$D = \frac{T}{N(N-1)} \tag{1}$$

- **Average Path Length:** This is calculated by finding the shortest path between a pair of nodes and then dividing by the total number of pairs. This shows, on an average, the number of steps it takes to reach one author from the other. The average path length of 7.329 and diameter of 22 indicates that the SEKE network does not exhibit small-world properties, where the average path length and diameter is around 2.0 and 6 respectively. This is surprising because we would expect stronger homophily between SEKE authors, who are mostly computer scientists, than other types of shared interests based on geographic proximity, or organizational affiliation, which typically lead to small-world properties in social networks.

- **Clustering Coefficient:** This measures the degree to which the authors group together so that the probability of a tie between two authors in a cluster is greater than the probability of a tie between any two random authors. The clustering coefficient is defined as the average clustering coefficient of all the nodes [5], where the clustering coefficient C_v for a node v is the proportion of all possible edges between the neighbors of a node that actually exist. The clustering coefficient is based on the number of triangles or closed and open triplets. The average clustering coefficient is a real number between 0 and 1, with the SEKE value being 0.8268. 734 nodes have been excluded from this computation because they have only one edge, and the network has 4268 triangles. The value closer to 1 suggests the presence of a small yet, close community and a large number of isolated groups. The distribution of the clustering coefficient in Figure 6 supports this conjecture, with a large peak at a very high value.

- **Component Sizes:** Similar to other co-author networks, the SEKE network is not a single connected graph. Therefore, to measure the degree of connectivity, we measure the relative size of the largest connected component as its actual size divided by the size of the network, which is approximately 38%. Previously, the relative sizes of the largest connected components were observed to be 20% for the library and communication science [15], 60% for SIGMOD [21], 92.6% for Medline, and 57.2% for NCSTRL [23]. The relative size for SEKE is consistent with Kretschmer's [12] observation that the largest components usually have a ratio of around 40%. This relative size may also be impacted by the nature of the disciplines, experimental sciences such as biology and physics may have larger connected components compared to computational disciplines such as SEKE and library and communication sciences. The distribution of the connected component sizes shown in Figure 7 has a peak at component size of two, followed by a size of three. This suggests that most authors collaborate within their comfort zone of friends, colleagues or members within their research group, rather than seeking out completely new partners.

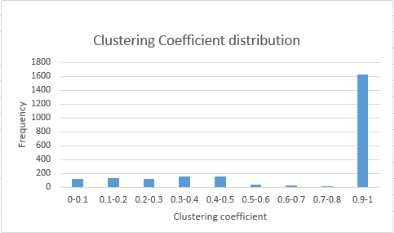

Figure 6: Distribution of Clustering Coefficient

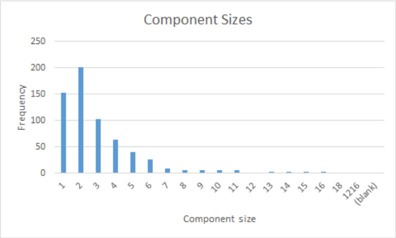

Figure 7: Distribution of Component Sizes

4 Ego-centric Analysis

In this section, we discuss ego-centric measures of centrality which is an important structural attribute that indicates an author's formal power or prominence in the network relative to the others [4]. We study the distribution of these centrality scores, computed for each node in the largest connected component.

4.1 Degree Centrality

The degree centrality $C_D(v)$ of an author v is given by the degree of v and is defined by Equation (2). It measures the number of ties of an author with others. Authors with more connections or a higher degree are more central to the structure, and tend to have a greater capacity to influence others. An author may have a high degree because he or she appears as a co-author on many articles, but each paper has a short list of authors. Alternatively, a highly connected author may appear as a co-author on a few articles, but each article is authored by many. While this measure does not consider connection strength, it does capture an author's collaborative scope within the network.

$$C_d(v) = deg(v) \tag{2}$$

4.2 Closeness Centrality

Closeness centrality is defined as the mean shortest distance by which a given author is separated from all the others [18]. It is measured as the average of the total reciprocal distance of an author from each of the other authors. Closeness centrality of an author v is given by Equation (3), where $d(i, j)$ is the distance between the two authors i and j, and N is the number of authors. A message originating in the most central position (i.e. from the author with the highest closeness centrality) would spread throughout the network in minimum time. Moreover, an author with high closeness centrality could access or obtain the resources owned by others more efficiently than any other author. Therefore, closeness centrality is a surrogate measure of an author's efficiency in communicating with others.

$$C_c(v) = \sum_{k=1}^{N} \frac{1}{d(i,j)} \tag{3}$$

4.3 Betweenness Centrality

Betweenness centrality is defined as the proportion of the shortest paths between all pairs that pass through a given author [3]. It represents an author's ability to control the flow of resources or information, which enables the author to broker information and resources to others [8]. Betweenness centrality of an author v is given by Equation (4), where $g_{j,v,k}$ is all the geodesics linking authors j and k which pass through author v and $g_{j,k}$ is the geodesic distances between authors j and k. Authors with high betweenness centrality play the role of a "middleman" or a "bridge" and could gain different resources and information from different groups. Also, when authors with high betweenness are removed, it typically results in the largest increase in the distance between others. It thus measures authors' importance to others' virtual communication.

$$C_B(v) = \sum_{j,k \neq v} \frac{g_{j,v,k}}{g_{j,k}} \tag{4}$$

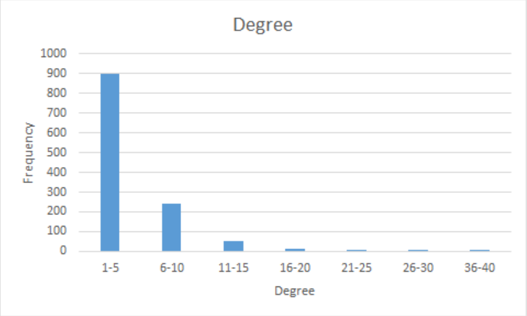

Figure 8: Distribution of Degree Centrality

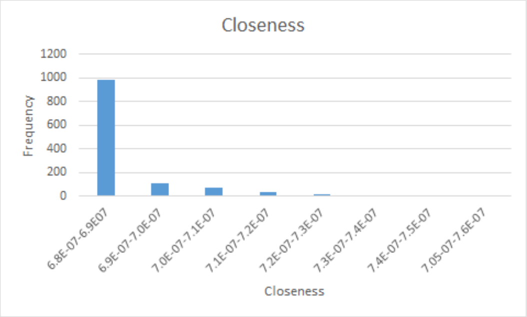

Figure 9: Distribution of Closeness Centrality

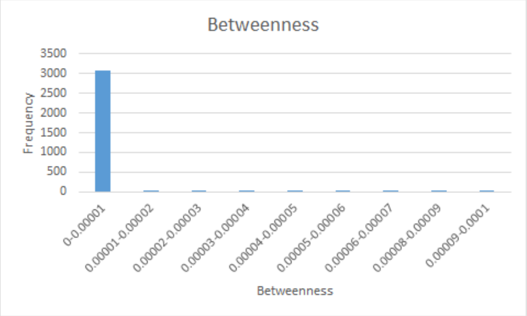

Figure 10: Distribution of Betweenness Centrality

Figures 8, 9, and 10 respectively show the distributions of the degree, closeness, and betweenness centralities for the SEKE co-author network. The spreads of these measures can be approximated using power-law distributions. The distribution of degree centrality in Figure 8 indicates that a large number of authors have a small number of collaborators, and only a fraction collaborate with a large number of others. The co-author network, color coded according to the node degree ranging from 1 to 37, depicted in Figure 11 confirms this distribution. In this network, blue nodes, which make up only a small fraction have the highest degree. Similarly, the distribution of closeness centrality in Figure 9 indicates that a very small number of authors are highly efficient in communicating with others and accessing their resources. Betweenness centrality distribution in Figure 10 suggests that after a large spike at the lowest value, the remaining values show the same proportions. Thus, a majority of the authors do not lie on the shortest paths between other pairs. Thus, in summary, although each centrality measures a different aspect of authors' embeddedness, we find that a very small fraction of SEKE authors lie in prominent positions. These authors sport a large number of collaborators, lie on the shortest paths between other pairs, and are highly efficient communicators.

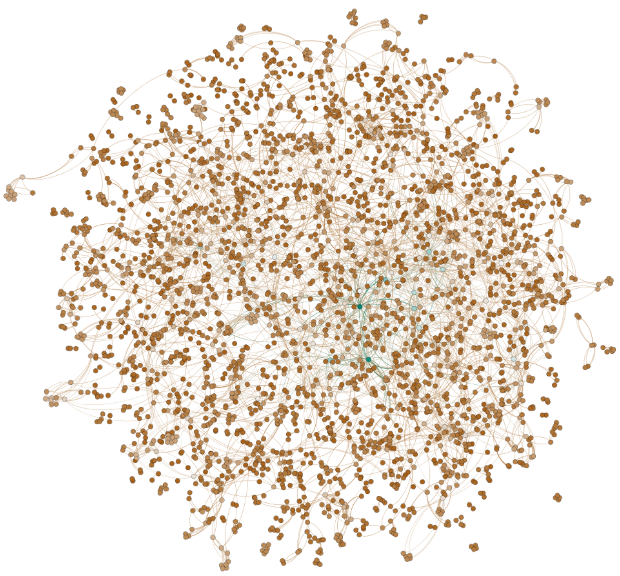

Figure 11: SEKE Co-Author Network

5 Related Work

Social network measures have been used to study the properties of co-author networks in various fields including mathematics, biology, physics, and computer science. Some authors also study how the network structure impacts local or micro-level properties including citation counts. These works, their measures and data, and their key objectives and findings are summarized in Table 2.

Most of the works in Table 2 study readily available, archived data from large communities such as Medline, Physics or Mathematics authors. They also assess the impact of network structure on micro-level properties of individual authors or articles, mostly captured in the form of citation counts. Our work can be distinguished in the following two ways: (i) we extract and process the co-author network of the SEKE conference from three sources; and (ii) we corroborate socio-centric and ego-centric measures to offer recommendations on how the SEKE conference could strategically improve its stature. The SEKE conference, by the virtue of its more than 25 years of history as a premier conference at the interplay of SE and KE, affords this unique opportunity.

6 Conclusions and Future Work

In this paper, we describe the process of extracting the network of SEKE co-authors over the entire history of the conference. We analyze this network using socio-centric and ego-centric network analysis methods to understand patterns of author involvement and collaboration. Corroborating the results from both these analyses reveals that the SEKE conference is characterized by a large percentage of authors who publish one or two papers, and who collaborate in tightly knit circles. A small fraction of the authors enjoy structural dominance in the network, and control and influence the flow of information and communication. Based on these findings, we offer recommendations that could strategically benefit SEKE.

Our future research involves longitudinal analysis to understand how the patterns of collaboration have evolved since the early editions of the conference.

References

[1] A. Abbasi, K. S. K. Cheung, and L. Hossain. “Egocentric analysis of co-authorship network structure, position and performance”. *Information Processing and Management*, 48:671–679, 2012.

[2] J. Bollen, M. A. Rodriguez, and H. Van De Sompel. Journal status. *Scientometrics*, 69(3), 2006.

[3] S. P. Borgatti. “Centrality and network flow. *Social Network*, 27(1):55–71, 2005.

[4] M. E. Burkhardt and D. J. Brass. “Changing patterns or patterns of change? The effects of a change in technology on social network structure and power”. *Administrative Science Quarterly*, 35(1):104–127, 1990.

[5] X. Cheng, C. Dale, and J. Liu. “Statistics and social network of YouTube videos”. In *Proc. of Intl. Workshop on Quality of Service*, pages 229–238, 2008.

[6] R. P. Dellvalle, L. M. Schilling, M. A. Rodriguez, H. Van de Sompel, and J. Bollen. “Refining dermatology journal impact factors using PageRank”. *Journal of the American Academy of Dermatology*, 57(1):116–119, 2007.

[7] Y. Ding. “Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks”. *J. Informetr*, 5(1):187–203, 2011.

[8] L. C. Freeman. “Centrality in social networks: Conceptual clarifications”. *Social Network*, 1(3):215–239, 1979.

[9] C. N. Gonzalez-Brambila, F. M. Veloso, and D. Krackhardt. “The impact of network embeddedness on research output”. *Research Policy*, 42:1555–1567, 2013.

[10] B. He, Y. Ding, and E. Yan. “Mining enriched contextual information of scientific collaboration: A meso

Table 2: Research Summary: Co-Authorship Networks

Citation	Measures	Data	Objectives
Mutschke [20]	Centrality	Digital Libraries	Collaboration patterns
Liu *et. al.* [17]	Centrality	Joint Conf. on Digital Libraries	Citation Counts
Yan *et. al.* [15]	Centrality	Library & Commn. Science	Citation Counts
Bollen *et. al.* [2]	Weighted Page Rank	ISI Journals	Prestige and Status
Dellvale *et. al.* [6]	Weighted Page Rank	Dermatology	Prestige, Status
Leydesdroff [14]	Centrality	Journal Citation Reports	Interdisciplinarity
Gonzalez *et al.* [9]	Centrality	Mexican Researchers	Research Productivity
Abbasi *et. al.* [1]	Degree Centrality Structural Holes	Library & Info. Science	G-index
Sarigol [24]	Centrality	CS publications	Citations
Newman [22]	Centrality	Medline Physics arXiv Mathematical Rev.	Collaboration Patterns
D'Amour *et. al.* [16]	Centrality	Patents	
Ding [7]	Topic Modeling Path Analysis	Information Retrieval	Topics, Citation

perspective". *Journal of the American Society for Information Science and Technology*, 62(5):831–845, 2011.

[11] Knowledge Systems Institute. `http://www.ksi.edu/seke/seke15.html`.

[12] H. Kretschmer. "Author productivity and geodesic distance in bibliographic co-authorship networks and visibility on the Web". *Scientometrics*, 60(3):409–420, 2004.

[13] S. Lee and B. Bozeman. "The impact of research collaboration on scientific productivity". *Social Studies of Science*, 35(5):673–702, 2005.

[14] L. Leydesdroff. "Betweenness centrality as an indicator of the interdisciplinarity of scientific journals". *Journal of the American Society of Information Science and Technology*, 58(9):1303–1319, 2007.

[15] E. Y. Li, C. H. Liao, and H. R. Yen. "Co-authorship networks and research impact: A social capital perspective". *Research Policy*, 42:1515–1530, 2013.

[16] G. C. Li, R. Lai, A. D'Amour, D. M. Doolin, Y. Sun, V. I. Torvik, A. Z. Yu, and L. Fleming. "Disambiguation and co-authorship networks of the U.S. patent inventor database (1975-2010). *Research Policy*, 43(6):941–955, 2013.

[17] X. Liu, J. Bollen, M. L. Nelson, and H. V. Sompel. "Co-authorship networks in the digital library research community". *Information Processing and Management*, 41:1462–1480, 2005.

[18] H. Lu and Y. Feng. "A measure of authors' centrality in co-authorship networks based on the distribution of collaborative relationships". *Scientometrics*, 81(2):499–511, 2009.

[19] T. Martin, B. Ball, B. Karrer, and M. E. J. Newman. "Coauthorship and citation in scientific publishing. *arXiv preprin arXiv:1304.0473*, 2013.

[20] P. Mutschke. "Mining networks and central entities in digital libraries: A graph theoretic approach applied to co-author networks". *Advances in Intelligent Data Analysis*, 2810:155–166, 2003.

[21] M. A. Nascimento, J. Sander, and J. Pound. "Analysis of SIGMOD's co-authorship graph". *SIGMOD Record*, 32(3):8–10, 2003.

[22] M. E. J. Newman. "Coauthorship networks and patterns of scientific collaboration". *Proc. of the National Academy of the Sciences of the United States of America*, 101(1):5200–5205, April 2001.

[23] M. E. J. Newman. "The structure of scientific collaboration networks". *Proc. of the National Academy of Science of the United States of America*, 98(2):404–409, 2001.

[24] E. Sarigol, R. Pfitzner, I. Scholtes, A. Garas, and F. Schweitzer. "Predicting scientific success based on coauthorship networks. *arXiv:1402.7268*, 2014.

[25] H. A. Simon. *Administrative Behavior*. Free Press, New York, 1976.