Recommending Intra-Institutional Scientific Collaboration through Coauthorship Network Visualization

Gustavo A. Parada
Tecnologico de Monterrey
Campus Monterrey
Monterrey, Mexico
A00812667@itesm.mx

Hector G. Ceballos
Tecnologico de Monterrey
Campus Monterrey
Monterrey, Mexico
ceballos@itesm.mx

Francisco J. Cantu
Tecnologico de Monterrey
Campus Monterrey
Monterrey, Mexico
fcantu@itesm.mx

Lucia Rodriguez-Aceves
Tecnologico de Monterrey
Centro de Sistemas de Conocimiento
Monterrey, Mexico
lucia.rodriguez@itesm.mx

ABSTRACT

For improving research productivity, quality and dissemination, we propose the development of a visual recommendation tool summing up scientific collaboration best-practices found in literature. Social Network Analysis are applied to a coauthorship network for generating a Potential Collaboration Index (PCI) based on productivity, connectivity, similarity and expertise. This work is evaluated by recommending intra-institutional collaboration in a comprehensive university. The accuracy of PCI is documented, along with suggestions and comments from 27 interviewed researchers.

Categories and Subject Descriptors

E.1 [**Data Structures**]: Graphs and networks; J.4 [**Social and Behavioral Sciences**]: Sociology

Keywords

Social Network Analysis; Scientific Collaboration; Coauthorship Networks; Network Visualization; Potential Collaboration Index.

1. INTRODUCTION

The collaboration of scientists in research activity has become the norm [5]. The increasingly interdisciplinary, complex, and costly characteristics of modern science encourage scientists to get involved in collaborative research. The growing number of collaborations among academicians [16], shows how they are embedded in networks of relationships where they exchange information, ideas, and resources [18].

CompSci'13, October 28, 2013, San Francisco, CA, USA.

ACM 978-1-4503-2414-4/13/10
http://dx.doi.org/10.1145/2508497.2508499 ...$15.00.

The most commonly methods used for studying collaboration networks have been: network analysis, bibliometrics, and qualitative methods [9]. Regarding quantitative methods, Social Network Analysis is highly used. Specifically co-authorship is the most visible and accessible indicator of scientific collaboration [19]. These studies correlate scientist attributes and their position in the co-authorship network with high scientific productivity [2, 21, 14, 11].

2. BACKGROUND

In this section we document those patterns on scientific collaboration associated to high researcher's performance. Next we provide a definition of co-authorship networks and present some SNA tools that can be used for analyze them.

2.1 Patterns on Scientific Collaboration

Scientific collaboration has been studied by observing and quantifying the dynamic structure of co-authorship and citation networks, as well as the access to resources and knowledge gained through collaborators.

2.1.1 Collaboration as competitive advantage

Newman's studies on co-authorship networks demonstrates that higher levels of collaboration, measured in terms of coauthors, are correlated with higher productivity (number of published papers) [19]. Additionally, Waltman et al. observed that collaborative publications tend to be cited more frequently than non-collaborative publications [22].

2.1.2 Collaboration network

According to McDonald, individual's social network strongly influences information seeking and collaboration behavior [17]. In this sense, Newman discovered that the probability of an individual gaining a new connection is proportional to the number of connections he already has [19]. Newman calls *preferential attachment* to the effect that a high number of coauthors and citations has over author's visibility.

2.1.3 Popularity and prestige in science

In science, *popularity* of an author can be expressed by the number of coauthors he has, whereas author *prestige* is denoted by the number of citations received by his papers. It has been observed that popularity and prestige of publication's authors impact positively their citation, independently of the Journal impact factor [21].

The preferential attachment referred by Newman is also described by Kas, Carley & Carley: authors receive credit/wealth in proportion to what they have already accumulated. In other words, a paper that has already been cited by many papers holds a visibility advantage over other less-cited papers, leading to the emergence of power laws [14].

2.1.4 Temporal, geographical and social closeness

Kas, Carley & Carley also explain *recency bias* on citation as the tendency of authors for citing recent papers; even highly cited papers stop receiving citations over time [14].

Also, people tend to work or continue working with those persons with who we collaborate with more frequently (recent contacts). It is more likely that *close friends* involve in collaboration, than weak ties (acquaintances) [11].

Katz says that scientific collaboration is the result of a process starting informally through casual conversations, and the communication leads to successively greater commitments of cooperation [15]. Geographical, social and political factors seem to lead to more collaboration given the high probability to start communication.

2.1.5 Similar and reachable collaborators

According to *homophyly* theories [8, 20], similarity in one or more attributes among researchers makes them suitable for collaboration. On this basis, McDonald studied an expert recommender system for identifying suitable collaboration candidates [17]. Evaluated users shown a "more is better" attitude, they preferred more candidates instead of exactness: users preferred reachability over expertise.

The *Triangle-Closing Model* states that new nodes in a social network have a tendency to complete triangles (cliques of 3); this is, they connect to a node and then to some of its neighbors [6]. Kas, Carley & Carley explained that this concept, also known as *transitivity*, is a frequent phenomenon of social networks [14]. In co-authorship, the triangular closure implies that authors develop relationships with existing co-authors of their immediate contacts or closest inner circle.

2.2 Co-authorship Networks

Co-authorship Networks are considered affiliation networks, where actors are linked by their common contribution as authors of a paper. Figure 1 (a) shows an authorship network where authors are represented by ovals (A1, A2, A3), papers are depicted by boxes (P1, P2, P3), and directed edges denote authorship. In co-authorship relations, the number of papers on which every couple authors participate can be calculated. The corresponding coauthorship network is shown in Figure 1 (b), where undirected arcs denote collaboration relationships and labels show the list and count of co-authored papers. Moreover, nodes can be annotated with author attributes obtained from other sources or from the network itself (c.f. [21]).

2.3 SNA Tools for Visualization

SNA software facilitates the analysis of social networks by providing deep analysis metrics of the network and the position that occupy their participants, as well as detection

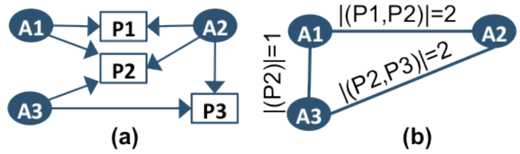

Figure 1: Authorship (a) and Coauthorship (b) Networks.

of communities. UCINET, Pajek and Gephi[1] are three SNA tools that provide facilities for analysis and visualization.

Among the metrics that can be calculated by these tools we have: *degree*, which is a centrality measure that indicates the number of ties (edges) to other actors in the network, *weighted degree*, which sums the weight of each edge outgoing from a node, and the *Erdös number*, which measures the shortest path from a node to a reference node.

3. RECOMMENDING SCIENTIFIC COLLABORATION

In order to suggest scientific collaboration we identify seven best-practices in literature that improve researcher's productivity, we encoded them in a potential collaboration index that is calculated for each researcher, and a network diagram is shown to facilitate the evaluation of the recommendations.

3.1 Scientific collaboration best-practices

From studies described above, we extract the following scientific collaboration best-practices:

Join to groups. Collaboration improves scientific productivity; intellectual isolation is not alright [19, 22].

Expand your collaboration network. Individual's social network strongly influences information seeking and collaboration behavior [17, 19].

Look for popularity & prestige. In *preferential attachment* popularity is given by number of coauthors, it can benefit the expansion of newcomer's network. On the other hand, a high citation index indicate greater visibility: reputation is contagious [14, 21].

Keep in touch. Keep working with previous collaborators as often as possible. Recent papers tend to be cited more than highly cited papers from the past [14, 11].

Rank your choices. Expert researchers become unreachable over time, but a second-best option is still a good choice to collaborate with [17].

Close the triangles. Researchers usually connect to collaborator's collaborators, completing triangles in their relationships [6, 14].

Choose close collaborators. Within a country's scientific community, number of collaborations varies as a function of the geographical distance between partners [15].

3.2 An Intra-Institutional Potential Collaboration Index

We introduce a *Potential Collaboration Index* (PCI) as a first approach for summarizing these collaboration practices

[1]Gephi: Open Graph Viz Platform. `http://gephi.org/`

as follows:

$$PCI = (0.2 \cdot W_C + 0.2 \cdot W_P + 0.2 \cdot W_H + 0.2 \cdot W_K + 0.2 \cdot W_O) \cdot W_T$$

where W_C is a *popularity factor*, W_P is a *productivity factor*, W_H is a *prestige factor*, W_K is a *keyword-matching factor*, W_O is an *organizational distance factor*, and W_T is a *topic similarity factor*.

This proposal allows to assign different percentages to each factor, for a first evaluation we started with a balanced combination of the first five factors (20% each one), whereas the topic similarity factor (W_T) affects the overall score as long as it is more likely to collaborate with people from the same field. All factors are normalized in the range 0 to 1, hence PCI too, from 0 indicating null to 1 indicating maximum compatibility. These factors are depicted in Figure 2.

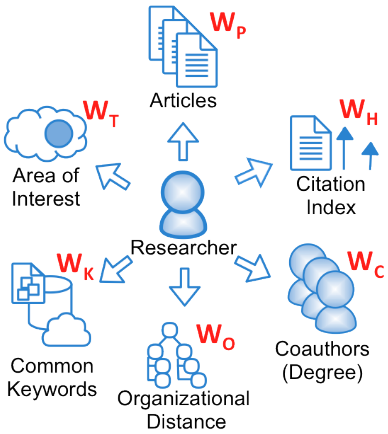

Figure 2: Recommendation criteria for intra-institutional collaboration.

The PCI indicates how likely or desirable is for an author A_i collaborating with another author A_j considering a set of potential collaborators A and their co-authorship network $A \times A$. W_C is given by the number of coauthors (Degree) of A_j normalized by the maximum degree in A. W_P is given by the number of accumulated publications of A_j divided by the maximum number of publications authored by some member of A. W_H is the H-index [13] of A_j divided by the maximum H-index in A. W_O is given by the geodesic distance between A_i's and A_j's respective departments with respect to the organizational hierarchy, which is represented by a tree having campuses and schools on the upper levels and departments and centers in the lower ones, it also reflects the geographic distance between different campuses.

Topic similarity is calculated through two factors: common keywords (W_K) and common fields (W_T). W_K is given by the number of common keywords referred by A_i and A_j in their respective publications, and divided by the minimal number of keywords identified for A_i or A_j. On the other hand, W_T is given by the subject area and the subject category of the journals where A_i and A_j have publications[2]. If both authors share at least one subject category then $W_T = 1$, if they share an area but not a category then $W_T = 0.5$, and if they do not share any area or category then then $W_T = 0$. Note that a subject area contains multiple subject categories.

3.3 A Scientific Collaboration Diagram

Beyond a list or table with the ranked list of potential collaborators, our recommendation is supported by a collaboration diagram that highlights the most compatible researchers in the institutional coauthorship network. In this diagram, nodes represent authors having written an article, a book or a book chapter, and edge's weight reflects the collaboration strength (i.e. the number of papers written in coauthorship).

Besides, the diagram is customized for the researcher receiving the suggestions (denoted as *pivot*). The pivot researcher is colored in red, his coauthors in blue, authors with highest PCIs in green and the rest of authors in gray. Node's size reflects recommendation, i.e. the bigger the node, the greater the PCI of that author with respect to the pivot. Edge's color depends on the nodes' colors it links.

Figure 3 is an example of this collaboration diagram. It displays 400 authors arranged with the Fruchterman Reingold force-directed layout algorithm[10], where the 18 researchers with the highest PCIs for the pivot researcher are highlighted. This image was generated using the open-source tool Gephi[4], which facilitates applying the filters and formats described above. In this image, the pivot researcher is labeled as Pivot, the top 5 recommendations are labeled as Recom_#, recommendations between positions 6 and 18 are labeled as R_# and the rest of the nodes are labeled as Author. In a real example the abbreviated name of each author is used as label.

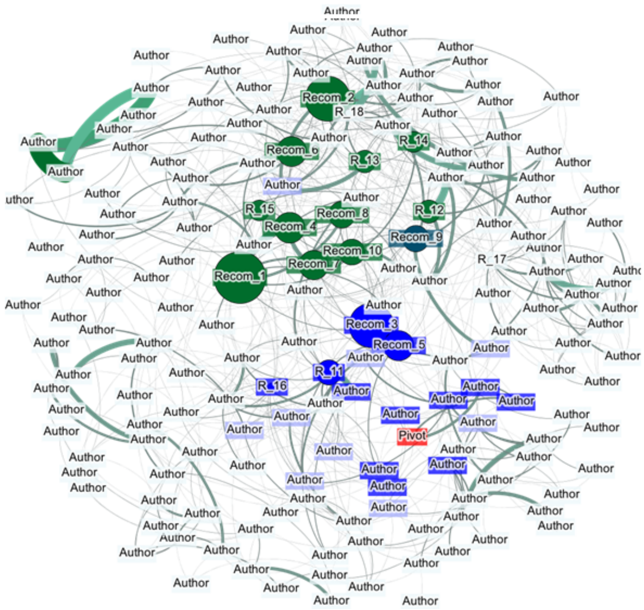

Figure 3: An example of the Scientific Collaboration Diagram.

4. CASE STUDY

For validating the accuracy of the Potential Collaboration Index and the usefulness of the collaboration diagram

[2]We considered subject areas and subject categories from the Scimago Journal Ranking (`http://www.scimagojr.com/journalrank.php`).

we performed a case study among researchers of a multi-campus comprehensive university. Using bibliographic data collected by the university and with information from other data sources, a sample of 926 authors was built for identifying potential collaborations.

4.1 Collaboration and expertise data sources

The information used for identifying professors publishing papers and for building the coauthorship network was obtained from our university's research corporate memory. This database contains information of professors, students, external authors, research groups, thesis and publications. The construction of the coauthorship network from this database was possible due to the normalization of person names (supported by unique IDs) in its bibliographic records.

The coauthorship network built for the case study was constituted by professors publishing articles (in journal or conference), books and book chapters, during the period 2003–2012. In order to capture high-value internal collaboration, other publications and activities such as thesis advisory were discarded from the sample, as well as the participation of students and external authors. Nodes (authors) were annotated with the total number of publications in the period and their degree centrality in the network.

Additionally we obtained information from other data sources. Keywords of publications authored by our professors were obtained from the Thomson Reuters InCites database[3], which comprises information from ISI Web of Knowledge. Subject areas and categories, as well as the H-index of our professors was obtained from a study that SCImago Lab[4] elaborated with information obtained from the Scopus database[5].

Additionally we manually enriched participant's keywords of interest with synonyms and translations between English and Spanish for ameliorating semantic issues caused by exact keyword matching.

4.2 Sample composition

Approximately 40% of the 926 authors considered in the sample are located in our Campus, but the rest of them are distributed in 8 different campuses across the country. Being a comprehensive university, our professors' publications are distributed in 24 of the 27 subject areas identified by the Scimago Journal Ranking.

The average degree in the coauthorship network was 3.85 authors and the average weighted degree was 14.45 collaborations. The top 10 degrees ranged between 32 and 20; remember that this degree only considers coauthorship with other professors. The average number of publications of all types was 37.8 and the average years of publishing experience was 11 years.

4.3 Evaluation methodology

115 professors from the Engineering School of our Campus were invited to participate in the study, but only 27 of them provided keywords and subjects of interest and were available for interviews. The PCI of the 926 researchers in the sample was calculated with respect to each participant of the study. The respective collaboration diagram was generated in Gephi by integrating author's PCI along with the other aforementioned author attributes, and following the specification given in section 3.3 (see Figure 3).

At the beginning of the interview, the PCI and the diagram composition were explained to the participant. Then, the participant was asked to evaluate the top 18 potential collaborators ranked according to the PCI calculated with his research interest and other attributes. Note that top recommendations are colored in green and their node size is proportional to their PCI in the diagram. The participant qualified each recommendation with one of values shown in Table 1.

Value	Description
+2	Known person, valuable unexpected recommendation.
+1	Known person, expected recommendation.
0	Unknown person, I need more information.
-1	Known person, but not similar nor compatible.

Table 1: Recommendation qualifications.

During the evaluation, the participant could request additional information of the recommended author and it is only provided the information considered in the composition of the PCI. At the end of the interview the user was asked to list those persons he was expecting for, whether they were in the top 18 list or not, i.e. people he knows and is willing to collaborate with. Finally the user was asked for the usefulness of the tool in a scale 1 to 5, where 1 represents Very poor and 5 represents Very High. Some participants also provided feedback and suggestions for improving both the PCI and the diagram.

5. RESULTS

Recommendation's qualification made by study participants were analyzed statistically in order to measure the PCI accuracy. On the other hand, comments and suggestion about the collaboration diagram were synthesized.

5.1 Collaboration index accuracy

Figure 4 shows the proportion of qualifications given by 27 users for the top 18 recommendations according to the PCI. As can be seen, 51.8% of the suggestions were considered good (+1 and +2). Only 18.3% of the top 18 recommendations were considered bad and 29.8% were unknown people.

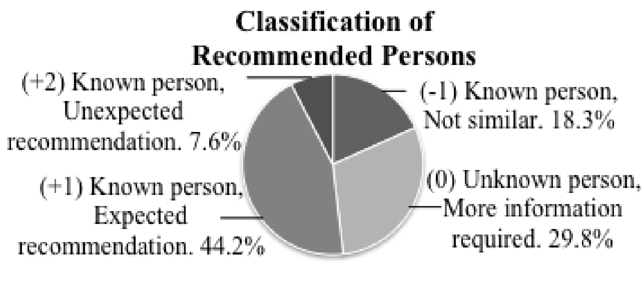

Figure 4: User consideration for Top 18 suggestions

For determining if the PCI provides good recommendations we calculated precision and recall of the top 3, 10 and

[3]Thomson Rtrs InCites ®. http://incites.isiknowledge.com/
[4]Scimago Lab. http://www.scimagolab.com/
[5]Elsevier. SciVerse Scopus ®. http://www.scopus.com/

18 recommendations. Results are shown in Table 2. It is noteworthy that this measures the perception of the usefulness of the tool as a function of the knowledge that the user has about the network.

Selection	Avg. Relevant Persons (+2)	Avg. Listed Expected Persons	(a) Avg. Precision	(b) Avg. Recall	F-measure on (a) and (b)
Top 3	20.0%	13.0%	0.538	0.219	0.311
Top 10	18.9%	18.9%	0.404	0.459	0.430
Top 18	7.6%	32.5%	0.336	0.665	0.446

Table 2: Precision, Recall and F-measure.

As can be seen in Table 2, the best precision is for the top 3 (0.538), i.e. in average 1.5 of the persons listed in the three first positions were considered good recommendations. On the other hand, when the user evaluates the first 18 recommendations he finds 66% of the people he was expecting. The F-measure, which mediates precision and recall, is good for the first 10 recommendations and it improves only slightly for the top 18. This result indicates that recommending only the top 10 would be enough.

5.2 Collaboration Diagram Assessment

Regarding the usefulness of the diagram, 19 participants of the study (70%) considered it high or very high, 5 participants (19%) consider it regular, and 3 participants (11%) consider it poor or very poor.

The main suggestions from the participants were: 1) including other collaborations such as joint participation in thesis committees and research projects; 2) including external authors and PhD students; and 3) letting users to set up their *own potential collaborator profile*, since eventually researchers are not looking for similarity with all of their own keywords, or all of their subject areas, but are looking for collaboration in some specific area of application or a new particular keyword of interest.

6. DISCUSSION AND RELATED WORK

Next we compare our approach with other scientific productivity and collaboration indexes, and discuss the advantages of using a network visualization for supporting the recommendation.

6.1 Productivity and Collaboration indices

There exist some indices that measure the productivity and the collaboration of scientist based on bibliographic data. Productivity that originally was measured in terms of publications is now measured as a function of how many citations these works receive (their impact). And by acknowledging the leverage that collaboration provides to the scientific labor, this has recently gained more attention.

Indices like the H-index [13] and the RP-index [3] that measures the impact and productivity of an author are designed for measuring the individual performance of a scientist based on his actual production and citation. Both indices mediates between the number of publications and the number of cites received by each one. However they do not consider the collaboration's effect on his performance, and in the case of the RP-Index it can discourage collaboration by dividing the credit of a publication between the number of coauthors.

The RC-Index [1] captures the collaboration of a researcher within a coauthorship network and qualifies it using the RP-Index. The RC-index indicates how collaborative is an author and how productive is a collaboration with him. This index basically considers the degree and the weighted degree of an author, and it has been used for comparing researchers within of research communities focused in a single subject area.

Like the RC-index, the PCI considers the collaboration of an author within his institution but it uses the widespread and better known H-index for measuring his impact or productivity. Additionally the PCI considers topic similarity for privileging collaborations with similar partners as suggested by homophyly theories.

We recognize that despite the PCI already considers degree centrality, it does not consider betweenness centrality which positive influence has just been demonstrated on citation count and on the strength of scientific collaboration [21].

6.2 Advantages of Network Visualization

Whereas most of the scientific collaboration best-practices proposed in section 3.1 are summarized in the PCI, some of them must be addressed by other means. We found that our network visualization facilitates both, presenting the recommendation and considering additional factors such as: the distance to potential collaborators, previous collaborators, and groups with intense collaboration.

Our collaboration diagram incorporates advices of McDonald's recommender system [17] by showing not only to the top 3, 10 or N best matches, but providing a wide network where the most compatible partners are highlighted but other options are also shown. And, despite the fact we displayed 400 potential collaborators in a single diagram, it was legible and manageable as validated in our case study.

Other benefits of the collaboration diagram is that it creates a quick reaction on the user and provides awareness of the own position in the scientific context of the University. Awareness relevancy is pointed out by structural holes theory, which states that certain actors can identify opportunities for cooperating and they mostly turn to be good for exploiting their brokerage position in the network [7].

7. CONCLUSIONS

We presented a mixed approach for recommending scientific collaboration, constituted by a Potential Collaboration Index (PCI) and a collaboration diagram that supports and visually enriches the recommendation. Both tools are grounded on Scientific collaboration best-practices extracted from Social Network Analysis studies performed on coauthorship networks and associated to scientific proficiency.

The Potential Collaboration Index (PCI) and the collaboration diagram proposed in this work showed a good acceptance in our case study. 70% of the participants considered it useful and only 11% considered that it does not add value to the actual user's knowledge. 81.6% of the top 18 recommendations were people already known by the participants and 63.5% of them were considered good choices. The other 18.3% of the recommended people that was not known by the participants can be explained by inaccuracy in the PCI, by the distribution of researchers across multiple campuses and disciplines, or by the lack of individual knowledge about other similar researchers.

We also provided a methodology for measuring the effectiveness of a collaboration recommendation metric. In our case study, participant's feedback indicated that the top 10 recommendations based on the proposed PCI provides a good balance between precision and recall (Table 2).

7.1 Future Work

The dynamic observed during interviews on which participants requested additional information from recommended authors, as well as the observation of an 18% of unknown persons, indicates that an interactive mechanism for reading both the network and author profiles would be quite useful. For this reason we recommend using a social network visualization system like Vizster, which provides this functionality and additionally allows filtering nodes and making textual search on author attributes [12].

Further analysis must be done for improving the accuracy of the proposed PCI. On one hand, considering other collaboration relationships will provide more potential collaborators. On the other, PCI weights can be adjusted according to author characteristics like: maturity (years of research experience), changes on their current research interests, and participation on interdisciplinary research (which evenly distributes their work among different subject areas). Topic similarity is an important issue that must be also addressed.

8. ACKNOWLEDGEMENTS

We would like to thank to the Tecnologico de Monterrey for funding this research through the research chair PRY058, Semantic Technology for Intelligent Organizations.

9. REFERENCES

[1] A. Abbasi, J. Altmann, and J. Hwang. Evaluating scholars based on their academic collaboration activities: two indices, the rc-index and the cc-index, for quantifying collaboration activities of researchers and scientific communities. *Scientometrics*, 83(1):1–13, 2010.

[2] A. Abbasi, J. Altmannand, and L. Hossain. Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures. *Journal of Informetrics*, 5(4):594–607, 2011.

[3] J. Altmann, A. Abbasi, and J. Hwang. Evaluating the productivity of researchers and their communities: the rp-index and the cp-index. *International Journal of Computer Science and Applications*, 6(2):104–118, 2009.

[4] M. Bastian, S. Heymann, and M. Jacomy. Gephi: An open source software for exploring and manipulating networks. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 2009.

[5] D. D. Beaver and R. Rosen. Studies in scientific collaboration Part III. Professionalization and the natural history of modern scientific co-authorship. *Scientometrics*, 1(3):231–245, 1979.

[6] F. Bonchi, C. Castillo, A. Gionis, and A. Jaimes. Social Network Analysis and Mining for Business Applications. *ACM Transactions on Intelligent Systems and Technology*, 2(3):1–37, 2011.

[7] R. S. Burt. *Structural Holes.* Harvard University Press, Cambridge, Ma., 1992.

[8] D. Byrne. *The attraction paradigm.* Academic Press, 1971.

[9] Y. Ding. Scientific collaboration and endorsement: Network analysis of coauthorship and citation networks. *Journal of informetrics*, 5:187–203, 2012.

[10] T. M. J. Fruchterman and E. M. Reingold. Graph drawing by force-directed placement. *Software: Practice and Experience*, 21(11):1129–1164, 1991.

[11] M. Granovetter. The strength of weak ties: A network theory revisited. *Sociological Theory*, 1:pp. 201–233, 1983.

[12] J. Heer and D. Boyd. Vizster: Visualizing online social networks. In *IEEE Information Visualization (InfoVis)*, pages 32–39, 2005.

[13] J. Hirsch. An index to quantify an individual's scientific research output. *Proc Natl Acad Sci U S A.*, 102(46):16569–16572, 2005.

[14] M. Kas, K. M. Carley, and L. R. Carley. Trends in science networks: understanding structures and statistics of scientific networks. *Social Network Analysis and Mining*, 2012.

[15] J. Katz. Geographical proximity and scientific collaboration. *Scientometrics*, 31(1):31–43, 1994.

[16] J. S. Katz and B. R. Martin. What is research collaboration? *Research policy*, 26:1–18, 1997.

[17] D. W. McDonald. Recommending collaboration with social networks. In *Proceedings of the conference on Human factors in computing systems - CHI '03*, page 593, New York, New York, USA, Apr. 2003. ACM Press.

[18] J. Moody. The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999. *American Sociological Review*, 69(2):213–238, 2004.

[19] M. Newman. Who is the best connected scientist? A study of scientific coauthorship networks. In E. Ben-Naim, H. Frauenfelder, and Z. Toroczkai, editors, *Complex Networks*, volume 650 of *LN in Physics*, pages 337–370. Springer Berlin / Heidelberg, 2004.

[20] J. Turner. *Rediscovering the social group: a self-categorization theory.* Basil Blackwell, Oxford, U.K, 1987.

[21] S. Uddin, L. Hossain, and K. Rasmussen. Network Effects on Scientific Collaborations. *PLoS ONE*, 8(2), 2013.

[22] L. Waltman, C. Calero-Medina, J. Kosten, E. C. Noyons, R. J. Tijssen, N. J. van Eck, T. N. van Leeuwen, A. F. van Raan, M. S. Visser, and P. Wouters. The leiden ranking 2011/2012: data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology*, 63(12):2419–2432, 2012.