Exploring Gender Role in Co-Authorship Networks for Computing Books: A Case Study in DBLP

Kosovare Sahatqija
South East European University, Faculty of Contemporary Sciences and Technologies, Tetovo, Macedonia
Arbana Kadriu
South East European University, Faculty of Contemporary Sciences and Technologies, Tetovo, Macedonia

Abstract

Social network analysis and mining intend to explore for certain, previously unknown, and probably useful relational information from social and information networks. In our case, the research paper is about identifying collaborative networks between the authors (co-authors) of Computer Science books with the highlighted focus on the women computer scientist's community. Often the hardest part of collaborating is knowing whom you should be collaborating with. Hence, this study will tackle this issue and will identify, and present a visualization of the co-authors which have already collaborated and how often they have collaborated. In this way, we are going to distinguish the successful collaboration between co-authors, the trend of further collaboration between them and the participation of women on these collaborations. This paper is research which is based on detailed and intensive analysis of the different ways of identifying these kinds of connections through secondary material.

Keywords: Gender role, co-authorship, DBLP, CS community
JEL classification: C00, C30

Introduction

Nowadays, is the area of data and in particular data mining. So, mining data using different datasets can result in very valuable findings. The techniques which are currently used can provide insights on many unexpected results and interesting at the same time. One of the data mining techniques which can serve us to have insights into the relations between two different entities (nodes, individuals, organizations, etc.) is Social Network Analysis.

The aiming of social network analysis and mining them is to discover some kind of implicit information which is unknown and can be valuable. Discovering these kinds of relationships can lead us to other conclusions which will allow us to see which author is usually collaborating with whom and then with further investigation we can understand their interests and specific field of their collaboration.

In this paper, we are interested in a special kind of social networks – coauthorship networks. Presenting the benefits of using social network analysis and the metrics to provide useful data which can be concluded after the calculations and visualizations of them and exploring gender role in co-authorship networks for computing books. For this purpose, we first crawl DBLP database to find the information we need and after that, we use social network analysis methods to explore the gotten dataset.

While the key purpose of this research is to explore the gender role in co-authorship for computing books many authors have published studies and presented why women scientists tend to publish fewer publications than their man colleagues. To achieve the aim of this research we have used DBLP Computer Science Bibliography and we will list some of the reasons which are considered a fundamental circumstance to these results.

The paper is organized as follows: after this section, respectively in the second section is the part of the related work of other authors regarding this topic and not only. Then a deeper explanation for the network analysis metrics which are going to be included in our calculations, the added value part is data analysis and interpreting their results. Finally, the conclusion of this paper research and future plans.

Related Work

The field of social network analysis has grabbed a lot of interest from different authors and it seems to play a crucial role in analysing and at the same time forecasting future relations. There are several authors who were interested to do research papers about this topic, at analysing books collaboration networks.

In one of the researches Sun et al. (2011) proposed an approach in order to predict future collaboration relationships, but this regarding heterogenous Networks. Their PathPredict model is based on the topological features in such networks and then a regression-based co-authorship prediction model to have a significant performance. A related approach used by Yan et al. (2010) is the weighted PageRank algorithm which is considered to provide reliable results in measuring author impact in collaborations.

Another research similarly as this one is made by Zhou et al. (2007). They propose a framework for co-ranking entities of different kinds in a heterogeneous network connecting researchers and publications they produce. Since the co-authorship domain got very huge interest, also Liu et al. (2005) have investigated the co-authorship network of the Digital Libraries of the research community in the conferences. They pointed out that the densest shapes (collaborations) include authors from the same institution or working on the same project. Moreover, Rodriguez et al. (2008) on their research claim based on the results that groupings of individuals characterized by similar scholarly expertise are more likely to have co-authorship.

Lu et al. (2009) have made a measure of authors' centrality in co-authorship networks. The experiment that they implemented showed that the extensity centrality and betweenness have a relatively high correlation and the correlation between extensity centrality and degree is low.

An advantage which we can get it as a result of examining these kinds of relations is finding the appropriate team members and their similar collaborative abilities. Cheatham et al. (2006) applied social network analysis to find out this issue. They collected the data of the people in their company and their collaborative relations. And using those relations they found out the way of creating teams which will have a helpful collaboration. On the other hand, Huang et al. (2006) visualized the co-authorship network via an internal tool that they developed which showed the weight of the co-authorship.

When considering the gender balance in these communities we tend to have a very disappointing result according to the authors who have previously made research on this. The reasons can vary to different ones but the main ones are like women scientists may have the struggle to possess equal profit, to stay updated on the market needs and furthermore to have a work-life balance, due to child-care responsibilities as it was highlighted in Baker et al. (2012) and Fox et al. (2005) journal.

Network Analysis Metrics

In our research, we have applied network analysis metrics to determine the relative position of individuals and clusters within a network (Hansen et al., 2011, Arif et al., 2012). We have applied five (5) metrics: degree centrality, closeness centrality, betweenness centrality, PageRank and eigenvector centrality. The explanation for these metrics is given below.

PageRank is a metrics which indicates or outputs the probability distribution that a random walk across hyperlinks will send us to a particular node. So, the higher value of PageRank is, the highest is the probability easily finding that node, regardless of the way of the walk.

Degree centrality is defined as the number of links occurrence upon a node. There are considered the number of unique edges that are connected to the node (Kadriu et al., 2013). The degree can be interpreted in terms of the immediate risk of a node for catching whatever is flowing through the network.

Between centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. It was introduced as a measure for quantifying the control of a human on the communication between other humans in a social network by Linton et al. (1977). In his conception, vertices that have a high probability to occur on a randomly chosen shortest path between two randomly chosen vertices have a high betweenness. The betweenness centrality of a node **v** is given by the expression:

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{1}$$

where **σ**st is the total number of shortest paths from node **s** to node **t** and **σst(v)** is the number of those paths that pass-through **v**.

Closeness centrality is the average length of the shortest path between the node and all other nodes in the graph. Thus, the more central a node is, the closer it is to all other nodes.

$$C(x) = \frac{1}{\sum_y d(y,x)} \tag{2}$$

*where **d(y,x)** is the distance between vertices **x** and **y***

Eigenvector centrality is a measure of the influence of a node in a network. It assigns relative scores to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. It is a very sophisticated view of centrality (Hansen et al., 2011). I.e. a person with few connections could have a very high eigenvector centrality if those few connections were themselves very well connected. It is a variant of the PageRank Google algorithm.

In the next section, we are going to explain the results on degree centrality, closeness centrality, betweenness centrality, PageRank and eigenvector centrality.

Data Analysis and Results

There are various online sources for collecting public data which we can use and further analyze them. In our case, we have used the DBLP computer science bibliography source to get the data on the Computer Sciences' books co-authorship.

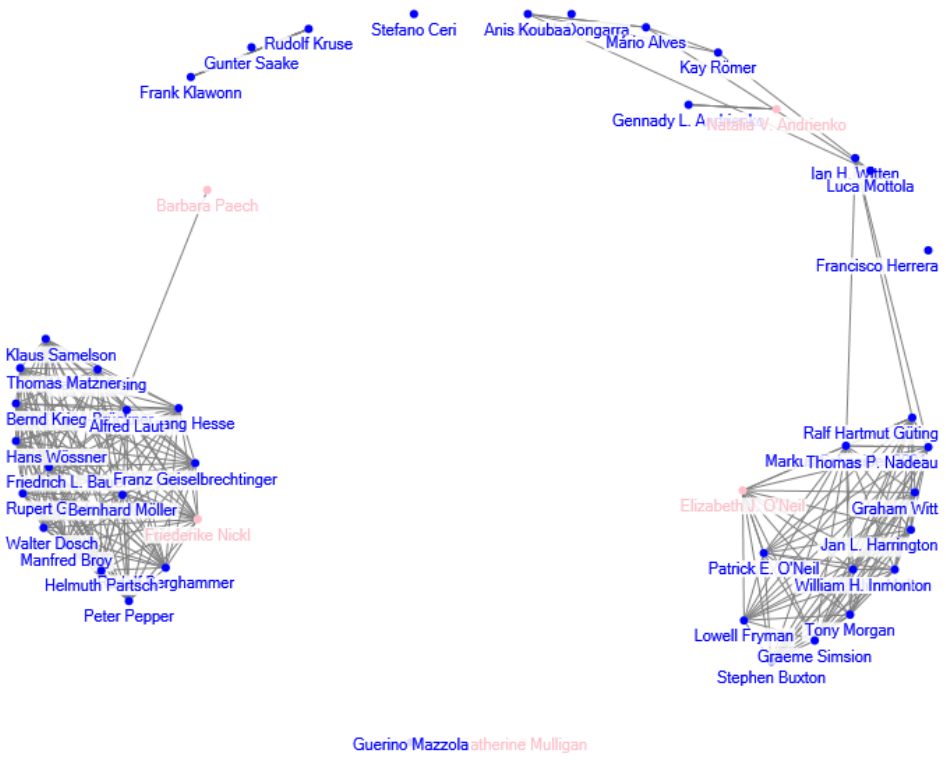
The dataset used to proceed with further analysis consists of 9510 co-authorship records or as we usually refer to them as edges. The application that we used to generate these analyses is the NodeXL template, an open-source Microsoft Excel Template.

After calculating the Overall Metrics of the dataset, we found out that besides the total number of edges (9510) we got to know more about our dataset. The number of authors/members which are part of this dataset is 7786. After merging the duplicate edges of collaboration, the number of the edges is reduced to 8331, which means that the number of the collaboration which was more repetitive is around 1200.

The authors are divided into two groups according to their gender. We can see there is a huge difference between these groups, where we consist of the members as:

- 1008 female (pink nodes)
- 6778 male (blue nodes)

Figure 1
The Nodes Regarding Gender (Female = Pink, Male = Blue)



Source: Authors' work

Since there is a lack of women on this community, there is also lack of collaboration between them (women) on the field of CS (noted in *Figure 1*).

Table 1 shows the top 10 authors for each measure performed respectively. They are listed from the highest value to the lowest. Below are described the findings related to the metrics calculated. From these results, we can also observe a huge gap of

women in this community of collaboration. From fifty (50) cells of this table just four (4) of them are women.

Table 1

Top 10 Authors in Five Different Social Network Analysis Metrics

PageRank	Degree	Between	Closeness	Eigenvector
Rolf Drechsler	Wolfgang Hesse	Manfred Broy	Lois Wakeman	Wolfgang Hesse
Gottfried Vossen	Anis Koubaa	Wolfgang Hesse	Angela Espinosa	Manfred Broy
Stefano Ceri Francisco Herrera	Manfred Broy	Barbara Paech	Nicola Bellomo	Friedrich L. Bauer
Horst Bunke	Ralf Hartmut Güting	Bernhard Rumpe	Jin Keun Seo	Bernhard Möller
Gunter Saake	Markus Schneider	Ian H. Witten	Daniela Calvetti	Helmuth Partsch
Abraham Kandel	Mário Alves	Klaus Voss	Eli Biham	Bernd Krieg-Brückner
Erik Cuevas	Bernd Krieg-Brückner	Anis Koubaa	Mansooreh Mollaghasemi	Martin Wirsing
Hartmut Ehrig	Martin Wirsing	Stefano Ceri	Rita Lewis	Friederike Nickl
Christoph Meinel	Ian H. Witten	Martin Wirsing	Karen L. McGraw	Rudolf Berghammer

Note: From the results only four(4) of them are women.

Source: Authors' work

The author with the highest value of PageRank is Rolf Drechsler with a value of 4.09. From this, we can conclude that Drechsler has the highest probability of easily finding him in collaborations.

From the results given in the table below we can conclude the Wolfgang Hesse is the author with the highest value of the degree, and we can say that Hesse has the highest influence in this co-authorship dataset.

Manfred Broy is the author with the highest value when it comes to acting as a bridge along the shortest path between two other nodes. He controls the communication between different nodes.

Regarding closeness centrality, Lois Wakeman scores are the highest. So, we can easily conclude that Wakeman is the node which is closer to other nodes in this co-authorship relationship.

The last metrics that we have calculated is about the eigenvector centrality. Even though in the literature we had some similarities of the eigenvector centrality with PageRank the results do not prove the same thing. The result is more similar to the results of the degree centrality. Where it shows that the author Hesse (node) proves that has the highest influence between all authors.

Some other metrics that we calculated for the overall network are:

- Maximum geodesic distance has its diameter length of 11
- Average geodesic distance is 1.51
- Graph density is 0.0002 (a very dense network)

Conclusion

The main idea of this paper was to present the benefits of data mining using social network analysis approaches to identify the co-authorship of the books related to the Computer Science field. Co-authorship social networks are based on the co-authorship relationship and are a result of the people collaborating to become a co-author. Using the data, the DBLP computer science bibliography we managed to analyze them using network analysis metrics. The focus was on the five metrics which are degree centrality, closeness centrality, betweenness centrality, PageRank and eigenvector centrality.

A very crucial substance that was added to this research was highlighting the gender role in co-authorship networks for computing books. As we noted a huge gap in this relation, we plan to do further research on the co-authorship between women among them and the most interesting field for them in the area of Computer Sciences.

References

1. Arif, T., Ali, R., Asger, M. (2012), "Scientific Co-authorship Social Networks: A case study of Computer Science Scenario in India", International Journal of Computer Applications, Vol. 52, No. 12, pp. 38-45.
2. Baker, M. (2012), Academic Careers and the Gender Gap, UBC Press, Vancuver, Toronto.
3. Cheatham, M., Cleereman, K. (2006), "Application of Social Network Analysis to Collaborative Team Formation", in the Proceedings of the International Symposium on Collaborative Technologies and Systems (CTS), Las Vegas, NV, USA, IEEE, pp. 306-311.
4. Fox, M. F. (2005), "Gender, family characteristics, and publication productivity among scientists", Social Studies of Science, Vol. 35, No. 1, pp. 131-150.
5. Hansen, D., Shneiderman, B., Smith, M. A. (2011), Analysing Social Media Networks with NodeXL: Insights from a connected world, Morgan Kaufmann.
6. Huang, T. H., Huang, M. L. (2006), "Analysis and Visualization of Co-authorship Networks for Understanding Academic Collaboration and Knowledge Domain of Individual Researchers", in the Proceedings of the International Conference on Computer Graphics, Imaging and Visualization, Sydney, Qld., Australia, IEEE, pp. 18-23.
7. Kadriu, A. (2013), "Discovering Value in Academic Social Networks: A Case Study in ResearchGate", in the Proceedings of the 35th International Conference on Information Technology Interfaces, Cavtat, Croatia, IEEE, pp. 57-62.
8. Linton, F. (1977), "A set of measures of centrality based upon betweenness", Sociometry, Vol. 40, No. 1, pp. 35-41.
9. Liu, X., Bollen, J., Nelson, M. L., Sompel, H. V. (2005), "Co-authorship networks in the digital library research community", Information Processing and Management, Vol. 41, No. 6, pp. 1462-1480.
10. Lu, H., Feng, Y. (2009), "A measure of authors' centrality in co-authorship networks based on the distribution of collaborative relationship", Scientometrics, Vol. 81, No. 2.
11. Rodriguez, M. A., Pepe, A. (2008), "On the relationship between the structural and socioacademic communities of a coauthorship network", Journal of Informetrics, Vol. 2, No. 3, pp. 195-201.
12. Sun, Y., Barber, R., Aggarwal, C. C., Han, J. (2011), "Co-author Relationship Prediction in Heterogeneous Bibliographic Networks", in the Proceedings of the International Conference on Advances in Social Networks Analysis and Mining, Kaohsiung, Taiwan, IEEE, pp. 121-128.
13. Yan, E., Ding, Y. (2010), "Discovering author impact: A PageRank perspective", Information Processing and Management, Vol. 47, No. 1, pp. 125-134.
14. Zhou, D., Orshanskiy, S. A., Zha, H., Giles, C. L. (2007), "Co-Ranking Authors and Documents in a Heterogeneous Network", in the Proceedings of the Seventh International Conference on Data Mining (ICDM), Omaha, NE, USA, IEEE, pp. 739-744.

About the authors

Kosovare Sahatqija, currently pursuing a Masters' degree in Data Engineering at the South East European University (SEEU). Prior to this, she received a Bachelor Degree in Computer Science at the Faculty of Contemporary Sciences and Technologies (SEEU). Kosovare has worked on numerous projects in software quality assurance, as well as analysis and project design, covering the entire life cycle from defining the project scope to feasibility assessment. While being an active member of different communities which encourage women in ICT fields, she contributed to data collection and analysis for numerous reports. The author can be contacted at **ks16575@seeu.edu.mk.**

Arbana Kadriu holds a Ph.D. degree in Computer Sciences from Ss. Cyril and Methodius University in Skopje from 2008, with a focus on natural language processing and information retrieval. She is an Associate Professor at Faculty of Contemporary Sciences and Technologies at SEE University in Macedonia. She has also background in artificial intelligence, machine learning, programming paradigms, software engineering, e-learning and social network analysis. She is author of more than 40 research papers. The author can be contacted at **a.kadriu@seeu.edu.mk.**