doi:10.2498/iti.2013.0566

Discovering Value in Academic Social Networks: A Case Study in ResearchGate

Arbana Kadriu
South East European University
1200 Tetove, Macedonia
E-mail: a.kadriu@seeu.edu.mk

Abstract. *The research presented in this paper is about detecting collaborative networks inside the structure of a research social network. As case study we consider ResearchGate and SEE University academic staff. First we describe the methodology used to crawl and create an academic-academic network depending from their fields of interest. We then calculate and discuss four social network analysis centrality measures (closeness, betweenness, degree, and PageRank) for entities in this network. In addition to these metrics, we have also investigated grouping of individuals, based on automatic clustering depending from their reciprocal relationships.*

Keywords. Research social networks, SNA metrics, collaborative networks

1. Introduction

The main idea behind specialized networks created for specific category of experts and researchers is to group them around a topic which could be of help to each other. In such a network, professionals can communicate with each other and share their achievements. They can follow different experts in their field of interest and subfields in that area, can upload papers or make/answer questions. This is all about knowledge sharing.

However, there is some skepticism about academics agreeing to share their knowledge. The scientific world, in addition to its human and progressive side, is also competitive and there is a kind of disbelief in the good faith distribution of research articles, projects and ideas.

So, having in mind these limitations and skepticism, can there be any other values discovered in these "limited" academic and research networks? In this paper, we try to find out another dimension to the usefulness of research networks. We investigate the "position" of group of academics in ResearchGate – generally identified amongst the worldwide scientific group as the world's largest professional social network for scientists. We try in a way to estimate the knowledge flow with focus in the South East European University academic staff.

2. Related work

A lot of work has been done at analyzing scientific collaboration networks. From the time of the pioneering article of Price [2] and Beaver & Rosen [3][4][5], a considerable amount of research articles have put emphasis on different ways and functions of scientific collaboration in particular scientific areas.

Using social network analysis, Newman [12] began to construct the actual collaboration networks between scientists in the field of physics, biomedical research, and computer science. However, Newman [13] pointed out that bibliometric analysis examining networks of individuals is not so easy to find.

These investigations about collaboration networks of scientists usually interpret that two scientists are considered connected if they have coauthored a paper.

Many papers have been published regarding collaborative networks in scientific conferences and journals, resulting in a number of significant investigations. According to Yanl&Ding (2011), collaborative network analysis, with its sound theory and methodology derived from physics, mathematics, graph theory, and social sciences, is expected to serve as the complement to traditional citation analysis [8]. This analysis very often involves use of centrality measures like degree, closeness centrality, betweenness, clustering coefficient and PageRank. Rodriguez and Pepe (2008) have studied the relationship between the structural and socio academic communities of a coauthorship network [14]. Mutschke (2003) utilized centrality to the coauthorship network of digital libraries research

[15]. Liu et al. (2005) used centrality measures to analyze the Joint Conference on Digital Libraries research community, comparing three kinds of centrality measures with the ranking of JCDL program committee membership [10]. Estrada and Rodriguez (2005) suggested a new centrality metric that typifies the participation of each node in all subgraphs in a coauthorship network [7]. Chen (2006) used betweenness centrality to detecting and visualizes emerging trends and transient patterns in scientific literature [1]. Yin et al. (2006) applied three centrality measures to COLLNET community coauthorship network [9]. Vidgen, Henneberg, and Naude (2007) applied five centrality measures to rank information system community of the European Conference on Information Systems [16]. Nascimento et al. (2003) studied the SIGMOD's coauthorship graph.

Most of the studied related to the scientific collaboration networks are based on co-authorships of papers in a particular research area. In our research we try to investigate how our university academic staff is allocated in terms of research areas in the ResearchGate structure.

3. Network centrality measures

In this research, we apply four standard centrality measures: degree centrality, closeness centrality, betweenness centrality, and PageRank [6].

For a given graph (network) G with N nodes and E edges, the main centrality measures are defined as follows:

Degree centrality. Degree centrality of a node is the number of unique edges that are connected to it. The equation of it is as following where $d(n)_i$ is the degree of n_i:

$$C_B(n_i) = \sum_{j,k \neq i} \frac{g_{jik}}{g_{jk}} \quad (1)$$

where g_{ijk} is all geodesics linking node j and node k which pass through node i and g_{jk} is the shortest path between the nodes j and k.

Closeness centrality. Closeness centrality is a measure of the average shortest distance from each node of the network to each other node. If we denote the distance between node i and j with as $d(n_i, n_j)$, the closeness equation will be:

$$C_C(n_i) = \sum_{i} \frac{1}{d(n_i, n_j)} \quad (2)$$

Betweeness centrality. Betweeness centrality is a measure of a node's centrality in a network equal to the number of shortest paths from all vertices to all others that pass through that node. Authors with high betweenness are the brokers and connectors who bring others together. Being between means that a vertex has the ability to control the flow of knowledge between most others. The betweenness centrality of a node v is given by the expression:

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{Q_{st}(v)}{Q_{st}} \quad (3)$$

PageRank. PageRank of a page X is the limiting probability that a random walk across hyperlinks will end up at X, as we run the walk for larger and larger numbers of steps [6]. This idea is summarized in the PageRank equation as follows:

$$PR(p) = (1-r)\frac{1}{N} + r\sum_{i=1}^{k} \frac{PR(p_i)}{C(p_i)} \quad (4)$$

where N is the total number of pages, r is a reducing factor, $C(p)$ is the outdegree of page p, and p_i denotes the inlinks of p.

We define also the network metrics:

- **Distance** between a pair of nodes – is the edge length of the shortest path between them
- **Graph's diameter** - is the maximum of the shortest distances between node pairs
- **Graph's density** – measures how many edges are in graph compared to the maximum possible number of edges between vertices in that graph.

3. Methodology for creation of the collaborative network

ResearchGate offers suitable tools made purposely for scientists. In this framework, it is possible to create separate workgroups for particular academic institutions and scientific (sub) fields, too.

We use the first possibility to find our university staff that is already registered at ResearchGate. This step is straightforward, we

have just to select link that shows the members from our institution.

We parse the gained site to make a list of our university members. This is done using regular expressions, with the identified pattern where the member names show:

```
<div class="indent-content">
    <h5><a href="profile/Name_Surname/"> Name
    Surname</a>
```

After that, we want to use the second above mentioned possibility of ResearchGate to find fields of interest for every member separately. We iterate the gained list of academics, and again using regular expressions:

```
<a href="topic/name of the topic/"
     class="keyword-list-token-text">
     Name of the topic</a>
```

we make a hash with the *person's name&surname* as *key* and *list of topics* of interest as *value*.

The gained hash described above is converted to a so-called *bipartite graph*, which according to Newmann (2003) is a graph that contains vertices of two distinct types, with edges running only between unlike types [11]. *Afiliation networks* in which people are joined together by common membership of groups take this form, the two types of vertices representing the people and the groups. In our case, two types of vertices are academics and topics of interest. This kind of networks is called also *two-mode* networks.

We convert then this network to an *academic-academic* network (*one-mode* network), using 2-mode to 1-mode routine. Two academics are linked with an edge if they are interested in the same scientific field. We assign weights to these edges depending of the number of common interests.

With this methodology we gained a collaborative network of SEEU academics, which can be used for further processing.

4. Data analysis

With this procedure we gained that our university group has 42 members, which in total cover 176 different topics of interest. For better analysis, we (manually) divided them in two groups – academic staff from computer science department and the others. There are 17 members from the CS department only, with 64 different topics of interest, and 25 persons from the other five university departments, covering 112 research topics. The 17/25 ratio alone indicates that computer science academics are more willing to be part of research social networks, compared to the other disciplines.

We have 71 edges, where 51 are from the CS department. This is not much surprise, as the second group belongs to other four departments at SEEU and naturally their subjects are from a wider collection.

There are six top topics of interest identified:

- Computer Science
- Computational Intelligence
- Artificial Intelligence
- Economics
- Applied Linguistics
- Educational Research

For example, one conclusion from the above list and combinations therein could be that our university has the prerequisites to launch integrated studies in Computational Linguistics and Business Integellence.

Fig. 1. shows the gained academic-academic collaborative network, where two academics are linked if they have interest in the same research topic. As we previously mentioned, we did a manual split of computer science staff (in red) and the others (in black). The edge width shows how much two academics are near with their topic preferences and interests.

Fig. 2. offers an illustration of automatic clustered dataset using Girvan-Newman algorithm [10].

It can be seen that there are also some isolated nodes – their fields of interest do not overlap with those of their colleges.

We applied centrality measures to rank authors and to understand how our staff "fits in" the examined social network. Table 1 shows top 15 academics for each measure respectively.

In this network, to have high degree means that you are engaged in a lot of different topics which are attractive to most of your colleges. And this makes you more influential in the university structure.

From this, we can conclude that A. Dika has the highest influence in our academic university network.

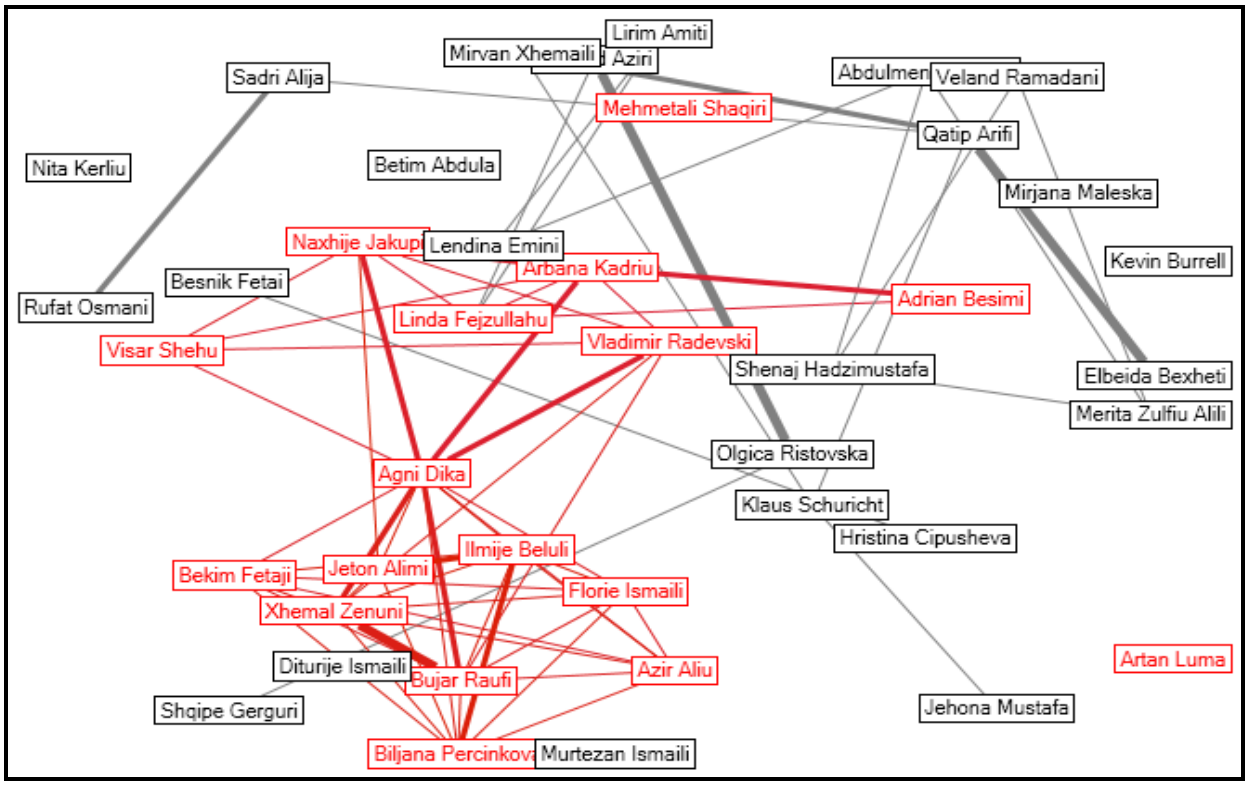

Figure 1. SEEU academic-academic collaborative network

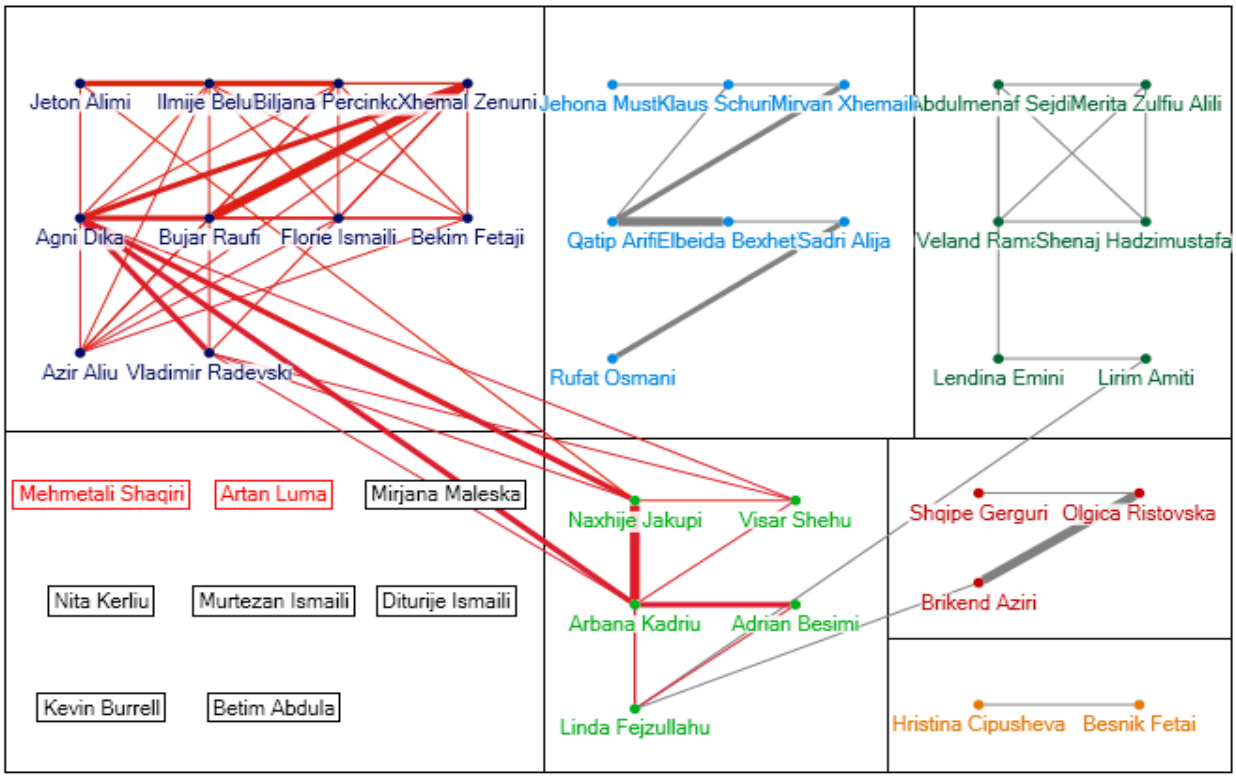

Figure 2. SEEU academic-academic collaborative network

Another conclusion from the first column of the shown table is that the first 14 academics are from the CS department, and only the last one is from other university departments.

L. Fejzullahu has the highest brokerage value, since her interests, except computer science topics, cover also some of economic area. She controls the flow of knowledge between most others.

Regarding closeness centrality, B. Fetai scores as the highest. With his choices, he is the closest to the others.

If we start randomly at any academic asking for a certain scientific issue, Q. Arifi has the highest probability to be our final person after some steps.

Table 1. Top 15 academics for particular SNA metrics

Degree	Betweenn	Closeness	PageRank
A.Dika	L.Fejzullah	B.Fetai	Q.Arifi
Xh.Zenuni	L.Amiti	H.Cipusheva	A.Dika
I.Beluli	A.Dika	Q.Arifi	K.Schuricht
B.Percink	L.Emini	K.Schuricht	L.Fejzullah
B.Raufi	N.Jakupi	S.Alija	A.Sejdini
F.Ismaili	A.Sejdini	M.Xhemaili	Xh.Zenuni
B.Fetaji	A.Kadriu	E.Bexheti	A.Kadriu
A.Aliu	B.Aziri	J.Mustafa	B.Raufi
N.Jakupi	O.Ristovsk	R.Osmani	I.Beluli
V.Radevs	V.Radevsk	N.Jakupi	B.Percinkov
A.Kadriu	Q.Arifi	A.Kadriu	N.Jakupi
J.Alimi	J.Alimi	L.Fejzullahu	Sadri Alija
L.Fejzulla	K.Schurich	A.Dika	F.Ismaili
V.Shehu	S.Alija	V.Radevski	B.Fetaji
A.Sejdini	Xh.Zenuni	J.Alimi	A.Aliu

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

- There are 11 connected components and 8 isolated nodes
- Maximum geodesic distance has length 7
- Average geodesic distance is 3.1
- Graph density is 0.084 (not a very dense network)

5. Conclusions and further work

We described a methodology which shows that research of social networks, beyond knowledge sharing, can provide us with extra information regarding particular research groups. This approach can be replicated for other research groups, such as reviewers/authors of a particular conference, potential academic employees, potential master/phd students. Further more, we think that research social networks could add value to the universities, since the methodology described together with the themes of published papers can be a contribution to profiling respective authors.

6. References

[1] C. Chen, CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. Journal of the American Society for Information Science and Technology, 57(3), 359- 377, 2006.

[2] D. J. de Solla Price, Little science, big science, Columbia University Press, New York, 1963.

[3] D. deB. Beaver and R. Rosen, Studies in scientific collaboration.1.Scientometrics, 1: 6584, Sage, 1978.

[4] D. deB. Beaver and R. Rosen, Studies in scientific collaboration. 2. Scientometrics, 2: 231245, Sage, 1979.

[5] D. deB. Beaver and R. Rosen, studies in scientific collaboration. 3. Scientometrics, 3: 133149, Sage, 1979.

[6] D. Easley and J. Kleinberg. Networks, Crowds, and Markets: Reasoning about a Highly Connected World. Cambridge University Press, 2010.

[7] E. Estrada, J. Rodriguez-Velazque, "Subgraph centrality in complex networks". Physical Review E, 71(5), 056103, 2005.

[8] E. Yanl and Y. Ding, "Discovering author impact: A PageRank perspective". Information Processing and Management, 47(1): 125-134, January 2011.

[9] L. Yin, H. Kretschmer, R. Hanneman, Z. Liu, Z, "Connection and stratification in research collaboration: An analysis of the COLLNET network". Information Processing and Management, 42, 1599-1613, 2006.

[10] M.Girvan and M. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA 99, 7821–7826, 2002.

[11] M. Newman. The Structure and Function of Complex Networks. SIAM Review 2003 45:2, 167-256

[12] M. Newman, Who is the best connected scientist? A study of scientific coauthorship networks, http://ideas.repec.org/p/wop/snfiwp/01-12-064.html, 2001.

[13] M. Newman, The structure of scientific collaboration networks, PNAS, 98:404409, the National Academy of Science, 2001

[14] M. Rodriguez and A. Pepe. “On the relationship between the structural and socioacademic communities of a coauthorship network”. Journal of Informetrics, 2(3), 195-201, 2008.

[15] P. Mutschke, “Mining networks and central entities in digital libraries. A graph theoretic approach applied to co-author networks”. Advances In Intelligent Data Analysis V, 2810, 155-166, 2003.

[16] R. Vidgen, S. Henneberg, P. Naude, “What sort of community is the European Conference on Information Systems? A social network analysis 1993-2005”. European Journal of Information Systems, 16(1), 5-19, 2007.

[17] X. Liu, J. Bollen, M. Nelson, H. Sompel “Co-authorship networks in the digital library research community”. Information Processing and Management, 41, 1462-1480, 2005.