The methodological status of co-authorship networks

Najko Jahn*

3 Jul 2008

Abstract

A powerful strategy within the study of collaboration in science is to posit that co-authorship patterns represent social networks.

It is prerequisite to an application of Social Network Analysis (SNA) to define the network entities. A network analysis of the inter-institutional collaboration in COLLNET on the basis of co-authorships was conducted. The study reveals that it is crucial whether the co-authorship itself is seen as an author's relational property or as a social event that brings the authors together. The former possibility is represented by a one-mode network in which each author can be related to each other author. Quite distinct from that are two-mode networks, the latter approach. They consist of two single data sets in which relations are only possible between different sets.

Different modes of representations require different network approaches. One is that co-authorship networks are seen as one-mode networks, which has the advantage of the application of a variety of measures. In contrast, two-mode networks, the other option, cannot be analysed by standard techniques but its distinctive features demand a new conceptualisation of measures. In conclusion, the two-mode perspective is more promising because it allows a dual perspective on collaboration in science which includes researchers as well as their scientific output.

1 Introduction

In most academic fields, inter-institutional collaboration is a common practice to promote lasting research improvements. Katz & Hicks (1997) showed with the aid of a bibliometric study that scientific publications that involve collaboration among different institutions gather more citations than papers written by authors from the same institution.

From a sociological perspective, this can be formulated in terms of Georg Simmel's notion of social circles. Instead of individual properties, he begs us to focus on the relational system someone is involved in. According to this, a particular event is identified by the multiple participations and memberships of the individuals that intersect with each other (Pizarro 2007, p. 769). In many cases of inter-institutional cooperation, these connections have a higher impact on the involved persons and institutions (Fennema & Schifj 1978). Because joint participation and overlapping memberships broad the information and knowledge flow among people and institutions, intersecting social circles count furthermore as an important source for achieving social recognition (Kadushin 2005).

In the case of inter-institutional collaboration in science, a jointly signed paper functions as a social circle as well as the institution as such. A scientometric approach on inter-institutional collaboration, therefore, has to focus on collaboration patterns and the effects of possible ties between institutions (Glänzel & Schubert 2004, p. 259).

Seeing the importance of social circles, Newman (2001a, 2001b) recommends to define scientific acquaintance with the help of co-authorship data; when writing a paper, the collaborators get to know each other. The research process, therefore, functions as a social circle in which people interact. Based on two different sets, the scientists on the one hand and the social events manifested in the co-authored papers on the other hand, Newman creates co-authorship networks with recourse to Social Network Analysis (SNA), a method based on the importance of interacting units (Wasserman & Faust 1994).

* Humboldt-Universität zu Berlin, Institute for Library and Information Science (IBI), Germany, najko@gmx.de

H. Kretschmer & F. Havemann (Eds.): Proceedings of WIS 2008, Berlin
Fourth International Conference on Webometrics, Informetrics and Scientometrics & Ninth COLLNET Meeting
Humboldt-Universität zu Berlin, Institute for Library and Information Science (IBI)

In adapting SNA, the scientometric debate attains a well proven method within social sciences which allows both processing and visualization of the gathered data.

According to Newman, co-authorship is represented by two-mode or affiliation networks. Throughout the paper, I use these terms interchangeably. In such a network, a relation within one set is only possible via a connection to the other set and vice versa. Its structure is at heart bipartite and thus differs from Otte & Rousseaus (2002) account.

Contrary to Newman, they model co-authorship relations from a unipartite perspective. Accordingly, one-mode networks contain only one set of vertices representing the authors. Even though in both types of network the scientists are connected by the joint authorship, the two network-types differ in their epistemic access to the phenomenon of collaboration in science. In the case of two-mode networks we presume the collaboration as a social event which brings the scientists together. In the other case, the co-authorship is just predicated of papers.

What seems at first glance as a non-essential philosophical puzzle, because it might just result from different modes of representation, has on closer examination serious practical implications. For instance, in accordance with Newman, Li-chun et al. (2006) claim to examine social relations among the researchers by their affiliation network (p. 1600), whose structure is bipartite and yet use methods valid only for unipartite networks. This indeterminacy, as we will see in this paper undermines their interpretations.

This paper aims firstly to measure the inter-institutional collaboration within the COLLNET-Network on the basis of a bipartite co-authorship network. This will be the basis of assessing in a second step why an indeterminate methodological status of co-authorship networks puts the scientometric debate in jeopardy.

2 Method

Before measuring the inter-institutional collaboration, we have to reconsider the scientometric data. Contrary to intersubjective relations, inter-institutional collaboration can be constituted by only one person. Besides co-authorship data, a single-authored paper has to be counted as valid pattern of cooperation as well if the author's institutional addresses listed in the publication indicates that the author is a member of at least two institutions. This is so because the joint membership is often based on an agreement between different institutions to share a researcher (Katz & Martin, p. 13). So, not only co-authorship data but also the author's institutional addresses play an essential role in scientometric debate in determining collaboration patterns in science (Glänzel & Schubert 2004, p. 259).

After making this conceptual distinction between collaboration and co-authorship, we follow Newmans bipartite approach in the notion of Wasserman & Faust (1994, Ch. 8) and examine the structure of the inter-institutional collaboration within COLLNET.

The basis of the two-mode network is the affiliation matrix $A=\{a_{ij}\}$ allocating members of the first set $N=\{n_1,n_2,...,n_g\}$ to members of the second set $M=\{m_1,m_2,...,m_h\}$. In an affiliation network, the former set contains g actors, the latter h events. In the actor-perspective, $a_{ij}=1$ if and only if actor i can be assigned to event j and $a_{ij}=0$ if and only if actor i joined not event j.

Since intersubjective relations between researchers are manifested in a jointly signed publication, the first set is constituted by articles. The second set, in turn, consists of the institutions listed in the underlying publication. Note that I do not distinguish between the different forms of inter-institutional collaboration as described by Katz & Martin (1997).

At this point, the distinction between actors and events seems arbitrary because both papers and institutions function as a publicly observable collection of scientists.[1] But, as we will see, these sets still represent distinctive features. Writing a paper and being affiliated with an institution are different kinds of social circles. The terms “actors” and “events” refers to these distinctive occurrence in order to maintain the dual perspective in the affiliation network by which papers are linked by their common institutions and institutions by their joint publications. The

1 I am indebted to Frank Havemann for making this point.

former identifies to which extent institutional knowledge flows into one particular paper without reference to a scientist, the latter detects the institutional output in a scientific community.

What exactly the shift from authors to papers is supposed to alter is a surprisingly delicate question. It may be a further step to scrutinize the methodological status of co-authorship networks by means of the concept of authorship. However, I will not pursue this here, but will restrict myself on the crucial distinction between unipartite and bipartite networks.

After these remarks, the matrix A is the starting point to examine the networks' global and particular properties.

2.1 The two-mode network

A common beginning to analyse two-mode networks is to calculate properties of actors, i.e. the publications, as well as events, i.e. the institutions listed in the publications from the matrix A (Wasserman & Faust 1994, p. 312ff.).

Institutions per Paper (IpP) is the number of institutions to which an paper i is connected. For our purpose, it measures the number of institutions involved in the research process for each publication. Their number is given by the row totals of A ,

$$a_{i+} = \sum_{j=1}^{h} a_{ij} \quad . \qquad (1)$$

However, *IpP* is equivalent to the *Degree Centrality* of a node. It is assumed that a high degree held by a vertice reflects a central position within the community. Because this measure infers the influence and prestige of a node from the number of links one node receives, it is in accordance with our initial hypothesis. The more institutions that are involved in a research, the higher is the impact of the paper in the particular scientific community (Katz & Hicks 1997). So, we are able to examine which article profits more and which less by the various institutional knowledge and information of its discipline.

In order to compare our findings with future studies, it is suitable to calculate the average number of institutions joined by one article. From the matrix A this is computed as

$$\bar{a}_{i+} = \frac{\sum_{i=1}^{g} \sum_{j=1}^{h} a_{ij}}{g} \qquad (2)$$

and is equivalent to *Mean Degree Centrality*.

In addition to the perspective on publications, we might examine the number of contributions one institution brings into the scientific discourse. Analogously, P*apers per Institution (PpI)* is calculated by column total,

$$a_{+j} = \sum_{i=1}^{g} a_{ij} \qquad (3)$$

and the average number is given by

$$\bar{a}_{+j} = \frac{\sum_{i=1}^{g} \sum_{j=1}^{h} a_{ij}}{h} \quad . \qquad (4)$$

Just as *IpP* is equivalent to the *Degree Centrality* for articles, *PpI* is a centrality measure as well. Therefore, we decide whether an institution occupies an important place within the community by the numbers of published papers. Whereas high-productive institutions can be located in the centre of the network, low productive ones are detected in the networks periphery.

For analysing co-authorship relations, *PpI* is in accordance with the traditional approach as described by Otte & Rousseau and applied to previous investigations of the COLLNET collaboration network (Kretschmer & Aguillo 2004, Kretschmer & Aguillo 2005, Li-chun et al. 2006). They detect the influence of one author, institution or country within in a scientific community with regard to the relations to other authors etc. However, in studying the degree of institutional coverage, the dual perspective of our approach takes into account the bipartite nature of scientific collaboration networks as presumed by Newman.

2.2 One-mode perspective

A reason why two-mode approaches have little impact on the scientometric debate lies in the fact that their analysis is more ambitious. For two-mode networks, many measures still miss a graph-theoretical founding so that the literature

recommends to focus on one mode for the subsequent analysis (de Nooy et al. 2005).

Following Wasserman & Faust (1994) further, we have to process the affiliation matrix A. If we aim at a one-mode analysis of actors, we summarize the relations between them by

$$X^N = A \times A' \quad . \qquad (5)$$

The first striking difference to the classical one-mode perspective is that the diagonal is meaningful. In our case, it counts the total number of institutions listed by one publication. Therefore, it is equivalent to the already defined *Degree Centrality*. Moreover, the derived network is valued and demands a different approach in order to measure the relations between pairs of papers.

One well-known measure in SNA is *Density*, which determines the general level of connectedness of collaboration networks (Otte & Rousseau 2002, p. 442). Since the maximally possible number of institutions a pair of article belongs to depends on the numbers of institutions represented in A, the *Density* of X^N is calculated as,

$$\Delta_{(N)} = \frac{\sum_{i=1}^{g} \sum_{j=1}^{g} x_{ij}^N}{g(g-1)} \qquad (6)$$

where $i \neq j$ and $\Delta_{(N)}$ ranges from 0 to h.

Assessing the overlap between institutions by joint publications is done analogously. The only difference is that we focus on the relations between the institutions, which are defined by the publications:

$$X^M = A' \times A \quad . \qquad (7)$$

We study the mean value of publications a pair of institutions jointly worked on by

$$\Delta_{(M)} = \frac{\sum_{k=1}^{h} \sum_{l=1}^{h} X_{kl}^M}{h(h-1)} \qquad (8)$$

where $k \neq l$. $\Delta_{(M)}$ is limited by g.

$\Delta_{(N)}$ functions as an index of potential acquaintance of articles, whereas $\Delta_{(M)}$ suggests how likely it is that two institutions collaborate within the network. Whether a pair of articles is more likely than a pair of institutions within the network indicates if scientist prefer to collaborate within the same or with different institutions.

After introducing the method it is clearly visible to which extent our approach differs from the methods applied for analysing unipartite networks. Even though both regard one-mode networks, ours is derived by the affiliation matrix A. If we, as we will see, do not take the bipartite arrangement of the gathered data into account, then our interpretation will become invalid.

3 Data

For the sake of the argument, I perform a network analysis that includes all contributions to the 2006 *Proceedings on the International Workshop on Webometrics, Informetrics and Scientometrics & Seventh COLLNET Meeting Nancy*[2] except Ashafi & Osarehs paper because it was not possible to detect their affiliations.

51 articles listed 57 institutions and are examined with the method described below.

The data are processed by the R-Project SNA-Package[3] and visualized with the help of Pajek[4].

4 Results

Inter-institutional collaboration instantiated by one author is common practice in COLLNET. That 10 authors list more than one institutional address shows the importance of an author's multiple affiliations which constitute and uphold inter-institutional collaborations. More than two institutions can be traced back to Hildrun Kretschmer and Michael Meyer.

2 COLLNET (Ed.). 2006.Proceedings International Workshop on Webometrics, Informetrics and Scientometrics & Seventh COLLNET Meeting. From: http://eprints.rclis.org/view/conftitle/International_Workshop_on_Webometrics,_Informetrics_and_Scientometrics_=_Seventh_COLLNET_Meeting.html

3 http://cran.r-project.org/web/packages/sna/index.html

4 http://pajek.imfm.si/doku.php?id=pajek

4.1 COLLNET two-mode network properties

In a first step, we consider the connectedness of the gathered affiliation network. Since the network is not connected but contains 30 components, the inter-institutional information flow within COLLNET is interrupted. The biggest component contains 24 vertices, i.e. 22,3% of all vertices.

Figure 1 clarifies general properties of two-mode networks further. Between members of one set are no paths of length 1 but at least of length 2 because connections within one set are constituted by the other set.

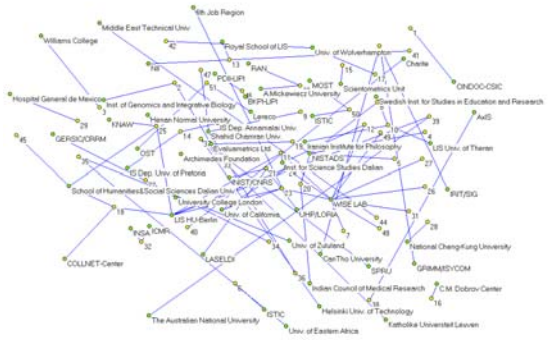

Figure 1: Affiliation network

We can now study both *IpP* and *PpI* from the matrix A. Table 1 reveals that whereas three publications are the respective product of four collaborating institutions the majority of contributions (58.8%) do not represent any inter-institutional collaboration according to our approach. In mean, each contribution results from 1.647 institutions.

Table 1: Institution per Paper

Institutes per Paper	*Contributions*
4	23 24 25
3	10 11 18 21 37 43
2	2 12 14 19 20 30 32 35 39 46 50 51
1	1 3 4 5 6 7 8 9 13 15 16 17 22 26 27 28 29 31 33 34 36 38 40 41 42 44 45 47 48 49

The other perspective on the two-mode network focuses on the degree to which institutions contribute to the scientific discourse within COLLNET. Table 2 shows the high degree of articles by members of Wise Lab as well as other departments located in Dalian (China). But LIS HU (Germany) and NISTADS (India) are engaged inversely proportional to the majority, too. Nine of 51 institutions participate at Nancy COLLNET Meeting with more than one paper, which, in turn, results in the Mean Degree Centrality 1.473.

Table 2: Number of papers for each institution

Papers per Institute	*Institution*
10	Wise Lab
6	LIS HU
5	NISTADS
3	School of Humanities&Social Sciences Dalian Univ.
3	Inst. for Science Studies Dalian
3	Shahid Chamran Univ.
2	Univ. of California
2	Univ. of Wolverhampton
2	Evaluametrics Ltd.

4.2 One-mode perspective

After identifying central publications and institutions within the network, we focus on the single sets and their relations with each other. First, we consider the relations between papers that are constituted by the second set, i.e. institutions. The processed one-mode network (Figure 2) is valued and determines how many institutions are shared by each pair of papers.

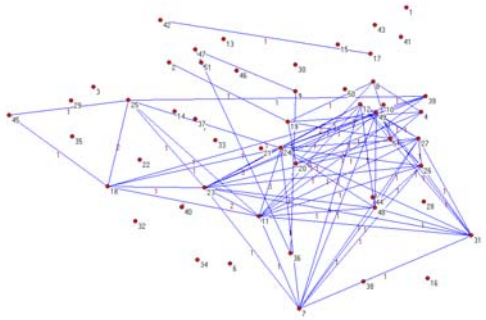

Figure 2: One-mode network – row perspective

The low degree of inter-institutional coverage is reflected in the one-mode perspective as well. The Density $\Delta_N = 0.0643$ means that, on average, a pair of paper shares only 0.0643 institutions. The maximally possible number is $h = 57$.

Analogously, in the second case, the relations between pairs of institutions are defined by the papers. Figure 3 visualizes the relations between the single institutions and the marked lines show the number of joint articles of a collaborating institutional pair.

Figure 3: One-mode network – column perspective

Even though Wise Lab (China) and NISTADS (India) belong to the highly productive institutions within COLLNET, they do not collaborate. The low degree of common inter-institutional cooperation validates $\Delta_{(M)} = 0.0301$. Thus, on average, each pair of institutions contributes 0.0301 papers during the conference.

4.3 Limits of one-mode perspective

It is tempting to investigate the derived unipartite networks further with regard to cliques since they resemble collaboration networks as described by Otte & Rousseau (2002, p. 443). But there are differences, which will be clearly visible in examining cohesion of subsets of actors or events.

Cliques are a basic concept in SNA. A clique consist of a subset of at least three vertices all of which are adjacent to one another (Wasserman & Faust 1994, p. 254). It represents a social relation which is based on complete mutuality.

Hence, one might detect such a clique in Figure 3 and claim that it represent a mutual collaboration between the institutions SPRU, Helsinki Universty of Technology and Katholike Universiteit Leuven.

$$n_1 \rightarrow \{m_1, m_2, m_3\}$$

Nevertheless, because of the bipartite arrangement of the data there are other scenarios possible.

$$n_2 \rightarrow \{m_1, m_2\}$$

$$n_3 \rightarrow \{m_1, m_3\}$$

$$n_4 \rightarrow \{m_2, m_3\}$$

For the reason that it is not decidable whether these institutions collaborated on the basis of the gathered unipartite network, we are in need of the affiliation matrix A which makes such relations clear. In our case it reveals that these three institutions in fact authored one contribution mutually (i.e. 37).

Thus, in the interpretation of the one-mode networks derived from a two-mode network, we have to restrict our inferences to pairs of actors or events.

5 Discussion

The findings reveal both patterns of inter-institutional collaboration in COLLNET and the benefits as well as the limits of a bipartite perspective on collaboration networks in science. Since the study was conducted for the sake of a

methodological argument, I will restrict myself to some remarks about the inter-institutional collaboration in COLLNET.

5.1 Inter-institutional collaboration in COLLNET

Even though some COLLNET-members are affiliated with more than one institution, they do not employ to the full all the theoretical possibilities inter-institutional collaboration has to offer. This is so because density calculations reveal a high discrepancy between the actual and the maximally possible values.

Because a pair of papers sharing one institution is more likely than a pair of institutions sharing one paper, we may conclude, that COLLNET-members prefer to collaborate with colleagues from the same rather than from different institutions.

Furthermore, our study reveals that only a few institutions play a central role, which is in accordance with earlier findings (Li-chun et al. 2006). Thus, both on intersubjective and inter-institutional level collaboration structures are sparse and depend upon an elite.

It is tempting to examine these structures further because they might have consequences in so far as they exclude other non-central members from the scientific discourse. But in order to investigate stratification processes the method described in this article is limited.

5.2 Benefits and Limits of the two-mode perspective

Limits of the two-mode perspective are due to the missing graph-theoretical techniques available for two-mode networks in the scientometric debate. Even though the adoption of well-known concepts like cliques which are introduced by Otte & Rousseau might be promising, their application prohibits the bipartite arrangement of our data.

But why should we continue to pursue the bipartite approach in spite of its limits? In order to assess this worry, suppose two methodological principles, the avoidance of ad-hoc hypothesis and the provision on an economical theory. In my opinion, these principles are well established and self-explanatory.

Let us confront both approaches with the former principle and decide whether they avoid ad-hoc explanations. Otte & Rousseau's account misses basic sociological concepts in so far as they only presume relations by means of the manifested co-authorship. In contrast, in introducing Simmel's notion of social circles, we do not just attribute a relation between authors, but we catch up with a theory which maintains our conclusion sociologically. This may be one of the reasons why Li-chun et al. (2006) refer to the term “affiliation network”.

Nevertheless, in claiming two sets the bipartite account seems overdetermined because co-authorship networks are represented in a unipartite network as well. Although bipartite networks are in so far overdetermined as they allow one more set, they are efficient from an economical point of view for the reason that they are not committed to unknown kinds of scientometric entities which are part of the co-authorship network. In our case, the bipartite arrangement of the data allows a dual perspective, which gives us a more profound insight in the structure of co-authorship network.

In conclusion, we should be aware of the different properties of unipartite and bipartite networks. Although SNA is commonly used, I argued that the methodological status of co-authorship networks is still indeterminate which, in turn, undermines the results of its application. Which approach to take cannot be decided yet, but I have shown that there are good reasons to opt for the bipartite representation of co-authorship networks.

Acknowledgement

I am deeply grateful to Hildrun Kretschmer, Frank Havemann and Michael Heinz for introducing me into the concepts of co-authorship networks as well as giving me always critical and helpful comments on my work. I am indebted to Karen Schumann, Franziska Singer and Stefanie Schweller, too.

References

de Nooy, W., Mrvar, A., & Batagelj, V. (2005). *Exploratory Social Network Ana-*

lysis with Pajek. Structural analysis in the social sciences, 27. Cambridge: Cambridge University Press.

Fennema, M. & Schifj, H. (1978). Analysing interlocking directorates: Theory and method. *Social Networks* 1(4), 297-332.

Glänzel, W & Schubert, A. (2004). Analyzing scientific networks through co-authorship. In: Moed, H.F. Et al. (eds.). *Handbook of Quantitative Science and Technology Research*, 257-276.

Kadushin, Ch. (2005). *The American intellectual elite : with a new introduction by the author*. Originally published 1974. New Brunswick, London: Transaction Publishers.

Katz, J. S., & Hicks, D. (1997). How much is a collaboration worth? A calibrated bibliometric model. *Scientometrics* 40(3), 541–554.

Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy* 26, 1–18.

Kretschmer, H., & Aguillo, I. F. (2004). Visibility of collaboration on the Web. *Scientometrics* 61(3), 405–426.

Kretschmer, H., & Aguillo, I. F. (2005). New indicators for gender studies in Web networks. *Inf. Process. Manage*. 41, 1481–1494.

Li-chun, Y., Kretschmer, H., Hanneman, R. A., & Ze-yuan, L. (2006). Connection and stratification in research collaboration: an analysis of the COLLNET network. *Inf. Process. Manage.* 42(6), 1599–1613.

Newman, M. E. J. (2001a). The structure of scientific collaboration networks. *Proceedings of he National Academy of Sciences* 98 (2), 404–409.

Newman, M. E. J. (2001b). Scientific collaboration networks I - Network construction and fundamental results. *Phys. Rev. E 64*, 016131 (2001). DOI: 10.1103/PhysRevE.64.016131.

Otte, E., & Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. *Journal of Information Science* 28(6), 441–453.

Pizarro, N. (2007). Structural Identity and Equivalence of Individuals in Social Networks. *International Sociology* 22(6), 767-792.

Wasserman, S., & Faust, K. (1994). Social network analysis: Methods and applications. Cambridge: Cambridge University Press.