Journal of Social Structure

JoSS

JoSS Article: Volume 12

Explorative Visualization of Citation Patterns in Social Network Research[*]

Ulrik Brandes and Christian Pich
Department of Computer & Information Science
University of Konstanz

Abstract

We propose a visual representation of bibliographic data based on shared references. Our method employs a distance metric that is derived from bibliographic coupling and then subjected to fast approximate multidimensional scaling. Its utility is demonstrated by an explorative analysis of social network publications that, most notably, depicts the genesis of an area now commonly referred to as network science. However, the example also illustrates some common pitfalls in bibliometric analysis.

1 Introduction

Bibliographic data are highly structured and highly relational. Hence it is not surprising that bibliometry (White and McCain, 1989) is an area in which data abounds and network analysis flourishes. Here we are interested in bibliographic data about social network research, and more specifically in its decomposition into subfields as reflected in the variation of citation patterns.

Social network research has been characterized as a *normal science* almost two decades ago (Hummon and Carley, 1993). In particular, the existence of an invisible college, a shared paradigm, and a primary journal are ascribed. Using main path analysis (Hummon and Doreian, 1989) on citations made in the first twelve volumes of the journal *Social Networks*, Hummon and Carley (1993) also identify six main research streams that shape the field. Neither do we want to replicate this study, nor continue it with newer data. Instead,

[*] A preliminary version of this work was presented in the 7th Viszards Session at the Sunbelt XXVIII Social Network Conference in St. Pete Beach, Florida, USA, 22-27 January 2008.

we want to demonstrate that diversity and scope have increased enough to facilitate the identification of more general thematic and behavioral clusters from citation data.

Our focus is on shared citation patterns rather than individual citations or citation counts, because quantitative analyses of citations rest on shaky grounds (Leydesdorff, 1998), and the purposeful evaluation of citations by actors outside of the peer review process affects the way in which citations are being made Weingart (2005). In particular, authors appear to adapt to the increasingly common attempt to identify important publications and research streams via citation counts (Lawrence, 2007; Todd and Ladle, 2008).

A form of bibliometric analysis that should be relatively stable with respect to such developments is the identification of clusters from similar citation behavior. In fact, increasing specialization and even strategic citations can be expected to lead to rather more prominent clusters. Whether these correspond to thematic topoi, schools of thought, or citation clubs, is a separate and substantive discussion that needs to take the specific context into account.

To explore the evolution of social network research after the above-mentioned study, we propose a novel visualization approach. While it is very similar to previous approaches (Small, 1999; Brandes and Willhalm, 2002; Chen and Hsieh, 2007), it differs in two important aspects: the way in which citation linkages are converted into a distance measure and the way in which a visualization is obtained from these distances.

The remainder of this contribution is therefore organized as follows. In Section 2 we provide background on different ways to represent bibliographic data in the form of networks. The data set selected for exploration is introduced in Section 3. Our proposed visualization approach is outlined in Section 4 and results from its application to the selected data are discussed in Section 5.

2 Bibliographic Networks

In this section we describe the underlying bibliographic data model. Bibliograpic networks can be derived from *objects*, which correspond to entities in bibliographic data, basic *relationships* among these objects as they are stored in bibliographic databases, and derived relationships constructed via the application of bibliographic *operators*.

We will use simple lower case letters such as m, n, p for numbers, indexed lower-case letters such as a_7, w_i to denote objects, script letters such as $\mathcal{C}, \mathcal{W}, \mathcal{A}$ for sets of objects, and upper-case letters such as G, C, A to denote matrices that represent relationships between objects.

Objects. The objects of interest in the present paper are sets

- $\mathcal{A} = \{a_1, \ldots, a_m\}$ of *authors*,
- $\mathcal{W} = \{w_1, \ldots, w_n\}$ of *works*, and
- $\mathcal{J} = \{j_1, \ldots, j_p\}$ of *journals*.

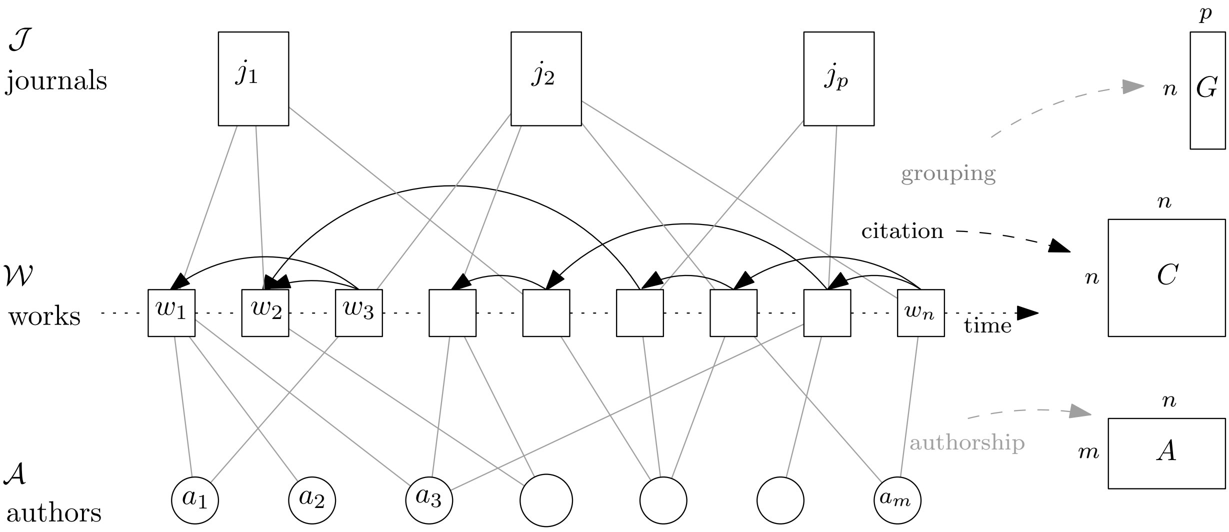


Figure 1: A basic data model for bibliographic networks. The matrices on the right represent basic relationships: a two-mode network G of journals containing works, a directed and (essentially) acyclic network C of citations between works, and another two-mode network A of works and their authors.

Other types of objects not considered here include institutions, scientific disciplines, and keywords.

Basic Relationships. Bibliographic objects are linked by relations such as citations between works or affiliations of authors with institutions. These give rise to bibliograpic networks and are conveniently represented in matrices. We refer to a relationship as basic, if it is stored explicitly in a bibliographic database.

We here consider three of the most common bibliographic relations, all of which give rise to unweighted directed graphs that can be represented in adjacency matrices with entries 1 and 0 for relations that are either present or absent. More precisely, our bibliographic model consists of the following basic relationships which are also depicted in Figure 1.

- *Citations* among works are stored in a square adjacency matrix $C \in \{0,1\}^{n \times n}$ that represents a directed graph with a directed edge from work w_1 to work w_2 if and only if w_1 cites w_2. Citation graphs are almost always acyclic, since citing works generally appear after cited works.

- *Authorship* is represented in a rectangular matrix $A \in \{0,1\}^{m \times n}$. It corresponds to a bipartite graph with an edge from author $a \in \mathcal{A}$ to work $w \in \mathcal{W}$ if and only if a is an author of w.

- *Grouping* relates works, for instance, by the journals or conference proceedings in which they are published, institutions to which their authors are affiliated, or scientific disciplines of the journals in which they are published. Assuming that every work appears in exactly one journal, every row of matrix $G \in \{0,1\}^{n \times p}$ contains exactly one 1.

Bibliographic Operators. Bibliographic operators can be applied to known relationships to derive new ones. The most common operators correspond to multiplication of the involved matrices. Each of them reflects a particular analytic perspective on basic relationships in the original bibliographic data. Note that the resulting matrix products are themselves matrices representing weighted graphs.

- *Bibliographic Coupling* (Kessler, 1963) between two works occurs if both cite some of the same references. By definition of the matrix product, the ij entry of $C^T C \in \mathbb{N}^{n \times n}$ gives the number of shared entries in the reference lists of works w_i, w_j.
- *Co-Citation* (Small, 1973) is dual to bibliographic coupling and measures how many times two works w_i, w_j are cited together in other works. These number can be read off matrix $CC^T \in \mathbb{N}^{n \times n}$.
- *Collaboration* counts the number of works that two authors have co-authored. Again, these numbers are conveniently obtained from a matrix product, in this case $A^T A \in \mathbb{N}^{m \times m}$.
- *Projection* is an example for a less commonly used operator. Instead of citations between works given in matrix $C \in \{0,1\}^{n \times n}$, one may be interested in citations between authors (White and Griffith, 1981). This information can also be obtained by concatenation of basic relations: the ij entry of product matrix $ACA^T \in \mathbb{N}^{n \times n}$ gives the number of times that a_i authors a paper in which another paper is cited that is authored by a_j.

In some contexts it may be necessary to post-process the entries of matrices obtained from bibliographic operators, for instance, by binarization or normalization with respect to row or column sums.

3 Data

The analysis of Hummon and Carley (1993) is based on papers which appeared in the journal *Social Networks* from 1978/79–1990. Instead of simply adding papers from subsequent years, we used more inclusive criteria to account for the increased scope the field.

The data set was compiled by Vladimir Batagelj (University of Ljubljana) on January 5, 2008, from *Web of Science*, a bibliographic database of ISI/Thomson.[1] Using a tool called WoS2Pajek,[2] the online server of the database was queried for publications that

- appeared in the journal *Social Networks* 1978/79–2007,

[1] http://scientific.thomson.com/products/wos/

[2] Freely available from http://vlado.fmf.uni-lj.si/pub/networks/pajek/WoS2Pajek/.

- contain the phrase "social network" or "social networks" or
- are authored by at least one of almost 100 hand-picked researchers in this area.

The query returned 5,463 publications written by 10,239 authors and published in 1,673 journals. In total, these publications contain 324,616 citations. Including all the works cited, the full data set, referred to as SN5, contains 193,376 publications written by 75,930 authors in 14,651 journals.[3]

Our analysis is focused on publications that were returned for the initial query *and* for which reference lists were available. The complete data are used, however, to determine the strength of linkages among them more accurately.

As can be expected, there are several issues with criteria-based bibliographic data such as this. Some problems such as scope and spelling variants are obvious and common to all bibliometric data sets. Other problems are more subtle and may lead to non-obvious misinterpretation. Examples of the latter are given in Section 5. We have no means to assess this, but do suspect that such or similar problems are fairly common as well.

4 Methods

Our approach can be considered an instantiation of the ISOMAP framework (Tenenbaum et al., 2000) with a distance measure derived from bibliographic coupling strength and an approximate representation of these distances in display space. Both components are described in detail in the following two sections, and the entire workflow is summarized in Figure 2.

4.1 Distance Measure

The transformation described here also applies to other basic relations among bibliographic objects. For simplicity and concreteness we restrict ourselves, however, to the citation relation between works.

Our goal is to generate a visualization in which works with similar bibliographies are placed close to each other. We therefore define a distance that is small if the overlap is large and vice versa.

From the bibliographic coupling operator described in the previous section we obtain the number of references shared by two works. While this can be considered a measure of similarity, it does not take into account the number of references in each paper. Having, for example, exactly three shared references may point to strong similarity between specialized papers in which only very few other works are cited, but may also point to strong dissimilarity between survey papers with lots of references.

Variant measures of bibliographic coupling therefore normalize by the number of references cited. In *Jaccard's index*, the number of joint entries is divided by the total number

[3] SN5 is available at `http://vlado.fmf.uni-lj.si/pub/networks/data/WoS/SN5.zip`.

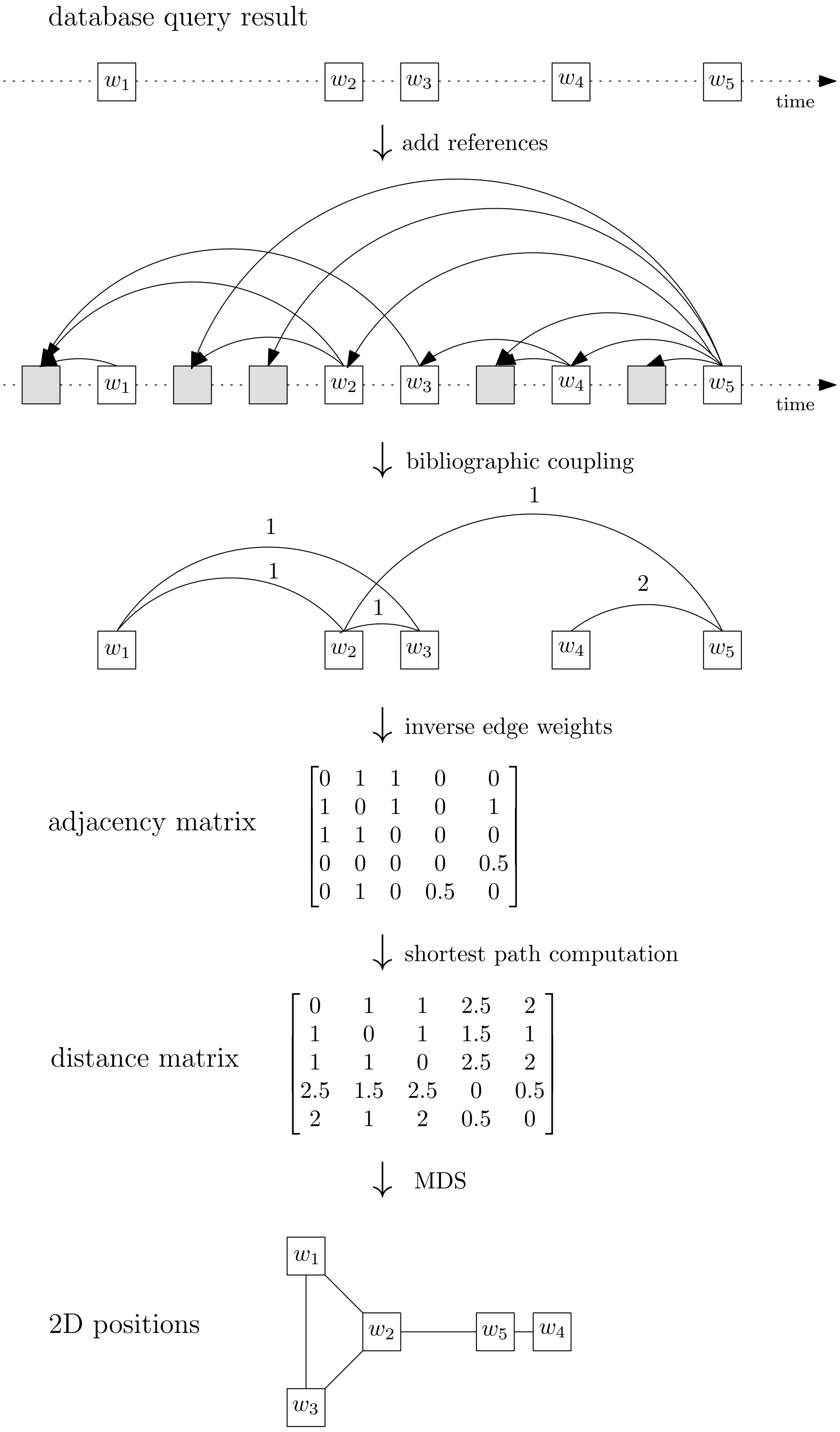


Figure 2: Basic layout workflow for a query result.

of references cited in either work,

$$J_{ij} = \frac{\left(CC^T\right)_{ij}}{|\mathcal{W}_i \cup \mathcal{W}_j|} = \frac{|\mathcal{W}_i \cap \mathcal{W}_j|}{|\mathcal{W}_i \cup \mathcal{W}_j|} \; .$$

An alternative is *Salton's index*,

$$S_{ij} = \frac{|\mathcal{W}_i \cap \mathcal{W}_j|}{\sqrt{|\mathcal{W}_i| \cdot |\mathcal{W}_j|}} \; ,$$

in which normalization is with the geometric mean. Note that this corresponds to *cosine similarity*, which is defined as the ratio between the inner product of two vectors divided by the product of their length and thus gives the cosine of the angle between the two vectors. The correspondence is via the characteristic vectors of the two sets. Since the number of references is usually of similar order, and the number of distinct references larger than the number of common references, Jaccard's and Salton's indices are often close to being multiples of each other (Hamersa et al., 1989).

Both indices are typically subjected to thresholding, i.e. the replacement of small values by zero, because the citation of some very general and highly cited references may otherwise create spurios similarities.

For our purposes, the choice of index is not of great importance. Hence, mostly for its straightforward interpretation, we stick to Jaccard's index and transform it into a dissimilarity measure by taking its inverse. While the inverse Jaccard index $\delta_{ij} = \frac{1}{J_i j}$ satisfies the requirement that it assigns small values to similar bibliographies and vice versa, it also has some drawbacks. Firstly, it is undefined if there are no shared references, since we would have to take the inverse of zero. Secondly, it is not a distance metric, because the triangle inequality is not satisfied: consider two works w_i, w_j with ten references each that share only one citation, and a third work w_k that also lists ten references, half of which are shared with w_i and the other half are shared with w_j. Then the dissimilarity between w_i and w_j is $\frac{19}{1}$, but the sum of the dissimilarities between w_i and w_k and between w_k and w_j is $\frac{15}{5+5} = \frac{3}{2}$, i.e. much smaller. And finally, as a consequence of both other drawbacks, the inverse Jaccard index is not stable against thresholding.

To be able to position works in two-dimensional space according their similarity, we therefore propose to construct a weighted graph with vertex set $\mathcal{W}$ and edges between every pair of works that have at least one common citation. For these edges, the inverse Jaccard index is well-defined. Then, we obtain a distance d_{ij} between pairs of works w_i, w_j by computing the shortest-path distance between their corresponding vertices, where the length of a path is the sum of the weights on its edges. This distance is a metric, and it is relatively stable against thresholding because edges of large weight (dissimilarity) are unlikely to be part of a shortest path.

The shortest-path distances d_{ij} obtained from the graph of finite dissimilarities $\delta_{ij} = \frac{1}{J_{ij}}$ can now be used to determine positions in two-dimensional space.

4.2 Multidimensional Scaling

The most common approaches to visualize bibliographic coupling or co-citation data are based on multidimensional scaling. See White and McCain (1998) or He and Hui (2001) for examples, and Börner et al. (2003) for a general overview.

Multidimensional scaling, or MDS for short, refers to a family of dimension-reduction techniques that turn a dissimilarity matrix into a two- or three-dimensional scatterplot. For a general introduction we refer to Borg and Groenen (2005) or Cox and Cox (2001). We first sketch the particular variant of MDS we are using here, and then our recent approximation technique which makes the approach suitable also for large data.

The input consists of distances d_{ij}, and the goal is to determine two-dimensional positions $p_1, \ldots, p_n \in \mathbb{R}^2$ with $p_i = (x_i, y_i)$ for all n objects such that the resulting Euclidean distances $\|p_i - p_j\| = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$ match the input distances d_{ij} as closely as possible. While there are various ways of achieving this objective, e.g. by numerical minimization of an objective function, we will concentrate on the earliest approach, which is often termed *classical scaling* (Torgerson, 1952).

Classical scaling is best understood by considering a hypothetical embedding in a metric high-dimensional space that is assumed to have given rise to the input distances. The product of the matrix of position vectors with its transpose is a matrix of inner products. From inner products, distances (actually, their squares) can be computed easily. Since inner products can, in turn, also be determined backwards from given distances, the idea is to compute an approximate low-rank factorization of the matrix of inner products. This yields low-dimensional coordinates with almost the same inner products, and therefore almost the same distances.

The four steps necessary are:

1. Construct a matrix $B \in \mathbb{R}^{n \times n}$ of squared distances d_{ij}^2.

2. Double-center B by subtracting from each entry the column and row means and adding the overall mean, so that the entries in each row and column sum to zero. (Note that this removes a degree of freedom stemming from the invariance of distances under translation).

3. Compute the two largest eigenvalues $\lambda_1, \lambda_2 \in \mathbb{R}$ and corresponding eigenvectors $u_1, u_2 \in \mathbb{R}^n$ of the double-centered matrix. Scale them equally, say, to unit length, $\|u_1\| = \|u_2\| = 1$.

4. Determine positions by setting $x = \sqrt{\lambda_1} u_1$ and $y = \sqrt{\lambda_2} u_2$.

Positions thus obtained are unique up to scaling (except in the empirically rare case that $\lambda_1 = \lambda_2$) and therefore reproducible.

Classical scaling is very appropriate for dense matrices of metric distances such as the ones we are dealing with. It is, however, computationally costly because it requires the entire double-centered distance matrix to be known. For large n, this may be prohibitive in terms of running time for the all-pairs shortest-path problem, but even more so in terms of the storage space needed for the quadratic matrix.

A computationally cheap alternative is our recently introduced technique for approximate classical scaling called PivotMDS (Brandes and Pich, 2006). Instead of the $n \times n$ distance matrix, it suffices to compute distances from a small number $k \ll n$ of pivots. The square matrix B is hence replaced by a rectangular $n \times k$ matrix C of squared distances to pivots only. Instead of eigenvectors of double-centered B, we compute left singular vectors of double-centered C which approximate the eigenvectors quite well at a fraction of the computational cost. Since, usually, $k \approx 100$ works reasonably well for any practically relevant size n, the method is at least an order of magnitude faster and requires only linear space.

5 Results

From the SN5 data we have generated networks of works, authors, and journals, respectively, by computing the Jaccard index of citation overlap and creating edges weighted by the inverse Jaccard index where the overlap exceeds a pragmatically chosen threshold. The following table summarizes the size of the giant components in networks resulting from these transformations.

domain	no. of nodes	no. of edges	max. dissimilarity
works	5 643	87 528	0.25
authors	9 273	111 068	0.10
journals	1 602	24 311	0.10

In these edge-weighted networks, shortest-path distances with respect to 200 pivots were computed and used to determine positions via PivotMDS. In each of the visualizations shown below, nodes represent bibliographic objects, and links represent the derived similarity measure of citation overlap. Node areas correspond to citation numbers as detailed below, and link widths and colors ranging from white to gray indicate the degree of similarity.

5.1 Works

Figure 3 shows the resulting visualization for bibliographic coupling among works. The diagram suggest that the field of social network analysis can be organized into several clusters, with the most prominent one in the middle-left. It contains many methodology-related articles and largely corresponds to the kind of research traditionally presented in the *Sunbelt Social Networks Conference* series. It appears to be appropriate to refer to it as *mainstream social network research.*

Opposite on the right we find works revolving around social support and public health. The diversified area in the upper left can be associated roughly with organizational networks, socio-economic status, social capital, and ego networks.

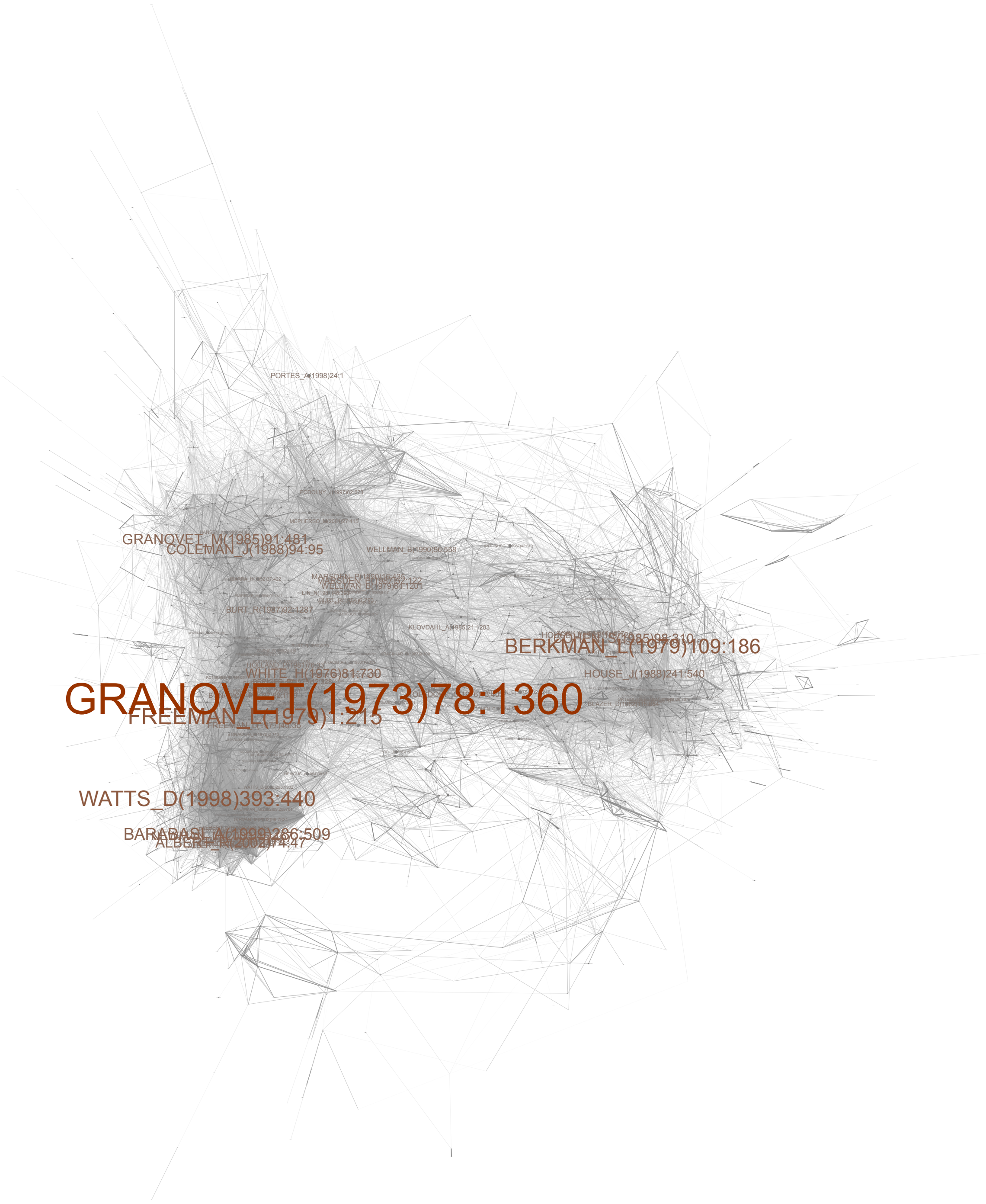


Figure 3: Giant component of the bibliographic coupling network among works in the SN5 data set. Node area and label font size are proportional to number of citations from within the data set, i.e. before and until 2007. Labels are centered. Note that bibliographies of books are not contained in the data which effectively excludes books from the network.

cited	label	reference
675	**WASSERMA_S(1994):**	**S. Wasserman and K. Faust: *Social Network Analysis.* Cambridge University Press, 1994.**
660	GRANOVET(1973)78:1360	M. Granovetter: The strength of weak ties. *American Journal of Sociology* 78(6):1360–1380, 1973.
396	**BURT_R(1992):**	**R.S. Burt: *Structural Holes.* Harvard University Press, 1992.**
344	FREEMAN_L(1979)1:215	L.C. Freeman: Centrality in social networks. *Social Networks* 1(3):215–239, 1978/1979.
326	WATTS_D(1998)393:440	D.J. Watts and S.H. Strogatz: Collective dynamics of 'small-world' networks. *Nature* 393:440–442, 1998.
313	BERKMAN_L(1979)109:186	L.F. Berkman: Social networks, host resistance, and mortality. *American Journal of Epidemiology* 109:186–204, 1979.
269	**FISCHER_C(1982):**	**C.S. Fischer: *To Dwell among Friends.* University of Chicago Press, 1982.**
250	BARABASI_A(1999)286:509	A.-L. Barabási and R. Albert: Emergence of scaling in random networks. *Science* 286(5439):509–512, 1999.
242	**COLEMAN_J(1990):**	**J.S. Coleman: *Foundations of Social Theory.* Harvard University Press, 1990.**
221	GRANOVET_M(1985)91:481	M. Granovetter: Economic action and social structure: The problem of embeddedness. *American Journal of Sociology* 91(3):481–510, 1988.
220	ALBERT_R(2002)74:47	R. Albert and A.-L. Barabási: Statistical mechanics of complex networks. *Reviews of Modern Physics* 74:47–97, 2002.
212	COLEMAN_J(1988)94:95	J.S. Coleman: Social capital in the creation of human capital. *American Journal of Sociology* 94(Supplement):S95–S120, 1988.
200	WHITE_H(1976)81:730	H.C. White, S.A. Boorman and R.L. Breiger: Social structure from multiple networks. *American Journal of Sociology* 81(4):730–780, 1976.
200	COHEN_S(1985)98:310	S. Cohen and T.A. Wills: Stress, social support, and the buffering hypothesis. *Psychological Bulletin* 98(2):310–357, 1985.

Table 1: Of the 14 works with at least 200 citations, four are books that hence do not appear in Figure 3.

The remaining prominent region, in the lower left, is very chohesive and its most highly-cited articles are more recent than those in the other areas. The authors in this area are mostly not from the social sciences and it could be labeled as *network science*. Bibliographies within this cluster are very homogeneous and otherwise related almost exclusively to the mainstream.

The most prominent works are highlighted by node area and label font size corresponding to the number of citations they receive from within the data set. While Granovetter's seminal paper on the strength of weak ties appears to dominate the field, Table 1 indicates that the book of Wasserman and Faust is actually the most cited reference.

Counting manually identified spelling variants, the citation count for GRANOVET(1973) 78:1360 increases to 808, but similar statements are true for other works as well. A particularly difficult entity is the well-known software tool UCINET, which is referenced in various ways and with various publication years. We have found more than 20 such variants for a total exceeding 300 citations. The program is thus likely to be among the TopTen of works relative to this data set, but does not figure prominently in the automated analysis. Of course it is open to debate how to compare a continuously developed piece of software with, say, revisions of books.

Since Figure 3 is based on bibliographic coupling, books do not even appear in the giant component because their bibliographies are not part of the data and thus cannot overlap with others. Similarly, some highly cited articles such as Milgram's small world article of 1967 or Erdős and Rényi's 1959 paper on random graphs are missing form the giant component because they are not part of the set of core articles (returned by the initial query) whose bibliographies have been evaluated.

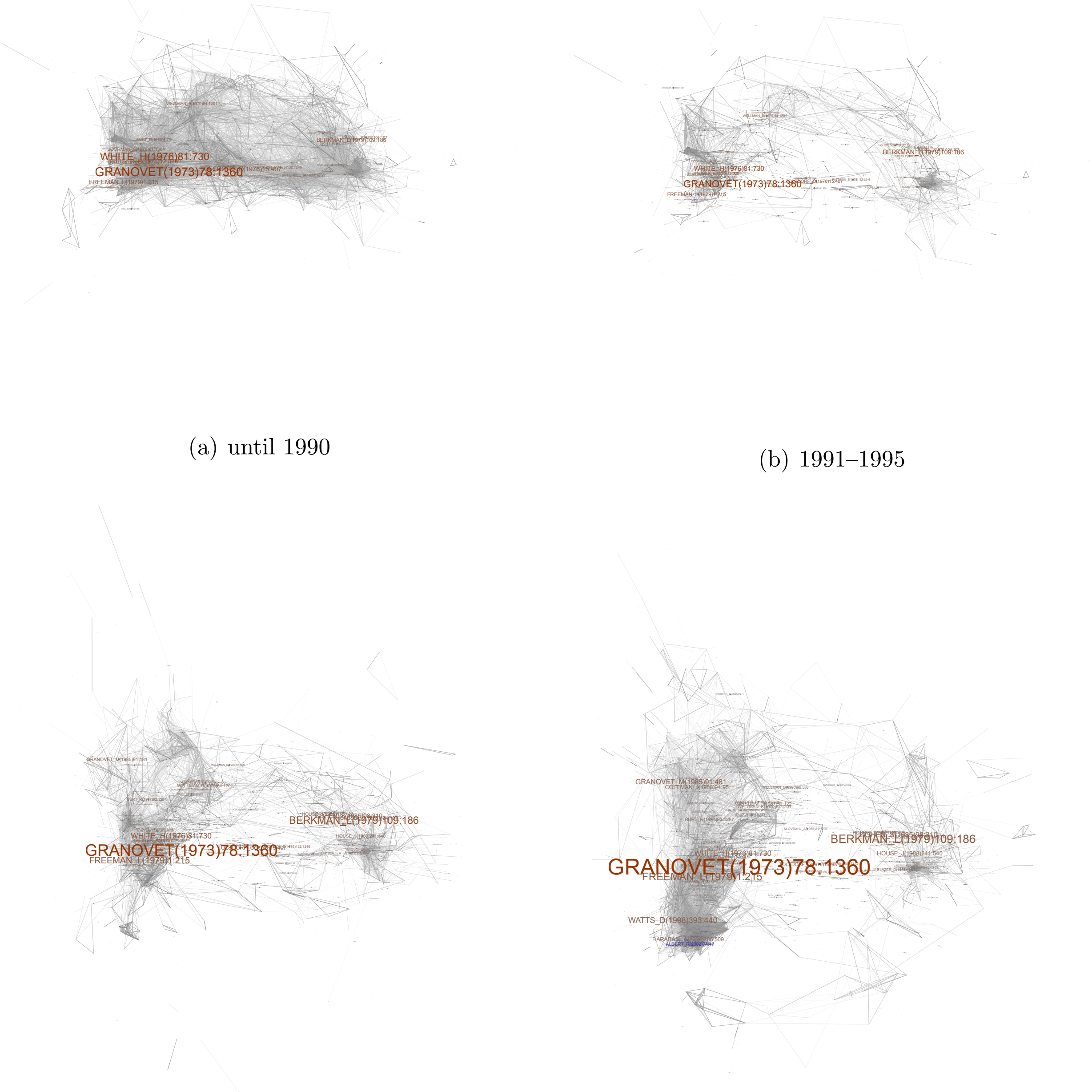
We finally take a look at how this situation developed. Figure 4 starts in 1990, i.e. with the situation reported in Hummon and Carley (1993), though with more inclusive data. Major parts of the 2007 structure with mainstream, health, and community clusters are already present. The subsequent diagrams show only similarity ties between works published in the respective interval. This serves to highlight where new and (in terms of citation patterns) similar work is being published.

It is interesting to follow the expansion and diversification over time. The most striking development, however, starts at the end of the 1995–2000 interval, when the first network science papers about small worlds, preferential attachment, and power laws appear. While much is going on in all areas, the most rapid development is happening in the lower left, with some works from the same and the previous period quickly catching up with the most cited papers over the entire covered time span.

A similar observation was made in a more focused analysis (Lazer et al., 2009). Note that, despite its title, this study is actually also based on bibliographic coupling, not co-citation.

5.2 Authors

By aggregating over works someone authored or co-authored, the concept of bibliographic coupling can be extended to authors. Two authors are considered to be coupled, if the sets



(a) until 1990

(b) 1991–1995

(c) 1996–2000

(d) 2001–2005

Figure 4: Evolution of the bibliographic coupling graph of works from 1990 to 2005. Positions are maintained from Figure 3, but a work is shown only if it was published by the respective year. Items new in a time interval are colored blue, and sizes are according to citations until then.

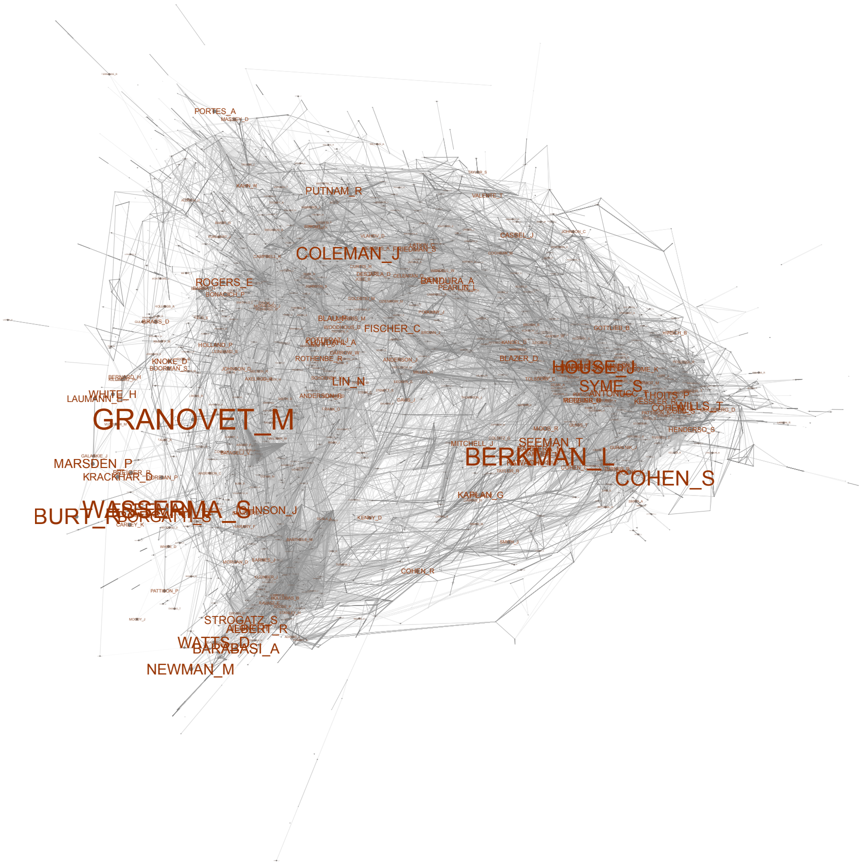


Figure 5: Bibliographic coupling network among authors in the SN5 data set. The area of each author node is proportional to the number of authors by which that author is cited.

of authors they cite overlap. As mentioned in Section 2, the number of times an author cites another author can be obtained from matrix ACA^T, where A is the author-by-work authorship matrix and C is the work-by-work citation matrix. Dichotomizing ACA^T yields a 0-1-matrix of authors citing authors corresponding to the work-citing-work matrix C that was used in the previous section to determine bibliographic coupling of works.

In Figure 5, authors are thus positioned according to how similar the sets of authors are that they refer to in their works. Note that this is different from scientific collaboration as investigated, for instance, in Moody (2004) because it does not require access to a potential co-author. Clusters in the coupling network could be due, for example, to the citation culture in a field, or author citation clubs.

While the general organization into compartments is not unlike that in Figure 3, it is interesting to note that many of the more prolific authors are rather peripheral. We assume that this is due to distinct specialties and the larger range and diversity of work they draw from, but did not check in detail.

Node area and label font size are proportional to the number of distinct authors from whom citations are received. This could be considered a measure of the breadth of influence.

citers	name
1718	Mark Granovetter
1475	Lisa F. Berkman
1431	Stanley Wasserman
1302	Ronald S. Burt
1227	Sheldon Cohen
1102	**Barry Wellman**
1079	James S. House
1030	James S. Coleman
1000	Linton C. Freeman

Table 2: Authors cited by at least 1000 distinct authors in the SN5 data set. Note that Wellman is not visible in Figure 5 because he is missing from the giant component.

The first thing to notice is the absence of Barry Wellman, who is the founder of the *International Network of Social Network Analysis* (INSNA) and certainly among the most widely cited authors (see Table 2) but missing from the giant component. Only at a threshold of 0.12 does he become connected to the giant component. This is only one of many conceivable difficulties in choosing an appropriate threshold.

Another interesting observation is that coupling among authors in the network science corner is much stronger than elsewhere. Aside from the health-related area on the right, this is the only visible clustering.

Observe, however, that the large variance in the number of publications and therfore also in citations made by the authors has an influence on the similarity measure. A severe data problem is that some authors appear under various labels. The most widely cited author, Granovetter, appears as GRANOVET with 1022 citing authors and as GRANOVET_M with 1152 citing authors. The union of these comprises the 1718 distinct labels of citing authors mentioned above. While we manually corrected severe distortion among the most widely cited authors, automatically doing so for all authors is a challenge in most data sets.

5.3 Journals

Finally, we use the method of the previous section once more, although this time to assess bibliographic coupling of journals rather than authors. The data is obtained in the exact same way, with the basic relation author-authored-work replaced by journal-published-work.

In Figure 6, positions once again represent how similar two journals are in terms of the sets of journals from which publications are cited in them, and sizes represent the number of distinct journals from which citations are received. In terms of being cited from within many other journals, *Social Networks* (343 citing journals) ranks fourth behind the *American Journal of Sociology* (518), *Social Science & Medicine* (404), and the *American Sociological Review* (379), and similar to some of the main authors in the previous section

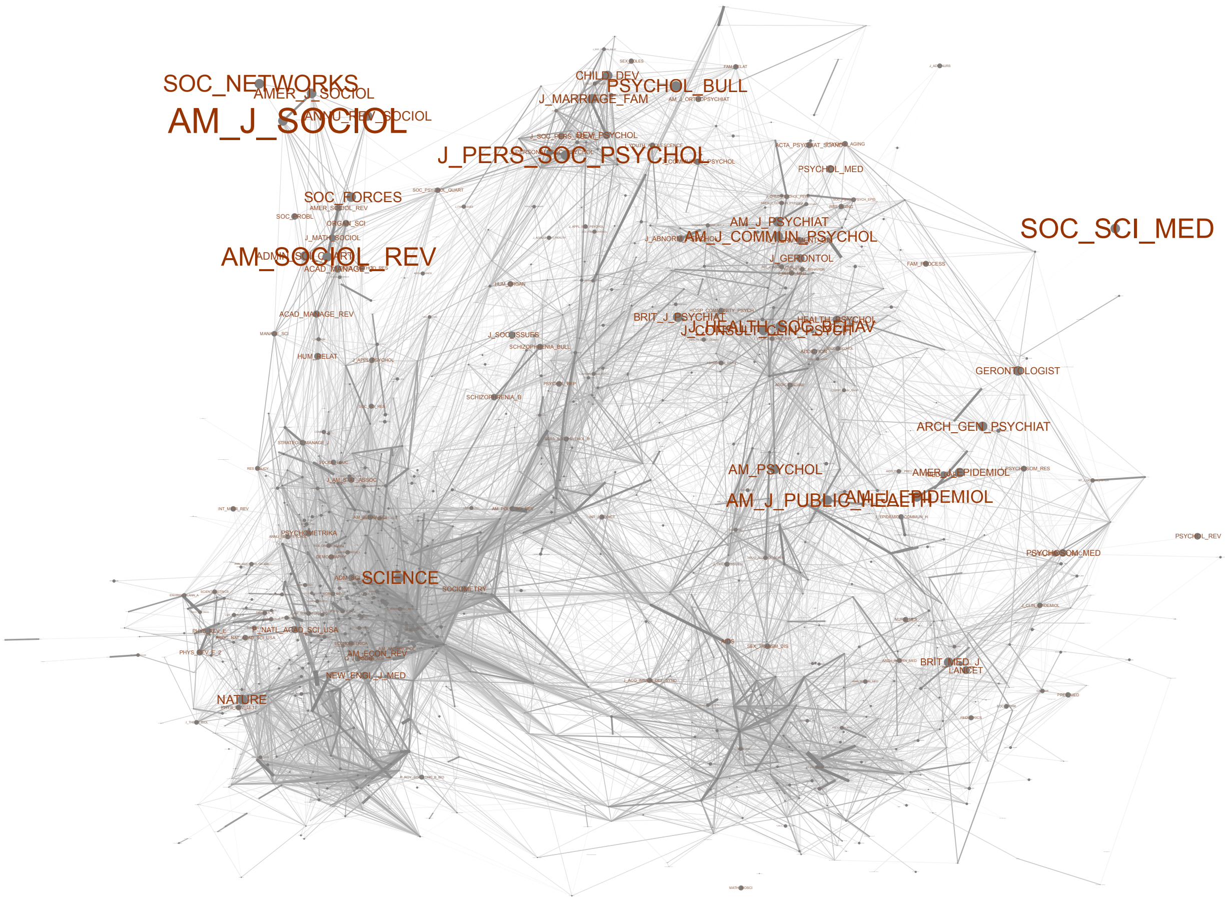


Figure 6: Bibliographic coupling network among journals in the SN5 data set. The area of each journal is proportional to the number of journals in which that journal is cited.

it apparently has a distinct profile. This prompts us to renew, almost twenty years later, the conclusion of Hummon and Carley (1993) that the field has its own specialty journal.

It is also obious that, for instance, the *American Journal of Sociology* appears under various labels. While AM_J_SOCIOL and AMER_J_SOCIOL are clearly visible, AM_J_SOCIOLOGY did not make it into the giant component. The set of journals citing AM_J_SOCIOL contains all journals citing any of the variants, though, and apparently the fairly large share of articles associated to AMER_J_SOCIOL is strongly coupled with those associated to AM_J_SOCIOL.

6 Conclusion

We proposed a method to visualize citation patterns in bibliographic data and applied it to explore a specific data set on publications about social network research.

The method uses shortest-path computations to turn given dissimilarities into a metric before subjecting them to classical MDS. Due to our recent efficient approximation technique, this rather generic approach can be applied to other and much larger data sets as well. For the present paper, layouts were computed in a matter of seconds using the implementation in a software tool called visone. [4]

It is most important, however, to be aware that the analysis we presented here is specific to the SN5 data set. As was hinted at several times, there are numerous pitfalls in quantitative bibliometric analysis. While some useful and plausible insight can be gained from the exploration, there are almost necessarily inconsistencies and inhomogeneities in the data. Together with the need for boundary specification, this may cause even those signals that appear to be strongest to be distorted and thus prone to misinterpretation.

Our exploration of the field of social network research based on the SN5 data set is but one illustration. A serious analysis would require much more care and use of domain knowledge.

References

Borg, I. and Groenen, P. (2005). *Modern Multidimensional Scaling.* Springer.

Börner, K., Chen, C., and Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology*, 37(1):179–255.

Brandes, U. and Pich, C. (2006). Eigensolver methods for progressive multidimensional scaling of large data. In *Proc. Graph Drawing.*

Brandes, U. and Willhalm, T. (2002). Visualization of bibliographic networks with a reshaped landscape metaphor. In *Proc. Joint Eurographics - IEEE TCVG Symposium on Visualization*, pages 159–164.

[4] Freely available from http:/www.visone.info/.

Chen, T. T. and Hsieh, L. C. (2007). On visualization of cocitation networks. In *Proc. Information Visualization*, pages 470–475.

Cox, T. and Cox, M. (2001). *Multidimensional Scaling.* CRC/Chapman and Hall.

Hamersa, L., Hemerycka, Y., Herweyersa, G., Janssena, M., Ketersa, H., and Rousseau, R. (1989). Similarity measures in scientometric research: The Jaccard index versus Salton's cosine formula. *Information Processing and Management*, 25(3):315–318.

He, Y. and Hui, S. C. (2001). Mining a web citation database for author co-citation analysis. *Information Processing and Management*, 38(4):491–508.

Hummon, N. P. and Carley, K. (1993). Social networks as normal science. *Social Networks*, 15:71–106.

Hummon, N. P. and Doreian, P. (1989). Connectivity in a citation network: the development of DNA theory. *Social Networks*, 11:39–63.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, 14:10–25.

Lawrence, P. A. (2007). The mismeasurement of science. *Current Biology*, 17(15):R583–R585.

Lazer, D., Mergel, I., and Friedman, A. (2009). Co-citation of prominent social network articles in sociology journals: The evolving canon. *Connections*, 29(1):43–64.

Leydesdorff, L. (1998). Theories of citation? *Scientometrics*, 43(1):5–25.

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

Small, H. (1973). Cocitation in the scientific literature: A new measure of the relationship between two document. *Journal of the American Society for Information Science*, 24:265–269.

Small, H. (1999). Visualizing science by citation mapping. *Journal of the American Society for Information Science*, 50(9):799–813.

Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323.

Todd, P. A. and Ladle, R. J. (2008). Hidden dangers of a 'citation culture'. *Ethics in Science and Environmental Politics*, 8(1):13–16.

Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and Method. *Psychometrika*, 17:401–419.

Weingart, P. (2005). Impact of bibliometrics upon the science system: Inadvertent consequences? *Scientometrics*, 62(1):117–131.

White, H. D. and Griffith, B. C. (1981). Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, 32:163–172.

White, H. D. and McCain, K. W. (1989). Bibliometrics. *Annual Review of Information Science and Technology*, 24:119–186.

White, H. D. and McCain, K. W. (1998). Visualizing a discipline: An author co-citation analysis of information science, 1972-1995. *Journal of the American Society for Information Science*, 49(4):327–355.