Fast Multidimensional Scaling of Network Distances and its Application to Large Bibliographic Networks

Ulrik Brandes and Christian Pich

Department of Computer & Information Science, University of Konstanz, Germany
{Ulrik.Brandes,Christian.Pich}@uni-konstanz.de

Abstract. The visual analysis of similarity, proximity, and distance in networks is a frequent task in many areas of application in which relational data occur. Often the large size of the involved data sets requires efficient and scalable computational tools. In this article we describe a method for the visualization of the proximity and distance in large networks, Based on classical multidimensional scaling, it converts network proximity into geometric proximity. We illustrate the use and efficiency of this method with an application to large bibliographic networks with thousands or tens of thousands of items and present some results.

1 Introduction

The applications in which data objects are related to or affiliated with other objects are countless. Networks are the most frequent model for this kind of data, and their analysis often relies on good visualizations, which can be helpful for understanding structural properties and designing further analytic tasks. For example, visualization can support or discard hypotheses about the network data with respect to clusters, patterns, linking behavior, importance analysis, and the underlying processes leading to an observed network.

Among the most popular analytic aspects of network data are proximity, similarity, and distance. Consequently, they are often the basis for network visualizations. While there is a plethora of such approaches, many of them are inherently limited to small or medium data sets and impractical for larger ones.

In this article we present an efficient method for the layout and visual analysis of networks, which scales even to very large input data sets. The fundamental notion is that of proximity among items in the network to be laid out. This network proximity, whose concrete definition depends on the particular network at hand, is converted into geometric proximity by positioning objects on a two-dimensional drawing area in a suitable way.

To illustrate the usefulness of our method, we apply it to networks of bibliographic objects and associations [15] and their visualization [3, 4, 12]. Nodes of various types, such as articles, authors, institutions, and journals, are linked by edges of different types, such as citations, authorship, and affiliation. The example data set is the result of a query to a large scale bibliography database.

2 Multidimensional Scaling of Network Distances

2.1 Network Distances

In our setting the essence of visualizing structural network properties of is to find a good layout, i.e., to assign two-dimensional positions to the involved objects. Our method constructs the positions in such a way that objects are proximate in the layout if they are proximate or similar in the network, and distant in the layout if they are distant or dissimilar in the network.

The definition of proximity found in most related works is *graph-theoretical distance*, i.e. the length of shortest paths. In a graph $G = (\mathcal{V}, \mathcal{E})$ of nodes $\mathcal{V} = \{v_1, \ldots, v_n\}$, the distance between nodes v_i and v_j is the minimum number of edges that have to be traversed on a path to get from v_i to v_j or vice versa. We will denote the distances as d_{ij}. The distances are easily computed using standard graph algorithms, such as breadth-first search [5].

Sometimes the edges are assigned *weight* labels $w_{ij} > 0$, possibly representing length, strength, or reliability. The definition of distance between nodes v_i and v_j is then slightly modified to be the minimum sum of edge weights on any path from v_i to v_j. Unweighted graphs can thus be considered weighted graphs with all edge weights equal to one. Distances in weighted graphs are computed, e.g., by the Dijkstra algorithm.

All networks visualized in this article are undirected, with the consequence that the distances satisfy the metric properties, namely, for all $v_i, v_j, v_k \in \mathcal{V}$, $d_{ij} \geq 0$, $d_{ii} = 0$, $d_{ij} = d_{ji}$ (symmetry), and $d_{ij} \leq d_{ik} + d_{kj}$ (triangle inequality). The metric properties of graph-theoretic distances in undirected graphs give rise to the following method of performing the abovementioned conversion from network distances to geometric distances.

2.2 Multidimensional Scaling

The basic visualization step is the transformation of network proximity into geometric proximity. In our method, this transformation is based on multidimensional scaling (MDS), a popular family of data analysis techniques turning input distances into a two-dimensional configuration in such a way that the resulting Euclidean distances in these configurations correspond to the input dissimilarities as closely as possible. Our visualization method is based on the classic variant of multidimensional scaling; we will give a short review and refer the reader to the standard references [1, 6] for a comprehensive overview.

For the sake of simplicity, we write the input distances d_{ij} as entries of a square matrix $D \in \mathbb{R}^{n \times n}$. In the two-dimensional case MDS seeks 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 $\|p_i - p_j\| = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$ fit the input distances d_{ij} as closely as possible. While there are various ways of achieving this objective, e.g. by numeric minimization of an objective function, we will concentrate on the earliest approach, which is therefore termed *classical multidimensional scaling* [13].

Basically, classical multidimensional scaling considers input distances as Euclidean, and looks for the two principal Euclidean dimensions that reconstruct as much of the variability of the observed distances as possible, by performing spectral analysis of matrices associated with the network. More specifically, it is equivalent to the computation of dominant eigenvalues and associated eigenvectors, an elementary technique from linear algebra which is closely related to principal component analysis [9]. It is done in four main steps:

1. For each pair (v_i, v_j) of nodes the distance d_{ij} is computed. The squared distance d_{ij}^2 is stored in matrix $B \in \mathbb{R}^{n \times n}$.
2. B is doubly-centered by subtracting from each entry the column and row means and adding the overall mean, so that each row and each column sums to zero.
3. The two dominant eigenvalues $\lambda_1, \lambda_2 \in \mathbb{R}$ and eigenvectors $u_1, u_2 \in \mathbb{R}^n$ (without loss of generality, $\|u_1\| = \|u_2\| = 1$) of B are computed. A simple and effective technique is power iteration [7].
4. Positions are obtained by setting $x = \sqrt{\lambda_1} u_1, y = \sqrt{\lambda_2} u_2$.

2.3 Fast Multidimensional Scaling

The positions thus obtained are reproducible and essentially unique. A drawback of classical multidimensional scaling is that the square matrix of distances between all $\binom{n}{2}$ pairs has to be built and decomposed, which is impractical for larger values of n above a few thousands – in which case the sheer size of the matrix exceeds the main memory even of high-end computers, making the method prohibitive.

This problem is solved by a speed-up technique called Pivot MDS [2]; instead of the $n \times n$ distance matrix, it suffices to compute distances from a small number $k \ll n$ (usually, $k \approx 100$ independently of n works reasonably well). The eigenvectors of a square matrix are replaced by singular vectors of a rectangular $n \times k$, which can be obtained significantly faster. Thereby the classical MDS variant is made scalable to even very large to citation databases and query results with several thousands of entries; for medium-sized instances, the computation is sped up by orders of magnitude, and even rendered possible only by using this Pivot MDS technique.

3 Bibliographic Networks

Citations and references in articles and other research contributions are of increasing interest in the structural analysis of scientific literature. Quantity and origin of the citations an article receives serve as important proxies for its influence on subsequent work; they are often used for assessing the impact and the evolution of this article, its authors, its authors' institutions, the journal in which it appears, and even its entire scientific discipline.

This section describes the underlying bibliographic data model for this type of network: The *objects*, which correspond to entities in bibliographic data; the basic *relationships* among these objects, as they are stored in bibliographic databases; and the bibliographic *operators* that transform relationships into other relationships dedicated to different analytic perspectives on the data.

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

3.1 Objects

Bibliographies are expressed in terms of sets of *objects* of different types, together with *relationships* between them. In this article the model and the bibliographic analysis encompasses sets of

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

This basic model may be extended by information about affiliations of authors to institutions, scientific disciplines, keywords attached to works, etc., but we will restrict the description to these three types.

3.2 Basic Relationships

The links connecting bibliographic objects, e.g. citations and affiliations, are conveniently modeled as matrices. Basic relationships are stored in bibliographic databases and may be obtained directly without additional effort.

In our bibliographic analysis we distinguish three principal relationship types and thus network types. Basic relationships are modeled as unweighted graphs associated with adjacency matrices having entries 1 if the two corresponding objects are related, and 0 otherwise. Figure 1 depicts the objects and basic relationships in our bibliography model.

- *Citations* among of works are stored in a square adjacency matrix $C \in \{0,1\}^{n \times n}$, representing 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 inherently acyclic, since cited works have to be published previously.
- *Authorship* is expressed in a rectangular matrix $A \in \{0,1\}^{m \times n}$. It represents a bipartite graph containing 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* is information about journals in which works are published; institutions to which authors are affiliated; or scientific disciplines assigned to journals. Assuming that every work appears in exactly one journals, every row of matrix $G \in \{0,1\}^{n \times p}$ contains exactly one 1.

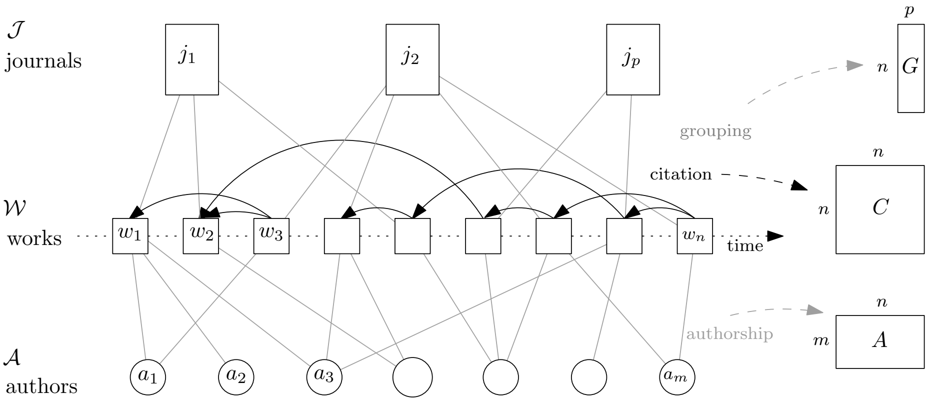


Fig. 1. The basic data model for bibliographic networks. The matrices on the right represent basic relationships: A bipartite journals–works graph; a directed acyclic graph of citations among works; and a bipartite works–authors graph.

3.3 Bibliographic Operators

Bibliographic operators combine the previously introduced basic relationships and transform them into a new relationships. In terms of matrices, our operators correspond to multiplication of the involved matrices.

In the following, we name the some frequent bibliographic operators, which will later be used for visualization. Each operator reflects a particular analytic perspective on the basic relationships in the original bibliographic data. Note that the resulting matrix products are themselves matrices representing graphs.

- *Bibliographic Coupling* between two works occurs if they have cited references in common [10]. The ij entry in $C^TC \in \mathbb{N}^{n\times n}$ is the number of common references of works w_i, w_j.
- *Co-Citation* is the dual to bibliographic coupling, measuring for two works w_i, w_j how many other works cite both of them, given in matrix $CC^T \in \mathbb{N}^{n\times n}$ [11].
- *Collaboration* counts the number of common works of two authors. $A^TA \in \mathbb{N}^{m\times m}$ is a well-known operator, giving the number of times two authors have collaborated in a common work.
- *Projection* Instead of citations of works, as given in matrix $C \in \{0,1\}^{n\times n}$, the analysis might be targeted at citations of authors [14]; the required information is obtained by the matrix product $ACA^T \in \mathbb{N}^{n\times n}$, which aggregates information about works to the affiliated authors.

In an optional post-processing step, it might be useful to discard, e.g., the counts of co-authorship, and to replace all positive entries in the resulting matrix products simply by ones, or to normalize entries such that rows sum to one, etc.

3.4 Bibliographic Proximity

Bibliographic relationships, as defined in Section 3, are the basis for calculating proximity. Traditionally, visualizations use similarity measures of binary attribute sets, such as the relative overlap of the number of citations [8, 16]. In the visualization setting which we will describe later, similarity measures have some shortcomings: Their typical range from 0 to 1 often leads to degenerate visualizations, since many of them assess works with no common references as most dissimilar (similarity 0). Moreover, there is no clear way of converting similarity into distances, which is crucial for our visualization method.

We will overcome these drawbacks by replacing the commonly used similarity of neighborhood sets with the graph-theoretical distances defined in Section 2. These distances are later the basis of the visual representation. Here and later in Section 4 we focus on bibliographic coupling, but the definitions are carried over in a straightforward way to the other operators defined above.

Coupling distance between works w_i, w_j is defined as the graph-theoretic distance between w_i and w_j in the graph corresponding to the co-citation matrix CC^T, i.e., the minimum number of coupling edges that have to be followed to get from w_i to w_j.

Coupling strength If desired, large numbers of common citations as given in matrix CC^T are emphasized by replacing uniform edge lengths in the co-citation graph with weights depending on the strength of co-citation. Let $\mathcal{W}_i$ denote the set of works cited by w_i. Because more common citations mean a smaller co-citation distance between two works, it makes sense to shorten the edge (w_i, w_j) by the *absolute frequency* $1/|\mathcal{W}_i \cap \mathcal{W}_j| = 1/(CC^T)_{ij}$, i.e., by regarding two works as more proximate if the absolute frequency of common citation is large.

Alternatively, edge lengths may be combined with the abovementioned similarity measures, such as the *relative frequency* $|\mathcal{W}_i \cup \mathcal{W}_j|/|\mathcal{W}_i \cap \mathcal{W}_j|$, whose inverse is referred to as the *Jaccard index.*

Significant ties To make distances more robust against degeneration, e.g. due to standard works cited in almost all other works, considering only significant co-citations is done by discarding “weak ties” whose weight is below a certain threshold. This makes the resulting distances more appropriate while reducing the set of edges, leading to more efficient distance computations in practice. Although thresholding may result in disconnected graphs, which requires every connected component to be analyzed and visualized separately, our experience with real-world citation data sets indicates that the largest connected component contains the vast majority of works.

Note that, especially for uniform edge lengths, it is often more computationally advantageous not to explicitly construct the inherently dense co-citation graph but to compute the distances in it by traversing the original sparse directed citation graph.

4 Application: A Bibliography of Social Network Analysis

We illustrate our approach with an example application to the visualization of citation data in Social Network Analysis literature. In this article we focus on visualizing the analytic aspect of bibliographic coupling, viz, the similarity of citation behavior, within a large set of several thousands of items. This data set, called *SN5*, was prepared by Vladimir Batagelj and is described in other articles in this volume.

4.1 Web Of Science Queries

The data set used in our application was generated from the Web of Science, a commercial citation database run by Thomson Reuters[1]. The WWW interface allows for queries to several thousands of scientific journals, with the usual search criteria, such as author name, title, year, etc. From a set of query results, e.g. all works with a title containing a given string, the query interface allows for retrieval of the resulting sets of citing or cited references, thus constructing bibliographic networks. Combinations of more than one query result set are obtained by union or intersection operators.

The initial query result set $\mathcal{W}$, consisting of a set of works, is extended by all works cited at least once by any work in $\mathcal{W}$. An edge (w_1, w_2) in the corresponding graph is weighted with the inverse of common references of w_1 and w_2, but excluding self-edges. Note that we are only interested in proximities within the original set $\mathcal{W}$, whereas all added works not already contained in $\mathcal{W}$ are not included in the visualization, but only serve to determine the strength of bibliographic coupling.

The graph-theoretic distances in the resulting weighted graph are then input to multidimensional scaling, giving the desired two-dimensional positions. The basic workflow of visualization is depicted in Figure 2.

4.2 Bibliographic Coupling Networks

From the *SN5* data set we have generated and visualized networks describing bibliographic couplings in different domains: Works, authors, and journals. The corresponding networks were obtained after computing the inverse relative frequency and discarding all edges with a relative bibliographic coupling frequency below the thresholds given in the right column. The following table gives the sizes of the largest connected components in 2007, for which we have produced visualizations.

domain	number of nodes	number of edges	threshold
works	5 463	87 528	0.25
authors	10 239	336 777	0.20
journals	1 673	60 263	0.15

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

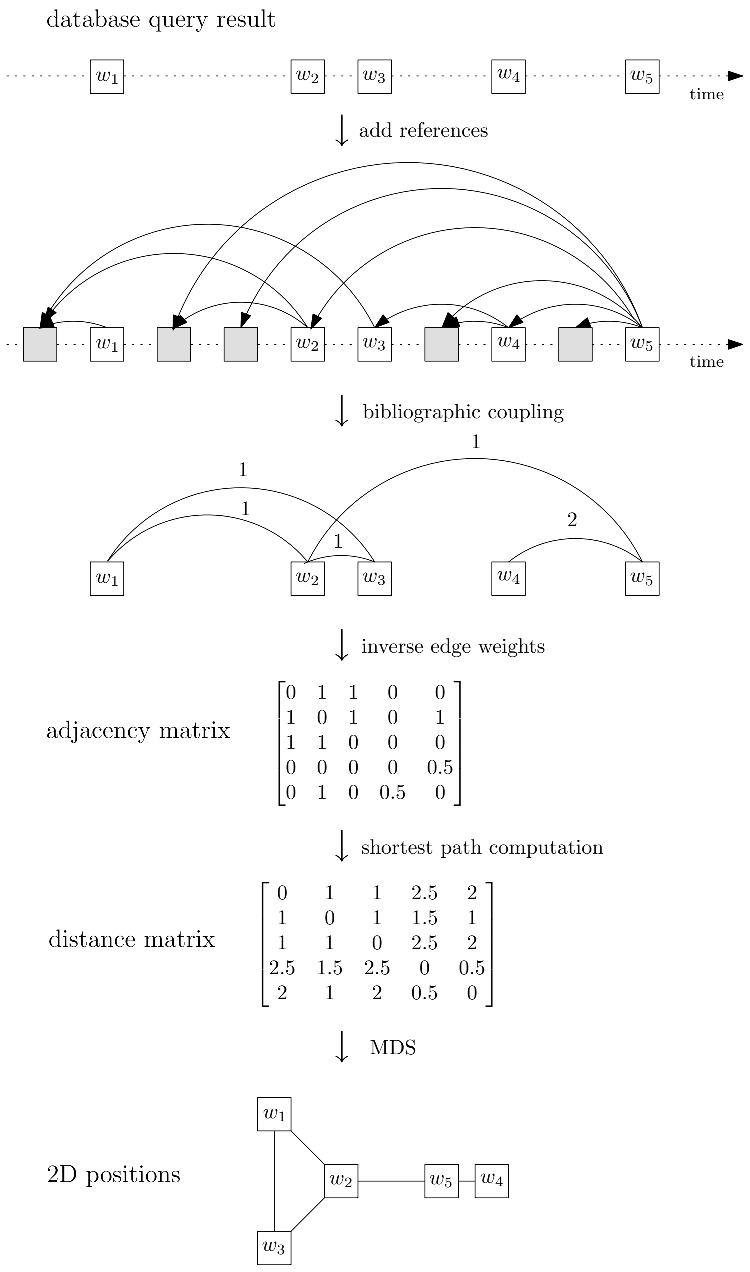

database query result
w_1
w_2
w_3
w_4
w_5
time
add references
bibliographic coupling
1
1
1
1
2
inverse edge weights
adjacency matrix
$\begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.5 \\ 0 & 1 & 0 & 0.5 & 0 \end{bmatrix}$
shortest path computation
distance matrix
$\begin{bmatrix} 0 & 1 & 1 & 2.5 & 2 \\ 1 & 0 & 1 & 1.5 & 1 \\ 1 & 1 & 0 & 2.5 & 2 \\ 2.5 & 1.5 & 2.5 & 0 & 0.5 \\ 2 & 1 & 2 & 0.5 & 0 \end{bmatrix}$
MDS
2D positions

Fig. 2. Basic workflow of the layout for a query result set $\mathcal{W} = \{w_1, \ldots, w_5\}$. The bibliographic coupling graph is generated by adding all cited works not already contained in $\mathcal{W}$ (gray). Using the distances in the resulting weighted bibliographic coupling graph, multidimensional scaling generates 2D positions.

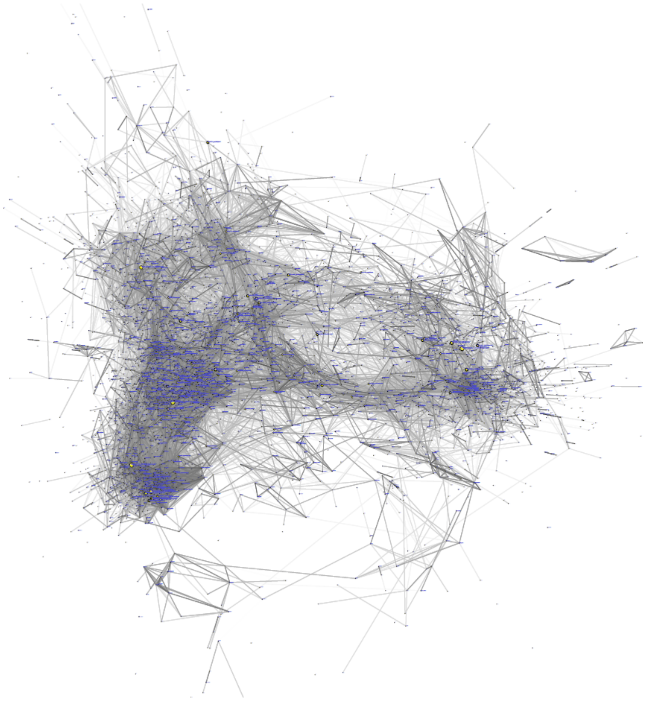


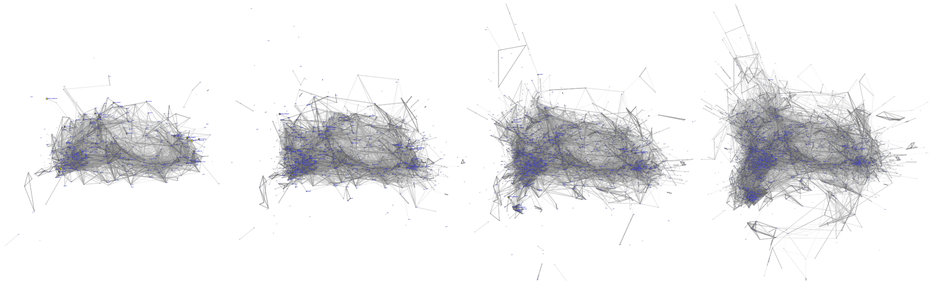
Fig. 3. Bibliographic coupling network among works in the SN5 data set. The area of each work node is proportional to the number of works in which that work is cited.

4.3 Results

In our visualizations we use varying graphical properties of bibliographic coupling edges. A striking spatial impression of visual clusterings is created by displaying and rendering edges after their strength, rendering the strongest couplings last and with thicker and darker edges. The layouts were generated by the fast variant of classical MDS described in Section 2, using $k = 200$.

Figure 3 shows the bibliographic coupling of works in 2007. The central component appears to be split into a left cluster of sociological articles, and a cluster on the right, with works originating in health sciences and psychology. These two clusters are connected by a "bridge" of works, which seem to be, not surprisingly, meta-studies of social network analysis methods from different fields, or works from social psychology. It is interesting to observe the small cluster in the lower left, which mainly contains physicists' works about network analysis.

The evolution of this bibliographic coupling network since 1990 is displayed in Figure 4. While the abovementioned center has stayed stable over the years, in 2000 a new nucleus of a cluster of works by physicists starts to form. While the main component does note change too much, in 2005 the attached physicists' cluster cluster has grown remarkably fast.



(a) 1990 (b) 1995 (c) 2000 (d) 2005

Fig. 4. Dynamic evolution of the bibliographic coupling graph of works, from 1990 to 2005. The layout is done with a flipbook approach, in which node positions reflect bibliographic proximity in 2007 and are fixed for all frames; in each frame the appearance of edges corresponds to the strength of bibliographic coupling at that time.

The bibliographic coupling among the same set of works in 2007, but projected to authors, is shown in Figure 5. Similar to 3 there is an apparent separation into two main clusters, again into rather sociology oriented and rather health or psychology oriented authors; even the peripheral cluster of physicists is present, this time in the lower right.

Figure 6 displays the bibliographic coupling projected to journals. While there is once again a separation into groups quite similar to those in Figure 3 and Figure 5, the sociology journals have a noticeably peripheral position in the upper right part.

It is important to note that all presented visualizations of citations and the resulting bibliographic coupling are specific to the SN5 data set, since only those citations are considered. Therefore the assumptions about works, authors, and journals that were made based on these visualizations might not hold for other data sets.

5 Discussion and Conclusion

We have described a method for the visualization of proximity and distances in large networks. Depending on the particular types of networks and corresponding proximities, the method conveys various structural properties from different analytic perspectives.

Thanks to fast and efficient algorithms, the method also allows for the analysis and visualization of larger networks and overcomes the limitations from which many other methods suffer. Moreover, it is easy to implement, since only basic graph traversals and matrix operations, but no sophisticated data structures, are required.

The presented visualizations of networks derived from bibliography data indicate that the method is successfully applied even to data sets of larger size.

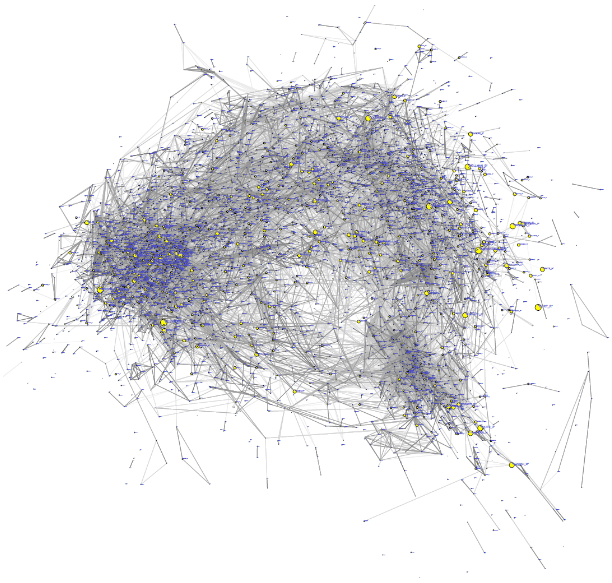


Fig. 5. Bibliographic coupling network among authors in the SN5 data set. The area of each author node is proportional to the number of works in which that author is cited.

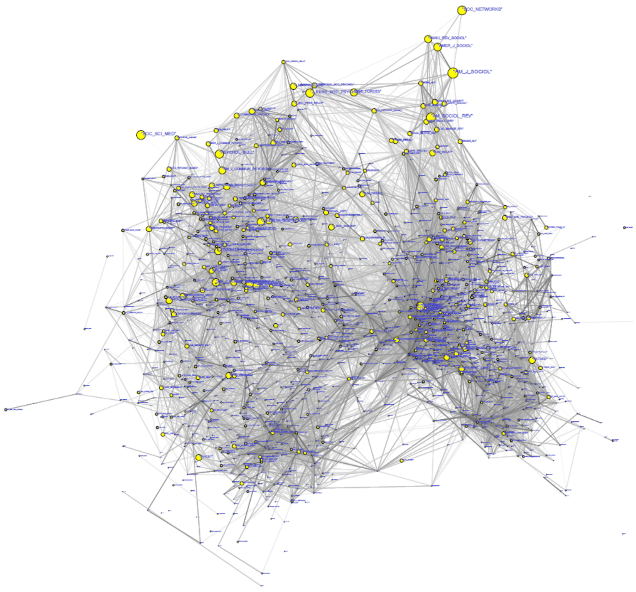


Fig. 6. Bibliographic coupling network among journals in the SN5 data set. The area of each journal is proportional to the number of works in which that journal is cited.

Note that there is no limitation to bibliographic networks, and the presented approach is, usually after some slight modifications, carried over to other types of bipartite or multipartite affiliation networks. Therefore, we hope that our method is not only useful for the visual summary of large bibliography data, but also as a visualization building block in other fields of application.

References

1. I. Borg and P. Groenen. *Modern Multidimensional Scaling.* Springer, 2005.
2. U. Brandes and C. Pich. Eigensolver methods for progressive multidimensional scaling of large data. In *Proc. Graph Drawing*, 2006.
3. Ulrik Brandes and Thomas Willhalm. Visualization of bibliographic networks with a reshaped landscape metaphor. In *Proc. Joint Eurographics - IEEE TCVG Symposium on Visualization*, pages 159–164, 2002.
4. Tsung Teng Chen and Liang Chi Hsieh. On visualization of cocitation networks. In *Proc. Information Visualization*, pages 470–475, 2007.
5. Thomas H. Cormen, Charles E. Leiserson, and Ronald L. Rivest. *Introduction to Algorithms.* MIT Press, 2001.
6. T. Cox and M. Cox. *Multidimensional Scaling.* CRC/Chapman and Hall, 2001.
7. G. H. Golub and C. F. van Loan. *Matrix computations.* Johns Hopkins University Press, 1996.
8. Yulan He and Siu Cheung Hui. Mining a web citation database for author co-citation analysis. *Information Processing and Management*, 38(4):491–508, 2001.
9. I. T. Joliffe. *Principal Component Analysis.* Springer, 1986.
10. M. M. Kessler. Bibliographic coupling between scientific papers. *American Documentation*, 14:10–25, 1963.
11. Henry Small. 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, 1973.
12. Henry Small. Visualizing science by citation mapping. *Journal of the American Society for Information Science*, 50(9):799–813, 1999.
13. W. S. Torgerson. Multidimensional scaling: I. Theory and Method. *Psychometrika*, 17:401–419, 1952.
14. H. D. White and B. C. Griffith. Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, 32:163–172, 1981.
15. Howard D. White and Katherine W. McCain. Bibliometrics. *Annual Review of Information Science and Technology*, 24:119–186, 1989.
16. Howard D. White and Katherine W. McCain. 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, 1998.