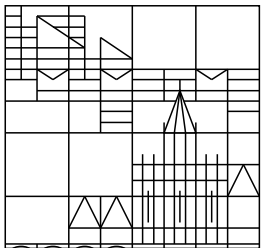


Universität Konstanz

Visualization of Bibliographic Networks with a Reshaped Landscape Metaphor

Ulrik Brandes
Thomas Willhalm

Konstanzer Schriften in Mathematik und Informatik
Nr. 170, April 2002

ISSN 1430–3558

Universität Konstanz
Fach D 188, 78457 Konstanz, Germany
Email: preprints@informatik.uni–konstanz.de
WWW: http://www.informatik.uni–konstanz.de/Schriften/

Visualization of Bibliographic Networks with a Reshaped Landscape Metaphor

U. Brandes and T. Willhalm
Department of Computer & Information Science,
University of Konstanz, Germany.

April 3, 2002

Abstract

We describe a novel approach to visualize bibliographic networks that facilitates the simultaneous identification of clusters (e.g., topic areas) and prominent entities (e.g., surveys or landmark papers). While employing the landscape metaphor proposed in several earlier works, we introduce new means to determine relevant parameters of the landscape. Moreover, we are able to compute prominent entities, clustering of entities, and the landscape's surface in a surprisingly simple and uniform way. The effectiveness of our network visualizations is illustrated on data from the graph drawing literature.

1 Introduction

Bibliographic analysis[24] uses publication data to structure and summarize a scientific field. These data are often given in the form of networks, with nodes representing authors, journals, or publications, and edges representing relations between these entities such as authorship, collaboration, or citation.

We present an approach to analyze and visualize bibliographic networks using uniform algorithms to determine the prominent entities in the network, to spatially represent the clustering of the network, and to compute a surface for a landscape visualization of results.

Since we propose an integrated method of analysis and visualization directed at particular aspects of bibliographic analysis, it may serve as a specialized component in more elaborate systems,[10, 5, 9] and in particular as a communication/exploration back-end for systems that specialize in extracting and presenting network data.[7, 23]

This paper is organized as follows. In Sect. 2 we recall the definition of Kleinberg's hubs & authorities indices[15] and sketch their use in the analysis of bibliographic data. Based on similar principles, a new method for two-dimensional layout of bibliographic networks preserving the scientific topography is presented in Sect. 3. In Sect. 4, index and layout are turned into a landscape visualization, again using the same algorithmic principles. An illustrative example comprised of publications in proceedings of Graph Drawing Symposia is given in Sect. 5.

2 Landmark Papers

To identify prominent entities in bibliographic networks, we determine the structural importance of vertices according to their position in the graph. Many concepts formalizing this notion are in use, but the concept of hubs & authorities,[15] though originally conceived to improve relevance ranking in Web search engines, appears to be particularly suitable for bibliographic networks. In this section, we present an alternative derivation of these indices to emphasize the similarity of their computation with those in later sections. We assume familiarity with basic matrix properties and computations.[12]

A straightforward notion of prominence in undirected graphs, commonly applied in the analysis of social networks,[22] is the idea that the importance of a vertex is determined by the importance of its neighbors. According to the following definition,

the importance assigned to a vertex is proportional to the total importance of its neighbors.

Definition 1 (eigenvector centrality[4]) *Let A be the adjacency matrix of a connected undirected graph $G = (V, E)$. Eigenvector centrality, $c(G) = c = (c_v)_{v \in V}$, is the (unique) solution of*

$$A \cdot c \quad = \quad \lambda \cdot c$$

subject to $c_v > 0$ for all $v \in V$ and $\sum_{v \in V} c_v = 1$, where λ is the (real, positive, and simple) largest eigenvalue of A.

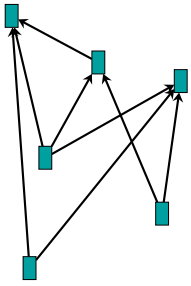
To simplify the presentation, we confine ourselves to the analysis of connected citation networks with respect to landmark publications. We thus consider as basic input connected directed graphs $G = (V, E)$, in which vertices $v \in V$ represent a publication, and directed edges $e = (u, v) \in E$ represent a citation of v in u. With straightforward modifications, our methods can be applied to other types of bibliographic networks and other types of analyses targeted, e.g., at surveys, prominent authors, or journals with high impact.

Two operators modeling two different aspects of positions in the directed graph are defined to transform it into a weighted undirected graph suitable for eigenvector centrality analysis. See Fig. 1 for an illustration.

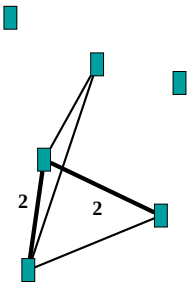
Definition 2 (bibliographic coupling[14] & co-citation[19]) *Let $G = (V, E)$ be a directed graph with adjacency matrix A. The weighted undirected graphs $\mathcal{B}(G)$ and $\mathcal{C}(G)$ induced by adjacency matrices $B = AA^T$ and $C = A^TA$ are called the* bibliographic coupling *and* co-citation *graph, respectively.*

It is interesting to note that bibliographic coupling of a bipartite graph in which vertices represent authors or publications, with edges from authors to their publications, yields a collaboration graph.

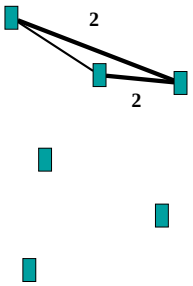
Designed to increase the effectiveness of Web search engines, hubs & authorities are formal notions of structural prominence of vertices in directed graphs. Intuitively, a Web page is considered a hub, if it links to many authorities, and a resource is an authority, if many hubs link to it. The implicit assumptions about the meaning of a



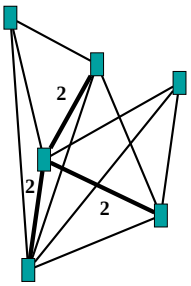
(a) citation network G



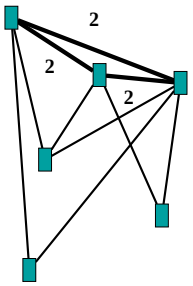
(b) coupling $\mathcal{B}(G)$



(c) co-citation $\mathcal{C}(G)$



(d) coupling similarity graph $\mathcal{S}_{\mathcal{B}}(G)$



(e) co-citation similarity graph $\mathcal{S}_{\mathcal{C}}(G)$

Figure 1: Operators transforming a citation network into weighted undirected graphs representing the essence of certain analytic perspectives.

link are generally the same as the ones made for citations. In fact, the Web can be considered the largest citation network there is.

Definition 3 (hubs & authorities[15]) *For a connected directed graph $G = (V, E)$, let B and C denote the adjacency matrices of $\mathcal{B}(G)$ and $\mathcal{C}(G)$, respectively. The* hub index, *$h(G) = h = (h_v)_{v \in V}$, and the* authority index, *$a(G) = a = (a_v)_{v \in V}$, are defined by*

$$\begin{aligned} B \cdot h &= \lambda_h \cdot h \\ C \cdot a &= \lambda_a \cdot a \end{aligned}$$

subject to $h_v, a_v > 0$ for all $v \in V$ and $\sum_{v \in V} h_v = \sum_{v \in V} a_v = 1$, where λ_h and λ_a are the (real, positive, and simple) largest eigenvalues of B and C, respectively.

Hubs & authorities are thus eigenvector centralities in the weighted undirected graphs constructed from a directed graph by means of bibliographic coupling and co-citation, i.e. $h(G) = c(\mathcal{B}(G))$ and $a(G) = c(\mathcal{C}(G))$. Starting from $a^{(1)} \leftarrow \frac{1}{n} \cdot \mathbf{1}$, the following interleaved version of power iteration is used to compute the indices without explicitly constructing the undirected graphs:

$$\begin{aligned} h^{(k)} &\leftarrow A \cdot a^{(k)} \\ h^{(k)} &\leftarrow h^{(k)} / \|h^{(k)}\| \\ \\ a^{(k+1)} &\leftarrow A^T \cdot h^{(k)} \\ a^{(k+1)} &\leftarrow a^{(k+1)} / \|a^{(k+1)}\| \end{aligned}$$

for $k > 0$, where n is the number of vertices in G. While the speed of convergence depends on the ratio between the largest and second-largest eigenvalue, convergence is usually rapid and we use stabilization of the eigenvalue approximation as our stopping criterion. Since bibliographic networks tend to be very sparse, with the number of edges linear in the number of vertices, each iteration takes time linear in the number of vertices in general.

3 Topics

We next describe a method to compute a two-dimensional positioning of the vertices of a bibliographic network that represents thematic clusters geometrically, but is technically very similar to the iterative computation of a prominence vector in the previous section.

The prominence analysis carried out in the previous section is based on an undirected graph in which weighted edges correspond to the extend of bibliographic coupling (hubs) or co-citation (authorities). Weights thus reflect similarity of entities with respect to the analytic perspective taken. However, if two vertices in a directed graph G are connected by just a single edge, they are adjacent in neither $\mathcal{B}(G)$ nor $\mathcal{C}(G)$. To incorporate similarity implicit in directed linkages, our definition of similarity contains an additional unit weight for each directed edge.

Definition 4 (similarity graphs) *Let $G = (V, E)$ be a directed graph with adjacency matrix A. The weighted undirected graphs $\mathcal{S}_\mathcal{B}(G)$ and $\mathcal{S}_\mathcal{C}(G)$ induced by adjacency matrices $S_B = AA^T + A + A^T$ and $S_C = A^T A + A + A^T$ are called* similarity graphs *with respect to bibliographic coupling and co-citation, respectively.*

Similarity graphs may be clustered geometrically using standard methods such as multidimensional scaling or force-directed graph layout algorithms. However, with these approaches optimum solutions are hard to obtain, and algorithms typically get stuck in local optima of varying quality. We therefore opt for spectral layout methods. We remark, though, that in comparison with other approaches,[10] our similarity graphs are special, and both the way we compute eigenvectors and the technique to avoid well-known defects of spectral layouts are different.

Spectral layout refers to the use of eigenvectors of graph-related matrices for positioning the vertices of the graph. The following matrix has fascinating applications in diverse areas.[17]

Definition 5 (Laplacian matrix) *Let $G = (V, E)$ be a (weighted) undirected graph with adjacency matrix A, and let D be the diagonal matrix of (weighted) degrees. The matrix $L(G) = L = D - A$ is called the* (weighted) Laplacian matrix *of G.*

Let us recall some fundamental facts about the Laplacian spectrum from algebraic graph theory.[11]

Lemma 1 *Let L be the Laplacian matrix of a (weighted) undirected graph G. The eigenvalues of L are non-negative real numbers, the smallest being zero (with multiplicity one if G is connected), and the largest being bounded by twice the maximum degree in G. Any two eigenvectors of L are either collinear or orthogonal, and the entries of an eigenvector associated with eigenvalue zero are all equal.*

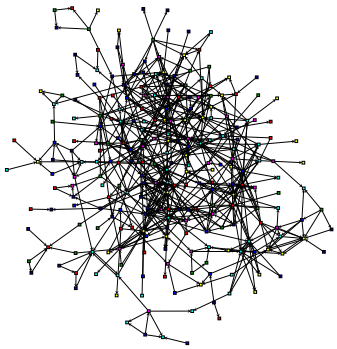
The reason for using eigenvectors of the Laplacian matrix for graph layout, in particular those associated with small eigenvalues, is the following. The value of the quadratic form $(x^T Lx)/(x^T x) = \sum_{e=\{u,v\}\in E} \omega_e \cdot (x_u - x_v)^2$ where ω_e is the weight of edge e, is called the *stress* resulting from x. The non-trivial eigenvectors of L are orthogonal to the trivial minimizer $\mathbf{1}$, i.e. centered around the origin, and their resulting stress is the associated eigenvalue of L. Therefore, pairwise orthogonal eigenvectors associated with the smallest non-zero eigenvalues yield balanced layouts of minimum stress.

If the underlying graph is not "round-shaped" (roughly, if the second-smallest eigenvalue is not large enough), Laplacian layouts yield clusterings which are too dense to be useful for visualization. This defect is well-known, and it has been suggested to use the Laplacian layout only to initialize a force-directed layout algorithm[10] which, however, results in significantly increased running times.

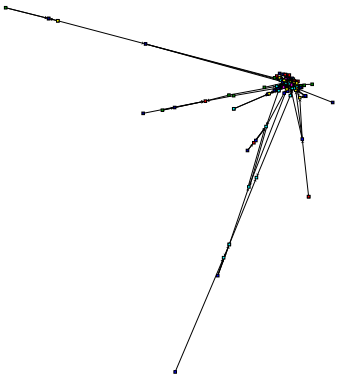
Instead, we propose to modify the Laplacian matrix by introducing a relaxation factor $0 \leq \rho \leq 1$. The matrix $L_\rho = (1-\rho) \cdot D - A$ compromises between the Laplacian and the adjacency matrix and thus avoids excessive displacement of loosely connected vertices. Figure 2 illustrates the effect.

To be able to compute eigenvectors of L_ρ with the same simple power iteration used for hubs & authorities, we reverse the order of its eigenvalues and repeatedly orthogonalize with $\mathbf{1}$.[6] Moreover, because of the potential loss of sparsity, we do not construct the similarity graphs explicitly, but proceed back and force along edge directions as in the computation of hubs & authorities.

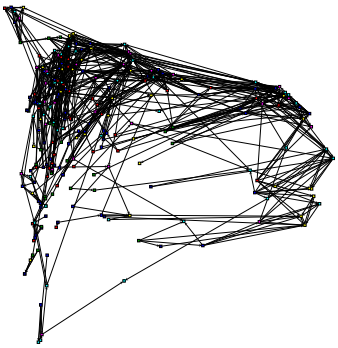
To compute a similarity clustering with respect to, say, co-citation, let A be the adjacency matrix of a directed graph G with n vertices, $D_{\mathcal{S}_\mathcal{C}(G)}$ the diagonal weighted degree matrix of $\mathcal{S}_\mathcal{C}(G)$, and Δ the maximum weighted degree of $\mathcal{S}_\mathcal{C}(G)$.



(a) Spring embedding of citation network G



(b) Laplacian layout of co-citation graph $\mathcal{S}_\mathcal{C}(G)$



(c) modified Laplacian layout ($\rho = 0.25$) of $\mathcal{S}_\mathcal{C}(G)$

Figure 2: Co-citation in citation network of Sect. 5; note that the Laplacian layouts are not primarily determined by citations, but the similarity of citation patterns.

$$
\begin{aligned}
x^{(k+1)} &\leftarrow A \cdot x^{(k)} \\
x^{(k+1)} &\leftarrow A^T \cdot x^{(k+1)} + (A + A^T) \cdot x^{(k)} \\
x^{(k+1)} &\leftarrow x^{(k+1)} + (2\Delta \cdot I - (1-\rho) D_{\mathcal{S}_C(G)}) \cdot x^{(k)} \\
x^{(k+1)} &\leftarrow x^{(k+1)} - \frac{1}{n} \sum_{v \in V} x_v^{(k+1)} \\
x^{(k+1)} &\leftarrow x^{(k+1)} / \|x^{(k+1)}\|
\end{aligned}
$$

A second dimension, y, is computed in much the same way, except that we orthogonalize with the first dimension by computing

$$y^{(k+1)} \leftarrow y^{(k+1)} - \frac{x^T \cdot y^{(k+1)}}{x^T \cdot x} x$$

at the end of each iteration. Again, we require only sparse matrix-vector and vector-vector multiplications, so that each iteration needs linear time and space.

4 Scientific Landscapes

The landscape metaphor is popular for visualizing bibliographic networks,[8, 10, 9] but in general the landscape is produced simply by overlaying a triangulated grid, where grid points are elevated according to the density of data points in their vicinity. The shape of the landscape thus conveys only one aspect in the network's analysis, namely clustering.

We define the shape of the landscape so as to display both clustering and prominence in the same visualization and to represent the underlying network structure more accurately. Intuitively speaking, we simplify a three-dimensional drawing of the network (in which two dimensions represent similarity between entities and the third is determined by a prominence index) by placing a table cloth over it. We next show how this table-cloth can be positioned with yet another variation of the iterative procedure used in the previous sections.

Assume we are given a connected undirected graph $G = (V, E)$ with n vertices and m edges together with a three-dimensional layout (x, y, z), in which each $v \in V$ is associated with a point $(x_v, y_v, z_v) \in \mathbb{R}^3$. In our particular application, x- and y-coordinates are the entries of eigenvectors of the modified Laplacian matrix of G, and z-coordinates are eigenvector centralities in G, i.e. $z = c(G)$, but a landscape could be generated in much the same way from any other three-dimensional layout as well.

We want to cover the layout from the top (z-direction) with a smooth surface to resemble a landscape in which elevations correspond to prominent entities. We therefore first generate a point set in the xy-plane, triangulate it, and finally compute z-coordinates for all points using this triangulation and the prominence of vertices.

The set of points defining the shape of the landscape is generated as follows. Consider the two-dimensional straight-line drawing of G defined by (x, y), and add $\Omega(\sqrt{n})$ equidistant horizontal and vertical lines each to the drawing. The set of points P that defines the landscape consists of all vertices of G and all intersections (between edges, grid lines, or edges and grid lines) thus created. Since $|P| \in \mathcal{O}(m^2)$, it may be desirable for very large graphs to reduce the number of points (at the cost of resemblance quality) by ignoring those induced by edges that cross other edges or grid lines.

Next, a Delaunay triangulation of P is computed, the resulting triangles of which are later used to render the surface. This triangulation may be restricted to include edges and grid lines.

It remains to determine z-coordinates for all $p \in P$ such that the surface covers the three-dimensional graph layout like a table cloth. Ideally, points created from vertices of the graph are placed at the z-coordinate of that vertex. On the other hand, for the surface to be smooth, points that are close in the xy-plane should also be close in z-direction. Hence consider the objective function

$$\sum_{p \in P} \sum_{q \in P} \omega_{pq} \cdot \|z_p - z_q\|^2$$

where ω_{pq} is a nonnegative weight measuring the influence of q on p, which will depend on the relative distance between them. We set $\omega_{pq} = 0$, if $p = q$ or p and q are not adjacent in the triangulation. Inspired by recent work on terrain modeling,[2, 1] we compute the remaining influence weights from Sibson's interpolant,[18] i.e. by temporarily removing p from the Voronoi diagram and setting ω_{pq} to the share of p's Voronoi cell that its Delaunay neighbor acquires through p's removal.

Minimization of the above objective function is straightforward. Note that it constitutes the

quadratic form associated with a Laplacian matrix, though this time of the triangulation graph with Sibson weights. Moreover, since the surface should cover the three-dimensional shape of the network, we have natural candidates for the z-coordinates of points stemming from a vertex or the intersection of an edge and use them as a lower bound for the elevation. Points on the convex hull (the border of the grid) are fixed to have z-coordinate equal to zero, i.e. at ground level. Subject to these constraints, the remaining coordinates are determined so as to minimize the above objective.

Since some points are already fixed, the minimization amounts to placing all other points in the weighted one-dimensional barycenter of their neighbors. The resulting system of linear equations has a unique solution,[21] which can be approximated quickly using an iterative equation solver. Let F be the edges of the Delaunay triangulation, then we iterate

$$z_p^{(k+1)} \leftarrow \sum_{q\,:\,\{p,q\}\in F} \frac{\omega_{pq}}{\sum_{q'\,:\,\{p,q'\}\in F} \omega_{p,q'}} \cdot z_q^{(k)}$$

for each $p \in P$ whose coordinate has not been fixed. These are once again sparse matrix computations, and since the matrix is weakly diagonally dominant, convergence is rapid.

5 Example

For proof of concept, we have implemented our approach in C++ using the Library of Efficient Data Types and Algorithms (LEDA)[16] and OpenGL, and tested it on a data set taken from the 2001 Graph Drawing Contest.[3] It consists of all papers published in proceedings of Graph Drawing Symposia 1994–2000 together with their mutual citations. The largest connected component is formed by 249 papers and 642 citations. It should be noted that this data cannot form the basis for valid conclusions about the relative importance of papers in the field of graph drawing as such. It was chosen simply because we are most familiar with the document corpus and could therefore evaluate much better the adequateness of our visualizations (relative to the given data set).

Using our reshaped landscape metaphor, the citation network suggests several hypotheses about the nature of citations in the area of graph drawing that are readily confirmed by inspection of the underlying data (see Figs. 3–6 in the color section). Peaks indeed indicate authoritative papers, and villages correspond to themes in graph drawing.

Consider, for instance, the mountain ridge stretching across the far end in Fig. 3. It is made up of subject areas, and peaks correspond to highly relevant papers within these subjects. A clear example are the two peaks on the right, where papers dealing with three-dimensional and orthogonal graph drawing cluster. At the Graph Drawing Symposium, many papers on three-dimensional layout deal with orthogonal representations.

Another interesting observation is the village formed by reports on the graph drawing contest itself (Fig. 4) which is hidden behind mainstream subjects.

Improved graphical design (e.g., richer glyphs), more sophisticated rendering (e.g., increased realism), and comprehensive means of user interaction (e.g., mouse-over labels, levels of detail) would certainly be useful for an actual system, but are beyond the scope of our work. The landscape visualization might further be extended by introducing topical area boundaries (based on implicit surface techniques[20]) or citation tracks (based on main path analysis[13]).

References

[1] M. Bertram and H. Hagen. Subdivision surfaces for scattered-data approximation. In D. Ebert, J. M. Favre, and R. Peikert, editors, *Data Visualization 2001. Proceedings of the 3rd Joint Eurographics and IEEE TCVG Symposium on Visualization (VisSym '01)*, pages 55–63. Springer, 2001.

[2] M. Bertram, S. E. Konkle, H. Hagen, B. Hamann, and K. I. Joy. Terrain modeling using voronoi hierarchies. In G. Farin, H. Hagen, and B. Hamann, editors, *Hierarchical Approximation and Geometrical Methods for Scientific Visualization*. Springer, 2001. To appear.

[3] T. C. Biedl and F. J. Brandenburg. Graph-drawing contest report. In P. Mutzel, M. Jünger, and S. Leipert, editors, *Proceedings of the 9th International Symposium on Graph Drawing (GD '01)*, volume 2265 of *Lecture Notes in Computer Science*, pages 513–522. Springer, 2002.

[4] P. Bonacich. Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, 2:113–120, 1972.

[5] K. Börner, A. Dillon, and M. Dolinsky. LVis – digitial library visualizer. In *Proceedings of the International Conference on Information Visualization (IV 2000)*, pages 77–81. IEEE Computer Society Press, 2000.

[6] U. Brandes and S. Cornelsen. Visual ranking of link structures. In F. Dehne, J.-R. Sack, and R. Tamassia, editors, *Proceedings of the 7th Workshop on Algorithms and Data Structures (WADS '01)*, volume 2125 of *Lecture Notes in Computer Science*, pages 222–233. Springer, 2001.

[7] A. Büggemann-Klein, R. Klein, and B. Landgraf. BibRelEx: Exploring bibiliographic databases by visualization of annotated content-based relations. *D-Lib Magazine*, 5(11), 1999.

[8] M. Chalmers. Using a landscape metaphor to represent a corpus of documents. In A. U. Frank and I. Campari, editors, *Proceedings of the European Conference on Spatial Information Theory (COSIT '93)*, volume 716 of *Lecture Notes in Computer Science*, pages 377–390. Springer, 1993.

[9] C. Chen and R. J. Paul. Visualizing a knowledge domain's intellectual structure. *IEEE Computer*, 34(3): 65–71, 2001.

[10] G. S. Davidson, B. Hendrickson, D. K. Johnson, C. E. Meyers, and B. N. Wylie. Knowledge mining with VxInsight: Discovery through interaction. *Journal of Intelligent Information Systems*, 11(3):259–285, 1998.

[11] C. Godsil and G. Royle. *Algebraic Graph Theory*, volume 207 of *Graduate Texts in Mathematics*. Springer, 2001.

[12] G. H. Golub and C. F. van Loan. *Matrix Computations*. Johns Hopkins University Press, 3rd edition, 1996.

[13] N. P. Hummon and P. Doreian. Connectivity in a citation network: The development of DNA theory. *Social Networks*, 11:39–63, 1989.

[14] M. M. Kessler. Bibliographic coupling between scientific papers. *American Documentation*, 14(1):10–25, 1963.

[15] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the Association for Computing Machinery*, 46(5):604–632, September 1999.

[16] K. Mehlhorn and S. Näher. *The LEDA Platform of Combinatorial and Geometric Computing*. Cambridge University Press, 1999.

[17] B. Mohar. Some applications of Laplace eigenvalues of graphs. In G. Hahn and G. Sabidussi, editors, *Graph Symmetry: Algebraic Methods and Applications*, NATO ASI Series C 497, pages 225–275. Kluwer, 1997.

[18] R. Sibson. A brief description of natural neighbor interpolation. In V. Barnett, editor, *Interpreting Multivariate Data*, pages 21–36. John Wiley & Sons, 1981.

[19] H. Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24:265–269, 1973.

[20] T. C. Sprenger, R. Brunella, and M. H. Gross. H-BLOB: A hierarchical visual clustering method using implicit surfaces. In *Proceedings of 11th Annual IEEE Visualization Conference (Vis 2000)*, pages 61–68. IEEE Computer Society Press, 2000.

[21] W. T. Tutte. How to draw a graph. *Proceedings of the London Mathematical Society, Third Series*, 13:743–768, 1963.

[22] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[23] D. R. White, J. Buzydlowski, and X. Lin. Co-cited author maps as interfaces to digital libraries: Designing pathfinder networks in the humanities. In *Proceedings of the International Conference on Information Visualization (IV 2000)*, pages 25–30. IEEE Computer Society Press, 2000.

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

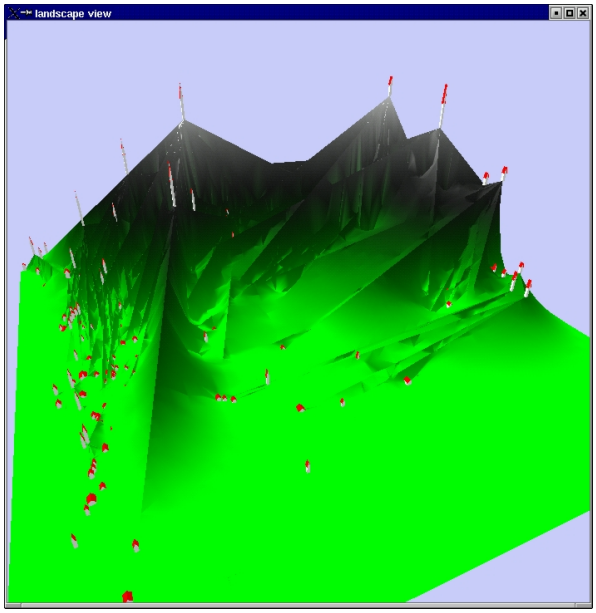


Figure 3: Simultaneous visualization of prominence (authority) and clustering (co-citation similarity) for Graph Drawing Proceedings citation network. Peaks correspond to landmark papers.

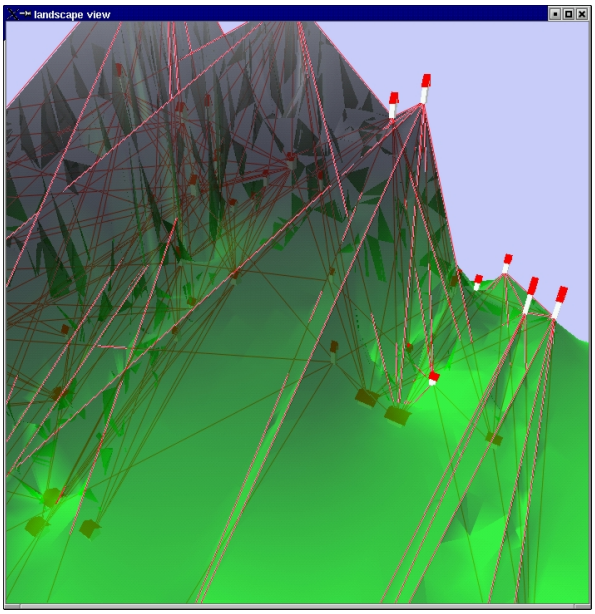


Figure 5: Similar citation patterns lead to close positions (citation edges shown, semi-transparent surface). Height and width of house depict the number of citations received and made.

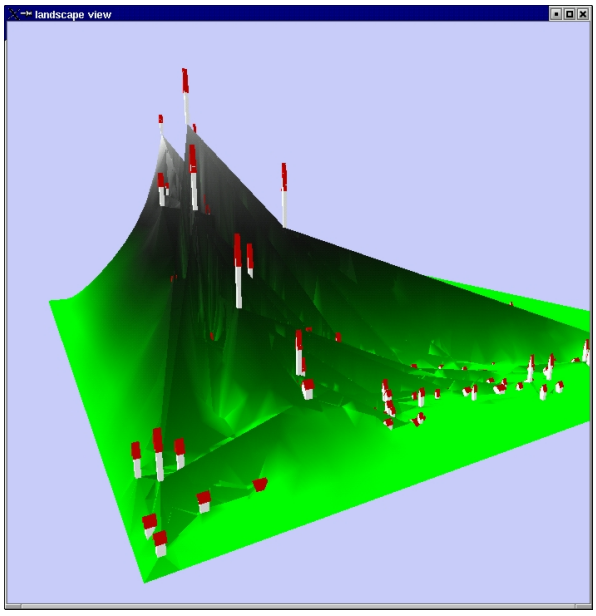


Figure 4: Graph-Drawing Contest Reports form a village hidden behind the mainstream ridge.

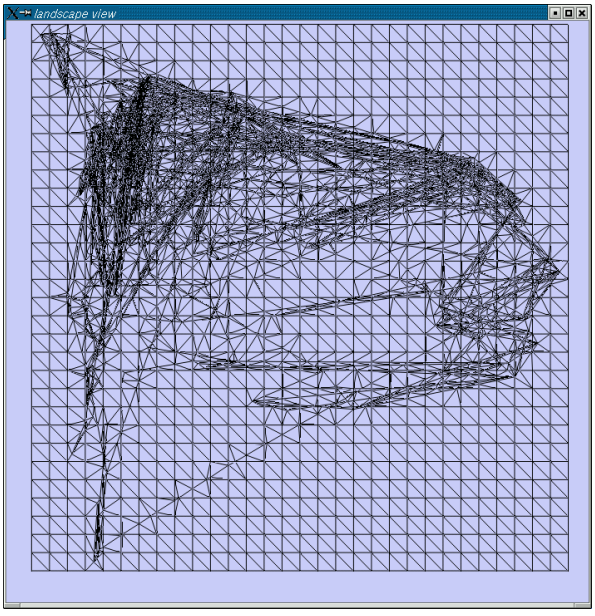


Figure 6: Restricted triangulation refining the layout in Fig. 2(c)