CentiBiN
Centralities in Biological Networks

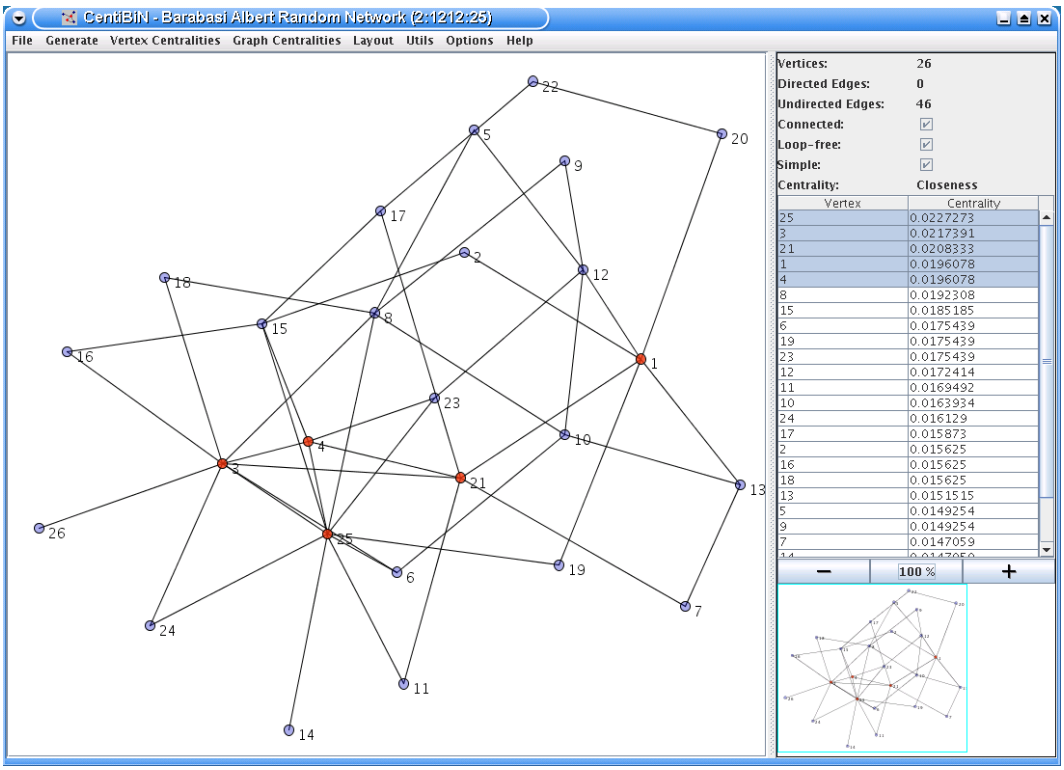
Dirk Koschützki
Research Group Network Analysis
Leibniz Institute of
Plant Genetics and Crop Plant Research,
06466 Gatersleben, Germany

23rd March 2006

1 Idea

CentiBiN allows the centrality analysis of given networks. The networks and the computed centrality measures can be visualised in a graphical user interface. Computed centrality values can be exported easily and used, for example, within R [19] for further analysis.

2 Graphical User Interface



A network under analysis is shown in the left part of the window. In the top right the major properties of the network are displayed. Centrality values of the vertices are shown in the middle on the right side. A birds eye view of the network is shown in the bottom right area.

Vertices can be selected in the list on the right to highlight them within the network view on the left side.

3 File Formats

CentiBiN supports different file formats for reading and writing networks.

3.1 Reading network files

CentiBiN is able to read networks in the Pajek `.net` format (see 3.3.1), DIPs tab separated files (see 3.3.4), Adjacency matrixes given in text files (see 3.3.2) and the GraphML file format (see 3.3.3). To load a network select the menu entry `File/Load Network` or press the keyboard shortcut `Control-o`.

3.2 Writing network files

CentiBiN is able to write networks into the Pajek `.net` format (see 3.3.1), a text file containing the adjacency matrixes (see 3.3.2) and the `.dot` format (see 3.3.5). To save a network select the menu entry `File/Save Network` or press the keyboard shortcut `Control-s`.

3.3 Network file formats

In the following sections the network file formats supported by CentiBiN are described.

3.3.1 Pajek network files

Pajek [3] is a program for network analysis by Vladimir Batagelj and Andrej Mrvar. For this program a file format for the representation of networks is specified. As the format is very simple we give a short example here. More information about the `.net` file format is available for example in the Pajek First Steps (http://vlado.fmf.uni-lj.si/pub/networks/pajek/howto/FirstSteps.PDF). Please note that directed edges are specified after the "arcs" section and undirected edges after the "edges" section.

```
*Vertices 4
1 "Eins"
2 "Zwei"
3 "Drei"
```

```
4 "Vier"
*Arcs
1 2
2 3
*Edges
3 4
```

3.3.2 Adjacency matrixes in text files

Networks can be specified as adjacency matrices. CentiBiN can read such a matrix from a file if the file contains exactly n rows containing n columns. Each column has to contain either a 0 or a 1. If the matrix is symmetric, then a graph containing undirected edges is constructed and in case of an asymmetric matrix a graph containing directed edges ("arcs") is constructed.

```
0 1 0 0
0 0 1 0
0 0 0 1
0 0 1 0
```

Additionally, labels may be included in the first column and first row. They will be used as vertex labels by CentiBiN:

```
"A" "B" "C" "D"
"A" 0 1 0 0
"B" 0 0 1 0
"C" 0 0 0 1
"D" 0 0 1 0
```

3.3.3 GraphML files

The GraphML file format is a XML based graph file format. We do not described this file format in this document and instead refer to the GraphML (http://graphml.graphdrawing.org) web page.

```
<?xml version="1.0" encoding="UTF-8"?>
<graphml xmlns="http://graphml.graphdrawing.org/xmlns"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://graphml.graphdrawing.org/xmlns
     http://graphml.graphdrawing.org/xmlns/1.0/graphml.xsd">
  <graph id="G" edgedefault="directed">
    <node id="n1"/>
```

```
    <node id="n2"/>
    <node id="n3"/>
    <node id="n4"/>
    <edge id="e1" directed="true" source="n1" target="n2"/>
    <edge id="e2" directed="true" source="n2" target="n3"/>
    <edge id="e3" directed="false" source="n3" target="n4"/>
  </graph>
</graphml>
```

3.3.4 DIP tab separated files

The Database of Interacting Proteins (DIP [21]) provides text files specifying interactions between proteins. These files can be downloaded from the DIP web page and imported into CentiBiN. As the file format is specific to DIP we do not explain it in this document.

3.3.5 DOT files

Graphviz [1] is an open source graph visualisation system. This system uses a file format called `DOT` files. We do not explain this format in this document and refer to the web page instead (The DOT Language (http://www.graphviz.org/doc/info/lang.html)).

4 Network Generators

CentiBiN provides methods for the generation of random networks based on five different algorithms such as Kleinberg's small-world generator[16] and the Barabási-Albert scale-free generator[2]. These networks can be analysed and visualised and may be used as reference models.

5 Implemented Centralities

Centralities are used to rank elements of a given network. Currently CentiBiN supports vertex and graph centralities. A vertex centrality assigns values to vertices. We use the convention that a high centrality value is assigned to the "central" vertices according to the centrality under investigation. Graph centralities are not further described in this user guide.

5.1 Preconditions for Centralities

Some of the centrality measures implemented in CentiBiN require that the network is connected, loop free, contains no parallel edges or is not "mixed", i.e., has only directed or undirected edges and not both.

5.1.1 Connected Network

A network is connected if every two vertices are connected by a path. A path is a chain of edges. Therefore a network is connected if we can "walk" from every vertex to every other vertex.

For networks with directed edges the direction of edges has to be considered. That means, that the we can "walk" only on the direction of the edge. If all vertices of a network containing directed edges are connected to each other on a path, then this network is called strong connected.

5.1.2 Loop Free Network

Loops are edges that connect a vertex to itself. Some of the centralities implemented in CentiBiN require that the network is loop free.

5.1.3 Parallel Edges

Between two vertices more than one edge may exist. This additional edge is usually not considered by the computation of centralities. Therefore CentiBiN can compute centralities only on network that are "simple".

5.1.4 Mixed Network

A network is called mixed if it contains both undirected and directed edges.

5.2 Vertex Centralities

CentiBiN supports the following vertex centralities for undirected networks: Degree, Eccentricity, Closeness, Radiality, Centroid Value, Stress, Shortest-Path Betweenness, Current-Flow Closeness, Current-Flow Betweenness, Katz Status Index, Eigenvector Centrality, Hubbell Index, Bargaining Centrality, PageRank, HITS Hubs, HITS Authority, Closeness Vitality

CentiBiN supports the following vertex centralities for directed networks: In-Degree, Out-Degree, Eccentricity, Closeness, Radiality, Centroid Value, Stress, Shortest-Path Betweenness, Katz Status Index, Eigenvector Centrality, Hubbell Index, Bargaining Centrality, PageRank, HITS Hubs, HITS Authority, Closeness Vitality

5.2.1 Definitions of the Centralities

Let $G = (V, E)$ be a undirected or directed, (strong) connected graph with $n = |V|$ vertices. $\deg(v)$ denotes the degree of the vertex v in an undirected graph. $\deg^-(v)$ and $\deg^+(v)$ denote the in- and out-degree in the directed case. $\text{dist}(v, w)$ is the length of a shortest path between the vertices s and t. σ_{st} denote the number of shortest paths from s to t and $\sigma_{st}(v)$ the number of shortest path from s to t that use the vertex v. A denotes the adjacency matrix of the graph G. Please note: References are given only for convenience and are not complete. For a more detailed description and further references please see [17] and for algorithms on computing some of the centralities given see [12].

5.2.2 Degree

$$\mathcal{C}_{deg}(v) := \deg(v)$$

For directed graphs in- and out-degree is used.

5.2.3 Eccentricity

$$\mathcal{C}_{ecc}(v) := \frac{1}{\max\{\text{dist}(v, w) \colon w \in V\}}$$

[10]

5.2.4 Closeness

$$\mathcal{C}_{clo}(v) := \frac{1}{\sum_{w \in V} \mathrm{dist}(v, w)}$$

[20]

5.2.5 Radiality

$$\mathcal{C}_{rad}(v) := \frac{\sum_{w \in V} (\Delta_G + 1 - \mathrm{dist}(v, w))}{n - 1}$$

Δ_G is the diameter of the graph G, defined as the maximum distance between any two vertices of G [24].

5.2.6 Centroid Value

$$\mathcal{C}_{cen}(v) := \min\{\mathrm{f}(v, w) \colon w \in V \setminus \{v\}\}$$

Where $f(v, w) := \gamma_v(w) - \gamma_w(v)$ and $\gamma_v(w)$ denotes the number of vertices that are closer to v than to w [23].

5.2.7 Stress

$$\mathcal{C}_{str}(v) := \sum_{s \neq v \in V} \sum_{t \neq v \in V} \sigma_{st}(v)$$

[22]

5.2.8 S.-P. Betweenness

$$\mathcal{C}_{spb}(v) := \sum_{s \neq v \in V} \sum_{t \neq v \in V} \delta_{st}(v)$$

$$\delta_{st}(v) := \frac{\sigma_{st}(v)}{\sigma_{st}}$$

[8]

5.2.9 C.-F. Closeness

$$\mathcal{C}_{cfc}(v) := \frac{n - 1}{\sum_{t \neq v} p_{vt}(v) - p_{vt}(t)}$$

Where $p_{vt}(t)$ equals the potential difference in an electrical network [7].

5.2.10 C.-F. Betweenness

$$\mathcal{C}_{cfb}(v) = \frac{1}{(n-1)(n-2)} \sum_{s,t \in V} \tau_{st}(v)$$

Where $\tau_{st}(v)$ equals the fraction of electrical current running over vertex v in an electrical network [7].

5.2.11 Katz Status

$$\mathcal{C}_{katz} := \sum_{k=1}^{\infty} \alpha^k (A^T)^k \vec{1}$$

Where α is a (positive) scaling factor [14].

5.2.12 Eigenvector

$$\lambda \mathcal{C}_{eigenvector} = A\mathcal{C}_{eigenvector}$$

The eigenvector to the dominant eigenvalue of A is used [4].

5.2.13 Hubbell index

$$\mathcal{C}_{hubbell} = \vec{E} + W\mathcal{C}_{hubbell}$$

Where $\vec{E}$ is some exogeneous input and W is a weight matrix derived from the adjancancy matrix A [11].

5.2.14 Bargaining

$$\mathcal{C}_{bargain} := \alpha(I - \beta A)^{-1} A\vec{1}$$

Where α is a scaling factor and β is the influence parameter [5].

5.2.15 PageRank

$$\mathcal{C}_{pagerank} = dP\mathcal{C}_{pagerank} + (1-d)\vec{1}$$

Where P is the transition matrix ($P := D^+A$) and d is the damping factor ($d \in [0,1]$) [18].

5.2.16 HITS-Hubs

$$\mathcal{C}_{hubs} = A\mathcal{C}_{auths}$$

Assuming $\mathcal{C}_{auths}$ is known [15].

5.2.17 HITS-Authorities

$$\mathcal{C}_{auths} = A^T \mathcal{C}_{hubs}$$

Assuming $\mathcal{C}_{hubs}$ is known [15].

5.2.18 Closeness-vitality

$$\mathcal{C}_{clv}(v) := \mathrm{WI}(G) - \mathrm{WI}(G \setminus \{v\})$$

Where $\mathrm{WI}(G)$ is the Wiener index of the graph G [17].

5.3 Graph Centralities

Additional to the vertex centralities 5.2 three graph centralities are supported by CentiBiN. The centralities can be computed via the Graph Centralities menu and are displayed in a separate dialog.

The following graph centralities are implemented in this version of CentiBiN:

- Graph Diameter
- Wiener Index
- Average Distance

5.4 Centrality Vectors

CentiBiN supports reading and writing of Pajek vectors containing centrality values. This file may then be used in other tools, e.g. Pajek, for further analysis of the network. Additionally, writing to a tabbed separated file format is supported.

5.4.1 `.vec` File Format

The Pajek `.vec` file format is supported. This file format specifies the number of vertices in the first line and the values for each vertex in the following n lines.

```
*Vertices 4
1.5
2.5
1.5
0.5
```

5.4.2 Tabbed Separated File Format

The computed centrality value can be exported into a file with a header line and n lines. Every line contains the position, the label of the vertex and the centrality value. The columns are separated by a tab stop.

```
Position VertexLabel Value
1 25 12.0
2 3 11.0
3 1 7.0
4 8 6.0
```

5.5 Analysing Centrality Vectors with R

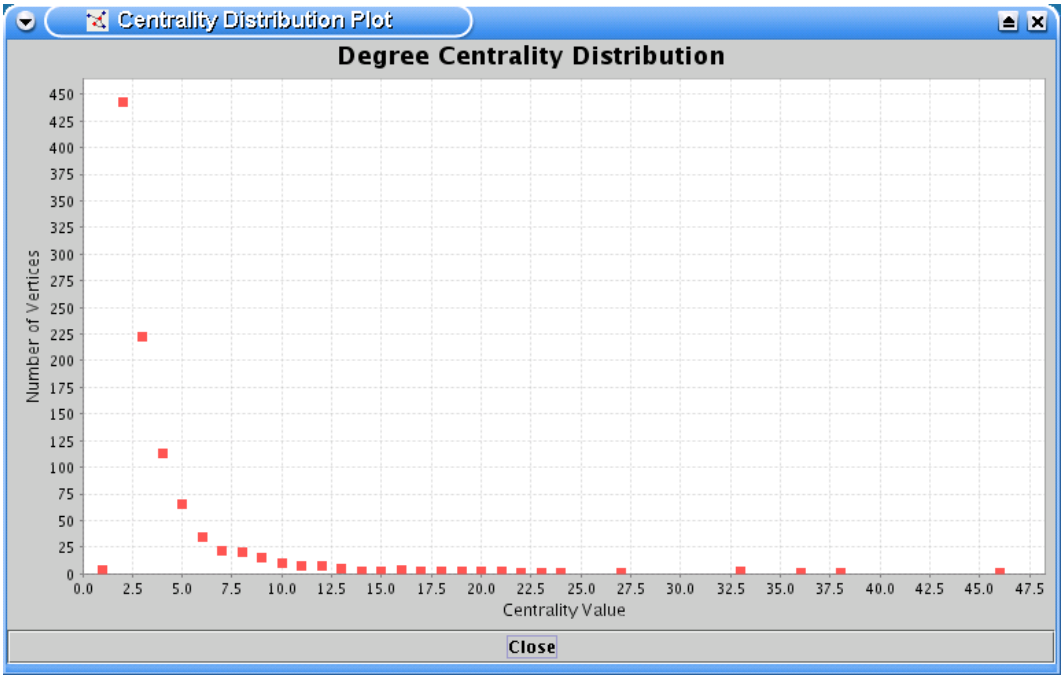
To load a centrality vector of the tsv file into R ([19]) use the following commands:

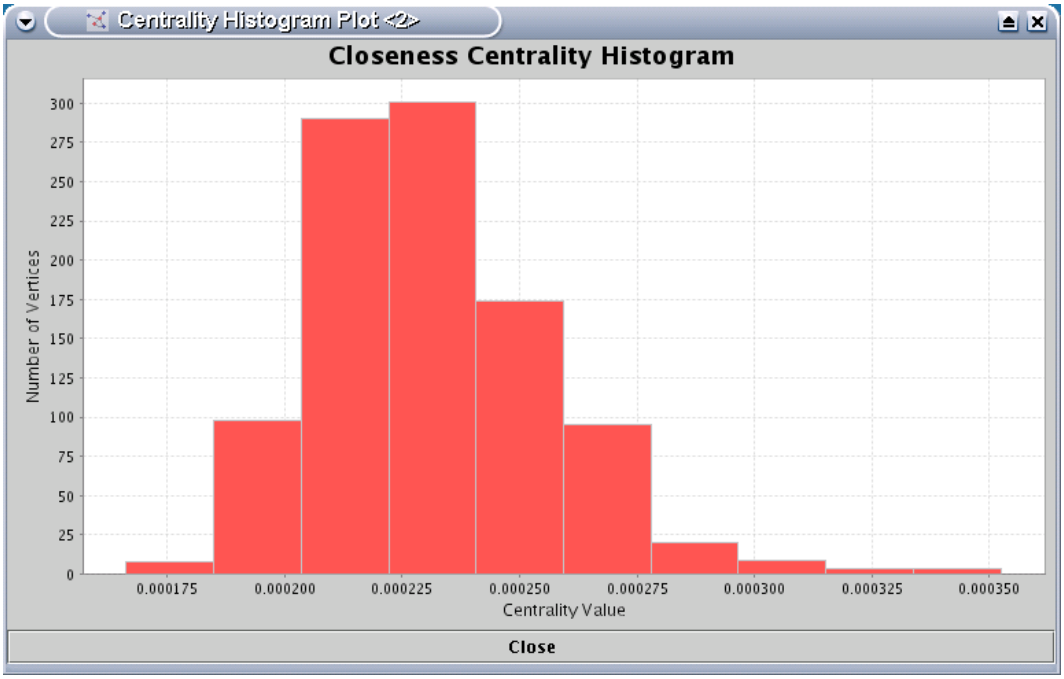
```
centralityVector <- read.table ("centralityVector.vec", skip=1)$V1
centralityVector
max (centralityVector)

centralityTable <- read.table ("centralityTable.tsv", header=TRUE, row.names=2)
centralityTable
max(centralityTable$Value)
```

5.6 Plotting the Distribution of Centrality Values

CentiBiN supports two plots for the centrality values: the distribution of centrality values and a histogram of centrality values. The following figures show both features.

The context menu available via the right mouse button provides several operations like zooming into the plot or saving the plot to a file.


Centrality Distribution Plot
Degree Centrality Distribution
Number of Vertices
Centrality Value
Close


Centrality Histogram Plot <2>
Closeness Centrality Histogram
Number of Vertices
Centrality Value
Close

6 Layout

CentiBiN offers five different layout algorithms for networks, reaching from simple circular to more advanced force directed layouts[13, 9]. Depending on the network, one or the other layout method results in a better visualisation.

7 Cleanup Operations

The "Utilities" menu contains several cleanup operations for the given network.

Some of these operations change the internal structure of the network and require that the network is saved to a file and reloaded afterwards. Only after performing a reload all internal data structures are rebuild correctly and centralities should be computed for the modified network.

7.1 Remove Loops

the Removes existing self loop from the network. Self loops are edges that connect a vertex to itself.

7.2 Remove Parallel Edges

Removes existing parallel edges from the network. In case of an undirected network at most one edge between two vertices may exist. In case of a directed network at most two (directed) edges may exist between two vertices, one edge for each of the possible directions.

7.3 Reduce to Giant Component

Several centrality measure require that the network is connected. This operation reduces the network to the largest connected component. In the case of a directed network the largest strong connected component is used.

7.4 Transform into Directed Network

The given network is transformed into a network containing only directed edges. All existing directed edges are copied and all existing undirected are transformed into two anti-parallel directed edges.

7.5 Transform into Undirected Network

The given network is transformed into a network containing only undirected edges. All existing undirected edges are copied and all existing directed are transformed into undirected edges. In the case of two anti-parallel directed edges only one undirected is created.

7.6 Prepare for Centralities – Directed

This operation performs the following step:

1. Removal of existing self loops (7.1)
2. Removal of existing parallel edges (7.2)
3. Reduction to the giant component (7.3)
4. Transformation into a directed network (7.4)

Only steps that are necessary for the given network are performed.

7.7 Prepare for Centralities – Undirected

This operation performs the following step:

1. Removal of existing self loops (7.1)
2. Removal of existing parallel edges (7.2)
3. Reduction to the giant component (7.3)
4. Transformation into an undirected network (7.5)

Only steps that are necessary for the given network are performed.

8 Options

The “Option” menu allows the activation of the automatic refresh of the birds eye view in the bottom right corner of the window and the activation of the tool tip system for vertices. Both features decrease the performance of CentiBiN and are therefore deactivated.

Additionally, the maximal size of graphs that shall be displayed can be adjusted in this menu. Graph of larger size can still be analysed but the graph visualisation is slow for larger graphs, therefore we suggest to disable the visualisation of graphs with several thousand vertices.

9 Getting more Help

CentiBiN is a ongoing project. Therefore do not hesitate to contact the author with questions, comments or bug reports.

Dirk Koschützki
koschuet@ipk-gatersleben.de

10 CentiBiN License

CentiBiN is available free of charge.

10.0.1 Disclaimer of Warranties

You acknowledge and agree that the use of CentiBiN is at your sole risk and that the entire risk as to satisfactory quality, performance, accuracy and effort is with you. It can not be guaranteed, that the functions contained in this software will meet your requirements, that the operation of CentiBiN will be uninterrupted or error-free, or that defects in the software will be corrected.

Bibliography

[1] Graphviz - graph visualization software [http://www.graphviz.org/].

[2] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[3] Vladimir Batagelj and Andrej Mrvar. Pajek - Analysis and Visualization of Large Networks. In *Graph Drawing Software*, pages 77–103. Springer, 2004.

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

[5] Phillip Bonacich. Power and centrality: A family of measures. *American Journal of Sociology*, 92(5):1170–1182, 1987.

[6] Ulrik Brandes and Thomas Erlebach, editors. *Network Analysis: Methodological Foundations*, volume 3418 of *LNCS Tutorial*. Springer, 2005.

[7] Ulrik Brandes and Daniel Fleischer. Centrality measures based on current flow. In *Proc. 22nd Symp. Theoretical Aspects of Computer Science (STACS '05)*, volume 3404 of *Lecture Notes in Computer Science (LNCS)*, pages 533–544. Springer-Verlag, 2005.

[8] Linton Clarke Freeman. A set of measures of centrality based upon betweeness. *Sociometry*, 40:35–41, 1977.

[9] Thomas M. J. Fruchterman and Edward M. Reingold. Graph drawing by force-directed placement. *Software - Practice and Experience*, 21(11):1129–1164, 1991.

[10] Frank Harary and Per Hage. Eccentricity and centrality in networks. *Social Networks*, 17:57–63, 1995.

[11] Charles H. Hubbell. In input-output approach to clique identification. *Sociometry*, 28:377–399, 1965.

[12] Riko Jacob, Dirk Koschützki, Katharina Anna Lehmann, Leon Peeters, and Dagmar Tenfelde-Podehl. *Network Analysis: Methodological Foundations*, chapter Algorithms for Centrality Indices, pages 62–82. Volume 3418 of Brandes and Erlebach [6], 2005.

[13] Tomihisa Kamada and Satoru Kawai. An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31(1):7–15, 1989.

[14] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

[15] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[16] Jon M. Kleinberg. Navigation in a small world. *Nature*, 406:845, 2000.

[17] Dirk Koschützki, Katharina Anna Lehmann, Leon Peeters, Stefan Richter, Dagmar Tenfelde-Podehl, and Oliver Zlotowski. *Network Analysis: Methodological Foundations*, chapter Centrality Indices, pages 16–61. Volume 3418 of Brandes and Erlebach [6], 2005.

[18] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[19] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2005. ISBN 3-900051-07-0.

[20] Gert Sabidussi. The centrality index of a graph. *Psychometrika*, 31:581–603, 1966.

[21] Lukasz Salwinski, Christopher S. Miller, Adam J. Smith, Frank K. Pettit, James U. Bowie, and David Eisenberg. The database of interacting proteins: 2004 update. *Nucleic Acids Res*, 32(1):449–451, 2004.

[22] Alfonso Shimbel. Structural parameters of communication networks. *Bulletin of Mathematical Biophysics*, 15:501–507, 1953.

[23] Peter J. Slater. Maximin facility location. *Journal of National Bureau of Standards*, 79B:107–115, 1975.

[24] Thomas W. Valente and Robert K. Foreman. Integration and radiality: measuring the extent of an individual's connectedness and reachability in a network. *Social Networks*, 1:89–105, 1998.