Bornholm network analysis

Finn Årup Nielsen

Informatics and Mathematical Modelling
Technical University of Denmark
DK-2800 Lyngby, Denmark

Email: fn@imm.dtu.dk
WWW: http://www.imm.dtu.dk/~fn

OVERVIEW

- Applications.
- The network as a matrix
- Centrality
- Iterative centrality
- Network measures
- Max flow-min cut

APPLICATIONS

- "Link analysis" on the Internet, e.g., Google: webpage (Brin and Page, 1998).
- Connectivity analysis in the brain (Young, 1992).
- Social network analysis, sociometry: people, organizations. E.g., (Scott, 1991).
- Bibliometric network analysis, scientometrics: citation, articles (Garfield, 1955).
- "Graph-based image analysis": voxels, objects. E.g., (Sonka et al., 1993).

THE NETWORK AS A MATRIX

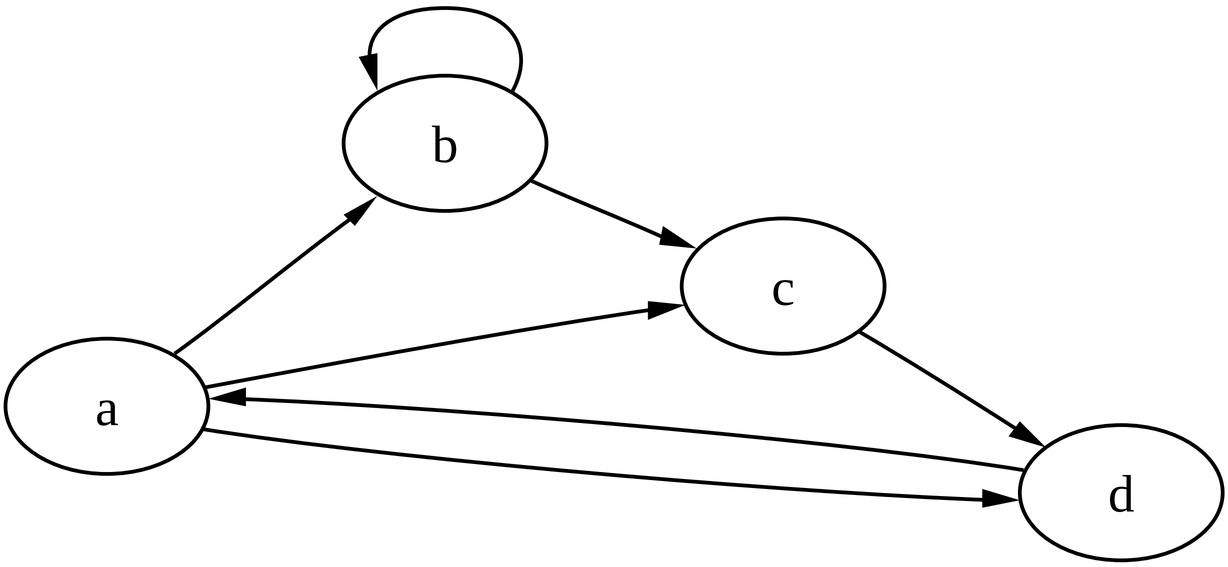

Figure 1: Network example.

- Adjencency matrix $\mathbf{A}(N \times N)$ with from node $\times$ to node. Example

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \tag{1}$$

- Terminology:
  - Node, vertex, e.g., a web-page.
  - Links, edges, arcs, bonds, e.g., a hyperlink.
  - Directed graph: in-links, out-links

CENTRALITY

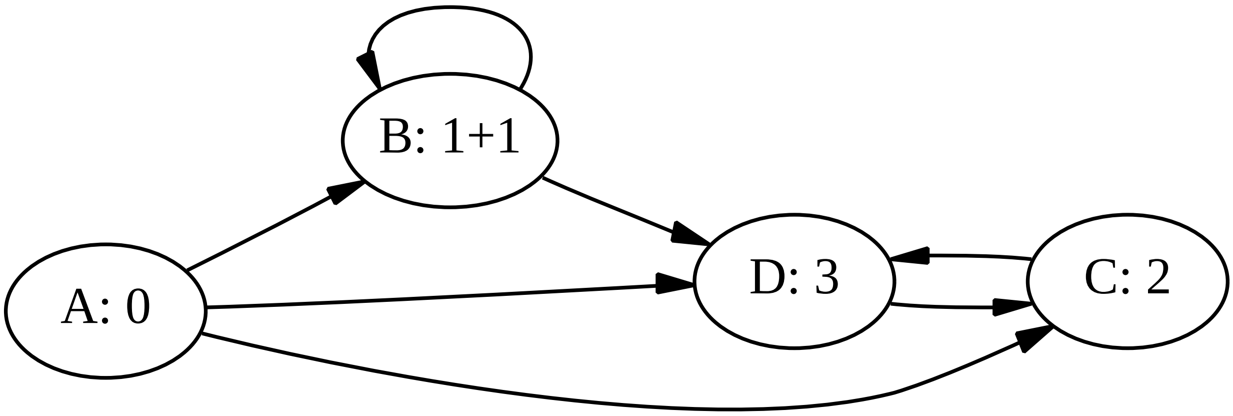

Figure 2: Centrality. Four nodes with ranks. Numbers are the centrality.

- Centrality, ranking, prestige, power, status, influence.
- Simple measure of centrality: Count the in-links to a node.

$$\mathbf{c} = \mathsf{sum}[\mathbf{A}] = \mathbf{1}^{\mathsf{T}}\mathbf{A} \tag{2}$$

- ResearchIndex distinguishes between self-citations and external citations

$$\mathbf{c} = \mathsf{sum}[\mathbf{A}] - \mathsf{diag}[\mathbf{A}] \tag{3}$$

EXAMPLE ON CENTRALITY

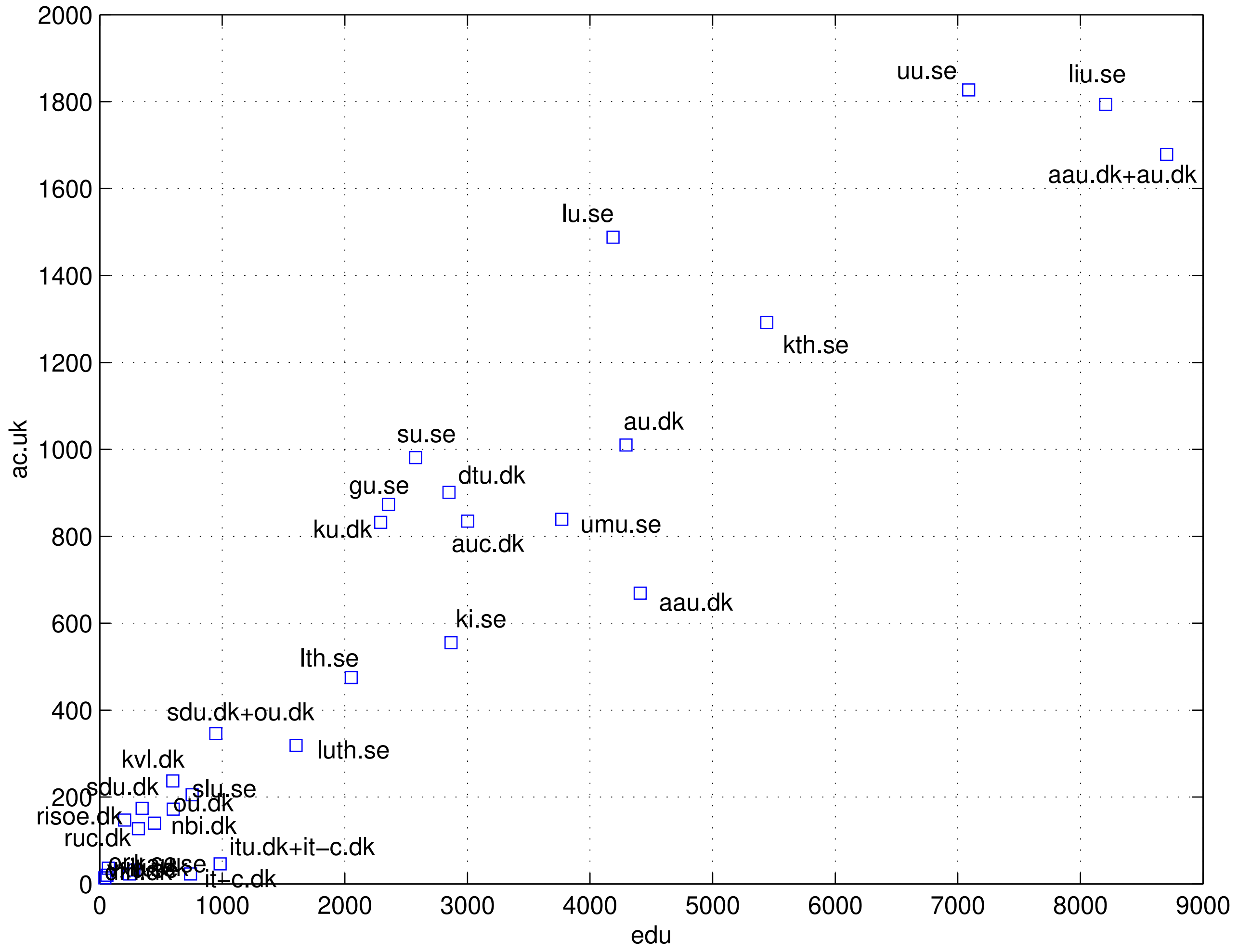

Figure 3: Centrality. Most cited Danish and Swedish universities from the "edu" and "ac.uk" domains. Number of links obtained with AllTheWeb search engine. Normalization? Students: 18220 for AU, 28356 for KU, 5755 for DTU.

- Citation count on Danish and Swedish universities from "edu and "ac.uk" domains.
- Conclusion(?): Århus University and Swedish (Linköping, Uppsala, Lund) universities "better" than Copenhagen (note diku.dk and nbi.dk).

NORMALIZED CENTRALITY

- Normalization count in scientometrics: *Journal impact factor* for journal evaluation with *Science Citation Index* (SCI) (Garfield, 1972)

$$\text{Journal impact factor} = \frac{\text{Times cited}}{\text{Articles published}} \tag{4}$$

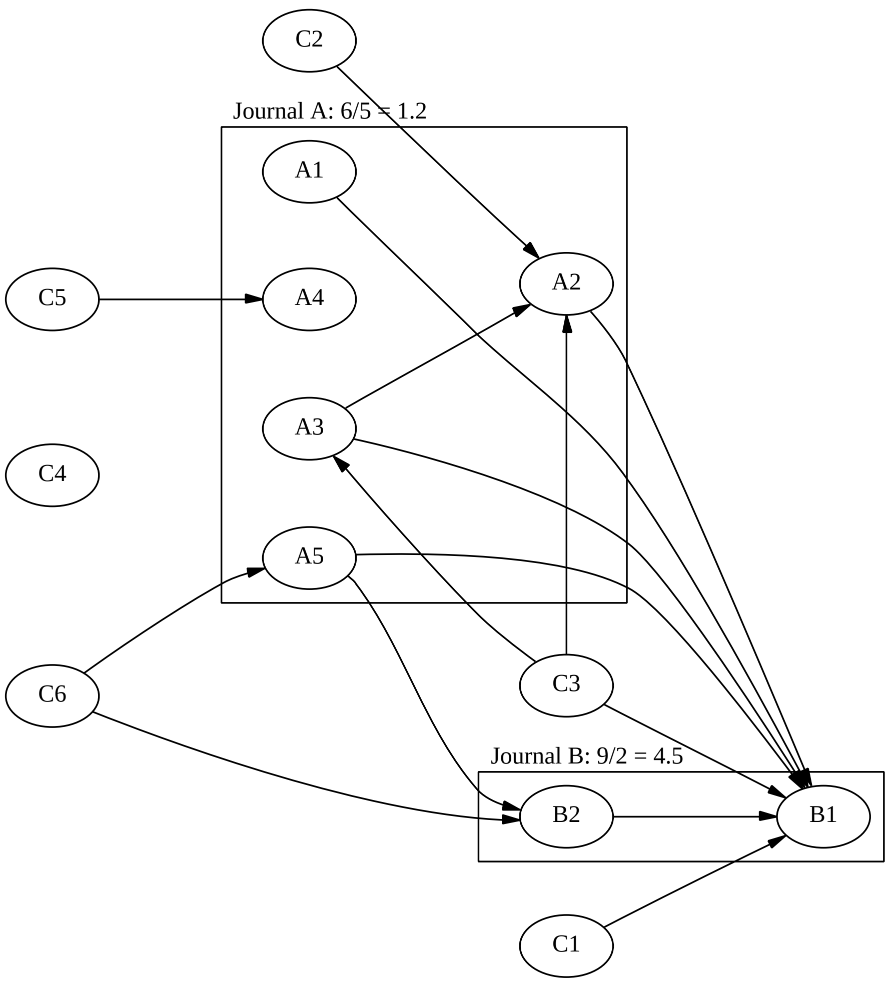

Figure 4: Normalized centrality. Which one has the highest rank: A2 or B2?

ITERATIVE-BASED CENTRALITY

- Inherite the centrality of in-links (Bonacich, 1972), (Scott, 1991, p. 91)

$$c_j = \sum_i a_{ij} c_i \tag{5}$$

where c_i is the centrality.

- Scientometrics (Pinski and Narin, 1976): "Influence weights" w_i for item i

$$w_i = \sum_{k=1}^{N} \frac{w_k a_{ki}}{s_i} \tag{6}$$
$$s_i = \sum_{j=1}^{N} a_{ij} \tag{7}$$

where s_i is the number of references.

"Influence" is the product of the number of references and the influence weight

$$I_i = s_i w_i \tag{8}$$

- Eigenvalue solution

MARKOV CHAIN CENTRALITY

- Adjacency matrix converted to a *stochastic matrix* $\mathbf{P}$ (non-negative and row sums are unity) (Cox and Miller, 1965, sec. 3.3).
- Eigenvalue $\lambda_1 = 1$ since row sum unity: $\mathbf{P1} = \mathbf{1}$ with a positive row (left) eigenvector statisfying

$$\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi} \tag{9}$$

  according to the Perron-Frobenius theorem (Cox and Miller, 1965)

- As infinitely many steps are taken:

$$\lim_{n \to \infty} \mathbf{P}^n = \mathbf{1}\boldsymbol{\pi}. \tag{10}$$

- Weight interpretable as the mean time in state i

"REGULARIZED" ITERATIVE-BASED CENTRALITY

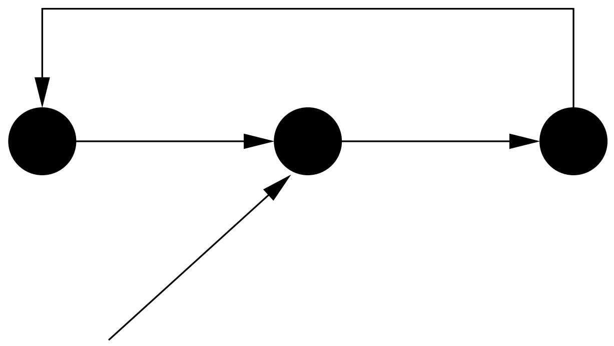

Figure 5: Absorbtion problem.

- Problem with absorbing states, i.e., some states are transient, nodes with no out-links, and network is associated with a reducible stochastic matrix: Infinite influence.
- PageRank (Page, 1997; Brin and Page, 1998; Page, 2001)

$$w_j = (1-d) + d \sum_{i=1, i \neq j}^{N} a_{ij} \frac{w_i}{s_i} \tag{11}$$

  where $d = 0.9$ and w_j is the "page rank" for the j page.
- "Random surfer": random follows links (with probability d) and occasionally jumps to a random page (with probability $1-d$).

DIFFERENT REFERENCING

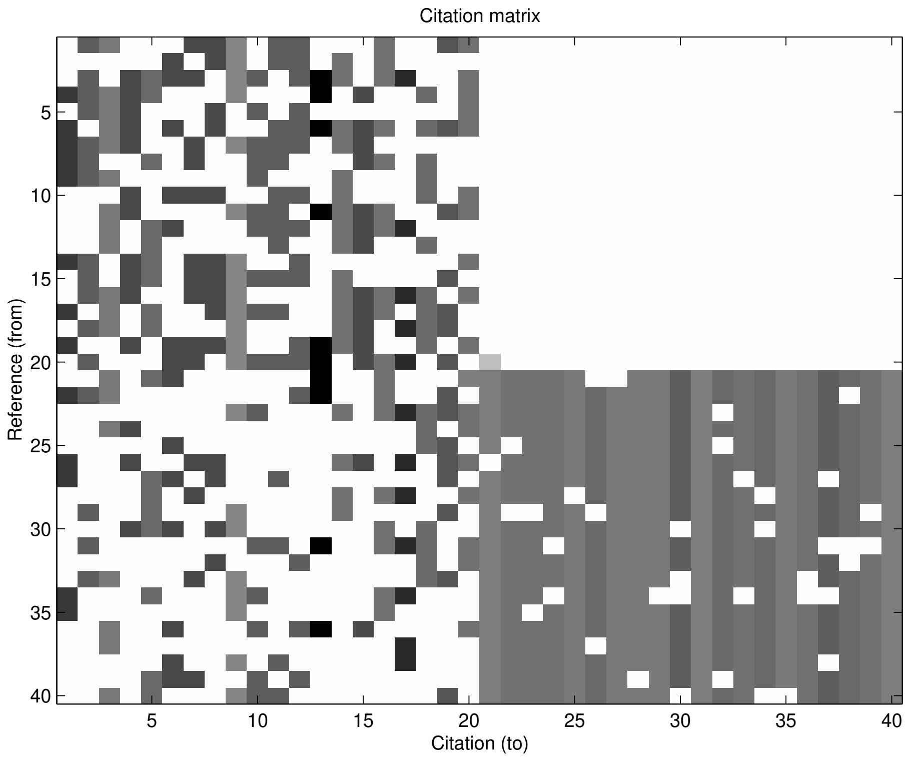

Figure 6: Two different segments: One segment with a small number of references (P=0.5), a second with a large a number of references (P=0.9), — and that references the first segment (P=0.1).

- "No normalization for the different referencing characterstics of different segments of the litterature" (Pinski and Narin, 1976, p. 298), e.g., many references in biochemistry, few in astronomy.
- Raw count ordering: $\{19, 25, 36, 17, 31, 32, 35, 10, ...\}$: Second segment most central.
- Markov chain ordering: $\{10, 17, 12, 18, 19, 5, 6, 16, ...\}$: First segment most central.

HUBS AND AUTHORITIES

- Hyperlink induced topic search (HITS), (Kleinberg, 1997; Kleinberg, 1998; Kleinberg, 1999)
- Two-way algorithm
  - *Hubs* links to many authorities.
  - *Authorities* linked to by many hubs.
- Iterative algorithm

$$a(j) := \sum_{i \to j} h(i) \qquad (12)$$

$$h(i) := \sum_{i \to j} a(j) \qquad (13)$$

$$\text{Normalize } a, h \qquad (14)$$

- Singular value decomposition svd$(\mathbf{A})$ or two eigenvalue problem on $\mathbf{A}^\mathsf{T}\mathbf{A}$ and $\mathbf{A}\mathbf{A}^\mathsf{T}$
- Higher components used for clustering, identifying communities (Gibson et al., 1998).
- Original description worked with a small subset of documents:
  1. Using a search engine to obtain a set (say 200) of documents (root set).
  2. Augmenting the set with documents pointing to and pointed to by the root set.
  3. Perform HITS on the combined data set.

PEAK AND BRIDGE

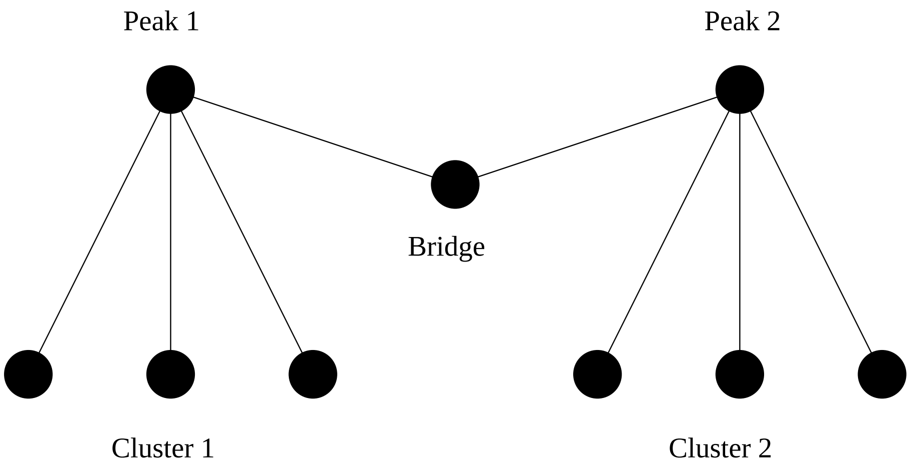

Figure 7: Peak and bridge. Derived from (Scott, 1991, fig. 5.5),

- A *peak* is a node which is more central than any other nodes it is connected to (Scott, 1991, p. 101).
  - Local centrality.
  - Similar to SPM99 definition of maximum point in an image volume (3D regular lattice).
- A *bridge* is a node which connects two peaks (Scott, 1991, p. 101) (Mintz and Schwartz, 1985) (Mizruchi et al., 1986).
  - High *derived* centrality: Related to authority

NETWORK MEASURES

- *Density* measures the connectedness of a network

$$d = \frac{l}{n(n-1)/2} \tag{15}$$

  where n is the number of nodes and l is the number of links (Scott, 1991, p. 74).

  - Local/global (ego-centric/socio-centric). Local density
  - A *strongly connected components* is a set of nodes where the density is one (for the local network).

- *Diameter* defined as the largest minimum distance between any two nodes (Scott, 1991).

  - Diameter of the web: Average 19 clicks between to documents (Albert et al., 1999).

"SMALL-WORLD" PHENOMENON

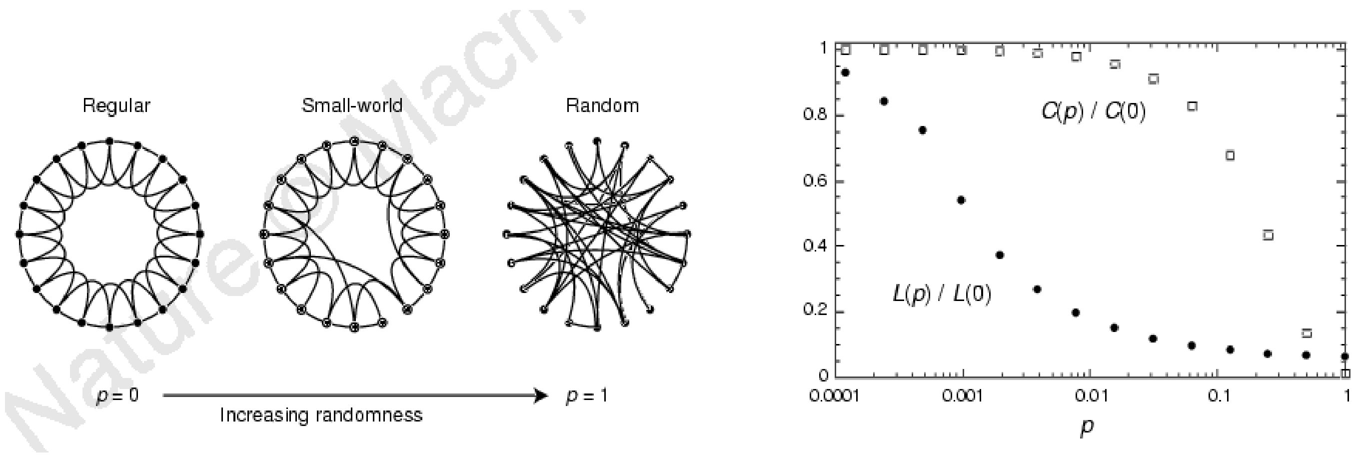

Figure 8: From (Watts and Strogatz, 1998, fig. 1 and 2).

- "Small-world" network (Watts and Strogatz, 1998).
- Nets between random graphs and regular lattices.
- Short average paths while still being clustered.
- Appear in "natural" networks: Film actors, power grid, neural network of worm.
- Small-world coefficients
  - Characteristic path length L: Number of edges in the shortest path between two vertices, averaged over all pairs of vertices.
  - Clustering coeffient C: The fraction between the (number of) actual connection and the possible connection among a vertex' neighbors (averaged over all vertices), cf. local density.
- Does not regard the distribution of degrees.

MAXIMUM FLOW AND MINIMAL CUT

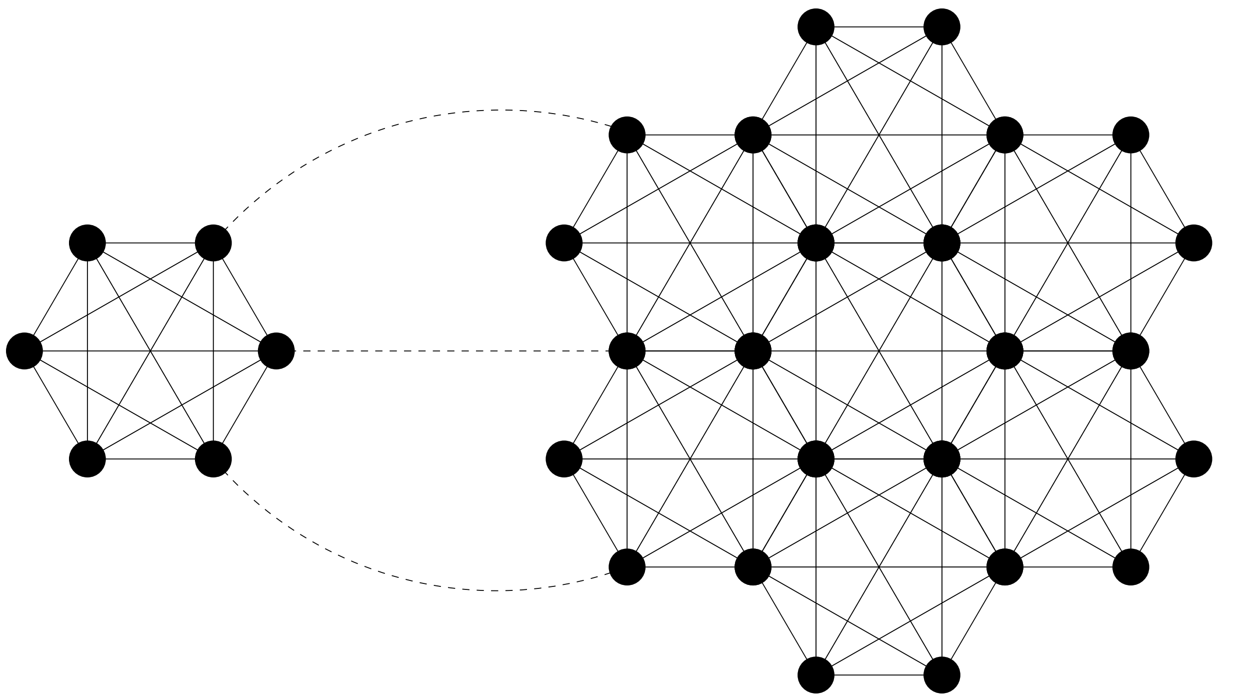

Figure 9: Maxmimum flow algorithm will separate the three dashed lines. From (Flake et al., 2002, figure 1).

- Max-flow-min cut (Ford and Fulkerson, 1956): Edges water pipes and vertices pipe junctions.
- Identification of web communities without text analysis (Flake et al., 2002; Flake et al., 2000).
- Maximum flow algorithms: "Incremental shortest augmentation" (ISA) (Flake et al., 2000, table 5), "pre-flow push" (Goldberg, 1986).

FURTHERMORE...

- Types of links
  - Negative/positive.
  - Valued graph, with numerical values for links.
  - Types of link identified by analyzing text around link, e.g., scientific citations can have reasons for citation: "giving credited for related work", "identifying methodology, equipment, and so on", "criticizing previous work", "disclaiming work or ideas of others" (Nanba et al., 2000).
- Text and link combination
  - Combination of page content (terms) and link structure (PageRank) (Richardson and Domingos, 2002), i.e., invidual rank for every word.
  - Determine what a page is known for (Rafiei and Mendelzon, 2000; Mendelzon and Rafiei, 2000).
  - HyPursuit (Weiss et al., 1996).
  - PLSA/PHITS (Cohn and Hofman, 2001).
  - Text, link and usage (Chen, 1997).

COCITATION ANALYSIS

- Science citation cotation analysis typically use data from ISI. Special (expensive) queries are required ("Dialog"). Only working with preselected key authors.
- Author cocitation analysis (White, 1981; White and Griffith, 1981; McCain, 1990). Construction of author $\times$ author matrices
- Example analysis types: Nonlinear multidimensional scaling, (hierarchical) cluster analysis.
- Example studies
  - Author cocitation analysis in information science (White and McCain, 1998).
  - (McCain, 1998) Dynamic (longitudinal) analysis of 36 journals in "neural networks research and related disciplines".
  - (Larson, 1996) Examined 34 earth science related web-sites. Obtained cocitation count from AltaVista "link:". Multidimensional scaling with groups "Geography/GIS", "Ecology/environment", "Geophysics", "Remote sensing", "Wheather Climate".

References

Albert, R., Jeong, H., and Barabasi, A.-L. (1999). diameter of the world-wide web. *Nature*, 401:130–131.

Bonacich, P. (1972). Technique for analysing overlapping memberships. In *Sociological Methodology*. Jossey-Bass, San Francisco.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117. http://www-db.stanford.edu/pub/papers/google.pdf. ResearchIndex: http://citeseer.nj.nec.com/brin98anatomy.html. Describes the Google Internet search algorithm and the PageRank algorithm.

Chen, C. (1997). Structuring and visualising the world-wide web with generalised similarity analysis. In *Proceedings of the 8th ACM Conference on Hypertext (Hypertext '97)*. ACM.

Cohn, D. and Hofman, T. (2001). The missing link — a probabilistic model of document content and hypertext connectivity. In Leen, T. K., Dietterich, T. G., and Tresp, V., editors, *Advances in Neural Information Processing Systems*, Boston, MA. MIT Press. NIPS-13.

Cox, D. R. and Miller, H. D. (1965). *The Theory of Stochastic Processes*. Chapman & Hall, London, UK. ISBN 0412151707, [defkat.dk — bibliotek.dk — amazon.com — bn.com].

Flake, G. W., Lawrence, S., and Giles, C. L. (2000). Efficient identification of web communities. In *Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 150–160.

Flake, G. W., Lawrence, S., Giles, C. L., and Coetzee, F. (2002). Self-organization and identification of web communities. *IEEE Computer*, 35(3):66–71. http://computer.org/computer/co2002/r3066abs.htm.

Ford, L. R. and Fulkerson, D. R. (1956). Maximal flow through a network. *Canadian Journal of Mathematics*, 8:399–404.

Garfield, E. (1955). Citation indexes for sciences. *Science*, 122(3159):108–111.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, (4060):471–479. PMID: 5079701. http://www.garfield.library.upenn.edu/essays/V1p527y1962-73.pdf.

Gibson, D., Kleinberg, J., and Raghavan, P. (1998). Inferring web communities from link topology. In *Proceedings of the ninth ACM conference on Hypertext and hypermedia: links, objects, time and space-structure in hypermedia systems*, pages 225–234. ACM.

Goldberg, A. V. (1986). A new approach to the maximum flow problem. In *Proceedings of the Eighteenth annual ACM Symposium on Theoty of Computing*, pages 136–146.

Kleinberg, J. M. (1997). Authoritative sources in a hyperlinked environment. Technical Report 10076(91892), IBM.

Kleinberg, J. M. (1998). Authoritative sources in a hyperlinked environment. In *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 668–677.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of ACM*, 46(5):604–632.

Larson, R. R. (1996). Bibliometrics of the world wide web: An exploratory analysis of the intellectual structure of cyberspace. In *Proceedings of the 1996 Annual ASIS Meeting, Baltimore*, pages 71–79.

McCain, K. W. (1990). Mapping authors in intellectual space: A technical overview. *Journal of the American Society for Information Science*, 41(6):433–443.

McCain, K. W. (1998). Neural networks research in context: A longitudinal journal cocitation analysis of an emerging interdisciplinary field. *Scientometrics*, 41(3):389–410.

Mendelzon, A. and Rafiei, D. (2000). What do the neighbours think? computing web page reputations. *Bulletin of the Technical Committee on Data Engineering*, 23(3):9–16. http://www.research.microsoft.com/research/db/debull/A00sept/mendelz.ps. ResearchIndex: http://citeseer.nj.nec.com/374160.html.

Mintz, B. and Schwartz, M. (1985). *The Power Structure of American Business*. University of Chicago Press, Chicago.

Mizruchi, M. S., Pariolis, P., Schwartz, M., and Mintz, B. (1986). Techniques for disaggretating centrality scores in social networks. In Tuma, N., editor, *Social Methodology*, pages 26–48. Jossey-Bass.

Nanba, H., Kando, N., and Okumura, M. (2000). Classification of research papers using citation links and citation types: Towards automatic review article generation. In *The 11th SIG Classification Research Workshop, Classification for User Support and Learning, 2000.11, in Chicago, USA*, pages 117–134. The American Society for Information Science (ASIS).

Page, L. (1997). PageRank: Bringing order to the web. Stanford Digital Libraries Working Paper 1997-0072, Stanford University.

Page, L. (2001). Method for node ranking in a linked database. U.S. Patent No. 6,285,999-B1.

Pinski, G. and Narin, F. (1976). Citation influence for journal aggregates of scientific publications: Theory with application to literature of physics. *Information Processing & Management*, 12:297–312.

Rafiei, D. and Mendelzon, A. O. (2000). What is this page known for? computing web page reputations. *Computer Networks*, 33:823–835.

Richardson, M. and Domingos, P. (2002). The intelligent surfer: Probabilistic combination of link and content information in pagerank.

Scott, J. (1991). *Social Network Analysis: A Handbook*. SAGE Publications, London, UK. ISBN 0803984804, [defkat.dk — bibliotek.dk — amazon.com — bn.com].

Sonka, M., Hlavac, V., and Boyle, R. (1993). *Image, Processing, Analysis and Machine Vision*. Chapman & Hall.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):397–498. PMID: 9623998. http://www.nature.com/cgi-taf/DynaPage.taf-?file=/nature/journal/v393/n6684/abs/393440a0_fs.html. Described so-called 'small-world' networks which are interpolations between regular (lattice) and random networks. They exhibits short averages paths while still being clustered. Examples are given by film actors, power grid and the neural network of a worm together with dynamics networks, e.g, multiplayer prisoner's dilemma.

Weiss, R., Velez, B., Sheldon, M. A., Nemprempre, C., Szilagyi, P., Duda, A., and Gifford, D. K. (1996). HyPursuit: A hierarchical network search engine that exploits content-link hypertext clustering. In *Proceedings of the Seventh ACM Conference on Hypertext, Washington, DC, March 1996*.

White, H. D. (1981). Cocited author retrieval online: An experiment with the social indicators litterature. *Journal of the American Society for Information Science*, 32:16–22.

White, H. D. and Griffith, B. C. (1981). Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for Information Science*, 32:163–171.

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.

Young, M. P. (1992). Objective analysis of the topological organization of the primate cortical visual system. *Nature*, 358(6382):152–155. PMID: 1614547.