Network analysis of texts

Vladimir Batagelj[*], **Andrej Mrvar**[†], **Matjaž Zaveršnik**[‡]

[*]University of Ljubljana, Faculty of Mathematics and Physics,
Jadranska 19, 1000 Ljubljana
vladimir.batagelj@uni-lj.si

[†]University of Ljubljana, Faculty of Social Sciences,
Kardeljeva ploščad 5, 1000 Ljubljana
andrej.mrvar@uni-lj.si

[‡]University of Ljubljana, Faculty of Mathematics and Physics,
Jadranska 19, 1000 Ljubljana
matjaz.zaversnik@fmf.uni-lj.si

Abstract

In the paper different ways to derive networks from the textual data and an overview of (possible) applications of network analysis to the analysis of texts is presented. Several examples of analyses of different text networks are given as illustrations.

1. Introduction

Different kinds of networks can be generated from already existing electronic sources. Text is a special and frequent form of such data. In the paper an overview of different ways to derive networks from the textual data and to analyse them are presented.

The obtained networks can be very large, having some ten or hundred of thousands of vertices. Therefore special algorithms are needed to analyze and visualize them.

All the analyses in the paper were done with Pajek, a program (for Windows) for large network analysis and visualization. It is freely available, for noncommercial use, at its site (Batagelj and Mrvar, 1998).

We shall assume that the reader is familiar with the basic notions of graph theory (see for example (Wilson and Watkins, 1990)).

2. Vocabularies

2.1. Transforming `water` into `wine`

In the recreational literature we find problems such as: transform the `water` into `wine` by a sequence of words obtained by changing one character (deleting, inserting or replacing it) each time. For example

`water` – `wader` – `wade` – `wane` – `wine`

or

`water` – `waver` – `wave` – `wive` – `wine`

Given a vocabulary of a language (we are using Knuth's vocabulary of English (Knuth, 1993)) we can construct the corresponding ***transformations graph*** $G = (V, E)$ in which the set of vertices V consists of words from the dictionary; and there is an edge $(u : v) \in E$ linking the words u and v iff v can be obtained from u by changing one character. On this graph the recreational problem turns into a problem of determining a path between given two words. Usually we try to find a shortest such path, also called a ***geodesic***. In Figure 1 the graph of all geodesics leading from `black` to `white` is presented.

Constructing the transformations graph is interesting also as a computational problem – how to do it efficiently? We used the following approach: for each word $w \in V$ a list of pairs (w', w) is produced, where w' is a transformation pattern in which the place of transformation is indicated by a star `*`. For example, for the word `brain` we obtain the list

(`*rain`, `brain`), (`b*ain`, `brain`),
(`br*in`, `brain`), (`bra*n`, `brain`),
(`brai*`, `brain`), (`*brain`, `brain`),
(`b*rain`, `brain`), (`br*ain`, `brain`),
(`bra*in`, `brain`), (`brai*n`, `brain`),
(`brain*`, `brain`)

It holds: $(u : v) \in E$ iff there exists a pattern p such that the union L of all lists contains both pairs (p, u) and (p, v). For example

(`rain`: `brain`) $\in E$

since (`*rain`, `rain`), (`*rain`, `brain`) $\in L$; and

(`train`: `brain`) $\in E$

since (`*rain`, `train`), (`*rain`, `brain`) $\in L$.

To identify efficiently all such pairs we first sort the list L on the first elements of its pairs. In this way pairs with the same pattern are grouped together. We have only to produce the corresponding edges for each such group. Note also that the list L can be viewed also as a 2-mode (bipartite) graph between patterns and words.

Using standard sorting algorithm the complexity of this procedure is of order $O(|V| \log |V|)$; it can be made linear by using bin sort.

Several transformations graphs in Pajek's format are available at Pajek's site.

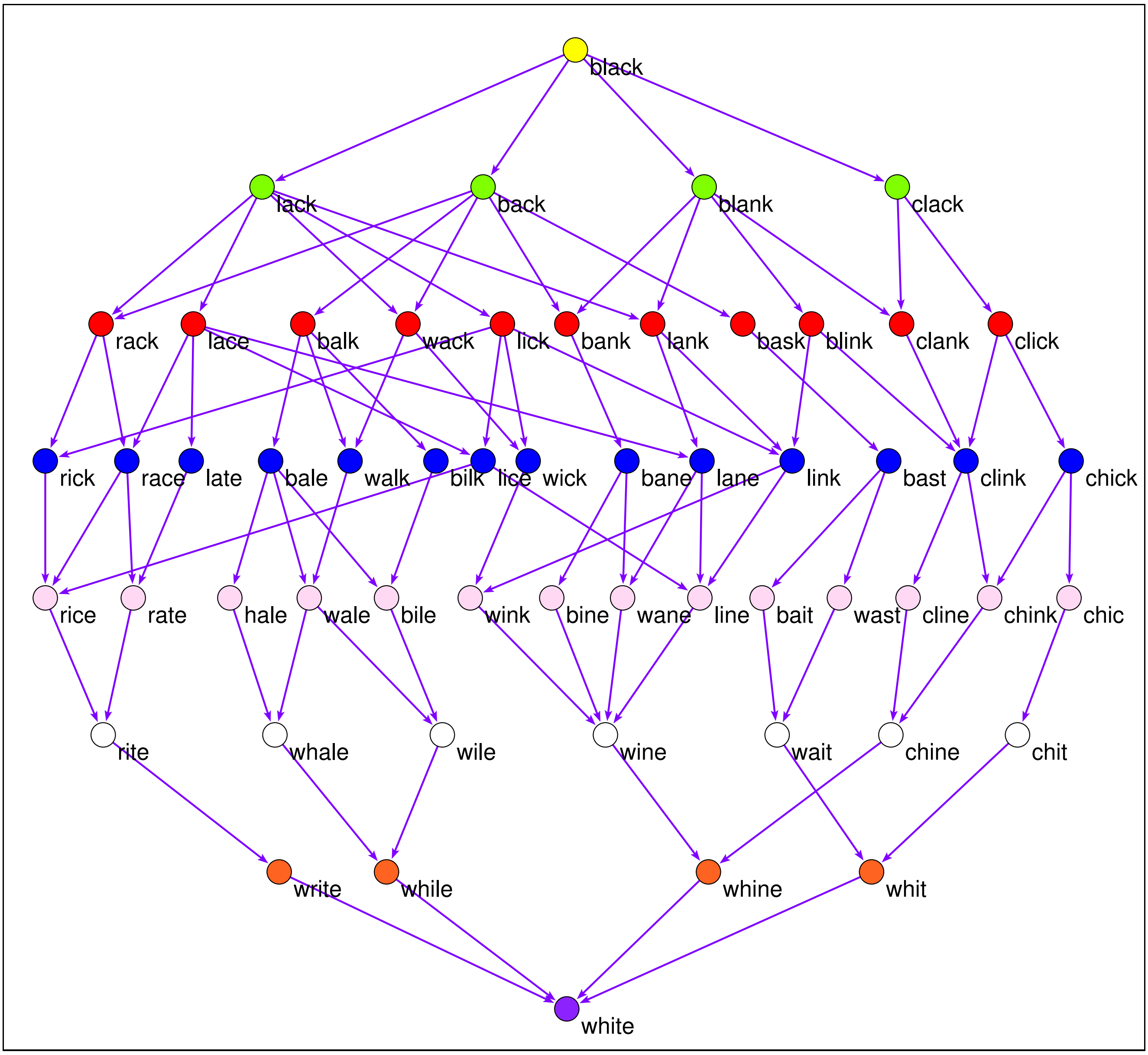

Figure 1: black – white

2.2. Things to do

The transformations graphs can be produced also for other languages, provided the language vocabulary is available. For Slovene language only a vocabulary of all word forms is freely available (NL, 2002; GNUsl, 2002) – it is not appropriate according to ’recreational rules’.

It is also possible to introduce additional transformations. For example a swap (interchange) of two characters (with empty character allowed):

`life – file` and `arc – car`

Other, linguistic relations between words are also interesting. An example of such data collection is the **WordNet**: a lexical database for the English language (WordNet, 2002; Lexical FreeNet, 2002). The Pajek’s version of WordNet data is in preparation.

3. Dictionaries

On the web several on-line dictionaries are available in which each term is described using other terms. For example: Online Dictionary of Library and Information Science (ODLIS, 2002), Free Online Dictionary of Computing (FOLDOC, 2002), and The GNU collaborative international dictionary of English (GCIDE_XML, 2002).

Such dictionary can be transformed into a directed graph $G = (V, A)$: the terms determine the set of vertices V; and there is an arc $(u, v) \in A$ from term u to term v iff the term v appears in the description of term u (as a marked term).

We present some approaches to analysis and visualization of dictionaries in a separate paper (Batagelj et al., 2002), demonstrating several options for analysis: searching for important, dense or in some other way interesting parts of network; searching for important (central) words in networks; and visualization of results.

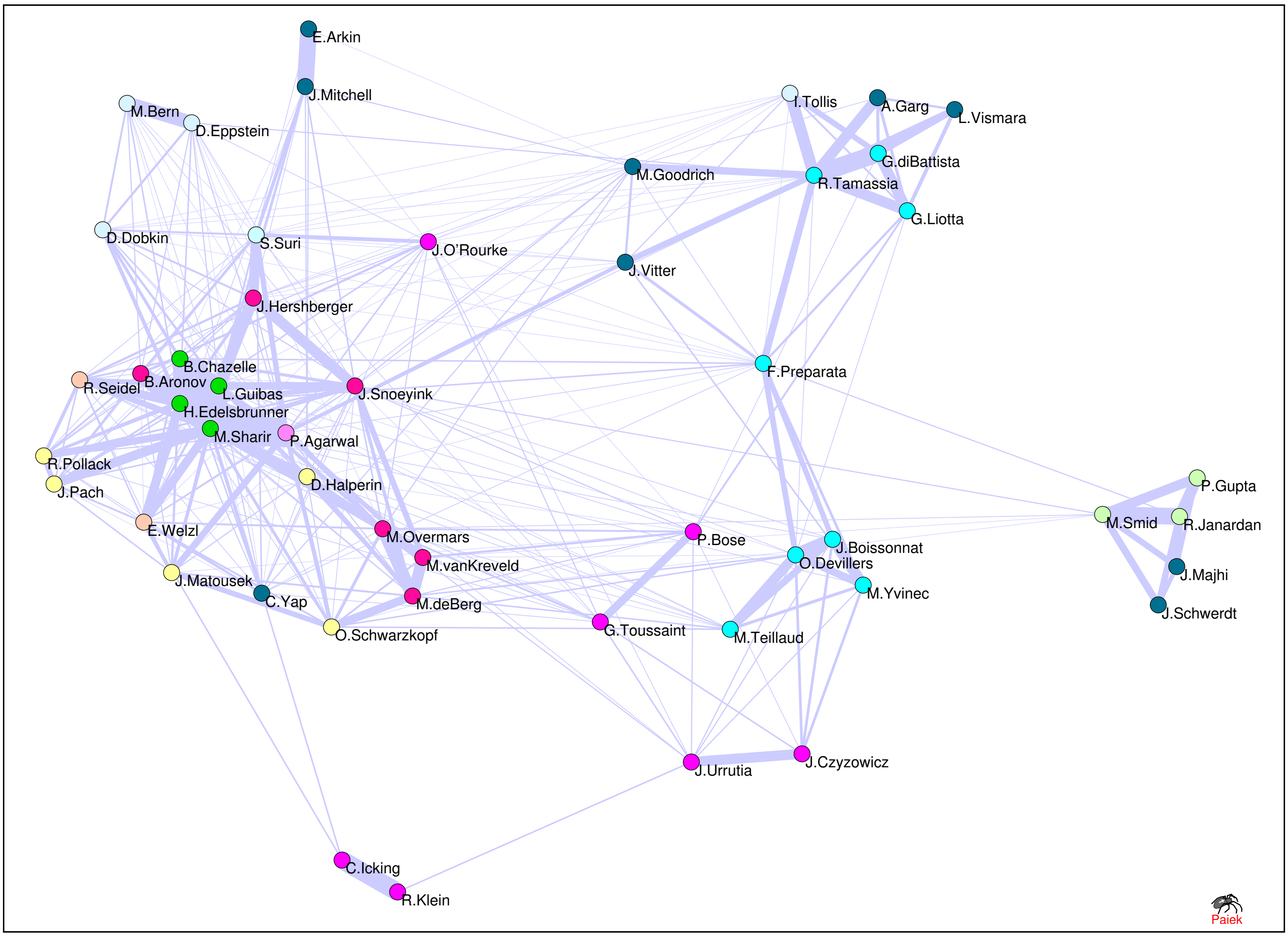

Figure 2: Valued core of the colaboration network of Computational Geometry at level 46.

4. Bibliographic networks

4.1. Collaboration networks

A 'classical' example of collaboration network is the ***Erdős network*** (Batagelj and Mrvar, 2000).

On the Internet many bibliographies in BiBTEX format are available (Beebe, 2002). From such a bibliography an ***authors colaboration network*** can be build. Its vertices represent different authors. Two authors are linked with an edge, iff they wrote a common paper. The weight of the edge is the number of publications they wrote together.

As an example we produced the authors collaboration network based on the bibliography obtained from the ***Computational Geometry Database*** geombib (Jones, 2002).

Using a simple program in Python, the BiBTEX data were transformed into the corresponding network, and output to the file in Pajek format. The obtained network has 9072 vertices (authors) and 22577 edges / 13567 edges as a simple network.

The problem with the obtained network is that it contains several vertices corresponding to the same author (Pankaj K. Agarwal, P. Agarwal, Pankaj Agarwal, and P.K. Agarwal) – that are easy to guess; but an 'insider' information is needed to know that O. Schwarzkopf and O. Cheong are the same person. We manually produced the ***name equivalence*** partition and then shrank the network according to it. The reduced simple network contains 7343 vertices and 11898 edges.

To this network we applied the algorithm for determining valued cores – vertex value is the sum of weights in a vertex (Batagelj and Zaveršnik, 2002). The cut at level 46 gave the network presented in Figure 2.

4.2. Citation networks

Another interesting type of networks that can be derived from the bibliographical data are ***citation networks***. Here the vertices are different publications from the selected area; two publications are connected by an arc if the first is cited by the second. The citation networks are almost acyclic.

A great source of the necessary data for building citation networks is the Web of Science (ISI, 2002) from where a selection of networks was constructed (Pajek's datasets, 2002).

The citation network analysis started with the paper (Garfield et al., 1964) in which, on the example of Asimov's history of DNA, it was shown that the analysis "*demonstrated a high degree of coincidence between an historian's account of events and the citational relationship between these events*". The next step was made by (Hummon and Doreian, 1989). They proposed three indices (NPPC, SPLC, SPNP) – weights of arcs that provide us with automatic way to identify the (most) important part of the citation network – the main path analysis. We developed algo-

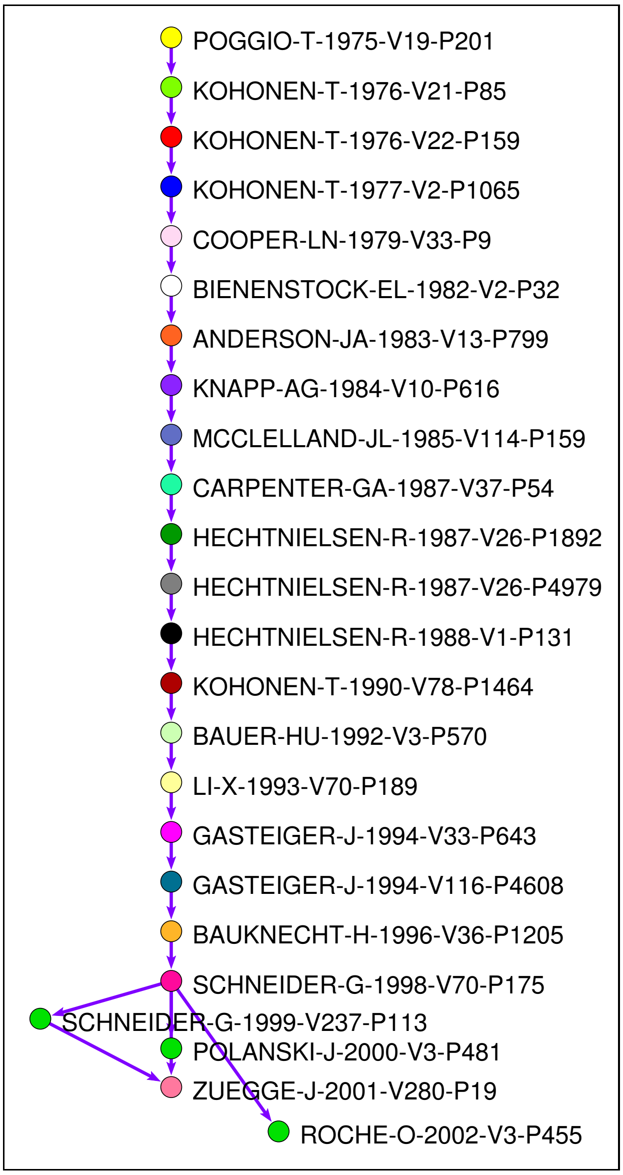
POGGIO-T-1975-V19-P201
KOHONEN-T-1976-V21-P85
KOHONEN-T-1976-V22-P159
KOHONEN-T-1977-V2-P1065
COOPER-LN-1979-V33-P9
BIENENSTOCK-EL-1982-V2-P32
ANDERSON-JA-1983-V13-P799
KNAPP-AG-1984-V10-P616
MCCLELLAND-JL-1985-V114-P159
CARPENTER-GA-1987-V37-P54
HECHTNIELSEN-R-1987-V26-P1892
HECHTNIELSEN-R-1987-V26-P4979
HECHTNIELSEN-R-1988-V1-P131
KOHONEN-T-1990-V78-P1464
BAUER-HU-1992-V3-P570
LI-X-1993-V70-P189
GASTEIGER-J-1994-V33-P643
GASTEIGER-J-1994-V116-P4608
BAUKNECHT-H-1996-V36-P1205
SCHNEIDER-G-1998-V70-P175
SCHNEIDER-G-1999-V237-P113
POLANSKI-J-2000-V3-P481
ZUEGGE-J-2001-V280-P19
ROCHE-O-2002-V3-P455

Figure 3: Main path in SOM citation network.

rithms to efficiently compute the Hummon and Doreian's weights (Batagelj, 2002), so that they can be used also for analysis of very large citation networks with several thousands of vertices.

In Figure 3 we present the main path determined in the SOM (self-organizing maps) citation network (4470 vertices and 12731 arcs).

For the 2001 Graph-Drawing Conference Contest the contest graph A was a self-citing network (GD01 contest, 2001) of GD Conference proceedings. There is a vertex for every paper in the proceedings of GD94 to GD2000, and an arc, if a paper refers to another GD paper.

5. Text analysis networks

5.1. Reuters terror network

Centering Resonance Analysis (CRA) is a new text analysis technique developed by Steve Corman and Kevin Dooley at Arizona State University (CRA, 2001). It uses natural language processing and network text analysis techniques to produce abstract representations of texts. For demonstration of CRA they produced and analyzed several networks. Among them also the ***Reuters terror news network*** that is based on all news released during 66 consecutive days by the news agency Reuters concerning the September 11 attack on the U.S., beginning at 9:00 AM EST 9/11/01. The vertices of a network are words; there is an edge between two words iff they appear in the same text unit (sentence). The weight of an edge is its frequency.

This network was selected by ***Viszards*** (network visualization group) as the case study network for a special visualization session on the Sunbelt XXII International Sunbelt Social Network Conference, New Orleans, USA, 13-17. February 2002. Different approaches to the analysis of the ***Reuters terror news network*** were presented.

We transformed the sequence of CRA networks into a single Pajek's temporal network and analyzed it using Pajek (Batagelj and Mrvar, 2002). It has $n = 13332$ vertices (different words in the news) and $m = 243447$ edges. We present here only two results.

We identified, using cores (Batagelj and Zaveršnik, 2001), in the total network the most important words and determined their layout. Then we produced a sequence of pictures (one for each day) displaying the changes of the news attention. In Figure 4 a picture for the 58th day is presented. The pictures were realized using SVG with the Javascript support for interactive viewing at different levels.

The second picture in Figure 5 presents a segment of the display of the total matrix of 1111 most important vertices (determined by a cut). To nevtralize the most frequent words we normalized the matrix using the ***geometric normalization***.

$$\mathrm{Geo}_{uv} = \frac{w_{uv}}{\sqrt{w_{uu}w_{vv}}}$$

Different stories appeared as connected components.

5.2. Other

Another source of temporal network data are the Keds encodings of news (KEDS, 2002).

For the 1999 Graph-Drawing Conference Contest the contest graph A was a temporal network representing different relations among characters in German TV series 'Lindenstrasse' (GD99 contest, 1999).

For some additional ideas on text analysis read (Popping, 2000).

6. Conclusions

The transformation of textual data into a corresponding network is much easier if the data are structured using some kind of ***markup*** such as (TEI, 2002; Reuters, 2002). The spread of XML based applications will contribute a lot in this direction.

We expect also many applications of network analysis in the implementations of the ***Semantic Web*** (Berners-Lee et al., 2001; Marko et al., 2002; SemanticWeb, 2002; W3C Semantic Web, 2002).

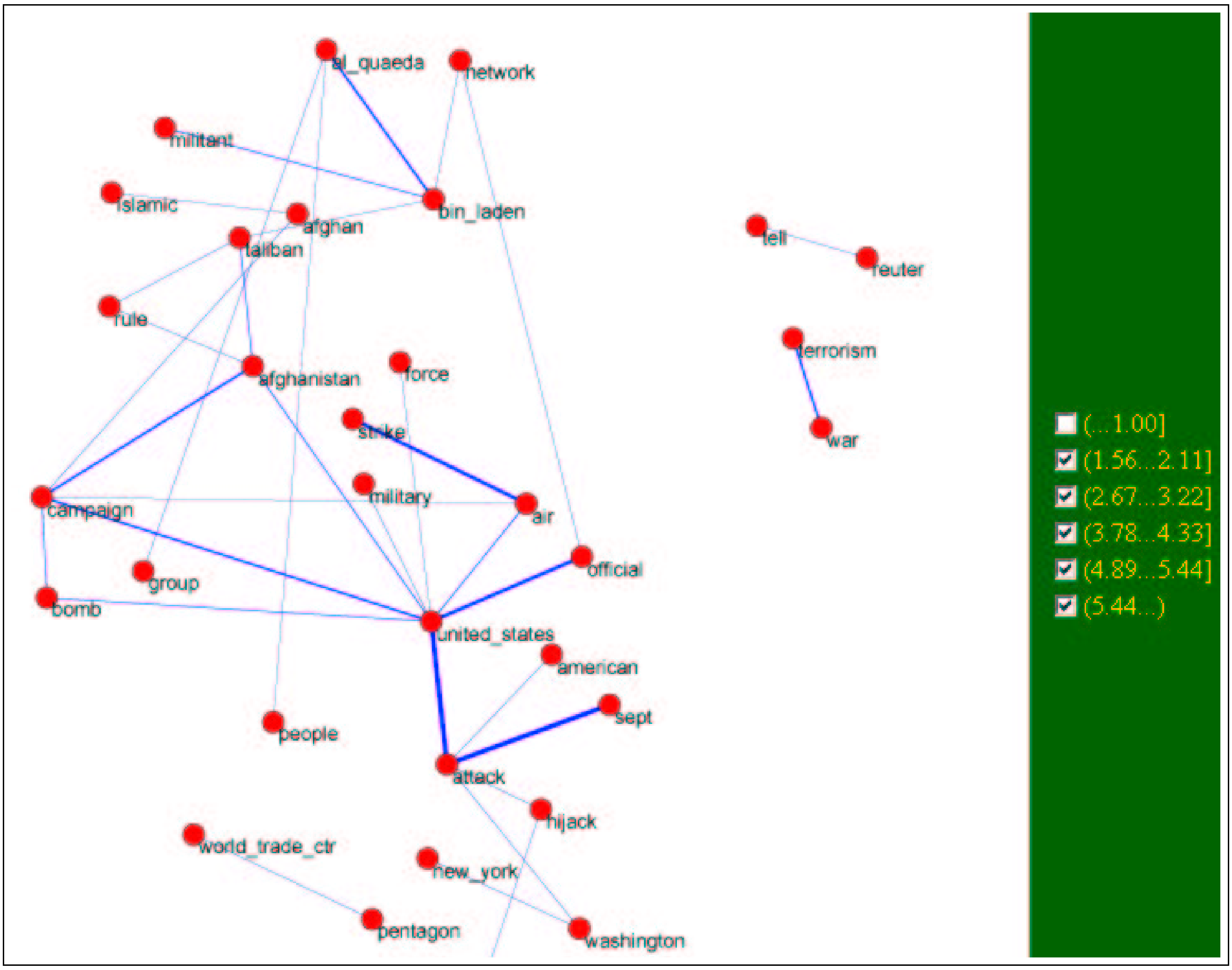

Figure 4: Main links in messages on the 58th day.

7. References

V. Batagelj and A. Mrvar. 1998. Pajek – A program for large network analysis. *Connections*, 21(2):47–57. http://vlado.fmf.uni-lj.si/pub/networks/pajek/.

V. Batagelj and A. Mrvar. 2000. Some Analyses of Erdős Collaboration Graph. *Social Networks*, 22:173-186. http://vlado.fmf.uni-lj.si/pub/networks/doc/erdos/.

V. Batagelj and A. Mrvar. 2002. Reuters terror news network analysis with pajek. To appear in JoSS http://www2.heinz.cmu.edu/project/INSNA/joss/.

V. Batagelj and M. Zaveršnik. 2001. An $O(m)$ algorithm for cores decomposition of networks. Submitted.

V. Batagelj and M. Zaveršnik. 2002. Generalized cores. Submitted.

V. Batagelj, A. Mrvar, and M. Zaveršnik. 2002. Network analysis of dictionaries. To appear in Information Society'02, Language technologies proceedings.

V. Batagelj. 2002. Efficient algorithms for citation network analysis. Submitted.

N. H. F. Beebe. 2002. Bibliographies page. http://www.math.utah.edu/~beebe/bibliographies.html.

T. Berners-Lee, J. Hendler, and O. Lassila. 2001. The Semantic Web. *Scientific American*.

CRA. 2001. Analyses of news stories on the terrorist attack. http://locks.asu.edu/terror/.

FOLDOC. 2002. Free on-line dictionary of computing. http://wombat.doc.ic.ac.uk/foldoc/.

E. Garfield, I. H. Sher, and R. J. Torpie. 1964. The use of citation data in writing the history of science. http://www.garfield.library.upenn.edu/papers/useofcitdatawritinghistofsci.pdf.

GCIDE_XML. 2002. The GNU version of the collaborative international dictionary of English, presented in the extensible markup language. http://www.ibiblio.org/webster/.

GD01 contest. 2001. GD proceedings self-citation network. http://www.infosun.fmi.uni-passau.de/GD2001/graphA/.

GD99 contest. 1999. 'Lindenstrasse' network. http://kam.mff.cuni.cz/conferences/GD99/contest/graphs/A.html.

GNUsl. 2002. prosto programje in slovenščina. http://nl.ijs.si/GNUsl/.

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

ISI. 2002. Web of science.

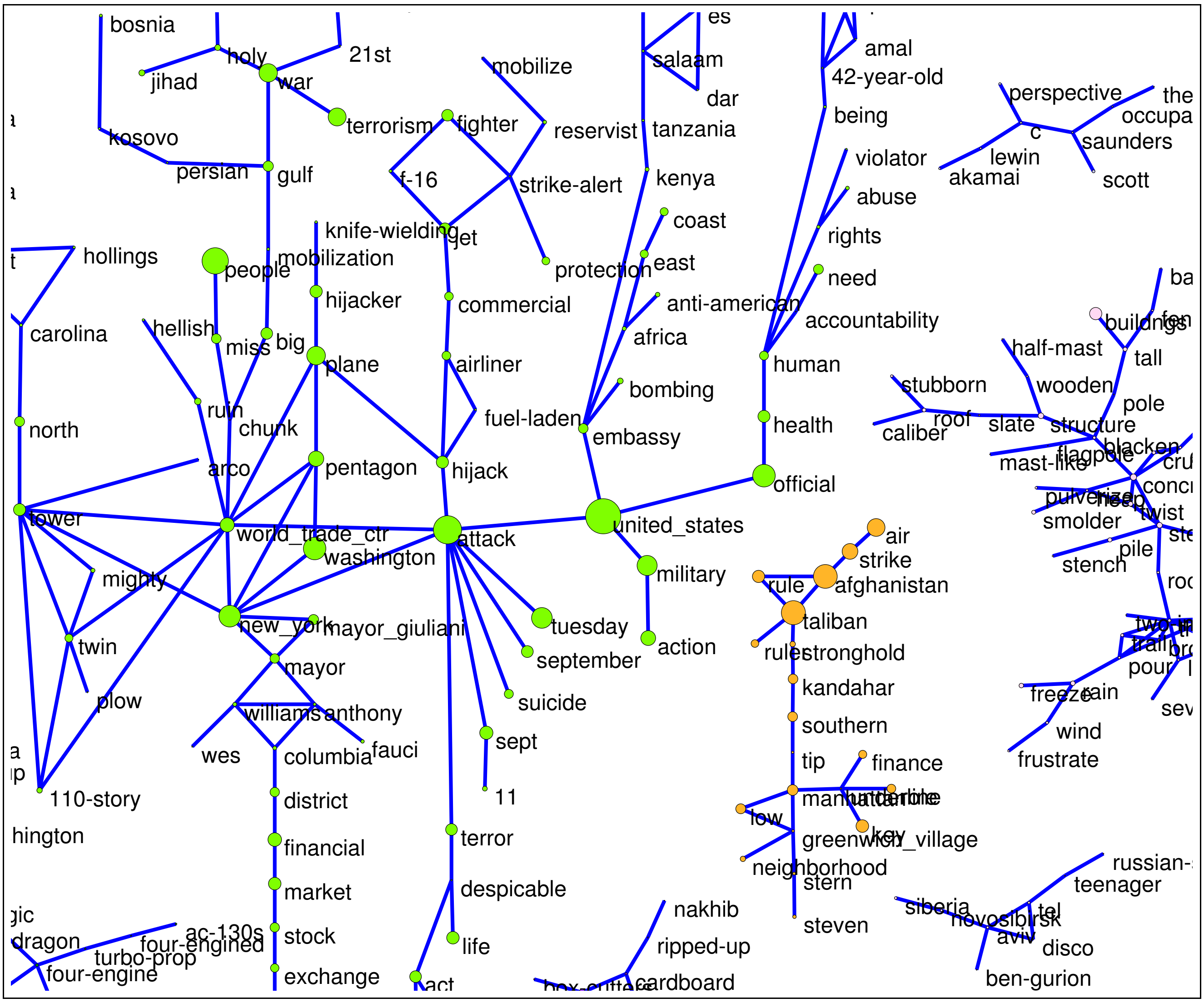

Figure 5: Geometric normalization.

http://www.isinet.com/isi/products/citation/wos/.

B. Jones. 2002. Computational geometry database. http://compgeom.cs.uiuc.edu/~jeffe/compgeom/ biblios.html, ftp://ftp.cs.usask.ca/pub/geometry/.

KEDS. 2002. Kansas event data system. http://www.ukans.edu/~keds/.

D. E. Knuth. 1993. *The Stanford GraphBase: A platform for combinatorial computing*. ACM Press and Addison-Wesley, New York. http://www-cs-faculty.stanford.edu/~knuth/sgb.html.

Lexical FreeNet. 2002. Connected thesaurus. http://www.lexfn.com/.

M. Marko, M. A. Porter, A. Probst, C. Gershenson, and A. Das. 2002. Transforming the World Wide Web into a complexity-based semantic network. http://arxiv.org/html/cs.NI/0205080.

NL. 2002. Natural language server at dept. of intelligent systems institute “Jožef Stefan”. http://nl.ijs.si/.

ODLIS. 2002. Online dictionary of library and information science. http://vax.wcsu.edu/library/odlis.html.

Pajek’s datasets. 2002. Citation networks. http://vlado.fmf.uni-lj.si/pub/networks/data/cite/.

R. Popping. 2000. *Computer-assisted Text Analysis*. Sage, London.

Reuters. 2002. Corpus XML. http://about.reuters.com/researchandstandards/corpus/.

SemanticWeb. 2002. The semantic web community portal. http://www.semanticweb.org/.

TEI. 2002. Consortium Website. http://www.tei-c.org/.

W3C Semantic Web. 2002. http://www.w3.org/2001/sw/.

R. J. Wilson and J. J. Watkins. 1990. *Graphs, An Introductory Approach*. Wiley. translation in slovene: DMFA RS, Ljubljana, 1997.

WordNet. 2002. A lexical database for the English language. http://www.cogsci.princeton.edu/~wn/.