Network analysis of dictionaries

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 the network analysis approaches to analysis of on-line dictionaries are presented. The proposed methods are illustrated by analyses of two such dictionaries: **ODLIS** – Online Dictionary of Library and Information Science; and **FOLDOC** – Free Online Dictionary of Computing.

1. Introduction

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); Dictionary of Visual Art (ArtLex, 2002); Online Dictionary of Musical Terms (Creative Music, 2002); Project Gutenberg (Thesaurus, 2002); Connected Thesaurus (Lexical FreeNet, 2002). In the paper the network analysis approaches to analysis of on-line dictionaries are presented.

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). In the network data collection at Pajek's site the ODLIS and FOLDOC networks are also available.

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

2. Dictionaries

In Figure 1 a segment of two descriptions from the **ODLIS** dictionary is presented.

Such dictionary can be transformed into a ***dictionary graph*** – a directed graph $G = (V, R)$: the terms determine the set of ***vertices*** V; and $R \subseteq V \times V$ is the ***described with*** relation on vertices: $uRv \equiv$ the term v is used to describe the term u. In other words – there is an arc $(u, v) \in R$ from term u to term v iff the term v appears in the description of term u.

For example from the segment in Figure 1 we see

$\{$`note area`, `area`, `catalog record`, `work`, ..., `notebook`, `loose-leaf`, `spiral`, ...$\} \subset V_{ODLIS}$

(`note area`, `contents`) $\in R_{ODLIS}$
(`notebook`, `cover`) $\in R_{ODLIS}$
...

A dictionary graph can be constructed for any given dictionary. In the case when the terms in the descriptions are not marked we can approximate the relation ***described with*** by assuming that all appearances of the terms in a description are marked.

Some approaches to analysis and visualization of on-line dictionaries will be presented, demonstrating several options for analysis:

- searching for important, dense or in some other way interesting parts of network;
- searching for important (central) terms in networks;
- visualization of results.

3. Analysis of ODLIS

The **ODLIS** graph has 2909 vertices (terms) and 18419 arcs, 5 of them ***loop***s (links to itself): `book`, `database`, `leaf`, `paper`, `subject`. The ***average (in/out)degree*** is 6.33.

3.1. Important parts of ODLIS graph

The **ODLIS** graph has 11 ***weakly connected components*** (disconnected pieces): one large (2898 vertices), one with 2 vertices (`use life`, `shelf life`), and 9 ***isolated*** vertices: `aristonym`, `bookstall`, `ESL`, `homily`, `literati`, `manifesto`, `patronymic`, `popular name`, `standing committee`.

In the following we shall analyse only the large component. It has 67 ***strongly connected components*** (terms explaining each other 'cyclically') of size at least 2. The large strong component has 1802 vertices, 951 are trivial (1

note area

The area of a catalog record following the physical description which gives the contents of the work, its relationship to other works, and any physical characteristics not included in preceding parts of the bibliographic description. Each note is given a separate paragraph.

notebook

A loose-leaf or spiral binder with flexible or inflexible cardboard or plastic covers, usually filled with blank ruled or unruled leaves for taking notes and/or filing syllabi, reading lists, assignments, and other material pertaining to a specific project or course of study. Also synonymous with laptop computer.

Figure 1: Segment from ODLIS dictionary

vertex) and the sizes of the remaining strong components are between 2 and 5.

The ***diameter*** (the distance between the most distant vertices) of the largest component is 16: (`hieronym` – `netspeak`), see Figure 2.

The core: The notion of core was introduced in (Seidman, 1983). Let $\mathbf{G} = (V, L)$ be a simple graph. A subgraph $\mathbf{H} = (C, L|C)$ induced by the set C is a ***k-core*** or a ***core of order*** k iff $\forall v \in C : \deg_H(v) \geq k$ and $\mathbf{H}$ is a maximum subgraph with this property. The core of maximum order is also called the ***main core***. The ***core number*** of vertex v is the highest order of a core that contains this vertex. Since the set C determines the corresponding core $\mathbf{H}$ we also often call it a core. There exists a very efficient algorithm, linear in number of links, to determine the core numbers of vertices (Batagelj and Zaveršnik, 2001).

The degree $\deg(v)$ can be the number of neighbors in an undirected graph or in-degree, out-degree, in-degree + out-degree, ... determining different types of cores.

The cores are used to identify the dense parts of a given graph.

The **`ODLIS`** graph main core is of order 14 on 94 vertices: each of the 94 vertices has at least 14 arcs to/from other 94 vertices. To obtain an insight into its internal structure we can use a hierarchical clustering on its vertices. Figure 3 presents a part of the obtained dendrogram (clustering tree).

Triangles: For a selected arc $a(u, v) \in R$ there are four different types of directed triangles: **cyc**lic, **tra**nsitive, **in**put and **out**put – see Figure 4.

For a directed simple graph $G = (V, R)$ we can produce the corresponding ***cyclic triangular network*** $\mathbf{N}_{cyc}(G) = (V, R_{cyc}, w)$ determined by G. Its support is a subgraph $G_{cyc} = (V, R_{cyc})$ of G which arcs are defined by $a \in R_{cyc}$ iff $a \in R$ and a is a base of a cyclic triangle. For $a \in R_{cyc}$ the weight $w(a)$ equals to the number of different cyclic triangles in G to which a belongs.

In a similar way we can define also the ***transitive triangular network*** $\mathbf{N}_{tra}(G) = (V, R_{tra}, w)$ (Batagelj and Zaveršnik, 2002).

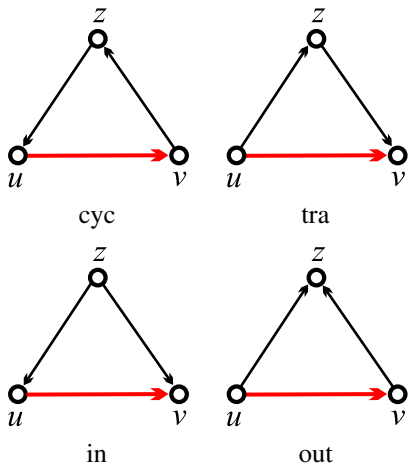

Figure 4: Types of directed triangles

To identify the 'important' parts of G we cut the network $\mathbf{N} = (V, Q, w)$ at selected ***level*** t obtaining the cut-network $\mathbf{N}(t) = (V, Q(t), w)$, where $Q(t) = \{a \in Q : w(a) \geq t\}$. This network is usually further reduced to its connected components of size at least k.

In Figure 5 the arcs belonging to at least 7 cyclic triangles; and in Figure 6 the arcs belonging to at least 11 transitive triangles, are displayed.

Symmetric subnetwork: In a directed network an interesting part is also its ***symmetric subnetwork*** obtained by transforming bidirected arcs to edges, and deleting the remaining arcs.

In the **`ODLIS`** graph its symmetric subnetwork has several connected components: one large (1107 vertices), one with 18 vertices, and one with 11 vertices.

3.2. Centrality measures

In network analysis there exist several measures used to sort out the most important vertices (Wasserman and Faust, 1994). We shall present the results of applying some of them to the **`ODLIS`** graph.

Because of the space limitations we transformed the ta-

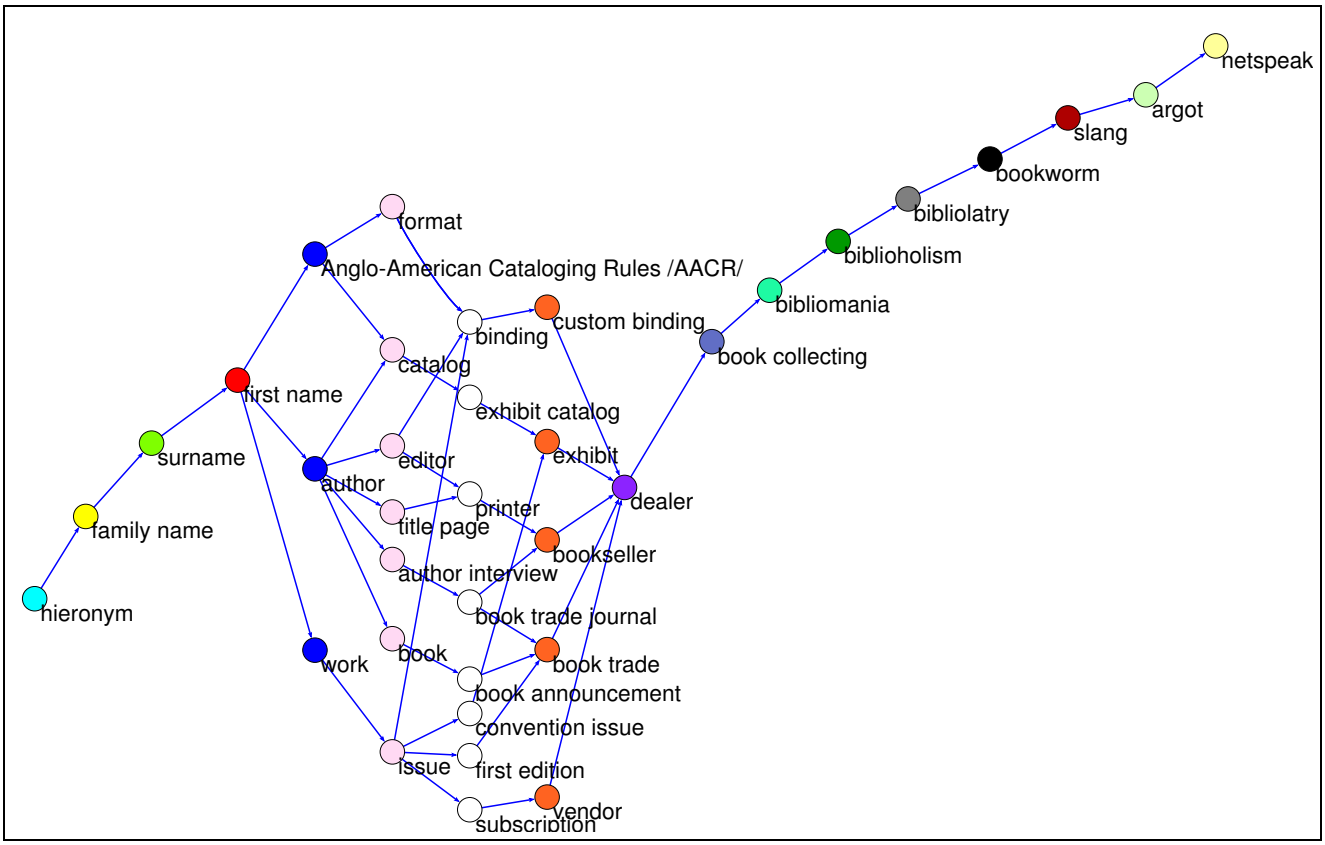

Figure 2: Diametric geodesics in ODLIS main component

bles into lists. The values in brackets are the corresponding values of the measure.

Input degree: identifies the terms most frequently used to explain the others.

book (593), library (557), work (388), printing (382), publishing (366), publication (240), text (237), author (224), information (221), publisher (201), catalog (185), page (177), paper (174), bibliographic item (174), binding (160), title (158), copy (142), homepage (141), issue (137), subject (135).

Output degree: identifies the terms that are not easy to explain (several others are used in the explanation), or complex terms that have several different meanings/ explanations.

periodical (42), catalog (40), bibliography (33), index (32), title (30), editor (29), journal (28), illustration (28), serial (28), parts of a book (27), browse (27), American Library Directory (27), binding (27), plate (27), American Library Association (26), URL (26), edition (25), title page (25), heading (25), series (25).

Hubs and authorities: In directed networks we can usually identify two types of important vertices: hubs and authorities (Kleinberg, 1998).

In the case of dictionaries each term explains something (is an *authority*, other terms point to it). But on the other hand, each term is explained by some other terms (is a *hub*, it points to other terms).

A vertex is a ***good hub***, if it points to many good authorities, and it is a ***good authority***, if it is pointed to by many good hubs.

The hubs and authorities are a refinement of input and output degrees: in the case of input degree we only count incoming lines while for authorities it is important also from whom the lines are coming (important or less important vertices/terms); the same holds for hubs and output degree.

Authorities $\times 1000$**:**

book (497), printing (327), library (325), work (278), publishing (266), publication (190), author (173), text (164), publisher (155), page (147), catalog (131), binding (130), paper (125), title (122), bibliographic item (107), information (102), issue (99), copy (93), periodical (90), illustration (87).

Hubs $\times 10000$**:**

edition (830), book (762), plate (730), ISBN (727), index (719), catalog (708), cover (701), publication history (696), illustration (680), table of contents (639), editor (634), collate (627), heading (622), insert (615), bibliography (612.3), Books in Print (611.6), title page (611), half-title (604), fascicle (602), periodical (600).

Closeness centrality: was introduced by Sabidussi (1966) and in normalized form (in the original definition

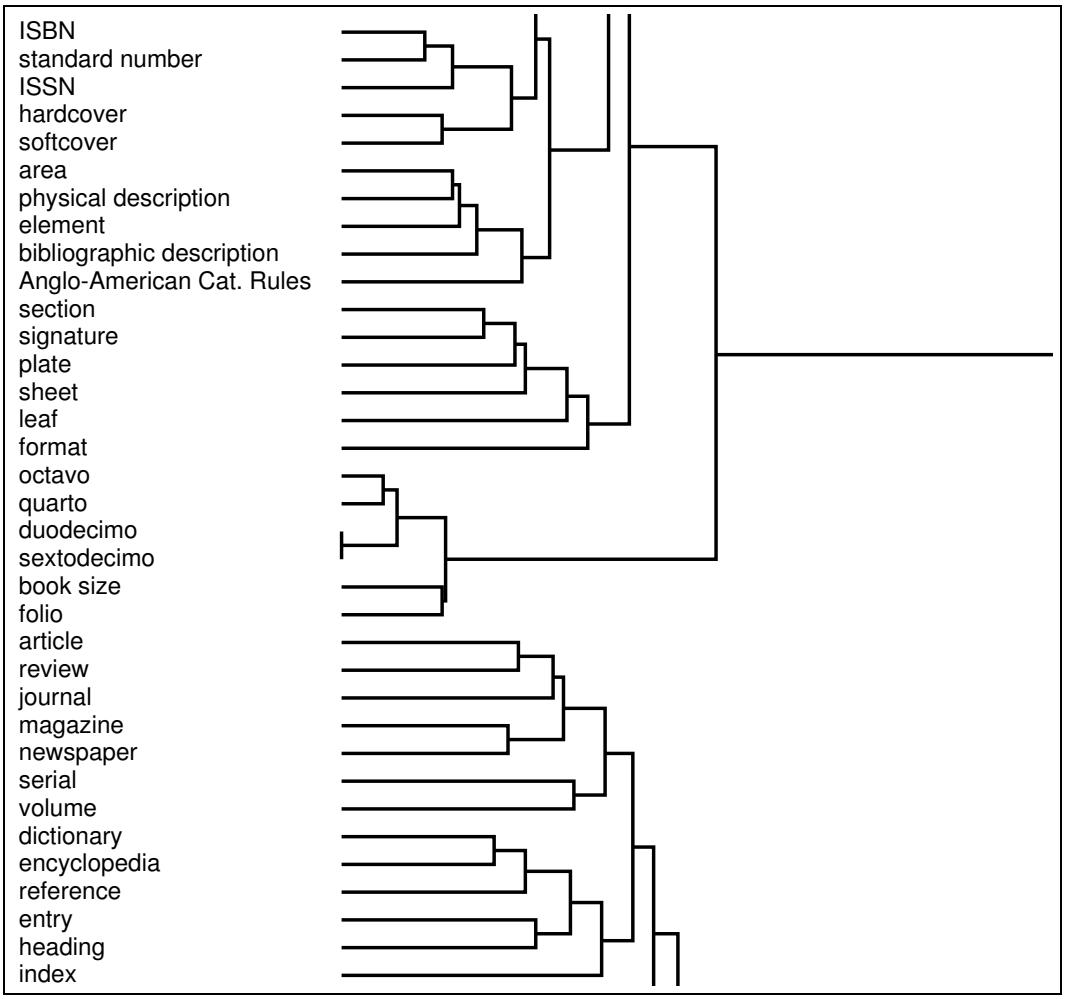

Figure 3: Part of the dendrogram of 14-core

the denominator was 1) by Freeman (1977):

$$cl(u) = \frac{n-1}{\sum_{v \in V} d(u,v)}$$

where $d(u,v)$ is the ***geodesic*** distance (the length of shortest path) from vertex u to vertex v. It is defined only for strongly connected graphs. It holds $0 \leq cl(u) \leq 1$ for every $u \in V$.

This measure is preferable to degree centrality, because it does not take into account only direct links but also indirect connections.

The following results are given for the largest strongly connected component (1802 vertices).

Input Closeness centrality $\times 1000$**:**
`library` (515), `book` (506), `work` (482), `printing` (464), `publishing` (456), `author` (422), `publication` (417), `publisher` (411), `information` (410), `data` (405.4), `issue` (405.3), `catalog` (404), `title` (403), `periodical` (399), `text` (396), `page` (392), `paper` (388), `volume` (378.5), `homepage` (378), `library collection` (377.6).

Output Closeness centrality $\times 10000$**:**
`catalog` (2725), `periodical` (2711), `index` (2666), `browse` (2638), `Books in Print` (2633), `issue` (2630), `text` (2615), `review` (2615), `editor` (2607), `parts of a book` (2600), `new book` (2598), `bibliography` (2596), `insert` (2590), `element` (2586), `Oak Knoll` (2585), `dust jacket` (2583), `masthead` (2581), `serial` (2580), `publication history` (2576), `bibliographic record` (2571).

Betweenness centrality: The idea of betweenness centrality Freeman (1977): vertex is central, if it lies on several geodesics among other pairs of vertices.

Formally the ***betweenness*** of a vertex u is defined by

$$b(u) = \frac{1}{(n-1)(n-2)} \sum_{\substack{v,t \in V: n(v,t) \neq 0 \\ v \neq t, u \neq v, u \neq t}} \frac{n(v,t;u)}{n(v,t)}$$

where $n(v,t)$ is the number of geodesics from v to t and $n(v,t;u)$ is the number of geodesics from v to t passing through u. It holds that $0 \leq b(u) \leq 1$.

In the case of a dictionary: vertices with high betweenness centrality are often used as intermediates for explaining other terms.

Betweenness centrality $\times 1000$**:**
`book` (78), `library` (68), `catalog` (39), `text` (38.5), `periodical` (34), `author` (32), `publishing` (29), `publisher` (26.5), `title` (26), `issue` (25), `illustration` (23), `paper` (22), `printing` (21), `homepage` (20), `library collection` (19.8),

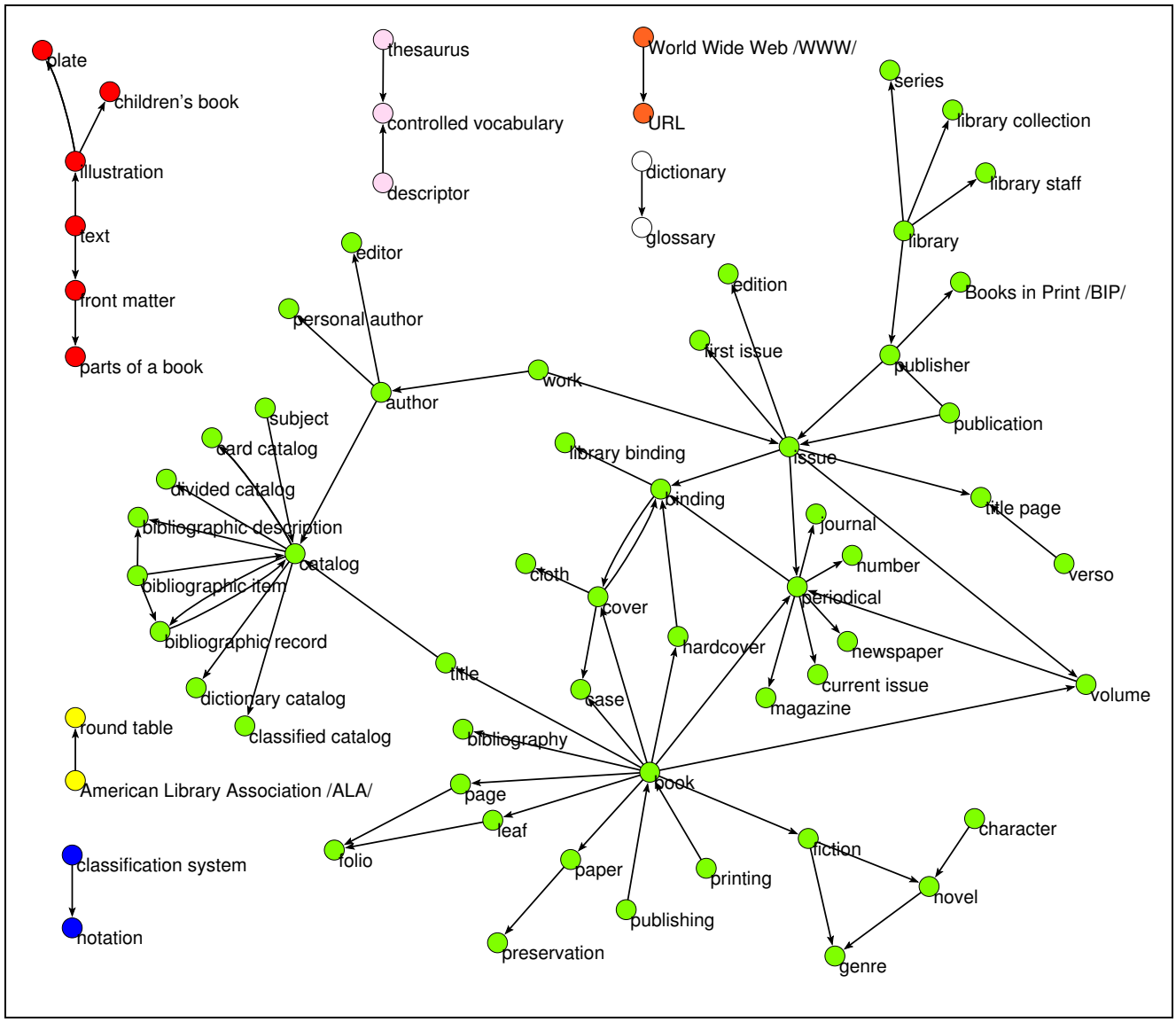

Figure 5: Arcs belonging to at least 7 cyclic triangles

`document` (18), `bibliographic record` (17.5), `copyright` (17.1), `binding` (17), `fiction` (16).

Lords: We call ***lords*** vertices that have 'strong influence' to their neighborhoods. Algorithm for determining lords is simple: at the beginning we assign to each vertex its degree as its initial power. The final distribution of power is the result of 'transferring' the power from weaker to stronger vertices. To determine this distribution we order vertices in the increasing order according to their degrees and in this order we deal the power of the current vertex to its stronger neighbors proportionally.

In the large component of the **ODLIS** graph the only lord is the `book`.

Clustering coefficient $C_2'(v)$: For a given undirected graph $G = (V, E)$ let $|E(G^1(v))|$ denotes the number of lines among vertices in 1-neighbourhood of vertex v, Δ maximum degree of vertex in a network, and $|E(G^2(v))|$ the number of lines among vertices in 1 and 2-neighbourhood of vertex v. Then the ***clustering coefficient*** is defined as

$$C_2(v) = \frac{|E(G^1(v))|}{|E(G^2(v))|}$$

We obtain a more sensitive measure by the following 'correction'

$$C_2'(v) = \frac{\deg(v)}{\Delta} C_2(v)$$

To apply this measure to the **ODLIS** graph we first determined its skeleton – the underlying undirected graph, and apply the measure on it.

Clustering coefficient $C_2'(v) \times 1000$:
`book` (199), `library` (145), `work` (73), `printing` (71), `publishing` (58), `publication` (28), `text` (24), `catalog` (23.4), `author` (23.2), `publisher` (22.8), `page` (18), `binding` (17), `bibliographic item` (15.4), `information` (14.8), `paper` (12.1), `title` (1182), `issue` (11.5), `cover` (11.2), `periodical` (10.5), `database` (9.5).

According to most of the measures the term `book` is absolutely dominating in the **ODLIS** graph.

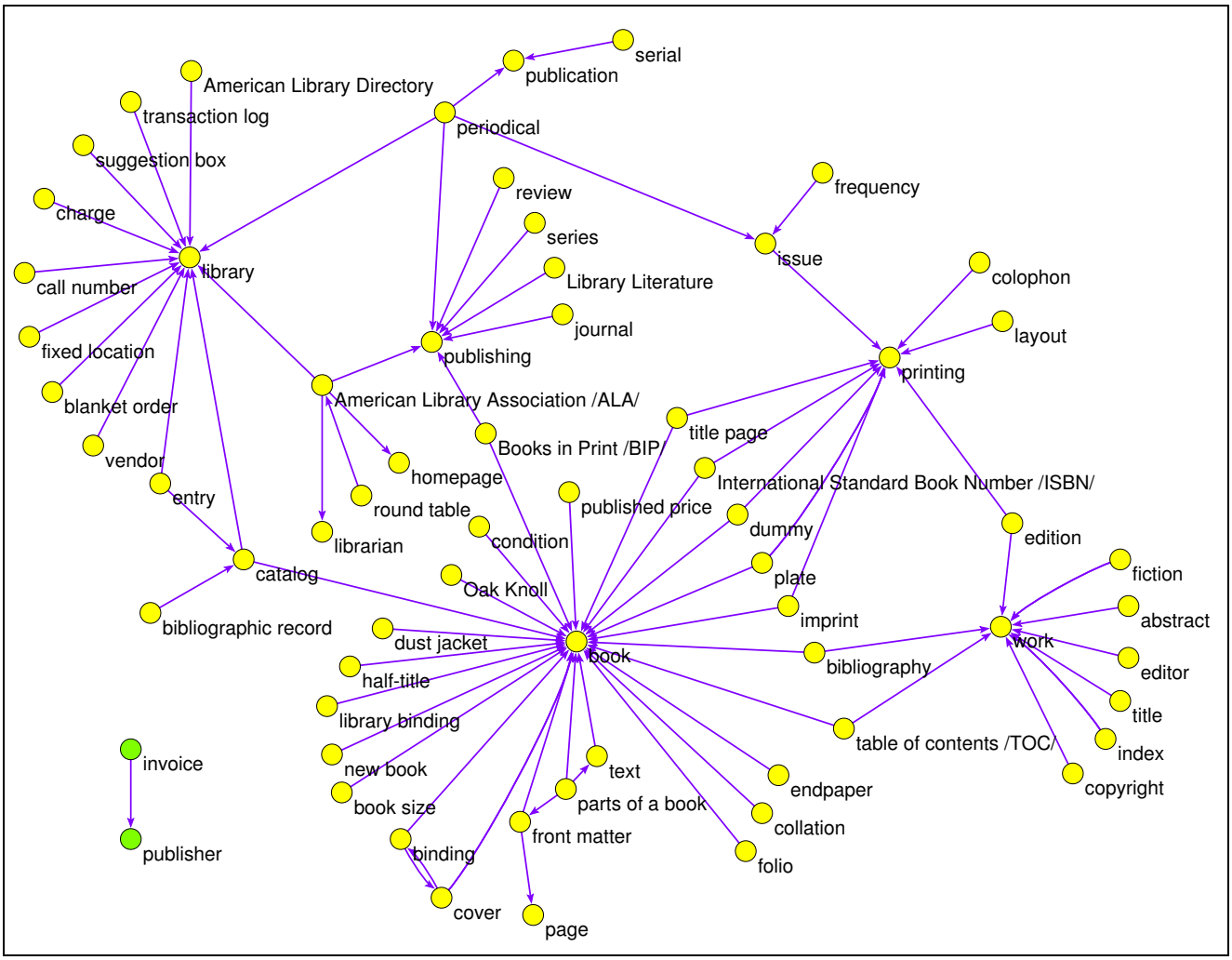

Figure 6: Arcs belonging to at least 11 transitive triangles

4. Analysis of FOLDOC

The **FOLDOC** dictionary is much larger (13425 terms) and we also had to do some 'cleaning' before the analysis.

Two vertices/terms were excluded: `Jargon File` and `TLAs (Three Letter Abbraviations)` since they were very frequent, but they were indicating only the source of the term and not explaining it.

Another problem were terms written in different ways but having the same meaning, such as:

- `American Standard Code for Information Interchange, ASCII`;
- `WWW, Web, World Wide Web`;
- `Win2K, W2K, Windows 2000, Windows NT 5, NT 5`;
- `Microsoft Word, MS Word`;
- `Microsoft Windows, MS Windows`;
- `Abstract Syntax Notation 1, ASN.1`;
- `Uniform Resource Locator, URL`;
- `System V, System 5`;
- `Intel486, 486, i486`;
- `assembly language, ASM`;
- `CD-ROM, Compact Disc`;
- `ISO seven layer, seven layer`;
- `computer virus, virus; ...`

Such terms were identified by preliminary analyses, a ***term equivalence*** partition was prepared and the network was shrunk according to it. Finally we got a cleaned version of the **FOLDOC** dictionary with 13356 vertices (terms), 120238 arcs (links), and no loops. Its average (in/out)degree is 9.003 and the diameter is 15 – connecting `RFC 1446` (Request For Comments) to `UDMA` (Ultra DMA).

4.1. Important parts of network

FOLDOC has only one weakly connected component – it is in one piece. It contains one large strongly connected component with 13247 vertices, and some smaller. The main core is of order 12 with 351 vertices. The symmetric subgraph has one large component (11459 vertices) and several small.

Triangles: The cuts of the cyclic triangular network produce several small components. At level 5 we obtain also some larger components: operating systems and processors (92 vertices, see Figure 7), internet (19), Commodore (13), Motorola (12), java (10), and socket (10). Similar structure we get also from the transitive triangular network cut at level 7. The largest components are: operating systems,

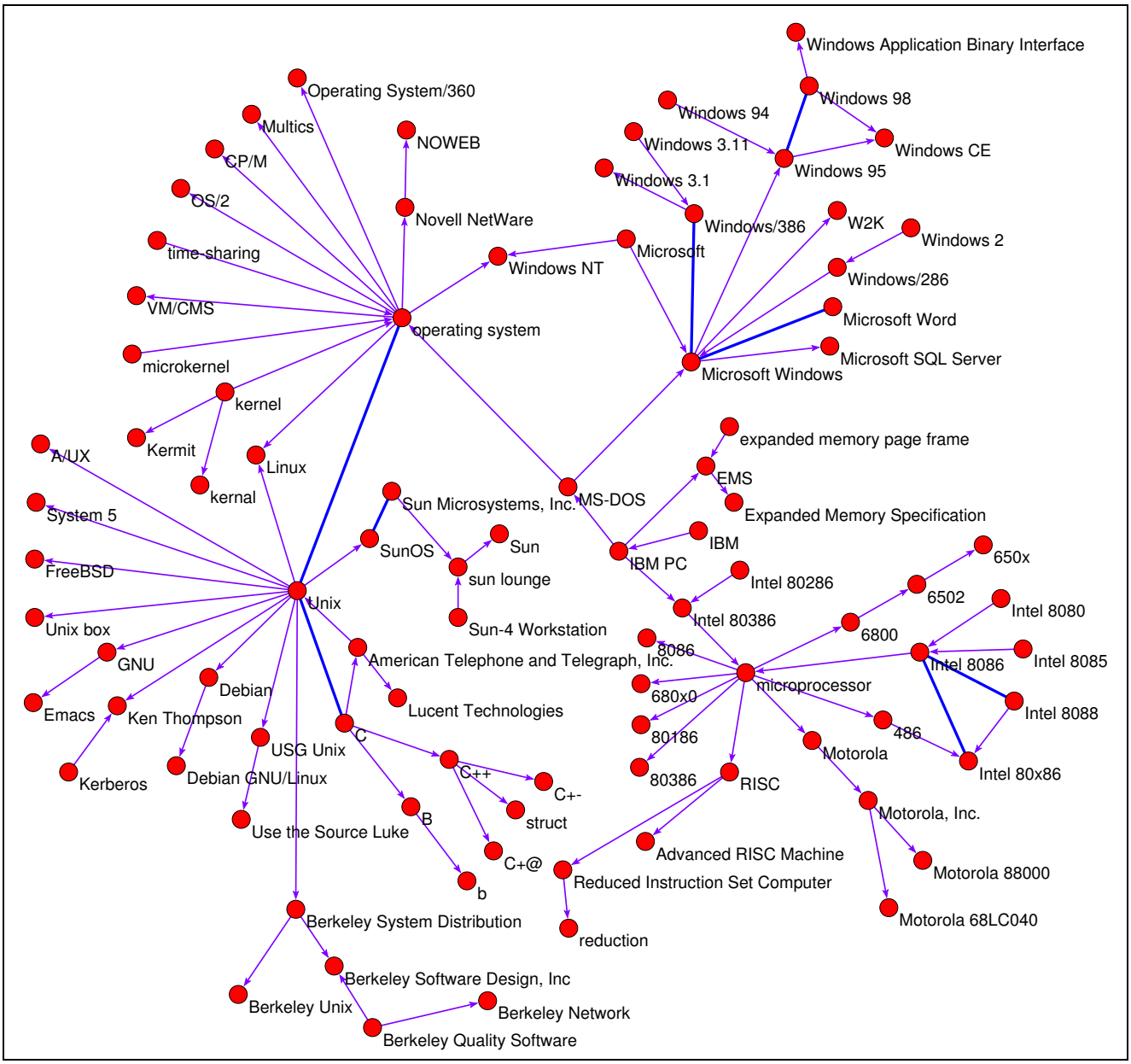

Figure 7: Main component in cyclic triangular network at level 5

Unix related terms, prog. languages (83), processors (20), standards and protocols (12), processor 386 (12), and java (12).

4.2. Centrality measures

Input degree:
Unix (720), C (394), Usenet (388), IBM (383), Internet (366), operating system (360), protocol (336), MS-DOS (259), standard (248), ASCII (241), Macintosh (209), algorithm (208), IBM PC (208), electronic mail (194), country code (190), WEB (189), object-oriented (180), database (173), Microsoft (171), CPU (168).

Output degree:
ASCII (102), operating system (62), Commonwealth Hackish (55), University of Edinburgh (50), chat (49), symbolic mathematics (48), Amiga (48), GCC (47), W2K (46), WEB (42), Java (39), Emacs (38), micro (37), Advanced RISC Machine (37), A# (37), Perl (37), microprocessor (36), computer security (36), NB (35), TeX (35).

Hubs and authorities

Authorities $\times 1000$:
Pentium (265), Internet Explorer (238), MB (237), Usenet (228), IBM (210), operating system (199), server (153), CPU (148), Unix (143), RAM (142), PowerPC (137.5), Linux (137.3), client (136), Solaris (133), Sun (132), AIX (131.7), MIPS (130), web browser (126), DEC Alpha (125.3), SGI (125).

Hubs $\times 1000$:
W2K (905), 486 (153), Emacs (71), CHRP (50), Microsoft Windows (46), AS400 (40.9), System 5 (40.7), WEB (39.8), Windows sockets (33), DEC (32.2), GCOS (32), A# (27.6), MUMPS (27.5), micro (26.8), Windows NT (26.6),

microprocessor(26.4), hs(26.3), Haskell(26.3), Pine(25.4), operating system(25.2).

Closeness centrality: on the largest strongly connected component (13247 vertices).

Input Closeness centrality $\times 1000$**:**
Unix(387), Usenet(368), operating system(353), C(352), Internet(344), IBM(340), MS-DOS(329), protocol(328), standard(325), IBM PC(322), WEB(321), server(319), electronic mail(316), CPU(313), algorithm(311), Macintosh(310), microprocessor(308), ASCII(304.2), Bell Laboratories(303.5), Linux(302).

Output Closeness centrality $\times 1000$**:**
ASCII(250), operating system(242), GCC(240), Emacs(239), A#(239), Java(239), TeX(239), Amiga(239), Perl(238), Commonwealth Hackish(237), computer security(236), FreeBSD(236), W2K(235), J(235), Advanced RISC Machine(235), siod(233), Moscow ML(233), Tcl(232), signature(232), Macintosh user interface(232).

Betweenness centrality $\times 10000$**:**
operating system(599), Unix(560), ASCII(555), C(348), Internet(284), Usenet(260), WEB(248), IBM PC(196), MS-DOS(185), chat(165), electronic mail(163), Macintosh(161), Amiga(152), Lisp(141), standard(132), American Telephone and Telegraph, Inc.(132), microprocessor(122), server(113), Prolog(112), IBM(110).

Lords:
Unix(200052), IBM(23669), Amiga(5433), object-oriented(3218), image(1470), Commonwealth Hackish(1234), LaTeX(1053), Mouse(940), CASE(558), set(471), suit(418), terminal(345), boot(158), Zermelo Frankel set theory(135), elegant(135), forward chaining(128), National Science Foundation(108), token(101), code management(92), ICL(90).

Clustering Coefficient $C_2'(v) \times 1000$**:**
Unix(483), IBM(197), Internet(195), protocol(174), C(169), operating system(167), ASCII(146), Usenet(137), microprocessor(103), IBM PC(103), standard(100), MS-DOS(99), electronic mail(83), WEB(77), OSI(74), Microsoft(74), ISO(71), Microsoft Windows(69), ITU-T(68), algorithm(68).

According to several criteria *Unix* is dominating in this network.

5. Conclusions

Using Pajek we can make also detailed inspections of selected parts of a dictionary network, such as: neighborhood of selected term, common neighbors of selected pair of terms, ...

The results of the analyses can be used also for checking the consistency of a dictionary.

Recent analyses of the frequency distributions and other characteristics of the network based on the Project Gutenberg (Thesaurus, 2002) and some other networks found out that these dictionary networks are also a ***small world*** networks (Motter et al., 2002; Steyvers and Tenenbaum, 2002; Albert and Barabasi, 2002).

6. References

R. Albert and A-L. Barabasi. 2002. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47–97. http://www.nd.edu/~networks/PDF/rmp.pdf.

ArtLex. 2002. Dictionary of visual art. http://www.artlex.com/.

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

V. Batagelj and M. Zaveršnik. 2002. Short cycles connectivity. Submitted.

Creative Music. 2002. Online dictionary of musical terms. http://www.creativemusic.com/features/dictionary.html.

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

J. M. Kleinberg. 1998. Authoritative sources in a hyperlinked environment. V: H. Karloff, ed., *Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms*. SIAM/ACM-SIGACT. http://www.cs.cornell.edu/home/kleinber/auth.ps.

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

A. E. Motter, A. P. S. de Moura, Y.-C. Lai, and P. Dasgupta. 2002. Topology of the conceptual network of language. http://arxiv.org/abs/cond-mat/0206530.

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

S. B. Seidman. 1983. Network structure and minimum degree. *Social Networks*, 5:269–287.

M. Steyvers and J. Tenenbaum. 2002. Small worlds in semantic networks. http://citeseer.nj.nec.com/438759.html.

Thesaurus. 2002. Project Gutenberg. ftp://ibiblio.org/pub/docs/books/gutenberg/etext02/mthes10.zip.

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

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