Evaluation of Internet resources: Bibliometric techniques applications.

How to map the Internet Web ?

An experiment for the biblio-scientometric hosts.

Hervé ROSTAING, Eric BOUTIN, Bruno MANNINA

CRRM, Université Aix-Marseille, France

Lepont, Université Toulon-Var, France

Evaluation of Internet resources

- Data collection
- Quantitative analysis and data set selection for mapping analysis
- Qualitative analysis : network mapping

Data collection

- Query to the Altavista search engine

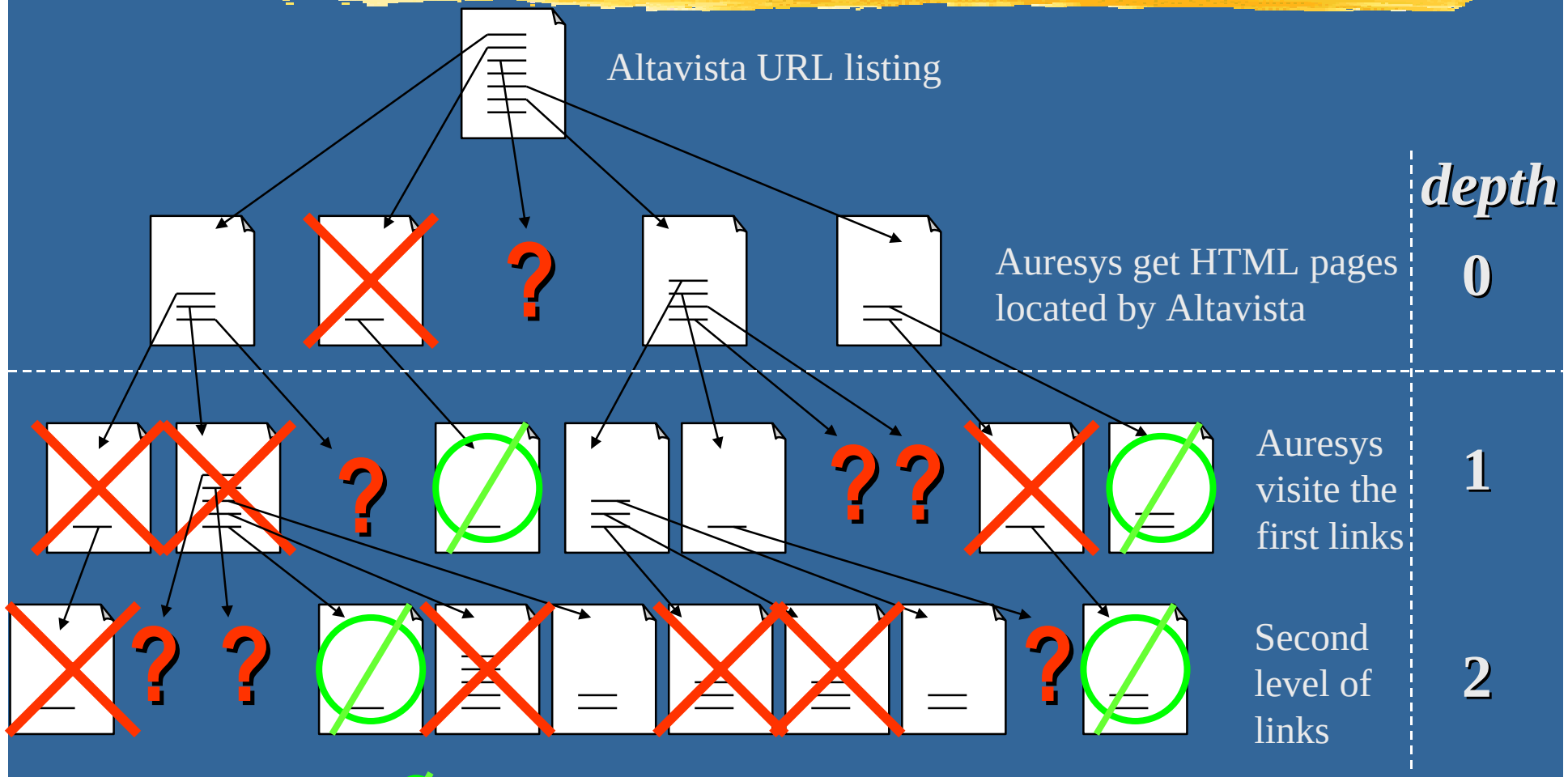
  scientometri or bibliometri* or scientometry or bibliometry*

  – > 3518 URL found
  – > 1010 URL display

  } HTML pages locations

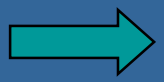
- Auresys robot *(CRRM, http://193.51.109.166)*
  – submit the same query to Altavista
  – propagating search from links present in the 1010 pages
  – build an local HTML pages database responding to the query
  – > collect the HTML pages and exceed the Altavista limit

Auresys propagating search



Auresys storage of HMTL pages

- Storage the HTML pages set after
  - only retaining pages responding to the query
  - removing duplicate pages visited
  - data organization : pages classified according to
    - hosts
    - internet domain
    - relevance index (wais index)
    - page categories : form action, directory, text, normal
  - extracting information for bibliographic reference creation



Bibliometric analysis

Auresys bibliographic reference

```
TIT  : PROGRAMME
IND  :  28
LNG  : english
NMP  : 2230
TAG  : Nom : GENERATOR<br> Content : Mozilla/4.03 [en] (Win95; I) [Netscape];
NIM  : 0
MCT  : scientometri;
DOM  : fr
NMO  : 1
DTR  : Mon Oct 26 21:59:33 1998
DMO  : Tue Aug  4 18:18:38 1998
NFS  : 436
URL  : http://cournot.u-strasbg.fr/divers/apr98.html
HST  : cournot.u-strasbg.fr
HXT  : www.business.auc.dk;
ABS  : and Public Science, Research Policy, Vol.26, pp.317-330. to Laurent BACH: Eval
       of large Research Programmes * Bach L. et al., 1995 "Evaluation of the economi
       effects of BRITE-EURAM programmes on the European industry" Scientometrics,...
AIN  : http://cournot.u-strasbg.fr/divers/;
AEX  : http://www.business.auc.dk/homepage.html;
IMG  : Aucune
PRF  : 0
TFI  : 13834
TST  : 12711
MLS  : Aucun
WAY  : cournot.u-strasbg.fr/divers/apr98.html;
```


Quantitative analysis

Selection of data set for qualitative analysis

Depth	Time spending	Visited pages	Found pages	Found hosts (A set)	Cited hosts	Cited hosts from A (B set)	Citing hosts from B
0	13 h	1010	421	299	1189	64	76
			80				
1	21 h	4029	501	315	1249	67	89
2	44 h	12612	597	321	1367	83	93
3	149 h	37529	783	331	1785	97	97

388 pages

Hosts holding numerous pages

Complete data set

CRRM.UNIV-MRS.FR	39
HUB.IB.HU-BERLIN.DE	17
WWW.THE-SCIENTIST.LIBRARY.UPENN.EI	16
WWW.CHEM.UVA.NL	16
GOPHER.RZ.UNI-DUESSELDORF.DE	15
WWW.INFORMATIK.UNI-TRIER.DE	12
WWW.CINDOC.CSIC.ES	12
TPAC.GCATT.GATECH.EDU	12
WWW.DGPS.DE	11
WWW.UNI-TRIER.DE	10
WWW.ASIS.ORG	10
SAHARA.FSW.LEIDENUNIV.NL	10
WWW.ENSSIB.FR	9
WWW.SRI.COM	8
SYY.OULU.FI	8
SHERLOCK.BERKELEY.EDU	8
AI.IIT.NRC.CA	8
others	562

Selected data set

unselected hosts	397
CRRM.UNIV-MRS.FR	39
HUB.IB.HU-BERLIN.DE	17
WWW.THE-SCIENTIST.LIBRARY.UP	16
WWW.CHEM.UVA.NL	16
WWW.INFORMATIK.UNI-TRIER.DE	12
WWW.CINDOC.CSIC.ES	12
WWW.ASIS.ORG	10
SAHARA.FSW.LEIDENUNIV.NL	10
WWW.ENSSIB.FR	9
WWW.SRI.COM	8
SYY.OULU.FI	8
SHERLOCK.BERKELEY.EDU	8
AI.IIT.NRC.CA	8
WWW.RAND.ORG	7
WWW.LIB.NCSU.EDU	7
WWW.ASLIB.CO.UK	7
WWW-SLIS.LIB.INDIANA.EDU	7
others	185

The most often cited hosts

Complete data set

none	506
CRRM.UNIV-MRS.FR	22
SAHARA.FSW.LEIDENUNIV.NL	13
WWW.PSYCHOLOGIE.UNI-TRIER.DE	11
WWW.PSYCHOLOGIE.UNI-FREIBURG.DE	11
WWW.PSYCHOLOGIE.HU-BERLIN.DE	11
UNION.NCSA.UIUC.EDU	11
EZINFO.UCS.INDIANA.EDU	11
XXX.LANL.GOV	10
WWW.ISINET.COM	10
WWW.ELSEVIER.NL	10
WWW.W3.ORG	9
WWW.INDIANA.EDU	9
WWW.DB.DK	9
WWW.ASIS.ORG	9
WWW.ADOBE.COM	9
WWW.YAHOO.COM	8
WWW.UMU.SE	8
WWW.NLC-BNC.CA	8
WWW.IIT.NRC.CA	8
WWW.CORPSERV.NRC.CA	8
WWW.CINDOC.CSIC.ES	8
INFO.LIB.UH.EDU	8
others	2437

Selected data set

none or unselected hosts	671
CRRM.UNIV-MRS.FR	22
SAHARA.FSW.LEIDENUNIV.NL	13
EZINFO.UCS.INDIANA.EDU	11
WWW.ISINET.COM	10
WWW.ASIS.ORG	9
WWW.UMU.SE	8
WWW.NLC-BNC.CA	8
WWW.CINDOC.CSIC.ES	8
INFO.LIB.UH.EDU	8
WWW.UNI-BIELEFELD.DE	7
WWW.THE-SCIENTIST.LIBRARY.U	7
WWW-SLIS.LIB.INDIANA.EDU	7
WWW.OCLC.ORG	6
WWW.IB.HU-BERLIN.DE	6
WWW.CHEM.UVA.NL	6
SHUM.CC.HUJI.AC.IL	6
SHERLOCK.BERKELEY.EDU	6
COOMBS.ANU.EDU.AU	6
others	157

Hosts having a great citation activity

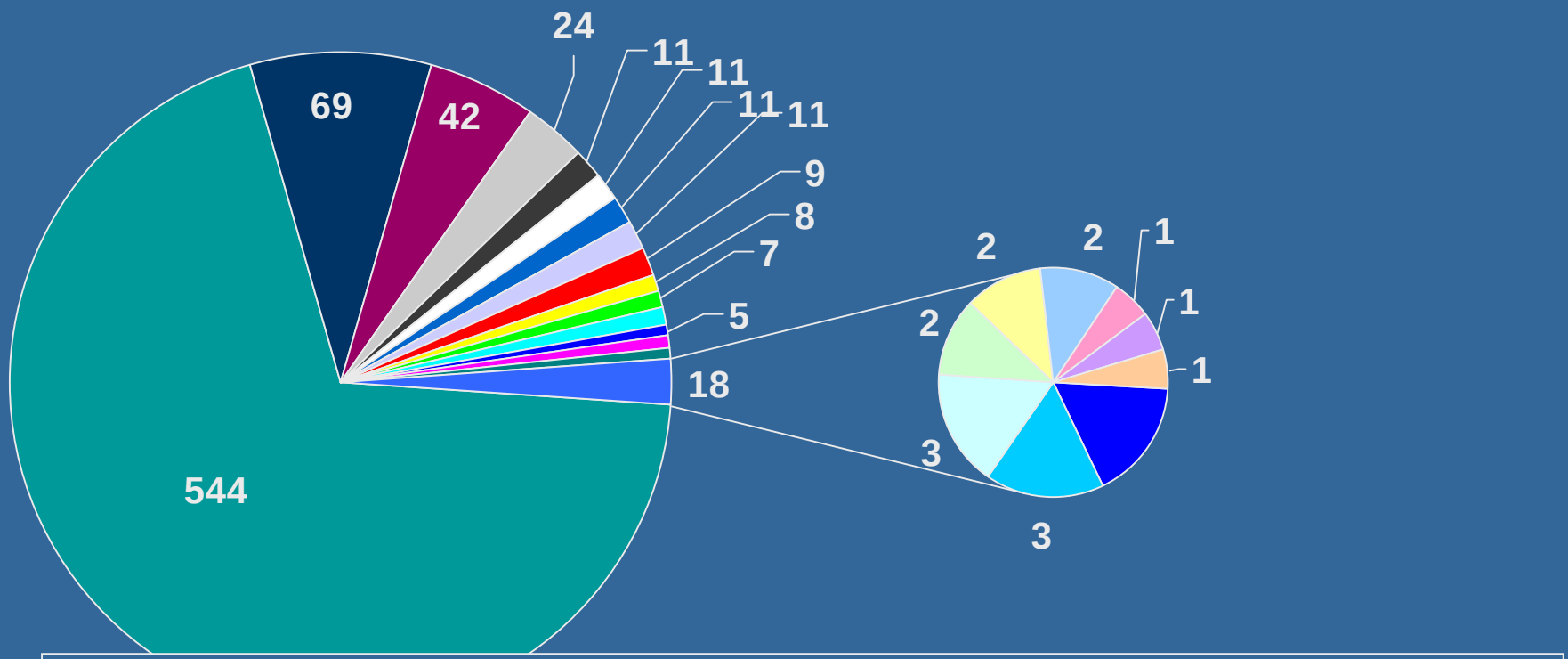
Complete data set

WWW.CINDOC.CSIC.ES	71
TORNADE.ERE.UMONTREAL.CA	31
CRRM.UNIV-MRS.FR	30
WWW.UNI-BIELEFELD.DE	21
OLYMPE.SCINFO.U-NANCY.FR	15
WWW.STEDETSOMIKKEER.DK	11
SUNSITE.INFORMATIK.RWTH-AACHEN.D	9
SHERLOCK.BERKELEY.EDU	8
LUCIEN.SIMS.BERKELEY.EDU	8
WWW.PASTEUR.FR	7
WWW.CHEMHERITAGE.ORG	6
WWW.ASIS.ORG	5
WWW.VALLESNET.ORG	5
WWW.CNAM.FR	5
WWW.INFORMATIK.UNI-TRIER.DE	4
WWW.BIOCHEMSOC.ORG.UK	4
WWW.YAHOO.COM.SG	4
WWW.PHILOS.RUG.NL	4
others cite less than 4 hosts	

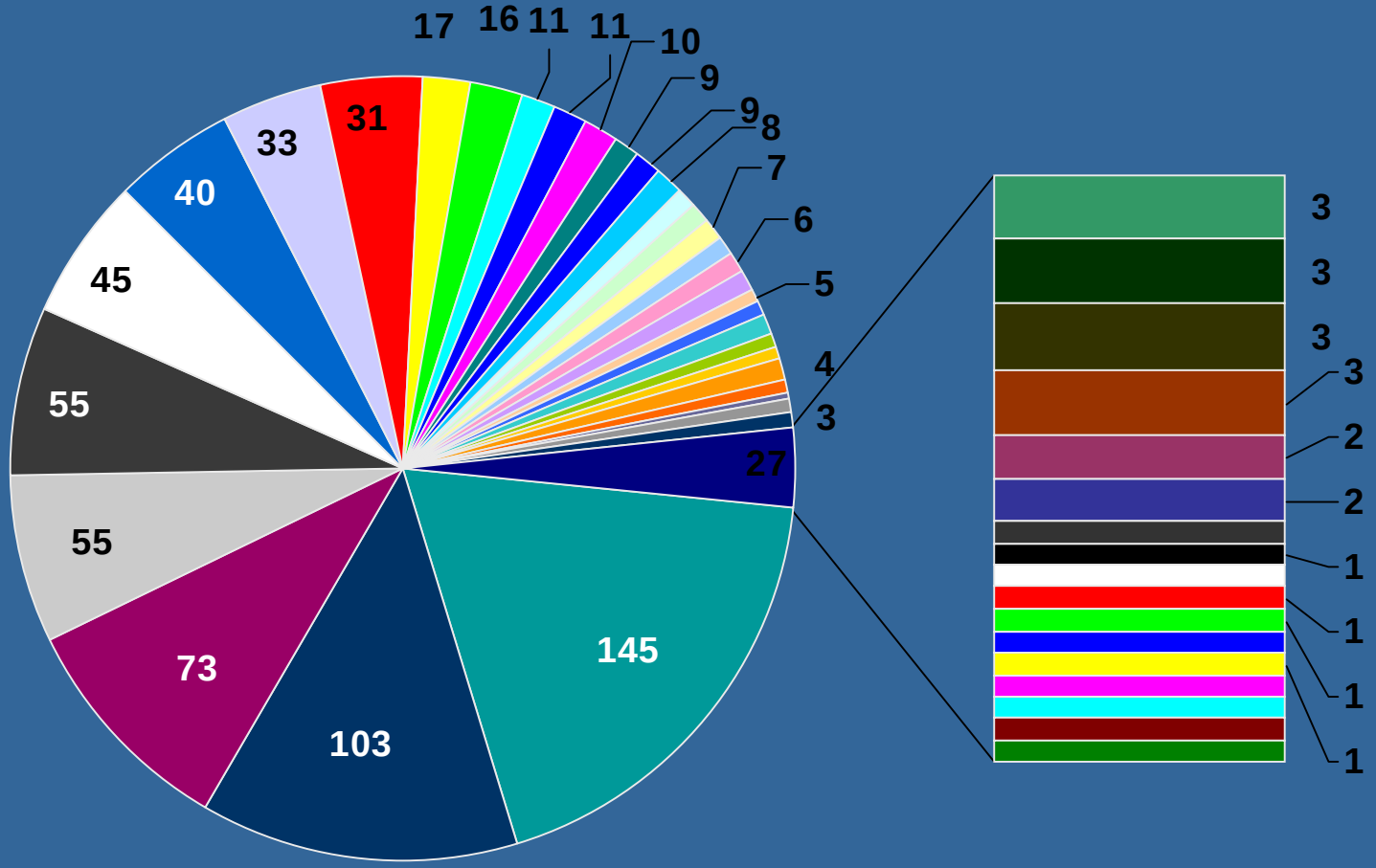
Selected data set

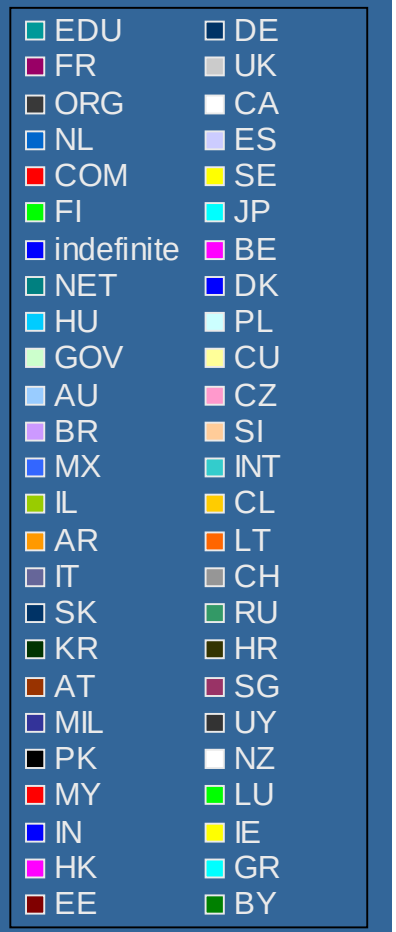
WWW.CINDOC.CSIC.ES	71
TORNADE.ERE.UMONTREAL.CA	30
CRRM.UNIV-MRS.FR	24
WWW.UNI-BIELEFELD.DE	21
SUNSITE.INFORMATIK.RWTH-AACHEN.DE	9
WWW.PASTEUR.FR	7
SHERLOCK.BERKELEY.EDU	6
WWW.ASIS.ORG	4
WWW.INFORMATIK.UNI-TRIER.DE	4
HUB.IB.HU-BERLIN.DE	3
WWW.IB.HU-BERLIN.DE	3
WWW.SLIS.INDIANA.EDU	3
WWW-SLIS.LIB.INDIANA.EDU	3
COOMBS.ANU.EDU.AU	2
EZINFO.UCS.INDIANA.EDU	2
WWW.ENSSIB.FR	2
WWW.LBORO.AC.UK	2
others cite only one host	

Distribution of languages for writing HTML pages in biblio-scientometrics



Distribution of internet domains for biblio-scientometrics hosts





Qualitative analysis :
mapping the Web for biblio-scientometrics field

- Main goal for mapping the Web
  - to create a map useful for web navigation
  - to understand the relationship between hosts belonging to a field of interests
  - to judge the central or peripheral character of hosts in relation to others
  - content analysis not possible : full text analysis in different languages

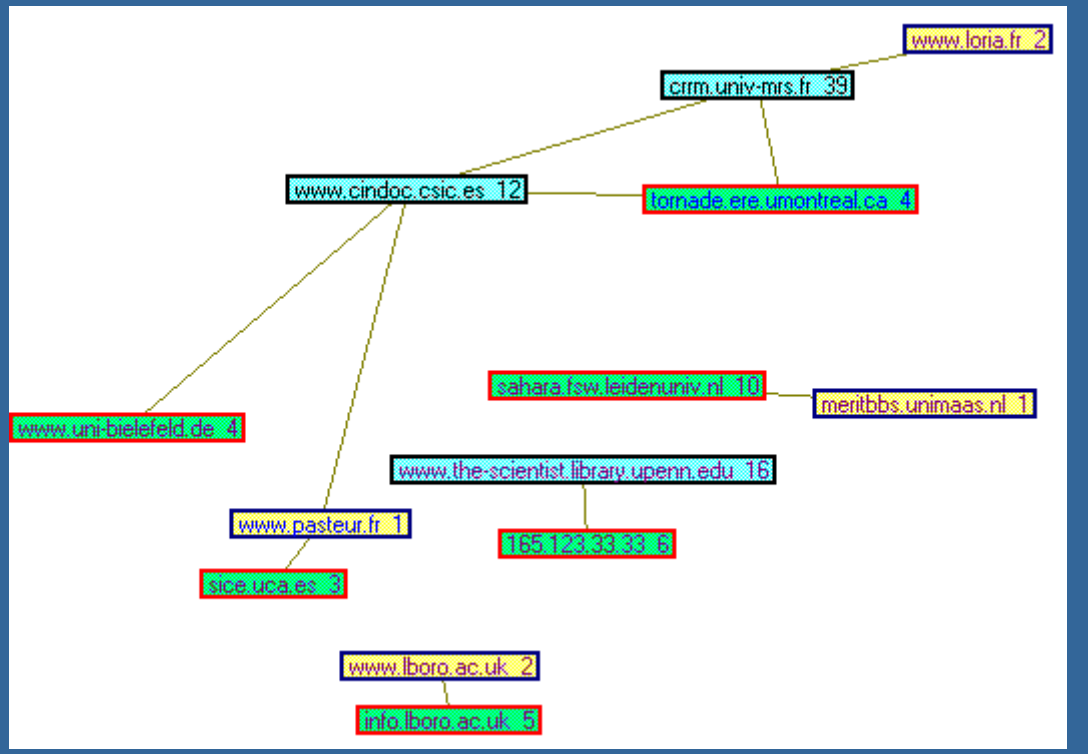
➡ **Mapping the links muddle for host cited and citing positions**

Network mapping : the components

- **node** = host name
- **arc** = relation measure between two nodes

  What relation to measure?

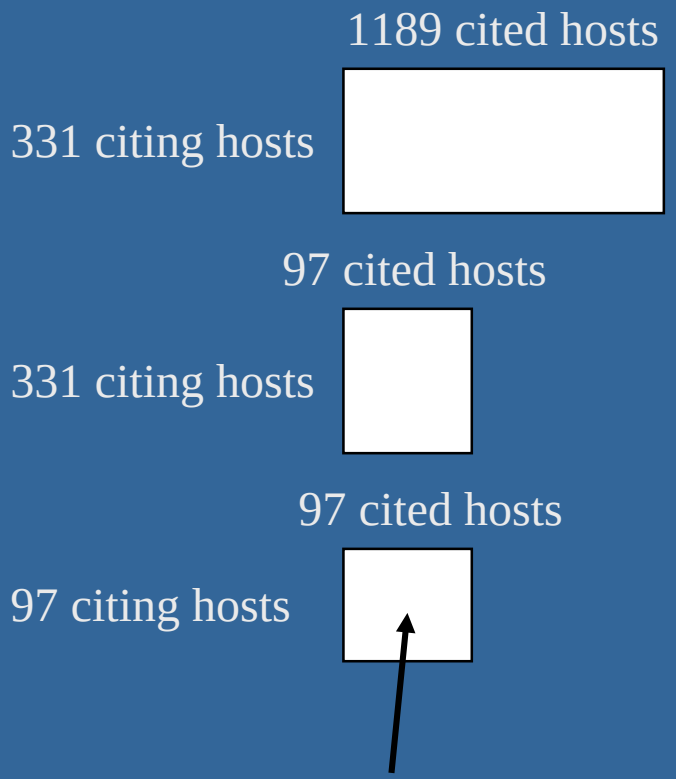
  - Hypertext linking to map the structure of the Web (central and peripheral hosts)
  - Citation phenomena to detect important hosts (valuable hosts)



Constraint : the software need a symmetrical matrix for input

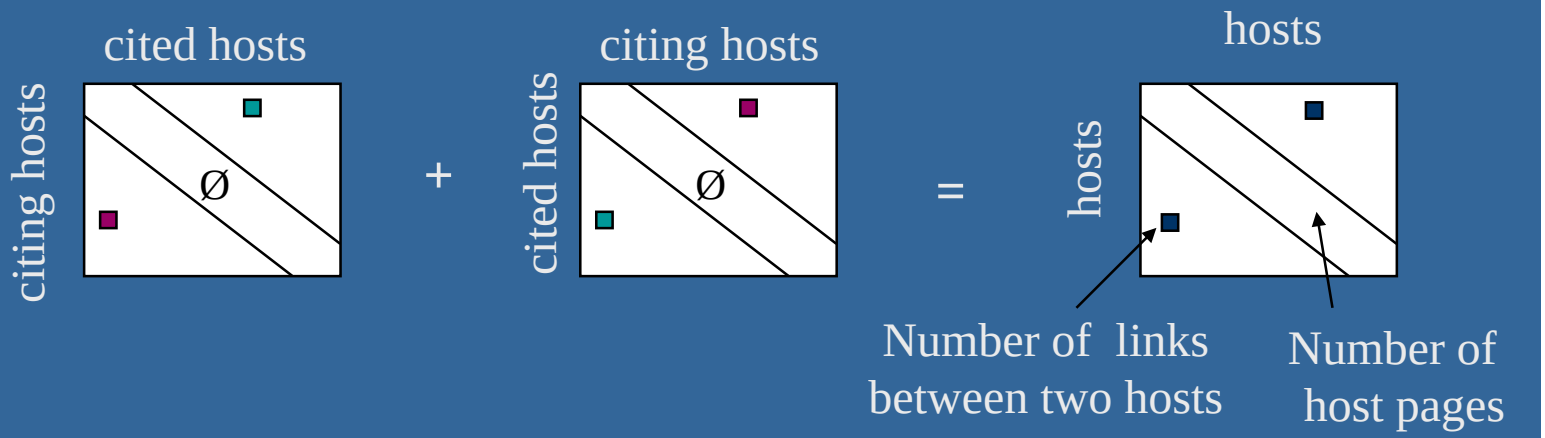
Matrix building

- Cross citation matrix
  - for the whole data set
  - only for biblio-sciento hosts
  - software constraint

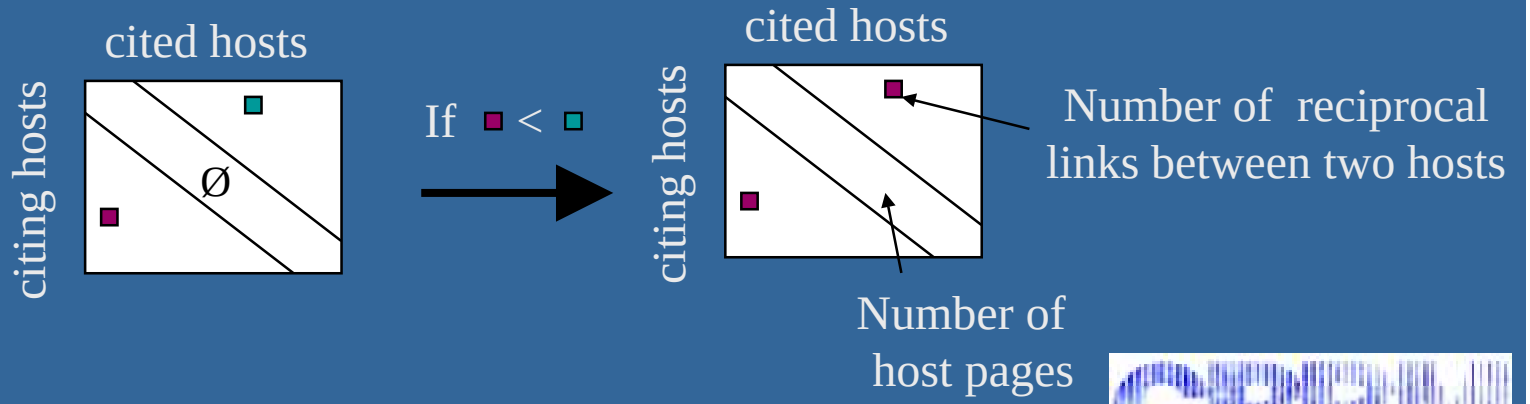


Measure calculation

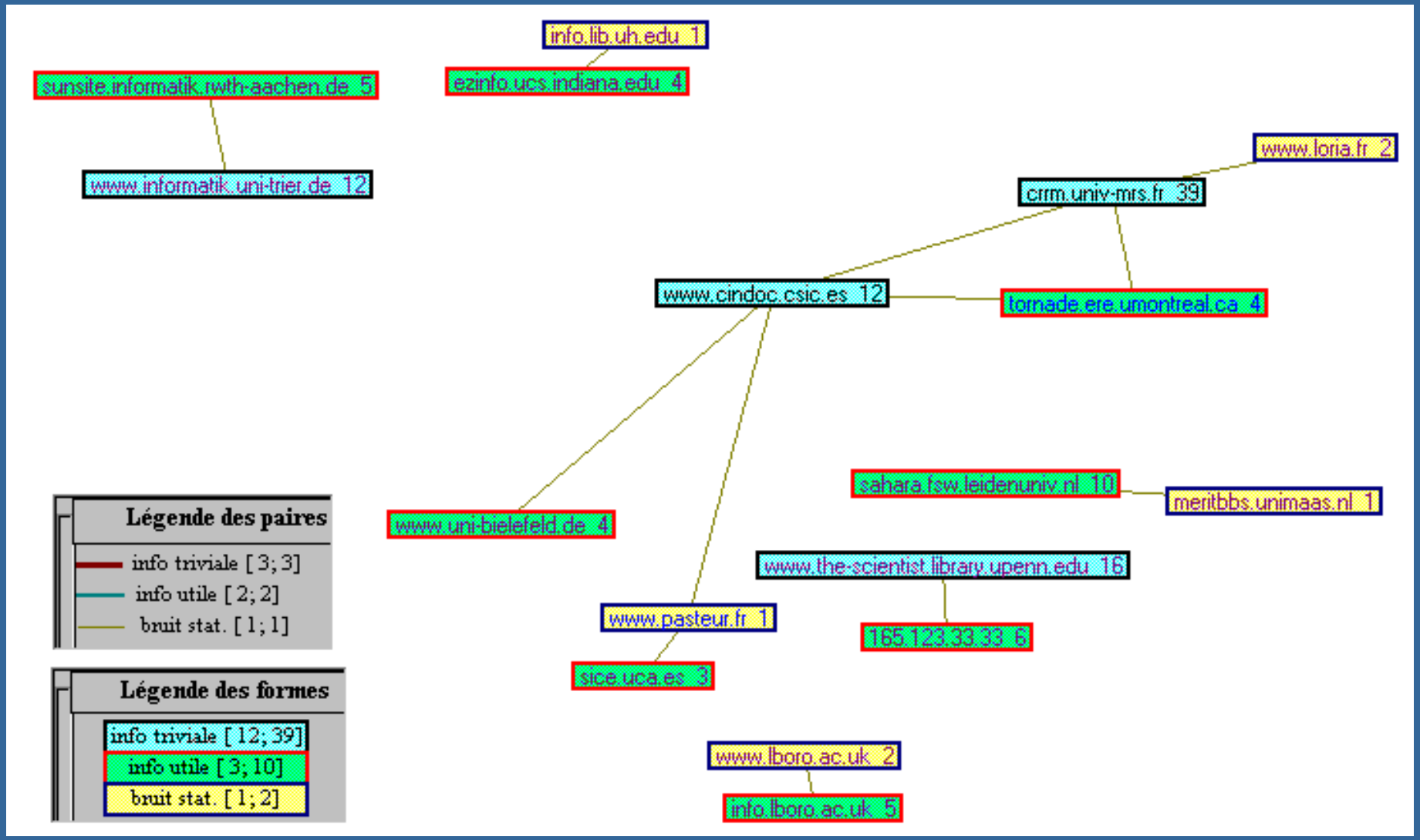
- $X + X^t$



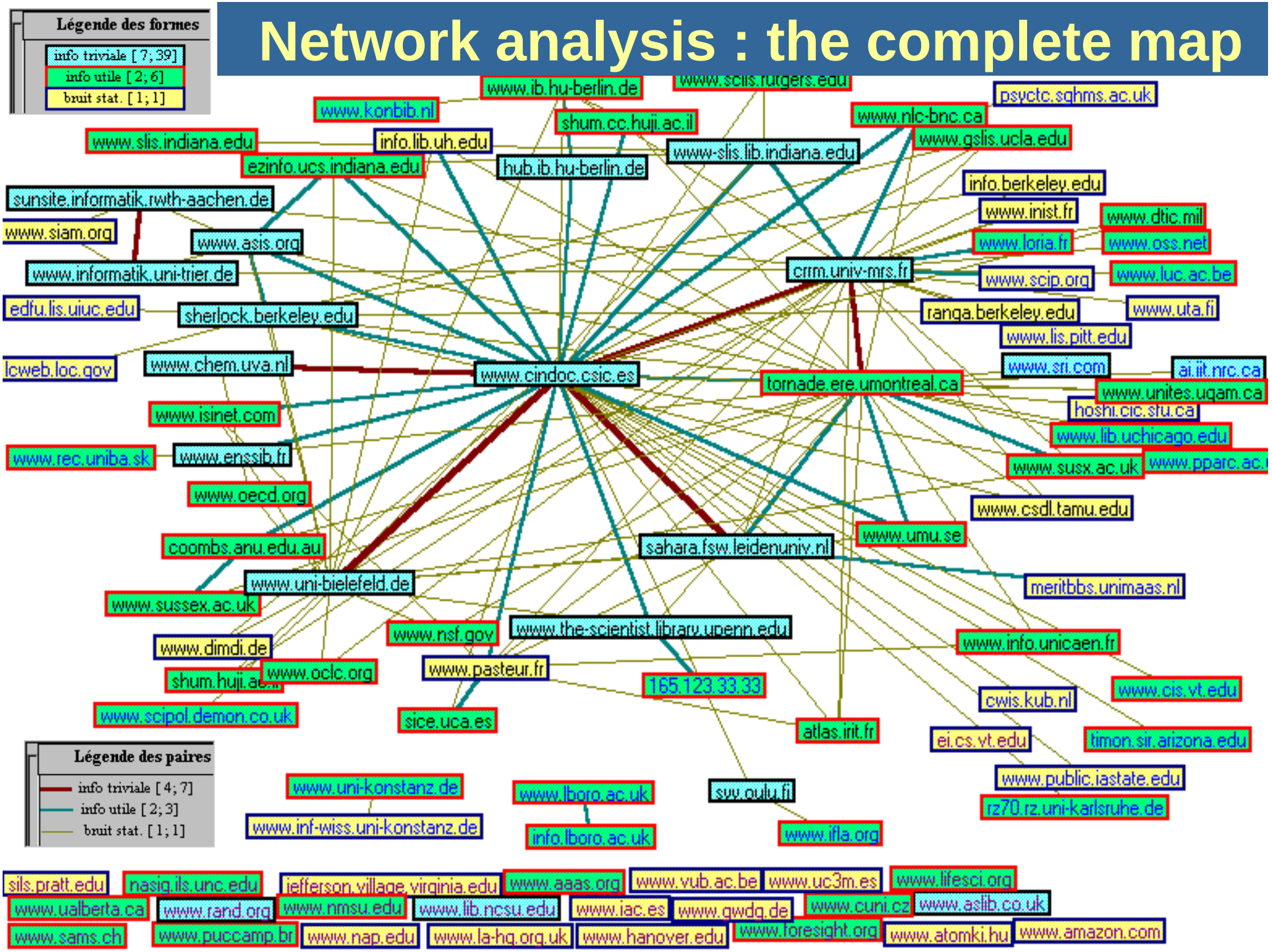
- $x_{ij} = x_{ji} = min \{x_{ij}, x_{ji}\}$



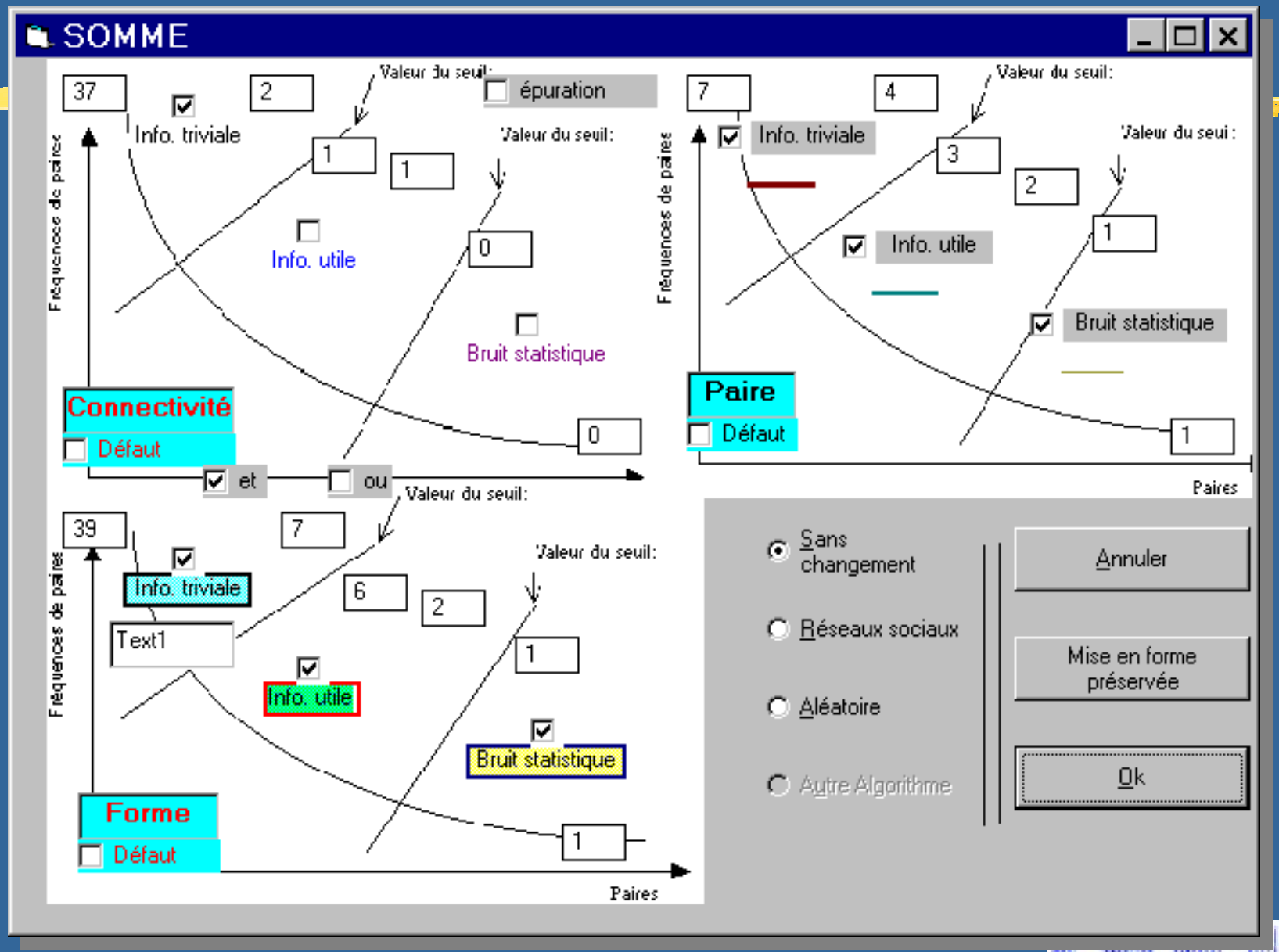
Map of hosts having reciprocal links



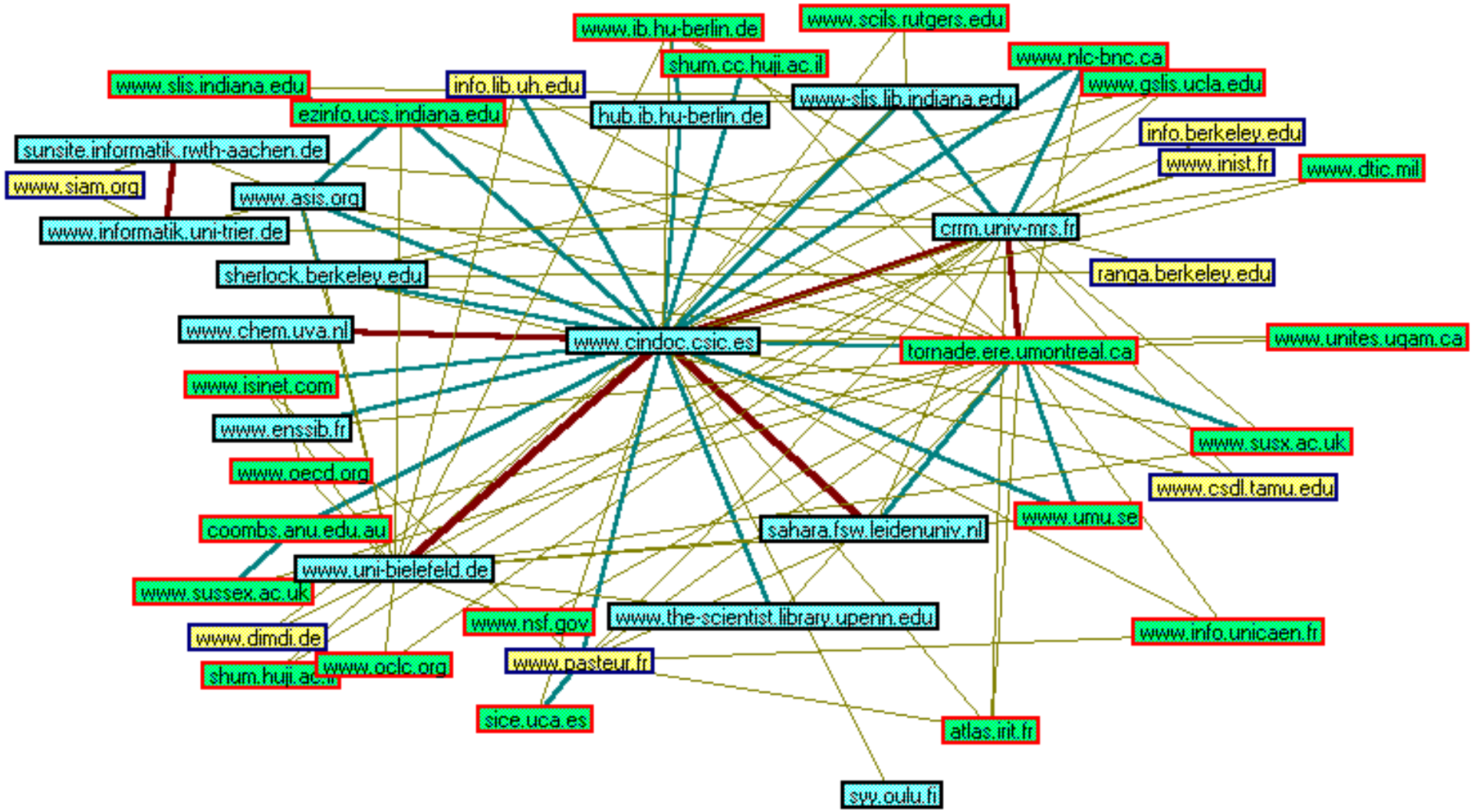
Network analysis : the complete map



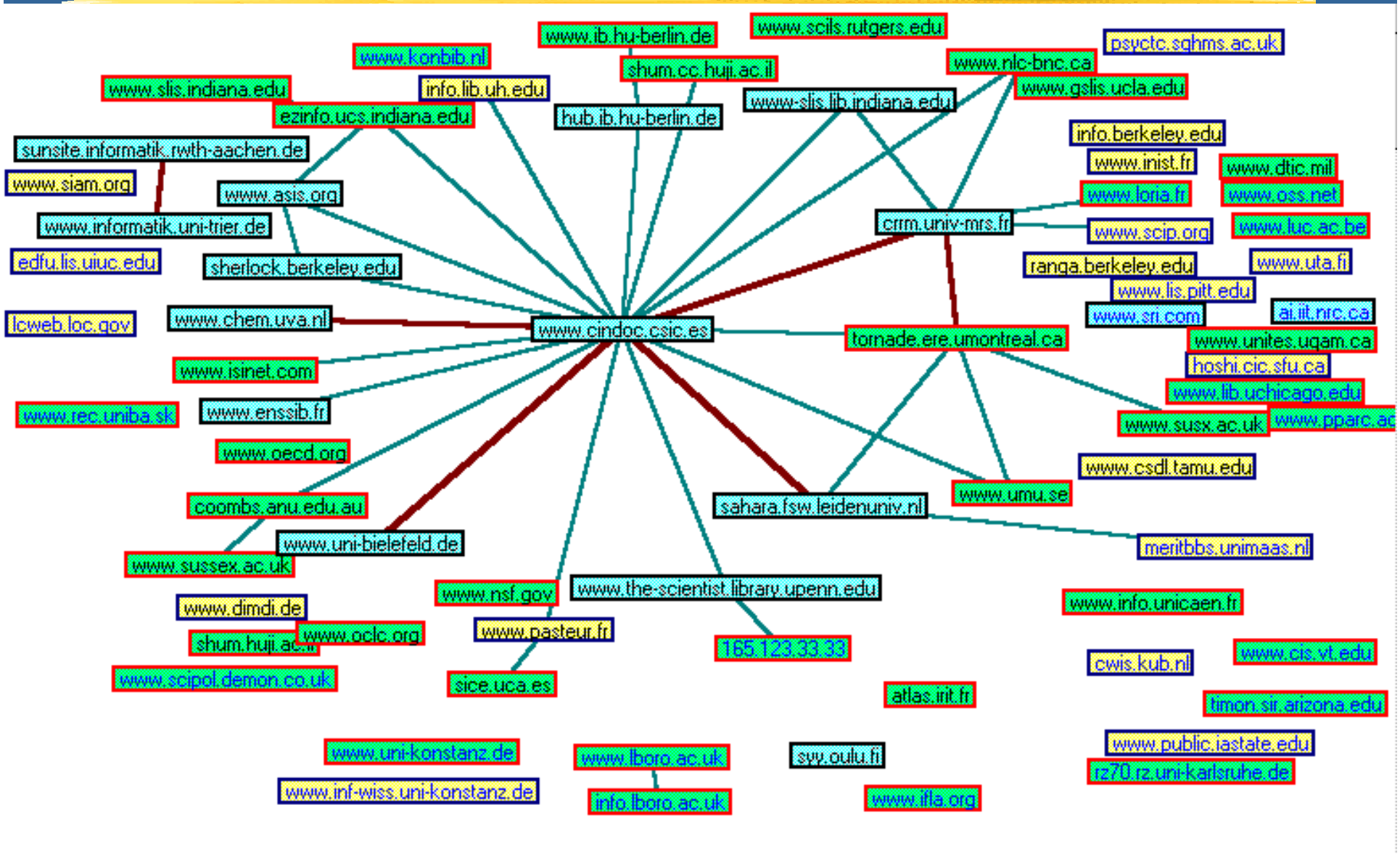
Filters setup



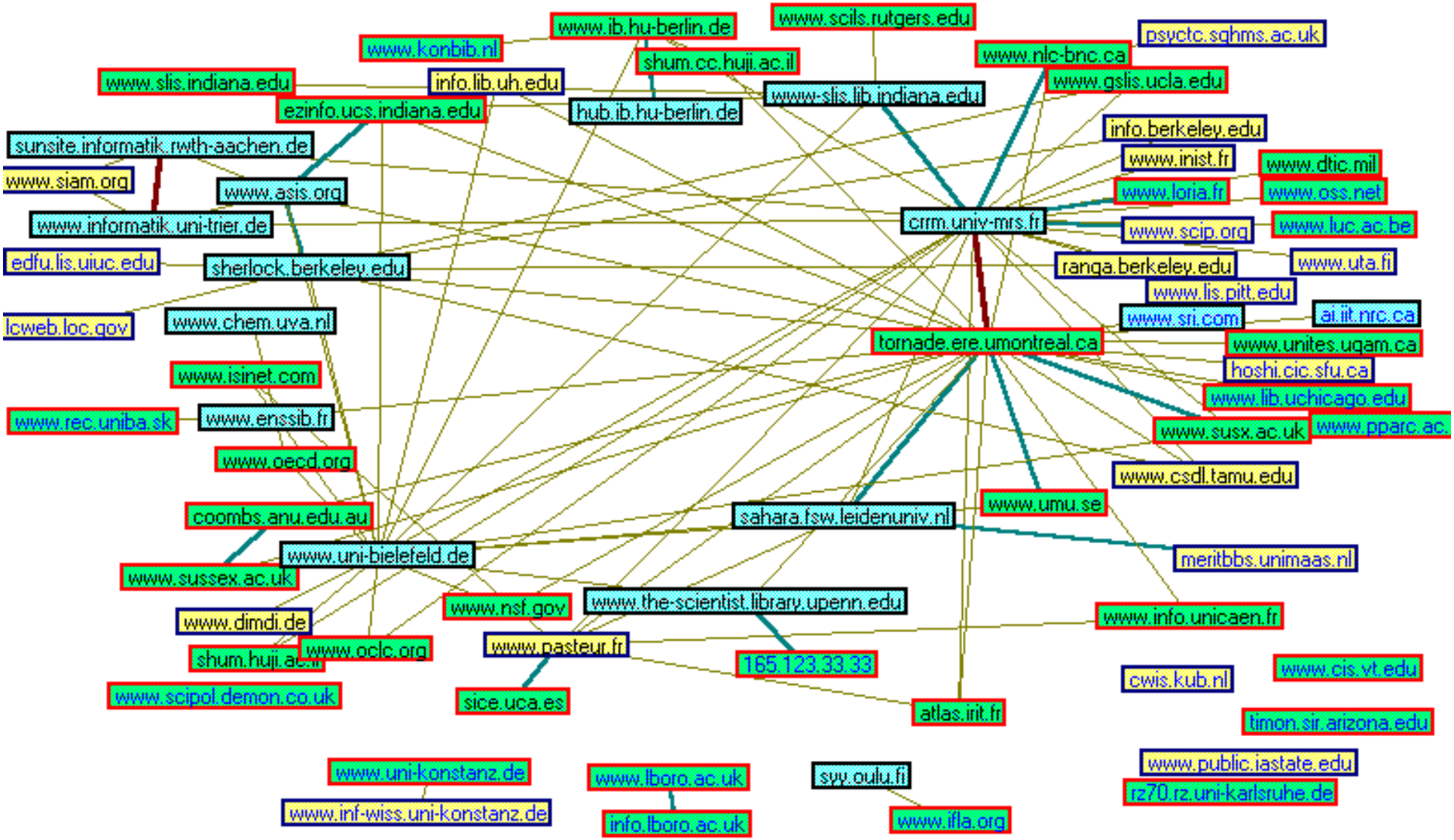
Map of hosts having links with 2 hosts at least



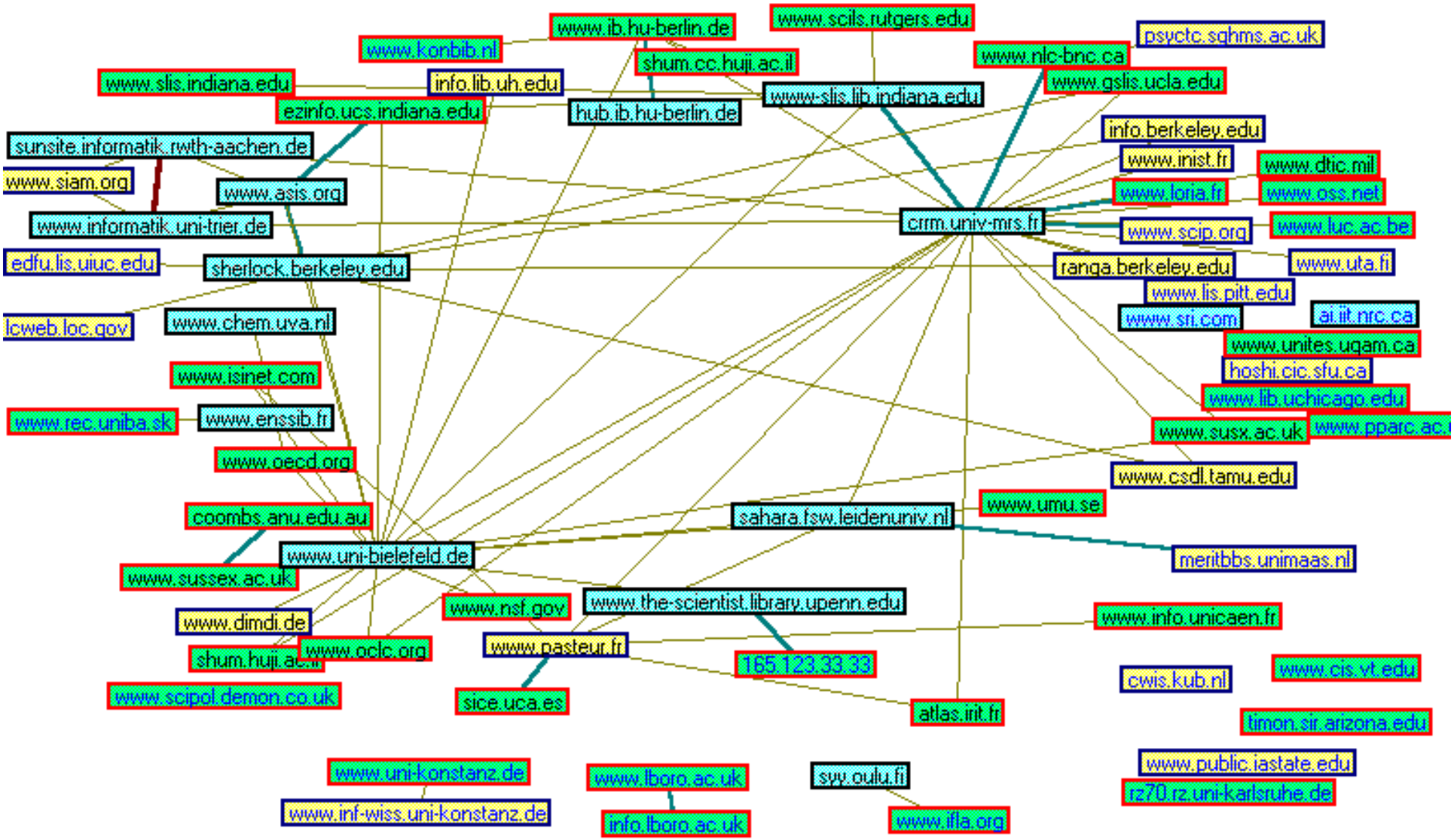
Map showing hosts having links with at least 1 host and showing links between two hosts greater than 1



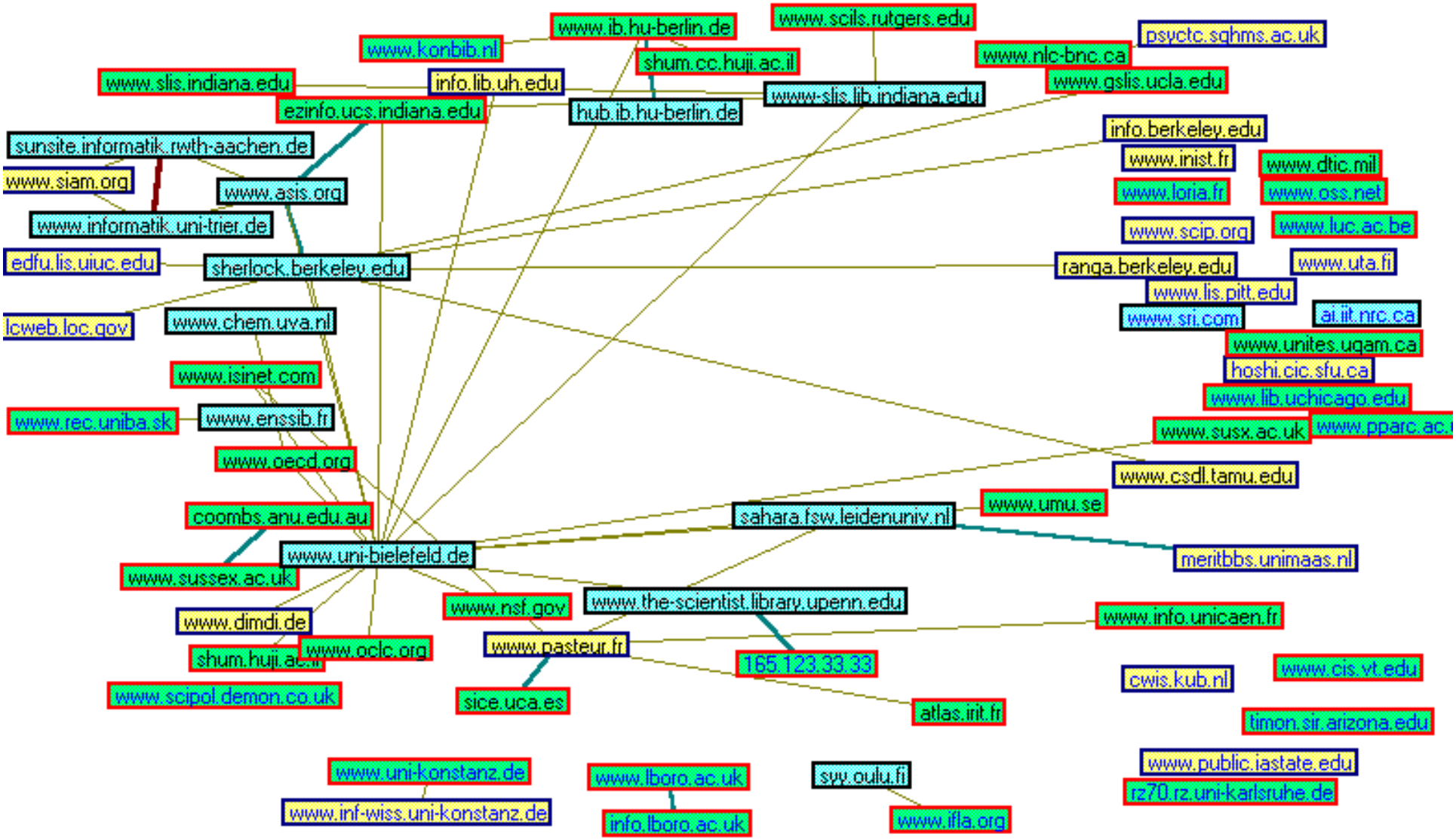
Partial map : 1 host deleted



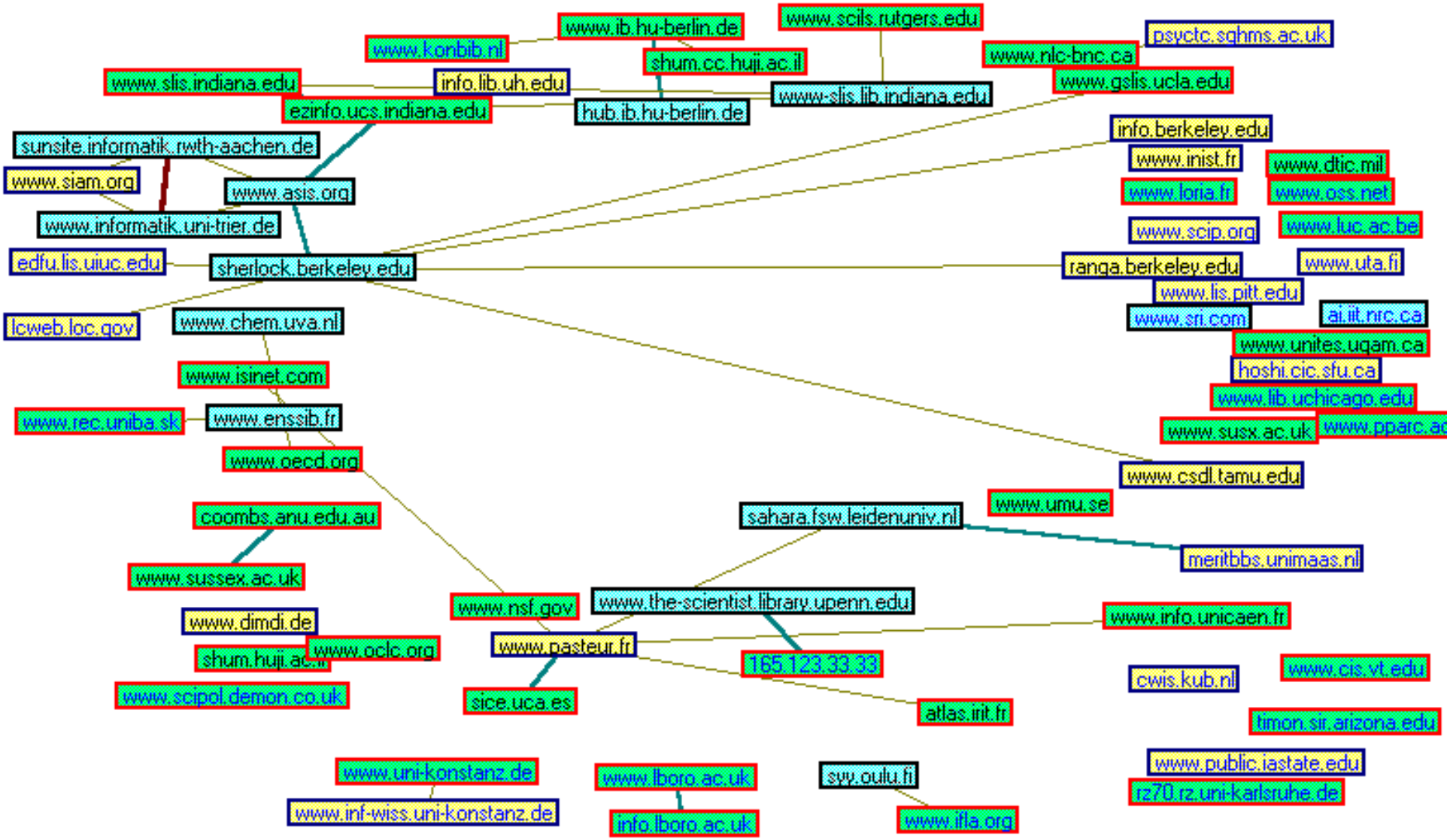
Partial map : 2 hosts deleted



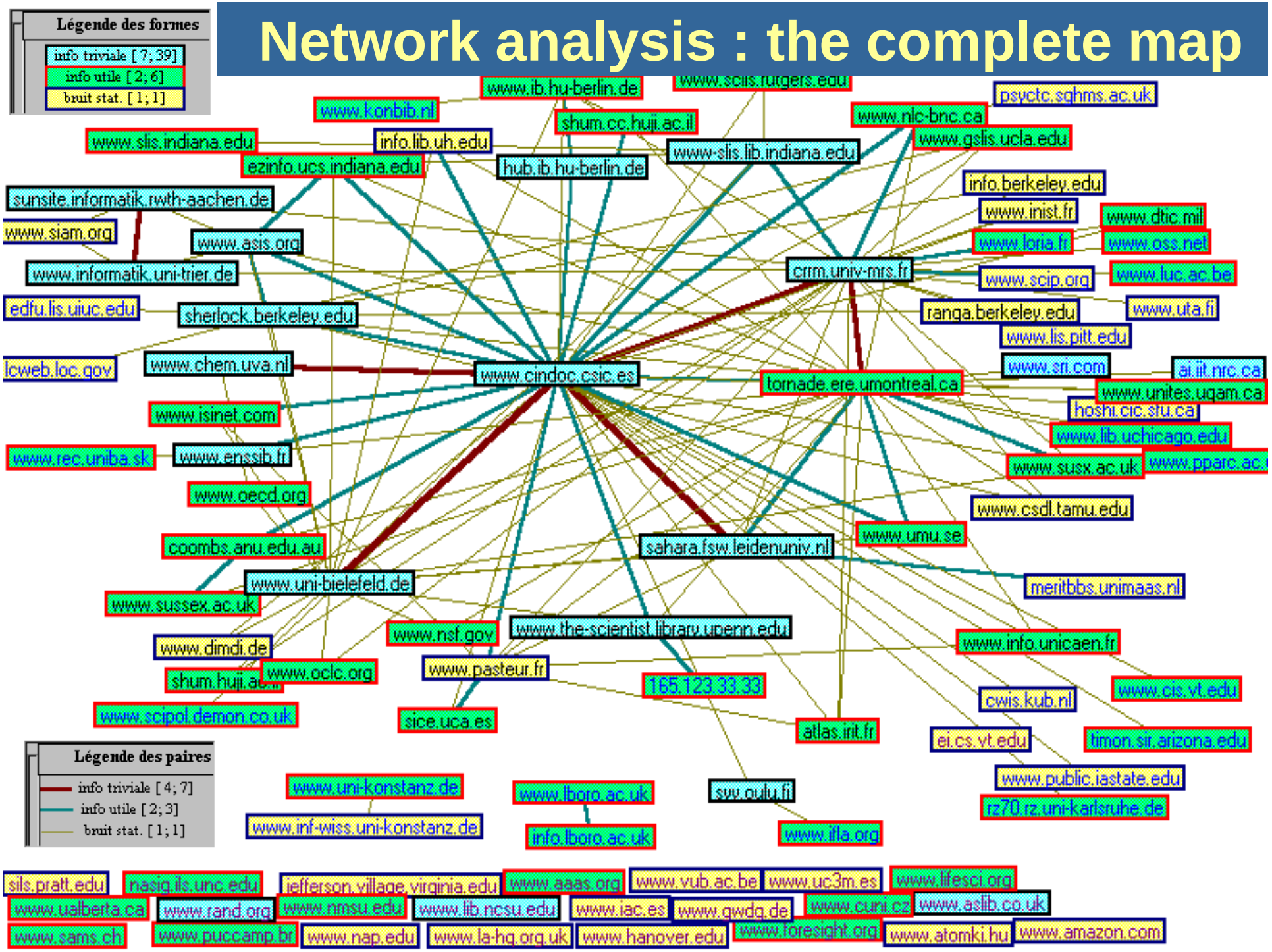
Partial map : 3 hosts deleted



Partial map : 4 hosts deleted



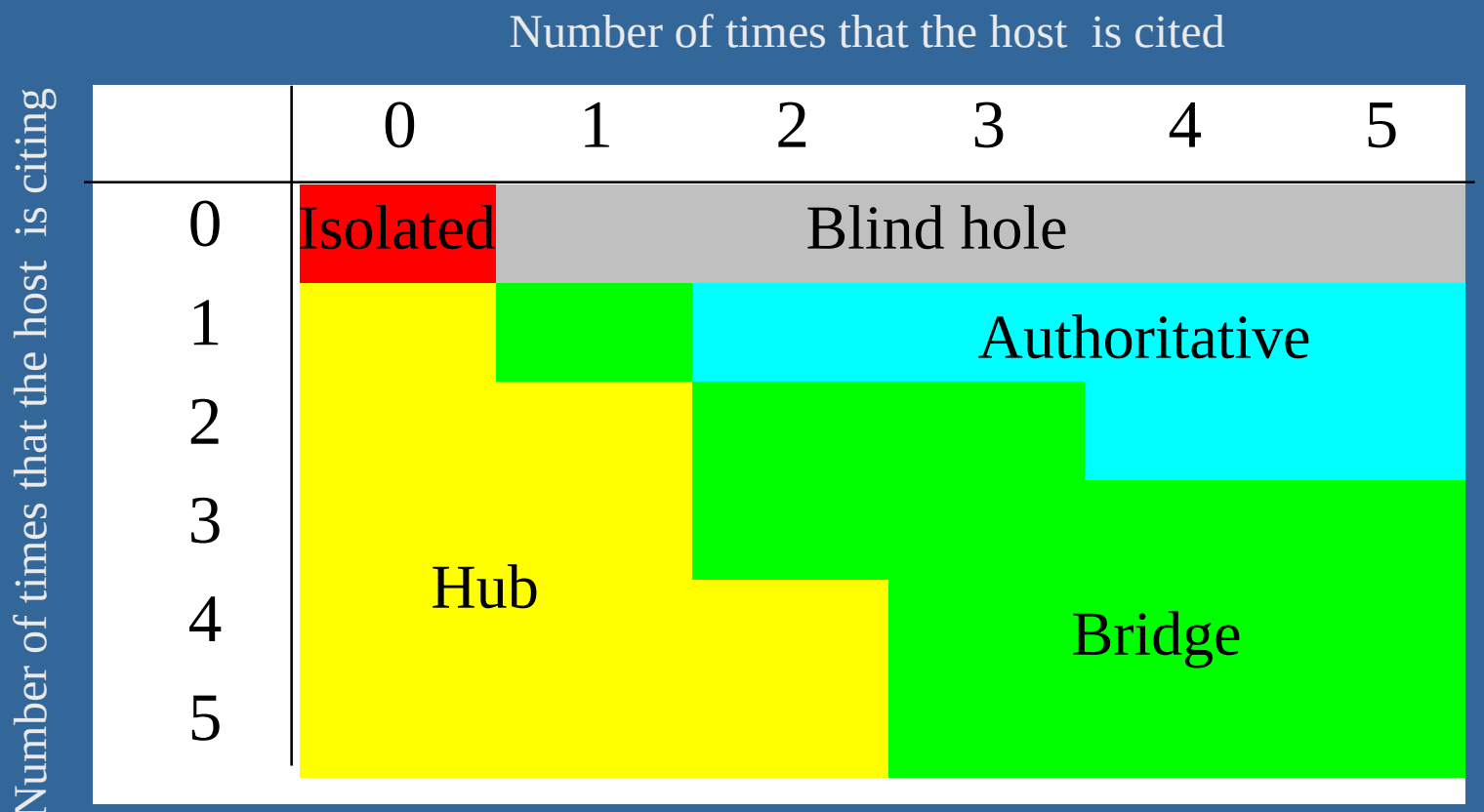
Network analysis : the complete map



How to be inform about the directions of the links

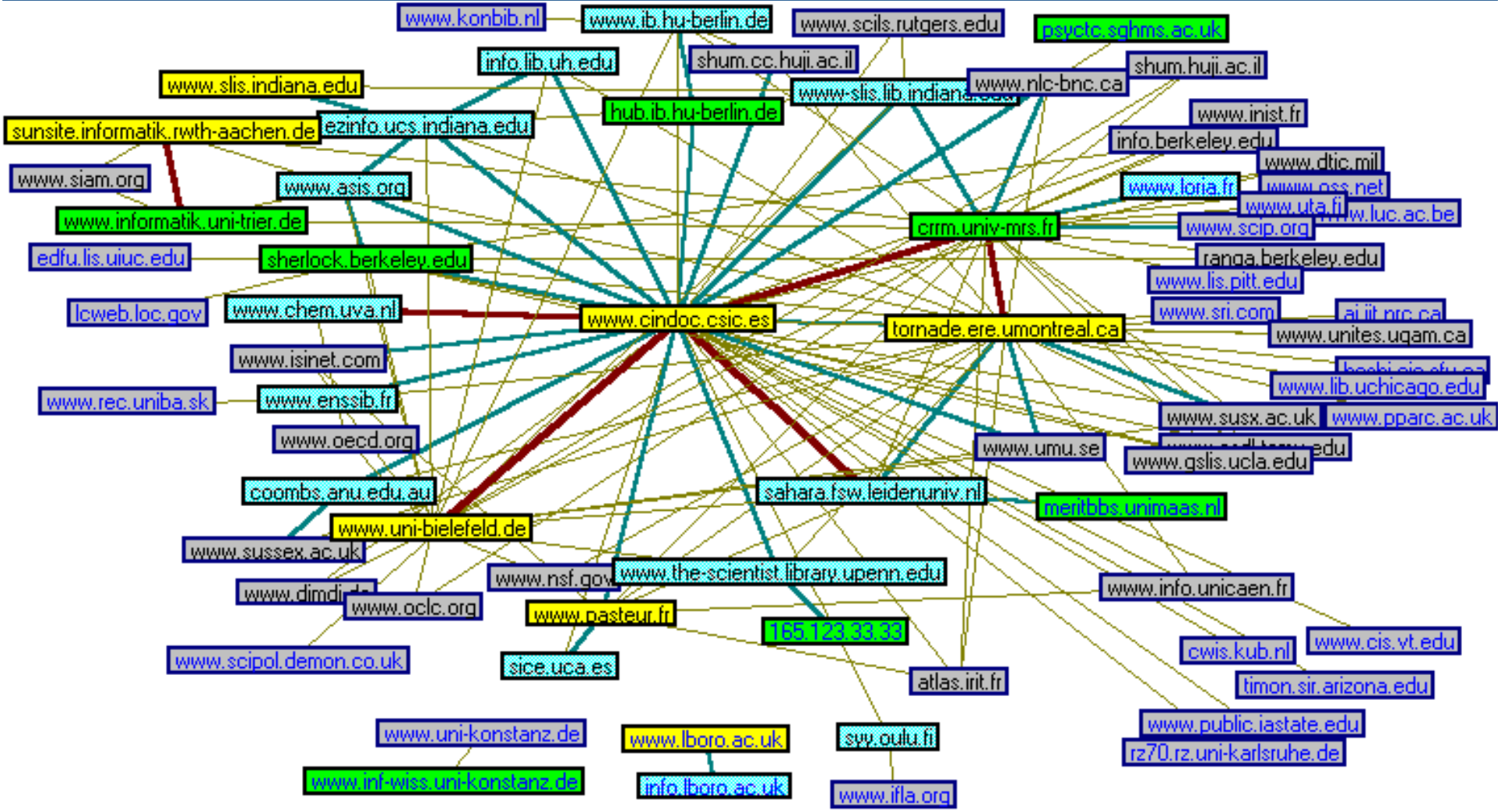
- Categorization of the hosts according to the "citation activity"
  - "blind hole host" = only receiving citations without any return
  - "authoritative host" = more often cited than citing
  - "hub host" = more often citing than cited
  - "bridge host" = nearly as many time cited as citing

Categorization table of hosts



The matrix diagonal is replaced by this categorization indicator

Complete map including categorization



Conclusion

- This first experiment must be improved
- The coupling of network analysis and clustering analysis must be investigated

- The treatment process nowadays used (three involved softwares : Auresys, Dataview, Matrisme) must be integrated into Auresys allowing interactivity possibilities (to switch from map to Auresys data organization)

Evaluation of Internet resources: Bibliometric techniques applications.

Hervé ROSTAING, rostaing@crrm.univ-mrs.fr

Eric BOUTIN, boutin@univ-tln.fr

Bruno MANNINA, mannina@crrm.univ-mrs.fr

http://crrm.univ-mrs.fr