Laboratoire
d'Histoire des Sciences
et de Philosophie
Archives Henri Poincaré

RTFM!
Scientific mailing lists as a means of communication: A novel kind of corpus

Camille Paloque Bergès
Alexandre Hocquet

European Society for the History of Science
Lisboa, Setembro 2014

Outline

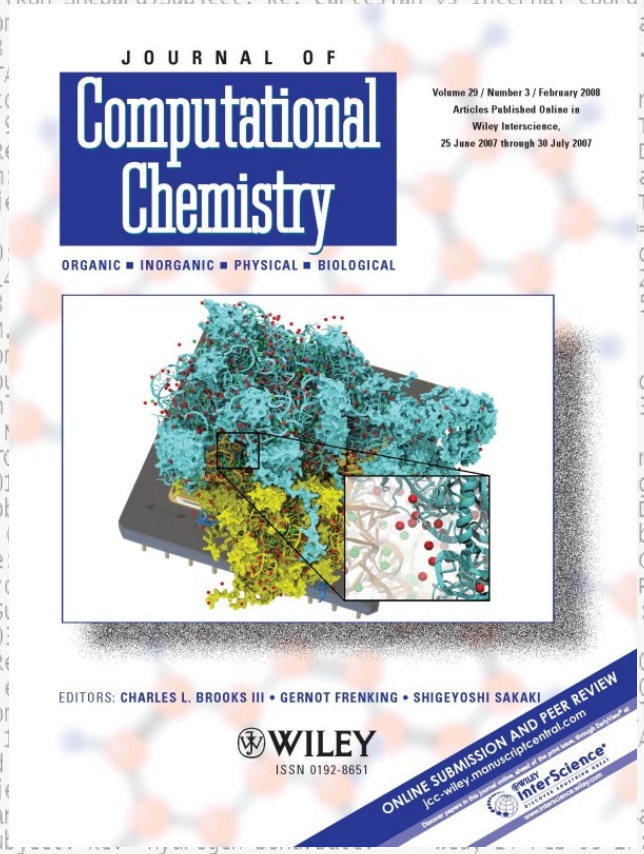
1. What is my story about? computational chemistry and software
2. What is my corpus? Computational Chemistry mailing List
3. Methodology: The flame wars
4. Methodology: The threaded conversation
5. Small data mining

Re: CCL:RTFM? What FM?

- *From*: "Peter Shenkin" <shenkin : at : still3.chem.columbia.edu>
- *Subject*: Re: CCL:RTFM? What FM?
- *Date*: Thu, 22 May 1997 15:29:27 -0400

I've gotten lots of private email responses on this subject, in
addition to the posted responses I've seen. I've been archiving
all of these.
After the dust settles, I'll post a summary to CCL.
-P.
--
**** "Deep Blue can't triumph in the game of life" (NY Times, 5/13/97)
****
* Peter S. Shenkin; Chemistry, Columbia U.; 3000 Broadway, Mail Code 3153 *
** NY, NY 10027; shenkin : at : columbia.edu; (212)854-5143; FAX: 678-9039
***
*MacroModel WWW page: http://www.columbia.edu/cu/chemistry/mmod/mmod.html *

1. Computational chemistry and software



Computational chemistry : use of computers to solve chemical problems

1. Computational chemistry and software

Computational chemistry & Hardware

- Supercomputers : 70s
- Personal Computers : 90s

Computational chemistry & Software

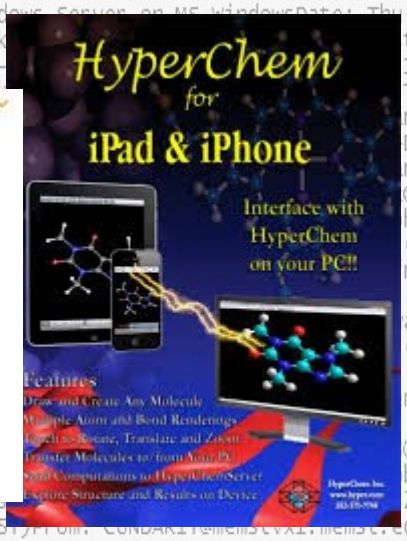
-Developer & User

GaussView

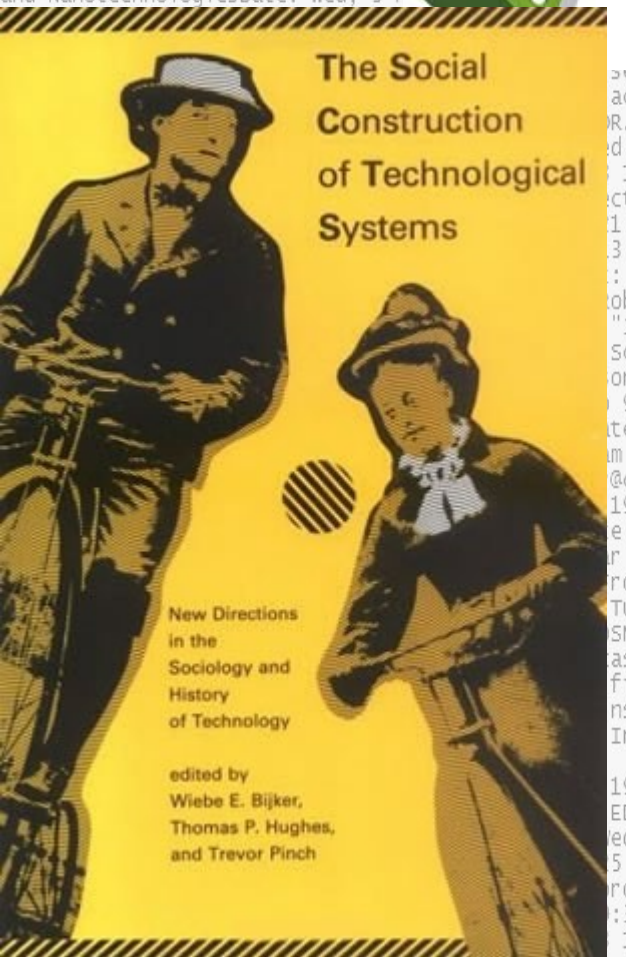
Gaussian, Inc.
Carnegie Office Park - Building 6
Pittsburgh PA 15106 USA

Spartan Student Edition

NWCHEM
HIGH-PERFORMANCE COMPUTATIONAL CHEMISTRY SOFTWARE



1. Computational chemistry and software



Mutual shaping

- Scientific modeling activity
- Technological, Economical and Political aspects of Software

2. The Computational Chemistry mailing List as a corpus

CCL: Computational Chemistry List

CCL.NET
computational chemistry list, ltd.

Wombat Bioactivity Database:

SUNSET MOLECULAR DISCOVERY LLC

http://www.ccl.net/index.shtml

CCL Home

About CCL

Rules, Instructions, Supporting CCL, Uploading file to CCL Site, Security Policy, Privacy Policy, About Us, Contact Us

Drug Discovery Training

Drug Discovery

Messages in a queue: 0 (approx. wait time: 00:00 hours)

Page supported by:

CSE ONLINE

Computational Science & Engin
... a laboratory without walls for scie

Welcome to the CCL Website!

ACS COMP Division supports CCL:

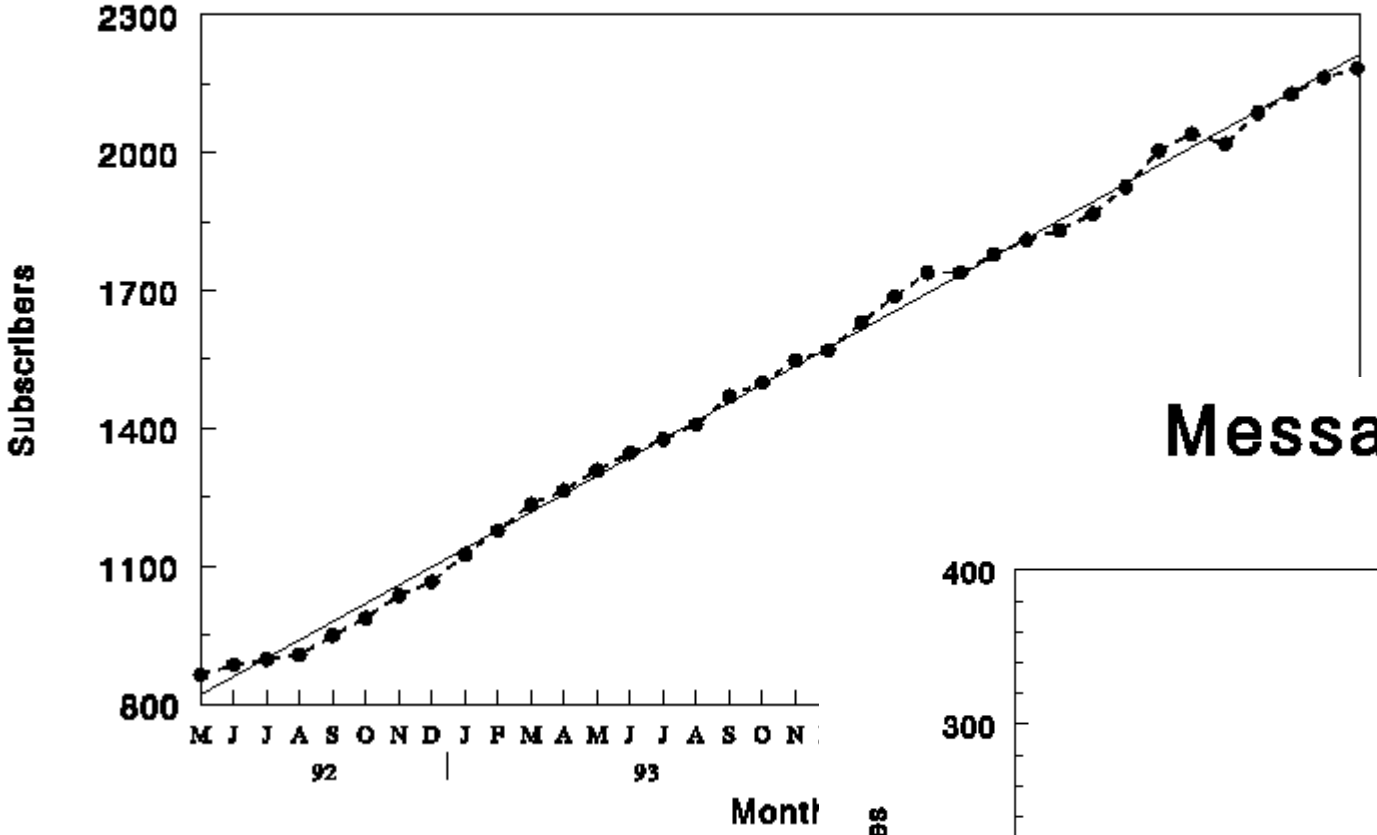
Attack of the killer workstations

- *From*: jle $#at#$ world.std.com (Joe M Leonard)
- *Subject*: Attack of the killer workstations
- *Date*: Tue, 10 Dec 91 20:04:44 -0500

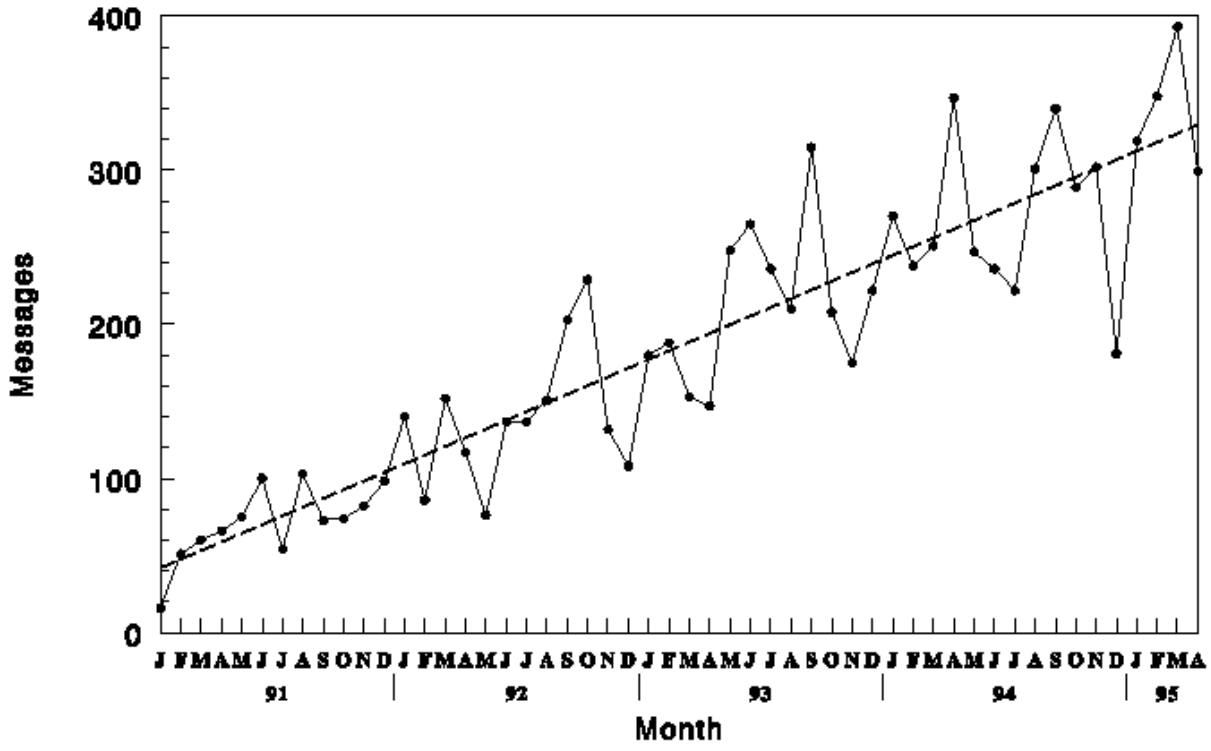
Reviewing the recent discussion regarding C vs. Fortran and structured vs. spagetti coding, isn't it reasonable to consider the time of the programmer as more important that the code's runtime (within reason)? Given that approx. $20k workstations are comparable to Cray machines, isn't 10-ish% overhead more than worth the human time savings through better-thought-out programming (structured, modular, OO, etc.)? I hope the day's past for really needing to count cycles in loops - as one can wait for the next version of workstation xxx to gain appreciable speedups.
Obviously, this isn't the case for mpp machines, for which the consensus is that one must rethink and/or reimplement algorithm's in order to use the hardware to advantage. Still, this might prove to be an example where clearer coding styles might turn out to be a win after all...
Joe Leonard
jle $#at#$ world.std.com

2. The Computational Chemistry mailing List as a corpus

Number of subscribers



Messages per month



2. The Computational Chemistry mailing List as a corpus

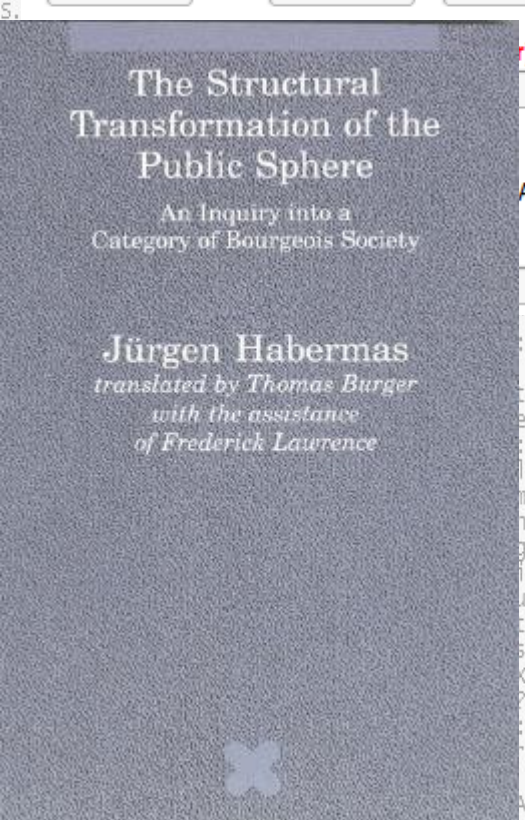
CCL.NET
computational chemistry list, ltd.

Send Message to CCL Subscribers or Administrator

For computer security reasons you have to retype this number: 567 in this box: . Your submission will not be accepted without
Please fill in all the fields below. Note: if you are sending a message to CCL Subscribers you will be asked for confirmation. After you click
URL in the confirmation e-mail or cut/paste the URL into your browser. This is the only way to verify that you entered your e-mail address
would be able to enter *your* e-mail address in the message, and impresonate you.
Click on the Submit button when finished.

Important message to participants with Yahoo email addresses. [Please click here]

First/Given Name **Middle Name** **Last/Sur/Family Name**



Administrator

full disclosure of methods?

- *From*: "DR. DOUGLAS A. SMITH, UNIVERSITY OF TOLEDO" <DSMITH -AatT- uoft02.ut
- *Subject*: full disclosure of methods?
- *Date*: Sat, 26 Jun 1993 11:21:56 -0500 (EST)

```
Mark Thompson recently wrote, regarding the SAM1 topic currently under
discussion:
>Let me address the more fundamental issue that this topic
>brings forth.  I share Graham Hurst's concerns. One of the
>basic tenets of good science is that of reproducibility,
>and independant verification.
This is, I think, universally true and accepted.  However, it is rarely
followed.  For example, there has been talk over the years that people
who use molecular mechanics for their research should publish the parameters
used for each study as part of the paper (or at least they should publish
the differences between their parameters and the "standard" parameters
```

2. The Computational Chemistry mailing List as a corpus

Rules

By posting to the list, you agree to adhere to the rules outlined in this document.
These rules, regarding what topics are acceptable/unacceptable and commercial announ
discussions. Do not reopen Pandora's box -- please.

Acceptable Topics

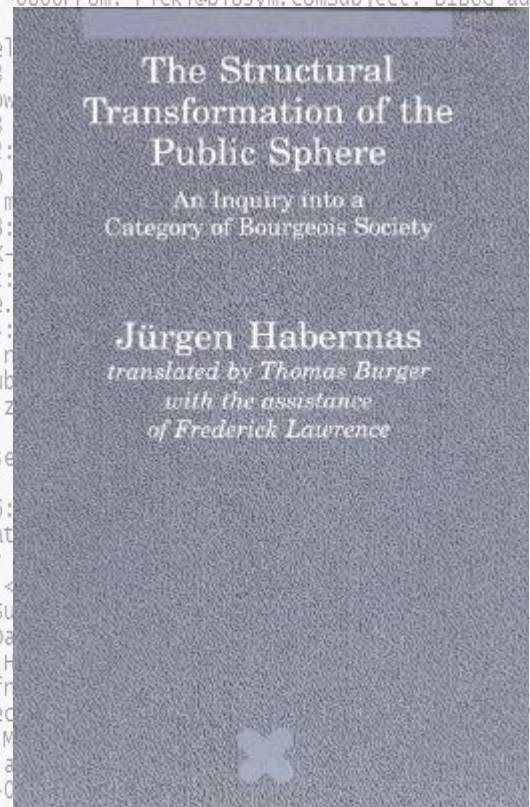
Topics which **may appear on the list** include:

- **software/hardware-related** issues
  - bugs: Reporting bugs in chemistry software and possible work-arounds
  - programs and utilities
  - availability: Information on the availability of particular software, data, etc.
  - hardware-related issues relevant to the computational chemistry community
- **announcements** of
  - new chemistry software
  - computational chemistry-related workshops and symposia
- **methods**
  - new methods and techniques in computational chemistry
  - Q&A: questions and answers about solving computational chemistry problems
- **opinions** about services and products

Commercial Announcements

Commercial software: software sold to anyone fo
to the ***list***, please respect the following rules:

- **Be brief** -- Short (25 lines or less) commercial p
- **Be straightforward** -- The subject line should be
- **No reposting** -- Reposts of software announcem
- **No "vaporware"** -- Announcements of "vaporwa
  are not allowed.



2. The Computational Chemistry mailing List as a corpus

IEEE Annals

A Social History of Bitnet and Listserv, 1985-1991

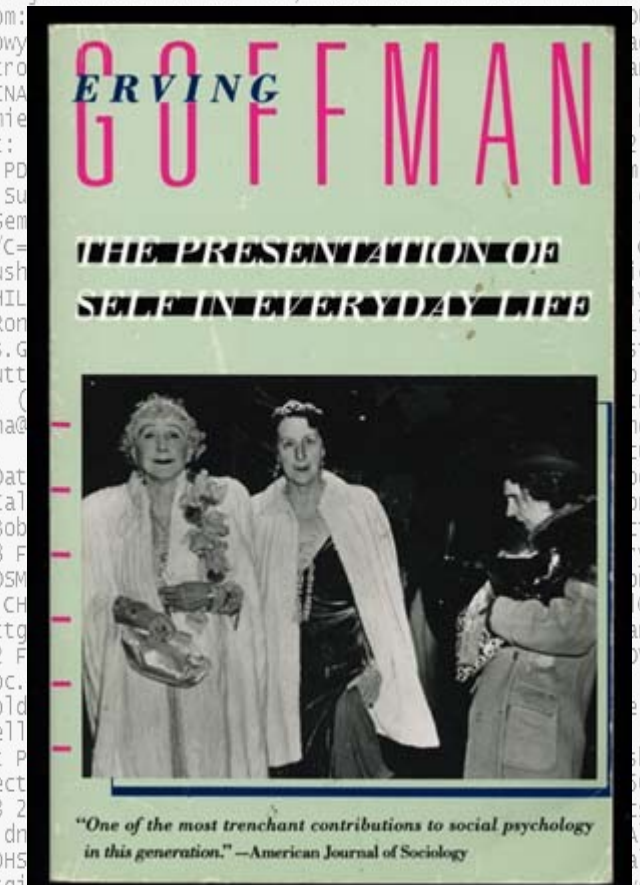
April-June 2000 (vol. 22 no. 2)
pp. 32-41

David Alan Grier
Mary Campbell

DOI Bookmark: http://doi.ieeecomputersociety.org/10.1109/85.841135

ABSTRACT

After the engineers built the computer networks, users had to build the social networks that made them useful. Listserv, the combined mailing list and file server, was an important tool for those interested in building network-based organizations. It first appeared on Bitnet, an academic network based on IBM computers. The early versions of Listserv became operational in the mid 1980s, and its early archives show how network users learned to use the software and, perhaps more importantly, how to manage network-based organizations.



Special Issue Articles

The Presentation of Self in the Age of Social Media: Distinguishing Performances and Exhibitions Online

Bernie Hogan[1]

Bulletin of Science, Technology & Society
30(6) 377–386
© 2010 SAGE Publications
Reprints and permission: http://www.sagepub.com/journalsPermissions.nav
DOI: 10.1177/0270467610385893
http://bsts.sagepub.com
SAGE

3. The flame wars

-Religious wars:

Flaming, Trolling

Sociologie des épreuves

full disclosure of methods?

- *From*: "DR. DOUGLAS A. SMITH, UNIVERSITY OF TOLEDO" <DSMITH -AatT- uoft02.ut
- *Subject*: full disclosure of methods?
- *Date*: Sat, 26 Jun 1993 11:21:56 -0500 (EST)

Re: full disclosure of methods?

Mark Thom
discussic
>Let me e
>brings f
>basic te

Re: full disclosure of methods?

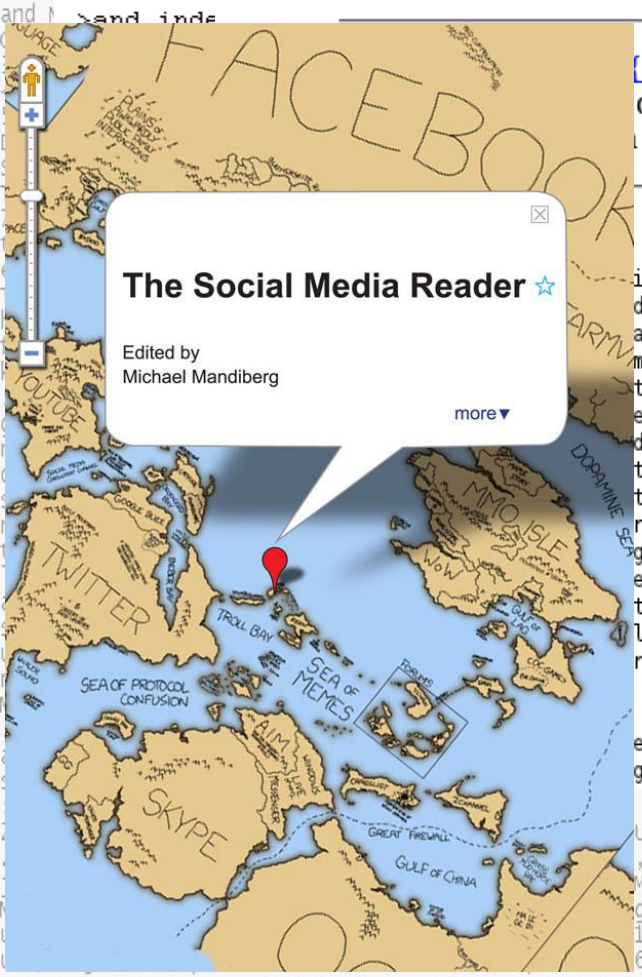
- *From*: st-amant { *at * } cgl.ucsf.edu (Alain St-Amant)
- *Su*
- *Da*

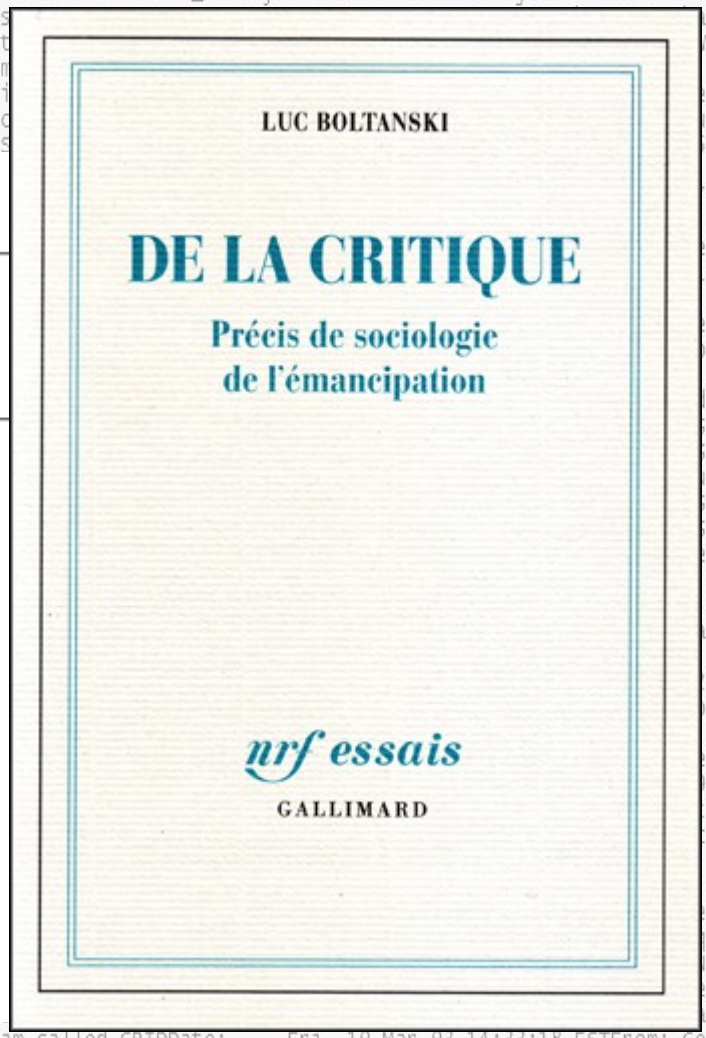
{ *at * } cgl.ucsf.edu (Alain St-Amant)
disclosure of methods?
93 12:21:56 -0700

wrote in the current discussion on the disclosure

id we had to reveal all our secrets and make them
d accessible? When software copyrights and patents
ate protection, maybe I will agree with that attitude.
mith is referring to the specific algorithms that are
the program more efficient but do not affect the final
e, I might agree.
disagree more if he is referring to some development
t actually affects the final results in any way.
ts me however, is the question of software copyrights
r. Smith alludes. I have been trying to get a feel
ghted and patented and I get a different answer from
ecific code be copyrighted or can the structure and
ted as well? How 'modified' should code be before
ly (and ethically if anyone is interested in expressing
ram? Or is it simply forbidden for code to

e net any e-mail sent to me, but I think that this would
g discussion and it would be as interesting to hear how





Hacker and Troller as Trickster (Gabriella Coleman)

4. The threaded conversations

full disclosure of methods?

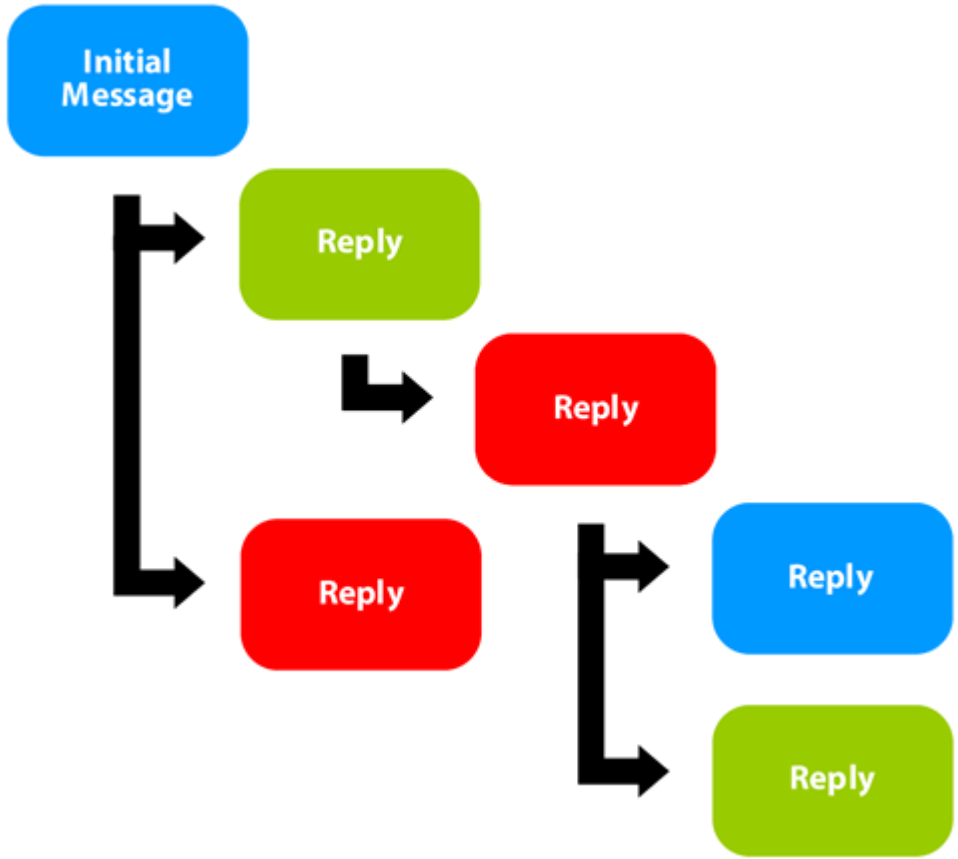
Re: full disclosure of methods?

Re: full disclosure of methods?

Re: full disclosure of methods?

- *From*: st-amant { *at * } cgl.ucsf.edu (Alain St-Amant)
- *Subject*: Re: full disclosure of methods?
- *Date*: Sat, 26 Jun 93 12:21:56 -0700

Douglas Smith recently wrote in the current discussion on the discl
of parameters:
> Besides, who ever said we had to reveal all our secrets and make
> readily available and accessible? When software copyrights and p
> really provide adequate protection, maybe I will agree with that
I'll assume that Dr. Smith is referring to the specific algorithms
implemented that make the program more efficient but do not affect
results. In which case, I might agree.
Of course, I couldn't disagree more if he is referring to some deve
in the methodology that actually affects the final results in any w
The point that interests me however, is the question of software co
and patents to which Dr. Smith alludes. I have been trying to get
for what can be copyrighted and patented and I get a different answ
everyone. Can only specific code be copyrighted or can the structu
algorithms be copyrighted as well? How 'modified' should code be b
it can be called legally (and ethically if anyone is interested in
an opinion) a new program? Or is it simply forbidden for code to
"evolve"
into a new program?
I will summarize to the net any e-mail sent to me, but I think that
make for an interesting discussion and it would be as interesting t

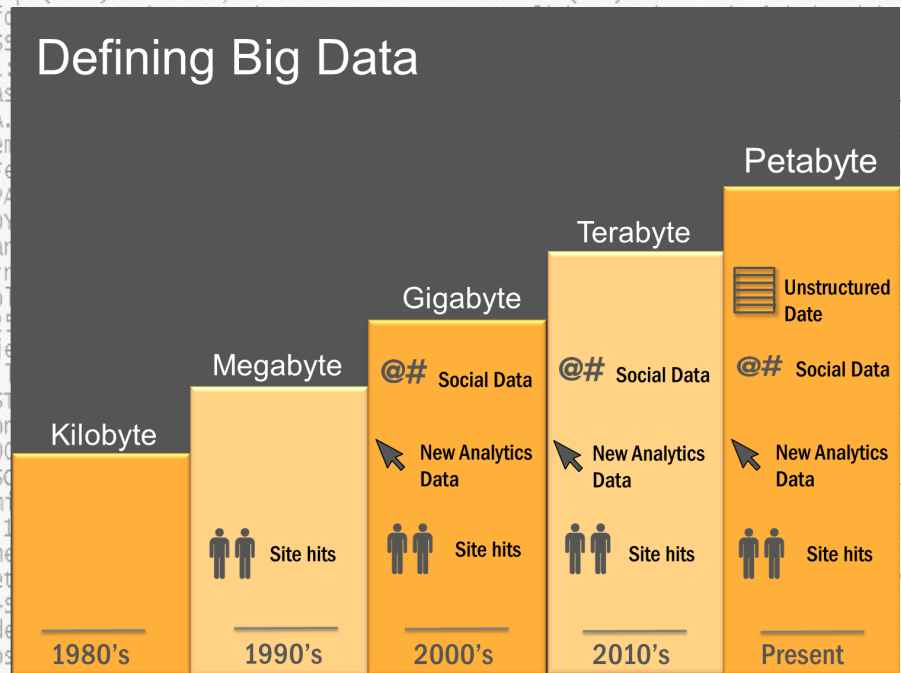


5. Small data mining

-a diachronic analysis of the debates : **Time series**

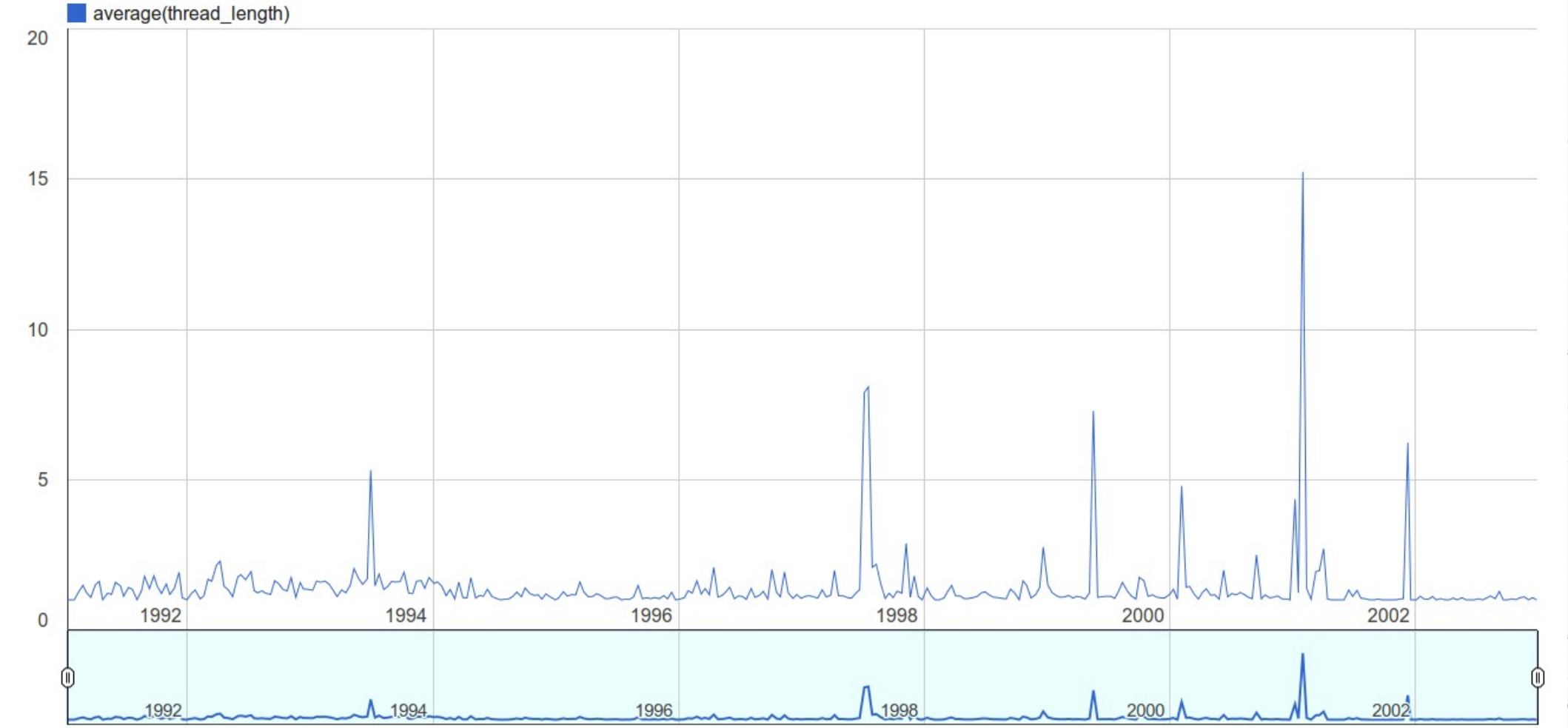
-a **social network analysis** : Bipartite graphs

-a literary analysis : **Text mining**



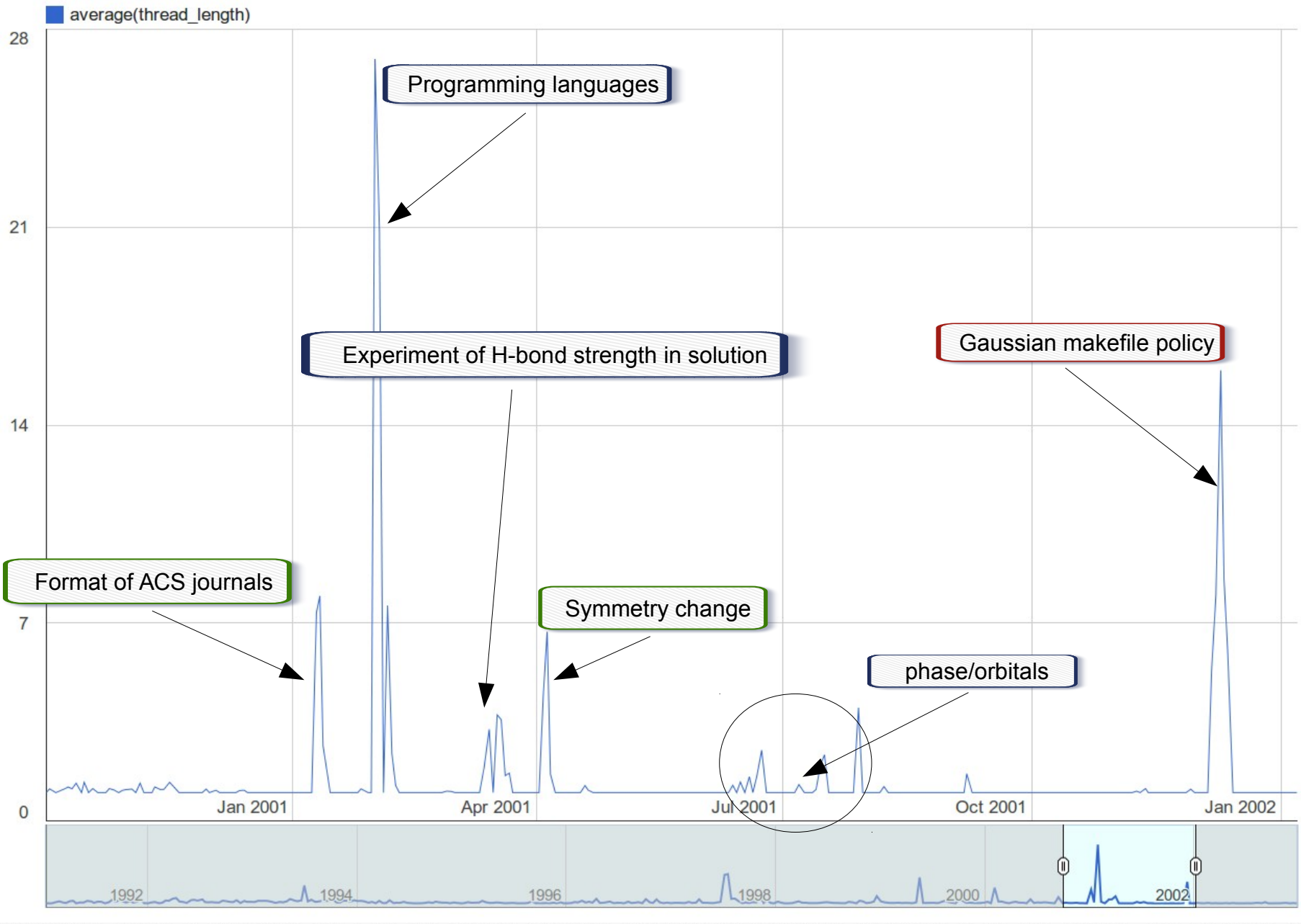
5. Small data mining: Time series

Averaged thread length vs. time : each peak represents an average of thread length for each message in a given period of time.
(if a message does not belong to a thread it has a thread length of 1)

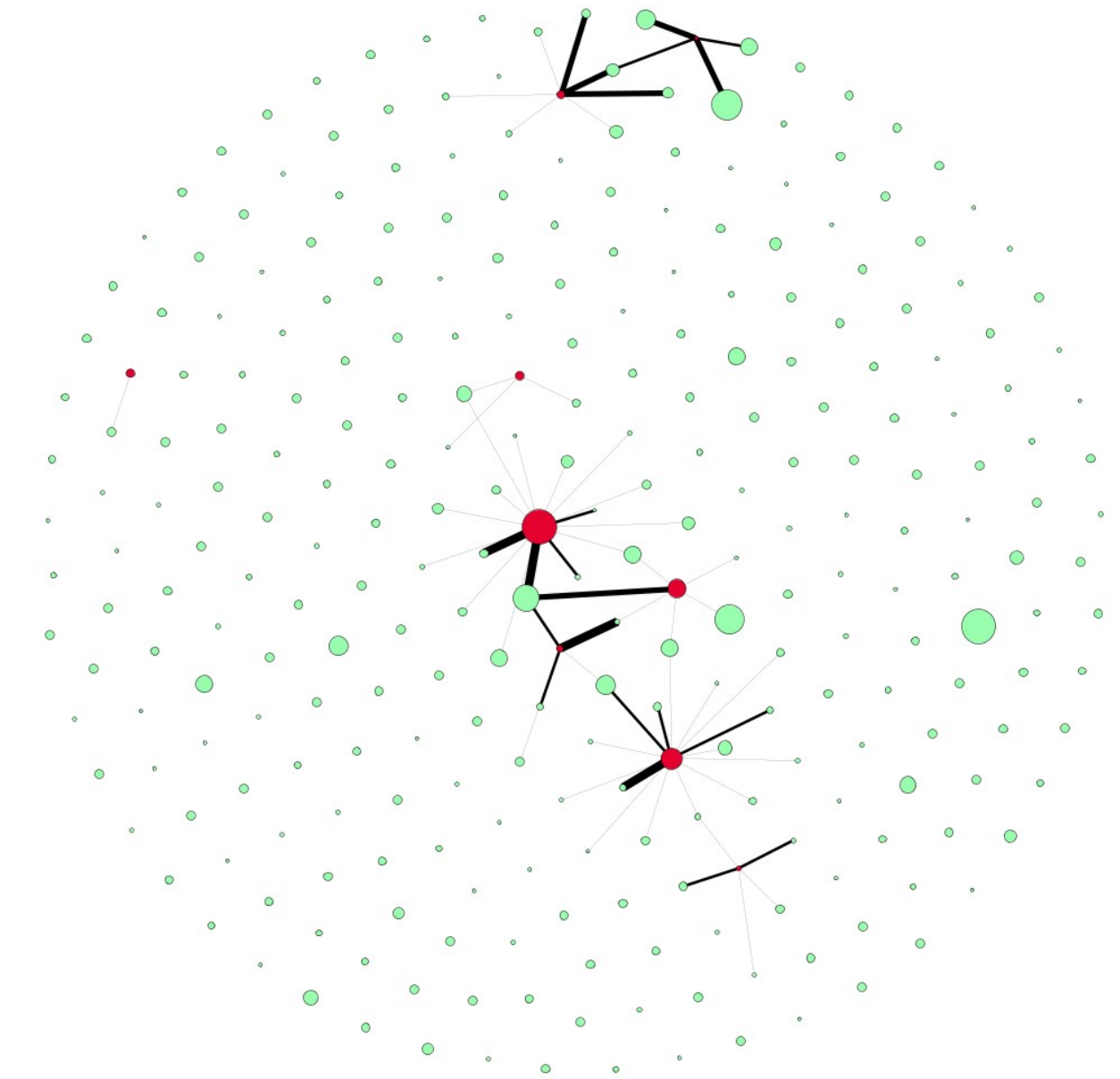


5. Small data mining: Time series

Year 2001 zoom : Six interesting threads extracted



5. Small data mining: Social Network Analysis



A yearly bipartite network :

Green dots are email addresses (size= # of messages)

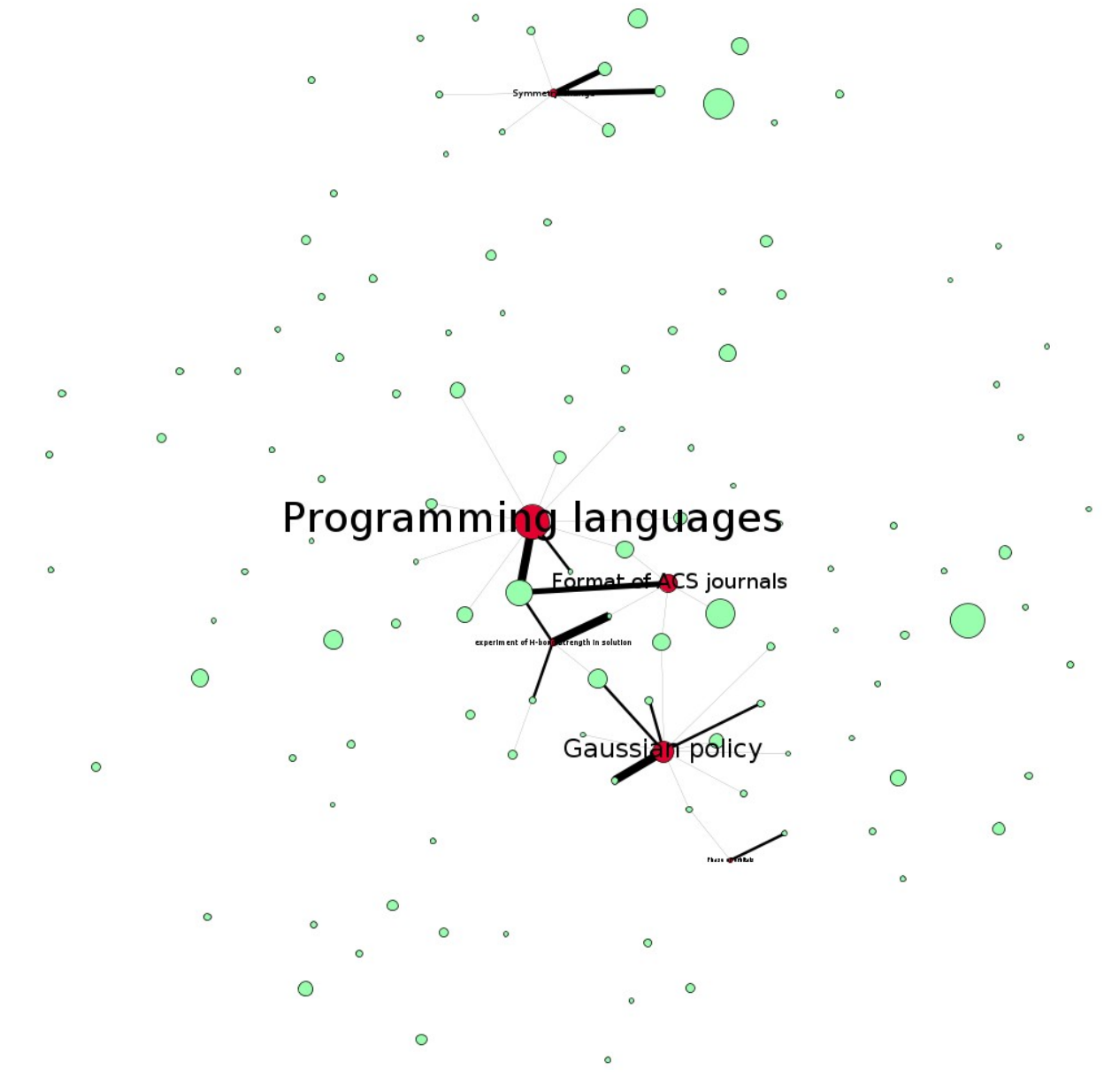
Red dots are threads/topics (size= # of contributors)

Edges link red and green dots when a contributor post in a topic (weighted by # of messages)

The majority of threads are made of one or two posts, a myriad of isolated components

The interesting part is how the big threads relate to one another : qualitatively, and by whom.

5. Small data mining: Social Network Analysis



Filtered 2001 network :

Green dots are email addresses (size= # of messages)

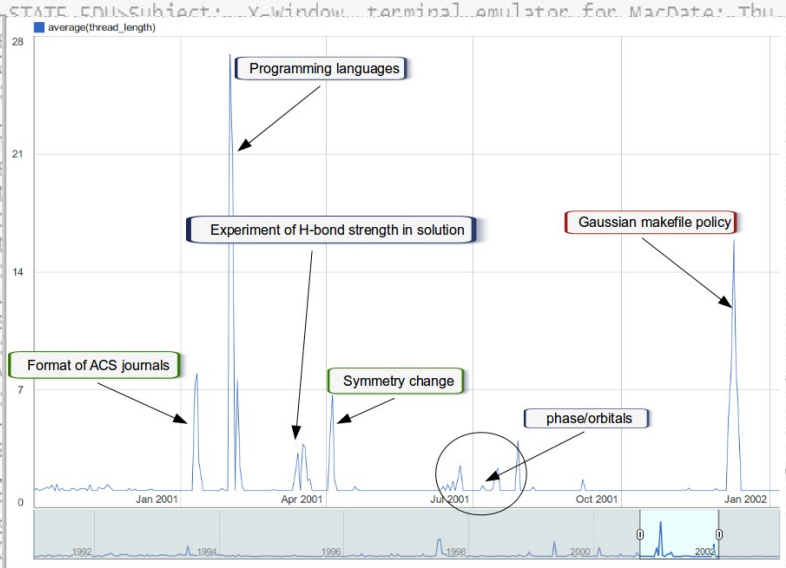
Red dots are threads/topics (size= # of contributors)

Edges link red and green dots when a contributor post in a topic (weighted by # of messages)

« symmetry change » is isolated : It does not attract the same contributors

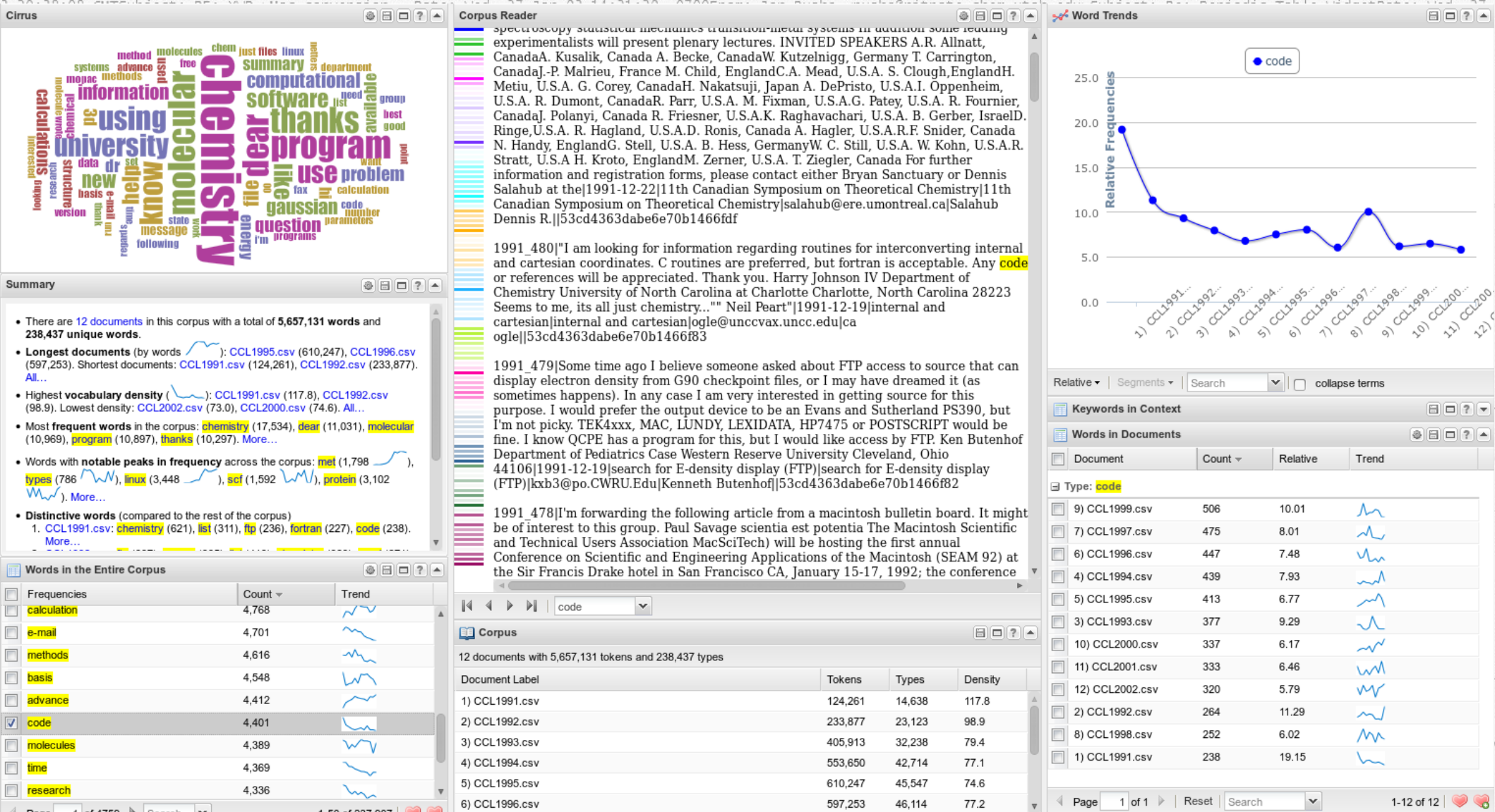
Argument : large edges with few people

Debate : numerous people, small edges

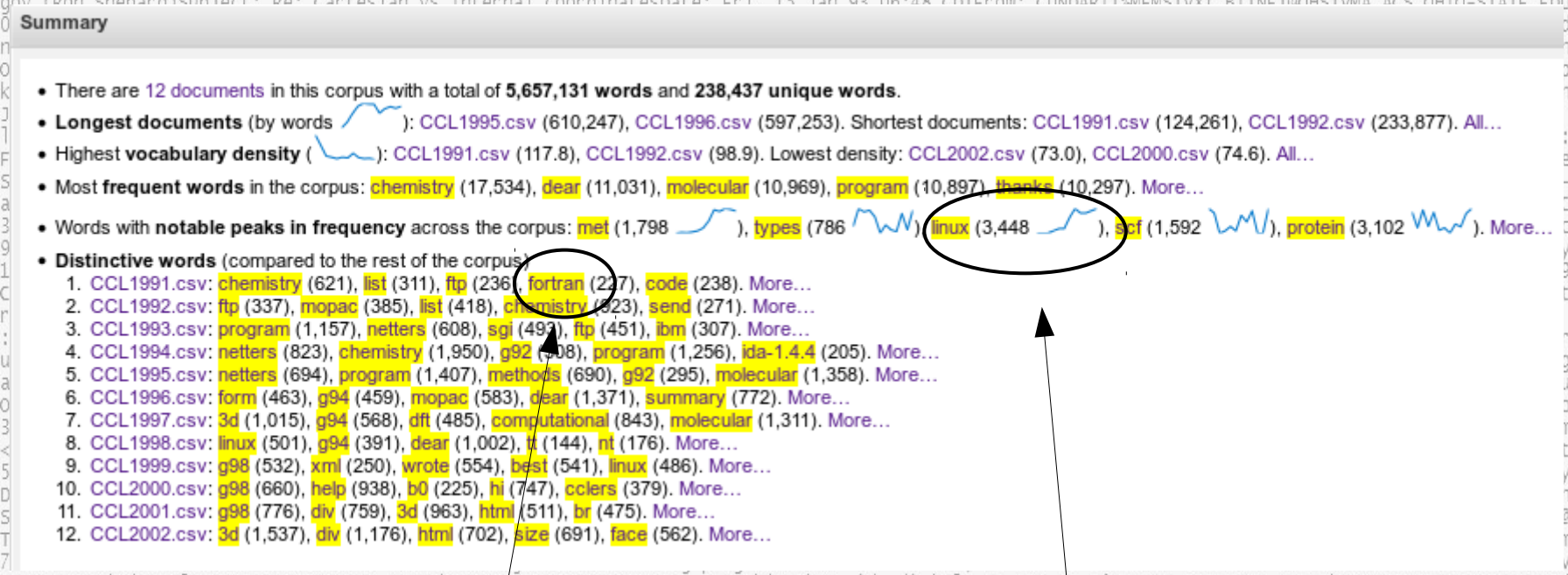


5. Small data mining: Text mining

Voyant tools as a corpus text miner.
The corpus from 1991 to 2002 is scissored into twelve yearly extracts, thus allowing for diachronic mining.



5. Small data mining: Text mining



Summary

- There are 12 documents in this corpus with a total of **5,657,131 words** and **238,437 unique words**.
- **Longest documents** (by words): CCL1995.csv (610,247), CCL1996.csv (597,253). Shortest documents: CCL1991.csv (124,261), CCL1992.csv (233,877). All...
- Highest **vocabulary density** ( ): CCL1991.csv (117.8), CCL1992.csv (98.9). Lowest density: CCL2002.csv (73.0), CCL2000.csv (74.6). All...
- Most **frequent words** in the corpus: chemistry (17,534), dear (11,031), molecular (10,969), program (10,897), thanks (10,297). More...
- Words with **notable peaks in frequency** across the corpus: met (1,798), types (786), linux (3,448), scf (1,592), protein (3,102). More...
- **Distinctive words** (compared to the rest of the corpus)
  1. CCL1991.csv: chemistry (621), list (311), ftp (236), fortran (227), code (238). More...
  2. CCL1992.csv: ftp (337), mopac (385), list (418), chemistry (323), send (271). More...
  3. CCL1993.csv: program (1,157), netters (608), sgi (493), ftp (451), ibm (307). More...
  4. CCL1994.csv: netters (823), chemistry (1,950), g92 (508), program (1,256), ida-1.4.4 (205). More...
  5. CCL1995.csv: netters (694), program (1,407), methods (690), g92 (295), molecular (1,358). More...
  6. CCL1996.csv: form (463), g94 (459), mopac (583), dear (1,371), summary (772). More...
  7. CCL1997.csv: 3d (1,015), g94 (568), dft (485), computational (843), molecular (1,311). More...
  8. CCL1998.csv: linux (501), g94 (391), dear (1,002), tt (144), nt (176). More...
  9. CCL1999.csv: g98 (532), xml (250), wrote (554), best (541), linux (486). More...
  10. CCL2000.csv: g98 (660), help (938), b0 (225), hi (747), cclers (379). More...
  11. CCL2001.csv: g98 (776), div (759), 3d (963), html (511), br (475). More...
  12. CCL2002.csv: 3d (1,537), div (1,176), html (702), size (691), face (562). More...

Notable peaks in frequency

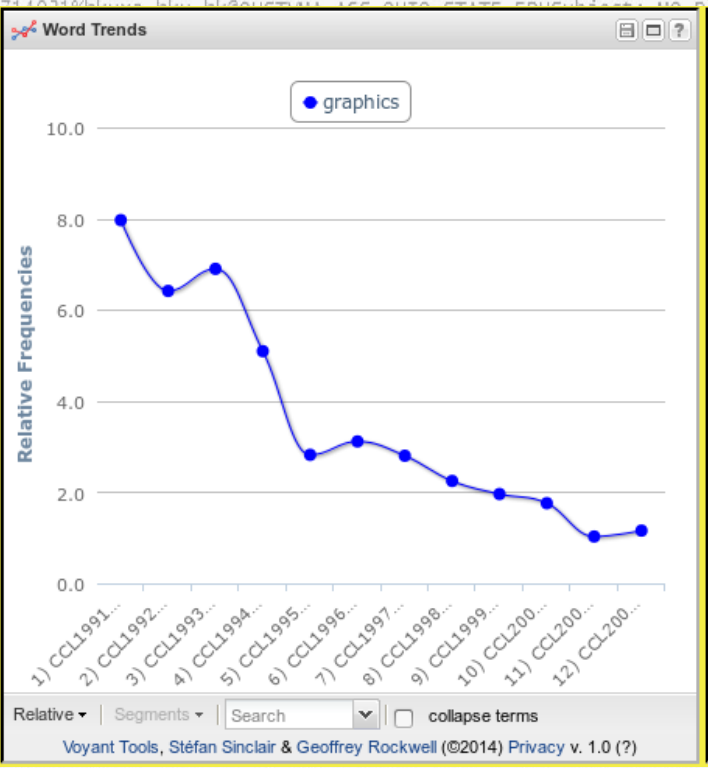
Distinctive words compared to the rest of the corpus

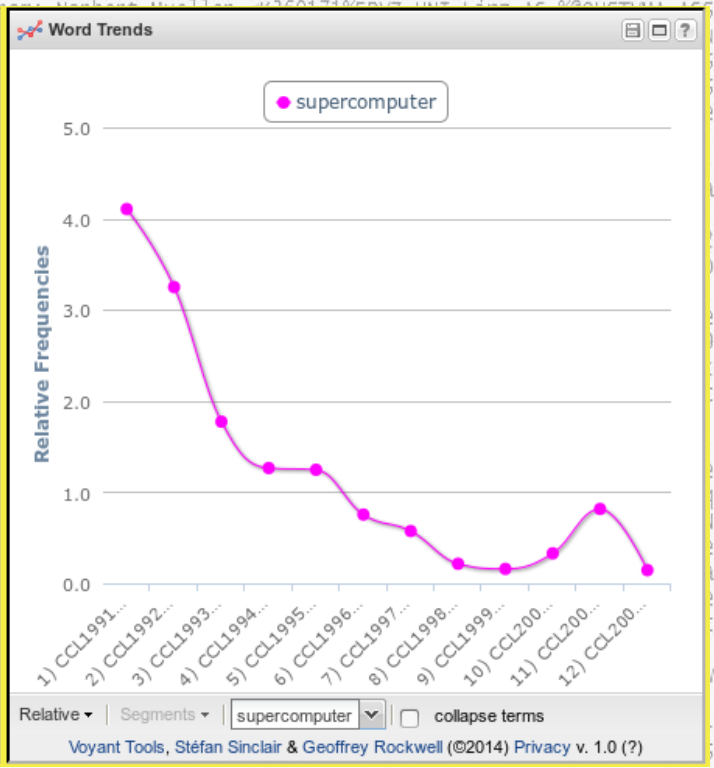
5. Small data mining: Text mining

Evolution of three keywords (graphics, linux, supercomputer) over time

« graphics » and « supercomputer » : typical words of a supercomputing era

« linux » : typical of a desktop era





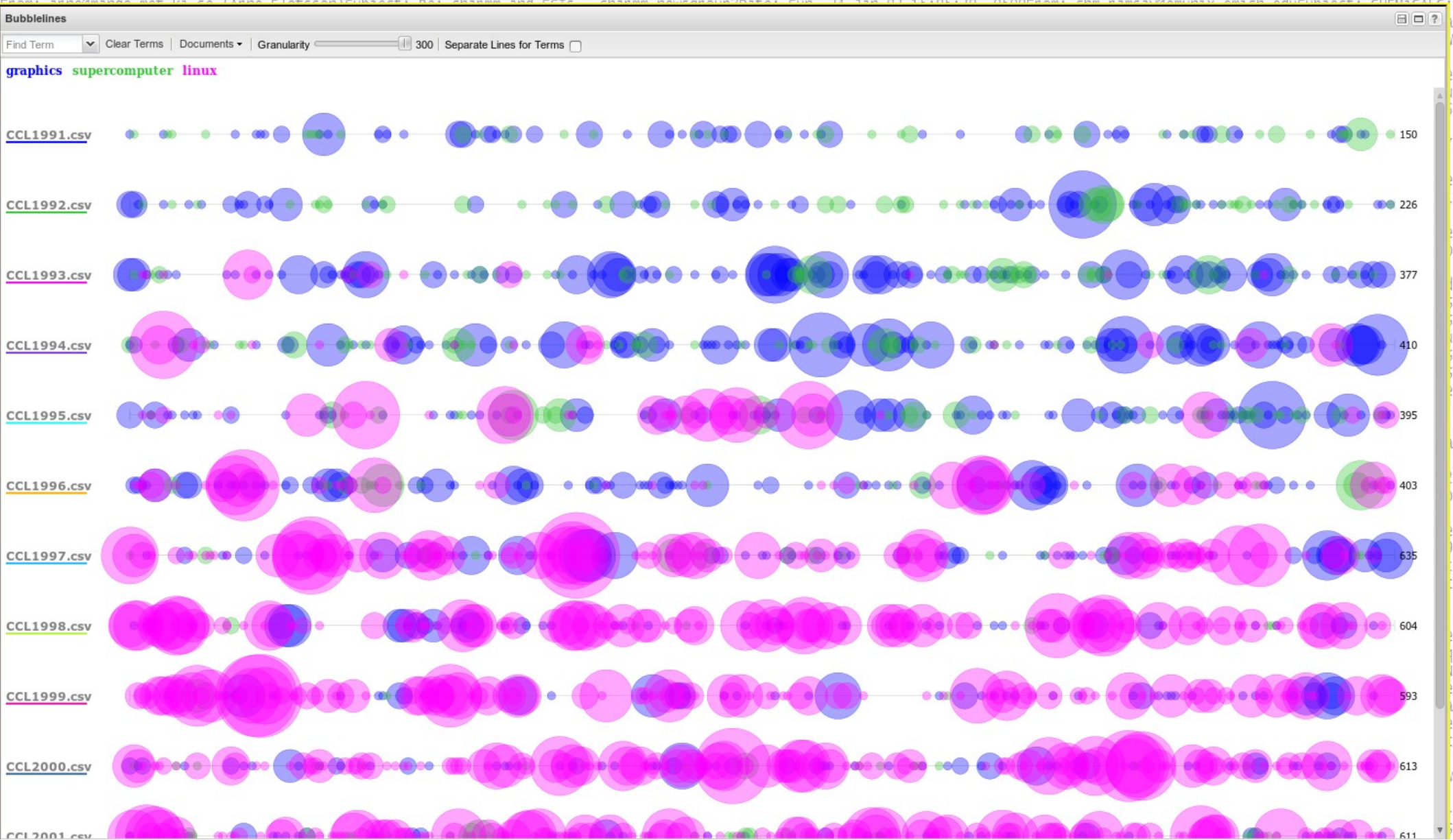
5. Small data mining: Text mining

Periodizing the corpus according to the nature of the debates.

« graphics » : blue

« supercomputer » : green

« linux » : pink

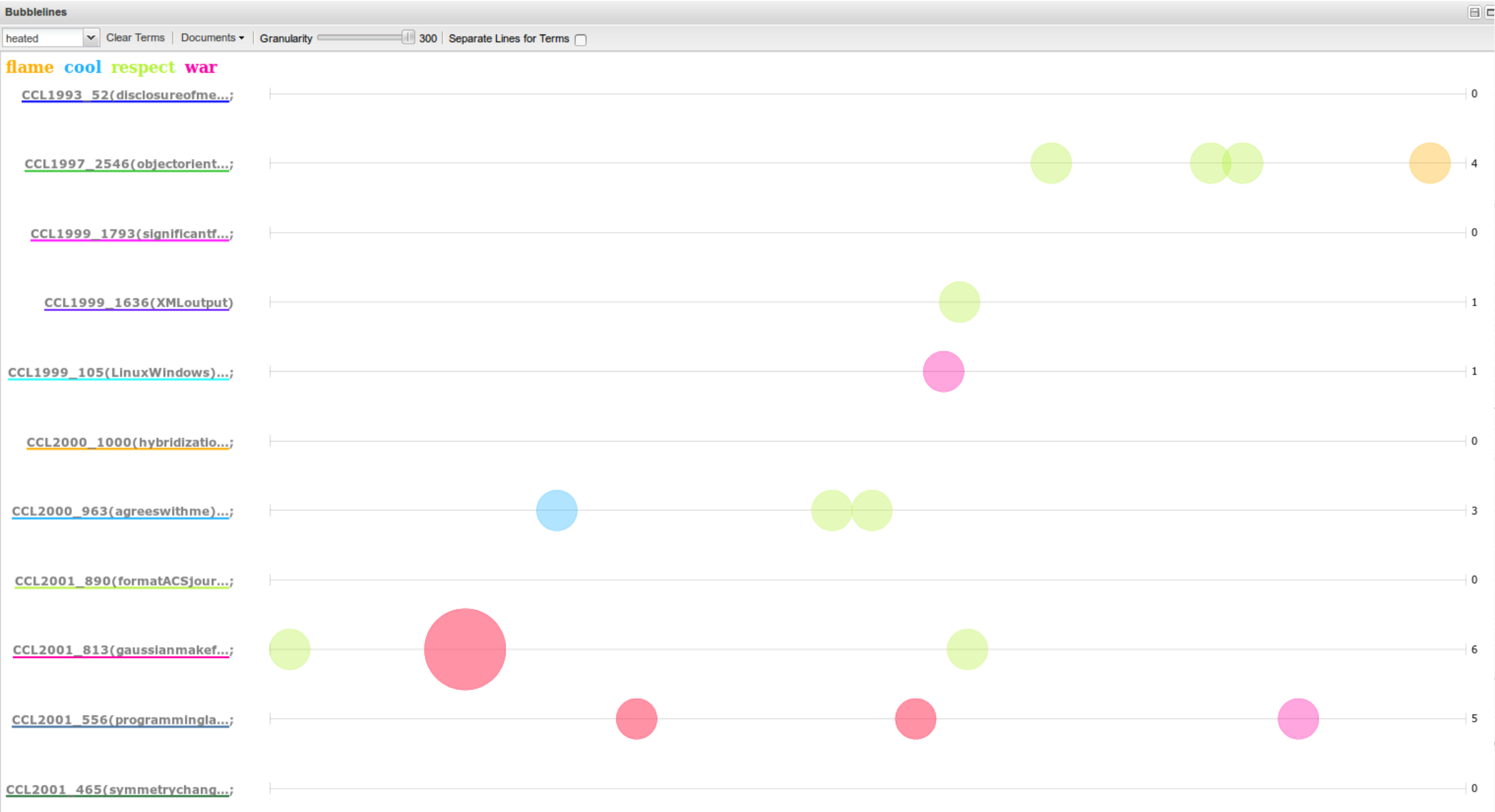


5. Small data mining: Text mining

Another kind of partitionning : according to the threads

5. Small data mining: Text mining

Locating the most passionate threads



5. Small data mining: Text mining

Four of the six threads highlighted in 2001 :

Allows to define proximity of topics according to keywords

Other possibility :

Identifying actors beyond their email addresses : Corpus partitionning according to emails for further text analysis

Bibliography

Boczkowski, P. J. 1999. “Mutual shaping of users and technologies in a national virtual community.” *Journal of Communication* 49(2):86–108.

Bury, Rhiannon. 2005. *Cyberspaces Of Their Own: Female Fandoms Online*. Peter Lang.

Flichy, Patrice. 1999. “Internet ou la communauté scientifique idéale.” *Réseaux* 17(97):77–120.

Grier, D.A., and M. Campbell. 2000. “A social history of Bitnet and Listserv, 1985-1991.” *IEEE Annals of the History of Computing* 22:32–41. Retrieved October 17, 2011.

Hansen, Derek L., Ben Shneiderman, and Marc Smith. 2010. “Visualizing threaded conversation networks: mining message boards and email lists for actionable insights.” Pp. 47–62 in *Proceedings of the 6th international conference on Active media technology*, *AMT’10*. Berlin, Heidelberg: Springer-Verlag

Kelty, Christopher M. 2008. *Two Bits: The Cultural Significance of Free Software*. Duke University Press.

Mandiberg, Michael. 2012 *The social media reader*. New York: New York University Press.

Marshall, Jonathan Paul. 2007. *Living on Cybermind: Categories, Communication, and Control*. Peter Lang.