Distributed Current Awareness Services

Thomas Krichel

2003-09-18

Thanks

- JISC, sponsor of Mailbase and JISCMail
- Mailman team
- WoPEc project
- Manchester Computing
- Bob Parks & Washington University of St. Louis
- CO PAH
  - Сергей И. Парнов
  - Татьяна И. Яковлева
- Heinrich Stammerjohans
- and the SINN03 organizers

What is current awareness?

- An old fashioned concept that implies a series of reports on
  - New items in a library
  - Per subject category
- Thus current awareness implies a two-dimensional classification on time and subject matter.

Is it useful in 7 A. Google?

- The time component is something that the search engines can not do easily
  - Can not divide items indexed according to types.
  - Do not understand subject matter.
  - Do not have a mode to find recent items.
- But generally can we trust computers to do it?

computers & thematic component

- In computer generated current awareness one can filter for keywords.
- This is classic information retrieval, and we all know what the problems are with that.
- In academic digital libraries, since the papers describe research results, they contain all “ideas” that have not been previously seen, therefore getting the keywords right is impossible.

Computers and time component

- In a digital library the “date” of a document can mean anything.
- The metadata may be dated in some implicit form.
  - Recently arrived records can be calculated
  - But record handles may be unstable
  - Recently arrived records do not automatically mean new documents.

We need human users!

- Cataloguers are expensive.
- We need volunteers to do the work.
- Junior researchers have good incentives
  - Need to be aware of latest literature
  - Absent in informal circulation channels of top level academics
  - Need to “get their name around” among researchers in the field.

History

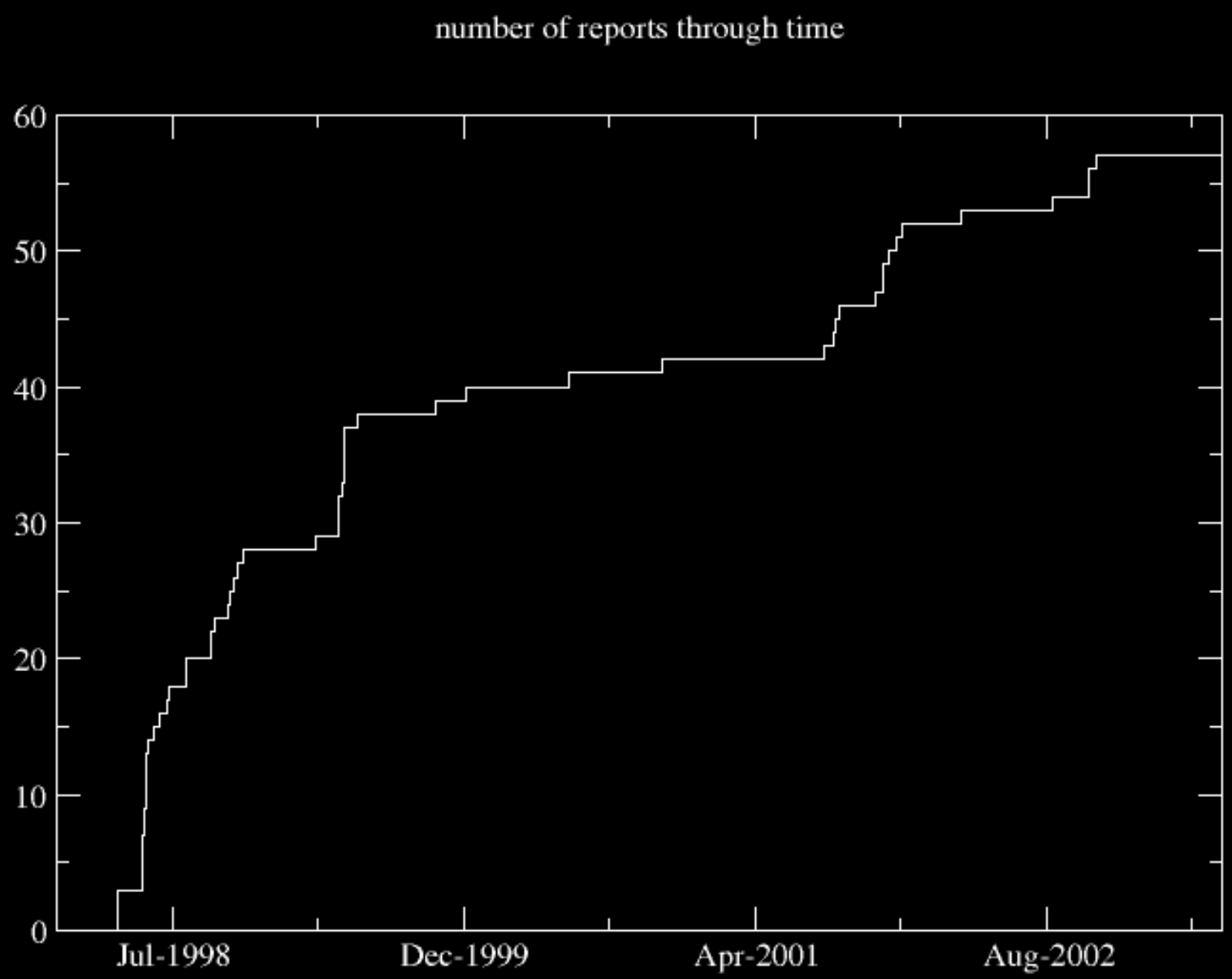
- We use the RePEc digital libray about economics
- System was conceived by Thomas Krichel
- Name “NEP” by Sune Karlsson
- Implemented by José Manuel Barrueco Cruz.
- Started to run in May 1998, has been expanding since…

General set-up

- General editor compiles a list of recent additions to the RePEc working papers data.
  - Computer generated
  - Journal articles are excluded
  - Examined by the General Editor (GE, a person)
- This list forms an issue of nep-all
- NEP-all contains all new papers
- Circulated to
  - nep-all subscribers
  - Editors of subject-reports

Subject reports

- These are filtered versions of nep-all.
- Each report has an editor who does the filtering.
- Each pertains to a subject defined by a one or more words
- Circulated by email.


number of reports through time
60
50
40
30
20
10
0
Jul-1998
Dec-1999
Apr-2001
Aug-2002

Report management

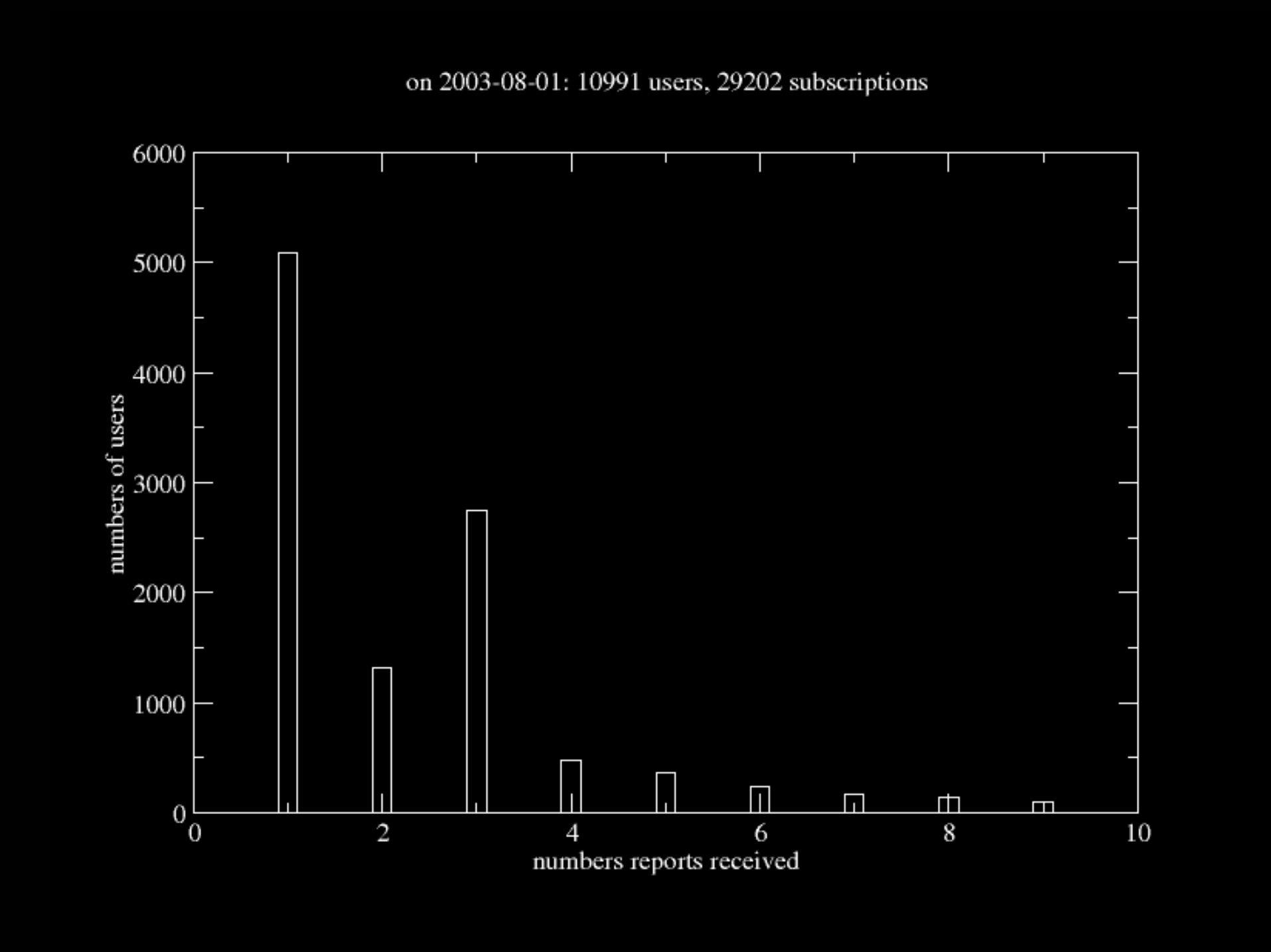
- Reports are in a flat space, without hierarchy.
- They have a varying size.
- Report creation has not followed an organized path
  - Volunteers have come forward with ideas.
  - If report creator retires as editor a volunteer among subscribers is easily found.
  - It has become practice for the GE to ask for CV before awarding an editorship.

NEP evaluation

- Ideally one would have a model of
  - Readers
  - Subjects
  - Resource constraints
- This model would predict values of observable variables in an optimum state.
- Distance between actual and optimum state can be calculated.

Data on readers

- Readers are people who have subscribed to reports.
- They are proxied by email addresses.
- Since 2003-02-01, Thomas Krichel has captured readership data
  - Once a month
  - For every report
- No historic readership data

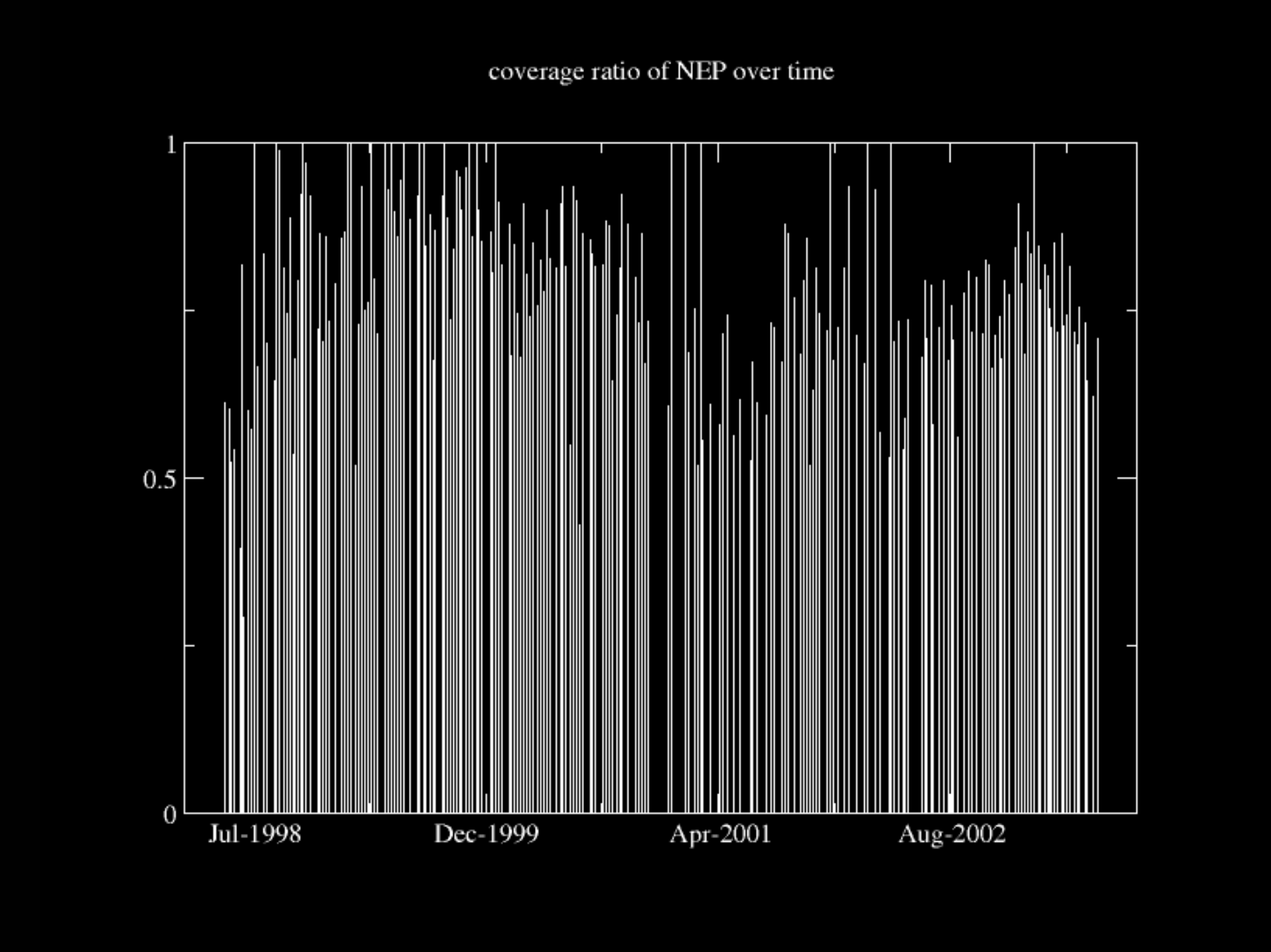

on 2003-08-01: 10991 users, 29202 subscriptions
numbers of users
numbers reports received
0
1000
2000
3000
4000
5000
6000
0
2
4
6
8
10

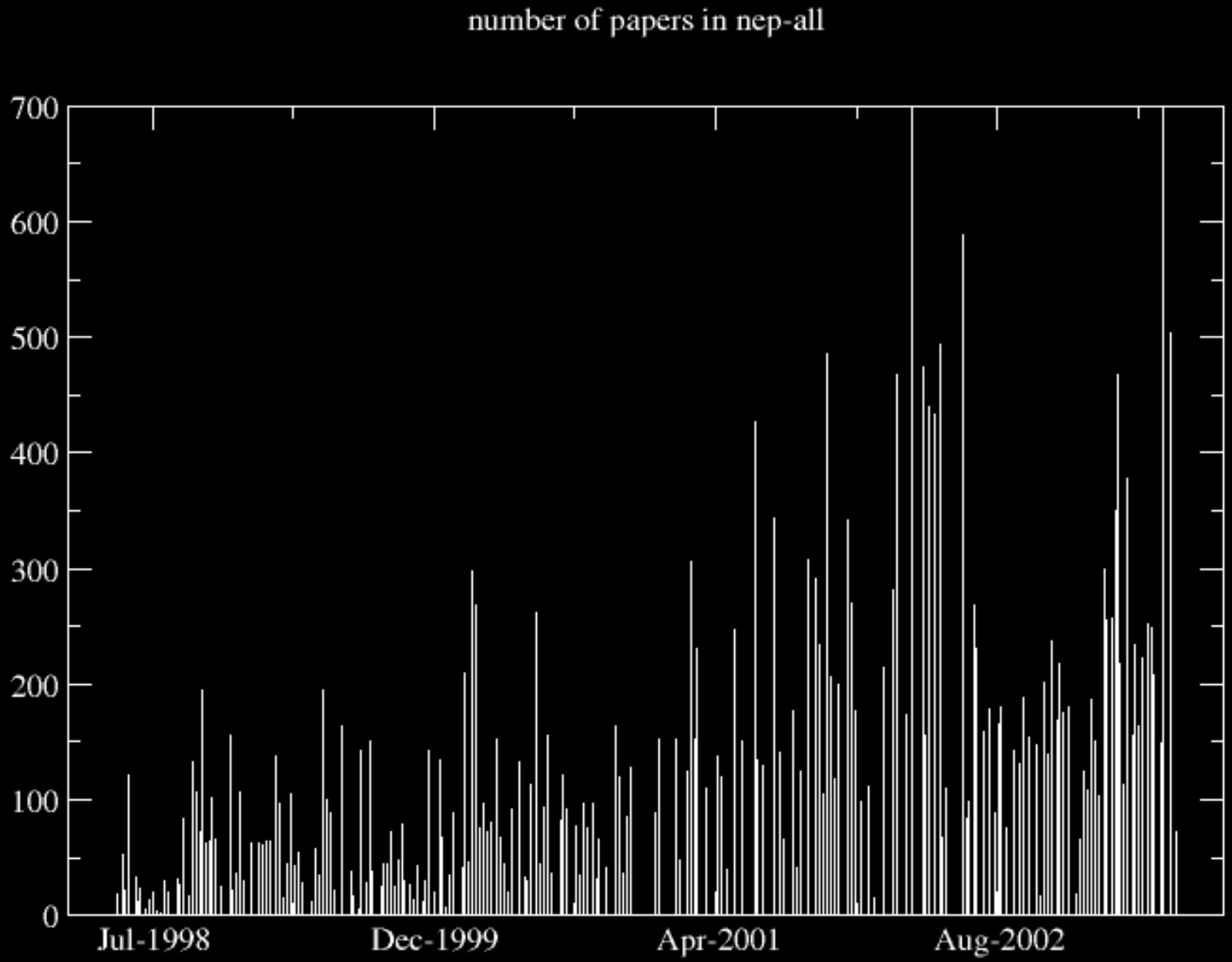
Data on papers difficult

- Logs of Mailbase, JISCMail and Mailman don't have detailed headers
  - Date information is difficult to parse and unreliable
  - Only reliable from 2003-01 with dummy subscriber set up
- Dates of issues (as opposed to mail dates) changed by editors
- Paper handles garbled up by
  - Mailing software
  - Editing software
- Report issue parser > 500 lines of Perl, growing!

Coverage ratio analysis

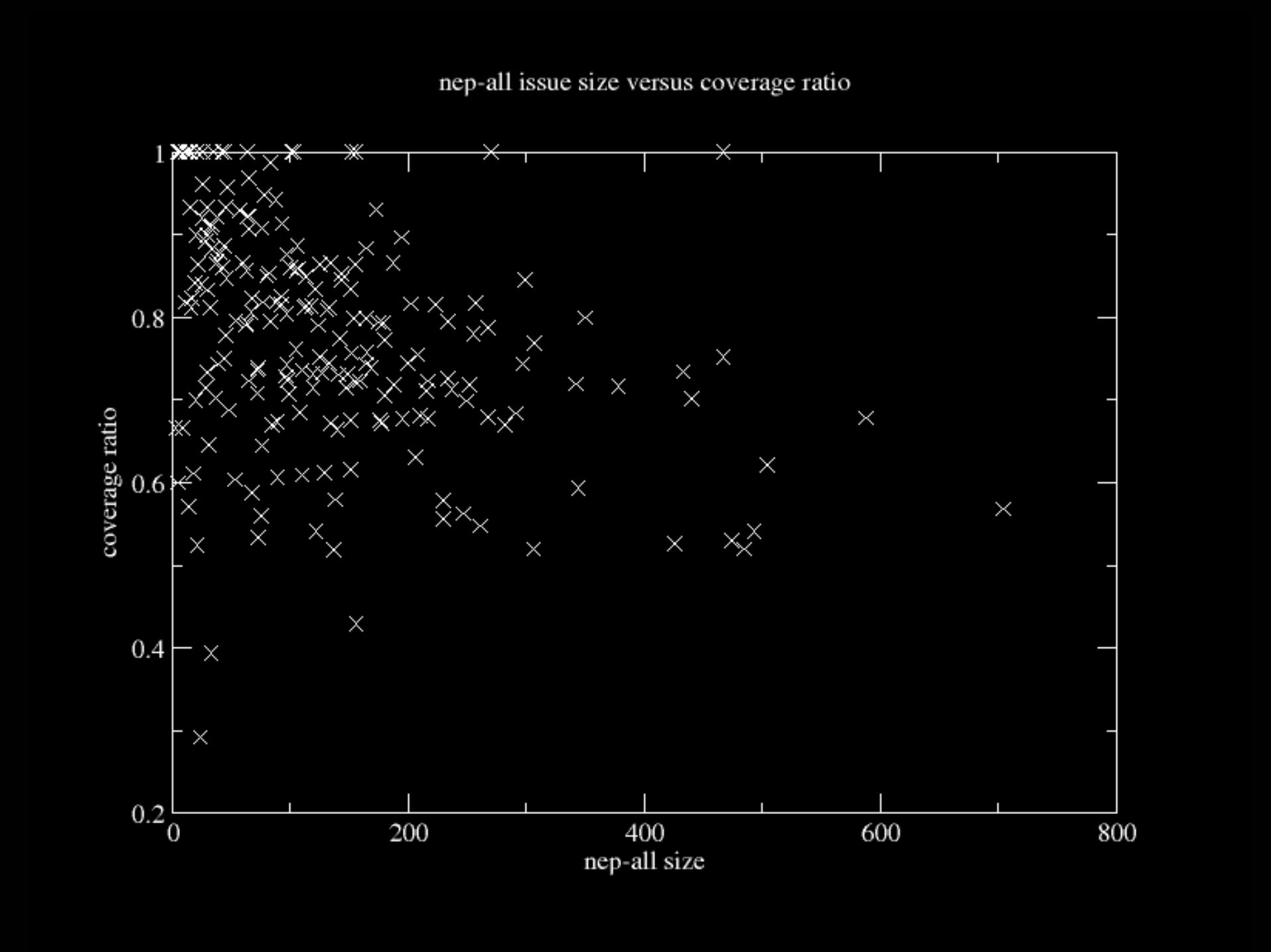
- Coverage ratio, is announced papers/size of nep-all
- It is a time varying characteristic of NEP as a whole.
- We expect it to increase over time because we have an expanding portfolio of reports.


coverage ratio of NEP over time
1
0.5
0
Jul-1998
Dec-1999
Apr-2001
Aug-2002


number of papers in nep-all
700
600
500
400
300
200
100
0
Jul-1998
Dec-1999
Apr-2001
Aug-2002

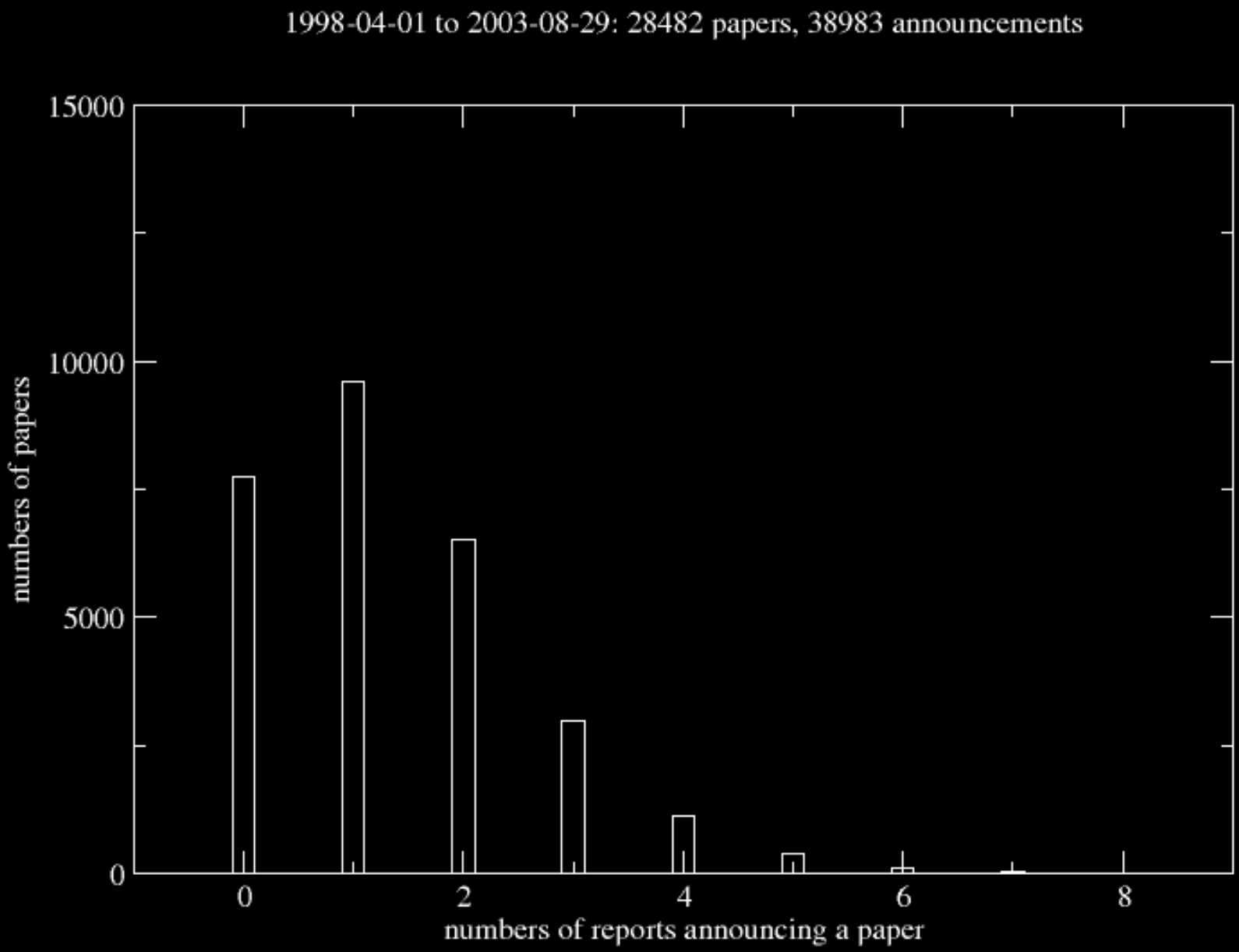
Target-size theory

- Subject concepts are fuzzy.
- Evidence of subject is flimsy at times.
- Editors have a target size for a report issue.
- Depending on the size of the nep-all issue, editors are more or less choosey.
- This theory should be most appropriate for medium-size reports. This could be confirmed by further research.


nep-all issue size versus coverage ratio
coverage ratio
nep-all size
1
0.8
0.6
0.4
0.2
0
200
400
600
800

Lousy paper theory

- Some papers in RePEc
  - are not good
  - are perceived not to be good
- They will never be announced.
- Editors dipute this theory but it may be possible to show that they are wrong.


1998-04-01 to 2003-08-29: 28482 papers, 38983 announcements
15000
10000
5000
0
numbers of papers
0
2
4
6
8
numbers of reports announcing a paper

Future developments

- Thomas Krichel sees NEP as a crucial tool for alternative peer review.
- Lousy paper theory supports that.
- But evaluation of papers is not enough. It is only a necessary step to the evaluation of the author.
- It will have to be done with respect to a neighborhood of the author.
- ACIS project is crucial.

Evaluation through downloads

- Data from Tim Brody shows that downloads data is strongly correlated with impact as measured by citations.
- But downloads of one have to be compared to a neighborhood of other documents
  - some areas of interest are more popular than others
  - logs accumulate over time
- NEP data crucial.

Download data manipulable

- If Tim´s work becomes more widely known, authors will rush to download
- This needs to be filtered.
- In addition, we need good filtering for search engine access.

Ticketing system to be done

- Ticketing is issueing a url for downloads that has an encrypted string encoding
  - report reader email address
  - report issue data
- This is not an access restriction tool.
- Repeated downloads with the same ticket will be discarded.

Aggregation

- The data is very rich
  - dissagregate per issue time
  - dissagregate per report
  - dissagregated by download time
- We need to merge with data on author using RePEc author service
- We need to produce numbers for authors. This can be done in many ways.

Conclusion

- NEP is an innovative digital library service.
  - model implementation
  - Generates rich and interesting data if properly monitored.
- Run by volunteers
  - No requirement for funding to run.
  - Technical infrastructure quite weak.
  - Needs an investment in specific software.

Thank you for your attention!

http://openlib.org/home/krichel