Network Text Analysis of R Mailing Lists

UseR! Rennes 2009

Angela Bohn, Ingo Feinerer, Kurt Hornik, Patrick Mair,
Stefan Theußl

7/10/2009

A mailing list social network

R-help mailing list:

Jan 2008 to May 2009

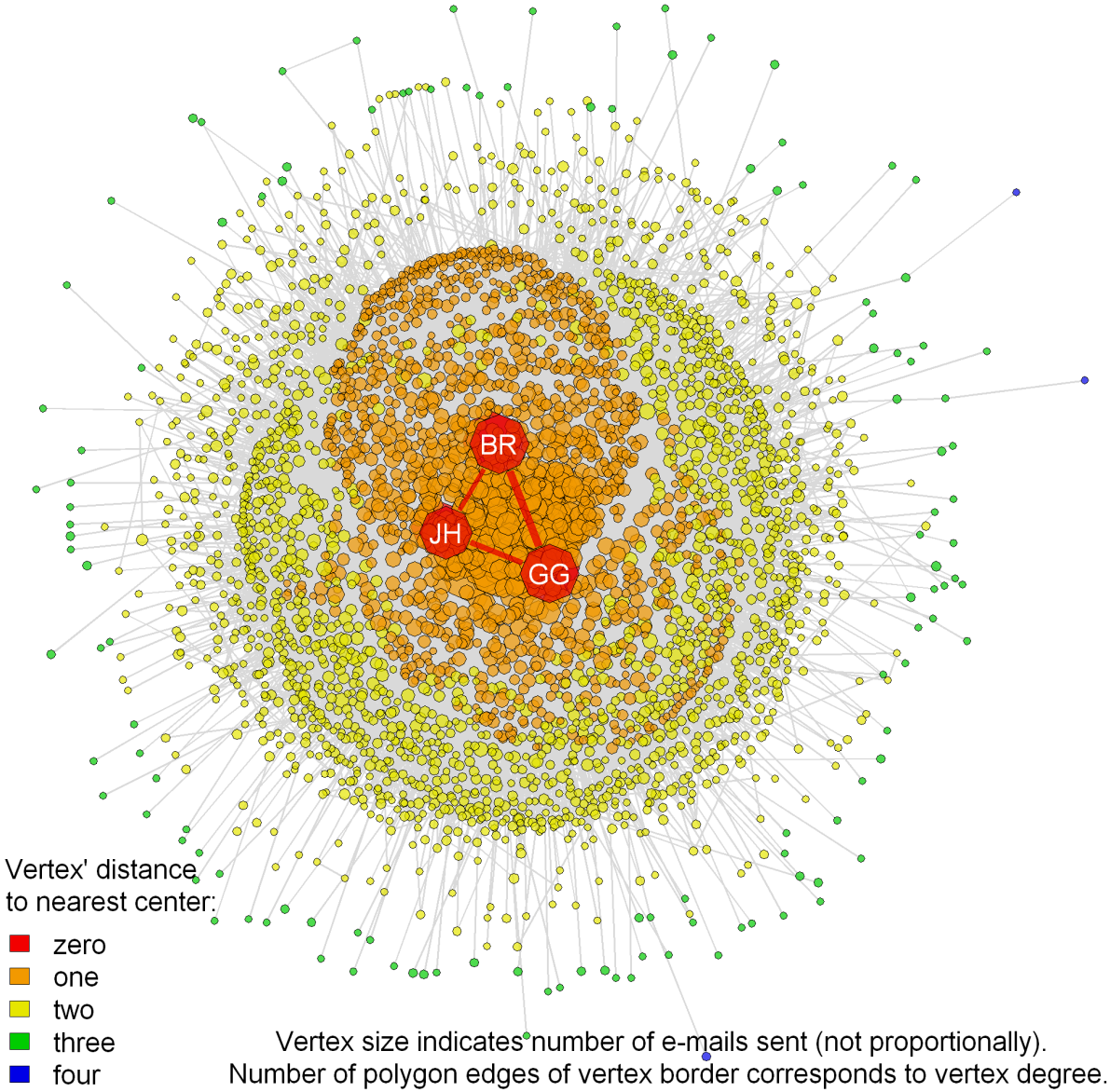
Number of authors: 5326

Number of mails: 41457

Avg. degree: 4.4

Diameter: 7

Legend:



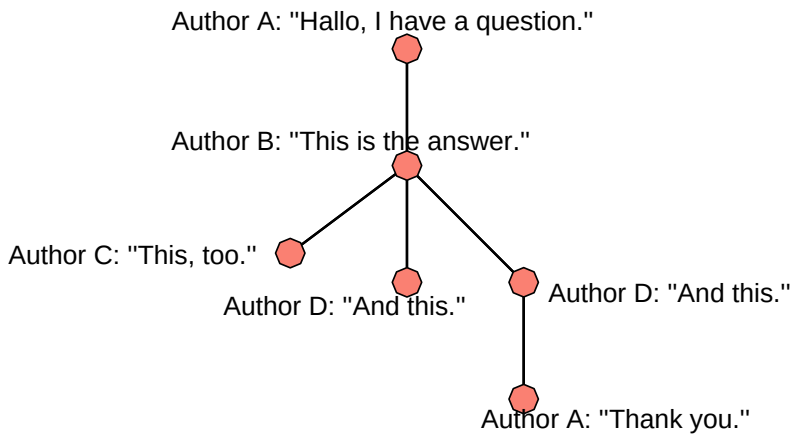
Combine SNA and TM

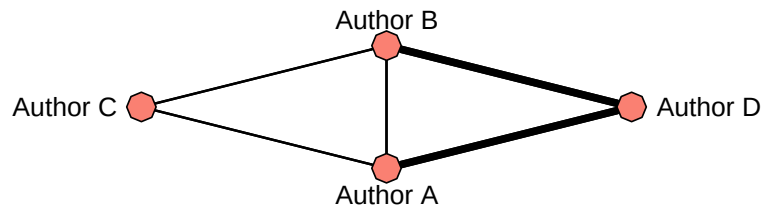
- Goal: Combine social network analysis (SNA) and text mining (TM) to find out more
- Data: Mailing lists R-help and R-devel
- Packages: `sna` and `tm`
- Results:
  - "Interest maps" of R users
  - Detection of bottlenecks in communication

Data preparation for social network analysis

- Create a social network from e-mail headers (`tm`):

From: dwinsemius at comcast.net (David Winsemius)
Date: Thu, 30 Apr 2009 18:49:55 -0400
Subject: [R] Extracting Element from S4 objects
In-Reply-To: <23302265.post@talk.nabble.com>
Message-ID: <A6039F4E-ABF4-41C5-B03E-FFF32E07C37A@comcast.net>





- Find aliases:

```
knoblauch at lyon.inserm.fr (knoblauch)
knoblauch at lyon.inserm.fr (Ken Knoblauch)
ken.knoblauch at inserm.fr (Kenneth Knoblauch)
ken.knoblauch at inserm.fr (Ken Knoblauch)
```

Levensthein Distance:
`agrep(base)`

Data preparation for text mining

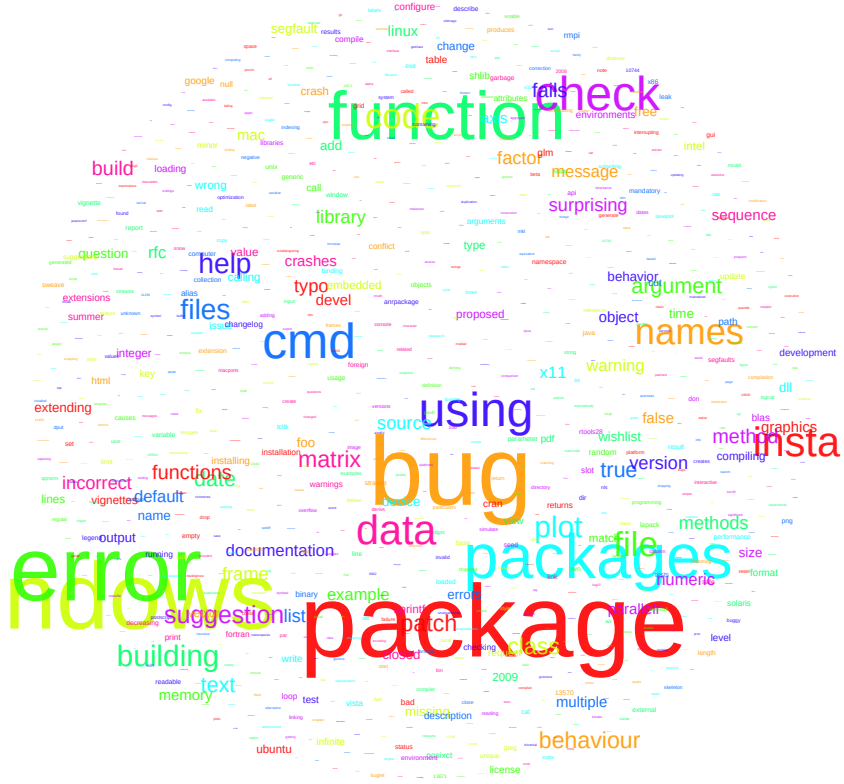
E-mail subjects:

```
[R] passing args from the command line
[R] navigating ggplot viewports
[R] how to go to a line in R
```

...

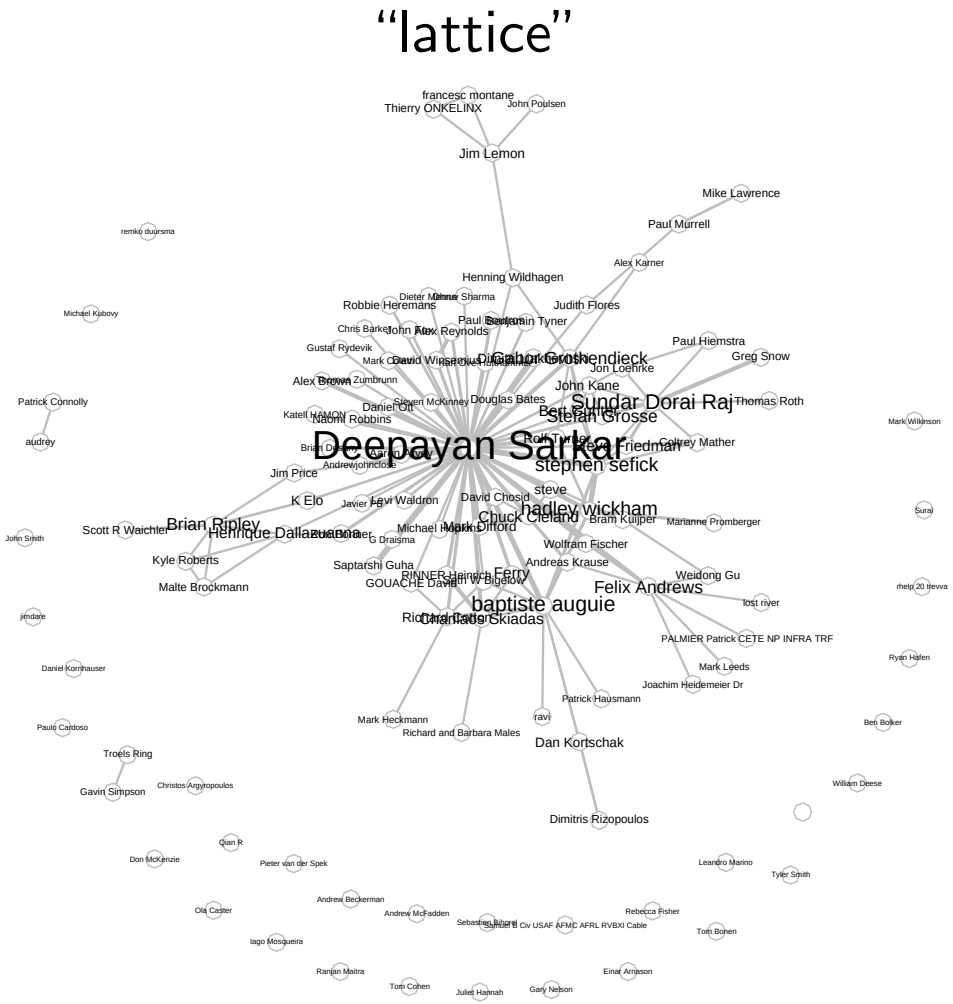
Term Frequencies (`termFreq(tm)`):

R-help

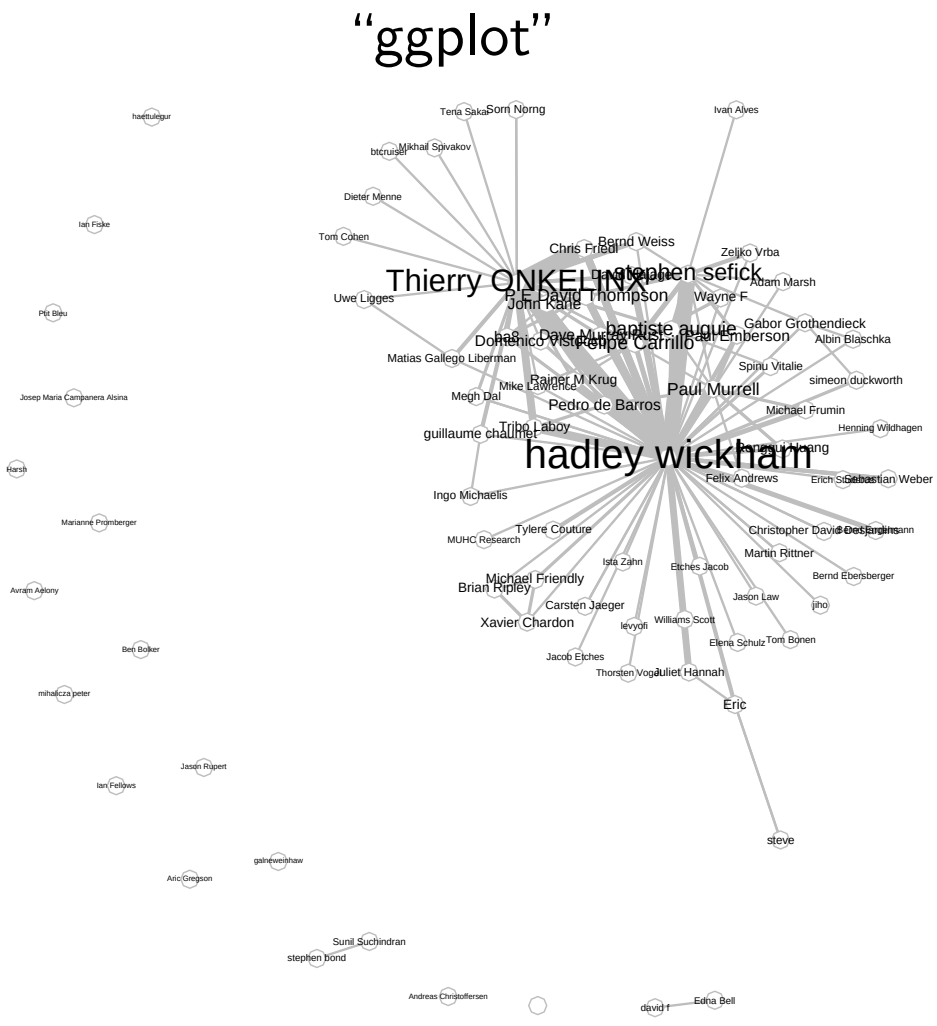


R-devel

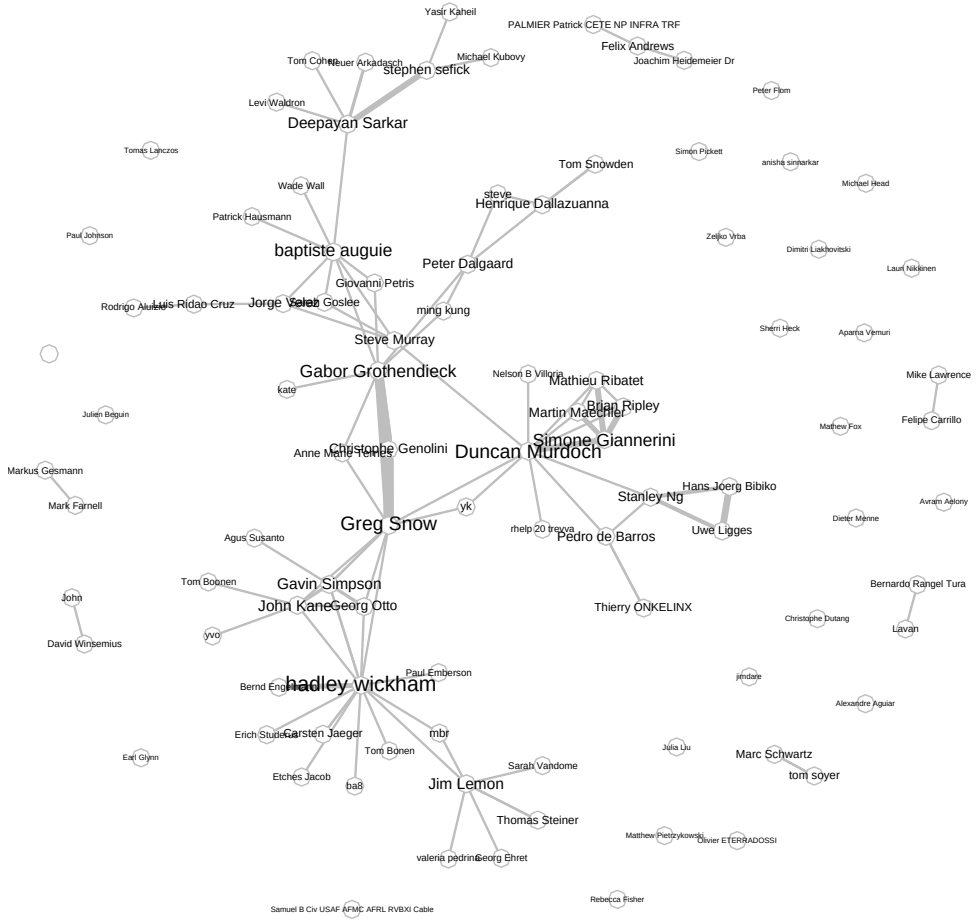
Word Networks



Word Networks

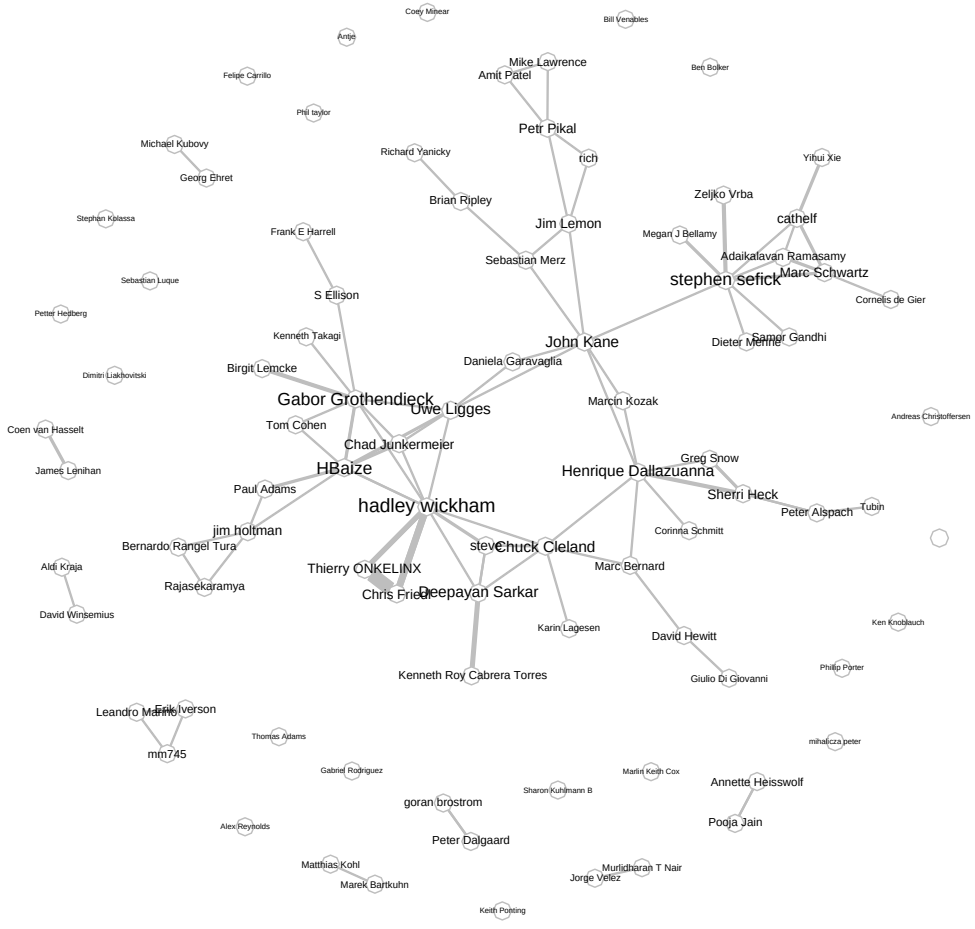


Word Networks

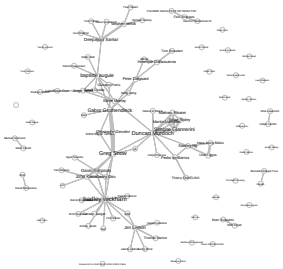
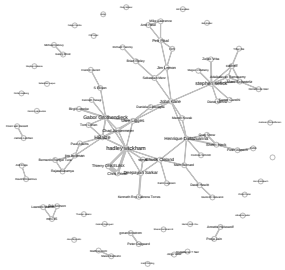


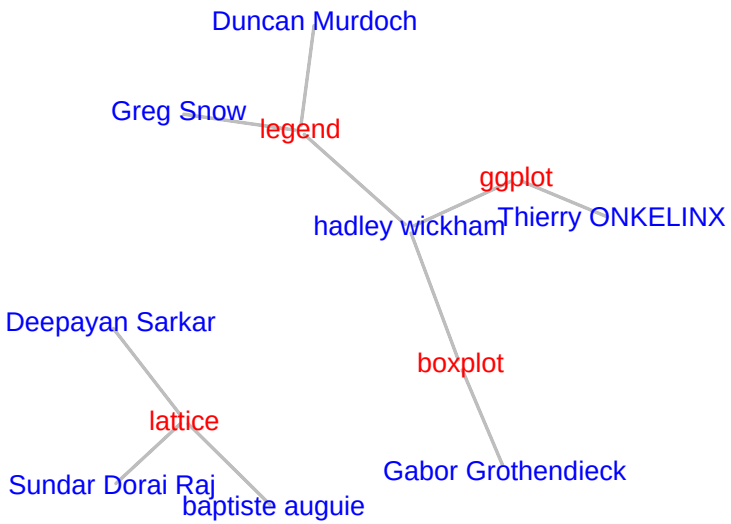
Word Networks

"boxplot"

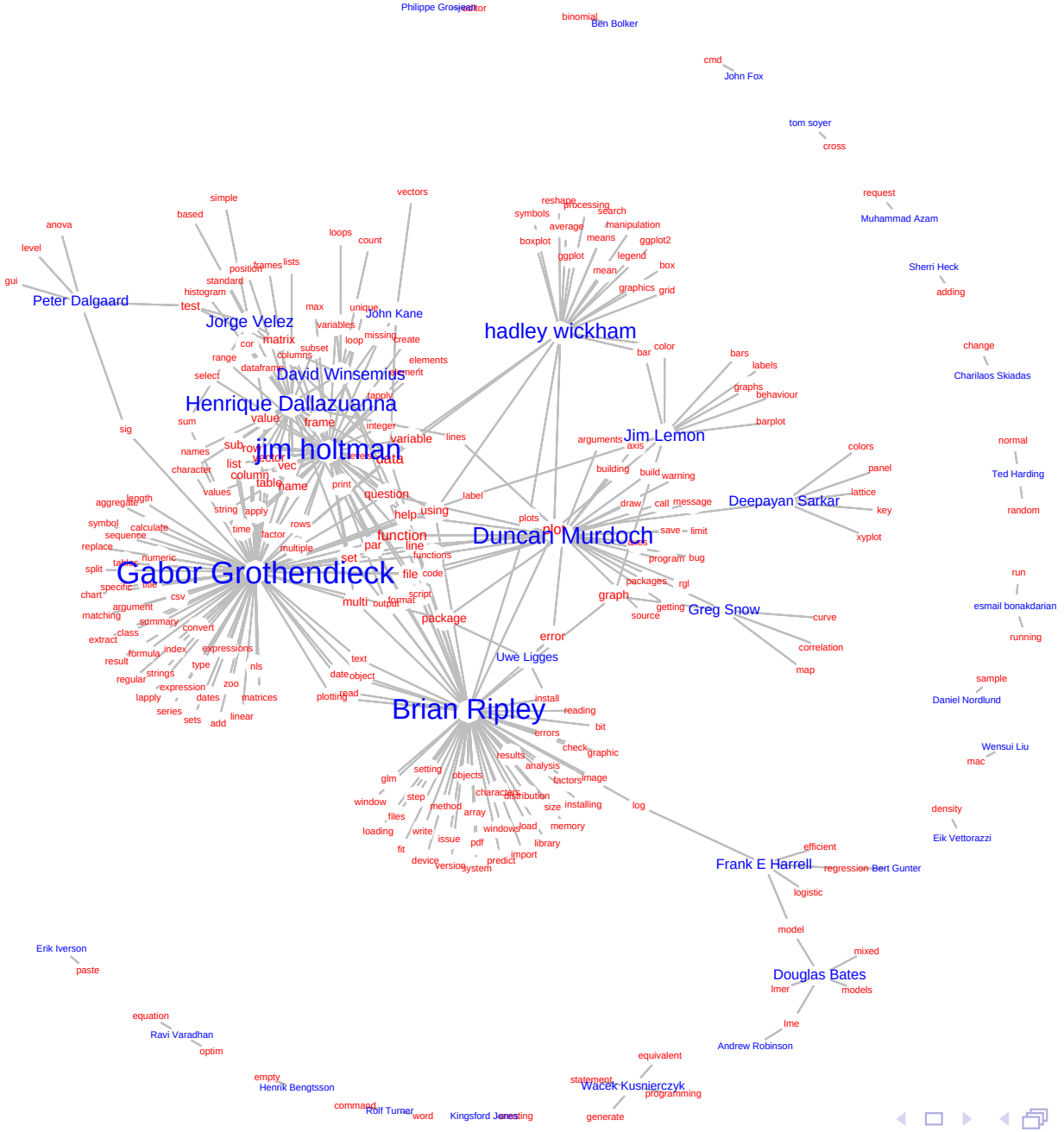


Centrality Measures

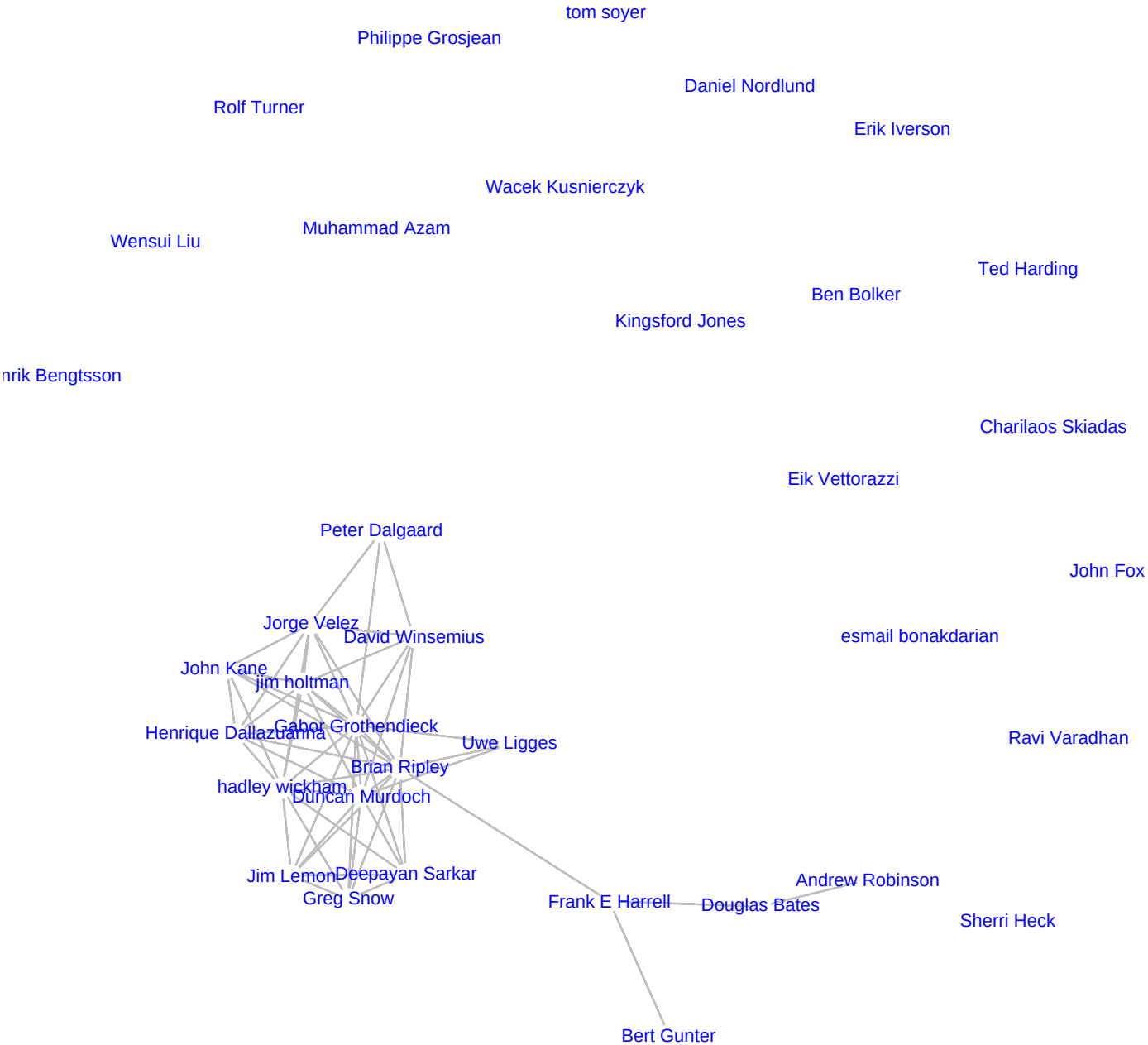
Notion	lattice	ggplot	legend	boxplot
				
Most central persons	Deepayan Sarkar, Sundar Dorai Raj, baptiste auguie	hadley wickham, Thierry ONKELINX	Duncan Murdoch, hadley wickham, Greg Snow	Gabor Grothendieck, hadley wickham



Results: Interest maps

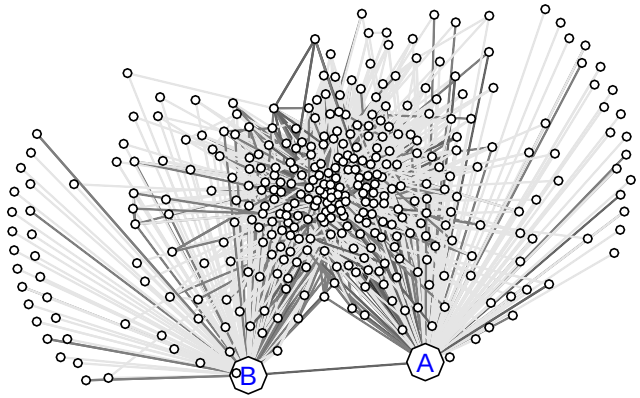


Results: Communication bottlenecks

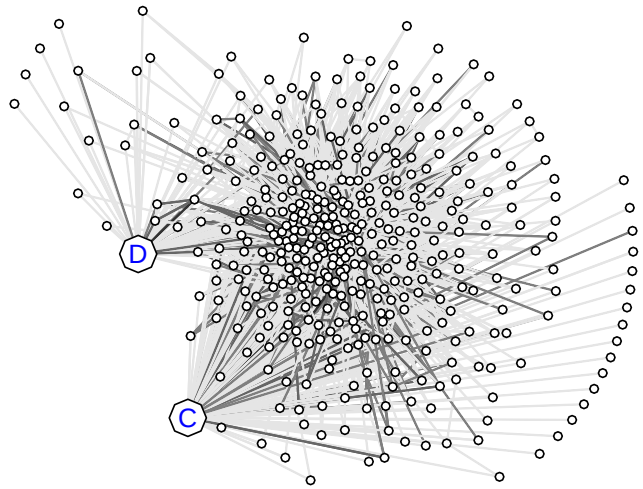


Results: Communication bottlenecks

Good.



Can be improved.



Thank you!

Packages:

`sna`: Carter T. Butts (2008). Social Network Analysis with `sna`. Journal of Statistical Software 24/6.

`tm`: I. Feinerer, K. Hornik, and D. Meyer (2008). Text Mining Infrastructure in R. Journal of Statistical Software 25/5.

References:

C. Bird, A. Gourley, P. Devanbu, M. Gertz, and A. Swaminathan. Mining email social networks. In Proceedings of the 2006 international workshop on Mining software repositories. ACM, New York, 2006.

Contact:

`angela.bohn@gmail.com`, `www.angela-bohn.de`