ETH zürich

When your social position predicts your success

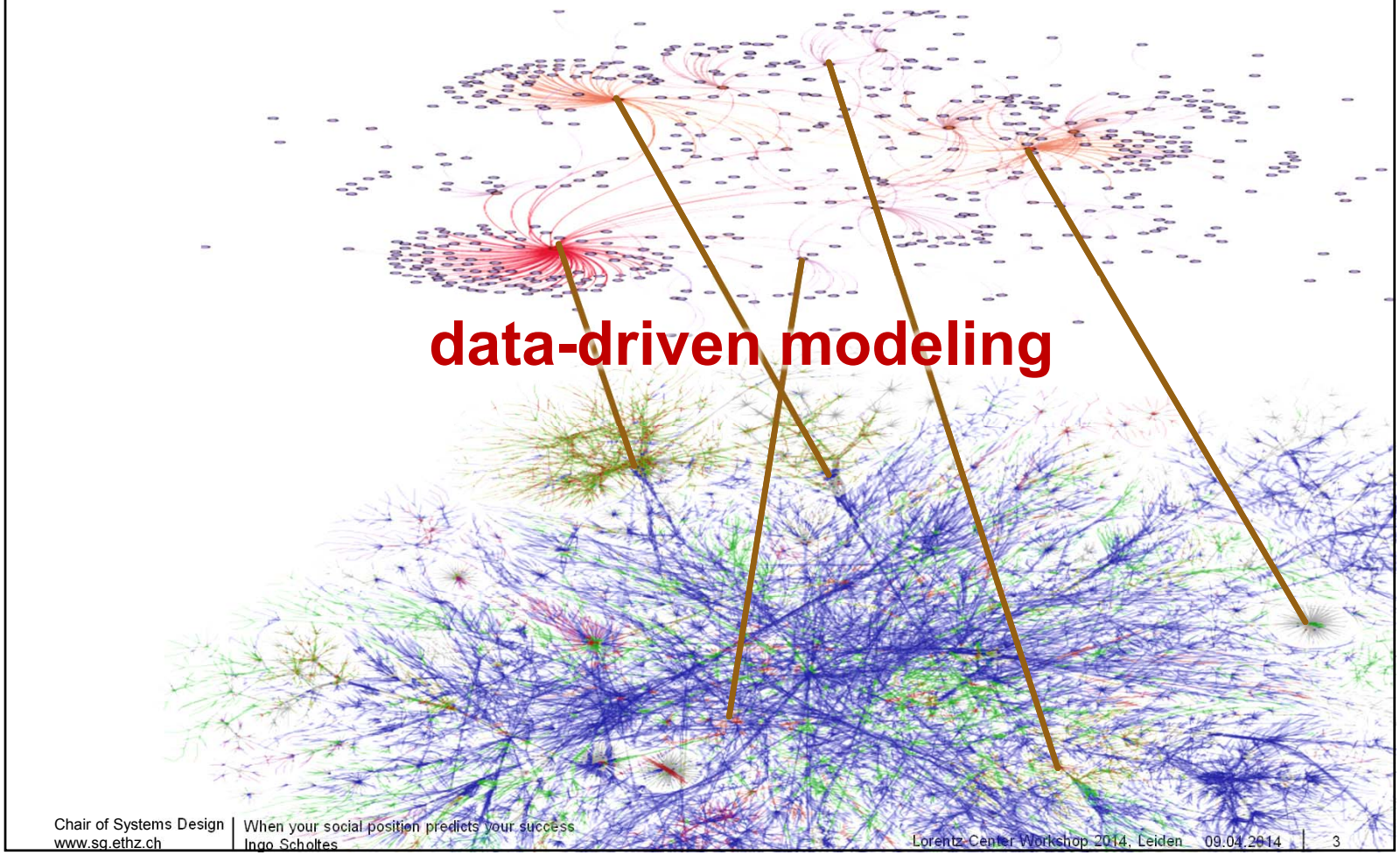
Lessons from Open Source communities and citations

Ingo Scholtes[1]
Chair of Systems Design
ETH Zürich
[1] in collaboration with René Pfitzner, Emre Sarigöl, Marcelo Serrano Zanetti, Antonios Garas and Frank Schweitzer

Chair of Systems Design @ ETH Zürich

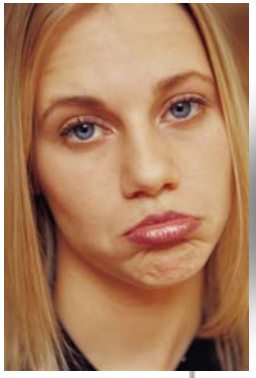
ETH zürich

Socio-technical systems

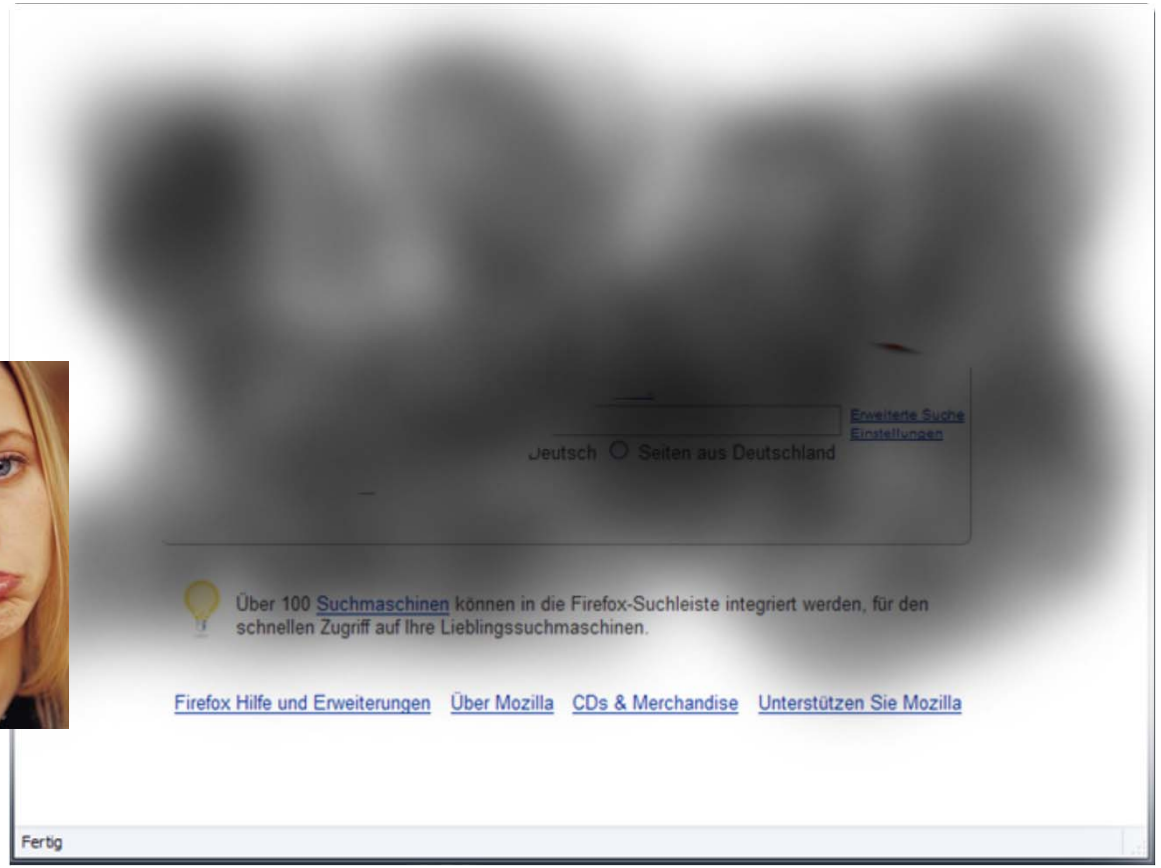


ETH zürich

Software engineering

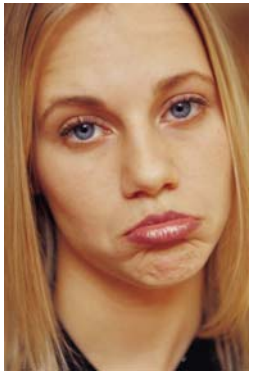


user



ETH zürich

Software defects: bug tracking



user

information overload!

developers

ETH zürich

	Firefox	Thunderbird	Eclipse	NetBeans	Total
Start date	April 2002	January 2000	October 2001	January 1999	
Bug reports	112,968	35,388	356,415	210,921	715,692
Changes	1,068,070	313,957	2,594,385	1,875,878	5,852,290
Valid	21 %	24 %	74 %	63 %	57 %
Faulty	79 %	76 %	26 %	37 %	43 %

Firefox: more than 79 % faulty bug reports

ETH zürich

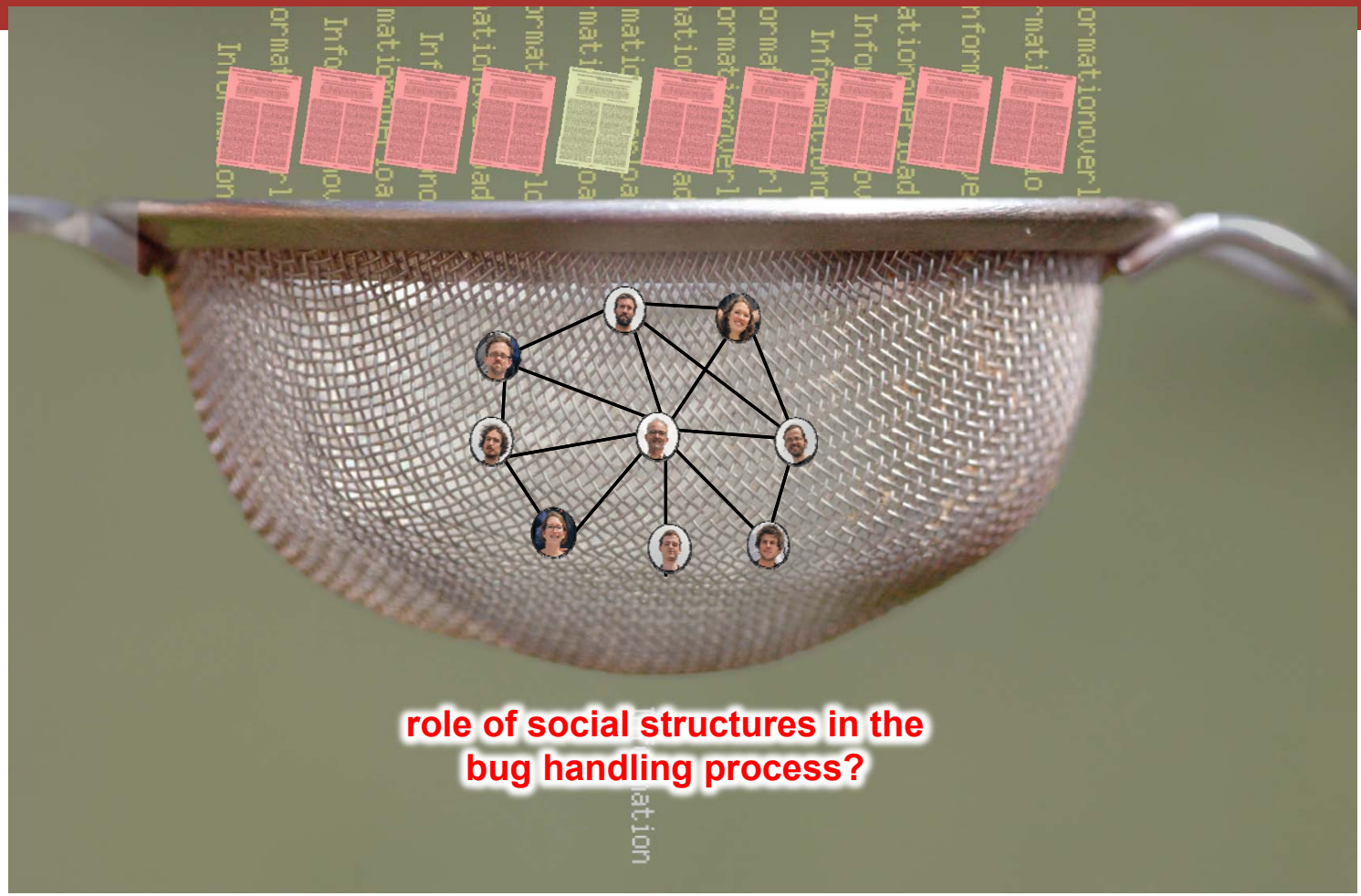
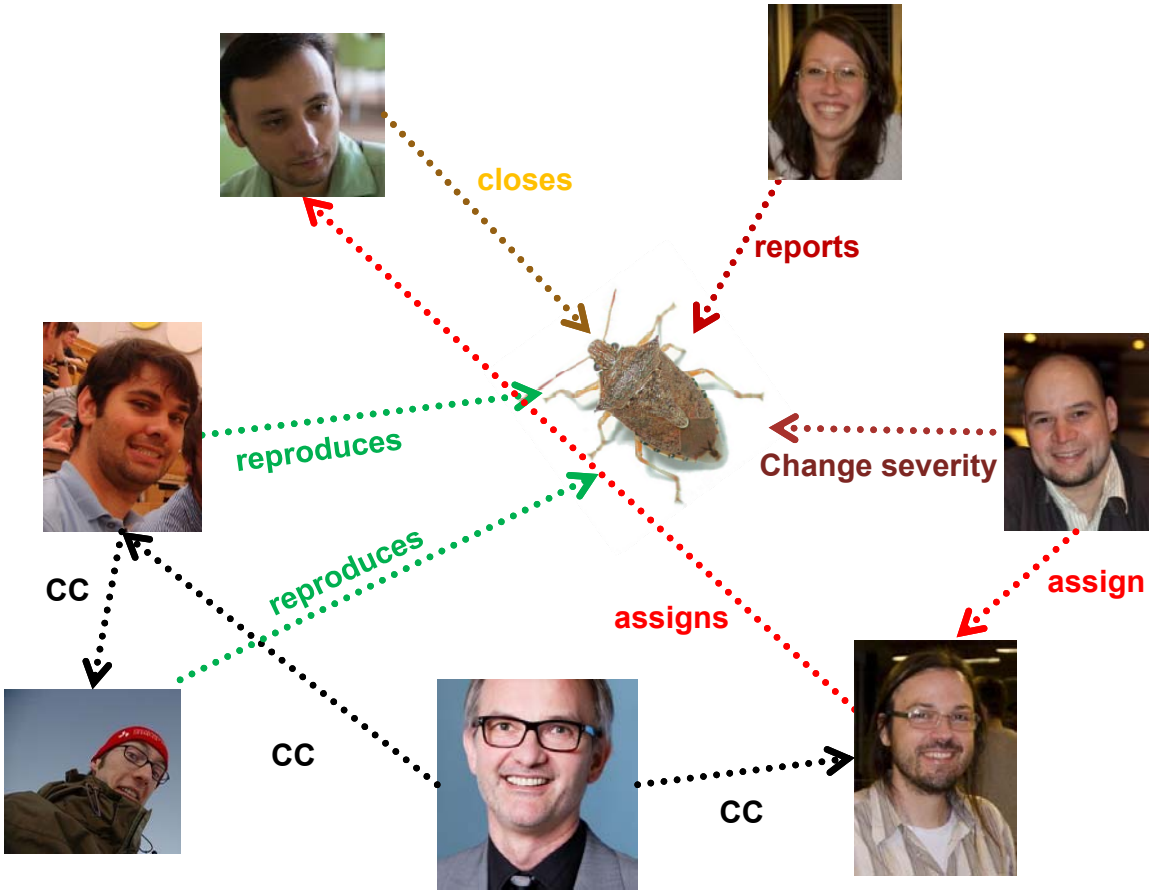


Image: Marina Noordegraaf, Flickr

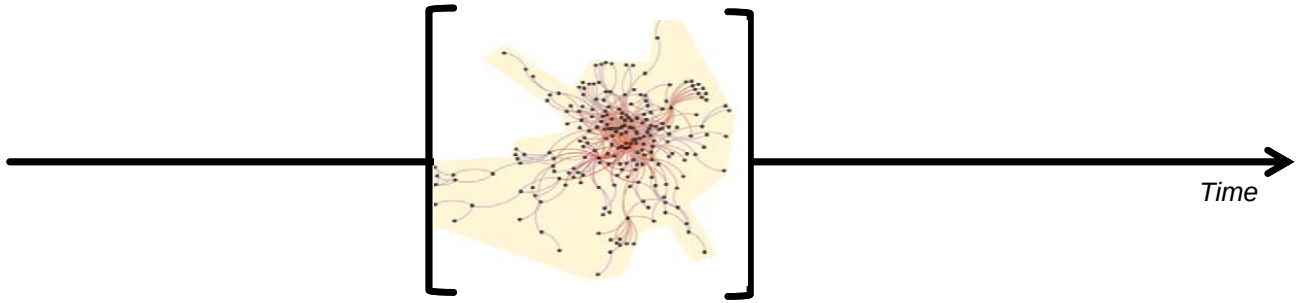
ETH zürich

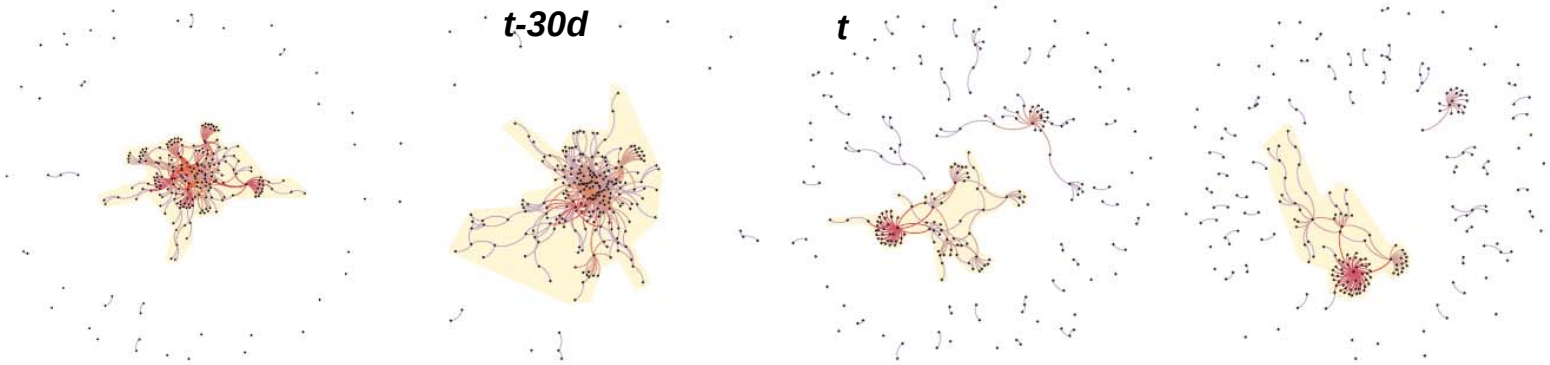
From bug interactions …



ETH zürich

… to evolving collaboration networks





(a) ECLIPSE (Dec. 2002)
244 nodes, 319 links

(b) NETBEANS (Jun. 2006)
246 nodes, 513 links

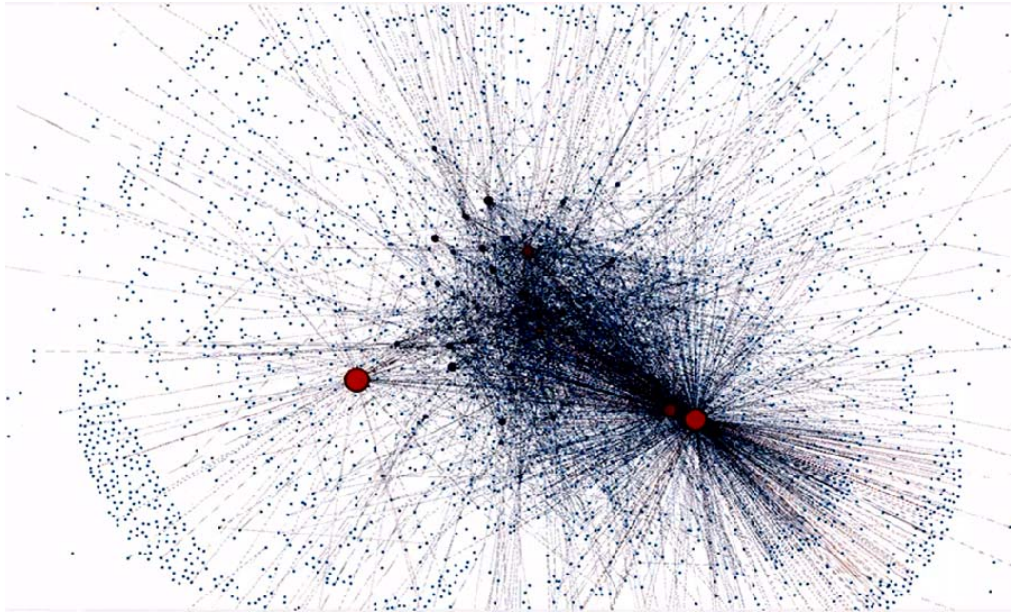
(c) FIREFOX (Oct. 2003)
241 nodes, 184 links

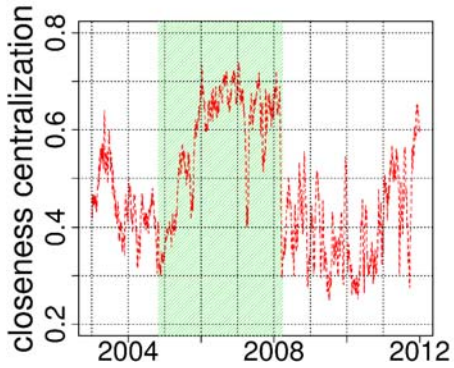
(d) THUNDERBIRD (Nov. 2004)
245 nodes, 170 links

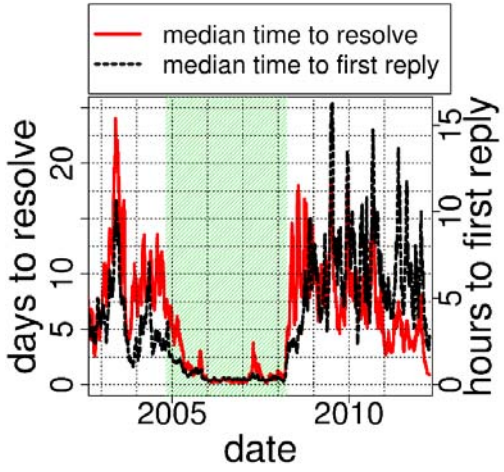
Marcelo S. Zanetti, Emre Sarigol, Ingo Scholtes, Claudio J. Tessone, Frank Schweitzer: **A Quantitative Study of Social Organization in Open Source Software Communities,** *In Proceedings of Imperial College Computing Student Workshop (ICCSW), Open Access Series in Informatics, pp. 116-122, London, UK, September 27-28, 2012*

ETH zürich

Evolution of social organization





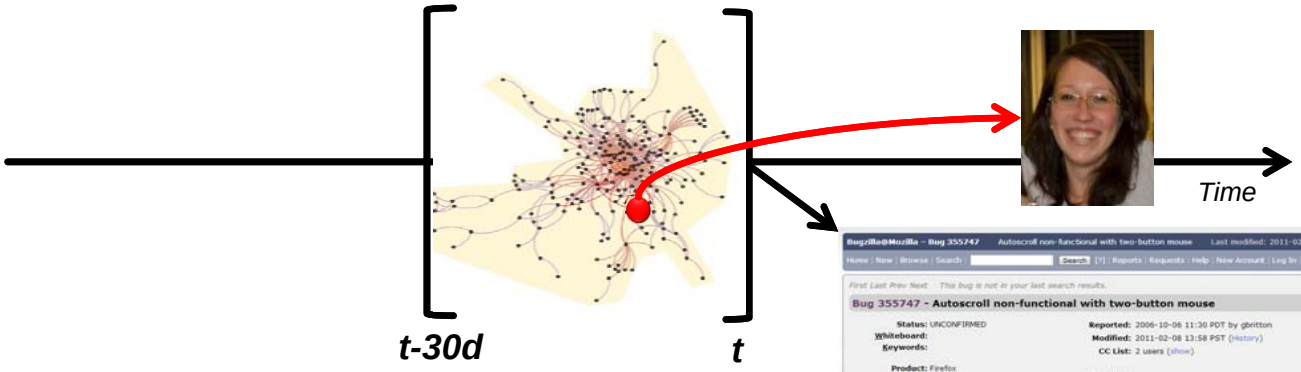


we observe how team structures influence team performance

Marcelo S. Zanetti, Ingo Scholtes, Claudio J. Tessone and Frank Schweitzer: **The Rise and Fall of a Central Contributor: Centralization and Performance in the Gentoo Community**, *In Proceedings of the 6th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE) held at ICSE 2013, San Francisco, CA, USA, 2013*

ETH zürich

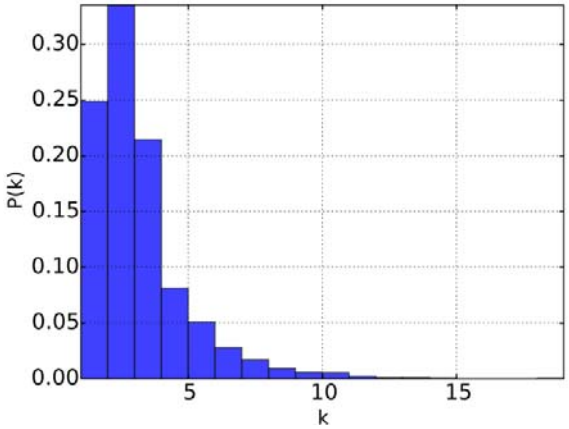
Back to information filtering



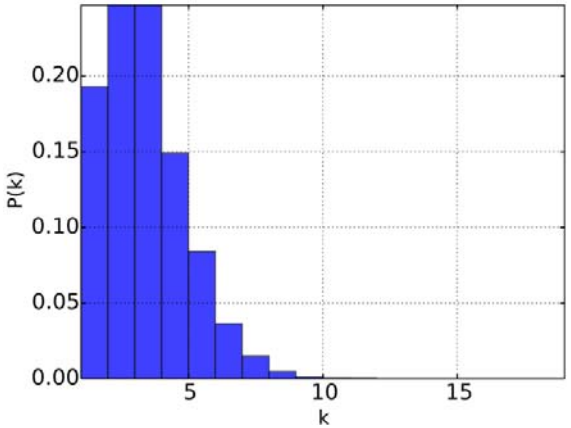
Hypothesis: position of bug reporter in the monthly collaboration network preceding the time of report is indicative for outcome of the bug handling process

ETH zürich

Statistical hypothesis testing



eigenvector centralities of users reporting **FIXED** bugs



eigenvector centralities of users reporting **INVALID** bugs

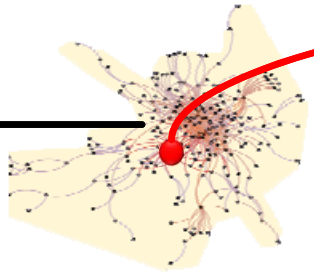
Hypothesis: position of bug reporter in the monthly collaboration network preceding the time of report is indicative for outcome of the bug handling process

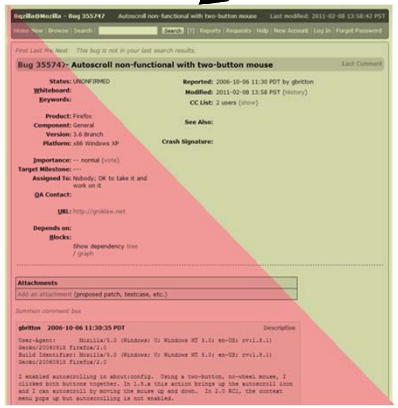
YES, it is!

ETH zürich

Prediction method

1. in largest connected component?
2. eigenvector centrality
3. closeness centrality
4. betweenness centrality
5. k-coreness
6. total degree
7. in-degree
8. out-degree
9. local clustering coefficient





Support Vector Machine

valid or faulty?

training set

ETH zürich

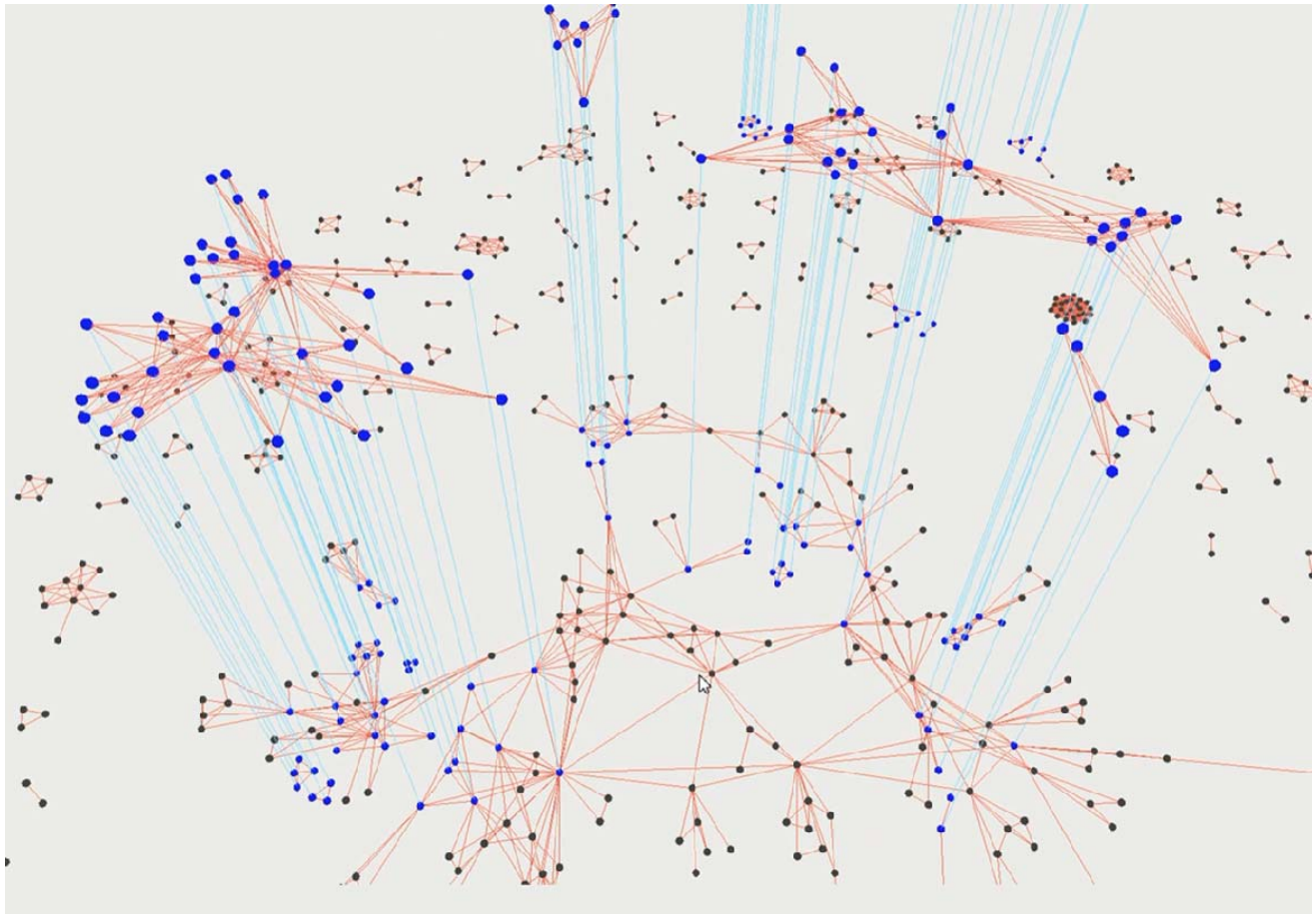
Prediction results

	Firefox	Thunderbird	Eclipse	NetBeans
Valid	21 %	23 %	74 %	63 %
Precision	82.5 %	90.3 %	88.7 %	78.9 %
Recall	44.5 %	38.9 %	91.0 %	87.0 %
F-Score	0.58	0.54	0.89	0.83

- ✓ **better prediction than previous approaches on same data**
- ✓ **the social network matters!**

ETH zürich

Collaboration–Citation-Multiplex



ETH zürich

Collaboration–Citation-Multiplex

arXiv:1402.7268v1 [physics.soc-ph] 28 Feb 2014

Predicting Scientific Success Based on Coauthorship Networks

Emre Sarigöl, René Pfitzner*, Ingo Scholtes, Antonios Garas, Frank Schweitzer
Chair of Systems Design, ETH Zurich
CH-8092 Zurich, Switzerland
*rpfitzner@ethz.ch

Abstract

We address the question to what extent the success of scientific articles is due to social influence. Analyzing a data set of over 100 000 publications from the field of Computer Science, we study how centrality in the coauthorship network differs between authors who have highly cited papers and those who do not. We further show that a machine learning classifier, based only on coauthorship network centrality measures at time of publication, is able to predict with high precision whether an article will be highly cited five years after publication. By this we provide quantitative insight into the social dimension of scientific publishing – challenging the perception of citations as an objective, socially unbiased measure of scientific success.

1 Introduction

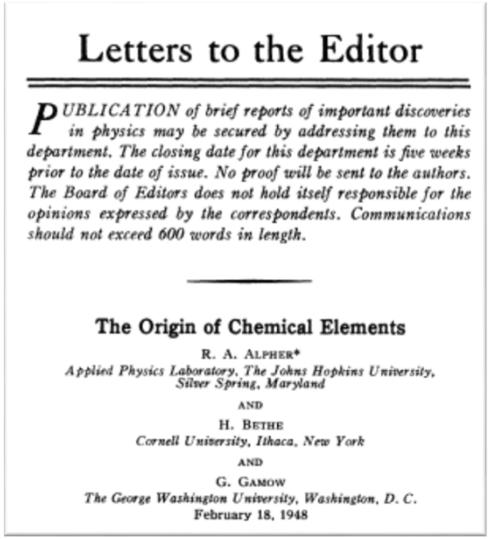
Quantitative measures are increasingly used to evaluate the performance of research institutions, departments, and individual scientists. Measures like the absolute or relative number of published research articles are frequently applied to quantify the *productivity* of scientists. To measure the *impact* of research, citation-based measures like the total number of citations, the number of citations per published article or the h-index [9], have been proposed. Proponents of such citation-based measures or rankings argue that they allow to quantitatively and objectively assess the *quality* of research, thus encouraging their use as simple proxies for the *success* of scientists, institutions or even whole research fields. The intriguing idea that by means of citation metrics the task of assessing research quality can be "outsourced" to the *collective intelligence* of the scientific community, has resulted in citation-based measures becoming increasingly popular among research administrations and governmental decision makers. As a result, such measures

ETH zürich

Evolving coauthorship networks

α

Ralph **Alpher**



Letters to the Editor

PUBLICATION of brief reports of important discoveries in physics may be secured by addressing them to this department. The closing date for this department is five weeks prior to the date of issue. No proof will be sent to the authors. The Board of Editors does not hold itself responsible for the opinions expressed by the correspondents. Communications should not exceed 600 words in length.

The Origin of Chemical Elements

R. A. ALPHER*
Applied Physics Laboratory, The Johns Hopkins University, Silver Spring, Maryland
AND
H. BETHE
Cornell University, Ithaca, New York
AND
G. GAMOW
The George Washington University, Washington, D. C.
February 18, 1948

γ

George **Gamow**

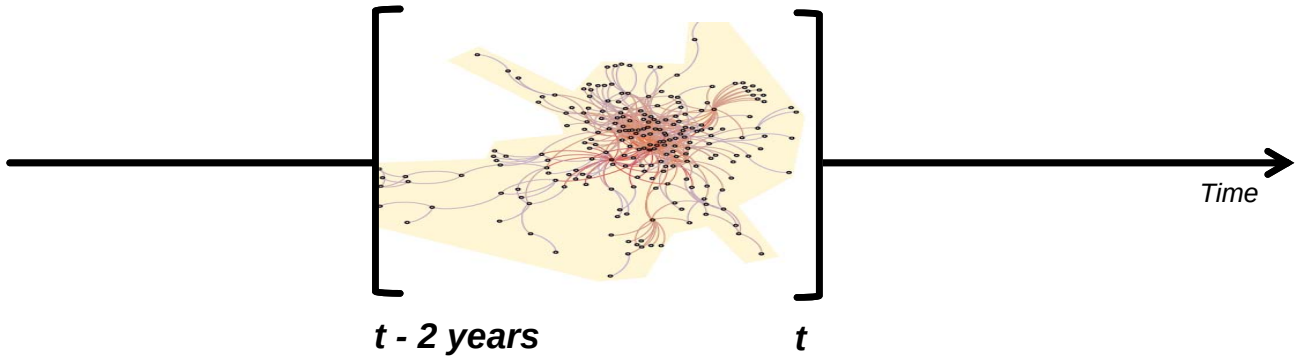
β

Hans **Bethe**

Image credit: Wikimedia Commons

ETH zürich

Evolving coauthorship networks



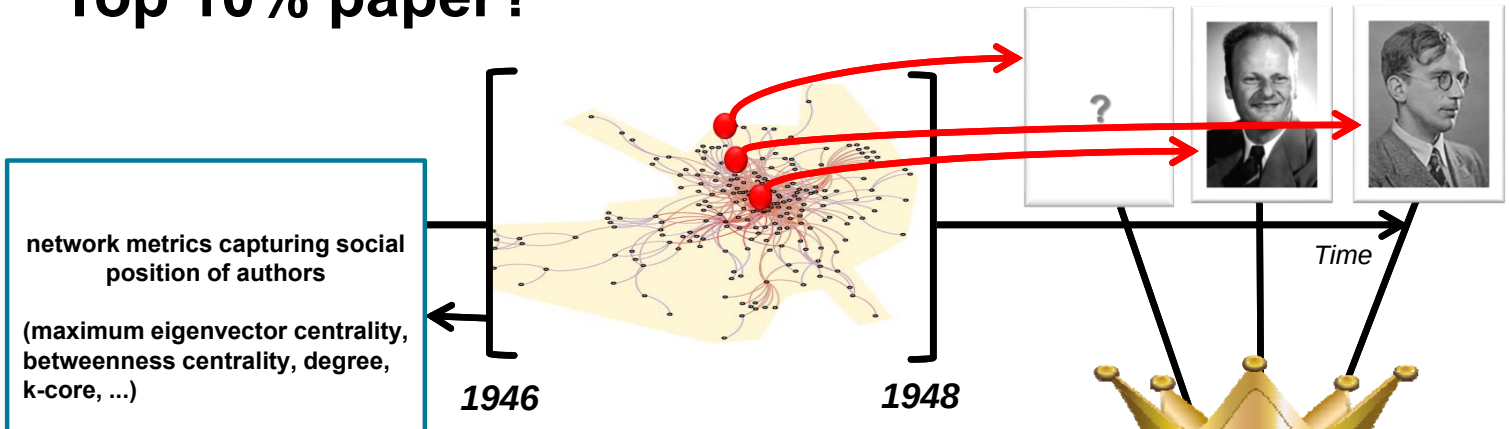
Hypothesis: position of authors in the coauthorship network up to two years before time of publication is indicative for citation success of a paper

YES, it is!

E Sarigöl, R Pfitzner, I Scholtes, A Garas and F Schweitzer: **Predicting Scientific Success Based on Coauthorship Networks,** February 2014, http://arxiv.org/abs/1402.7268

ETH zürich

Top 10% paper?



network metrics capturing social position of authors

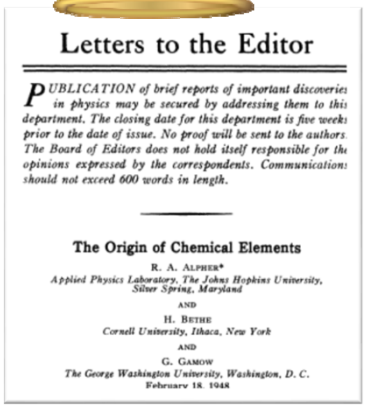
(maximum eigenvector centrality, betweenness centrality, degree, k-core, ...)

Random Forest Classifier

top 10% paper by the year 1953?

training set

data set
108,758 CS publications
160,891 researchers
5,3 Million collaborations



Letters to the Editor

PUBLICATION of brief reports of important discoveries in physics may be secured by addressing them to this department. The closing date for this department is five weeks prior to the date of issue. No proof will be sent to the authors. The Board of Editors does not hold itself responsible for the opinions expressed by the correspondents. Communications should not exceed 600 words in length.

The Origin of Chemical Elements

R. A. ALPHER*
Applied Physics Laboratory, The Johns Hopkins University, Silver Spring, Maryland
AND
H. BETHE
Cornell University, Ithaca, New York
AND
G. GAMOW
The George Washington University, Washington, D. C.
February 18, 1948

Image credit: Wikimedia Commons

ETH zürich

Social position correlates with future success

Results (CS)	
Precision	60 %
Recall	18 %
F-Score	0.28

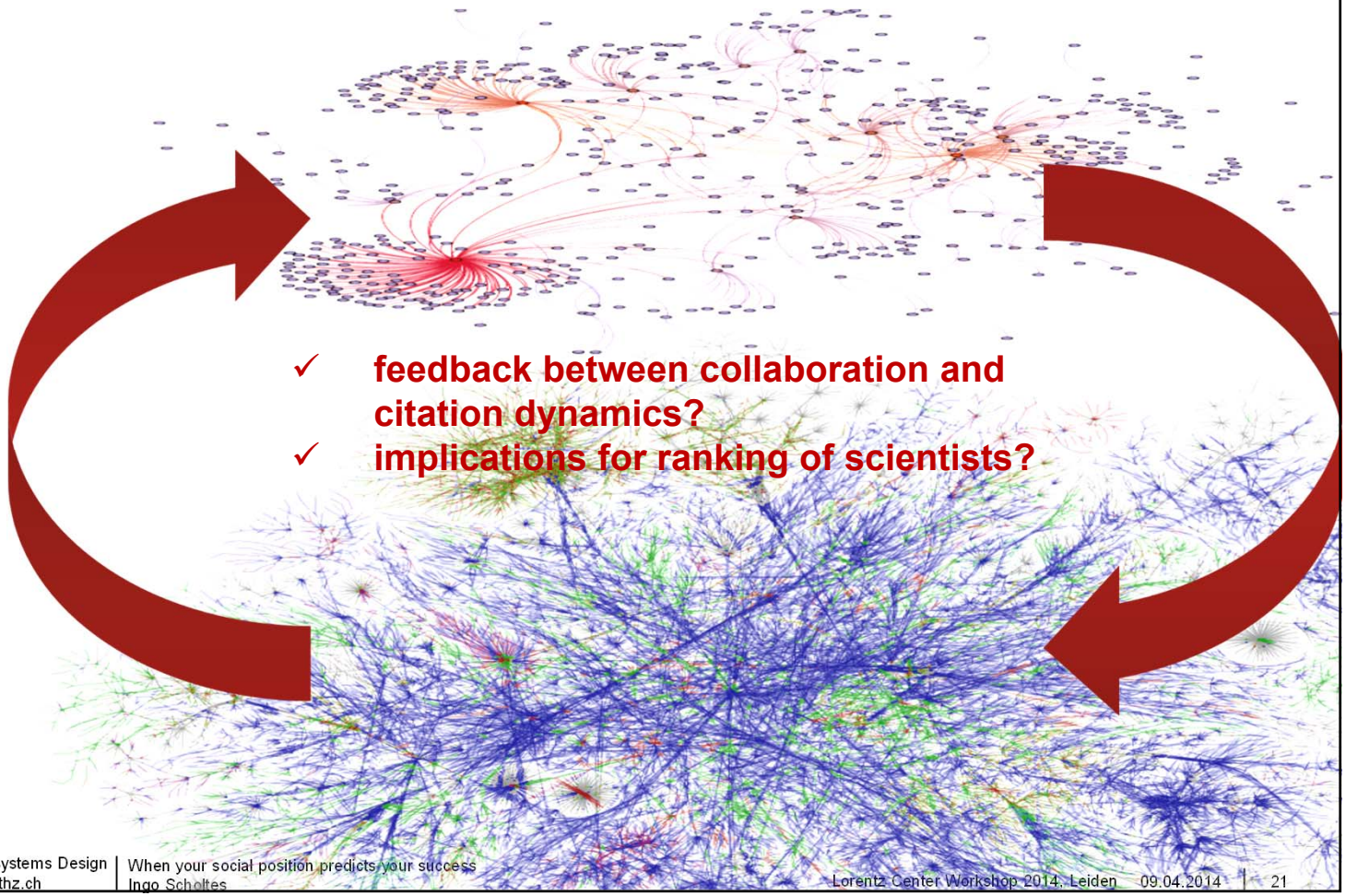
✓ six times better than random guess

✓ social position of authors alone sufficient to predict citation success of 2,000 out of 10,800 top papers

- ✓ data mining + network analysis + predictive analytics
- ✓ highlights influence of social structures in information filtering
- ✓ reputation is multi-dimensional and time-varying concept

ETH zürich

Closing the loop?



- ✓ feedback between collaboration and citation dynamics?
- ✓ implications for ranking of scientists?

ETH zürich

Thank you!

E Sarigöl, R Pfitzner, I Scholtes, A Garas and F Schweitzer: **Predicting Scientific Success Based on Coauthorship Networks,** February 2014, **http:**//arxiv.org/abs/1402.7268

MS Zanetti, I Scholtes, CJ Tessone and F Schweitzer: **Categorizing Bugs with Social Networks: A Case Study on Four Open Source Software Communities**, In Proceedings of the 35th International Conference on Software Engineering (ICSE 2013), SEIP track, San Francisco, CA, USA, 2013, http://arxiv.org/abs/1302.6764

MS Zanetti, I Scholtes, CJ Tessone and F Schweitzer: **The Rise and Fall of a Central Contributor: Centralization and Performance in the Gentoo Community**, In Proceedings of the 6th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE) held at ICSE 2013, San Francisco, CA, USA, 2013, http://arxiv.org/abs/1302.7191

MS Zanetti, E Sarigol, I Scholtes, CJ Tessone, F Schweitzer: **A Quantitative Study of Social Organization in Open Source Software Communities,** In Proceedings of Imperial College Computing Student Workshop (ICCSW), Open Access Series in Informatics, pp. 116-122, London, UK, September 27-28, 2012, http://arxiv.org/abs/1208.4289

I Scholtes, N Wider, R Pfitzner, A Garas, CJ Tessone, F Schweitzer: **Slow-down vs. Speed-up of Information Diffusion in Non-Markovian Temporal Networks,** arXiv:1307.4030, July 15 2013, http://arxiv.org/abs/1307.4030

R Pfitzner, I Scholtes, A Garas, CJ Tessone, F Schweitzer: **Betweenness Preference: Quantifying Correlations in the Topological Dynamics of Temporal Networks,** Physical Review Letters, Vol. 110, 198701, May 10 2013, http://journals.aps.org/prl/abstract/10.1103/PhysRevLett.110.198701

http://www.sg.ethz.ch

+Chair of Systems Design

@ingo_s

ischoltes@ethz.ch