KONECT
The Koblenz Network Collection

Towards a Broad Analysis of Complex Systems

Jérôme Kunegis

based on work with Julia Perl, Christoph Carl Kling, Daniel Dünker, Eiko Yoneki, Valentin Dalibard and Jesús Cabello González

KoMePol

WeST
People and Knowledge Networks

Reveal

Institute for Web Science and Technologies · University of Koblenz-Landau, Germany

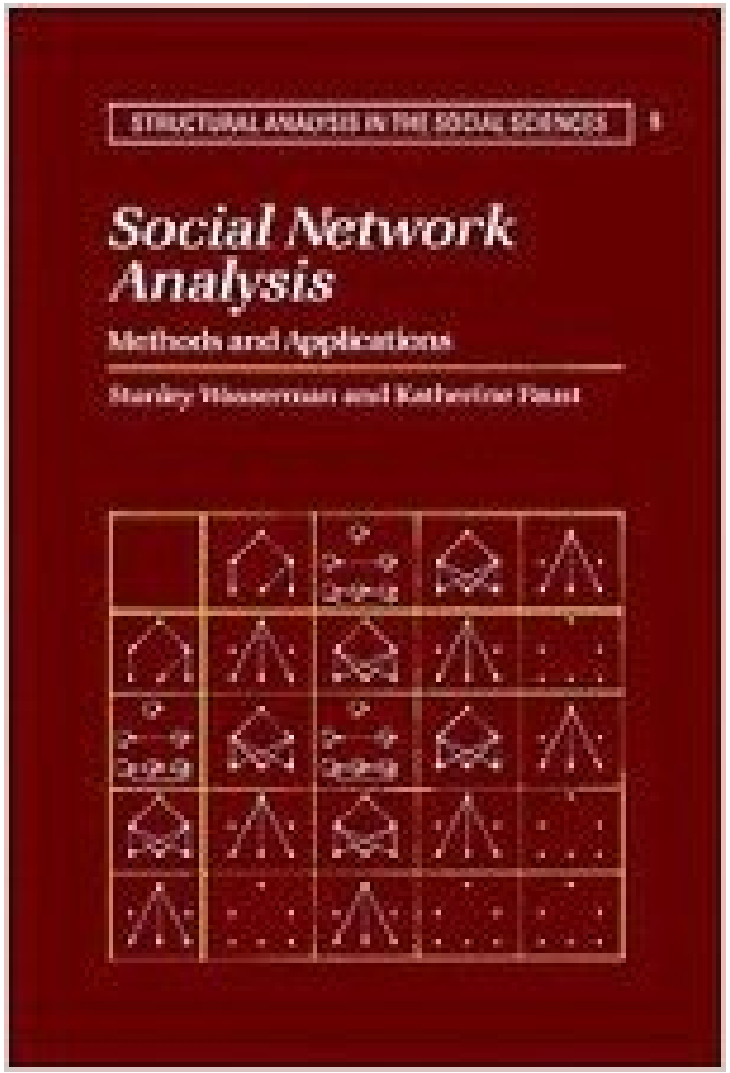
Table B.3. *"Reports to" relation between managers of Krackhardt's high-tech company*

Manager																					
1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
4	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
9	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
13	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
16	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
18	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Social Network Analysis without the Web

"Social Network Analysis" by
Wasserman & Faust first edition 1994

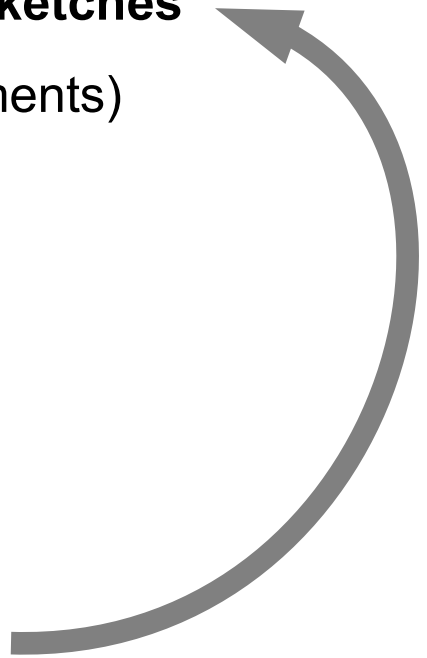
Contains 18 datasets (based on 19th printing)



WWW 2014 Best Paper Nominations

- **Community-Based Bayesian Aggregation Models for Crowdsourcing**
  - **4** datasets (crowdsourcing)
- **Efficient Estimation for High Similarities using Odd Sketches**
  - **5** real-world datasets + synthetic dataset (text documents)
- **Local Collaborative Ranking**
  - **3** datasets (rating networks)
- **Engaging with Massive Online Courses**
  - **1** dataset (case study)

The best paper award goes to…

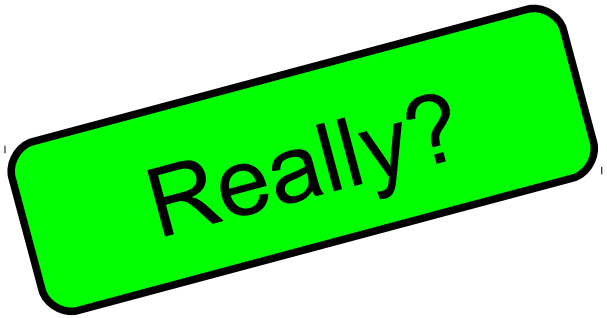


Why Do Researchers Use Multiple Datasets?

- To cover more application areas
- To show that results are generalizable
- To make results more statistically significant

Showing that Algorithm X is Better Than Y

- Setup: We want to compare two prediction algorithms X and Y
- Experiment: Apply X and Y to dataset A
- Result: X has higher precision than Y
- Conclusion: "Algorithm X performs better than algorithm Y"



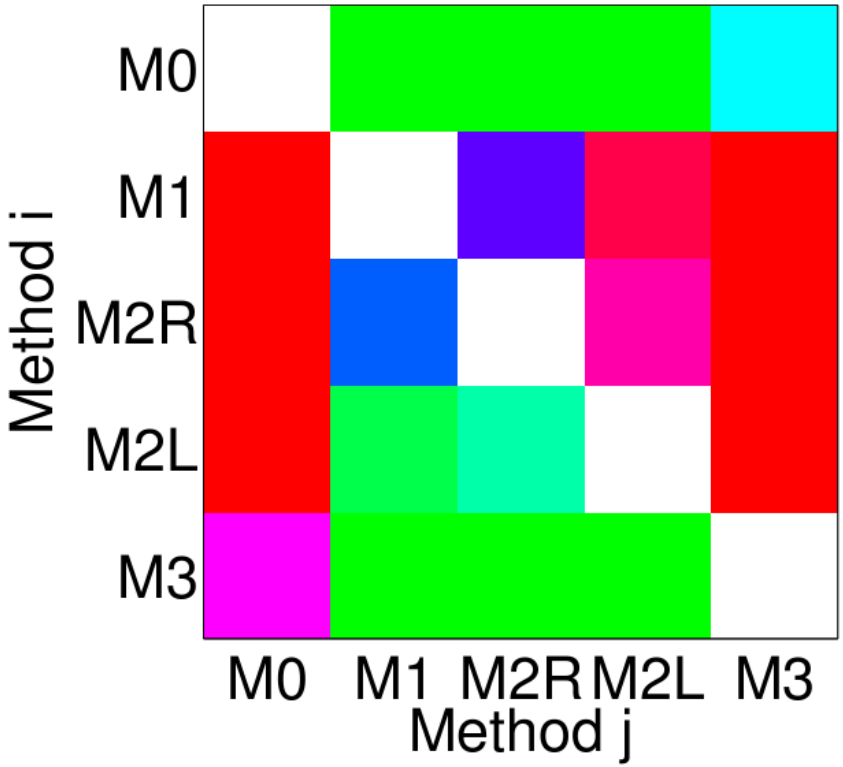
Let's Make More Experiments

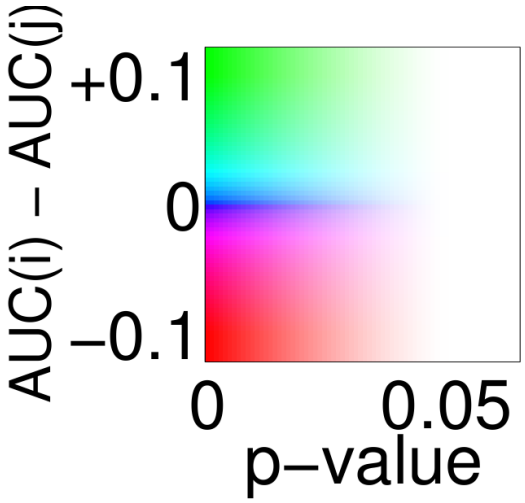
- Add a dataset B to the experiments
- On dataset B, Y performs better than X

More Datasets

- Add more datasets to the mix
- Algorithm X is better than algorithm Y with 6 out of 10 datasets
- Under Null hypothesis of equal probability of X performing better than Y on any one dataset and results on datasets being independent, the probability that this happens is 17%, i.e., not significant!
- Need about 65 datasets to get a statistically significant result at $p \leq 0.05$ for a 60% result.

Example: Link Prediction





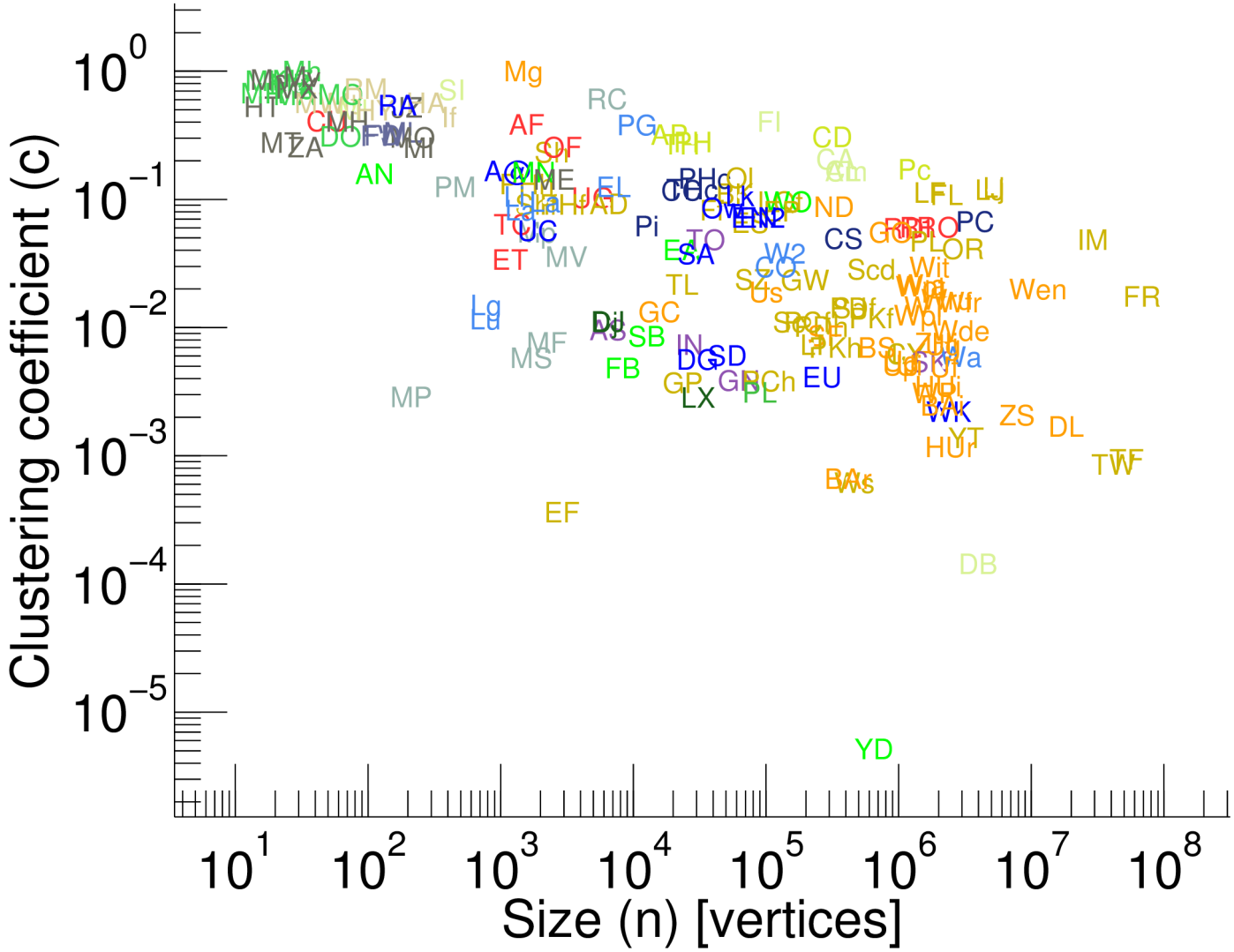
Measuring Properties of Networks

- Everyone knows: networks have high clustering coefficient

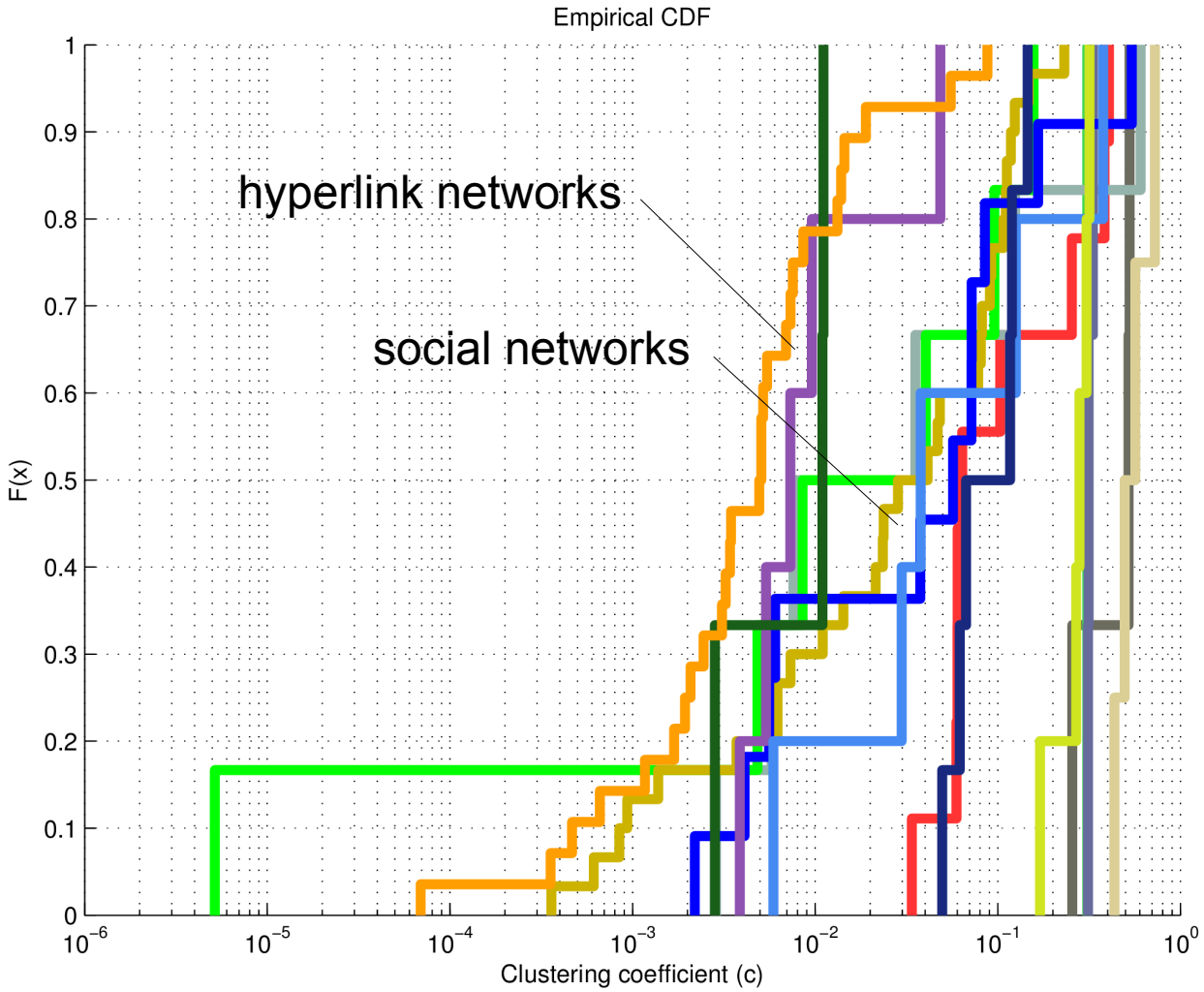
- What is the typical clustering coefficient of a social network?
- What is the typical clustering coefficient of a hyperlink network?

- How can one answer these types of questions?

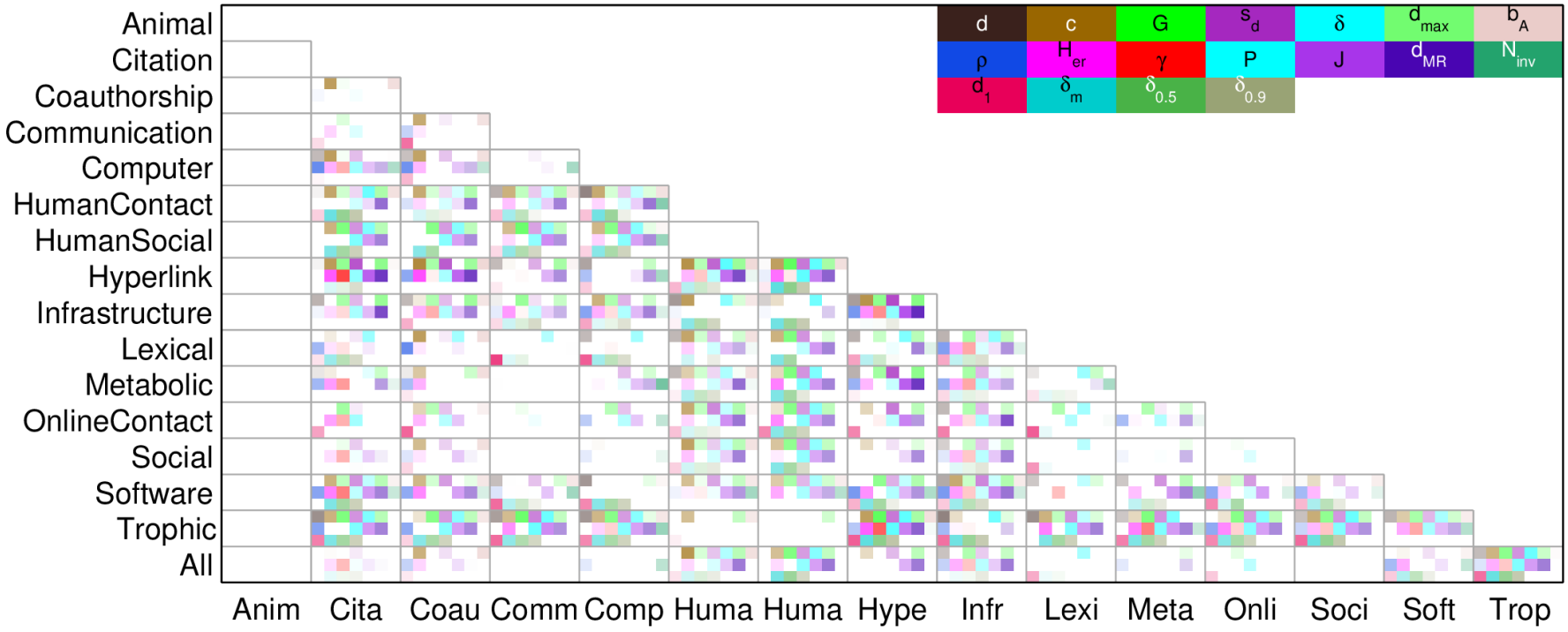
Clustering Coefficient



Can the Category of a Network Be Predicted?

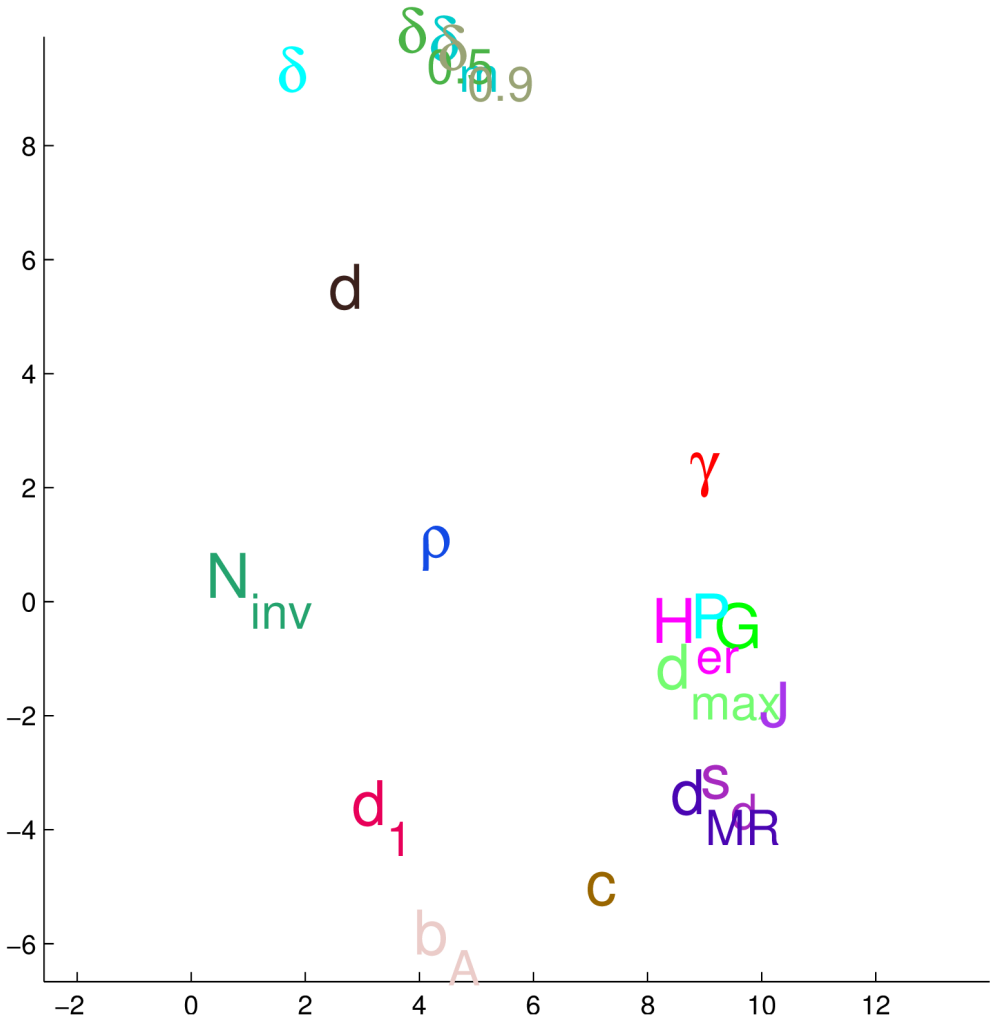


Predicting the Type of a Network



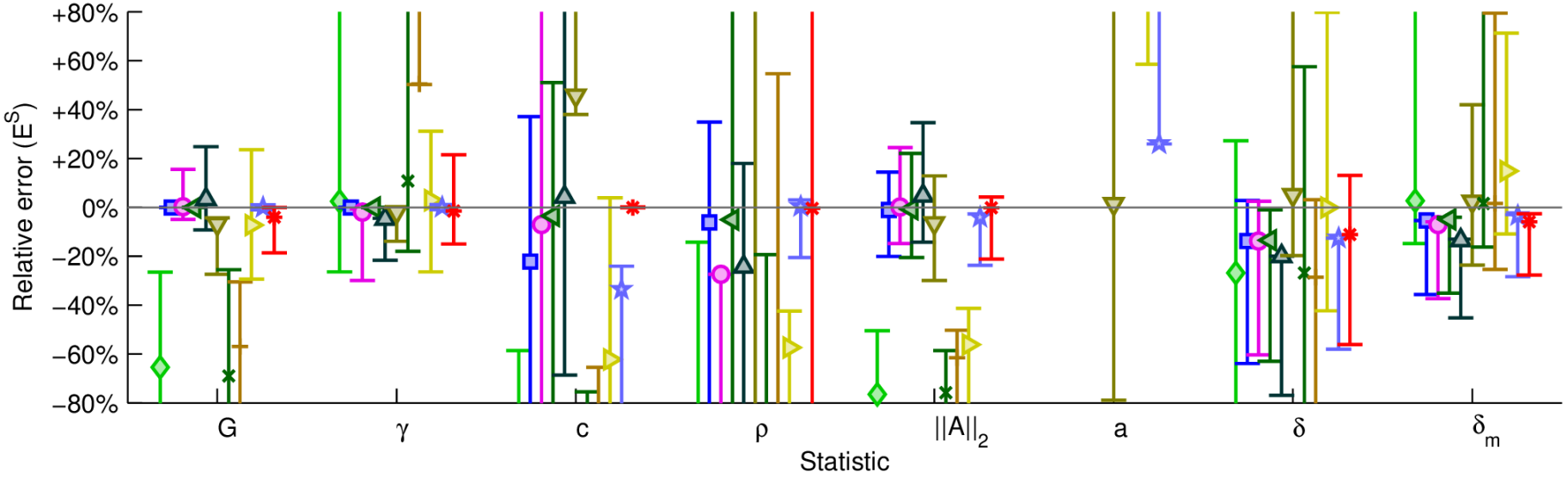
Cell shown as non-white when the given numerical graph statistic has values in the two given groups of graphs that significantly differ according to Kolmogorov–Smirnov test at p = 0.05. (Note: assumes normality of graph statistic per graph category)

Clustering of Network Statistics (PCA)



Verify Graph Models

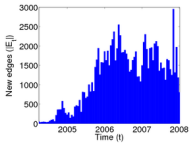
- Experiment: Generate graphs with properties matching those of given graphs
- Evaluation: Average relative error on 36 datasets

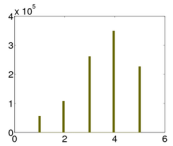


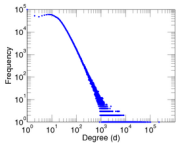
KONECT.UNI-KOBLENZ.DE

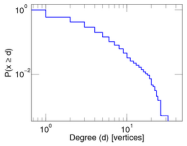
Code	Name ▲	Category	F.	W.	M.		n	m	c
CL	Actor collaborations	Misc	U	=			382,219	33,115,812	16.6%
ME	Adolescent health	HumanSocial	D	+			2,539	12,969	14.2%
AD	Advogato	Social	D	+	◒	123 abc	6,541	51,127	9.22%
TC	Air traffic control	Infrastructure	D	−	◒		1,226	2,615	6.39%
CA	Amazon (MDS)	Misc	U	−			334,863	925,872	20.5%
Am	Amazon (TWEB)	Misc	D	−			403,394	3,387,388	16.6%
AP	arXiv astro-ph	Coauthorship	U	−			18,771	198,050	31.8%
PH	arXiv hep-ph	Coauthorship	U	−		🕒	28,093	4,596,803	28.0%
PHc	arXiv hep-ph	Citation	D	−	◒		34,546	421,578	14.6%
TH	arXiv hep-th	Coauthorship	U	−		🕒	22,908	2,673,133	26.9%
THc	arXiv hep-th	Citation	D	−	◒		27,770	352,807	12.0%
BAi	Baidu internal	Hyperlink	D	=	◒		2,141,300	17,794,839	0.245%
BAr	Baidu related	Hyperlink	D	=	◒		415,641	3,284,387	0.0663%
BS	Berkeley/Stanford	Hyperlink	D	−			685,230	7,600,595	0.694%
MB	Bison	Animal	D	+			26	314	78.9%
Mg	Blogs	Hyperlink	D	−		123 abc	1,490	2,220,035	100%
BK	Brightkite	Social	U	−			58,228	214,078	11.1%
PM	Caenorhabditis elegans	Metabolic	U	=	◒		453	4,596	12.4%
IN	CAIDA	Computer	U	−		🔗	26,475	53,381	0.732%

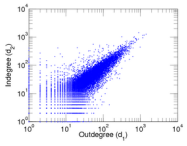
Square count	$q = \lvert\{u,v,w,x \mid u \sim v \sim w \sim x \sim u\}\rvert/8$
4-tour count	$T_4 = 8q + 4s + 2m$
Power law exponent	$\gamma = 1 + n\left(\sum_{u \in V} \ln \frac{d(u)}{d_{\min}}\right)^{-1}$
Gini coefficient	$G = \frac{2\sum_{i=1}^{n} i d_i}{n \sum_{i-1}^{n} d_i} - \frac{n+1}{n}$
Relative edge distribution entropy	$H_{\text{er}} = \frac{1}{\ln \lvert V \rvert} \sum_{u \in V} -\frac{d(u)}{D} \ln \frac{d(u)}{D}$

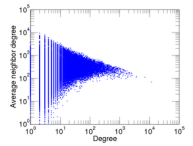

Temporal distribution

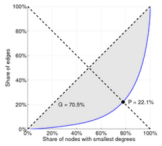

Edge weight and multiplicity distribution

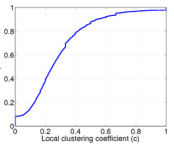

Degree distribution

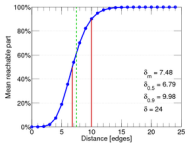

Cumulative degree distribution

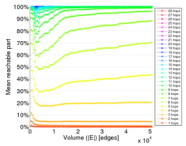

Out/indegree comparison

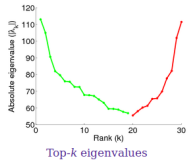

Assortativity plot

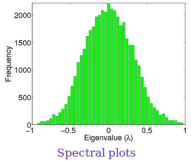

Lorenz curve

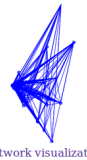

Clustering coefficient distribution

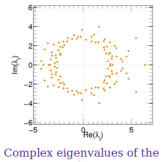

Distance distribution


Temporal distance distribution

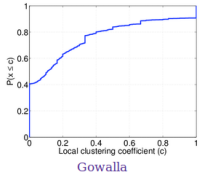

Top-k eigenvalues

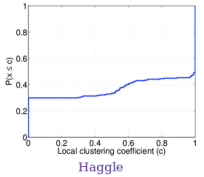

Spectral plots

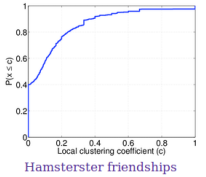

Network visualization

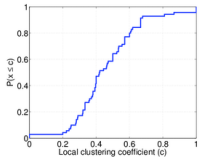

Complex eigenvalues of the asymmetric adjacency matrix

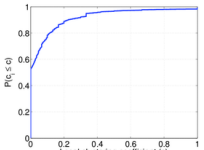

Layout

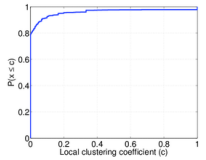

Gowalla

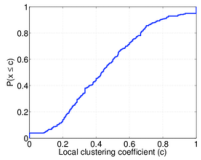

Haggle

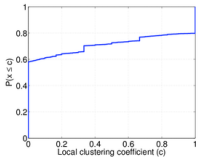

Hamsterster friendships

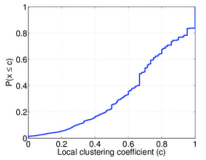

Highschool


Hudong internal links


Human protein (Figeys)


Infectious


Internet topology


JDK dependency

KONECT.UNI-KOBLENZ.DE

- 239 network datasets as of June 2015
- undirected / directed / bipartite, unweighted / multiple edges / signed / ratings / etc., loops, timestamps
- 32 categories: social, rating, text, contact, lexical, interaction, infrastructure, hyperlink, computer, citation, authorship, animal, etc.

- MATLAB Toolbox **konect.uni-koblenz.de/toolbox**

Thank you

Network dataset donations accepted at
KONECT.UNI-KOBLENZ.DE
kunegis@uni-koblenz.de

Jérôme Kunegis

WeST
People and Knowledge Networks