Part II
Web Science in Practice: Web Observatories

Jérôme Kunegis & Steffen Staab

WSTNet Web Science Summer School 2016

WeST
People and Knowledge Networks

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

Case Study

- Website: **http://konect.uni-koblenz.de/**
- Handbook: **http://konect.uni-koblenz.de/downloads/konect-handbook.pdf**
- Programming language: Octave / Matlab

WeST
People and Knowledge Networks

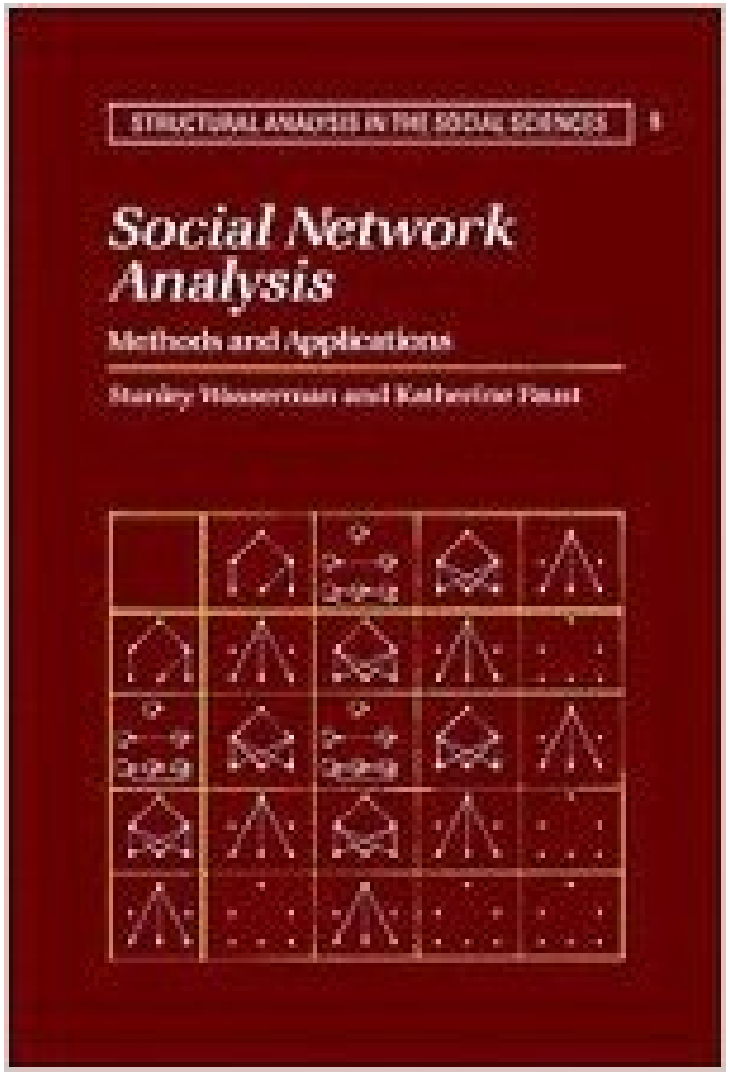
Table B.3. *"Reports to" relation between managers of Krackhardt's high-tech company*

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 1 0 0 0
11	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 1
13	0 0 0 0 0 0 0 0 0 0 0 0 1 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
15	0 0 0 0 0 0 0 0 0 0 0 0 1 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
17	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
19	0 0 0 0 0 0 0 0 0 0 0 0 1 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
21	0 0 0 0 0 0 1 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)

WWW 2014 Best Paper Nominations

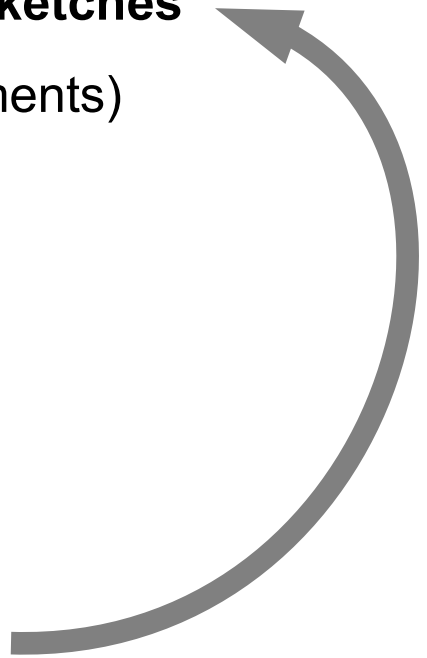
- **Community-Based Bayesian Aggregation Models for Crowdsourcing**
  - **4** datasets (crowdsourcing)
- **Efficient Estimation for High Similarities using Odd Sketches**
  - **5** real-world datasets + synthetic dataset (text documents)
- **Local Collaborative Ranking**
  - **3** datasets (rating networks)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

The best paper award goes to…

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

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

Showing that Algorithm X is Better Than Y

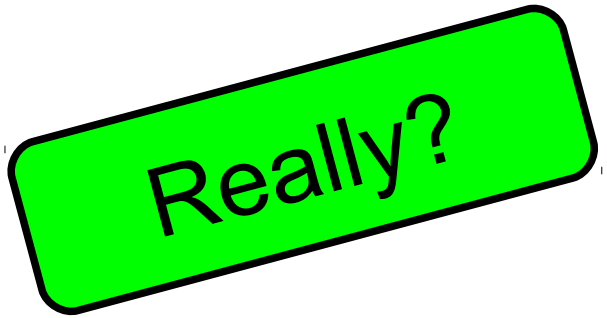
- Setup: We want to compare two prediction algorithms X and Y
- Experiment: Apply X and Y to dataset A
- Result: X has higher precision than Y

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"

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

Let's Make More Experiments

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

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

More Datasets

- Add more datasets to the mix
- Algorithm X is better than algorithm Y with 6 out of 10 datasets

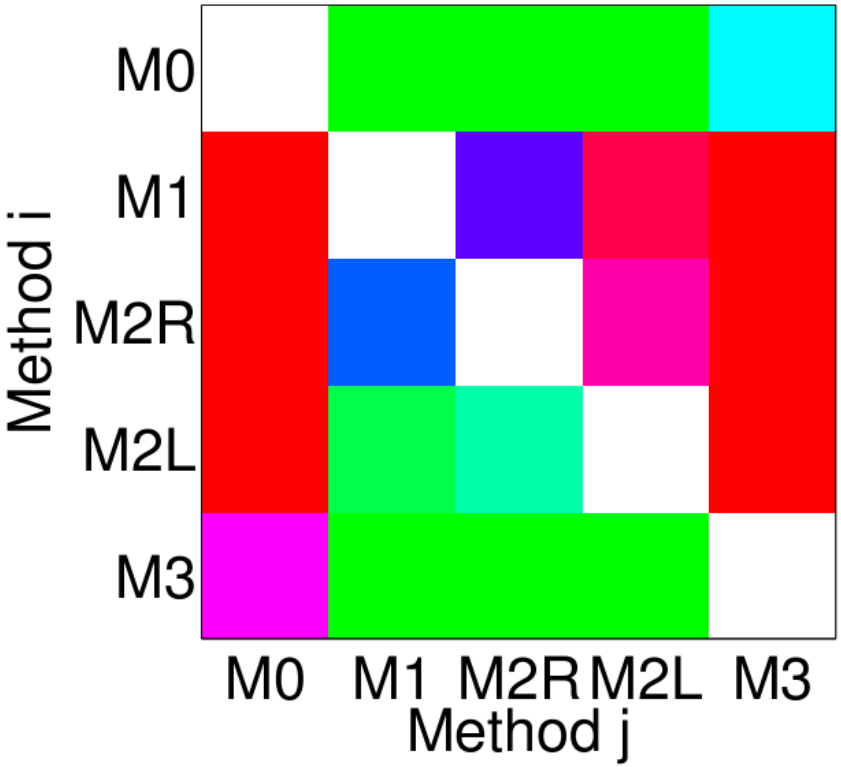
More Datasets

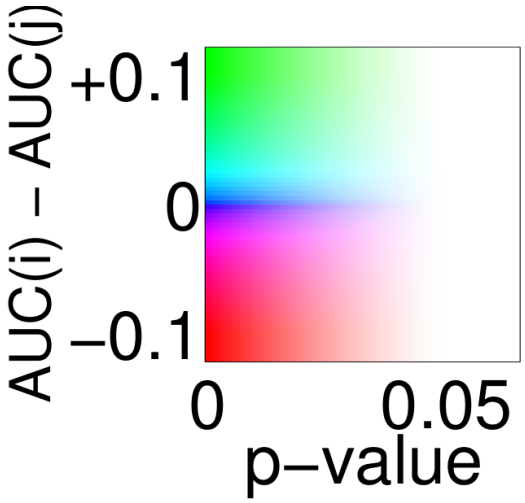
- Add more datasets to the mix
- Algorithm X is better than algorithm Y with 6 out of 10 datasets
- Under Null hypothesis of equal probability of X performing better than Y on any one dataset and results on datasets being independent, the probability that this happens is 17%, i.e., not significant!

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





Measurements in Datasets

- Everyone knows: networks have high clustering coefficient

Measurements in Datasets

- 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?

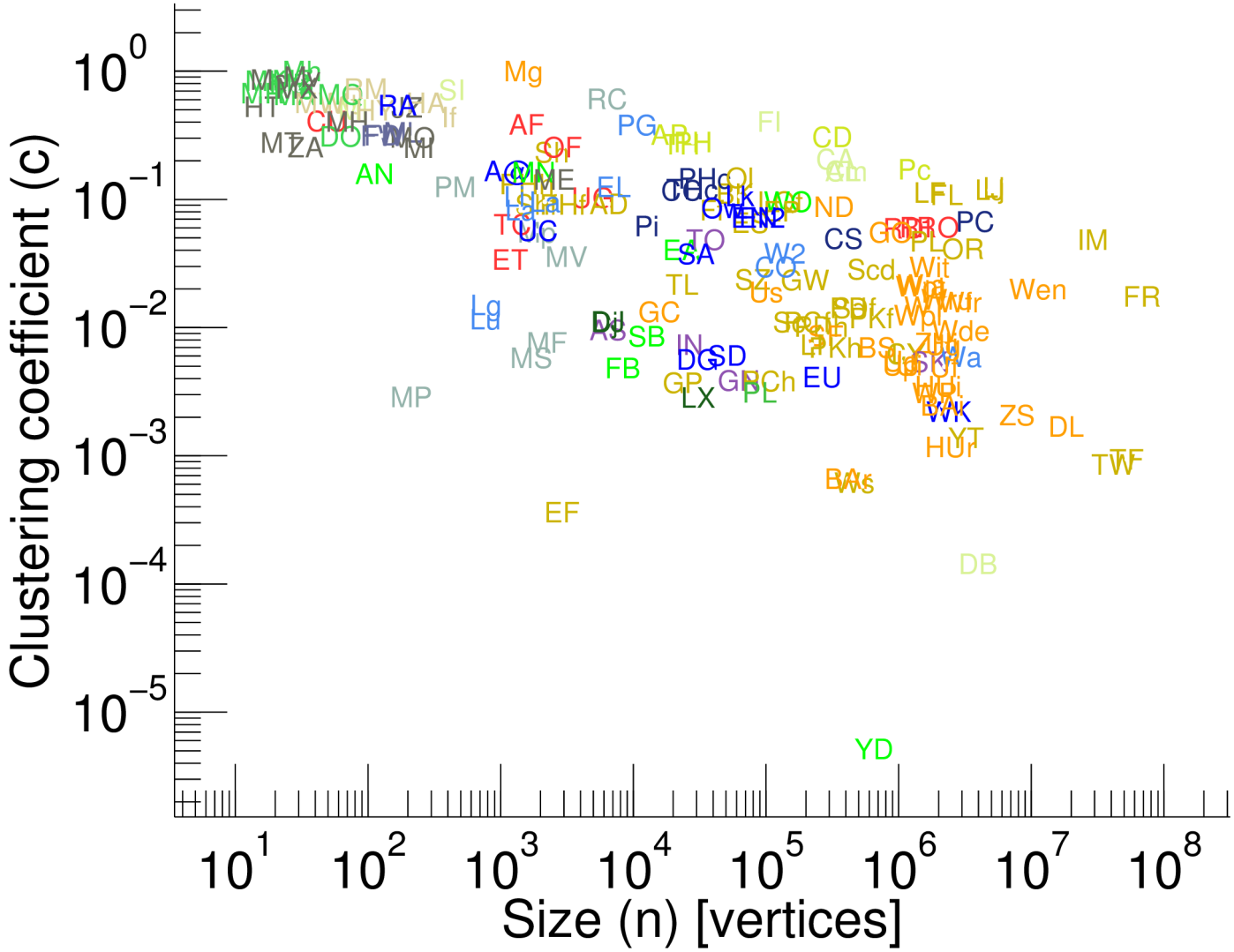
Measurements in Datasets

- Everyone knows: networks have high clustering coefficient

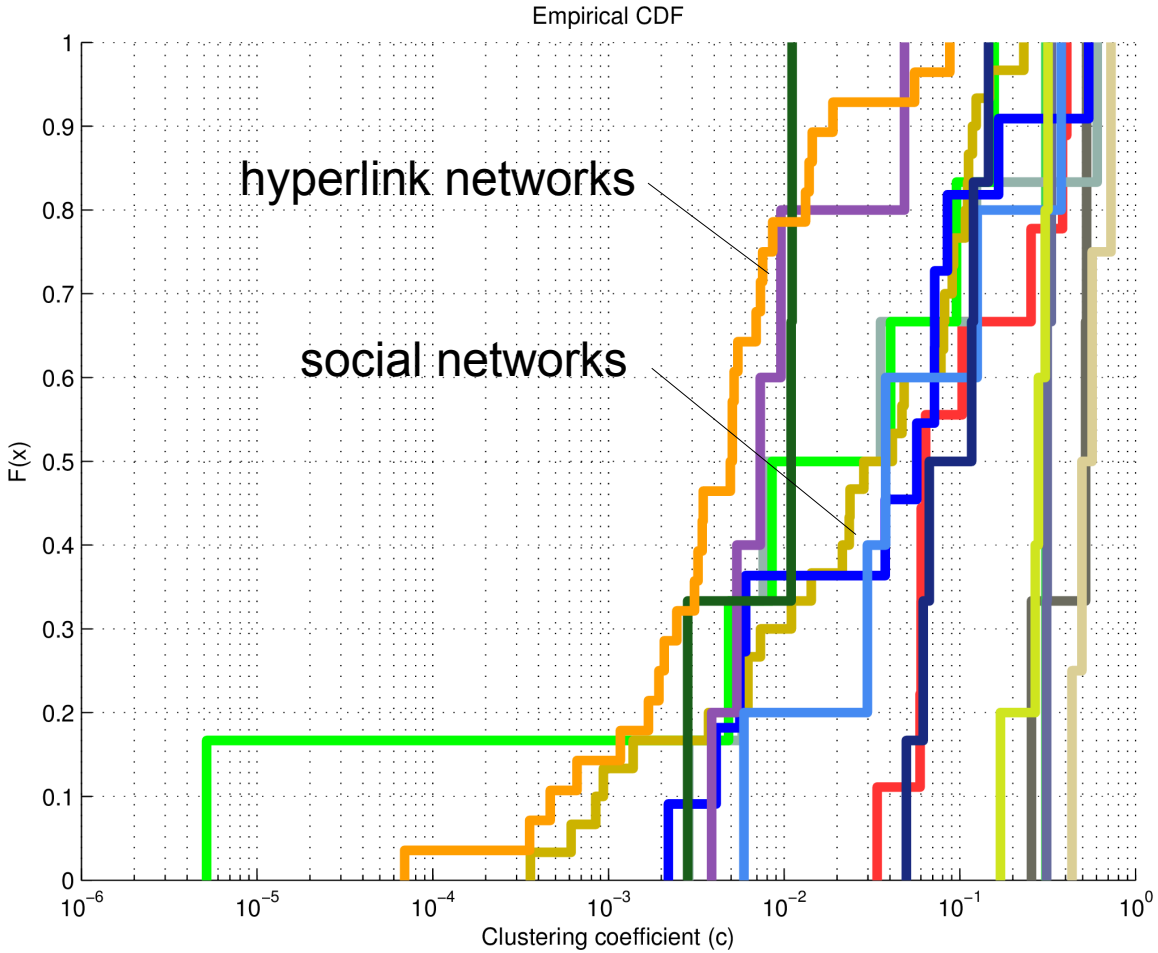
- What is the typical clustering coefficient of a social network?
- What is the typical clustering coefficient of a hyperlink network?

- How can one answer these types of questions?

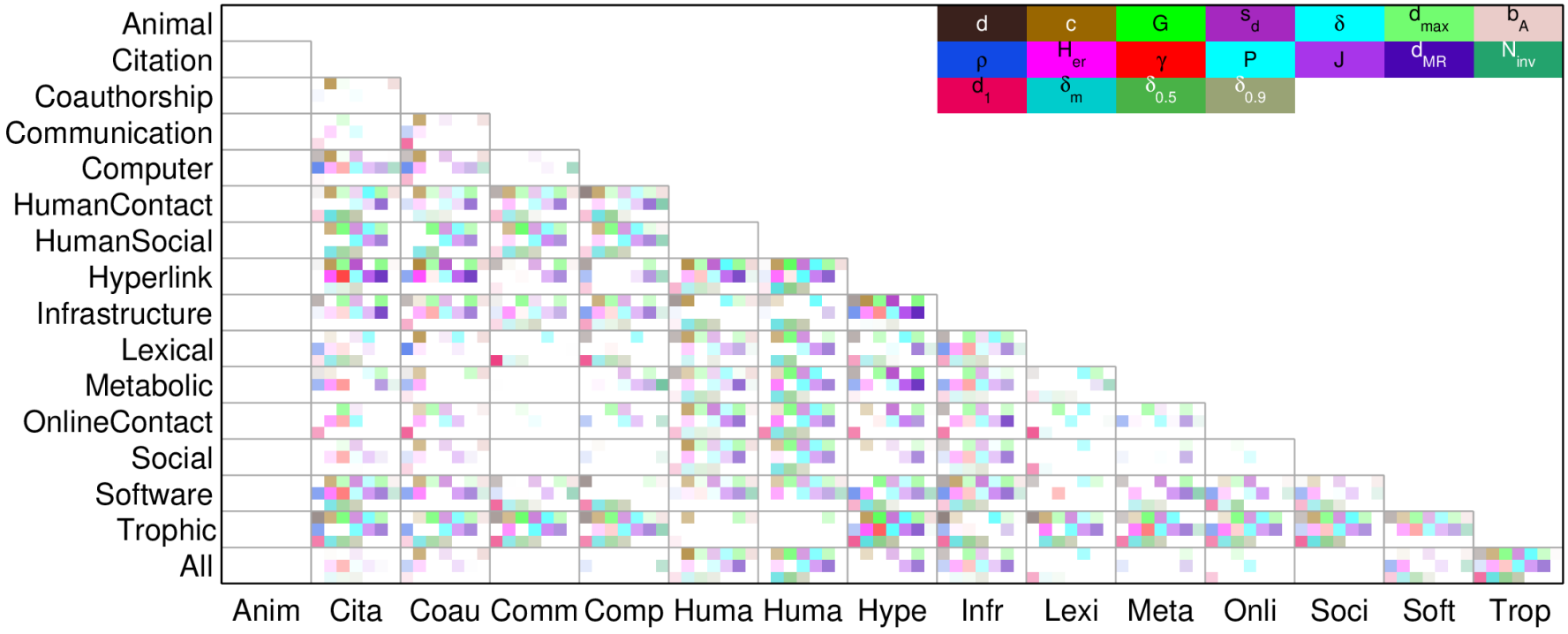
Clustering Coefficient



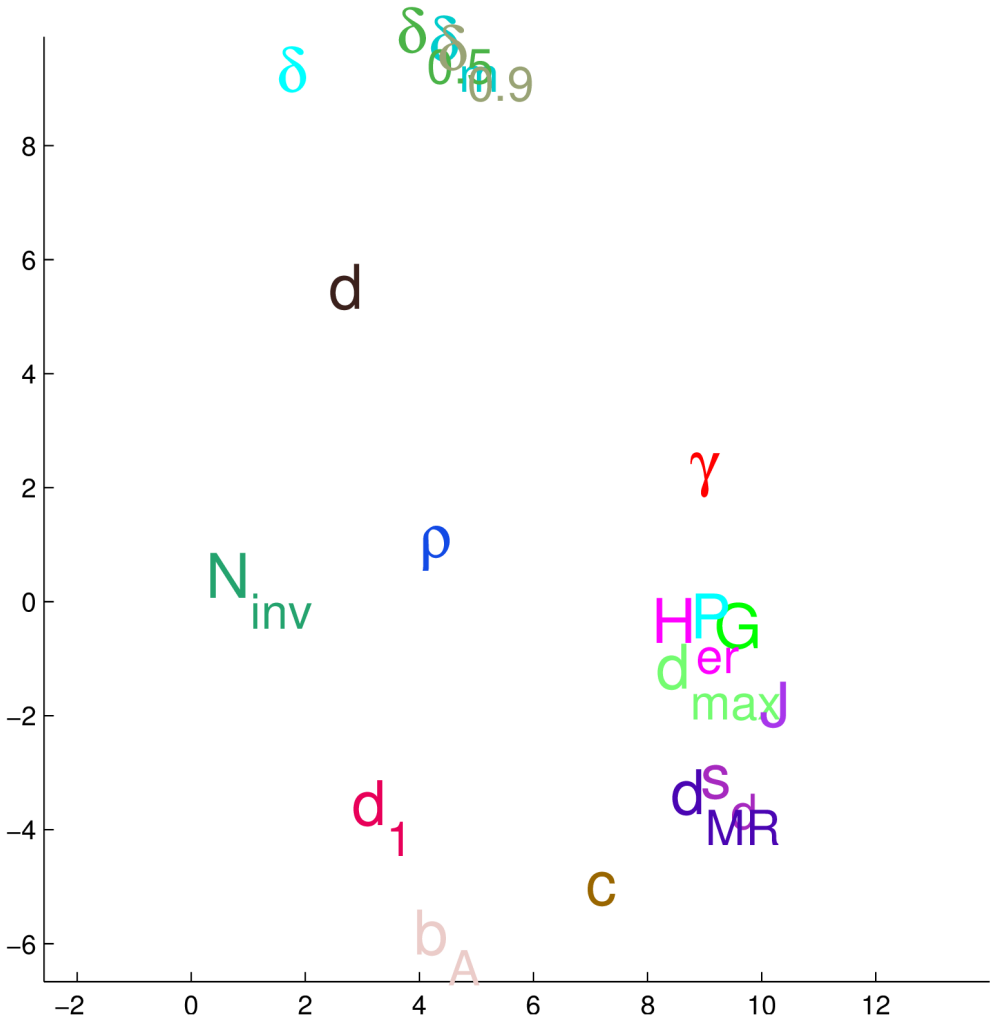
Can the Category of a Network Be Predicted?



Predicting the Type of a Network

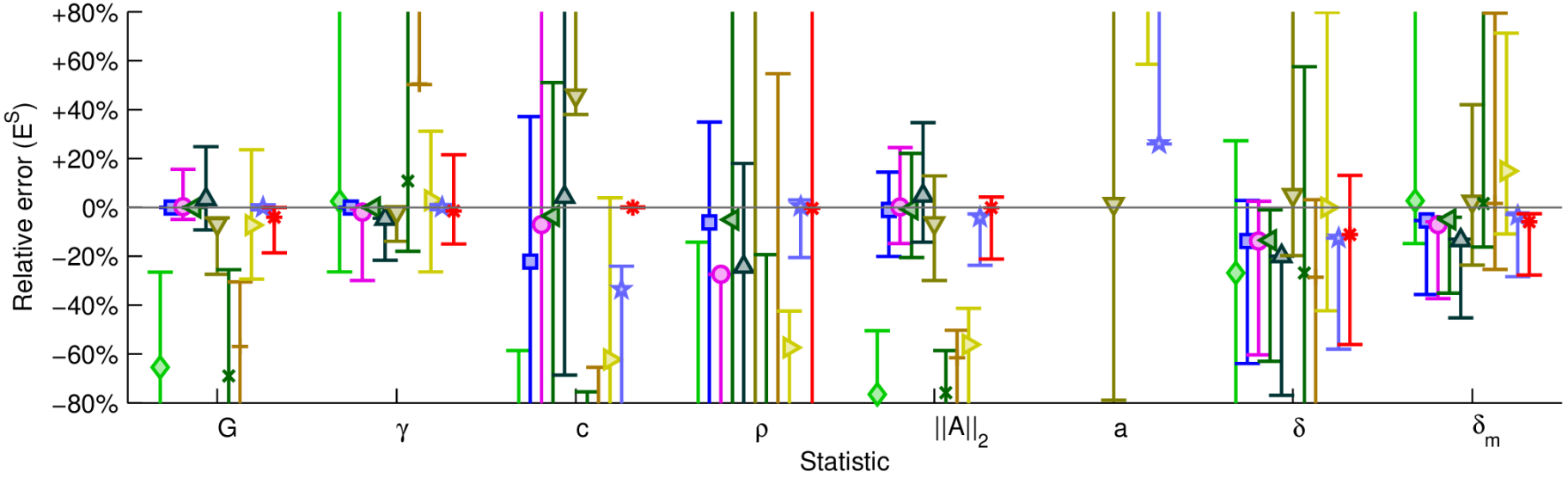


Clustering of Network Statistics (PCA)



Verify Graph Models

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

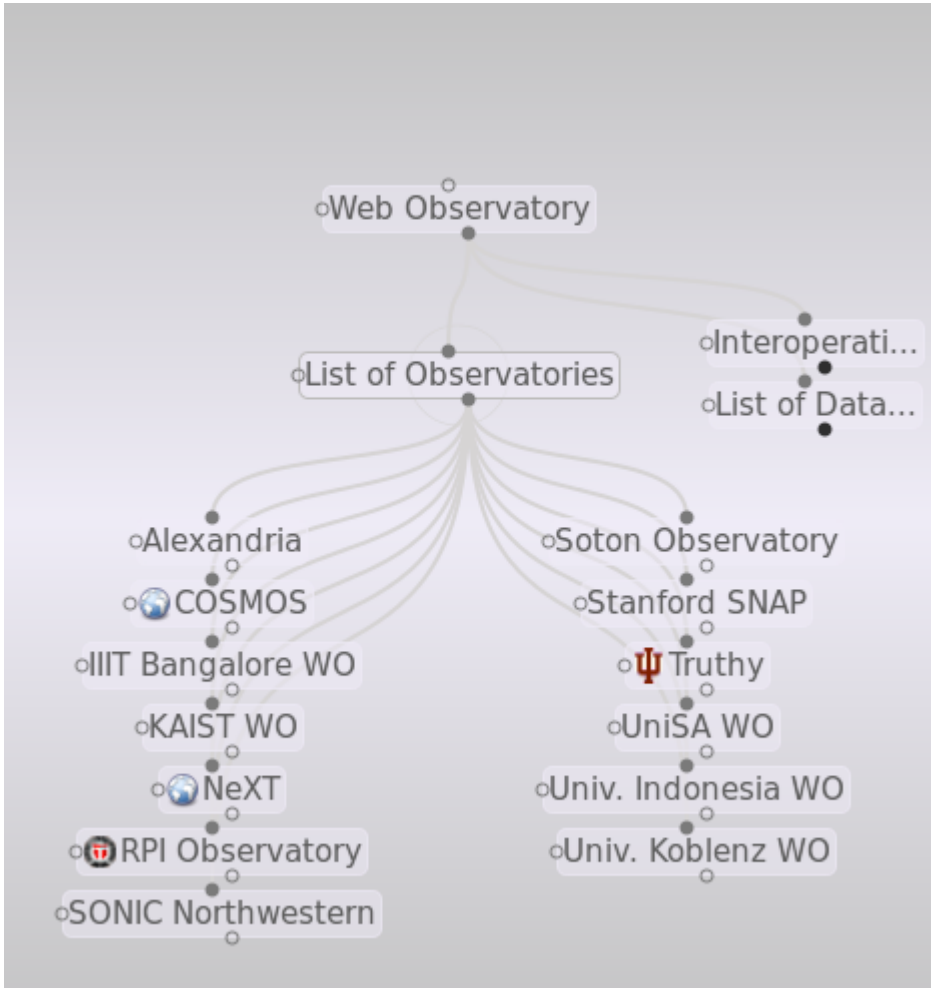


Web Observatories

"A Web Observatory is a system which gathers and links to data on the Web in order to answer questions about the Web, the users of the Web and the way that each affects the other."

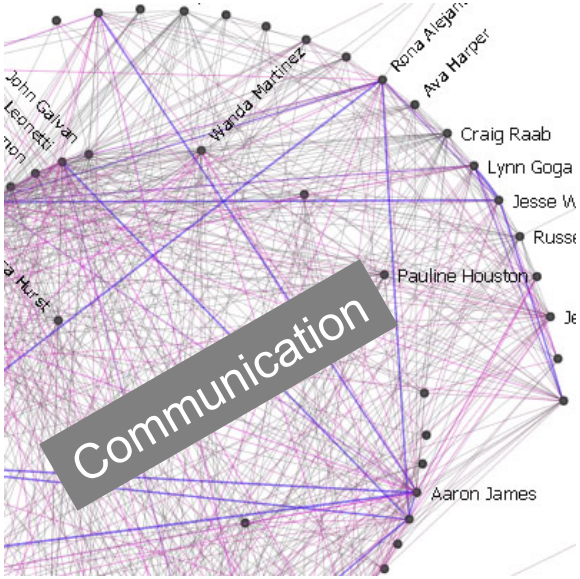
http://www.webscience.org/web-observatory/

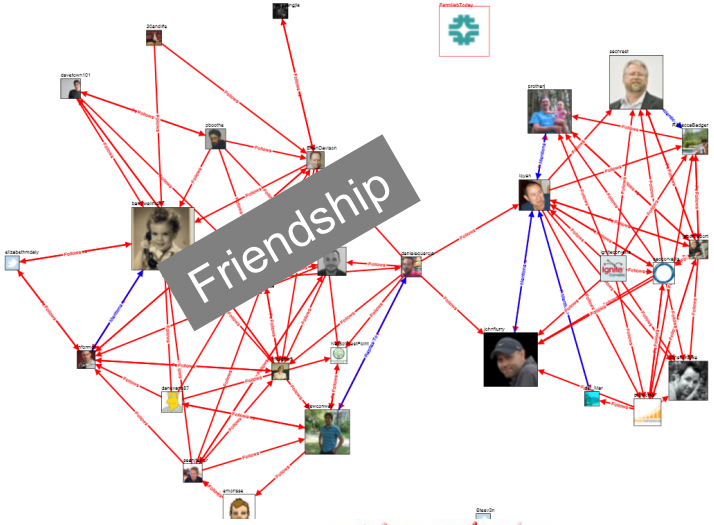
Web Observatories

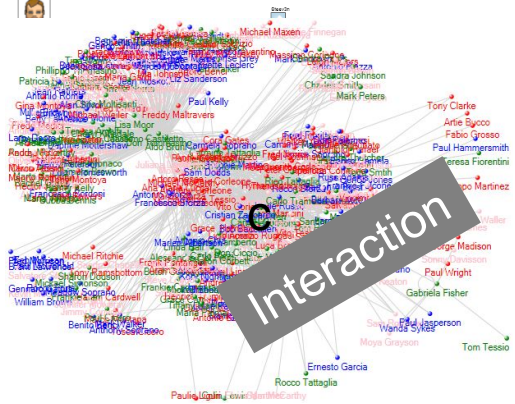


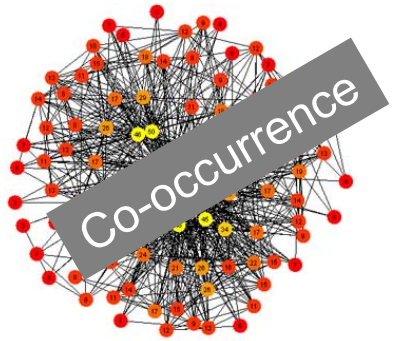
http://www.webscience.org/web-observatory/

“Everything Is a Network”









Web Observatories for Networks

Stanford Network Analysis Project

Index of Complex Networks

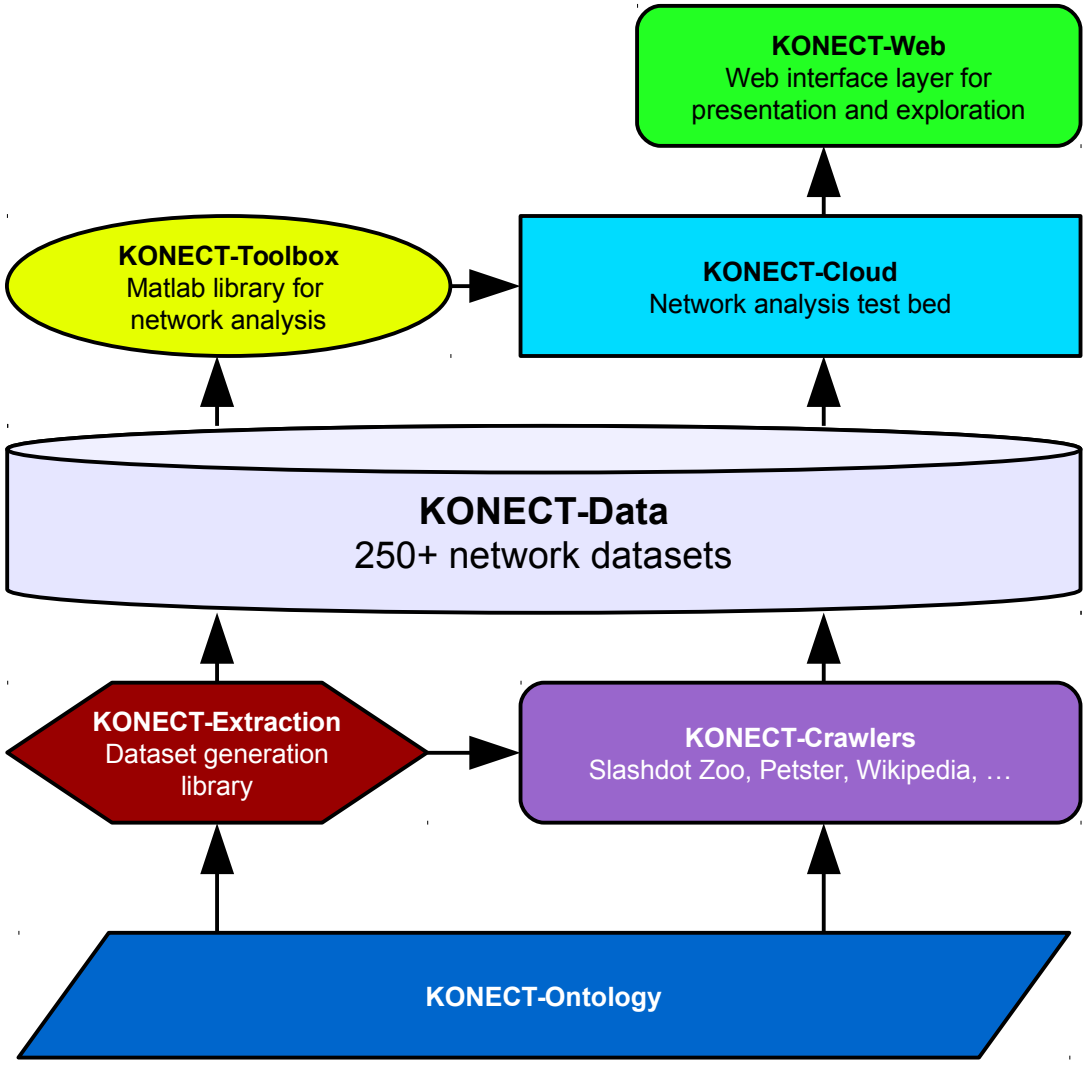
Index of Complex Networks, U. of Colorado Boulder

KONECT

Koblenz Network Collection

…And many smaller ones

KONECT – Typical “Observatory” Setup



Why an Ontology?

- Unipartite vs Bipartite ? Directed vs Undirected ?
- Weighted vs Unweighted vs Signed vs Ratings ?
- Multiple edges ? Loops ?
- Timestamps?
- Disappearing edges ?
- Node metadata ?
- Node types ? Edge types ?
- License ?
- Size ? Structural properties ?

Why an Ontology?

- Unipartite vs Bipartite ? Directed vs Undirected ?
- Weighted vs Unweighted vs Signed vs Ratings ?
- Multiple edges ? Loops ?
- Timestamps?
- Disappearing edges ?
- Node metadata ?
- Node types ? Edge types ?
- License ?
- Size ? Structural properties ?

Dataset Format

Table 1: The network formats allowed in KONECT. Each network dataset is exactly of one type.

#	Symbol	Type	Edge partition	Edge types	Internal name
1	U	Undirected	Unipartite	Undirected	`sym`
2	D	Directed	Unipartite	Directed	`asym`
3	B	Bipartite	Bipartite	Undirected	`bip`

Kinds of Relationship

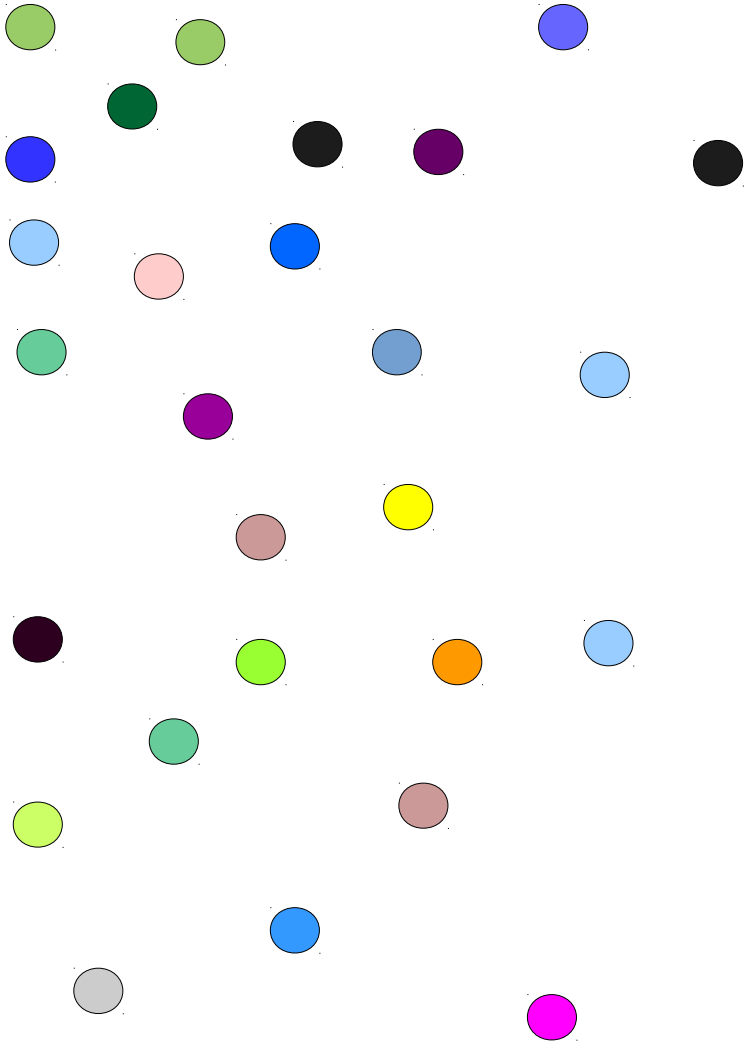
Table 2: The edge weight and multiplicity types allowed in KONECT. Each network dataset is exactly of one type. Note that due to historical reasons, networks with multiple unweighted edges have the internal name `positive`, while positively weighted networks have the internal `posweighted`. For signed networks and positive edge weights, weights of zero are only allowed when the tag `#zeroweight` is set.

#	Symbol	Type	Multiple edges	Edge weight range	Edge weight scale	Internal name
1	$-$	Unweighted	No	$\{1\}$	–	`unweighted`
2	$=$	Multiple unweighted	Yes	$\{1\}$	–	`positive`
3	$+$	Positive weights	No	$(0, \infty)$	Ratio scale	`posweighted`
4	$\pm$	Signed	No	$(-\infty, +\infty)$	Ratio scale	`signed`
5	$\overset{+}{=}$	Multiple signed	Yes	$(-\infty, +\infty)$	Ratio scale	`multisigned`
6	$*$	Rating	No	$(-\infty, +\infty)$	Interval scale	`weighted`
7	${}_*^*$	Multiple ratings	Yes	$(-\infty, +\infty)$	Interval scale	`multiweighted`
8	$\rightleftarrows$	Dynamic	Yes	$\{1\}$	–	`dynamic`
9		Multiple positive weights	Yes	$(0, \infty)$	Ratio scale	`multiposweighted`

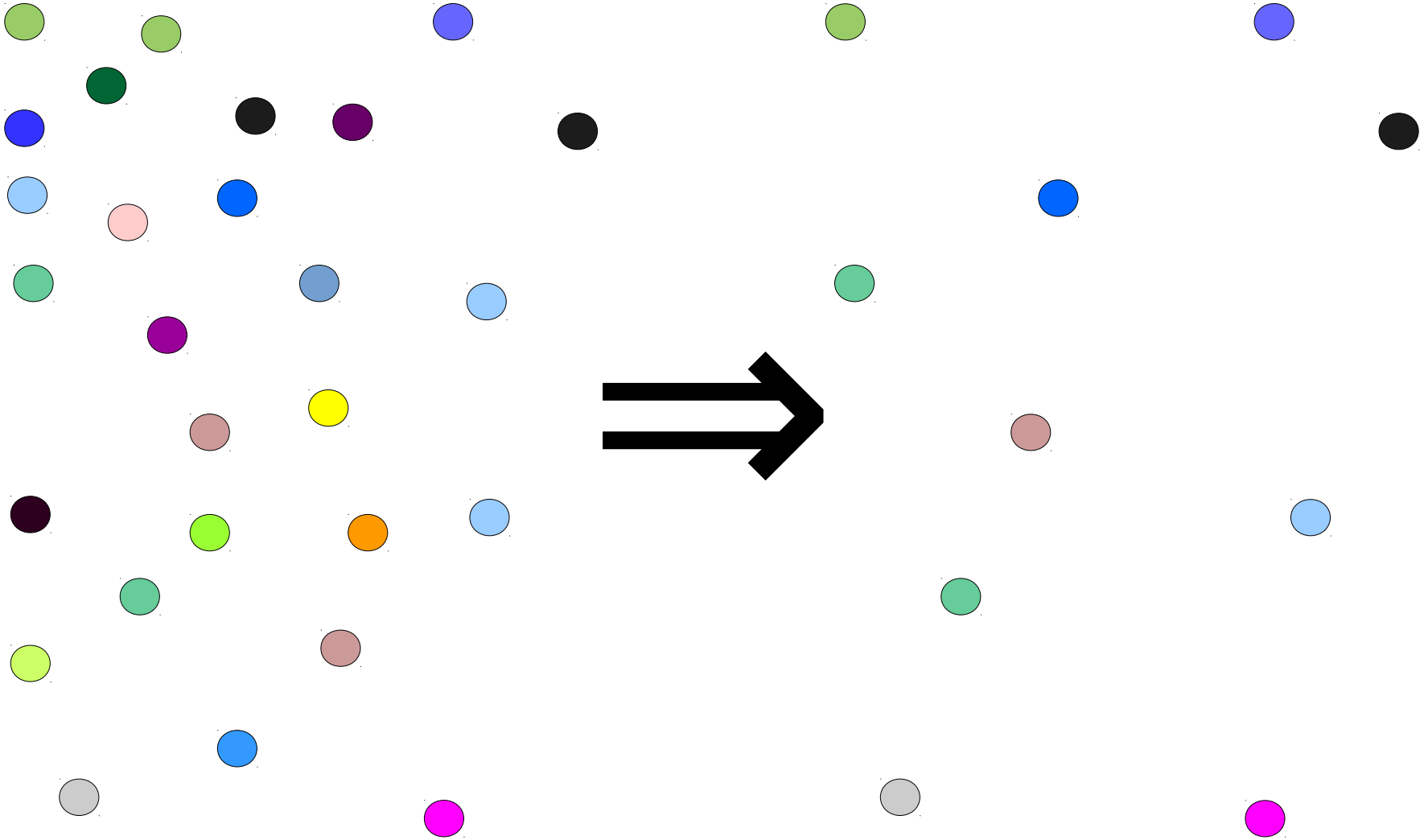
Data Prevenance and Integrity

- Node sampling
- Edge sampling
- *k*-core
- Missing edge orientations
- Reduction to the largest connected component
- Timestamps are rounded / binned
- etc.

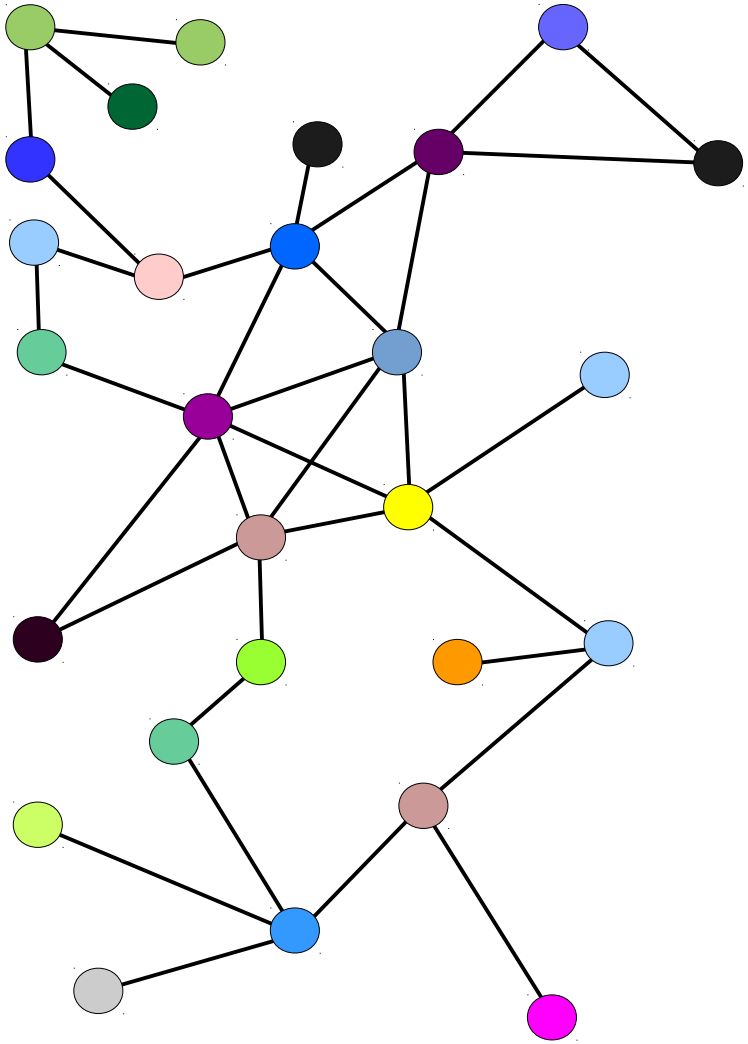
Data Integrity: Sampling



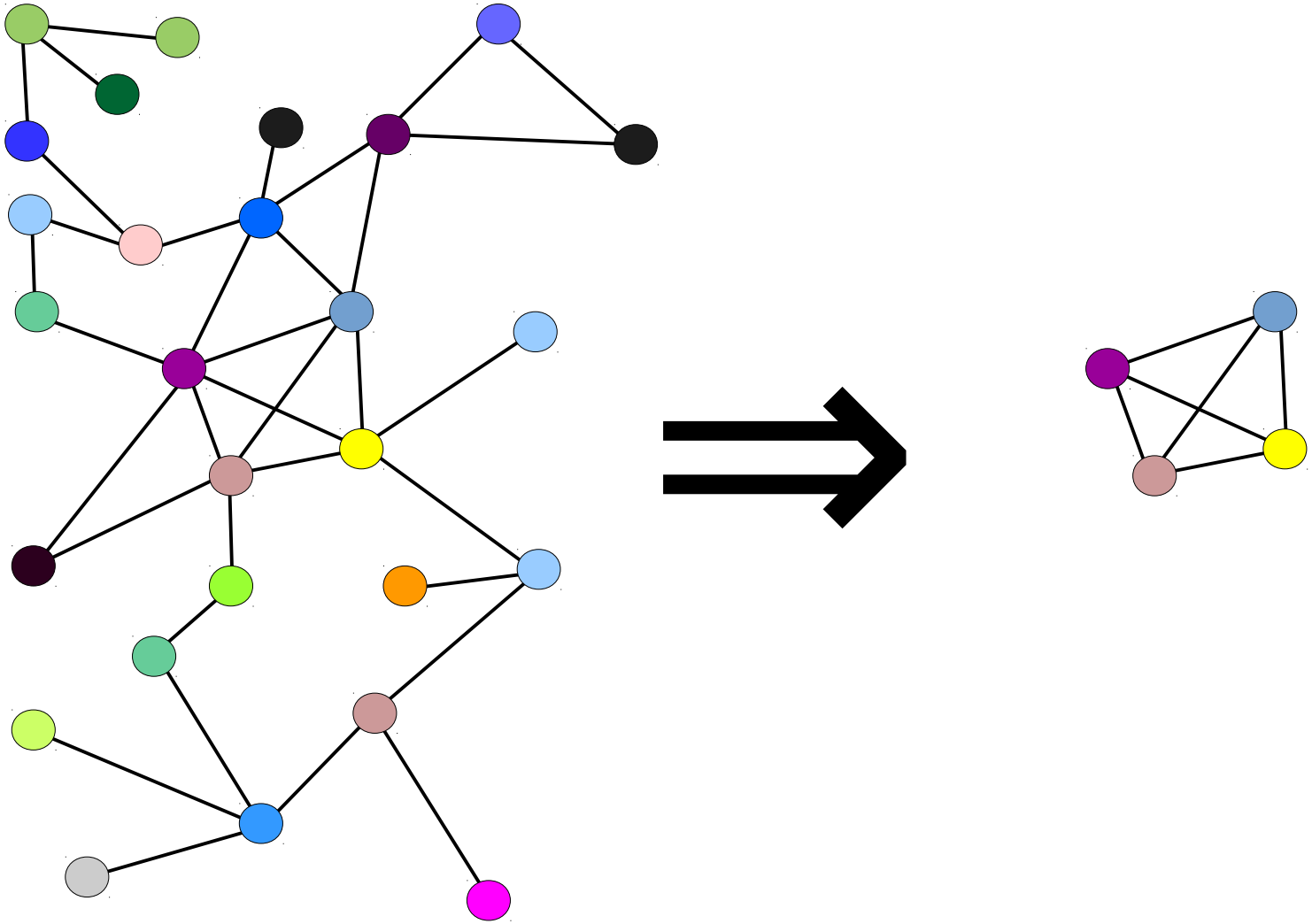
Data Integrity: Sampling



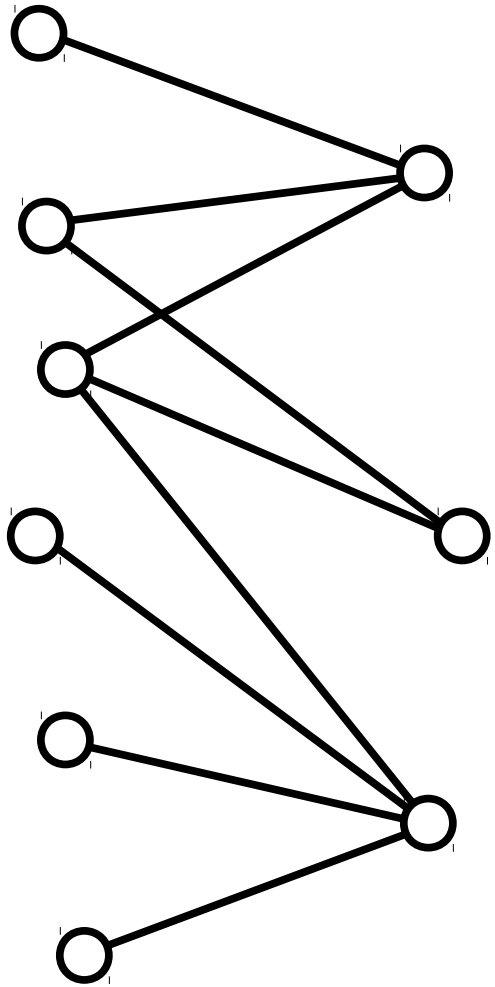
Data Integrity: *k*-Core



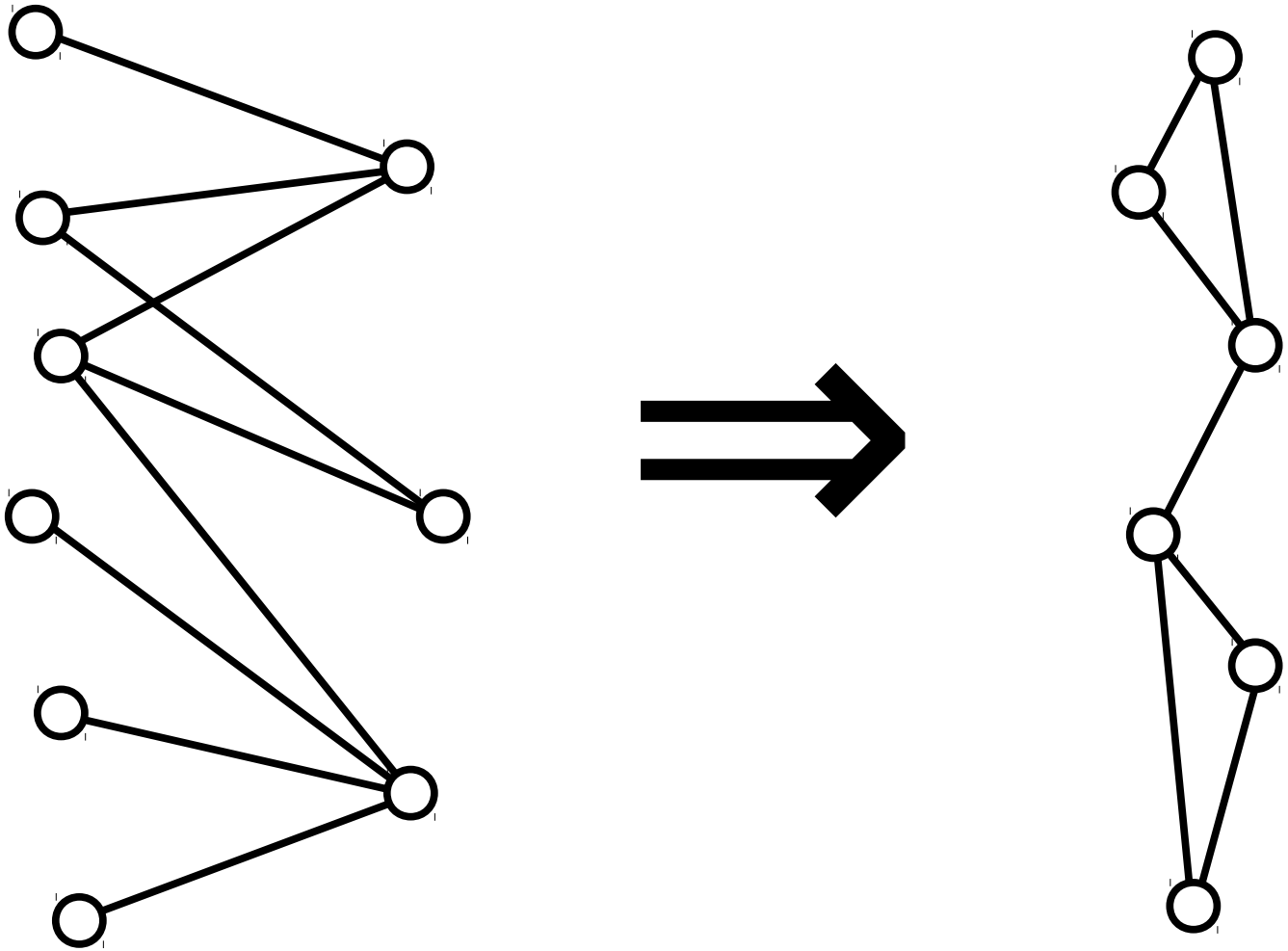
Data Integrity: *k*-Core



“Joined” Dataset



“Joined” Dataset



Other Semantics of Datasets

- Acyclic (only directed)
- Nonreciprocal (only directed)
- Loop-free
- etc.

Choosing a Dataset

http://konect.uni-koblenz.de/networks/

Code	Name ▲	Category	F.	W.	M.			n	m	Download
CL	Actor collaborations	Misc	U	=				382,219	33,115,812	
AM	Actor movies	Affiliation	B	−				511,463	1,470,404	
ME	Adolescent health	HumanSocial	D	+				2,539	12,969	
AD	Advogato	Social	D	+				6,541	51,127	
TC	Air traffic control	Infrastructure	D	−				1,226	2,615	
CA	Amazon (MDS)	Misc	U	−				334,863	925,872	
Am	Amazon (TWEB)	Misc	D	−				403,394	3,387,388	
AR	Amazon ratings	Rating	B	*				3,376,972	5,838,041	
Ar	American Revolution	Affiliation	B	−				141	160	
AP	arXiv astro-ph	Coauthorship	U	−				18,771	198,050	
AC	arXiv cond-mat	Authorship	B	−				38,741	58,595	
PH	arXiv hep-ph	Coauthorship	U	−				28,093	4,596,803	
PHc	arXiv hep-ph	Citation	D	−				34,546	421,578	
TH	arXiv hep-th	Coauthorship	U	−				22,908	2,673,133	
THc	arXiv hep-th	Citation	D	−				27,770	352,807	
BAl	Baidu internal links	Hyperlink	D	=				2,141,300	17,794,839	
BAr	Baidu related pages	Hyperlink	D	=				415,641	3,284,387	
BS	Berkeley/Stanford	Hyperlink	D	−				685,230	7,600,595	
MN	Bible	Lexical	U	+				1,773	9,131	
BtI	BibSonomy tag-publication	Folksonomy	B	=				972,120	2,555,080	

Choose by Data Size

http://konect.uni-koblenz.de/networks/

Code	Name	Category	F.	W.	M.	n	m ▼	Download
FR	Friendster	Social	D	−		68,349,466	2,586,147,869	
TF	Twitter (MPI)	Social	D	−		52,579,682	1,963,263,821	
TW	Twitter (WWW)	Social	D	−		41,652,230	1,468,365,182	
Wen	Wikipedia links, English	Hyperlink	D	−		12,150,976	378,142,420	
OG	Orkut	Affiliation	B	−		11,514,053	327,037,487	
Dui	Delicious user–URL	Folksonomy	B	=		34,611,302	301,186,579	
Dut	Delicious user–tag	Folksonomy	B	=		5,345,180	301,186,579	
Dti	Delicious tag–URL	Folksonomy	B	=		38,289,740	301,183,605	
en	Wikipedia (en)	Authorship	B	=		25,323,882	266,769,613	
YS	Yahoo songs	Rating	B	✶		1,625,951	256,804,235	
DL	Wikipedia, English	Hyperlink	D	=		18,268,992	172,183,984	
TR	TREC (disks 4–5)	Text	B	=		1,729,302	151,632,178	
WT	Web trackers	Hyperlink	B	−		40,421,974	140,613,762	
OR	Orkut	Social	U	−		3,072,441	117,185,083	
LG	LiveJournal	Affiliation	B	−		10,690,276	112,307,385	
Wfr	Wikipedia links, French	Hyperlink	D	−		3,023,165	102,382,410	
NX	Netflix	Rating	B	✶		497,959	100,480,507	
RE	Reuters	Text	B	=		1,065,176	96,903,520	
Wit	Wikipedia links, Italian	Hyperlink	D	−		1,865,965	91,555,008	
Ug	Wikipedia, de (dynamic)	Hyperlink	D	⇄		2,166,669	86,337,879	
Wru	Wikipedia links, Russian	Hyperlink	D	−		2,853,118	82,056,101	

Choose by Property

http://konect.uni-koblenz.de/statistics/diameter

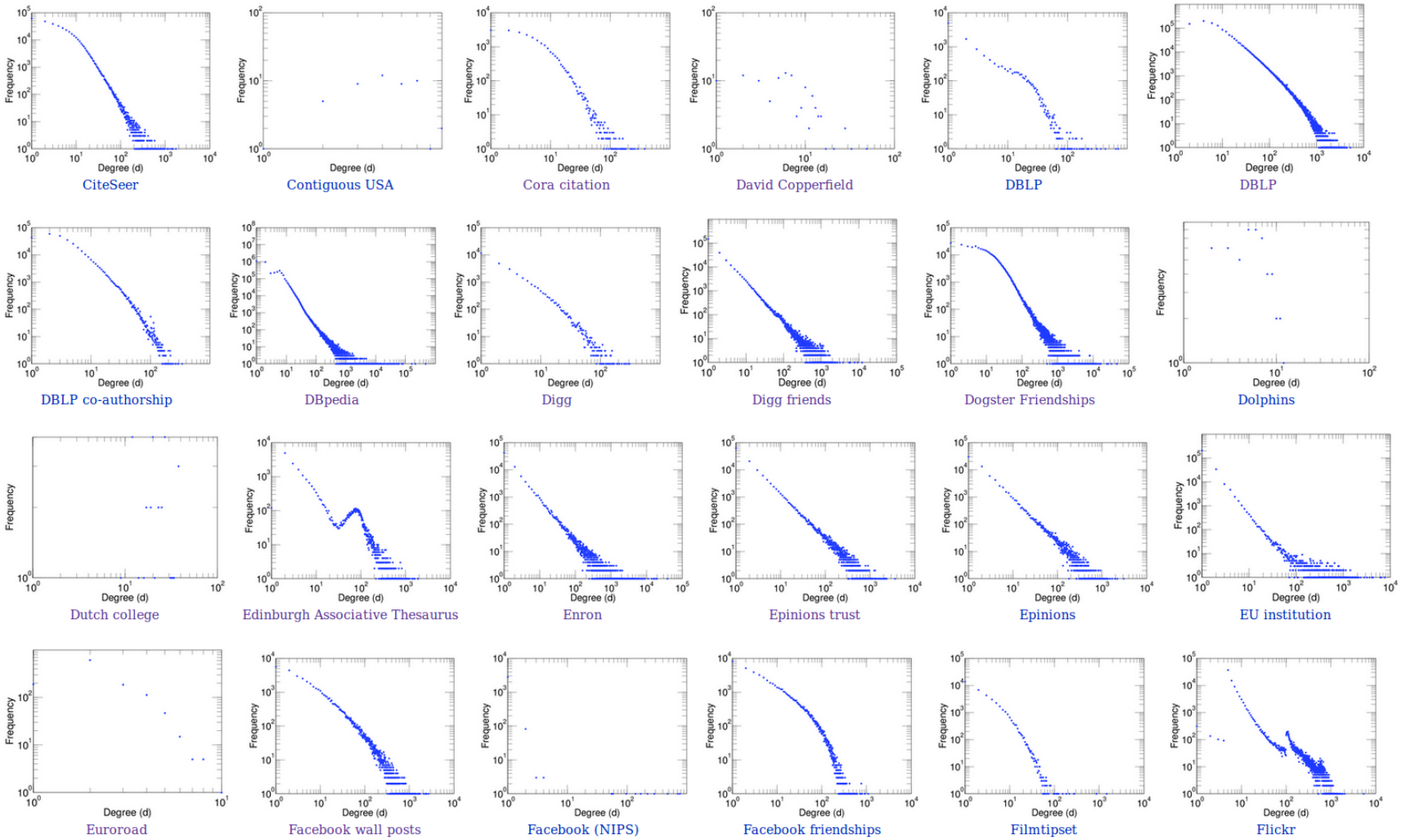
Code	Name	Category	F.	W.	M.		n	m	δ ▼	$\delta_{0.9}$	δ_{m}	$\tilde{\delta}$	δ_{M}
R1	Texas	Infrastructure	U	—			1,379,917	1,921,660	1,064	698.83	451.40	5.98472×10^{2}	
RO	California	Infrastructure	U	—			1,965,206	2,766,607	865	511.07	315.89	4.95040×10^{2}	
RD	Pennsylvania	Infrastructure	U	—			1,088,092	1,541,898	794	528.61	312.58	5.45006×10^{2}	
BS	Berkeley/Stanford	Hyperlink	D	—			685,230	7,600,595	208	9.79	7.21	1.03080×10^{1}	
SF	Stanford	Hyperlink	D	—			281,903	2,312,497	164	8.79	6.36	9.40826×10^{0}	
WT	TREC WT10g	Hyperlink	D	—			1,601,787	8,063,026	112	11.10	8.70	1.07196×10^{1}	
HUr	Hudong related	Hyperlink	D	—			2,452,715	18,854,882	108	4.96	4.41		
DB	DBpedia	Misc	D	=			3,966,924	13,820,853	67	6.27	5.19	6.50821×10^{0}	
Lk	Linux kernel mailing list replies	Communication	D	=			63,399	1,096,440	63	7.79	5.19		
ET	Euroroad	Infrastructure	U	—			1,174	1,417	62	33.34	19.18	3.23879×10^{1}	
WR	Writers	Authorship	B	—			135,569	144,340	60	20.95	15.10	2.09247×10^{1}	
PR	Producers	Authorship	B	—			187,677	207,268	50	17.20	11.31	1.86413×10^{1}	
Pa	DBLP	Authorship	B	—			5,425,963	8,649,016	50	13.93	11.47	1.39057×10^{1}	
CA	Amazon (MDS)	Misc	U	—			334,863	925,872	47	14.85	11.73		
WC	Wikipedia (en)	Feature	B	—			2,036,440	3,795,796	46	15.31	11.75	1.37377×10^{1}	
ND	Notre Dame	Hyperlink	D	—			325,729	1,497,134	46	8.92	6.96	9.68814×10^{0}	
UG	US power grid	Infrastructure	U	—			4,941	6,594	46	28.17	20.09	2.83547×10^{1}	
TM	Teams	Affiliation	B	—			935,627	1,366,466	41	7.69	6.37	7.17361×10^{0}	
ST	Movies	Feature	B	—			157,184	281,396	38	12.54	9.43	1.09744×10^{1}	
AC	arXiv cond-mat	Authorship	B	—			38,741	58,595	36	16.86	12.83	1.51815×10^{1}	
CS	CiteSeer	Citation	D	—			384,413	1,751,463	34	7.96	6.35	7.99549×10^{0}	
BAr	Baidu related	Hyperlink	D	=			415,641	3,284,387	32	8.54	6.11	8.88541×10^{0}	
MC	Crime	Interaction	B	—			1,380	1,476	32	19.47	13.37		
GE	DBpedia genre	Feature	B	—			266,717	463,497	32	7.44	5.08	6.19932×10^{0}	
YT	YouTube	Social	U	—			3,223,589	9,375,374	31	6.64	5.29		

Choose by Many Network Statistics

Statistic	Expression
Size	$n = \|V\|$
Volume	$m = \|E\| = \frac{1}{2}\|\mathbf{A}\|_{\mathrm{F}}^2$
Average degree	$d = \frac{1}{\|V\|}\sum_{u\in V} d(u) = \frac{2m}{n}$
Fill	$p = \begin{cases} 2m/[n(n+1)] & \text{when } G \text{ is undirected} \\ m/n^2 & \text{when } G \text{ is directed} \\ m/(n_1 n_2) & \text{when } G \text{ is bipartite} \end{cases}$
Maximum degree	$d_{\max} = \max_{u\in V} d(u)$
Reciprocity	$y = \frac{1}{m}\|\{(u,v)\in E \mid (v,u)\in E\}\|$
Negativity	$\zeta = \frac{\|\{e\in E \mid w(e)<0\}\|}{m}$
LCC	$N = \max_{F\subseteq C}\|F\|$
Wedge count	$s = \sum_{u\in V}\binom{d(u)}{2} = \sum_{u\in V}\frac{1}{2}d(u)(d(u)-1)$
Claw count	$z = \sum_{u\in V}\binom{d(u)}{3} = \sum_{u\in V}\frac{1}{6}d(u)(d(u)-1)(d(u)-2)$
Triangle count	$t = \|\{\{u,v,w\} \mid u\sim v\sim w\sim u\}\| \,/\, 6$
Square count	$q = \|\{u,v,w,x \mid u\sim v\sim w\sim x\sim u\}\|/8$
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	$H_{\mathrm{er}} = \frac{1}{\ln\|V\|}\sum_{u\in V} -\frac{d(u)}{D}\ln\frac{d(u)}{D}$
Assortativity	ρ
Clustering coefficient	$c = \frac{\|\{u,v,w\in V \mid u\sim v\sim w\sim u\}\|}{\|\{u,v,w\in V \mid u\sim v\neq w\sim u\}\|} = \frac{3t}{s}$
Diameter	$\delta = \max_{u\in E}\epsilon(u) = \max_{u,v\in E} d(u,v)$
Spectral norm	$\|\lambda_1[\mathbf{A}]\| = \|\mathbf{A}\|_2$
Algebraic connectivity	$a = \lambda_2[\mathbf{L}]$
Preferential attachment exponent	$\min_{\alpha,\beta}\sum_{u\in V}(\alpha + \beta\ln[1+d_1(u)] - \ln[\lambda + d_2(u)])^2$

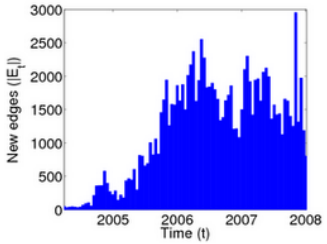
http://konect.uni-koblenz.de/statistics/

Choose by Degree Distribution

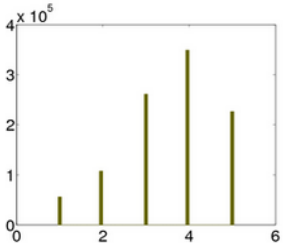


http://konect.uni-koblenz.de/plots/degree_distribution

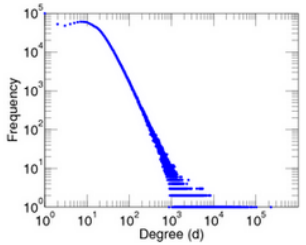
Choose by Any Plot Type



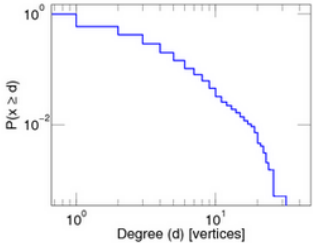
Temporal distribution



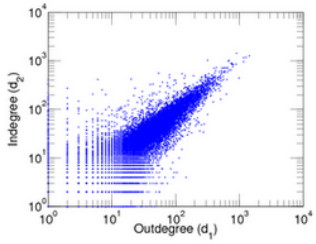
Edge weight and multiplicity distribution



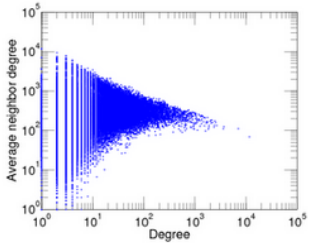
Degree distribution



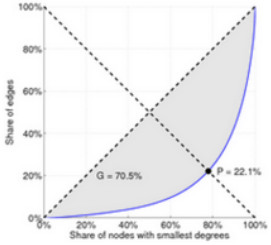
Cumulative degree distribution



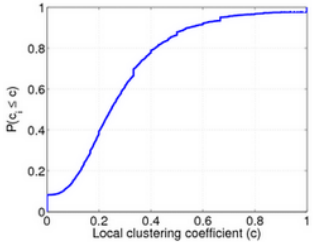
Out/indegree comparison



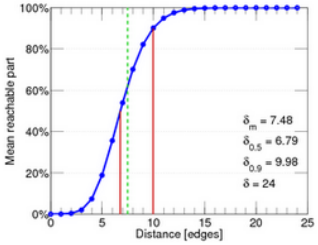
Assortativity plot



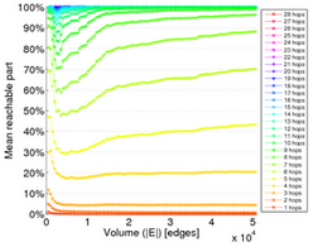
Lorenz curve



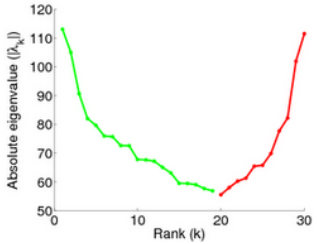
Clustering coefficient distribution



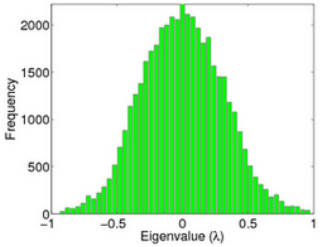
Distance distribution



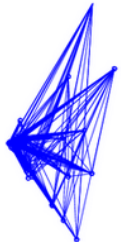
Temporal distance distribution

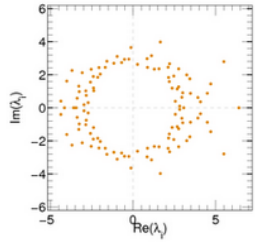


Top-*k* eigenvalues

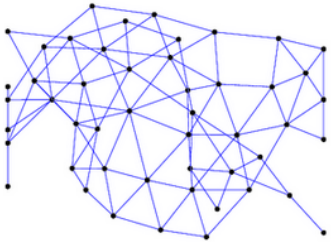


Spectral plots


Network visualization



Complex eigenvalues of the asymmetric adjacency matrix

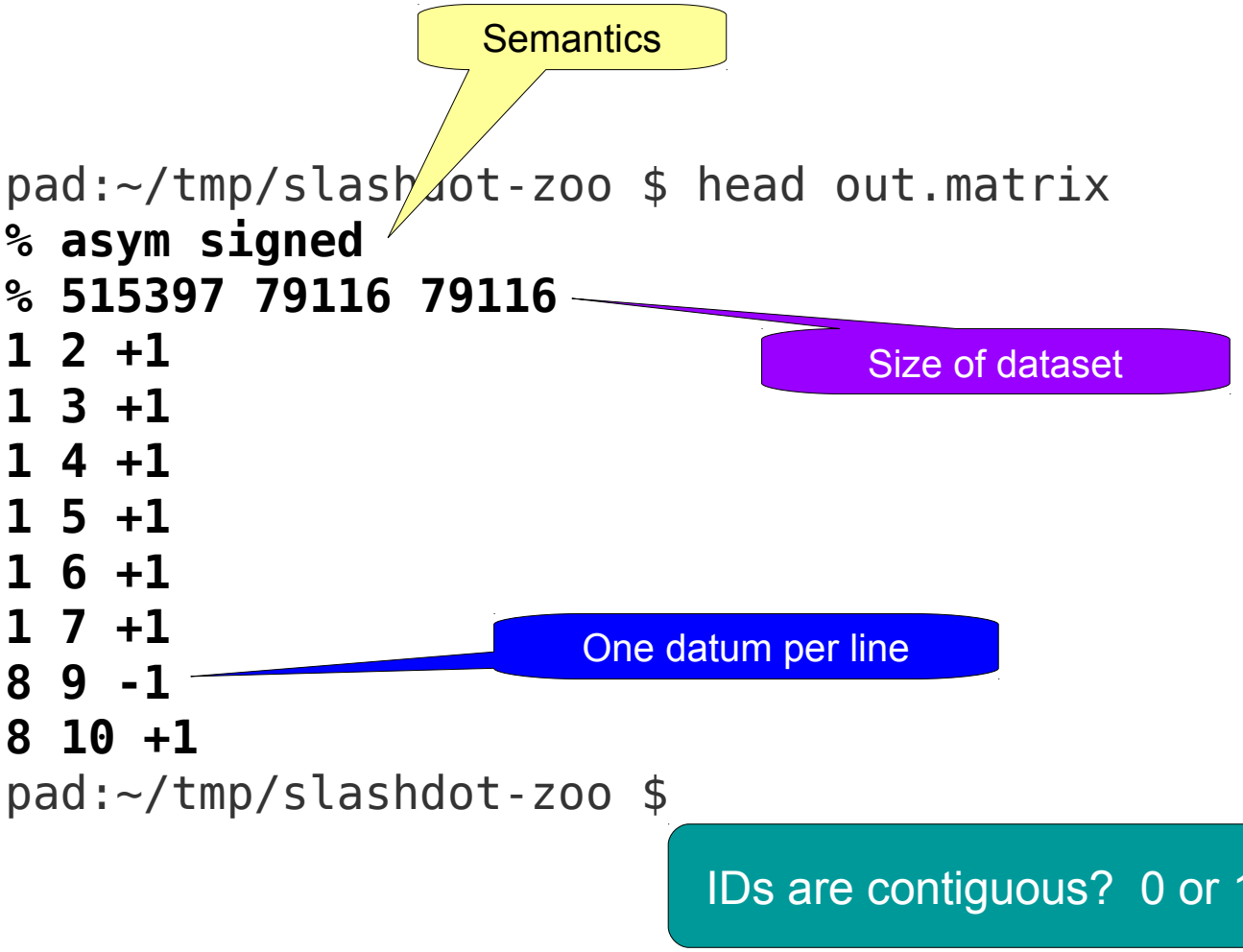

Layout

http://konect.uni-koblenz.de/plots/

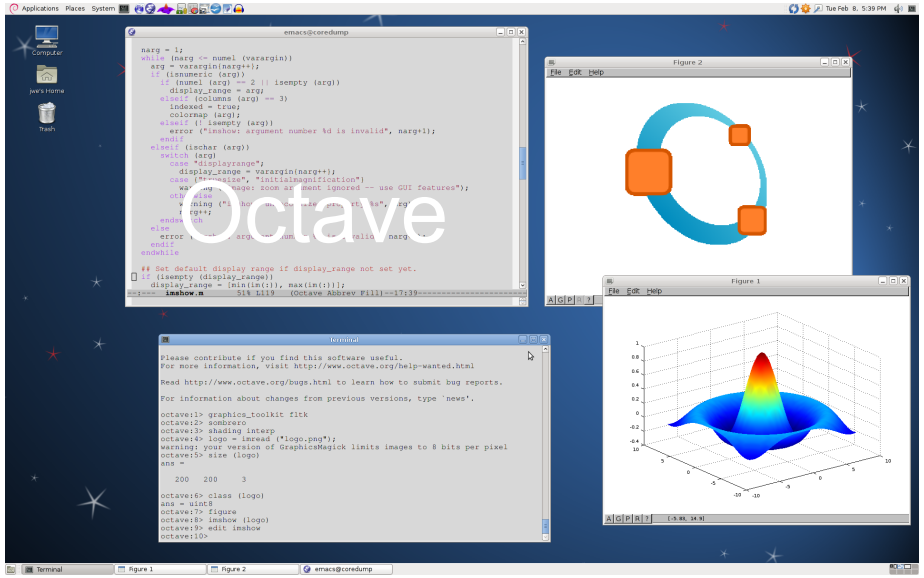
Reading Data Files

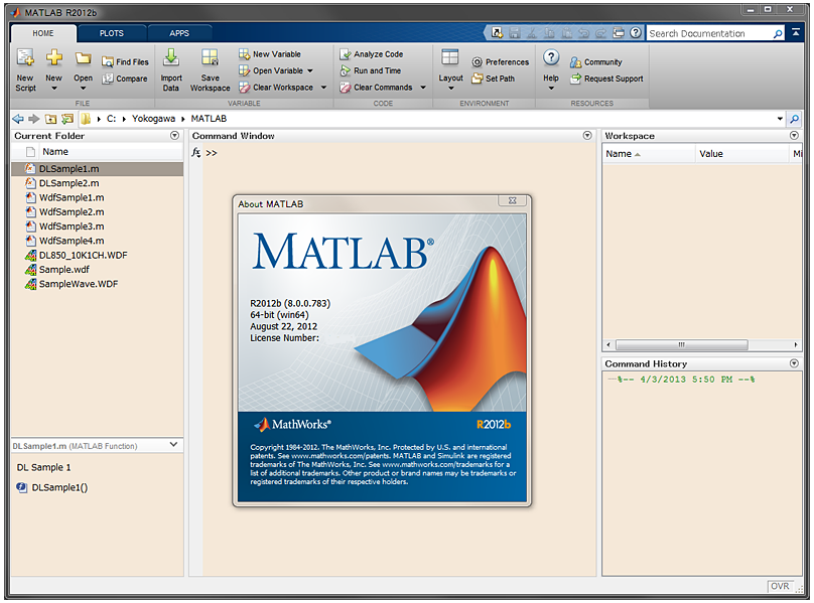
```
pad:~/tmp/slashdot-zoo $ head out.matrix
% asym signed
% 515397 79116 79116
1 2 +1
1 3 +1
1 4 +1
1 5 +1
1 6 +1
1 7 +1
8 9 -1
8 10 +1
pad:~/tmp/slashdot-zoo $
```

Reading Data Files



Numerical Computing Languages





C

python

FORTRAN

julia

C++

S

R

Java™

Thank you

Network dataset donations accepted at kunegis@uni-koblenz.de

Jérôme Kunegis & Steffen Staab

WeST
People and Knowledge Networks

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