Network Analysis in the Age of Large Network Dataset Collections – Challenges, Solutions and Applications

Tutorial

Jérôme KUNEGIS, Renaud LAMBIOTTE
with acknowledgments to many people

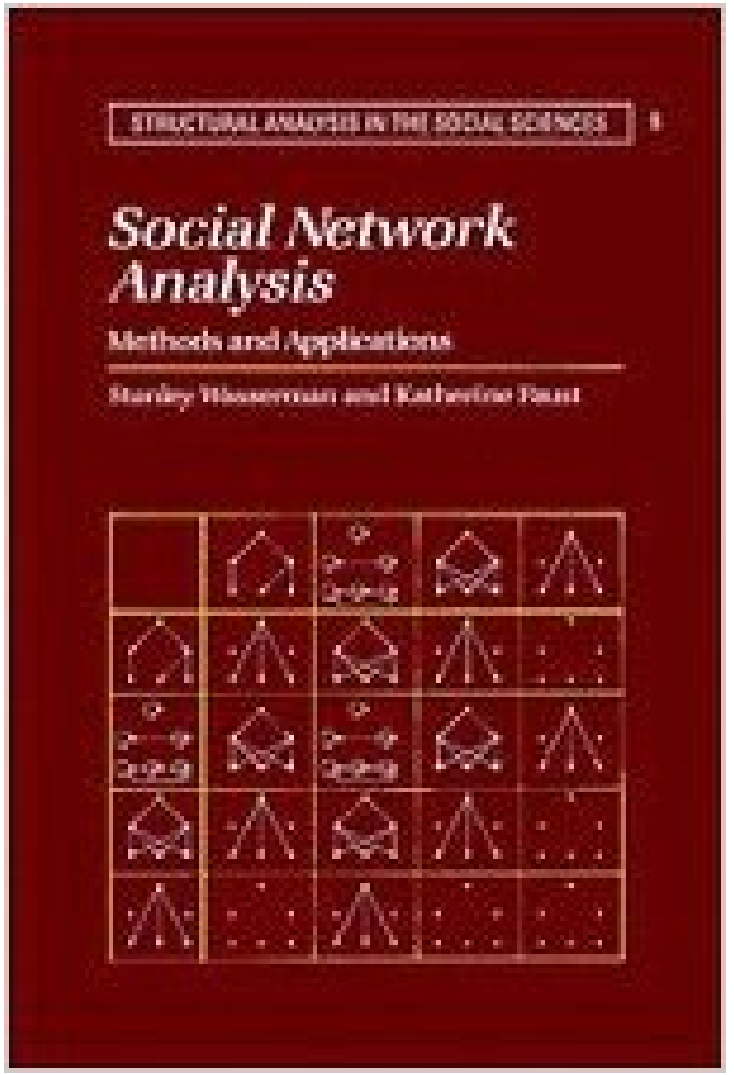
10 Nov 2017
CIKM'17

http://KONECT.cc/

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 0 1 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 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 0
15	0 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 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 0 1 0 0 0 0 0 0 0
20	0 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 0

Social Network Analysis without the Web

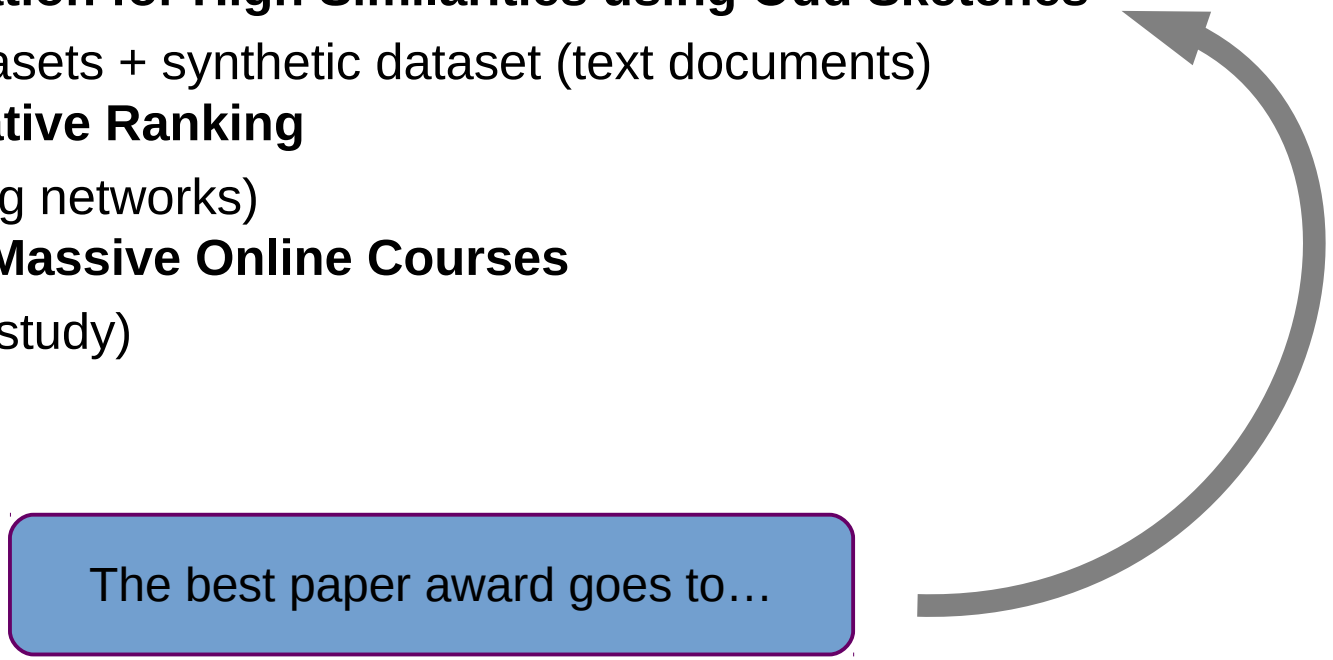


“Social Network Analysis” by
Wasserman & Faust first edition 1994

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)

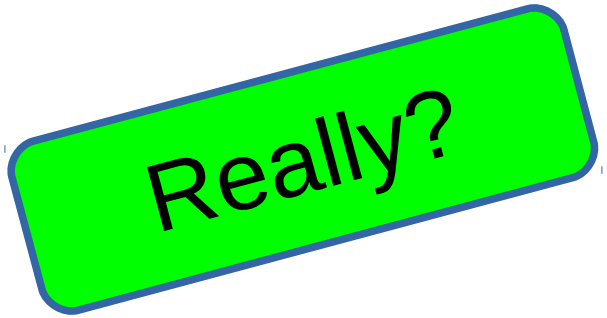


Why Do Researchers Use Multiple Datasets?

- 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.

Who We Are

Jérôme KUNEGIS
University of Namur
Belgium

Renaud LAMBIOTTE
University of Namur
University of Oxford

The KONECT.cc Project – Koblenz Network Collection

On the Spectral Evolution of Large Networks

Jérôme Kunegis

Institute for Web Science and Technologies
University of Koblenz-Landau
kunegis@uni-koblenz.de

November 2011

Vom Promotionsausschuss des Fachbereichs 4: Informatik der Universität Koblenz-Landau zur Verleihung des akademischen Grades

Doktor der Naturwissenschaften (Dr. rer. nat.)

genehmigte Dissertation.

PhD thesis at the University of Koblenz-Landau.

Datum der wissenschaftlichen Aussprache:	9. November 2011
Vorsitz des Promotionsausschusses:	Prof. Dr. Karin Harbusch
Berichterstatter:	Prof. Dr. Steffen Staab
Berichterstatter:	Prof. Dr. Christian Bauckhage
Berichterstatter:	Prof. Dr. Klaus Obermayer

https://kola.opus.hbz-nrw.de/frontdoor/index/index/docId/581


The trick
is...


Everything
is a
NETWORK!

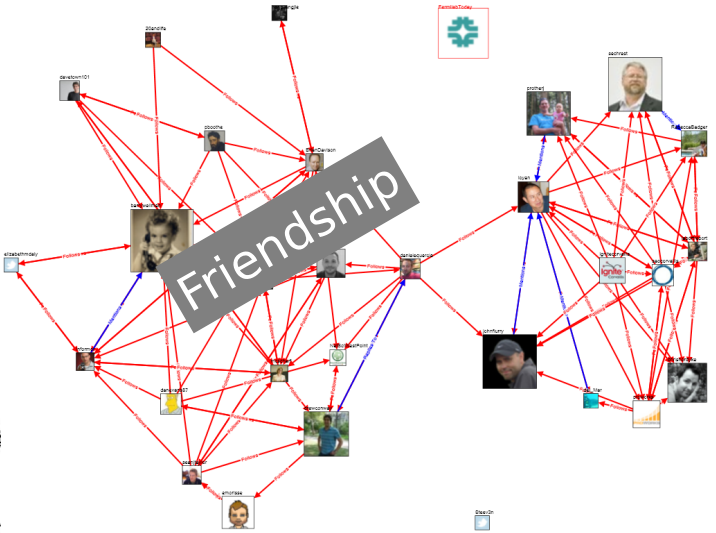
Explanation : Why *Everything Is a Network*

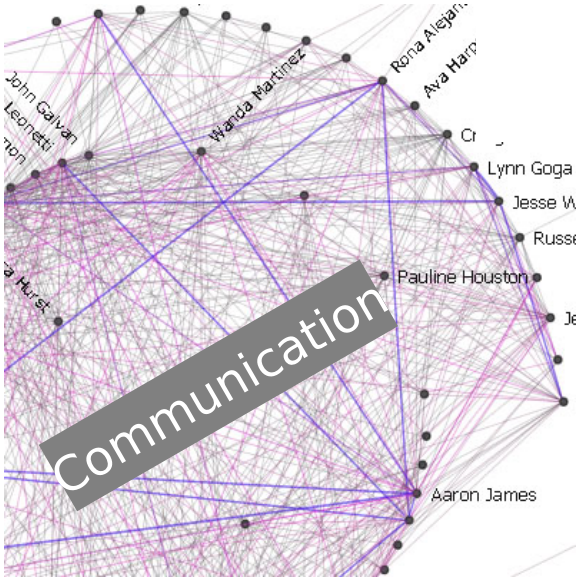
Argument by considering the structure of large systems :

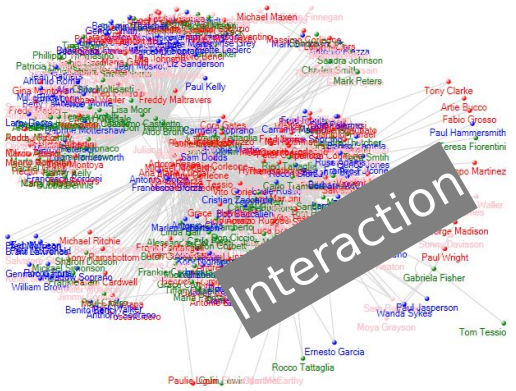
- Large systems consist of many parts
- Any part can only communicate with a finite number of other parts
- This gives rise to a sparse network structure

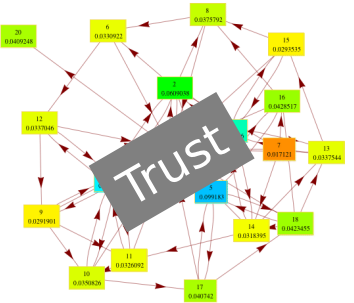
Caveat : May not be a complex system (e.g. a grid)

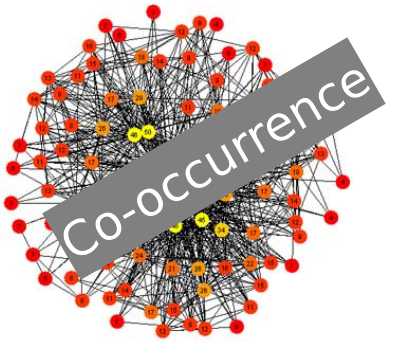
Well, Only *Almost* Everything Is a Network



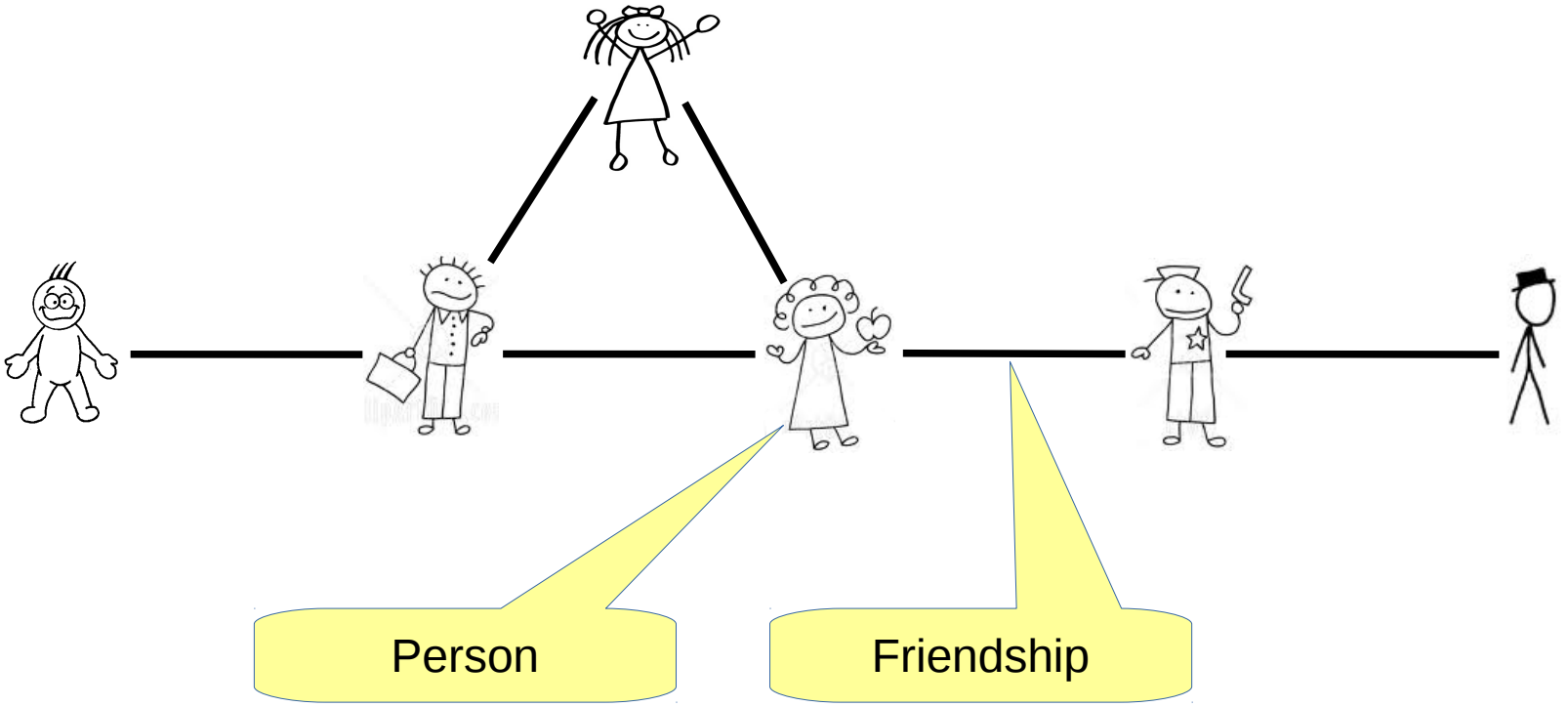






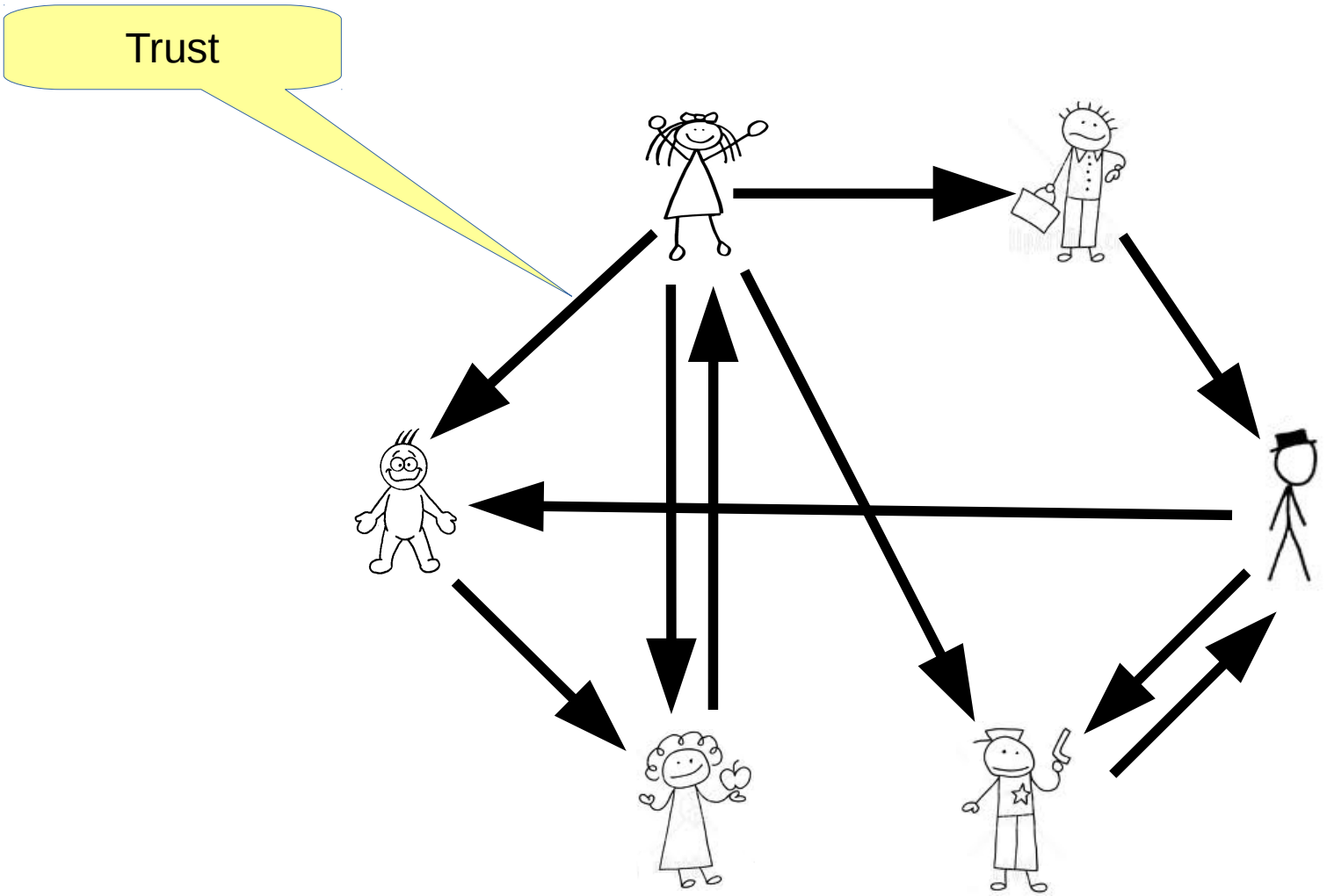


Social Network



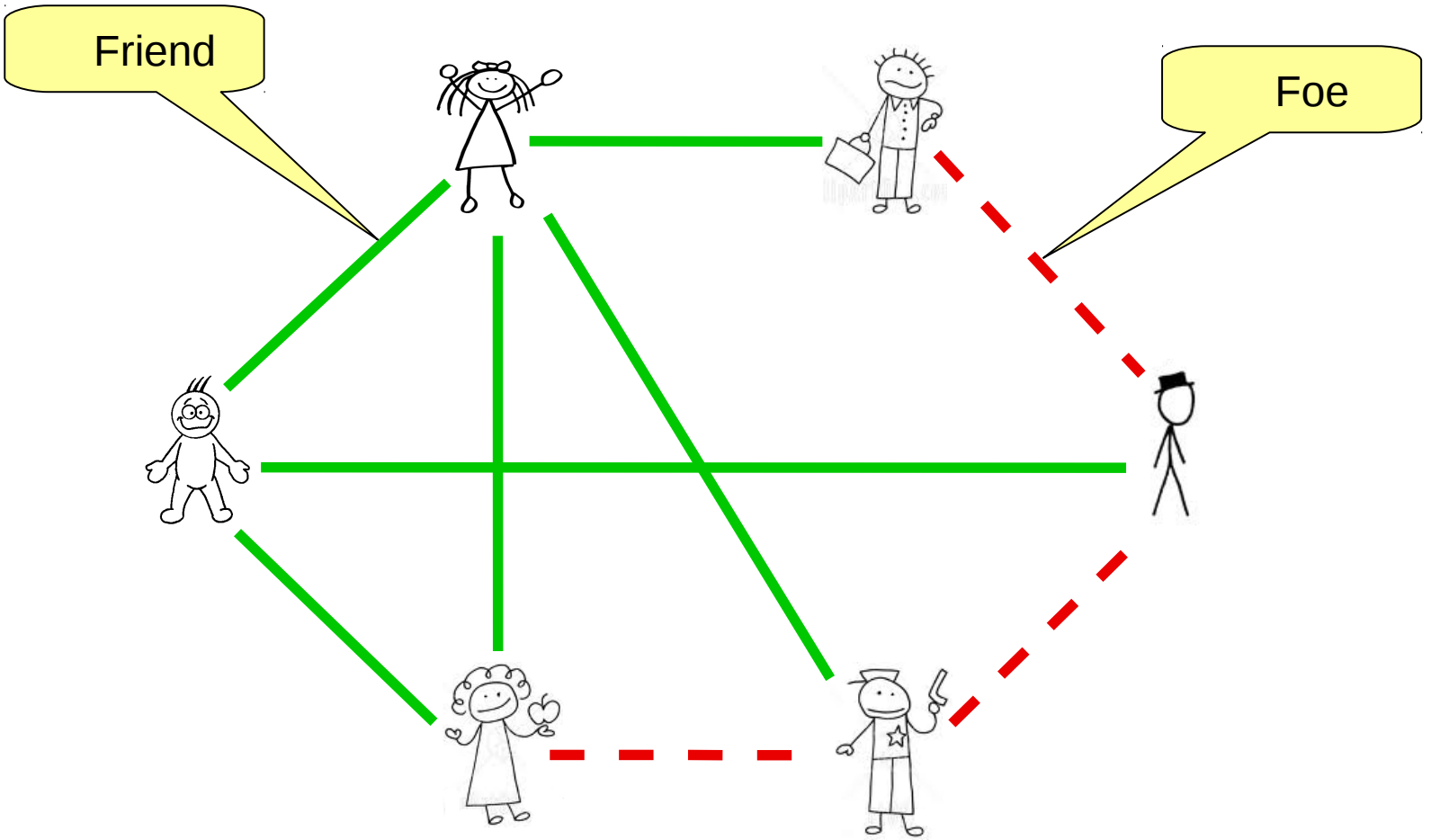
http://konect.cc/networks/facebook-wosn-links/

Trust Network



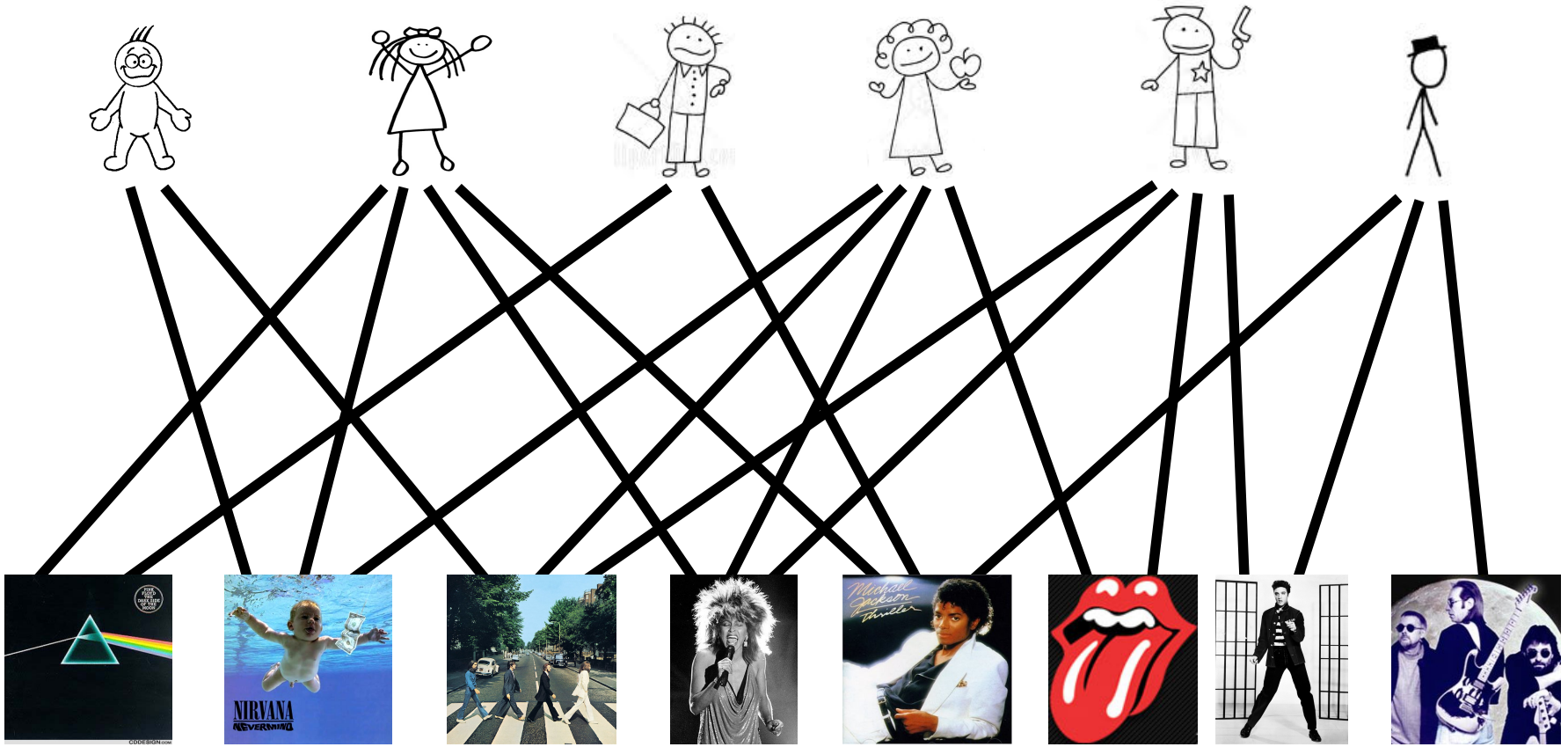
http://konect.cc/networks/advogato/

Signed Network



http://konect.cc/networks/slashdot-zoo/

Bipartite Network



http://konect.cc/networks/lastfm_band/

A Network Dataset Is Like a Gummi Bear

A Network Dataset Is Like a Gummi Bear

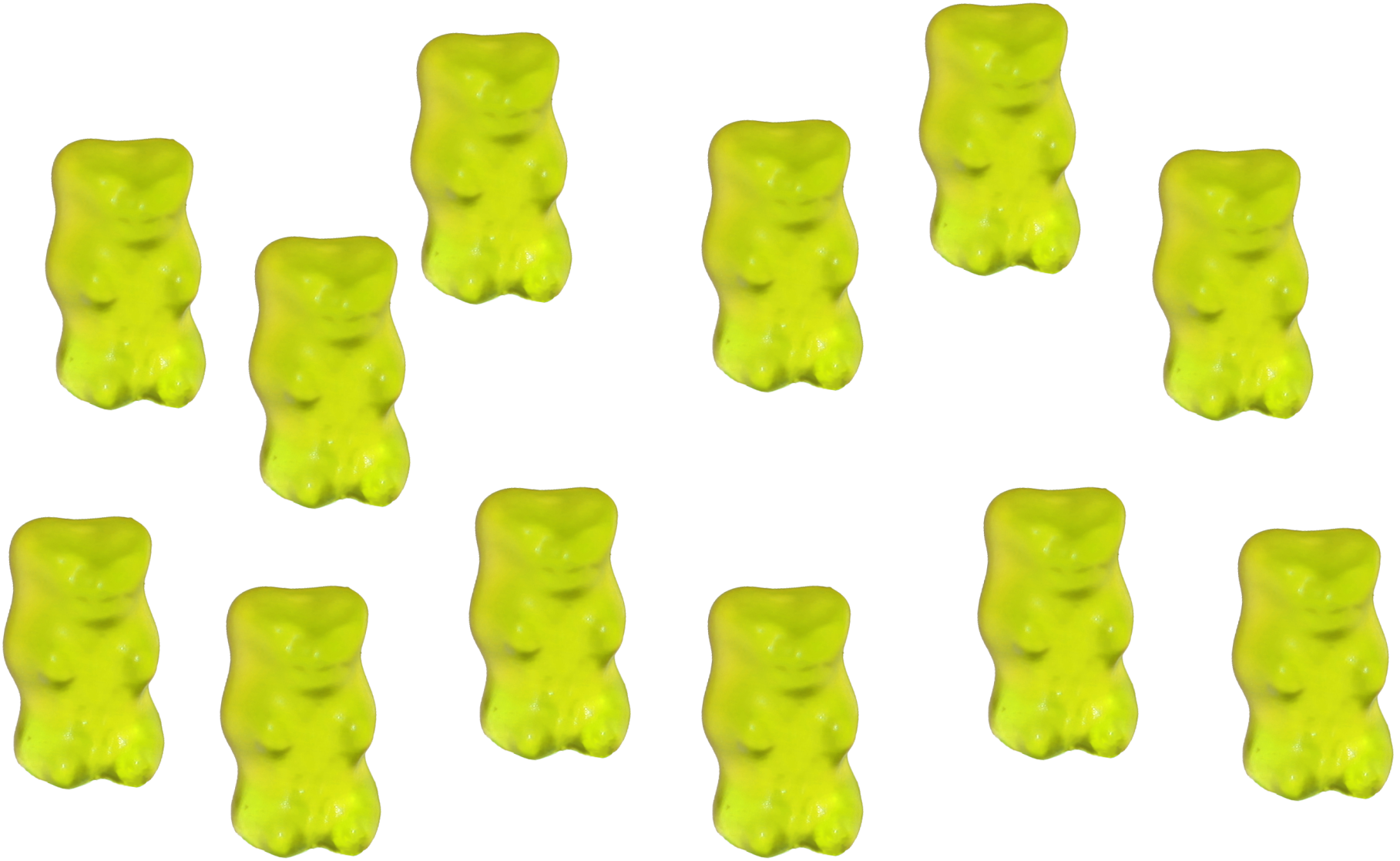
Lots of content to analyse

Test network models

Evaluate prediction algorithms

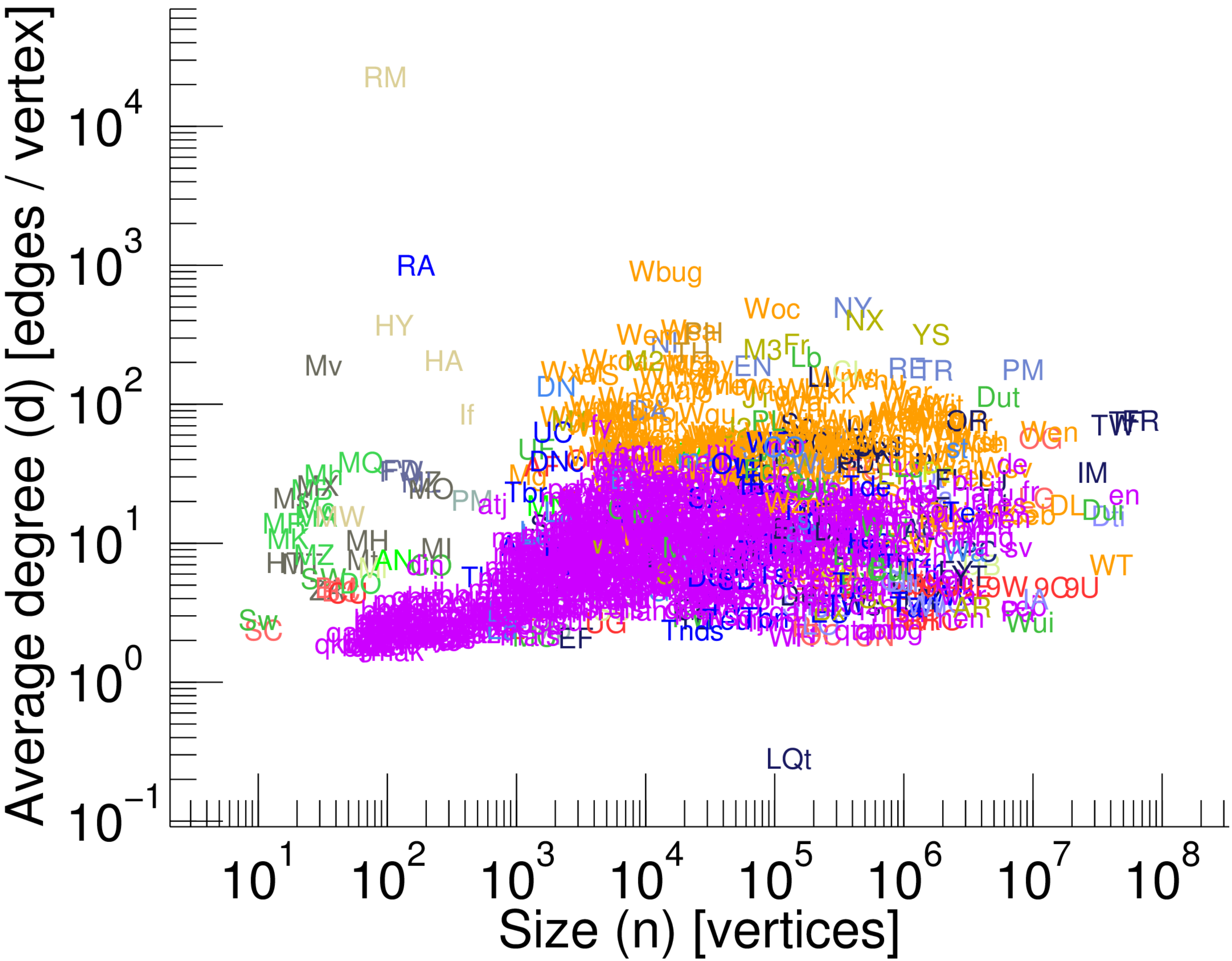
Test scalability of algorithms

When You Have Tested One, You Have Tested All ?!



Or Do You?

Diversity of Network Datasets



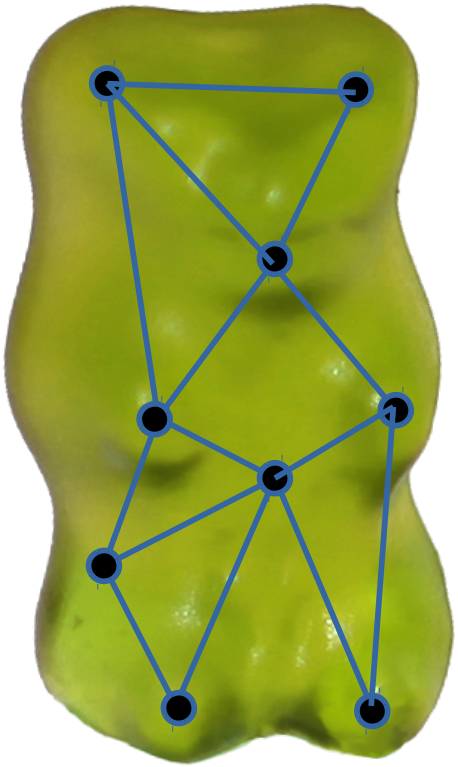
Network Categories

	Internal name	Vertices	Edges	Properties	Count
●	`Affiliation`	Actors, groups	Membership	B − =	16
●	`Animal`	Animals	Tie	U D − +	9
●	`Authorship`	Authors, works	Authorship	B − =	808
●	`Citation`	Documents	Citation	D −	7
●	`Coauthorship`	Authors	Coauthorship	U − =	3
●	`Cocitation`	Authors	Cocitation	U =	2
●	`Communication`	Persons	Message	U D − =	42
●	`Computer`	Computers	Connection	U D − =	13
●	`Feature`	Items, features	Property	B − = +	17
●	`HumanContact`	Persons	Real-life contact	U = +	5
●	`HumanSocial`	Persons	Real-life tie	U D − + ± $\stackrel{+}{=}$	12
●	`Hyperlink`	Web page	Hyperlink	D B − = ⇌	191
●	`Infrastructure`	Location	Connection	U D − = +	23
●	`Interaction`	Persons, items	Interaction	D B − = $\stackrel{+}{=}$	25
●	`Lexical`	Words	Lexical relationship	U D − =	5
●	`Metabolic`	Metabolites	Interaction	U D − =	7
●	`Misc`	Various	Various	U D − = +	11
●	`OnlineContact`	Users	Online interaction	U D − = ± $\stackrel{+}{=}$ ⇌	15
●	`Rating`	Users, items	Rating	B − = $*$ $_*^*$	15
●	`Social`	Persons	Online tie	U D − + ± $*$ ⇌	46
●	`Software`	Software Component	Dependency	D − =	3
●	`Text`	Documents, words	Occurrence	B =	10
●	`Trophic`	Species	Carbon exchange	D − +	3
	Total				**1288**

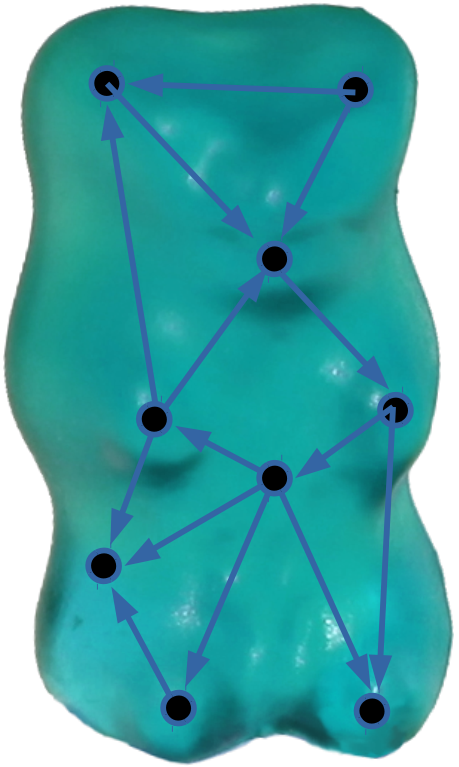
source: https://github.com/kunegis/konect-handbook/blob/master/konect-handbook.pdf

Network Formats

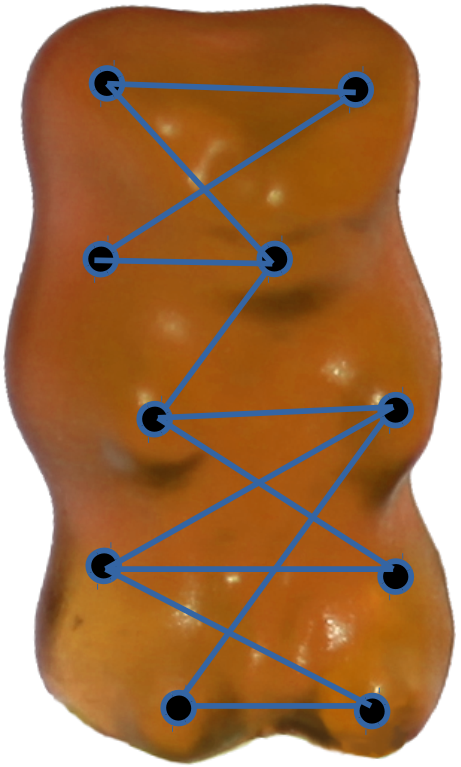
U • Undirected



D • Directed



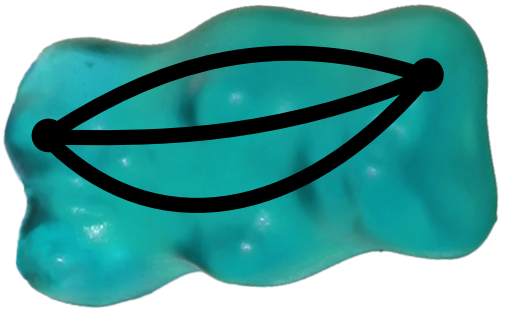
B • Bipartite



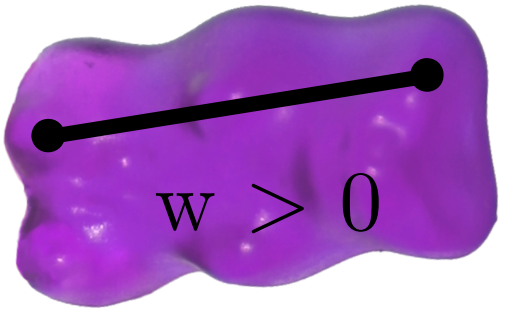
Edge Weight and Multiplicity Types

– • Unweighted

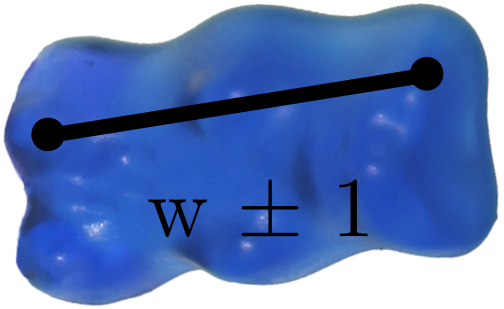
= • Multiple



+ • Positive



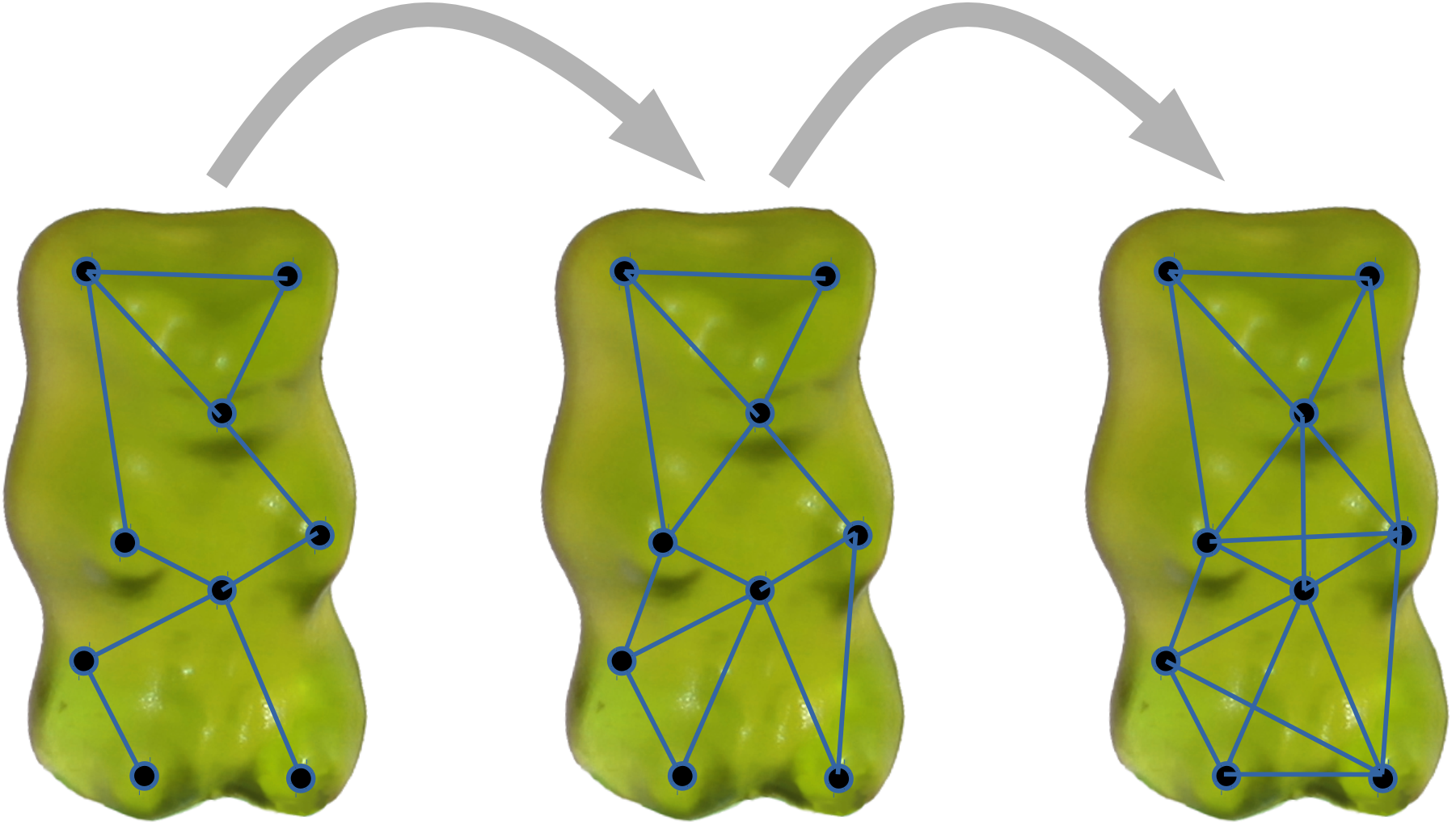
± • Signed



∗ • Rating

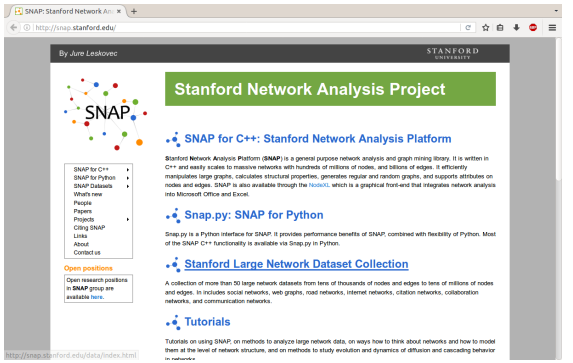
$\overset{*}{*}$ • Multiple Ratings

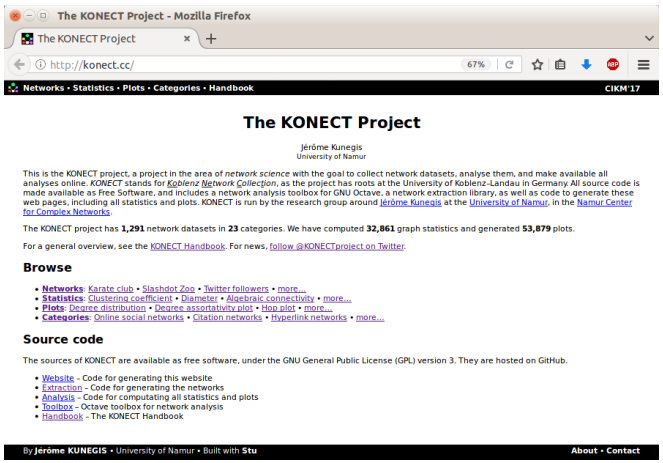
Timestamps

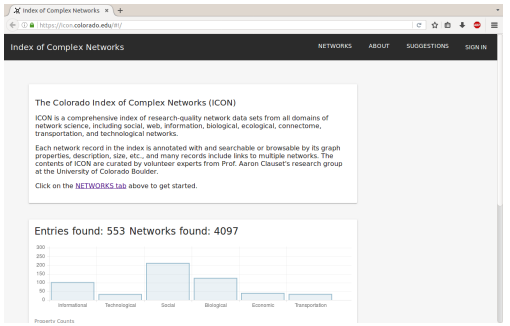


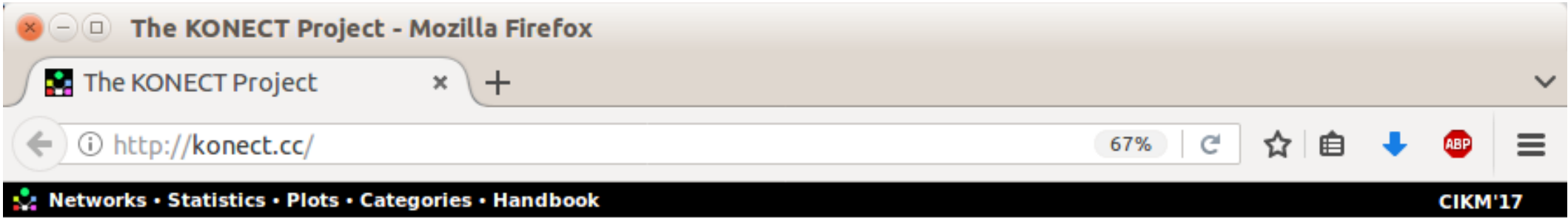
Collections of Network Datasets

- **SNAP**.stanford.edu
  - « Stanford Network Analysis Project »
  - by Jure Leskovec, Stanford Univ. (~2009)
  - several 100 networks; not systematic
  - Available for download
  - Some statistics available
- **KONECT**.cc
  - « Koblenz Network Collection »
  - by Jérôme Kunegis, Univ. of Namur (~2011)
  - 1200+ networks
  - Most networks available for download
  - Many statistics available
- **ICON**.colorado.edu
  - « Index of Complex Networks »
  - by Aaron Clauset, Univ. of Colorado (~2016)
  - 4000+ datasets
  - Not available for download (“index”)









The KONECT Project

Jérôme Kunegis
University of Namur

This is the KONECT project, a project in the area of *network science* with the goal to collect network datasets, analyse them, and make available all analyses online. *KONECT* stands for *Koblenz Network Collection*, as the project has roots at the University of Koblenz–Landau in Germany. All source code is made available as Free Software, and includes a network analysis toolbox for GNU Octave, a network extraction library, as well as code to generate these web pages, including all statistics and plots. KONECT is run by the research group around Jérôme Kunegis at the University of Namur, in the Namur Center for Complex Networks.

The KONECT project has **1,291** network datasets in **23** categories. We have computed **32,861** graph statistics and generated **53,879** plots.

For a general overview, see the KONECT Handbook. For news, follow @KONECTproject on Twitter.

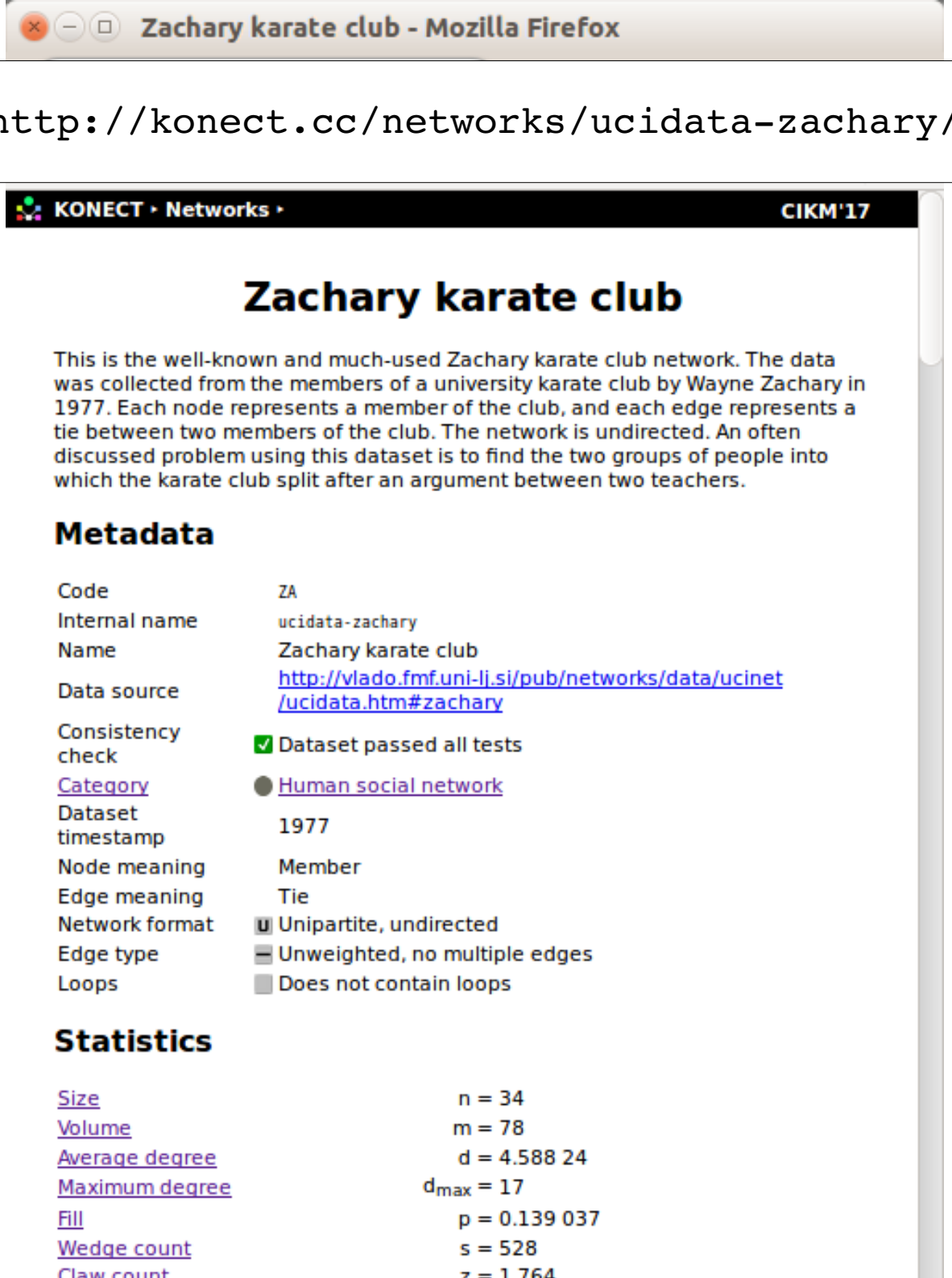
Browse

- **Networks**: Karate club • Slashdot Zoo • Twitter followers • more...
- **Statistics**: Clustering coefficient • Diameter • Algebraic connectivity • more...
- **Plots**: Degree distribution • Degree assortativity plot • Hop plot • more...
- **Categories**: Online social networks • Citation networks • Hyperlink networks • more...

Source code

The sources of KONECT are available as free software, under the GNU General Public License (GPL) version 3. They are hosted on GitHub.

- Website – Code for generating this website
- Extraction – Code for generating the networks
- Analysis – Code for computating all statistics and plots
- Toolbox – Octave toolbox for network analysis
- Handbook – The KONECT Handbook

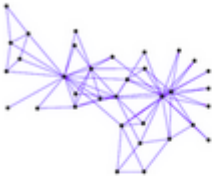

Zachary karate club - Mozilla Firefox
http://konect.cc/networks/ucidata-zachary/
KONECT ▸ Networks ▸
CIKM'17
Zachary karate club
This is the well-known and much-used Zachary karate club network. The data was collected from the members of a university karate club by Wayne Zachary in 1977. Each node represents a member of the club, and each edge represents a tie between two members of the club. The network is undirected. An often discussed problem using this dataset is to find the two groups of people into which the karate club split after an argument between two teachers.
Metadata
Code
ZA
Internal name
ucidata-zachary
Name
Zachary karate club
Data source
http://vlado.fmf.uni-lj.si/pub/networks/data/ucinet/ucidata.htm#zachary
Consistency check
Dataset passed all tests
Category
Human social network
Dataset timestamp
1977
Node meaning
Member
Edge meaning
Tie
Network format
Unipartite, undirected
Edge type
Unweighted, no multiple edges
Loops
Does not contain loops
Statistics
Size
n = 34
Volume
m = 78
Average degree
d = 4.588 24
Maximum degree
d_{max} = 17
Fill
p = 0.139 037
Wedge count
s = 528
Claw count
z = 1 764

Statistics

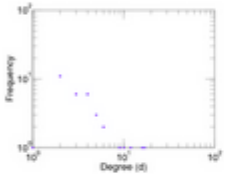
Size	$n = 34$
Volume	$m = 78$
Average degree	$d = 4.588\,24$
Maximum degree	$d_{max} = 17$
Fill	$p = 0.139\,037$
Wedge count	$s = 528$
Claw count	$z = 1{,}764$
Cross count	$x = 5{,}082$
Size of LCC	$N = 34$
Relative size of LCC	$N^{rel} = 1.000\,00$
Degree assortativity	$\rho = -0.475\,613$
Degree assortativity p-value	$p_\rho = 3.509\,45 \times 10^{-10}$
Spectral norm	$\|A\|_2 = 6.725\,70$
Gini coefficient	$G = 0.385\,370$
Power law exponent	$\gamma = 1.780\,96$
Tail power law exponent	$\gamma_t = 2.161\,00$
Relative edge distribution entropy	$H_{er} = 0.924\,709$
Clustering coefficient	$c = 0.255\,682$
Triangle count	$t = 45$
Diameter	$\delta = 5$
50-Percentile effective diameter	$\delta_{0.5} = 1.840\,54$
90-Percentile effective diameter	$\delta_{0.9} = 3.441\,98$
Mean distance	$\delta_m = 2.443\,26$
4-Tour count	$T_4 = 3{,}500$
Square count	$q = 154$
Algebraic connectivity	$a = 0.468\,525$

Plots

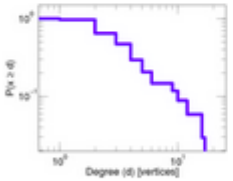
Fruchterman-Reingold graph drawing



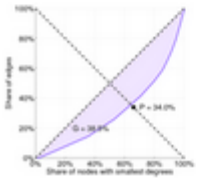
Degree distribution



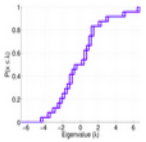
Cumulative degree distribution



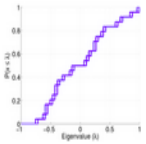
Lorenz curve



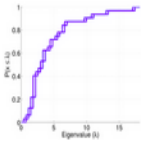
Spectral distribution of the adjacency matrix



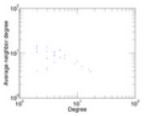
Spectral distribution of the normalized adjacency matrix



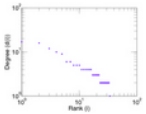
Spectral distribution of the Laplacian



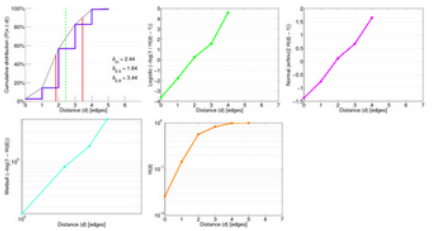
Degree assortativity



Zipf plot

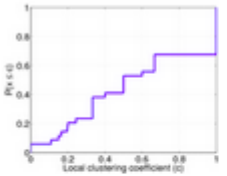


Hop distribution

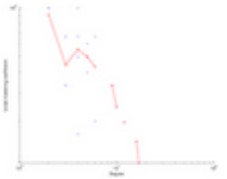


Double Laplacian graph drawing

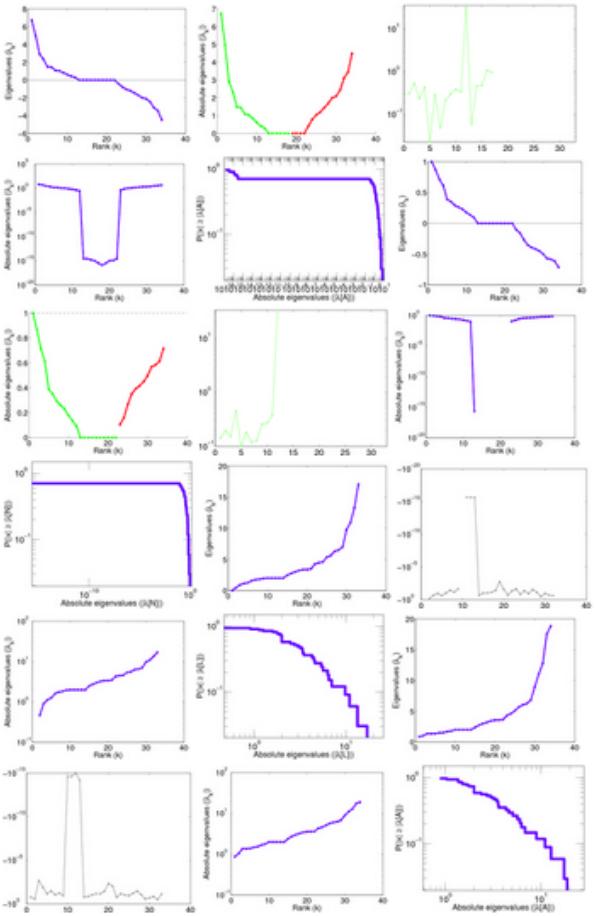
Clustering coefficient distribution



Average neighbor degree distribution



Matrix decompositions plots



http://konect.cc/plots/

Downloads

- Data as TSV (1,856 bytes)
- Extraction code (GitHub)

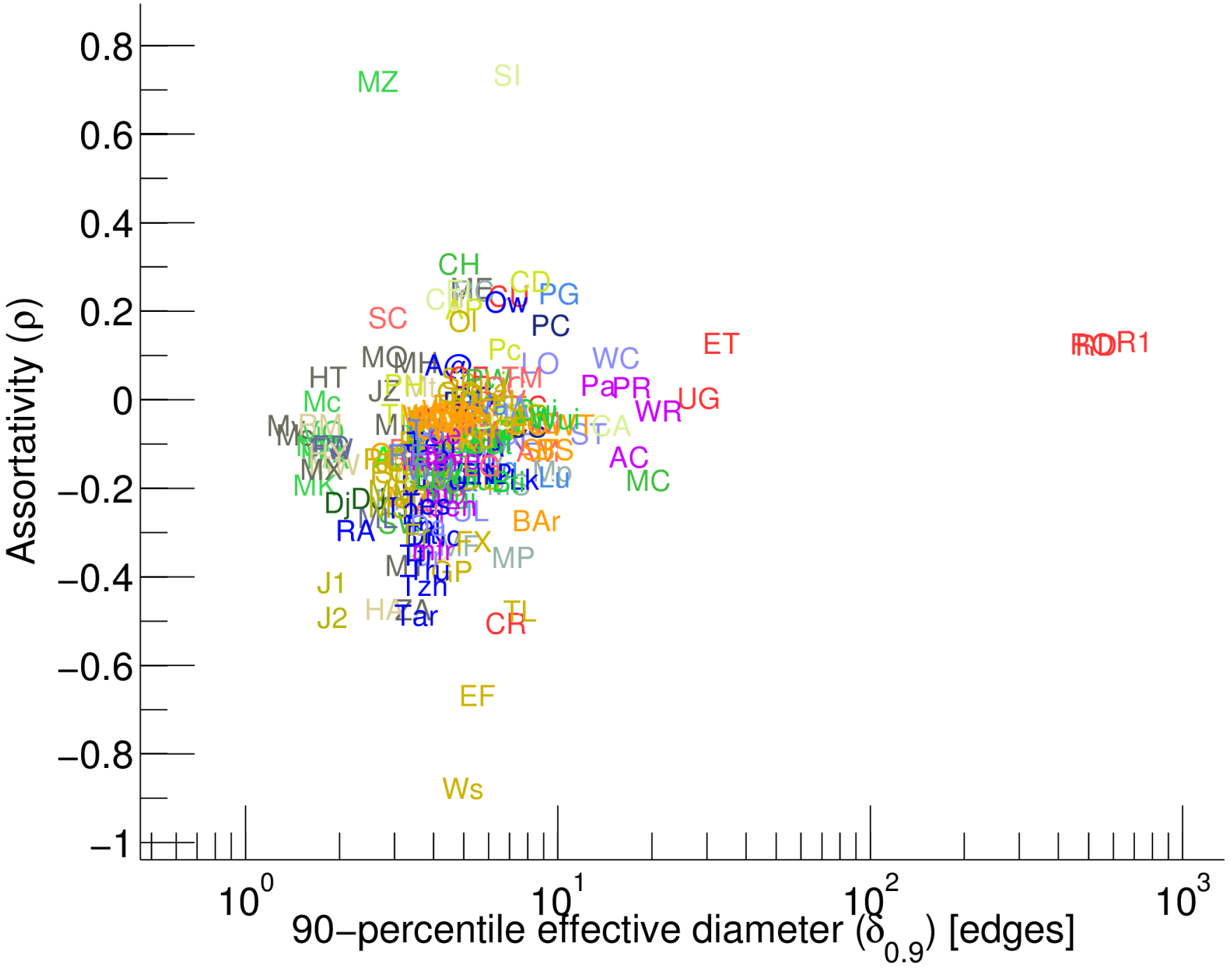
References

[1] Jérôme Kunegis. KONECT – The Koblenz Network Collection. In *Proc. Int. Conf. on World Wide Web Companion*, pages 1343–1350, 2013. [http]

[2] Wayne Zachary. An information flow model for conflict and fission in small groups. *J. of Anthropol. Res.*, 33:452–473, 1977.

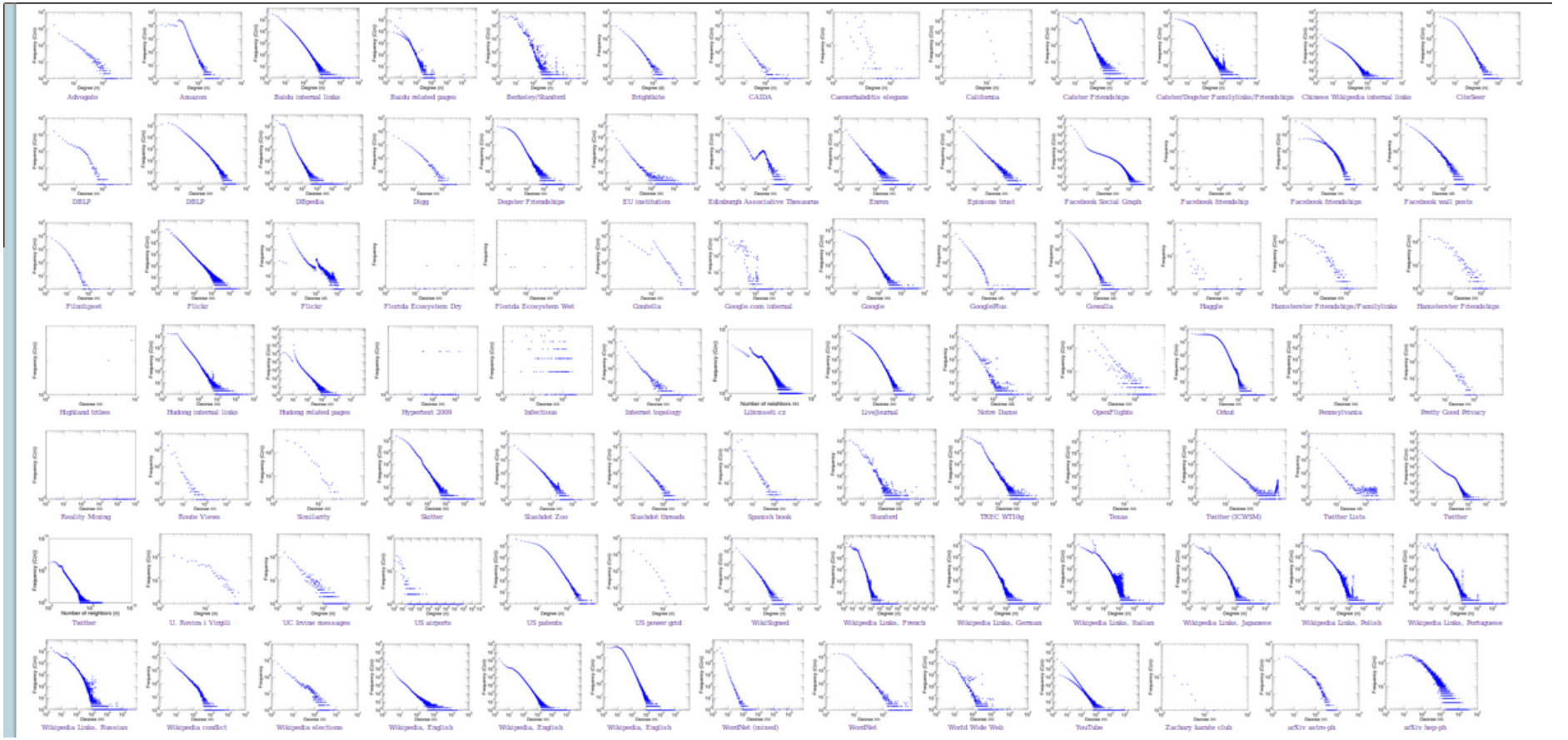
Network Comparison http://konect.cc/statistics/

Command: `stu @scatter.diameff90.assortativity`



Network Comparison: Plots

http://konect.cc/plots/

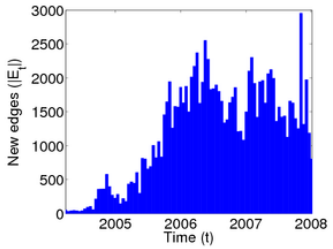


Example: Degree distribution

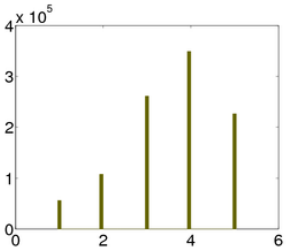
Command: `stu @degree`

More Plots

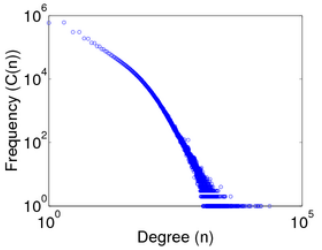
http://konect.cc/plots/



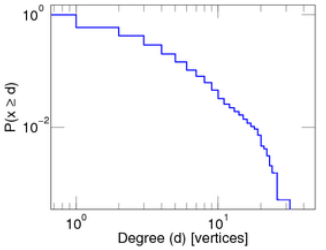
Temporal distribution



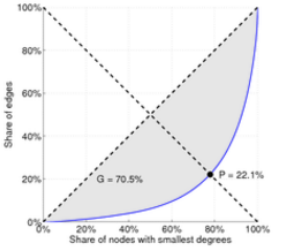
Edge weight and multiplicity distribution



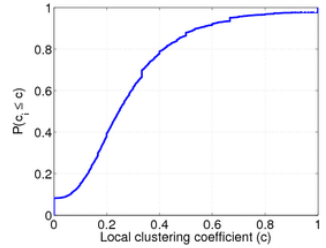
Degree distribution



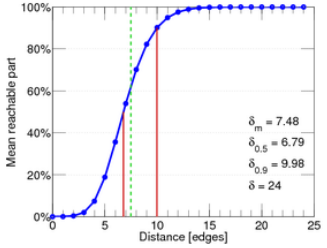
Cumulative degree distribution



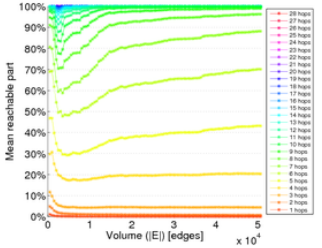
Lorenz curve



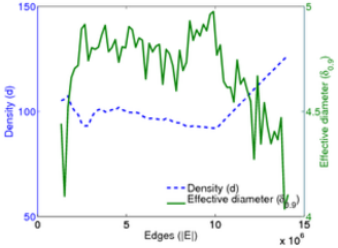
Clustering coefficient distribution



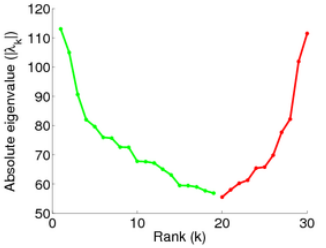
Hop plot



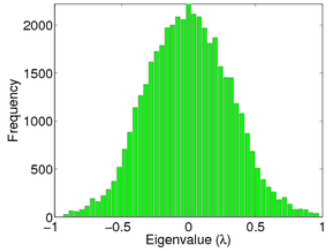
Temporal hop plot



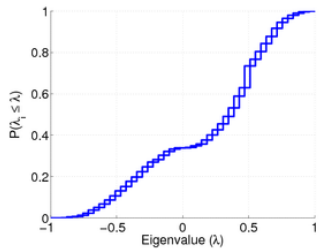
Evolution of average degree and diameter



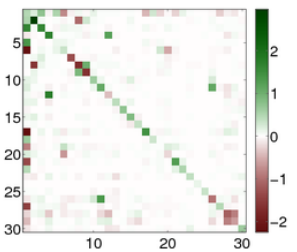
Top-*k* eigenvalues



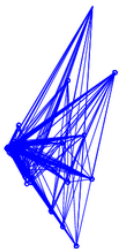
Spectral distribution



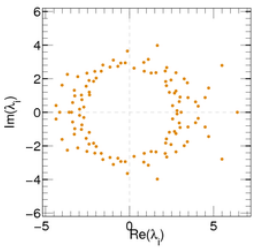
Cumulative spectral distribution



Spectral diagonality test



Network visualization



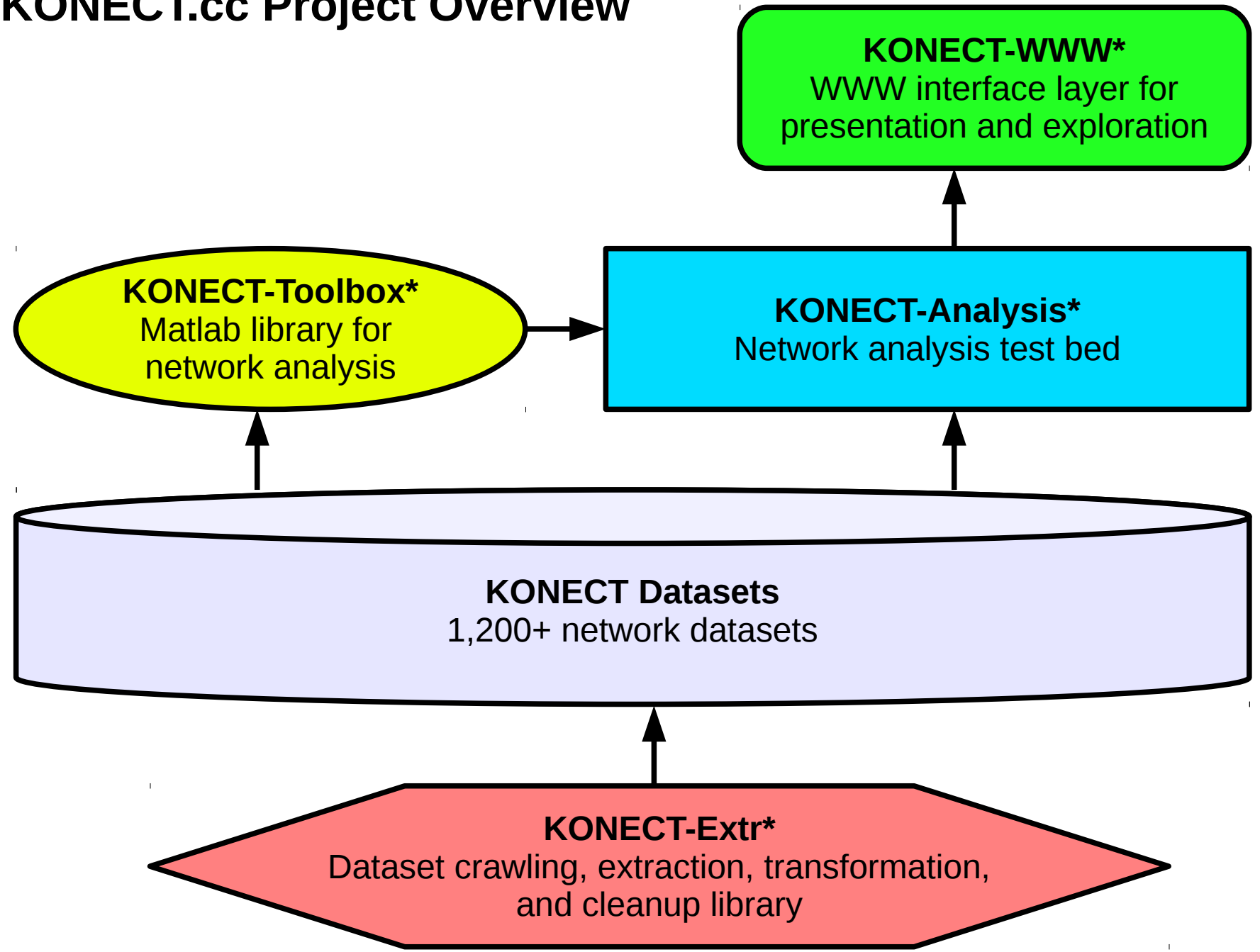
Complex eigenvalues of the asymmetric adjacency matrix

Download

http://konect.cc/networks/

Code	Name	Category	F.	W.	M.	Size	Volume	Avg. degree	Download
Ds	discogs_lstyle	Features	B	▬		244,147	5,255,950	21.56	
AN	Adjective-noun relationships	Lexical	B	−		194	425	3.94	
AD	Advogato	Social	D	+		6,551	51,332	7.84	
AM	Amazon	Contact	U	−		805,751	3,387,388	4.20	
AR	Amazon ratings	Ratings	B	*	⊘	3,376,972	5,838,041	2.72	
AP	arXiv astro-ph	Contact	U	−		37,544	396,160	10.55	
AC	arXiv cond-mat	Authorship	B	−		38,741	58,595	3.50	
PH	arXiv hep-ph	Contact	U	−		28,093	12,730,096	453.14	
PHc	arXiv hep-ph	Reference	D	−		60,366	421,578	6.98	
THc	arXiv hep-th	Reference	D	−		48,239	352,807	7.31	
TH	arXiv hep-th	Contact	U	−		22,906	11,209,366	489.32	
th	arXiv hep-th (KDD Cup)	Reference	D	−		27,770	352,807	12.70	
BAI	Baidu	Reference	D	▬		2,141,300	17,794,839	8.31	
BAr	Baidu	Reference	D	▬		415,641	3,284,387	7.90	
BS	Berkeley/Stanford	Reference	D	−		1,297,360	7,600,595	5.86	
Btl	BibSonomy ti	Folksonomy	B	▬	⊘	975,963	2,555,080	12.48	
Bul	BibSonomy ui	Folksonomy	B	▬	⊘	777,064	2,555,080	440.99	
But	BibSonomy ut	Folksonomy	B	▬	⊘	210,467	2,555,080	440.99	
BK	Brightkite	Social	U	−		58,228	214,078	3.68	
PM	Caenorhabditis elegans	Contact	U	▬		453	4,596	10.15	
IN	CAIDA	Physical	U	−		26,475	106,762	4.03	
RO	California	Physical	U	−		3,930,412	5,533,214	1.41	
Sc	Catster	Social	U	−		149,700	5,449,275	36.40	
Scd	Catster/Dogster	Social	U	−		624,127	15,705,337	25.16	
CS	CiteSeer	Reference	D	−		723,131	1,764,929	2.44	
Ctl	CiteULike ti	Folksonomy	B	▬	⊘	885,046	2,411,819	15.74	
Cul	CiteULike ui	Folksonomy	B	▬	⊘	734,464	2,411,819	106.18	
Cut	CiteULike ut	Folksonomy	B	▬	⊘	175,992	2,411,819	106.18	
CN	Countries	Affiliation	B	▬		512,761	557,587	1.09	
PI	DBLP	Reference	D	−		12,591	49,793	3.95	

KONECT.cc Project Overview



*** GitHub package**

Matlab Toolbox

https://github.com/kunegis/konect-toolbox

```
konect_dentropy.m
konect_diameff.m
konect_diammean.m
konect_effective_diameter.m
konect_eigl.m
konect_eign.m
konect_eigskew.m
konect_first_index.m
konect_fromto.m
konect_gini_direct.m
konect_gini.m
konect_hopdistr_ex.m
konect_hopdistr.m
konect_imageubu_complex.m
konect_imageubu.m
konect_jain.m
konect_join.m
konect_label_statistic.m
konect_map.m
konect_matrix.m
konect_mauc.m
konect_network_rank_abs.m
konect_normalize_additively.m
konect_normalized_entropy.m
konect_normalize_matrix.m
konect_normalize_rows.m
konect_order_dedicom.m
konect_own.m
```

Material

Handbook of Network Analysis

KONECT project

Jérôme Kunegis
University of Namur, Belgium
naXys – Namur Center for Complex Systems
konect.cc

November 4, 2017

Abstract

This is the handbook for the KONECT project, a scientific project to archive network datasets, compute systematic network theoretic statistics about them, visualize their properties, and provide corresponding data and Free Software tools to researc

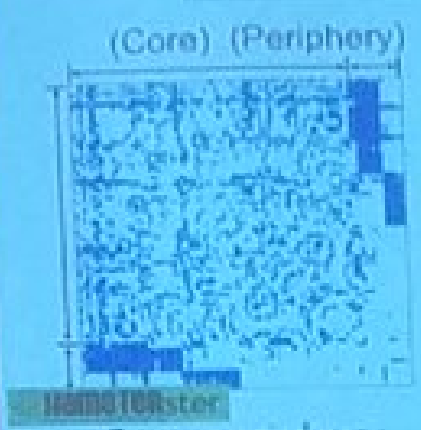
https://github.com/kunegis/konect-handbook

http://KONECT.cc/

KONECT Datasets in the Wild

Structural-Core Pattern: Pattern

Pattern 3: Structural-Core Pattern
Degeneracy-cores have structural patterns such as **core-periphery** and **communities**

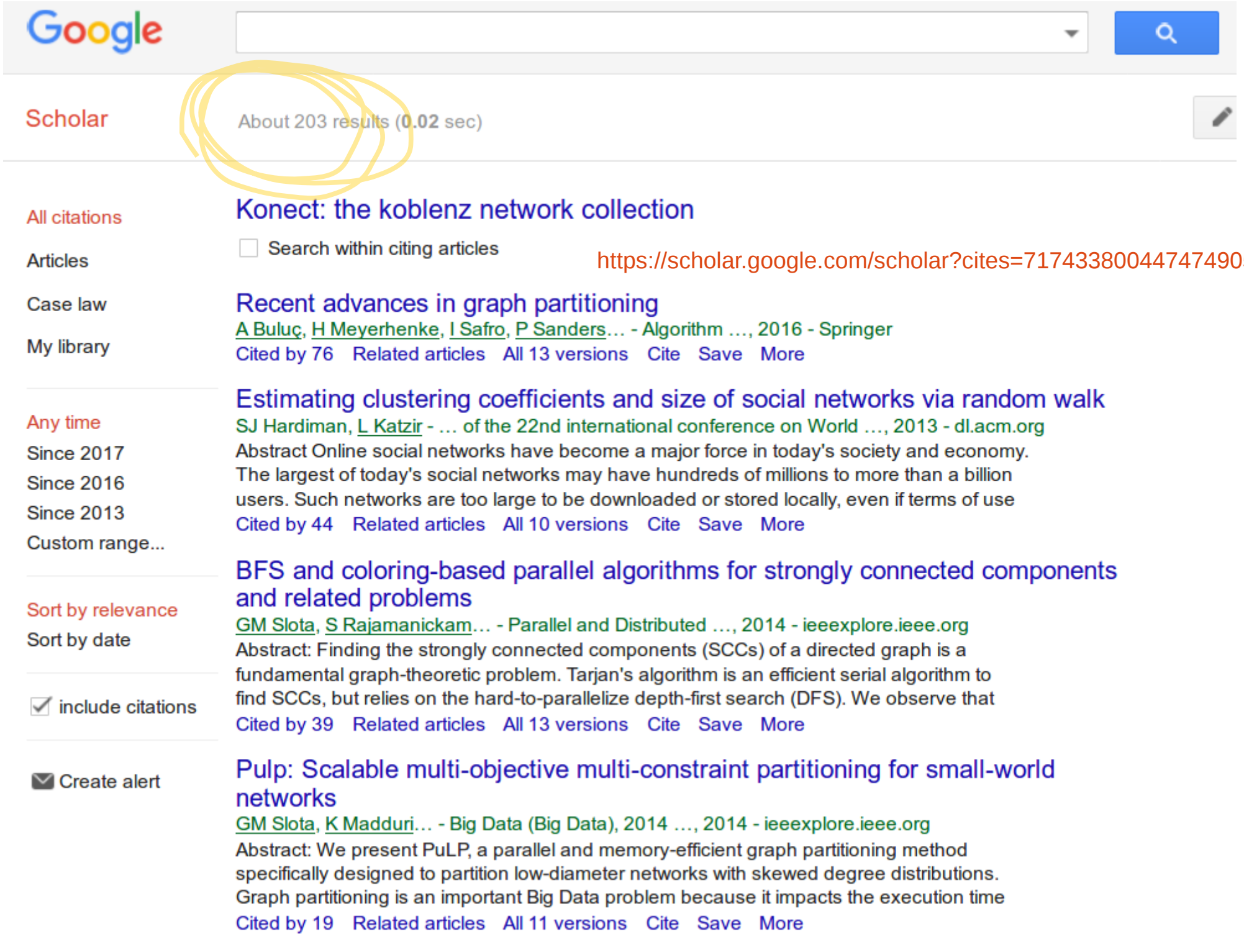


Core-periphery

Communities

Nodes in degeneracy-cores are not homogeneous

(Tina Eliassi-Rad, NetSci 2017)


Google
Scholar
About 203 results (0.02 sec)
All citations
Articles
Case law
My library
Any time
Since 2017
Since 2016
Since 2013
Custom range...
Sort by relevance
Sort by date
include citations
Create alert
Konect: the koblenz network collection
Search within citing articles
https://scholar.google.com/scholar?cites=7174338004474749050
Recent advances in graph partitioning
A Buluç, H Meyerhenke, I Safro, P Sanders... - Algorithm ..., 2016 - Springer
Cited by 76 Related articles All 13 versions Cite Save More
Estimating clustering coefficients and size of social networks via random walk
SJ Hardiman, L Katzir - ... of the 22nd international conference on World ..., 2013 - dl.acm.org
Abstract Online social networks have become a major force in today's society and economy. The largest of today's social networks may have hundreds of millions to more than a billion users. Such networks are too large to be downloaded or stored locally, even if terms of use
Cited by 44 Related articles All 10 versions Cite Save More
BFS and coloring-based parallel algorithms for strongly connected components and related problems
GM Slota, S Rajamanickam... - Parallel and Distributed ..., 2014 - ieeexplore.ieee.org
Abstract: Finding the strongly connected components (SCCs) of a directed graph is a fundamental graph-theoretic problem. Tarjan's algorithm is an efficient serial algorithm to find SCCs, but relies on the hard-to-parallelize depth-first search (DFS). We observe that
Cited by 39 Related articles All 13 versions Cite Save More
Pulp: Scalable multi-objective multi-constraint partitioning for small-world networks
GM Slota, K Madduri... - Big Data (Big Data), 2014 ..., 2014 - ieeexplore.ieee.org
Abstract: We present PuLP, a parallel and memory-efficient graph partitioning method specifically designed to partition low-diameter networks with skewed degree distributions. Graph partitioning is an important Big Data problem because it impacts the execution time
Cited by 19 Related articles All 11 versions Cite Save More

Learning Goals

- Everything is a network
- Take datasets
- Give datasets
- Increase the number of datasets you use
- Compare datasets
- Transfer knowledge across datasets
- Automatize everything

http://konect.cc/
https://github.com/kunegis/konect-handbook
https://github.com/kunegis/stu

We are on Twitter: @KONECTproject

Outline

- 1 Introduction
- 2 Statistics
- 3 Plots
- 4 Studies
  - (a) Diversity
  - (b) Graph generator
  - (c) Power laws
  - (d) Preferential attachment
  - (e) Conflict
- 5 Automatize everything : Stu

One More Note

This is a tutorial, not a paper talk – questions, comments and direct participation are welcome :)