COMPLEX NETWORKS ANALYSIS INTRODUCTION

Cazabet Rémy

PRESENTATION

- Remy Cazabet (“remi”)
- Associate professor in Computer Science at “Université de Lyon”
- Topics of research:
  - Network Science,
  - Data Mining,
  - Machine Learning,
  - Social Network Analysis,
  - Complex Systems, …

PRESENTATION

Lyon : 2nd city in France

UNIVERSITÉ DE LYON

UNIVERSITÉ DE LYON | Université Claude Bernard Lyon 1

PRESENTATION

What about you ?

COURSE ORGANIZATION

- Every day, 2h lectures, 2h practicals.
- We learn a new topic, we apply it on example graphs.
- You can come with your own data. There are many websites with repositories of "interesting" graphs,
  - http://networkrepository.com
  - Marvel, TV series, economics, soccer…

COURSE ORGANIZATION

- Gradation for every week
- End of first week:
  - Send a report on the analysis of a graph you have chosen according to what we have studied (What you think is relevant)
- End of last week:
  - Send a report on the analysis of a DYNAMIC graph according to what we have studied.
- One part of the report should be a Jupyter Notebook

INTRODUCTION

GRAPH OR NETWORKS

- What you have seen last week:
  - Graph theory => Efficient algorithms, complexity analysis, proofs…
- What we will see together:
  - How to make data "speak"
  - Not any kind of data: relational ones, modeled by **networks**

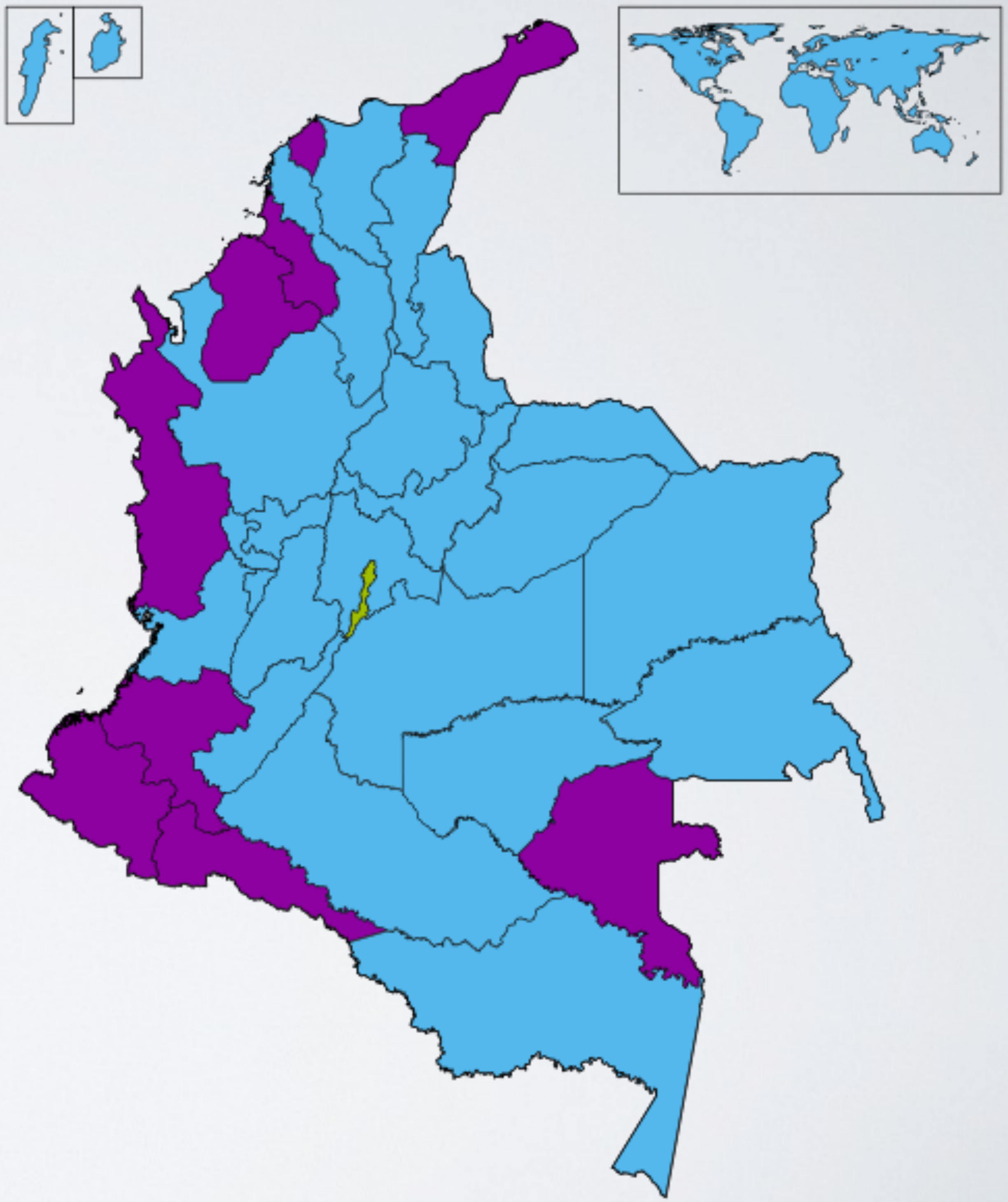
CONTEXT

- Big data, data science, data mining, machine learning, artificial intelligence ….
- Input: Data
- Output:
  - Knowledge
  - Model
  - Prediction

CONTEXT

- Let's take an example: Colombian elections
- Data:
  - Results (by geographical regions)
  - Polls before the vote
  - Surveys: Age, genre, income, marital status, etc.
  - …

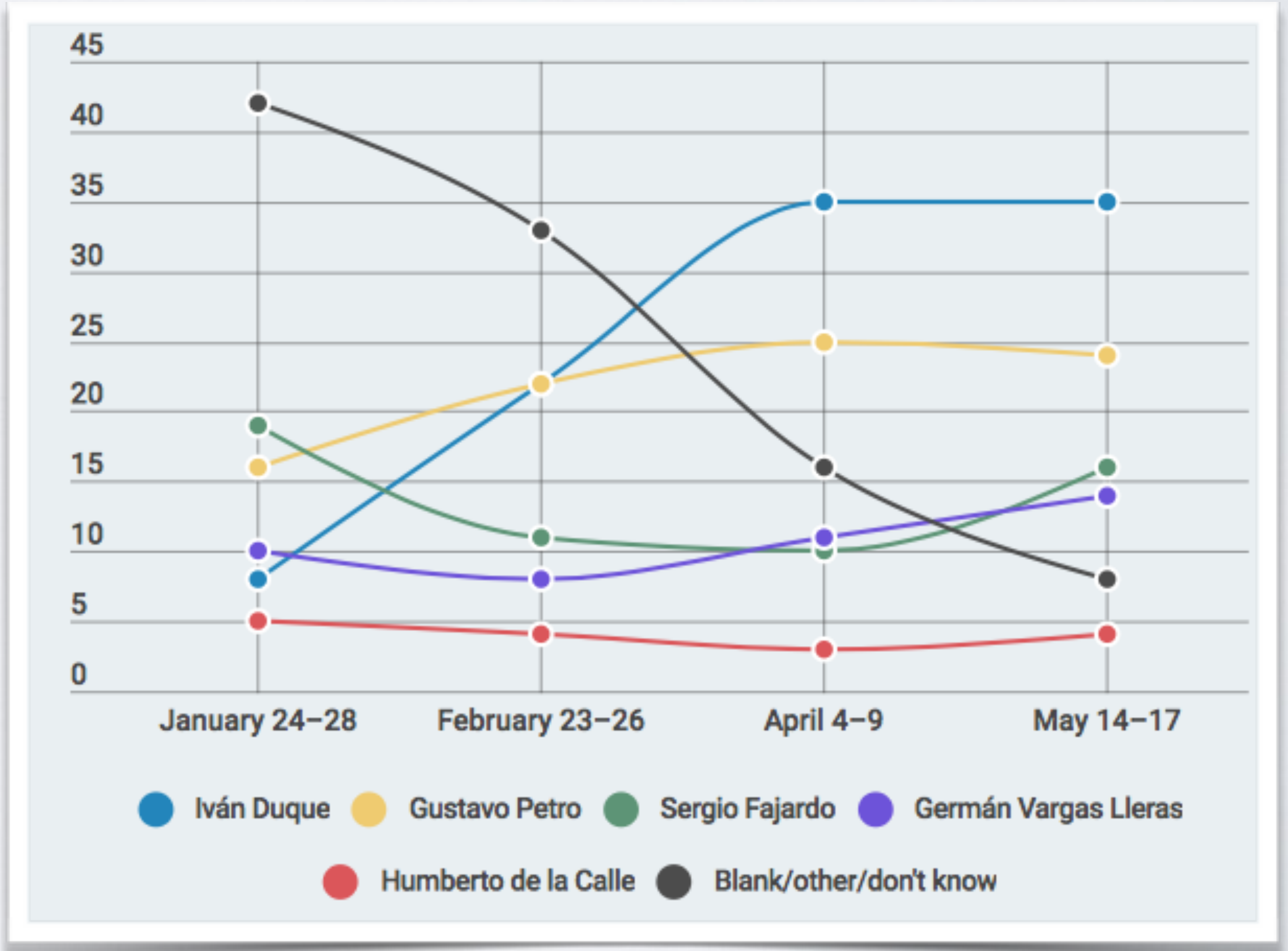
CONTEXT



- Acquiring Knowledge:
  - Geographical disparities
  - Opinions of social classes
  - Long term evolution of the society
  - …

CONTEXT

- Predicting:
  - Time series analysis: predict the futur given trends
  - Predict the vote of a person given its profile
  - Predict how societal evolutions will affect votes
  - …



CONTEXT

- Data oriented decision making/analysis is now ubiquitous:
  - Finance
  - Sport (money game…)
  - Industry (Predictive Maintenance, Supply chain optimisation…)
  - Politics (Cambridge analytica..)

- And Data-Oriented applications continues to expand
  - Self driving cars (data, data, data)
  - Smart cities
  - Physics, Biology, Medicine, …

GRAPHS ?

- Coming back to Colombian elections
  - What information could we add besides features describing each individual ?
  - =>Adding relational data
  - Who is a relative (daughter/sister/grandmother/…) of whom ?
  - Who is a friend of whom ?
  - Who works in the same company ?
  - …
- Tell me who your friends are and I'll tell you who you are
- Knowledge/Opinions propagates and form "social networks"

GRAPHS ?

- “But this information is much harder to obtain than individual ones… right ?”
- On the contrary ! Social Media !
- +, why not, cell phone, emails, WhattsApp, …

GRAPHS ?

- Graphs can also represent any type of data:
  - Step1) Compute correlations between elements
  - Step2) Filter out low values
  - Step3) You have a graph !
- Often used to scale algorithms (DBscan…)
- Or to apply network analysis tools
- (More on that later)

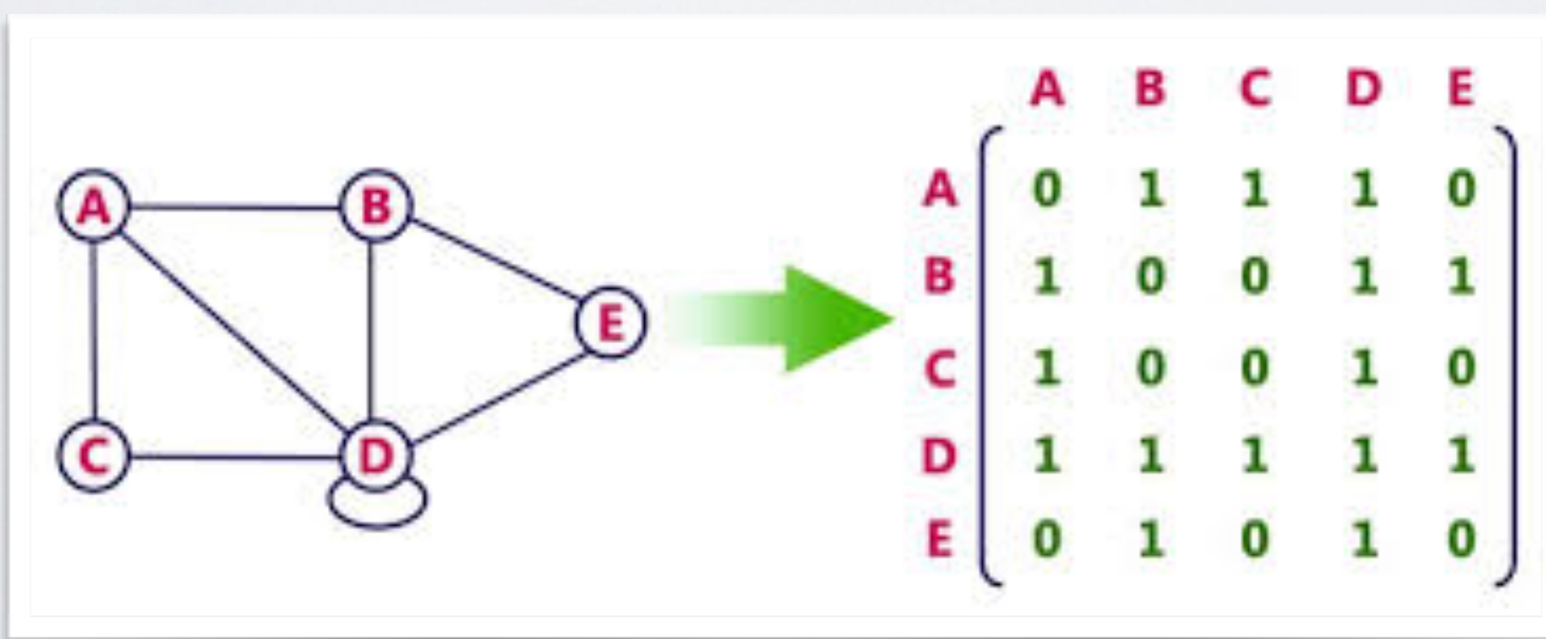
GRAPHS ?

- What is so special about graphs ? Isn't it a feature like any other ?

- Classical data mining/machine learning can be summarized as:
  - An item is described as a VECTOR: [x1,x2,x3,…,xN]
  - We learn sequences of operations on these vectors to predict something
    - IF age>X and income>Y and city in […] THEN Vote=Mr. XXX
  - If your feature is not numeric, you transform it to numbers.
    - For instance: department= NAME
    - Some methods can handle them directly (decision trees, …)
    - Or transformation to vector:
    - 30 departments: Each person has a vector with 29 zeros and a 1

GRAPHS ?

- A graph can be represented as:
  - A list of edges : [{v1,v2}, {v1,v3}, {v5,v7},…]
  - A neighborhood list: {v1:{v2,v3},v2:{v1},v5:{v7},…}
  - An adjacency matrix



GRAPHS ?

- We could use a line of the adjacency matrix as feature vector
- It does not work because:
  - Sparsity: too many 0s
  - Curse of dimensionality
  - Similar features means similar item. Not for adj. matrix:
    - It means connected to the same node
    - What is interesting in graphs is elsewhere: not only direct neighbors

GRAPHS ?

- Field of **Network Science**
- Contributions from physicists, computer scientists/Engineers and mathematicians (beyond traditional scientific fields)
- For me, a "tool" for all scientists, like probabilities, spectral analysis or machine learning
- For computer science: related to ML, DM. Same level as Natural Language Processing, maybe

GRAPHS ?

- Graphs or networks?
- I use both terms interchangeably
- **Graph theory:** older field (env. 70 years), mostly theoretical, studying properties of graphs (usually synthetic) and algorithms on graphs
- **Network Science:** born from graph theory (env. 10 years), interested in real networks, with both theory and applications
- **Social Network Analysis:** Older term than network science (env. 40 years), network science on SN

CHAPTER I
DESCRIBING A NETWORK AT THE GLOBAL SCALE

SIZE

- A network is composed of nodes and edges.
- Size: How many nodes and edges ?

	#nodes	#edges
Wikipedia HL	2M	30M
Twitter 2015	288M	60B
Facebook 2015	1.4B	400B
Brain c. Elegans	280	6393
Roads US	129k	165k
Airport traffic	3k	31k

DENSITY

Defined as:

Directed $D = \frac{|E|}{|V|\,(|V|-1)}$

Undirected $D = \frac{2|E|}{|V|\,(|V|-1)}$

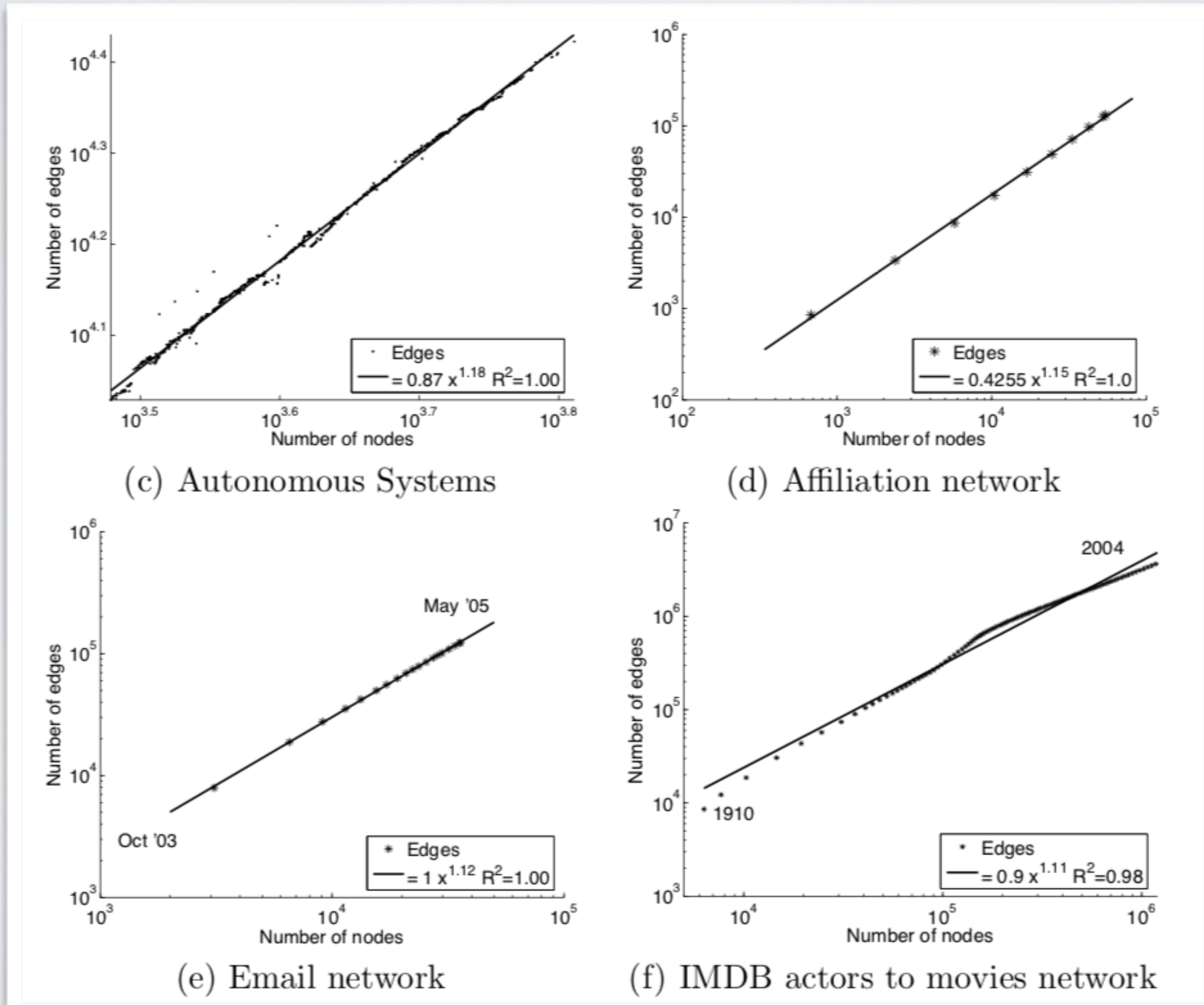
Often more relevant: average degree (2|E| / |V|)

	#nodes	#edges	Density	avg. deg
Wikipedia	2M	30M	1.5×10^{-5}	30
Twitter 2015	288M	60B	1.4×10^{-6}	416
Facebook	1.4B	400B	4×10^{-9}	570
Brain c.	280	6393	0.16	46
Roads Calif.	2M	2.7M	6×10^{-7}	2.7
Airport	3k	31k	0.007	21

DENSITY

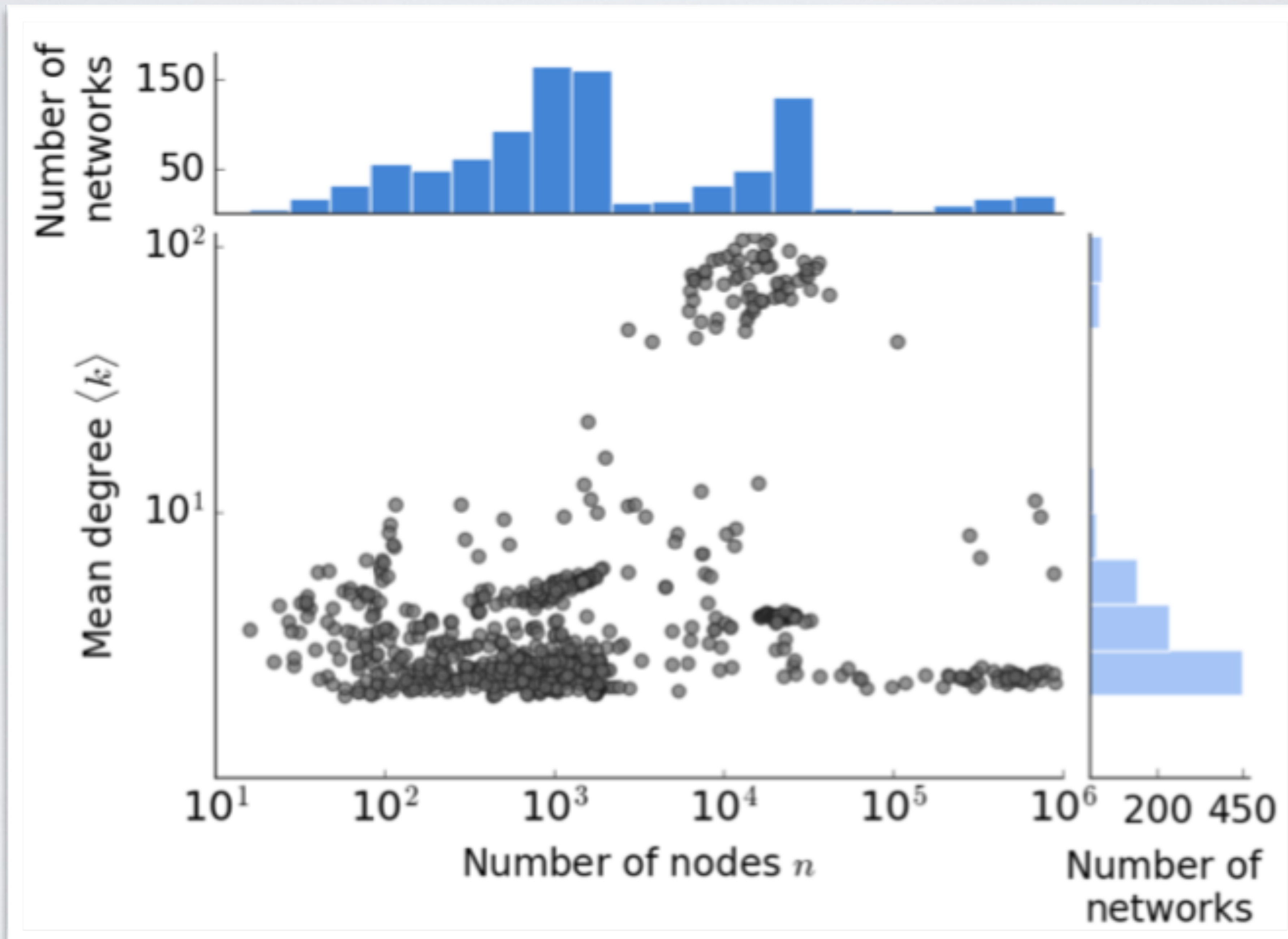
- It has been observed that: [Leskovec. 2006]
  - When graphs increase in size, the average degree increases
  - This increase is very slow
- Think of friends in a social network

DENSITY



Number of edges
Number of nodes
Edges
= 0.87 x^1.18 R^2=1.00
(c) Autonomous Systems
Edges
= 0.4255 x^1.15 R^2=1.0
(d) Affiliation network
May '05
Oct '03
Edges
= 1 x^1.12 R^2=1.00
(e) Email network
2004
1910
Edges
= 0.9 x^1.11 R^2=0.98
(f) IMDB actors to movies network

DENSITY

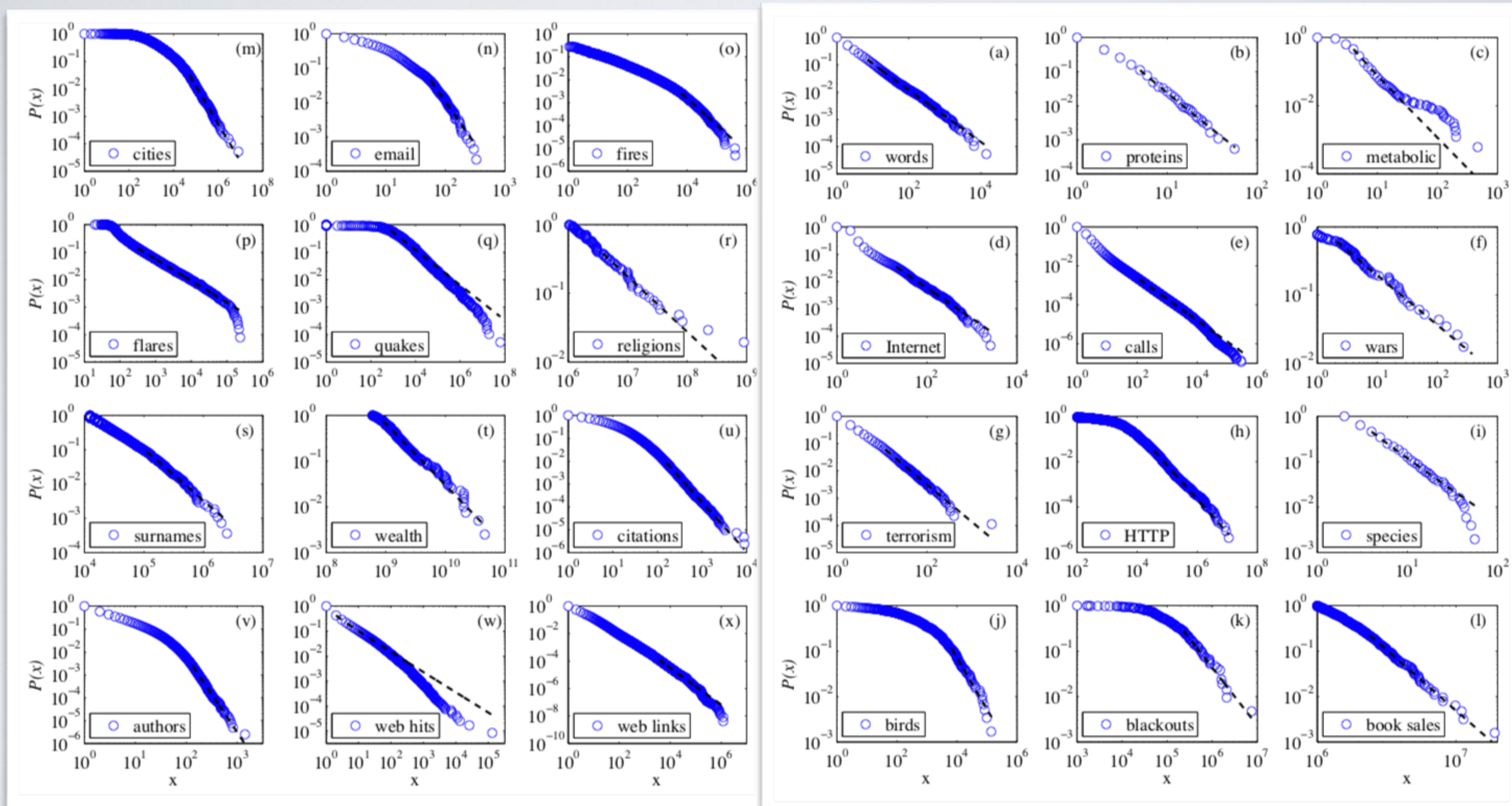


[Broido, Clauset 2018]

DEGREE DISTRIBUTION

- In a fully random graph (Erdos-Renyi), degree distribution is a normal distribution centered on the average degree
- In real graphs, in general, it is not the case:
    - A high majority of small degree nodes
    - A small minority of nodes with very high degree (Hubs)
- Often modeled by a **power law**

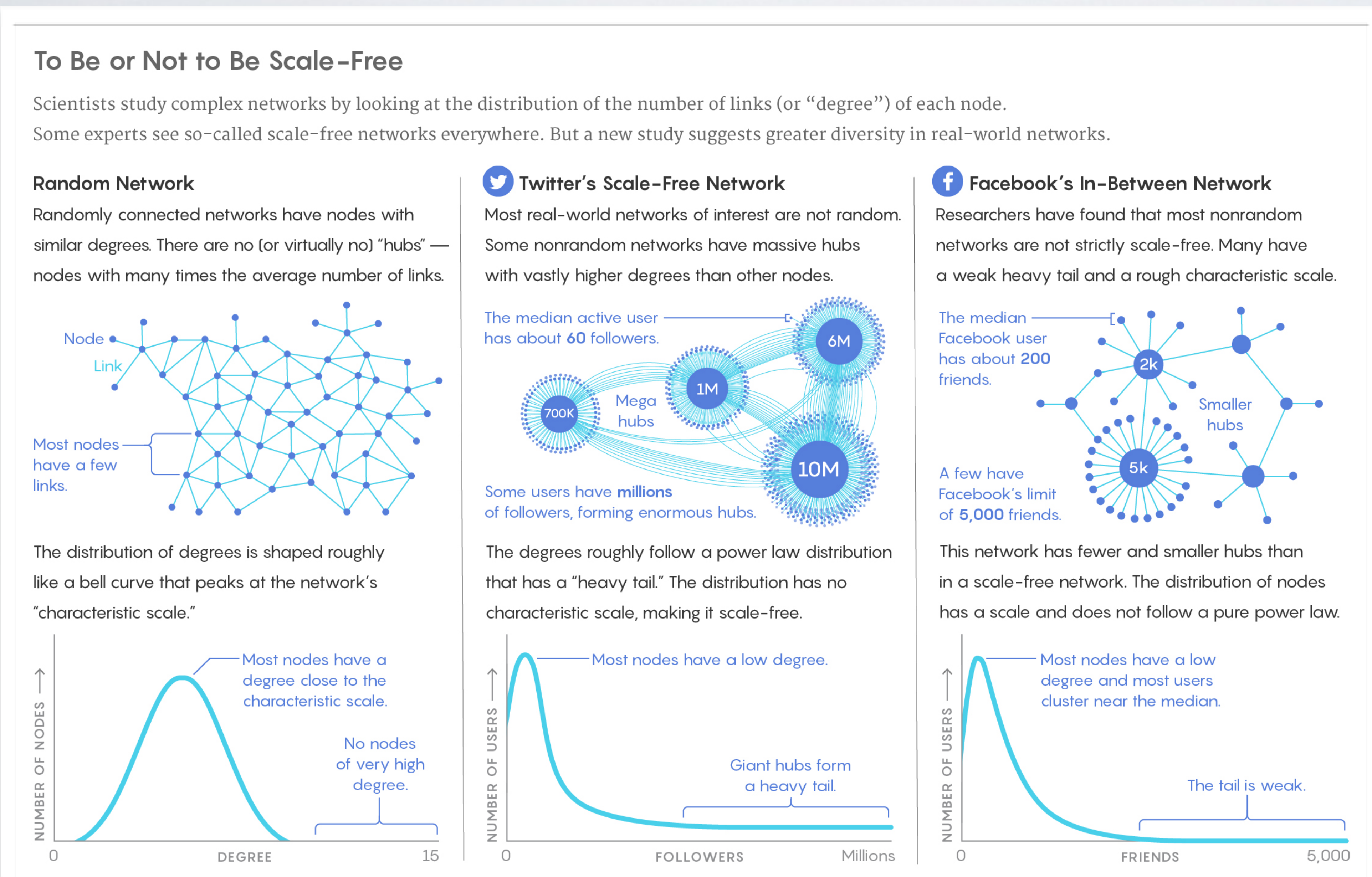
DEGREE DISTRIBUTION



[Clauset 2009]

DEGREE DISTRIBUTION

Power law/Scale free distribution: $f(x) = ax^{-k}$



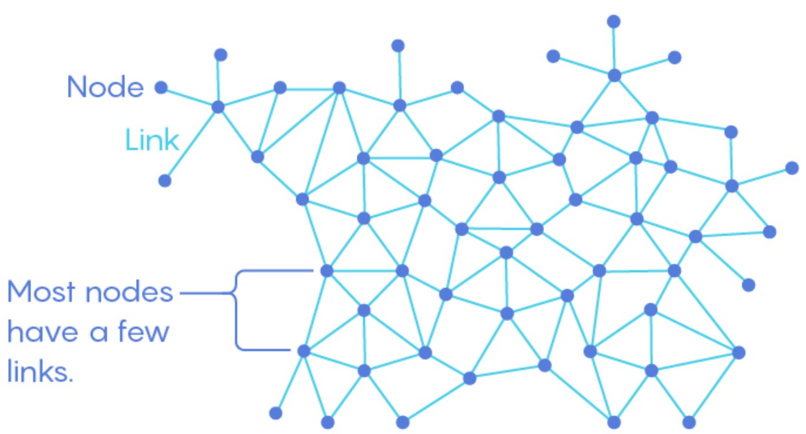
[Quanta magazine 2018]

To Be or Not to Be Scale-Free

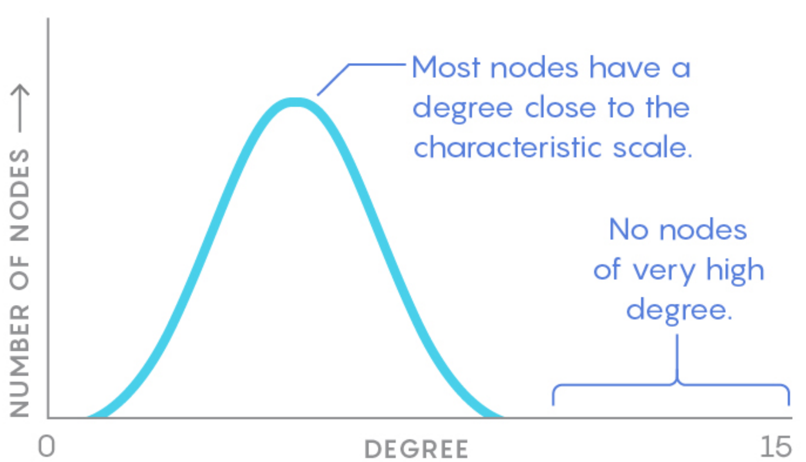
Scientists study complex networks by looking at the distribution of the number of links (or "degree") of each node. Some experts see so-called scale-free networks everywhere. But a new study suggests greater diversity in real-world networks.

Random Network

Randomly connected networks have nodes with similar degrees. There are no (or virtually no) "hubs" — nodes with many times the average number of links.

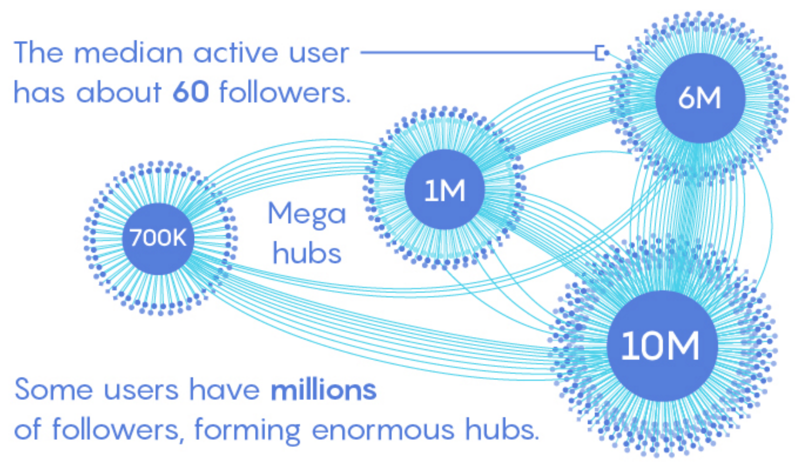


The distribution of degrees is shaped roughly like a bell curve that peaks at the network's "characteristic scale."

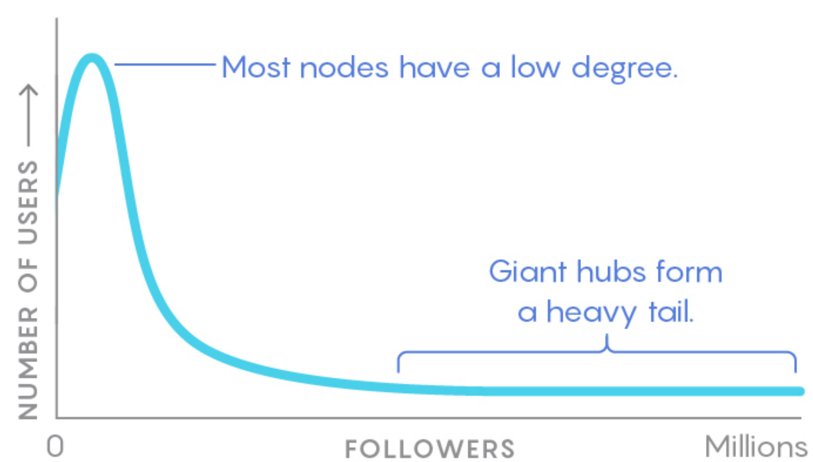


Twitter's Scale-Free Network

Most real-world networks of interest are not random. Some nonrandom networks have massive hubs with vastly higher degrees than other nodes.

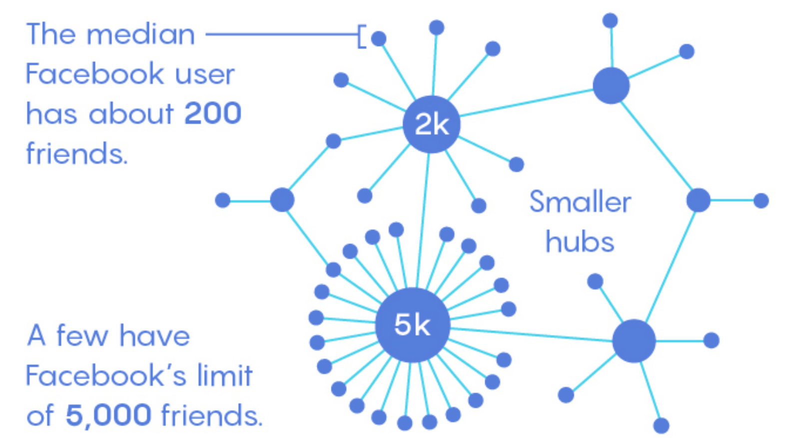


The degrees roughly follow a power law distribution that has a "heavy tail." The distribution has no characteristic scale, making it scale-free.

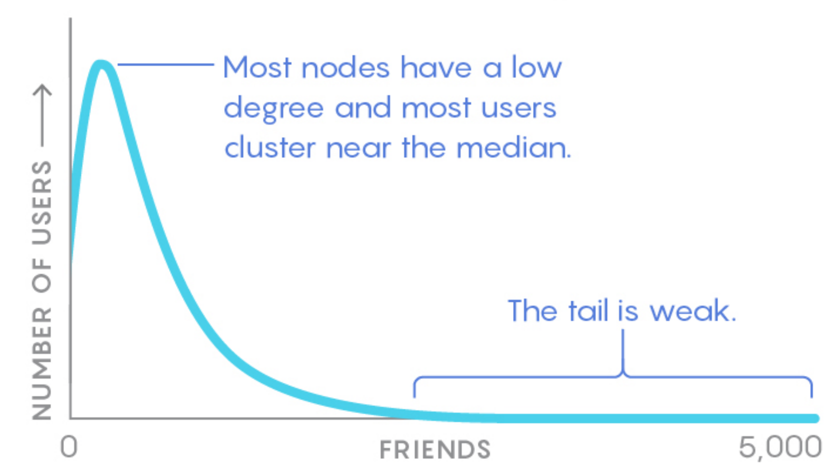


Facebook's In-Between Network

Researchers have found that most nonrandom networks are not strictly scale-free. Many have a weak heavy tail and a rough characteristic scale.



This network has fewer and smaller hubs than in a scale-free network. The distribution of nodes has a scale and does not follow a pure power law.



DEGREE DISTRIBUTION

- This has important implications:
  - There is no "scale" in the degree: the average degree is not representative
  - It is not realistic to use "random graphs" (ER) for evaluating algorithms performance
- If the degree distribution is not a power law, some algorithms might not behave as expected (spatial networks…)

CLUSTERING COEFFICIENT

Global clustering coefficient

$$C = \frac{\text{number of closed triplets}}{\text{number of all triplets (open and closed)}}.$$

Triplet: set of 3 nodes connected by 2 or 3 edges

Average Clustering Coefficient

Clustering coefficient of a node: $$C_i = \frac{2|\{e_{jk} : v_j, v_k \in N_i, e_{jk} \in E\}|}{k_i(k_i - 1)}$$

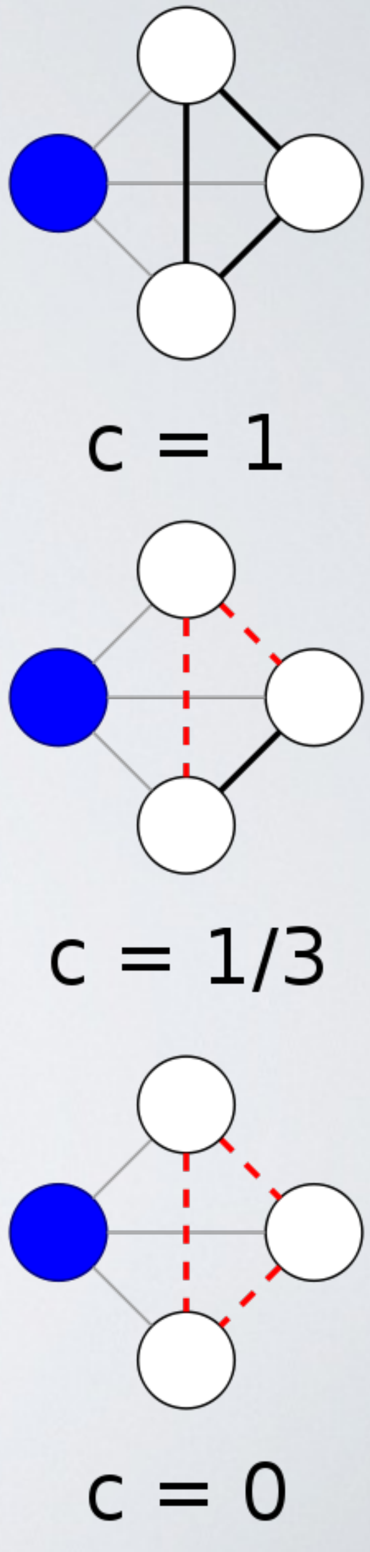
Average CC: $$\bar{C} = \frac{1}{n}\sum_{i=1}^{n} C_i$$

CLUSTERING COEFFICIENT

The higher the value,
the more **locally dense** is the network.

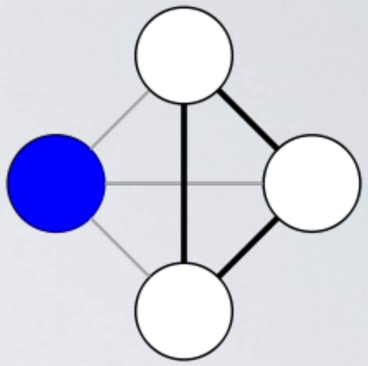
"Friends of my friends are my friends"

Higher in real networks than random

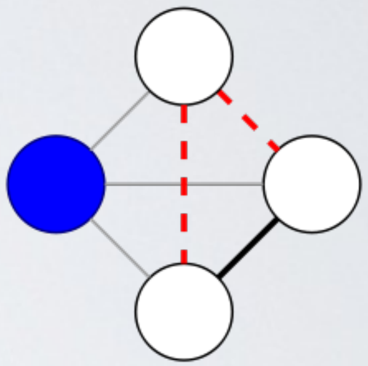


CLUSTERING COEFFICIENT

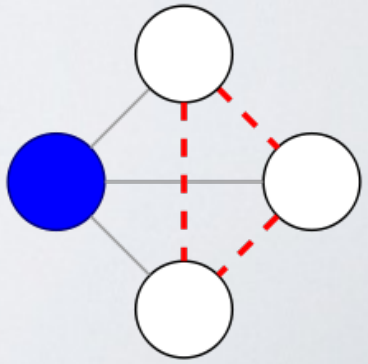
- Facebook ego-networks: 0.6
- Twitter lists: 0.56
- California Road networks: 0.04
- Random (ER): =density: very small for large graphs



c = 1



c = 1/3



c = 0

CONNECTED COMPONENTS

- A connected component: a group of nodes all mutually reachable
- Most real networks:
  - A "Giant connected component" including >99% nodes
  - A few small connected components
- E.g.: Facebook 2011: 99.91%

DIAMETER

- Shortest path between nodes *u* and *v*: minimal number of hops between them.
- Diameter: the longest shortest path in the network
- Very sensible to outliers, not reliable

AVERAGE PATH LENGTH

- Average shortest path between all pairs of nodes
- The famous 6 degrees of separation (Milgram experiment)
  - In fact 6 hops
  - (More on that next slide)
- Not too sensible to noise
- Tells your if the network is "stretched" or "hairball" like

SIDE-STORY: MILGRAM EXPERIMENT

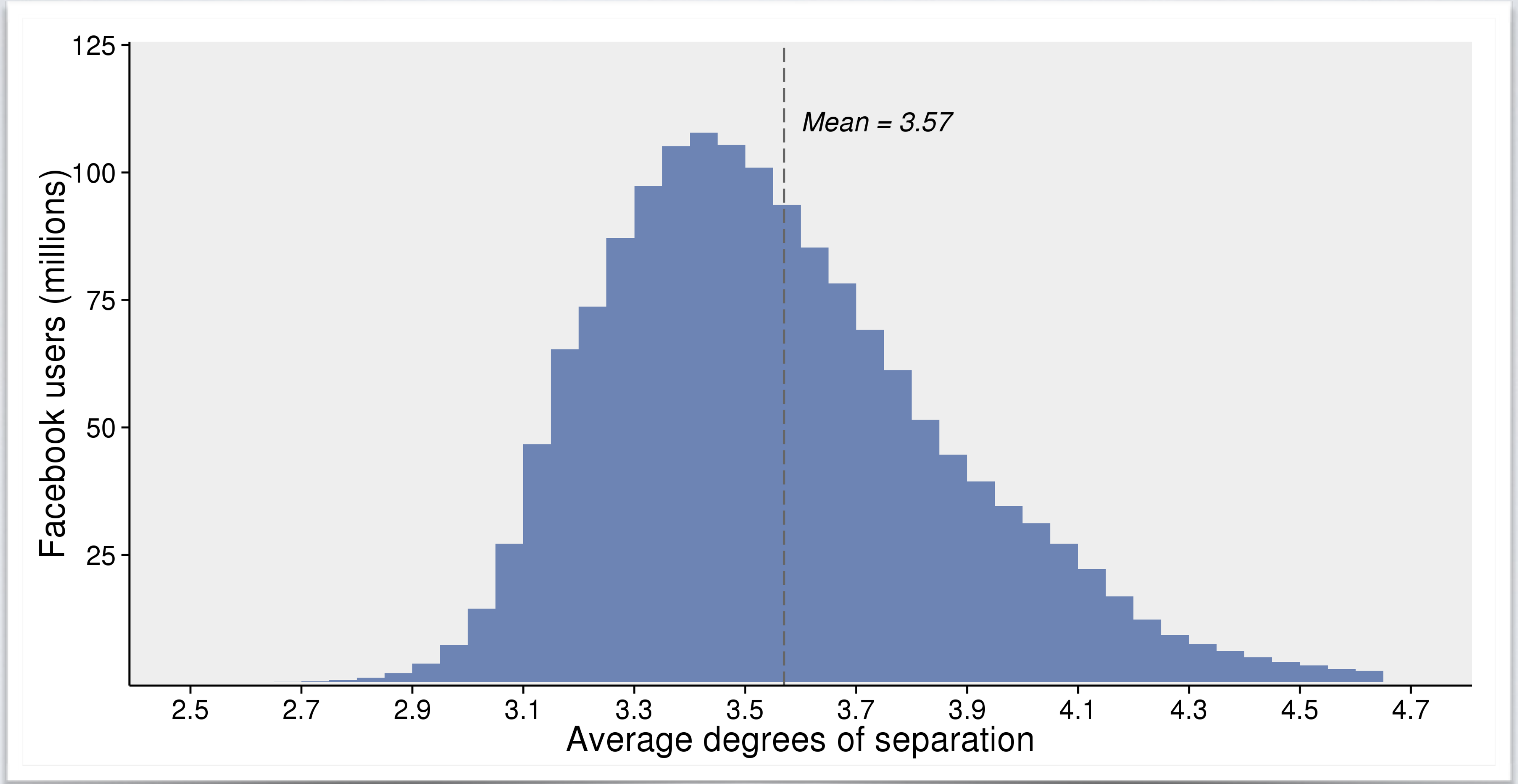
- Small world experiment (60's)
  - Give a (physical) mail to random people
  - Ask them to send to someone they don't know
    - They know his city, job
  - They send to their most relevant contact
- Results: In average, 6 hops to arrive

SIDE-STORY: MILGRAM EXPERIMENT

- Many criticism on the experiment itself:
  - Some mails did not arrive
  - Small sample
  - …

- Checked on “real” complete graphs (giant component):
  - MSN messenger
  - Facebook
  - The world wide web
  - …

SIDE-STORY: MILGRAM EXPERIMENT



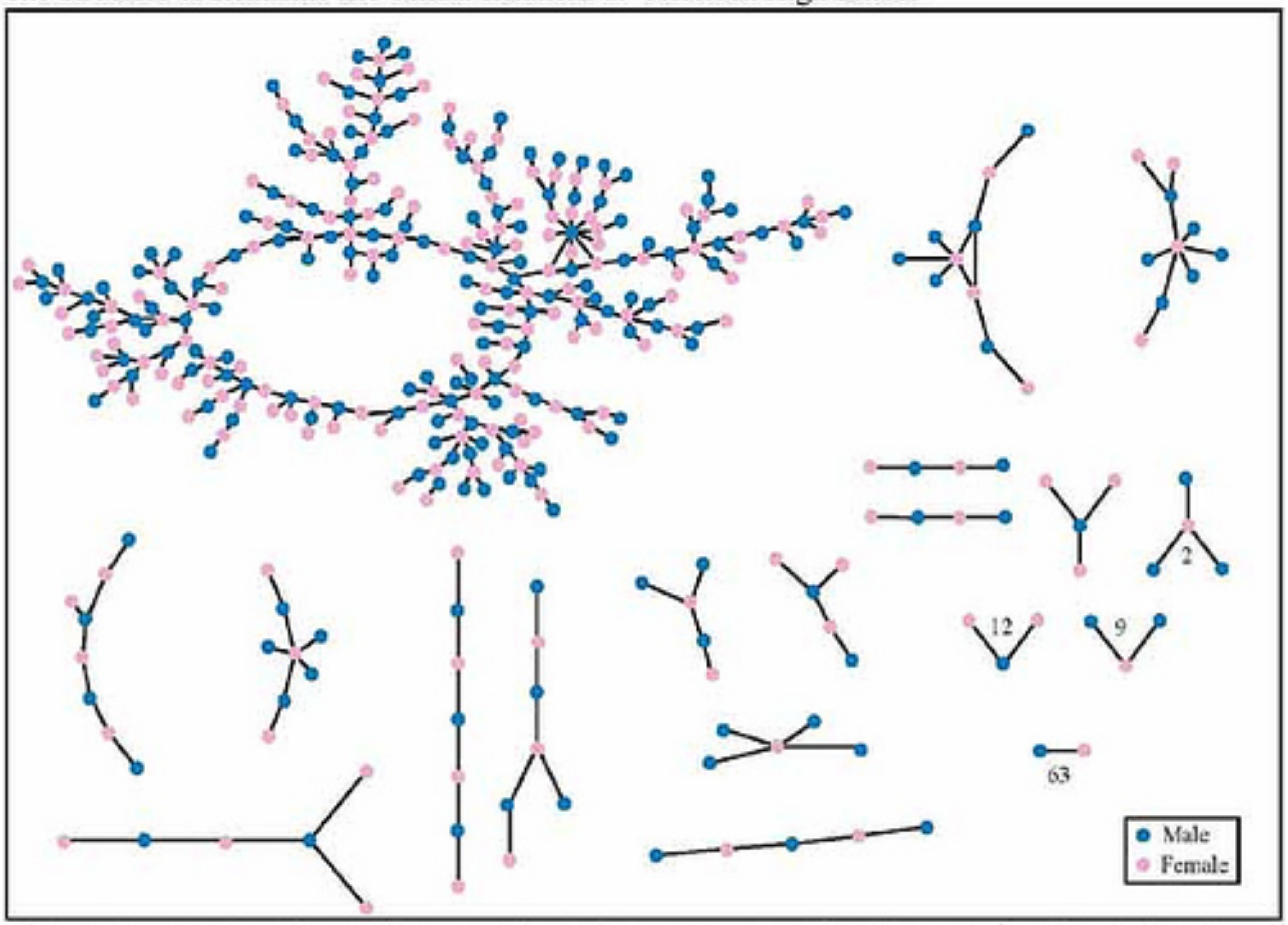
Facebook

HOMOPHILY/ASSORTATIVITY

- Nodes might have a preference for some other nodes
  - Similar nodes (age in social networks)
  - Different nodes (genre in sentimental networks (yes, it has been done!))
  - Nodes with a particular property
- "Assortativity" alone often used to mean "degree assortativity"
  - Large nodes are preferentially connected to large nodes
- All this implies: "compared with a random network"

HOMOPHILY/ASSORTATIVITY

The Structure of Romantic and Sexual Relations at "Jefferson High School"



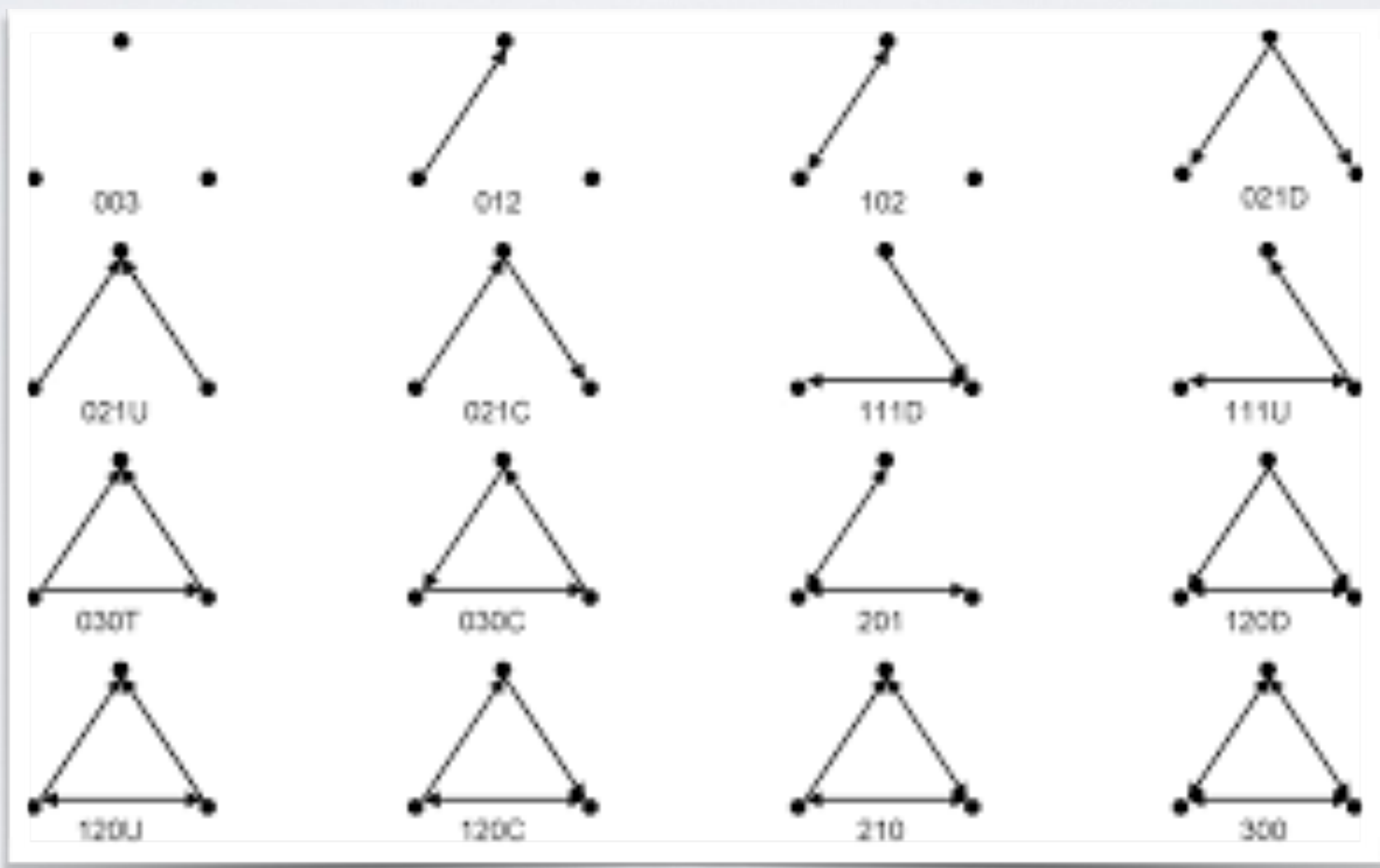
Each circle represents a student and lines connecting students represent romantic relations occuring within the 6 months preceding the interview. Numbers under the figure count the number of times that pattern was observed (i.e. we found 63 pairs unconnected to anyone else).

HOMOPHILY/ASSORTATIVITY

- Nodes might have a preference for some other nodes
  - Similar nodes (age in social networks)
  - Different nodes (genre in sentimental networks (yes, it has been done!))
  - Nodes with a particular property
- "Assortativity" alone often used to mean "degree assortativity"
  - Large nodes are preferentially connected to large nodes
- All this implies: "compared with a random network"

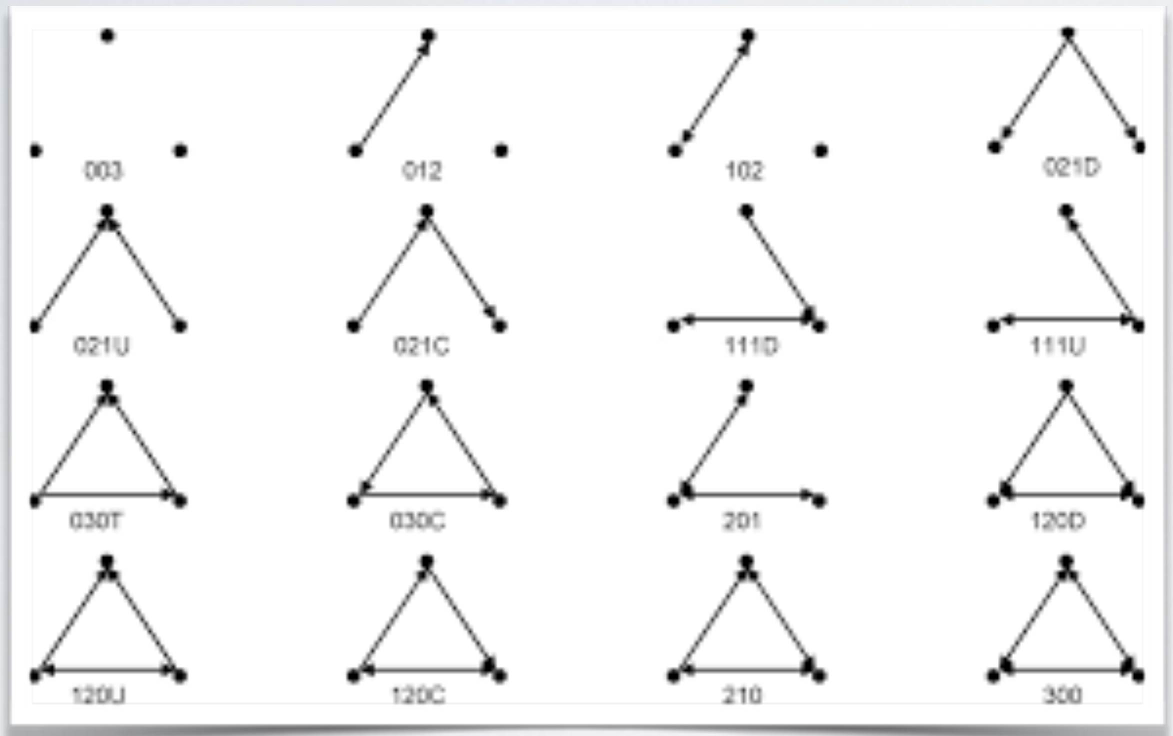
OTHER (A FEW EXAMPLES)

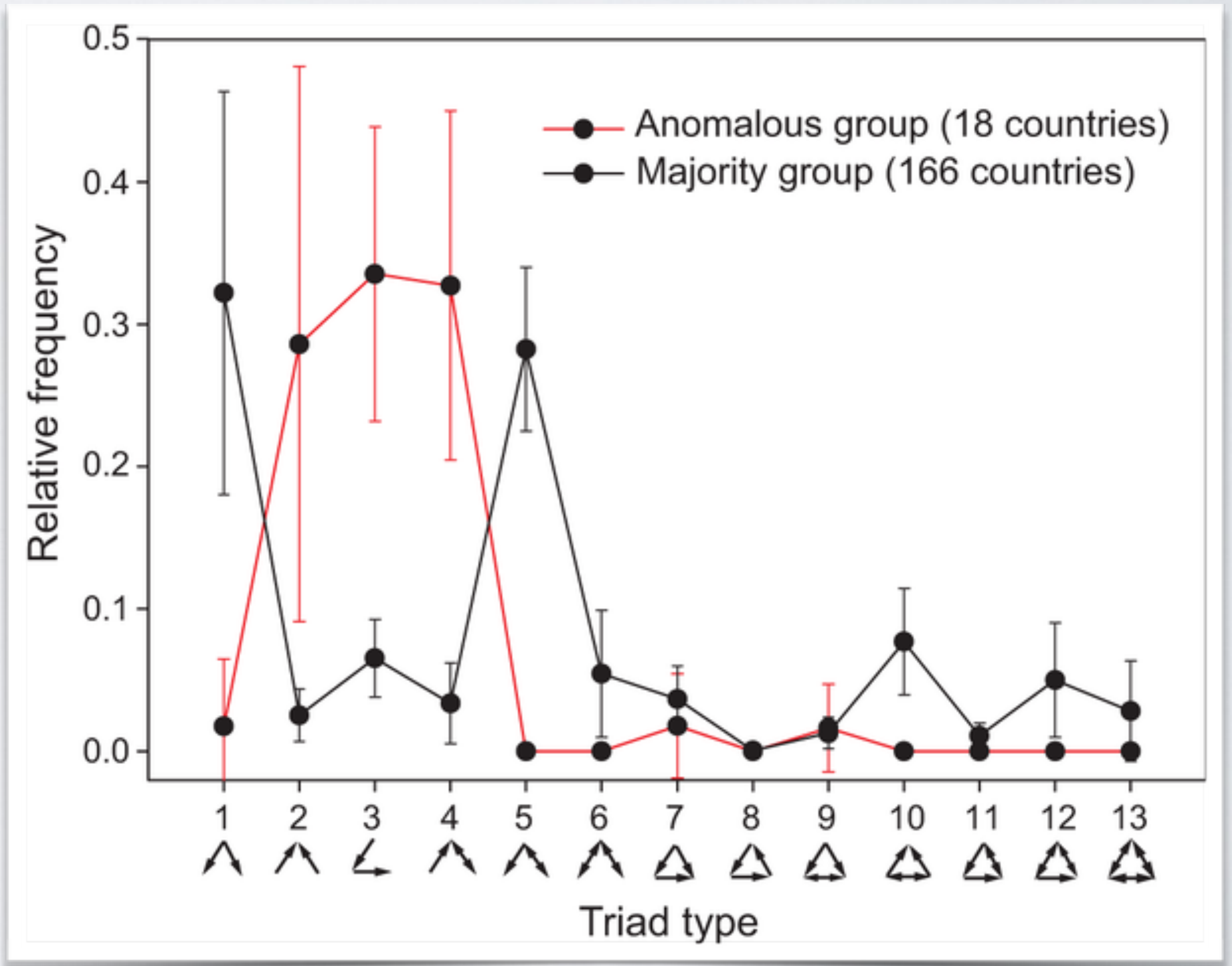
Triads counting



OTHER

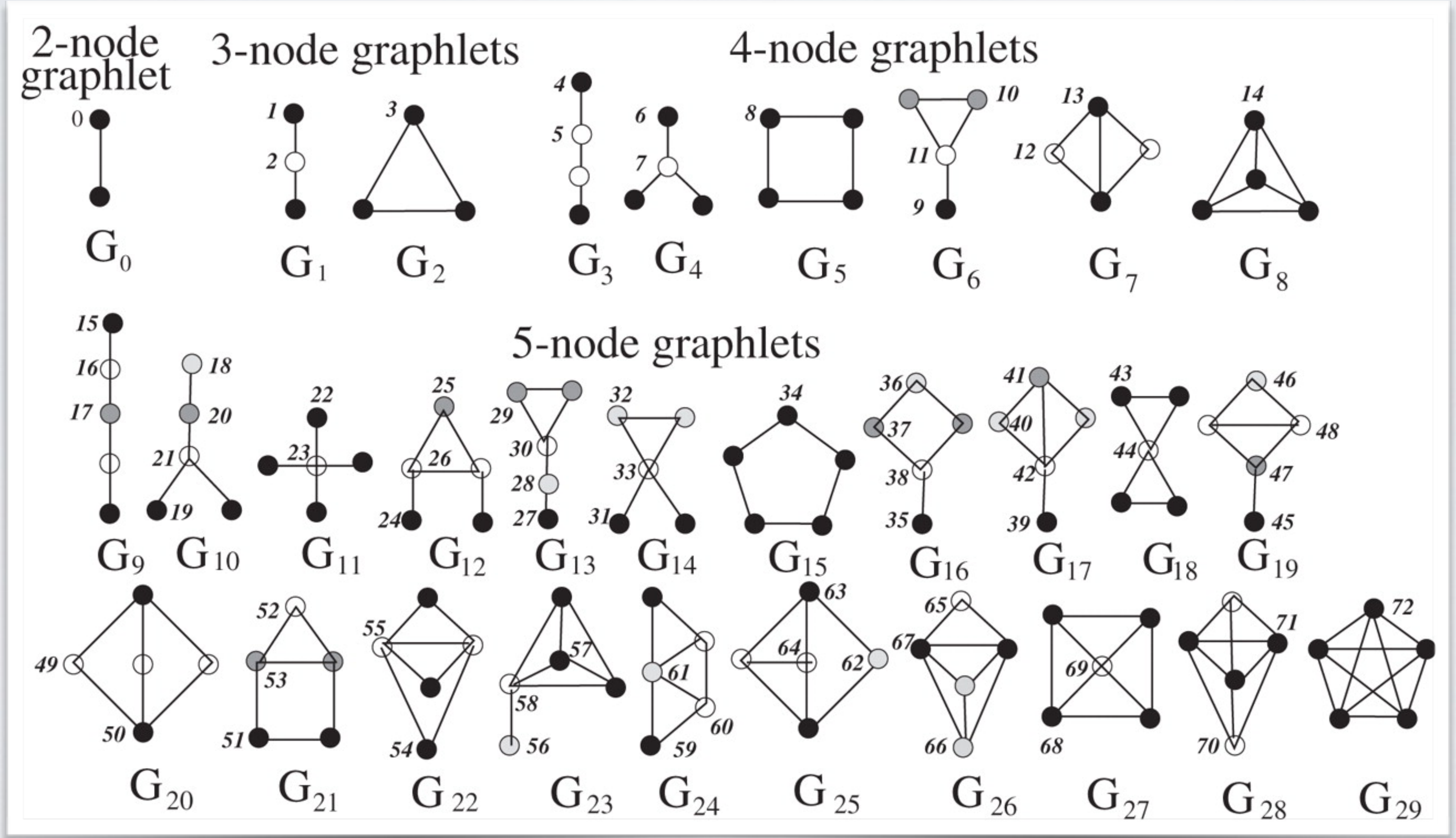
Triads counting



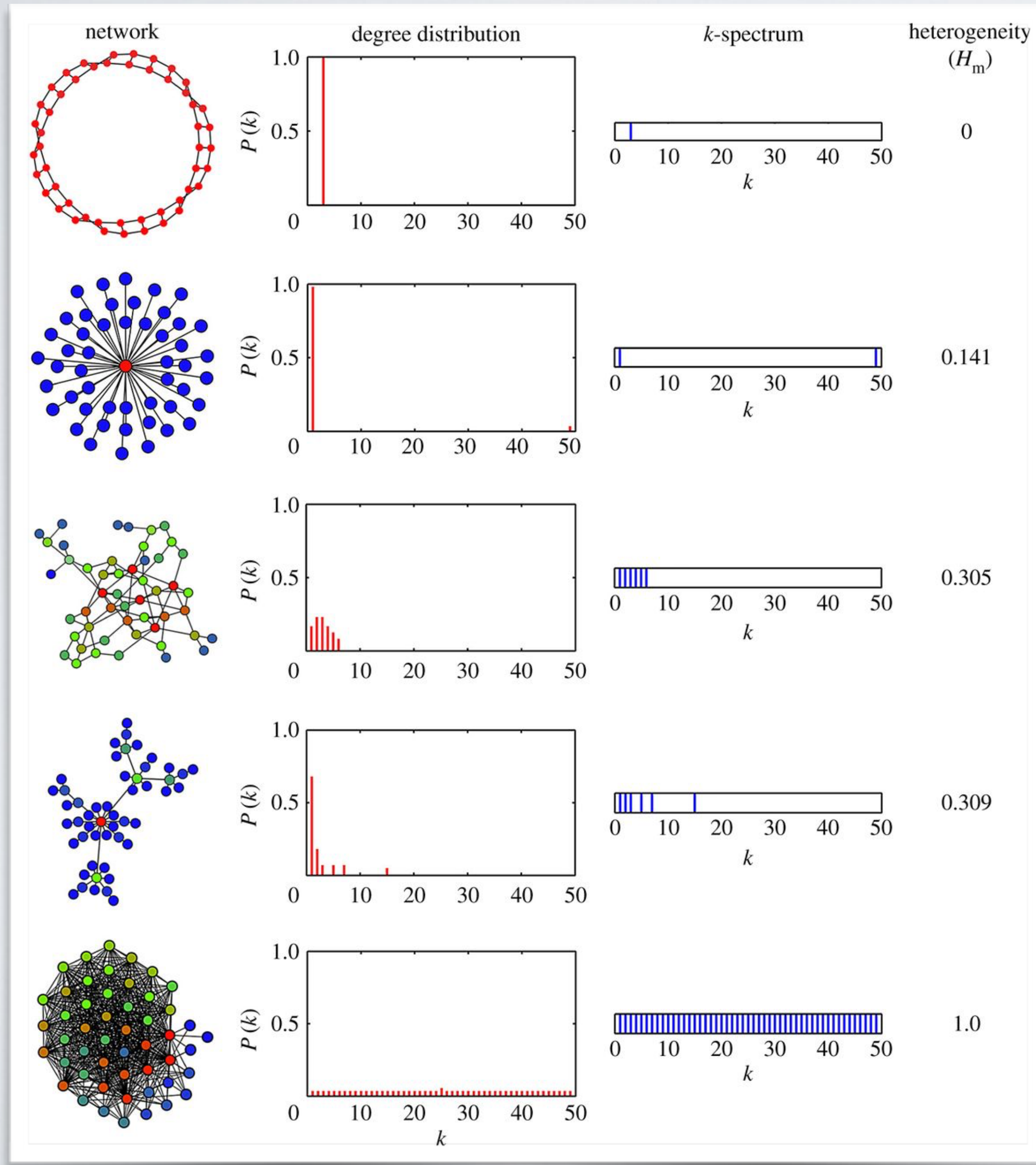


OTHER

Graphlets



OTHER



Spectral properties

Look for
Spectral graph theory

PROPERTIES OF REAL NETWORKS

SMALL WORLD NETWORK

- Not formally defined.
  - **Small** average distance (< log(N) ?)
  - **High** Clustering (>0.1 ?)

- Random networks (ER) have small avg. distance but low clustering

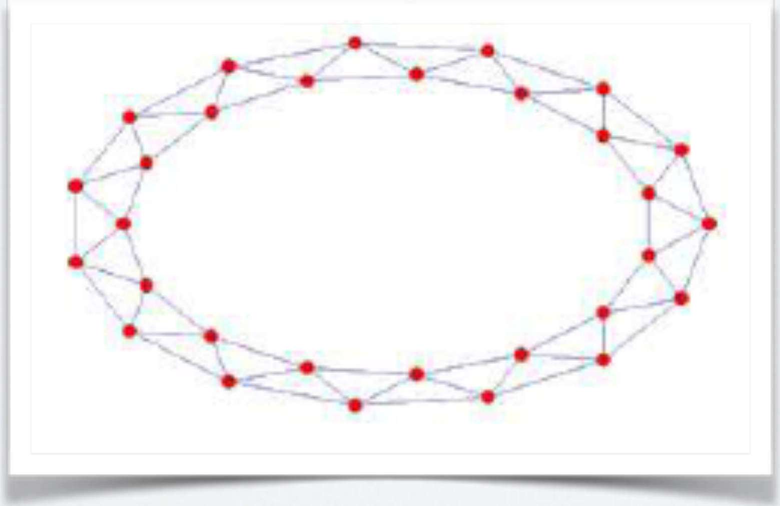
- Spatial networks have high clustering but high avg. distance

SMALL WORLD NETWORK

- Not formally defined.
  - **Small** average distance (< log(N) ?)
  - **High** Clustering (>0.1 ?)

- Random networks (ER) have small avg. distance but low clustering

- Spatial networks have high clustering but high avg. distance



CLASSIFYING NETWORKS

TABLE I. DISTRIBUTION OF NETWORKS OVER DOMAINS

Domain	Number of Networks
Social	25
Citation	20
Communication	28
Ecology	20
Biomolecular	32
Computer	21
Transportation	5

[Kantarci et al. 2013]

CLASSIFYING NETWORKS

TABLE II. OVERVIEW OF TOPOLOGICAL MEASURES RELATIVELY TO DOMAINS

	n	δ	$\langle k \rangle$	C	$\langle d \rangle$	D	R	Q
Social	[11, 1882] μ:143.88 σ: 448.52	[0.0004, 0.38] μ: 0,29 σ: 0,25	[1.85, 66.69] μ: 11.39 σ: 14.54	[0.01, 0.87] μ: 0.38 σ:0.25	[1.26, 9.33] μ: 2.80 σ: 1.68	[2, 305124] μ: 12212.12 σ: 61023.31	[2, 16] μ: 3.2 σ:4.07	[-0,03, 0.89] μ: 0.31 σ: 0.29
Citation	[35, 27779] μ:3424.53 σ: 7547.97	[0.0004, 0.26] μ: 0.07 σ: 0.09	[3.24, 516.80] μ: 39.81 σ: 104.77	[0.03, 0.69] μ: 0.23 σ: 0.17	[1.76, 8.46] μ: 3.88 σ: 1.55	[3, 37] μ: 13.93 σ: 0.26	[2, 49] μ: 8.29 σ: 13.67	[0.14, 0.93] μ: 0.41 σ: 0.20
Communication	[12, 3861] μ: 427.93 σ:103.822	[0.0004, 0.36] μ: 0.12 σ: 0.11	[1.83, 27.70] μ: 7.50 σ: 5.66	[0.01, 0.88] μ: 0.25 σ: 0.22	[1.21, 6.53] μ: 2.98 σ: 1.50	[3, 33] μ: 10.35 σ: 8.42	[2, 22] μ: 5.25 σ: 6.64	[0.01, 0.79] μ: 0.42 σ: 0.24
Ecological	[24, 128] μ: 65.38 σ: 35.00	[0.0816, 0.23] μ: 0.15 σ: 0.03	[5.13, 33.39] μ: 18.15 σ: 10.11	[0.25, 0.49] μ: 0.38 σ: 0.08	[1.81, 3.36] μ: 2.31 σ: 0.35	[8, 947493] μ: 133126.5 σ: 302590.7	[2, 11] μ: 3 σ: 2.16	[0.01, 0.53] μ: 0.04 σ: 0.12
Biomolecular	[23, 3839] μ 1099.44 σ:889.27	[0.0012, 0.34] μ: 0.02 σ: 0.06	[2.15, 15.88] μ: 5.34 σ: 2.37	[0.02, 0.57] μ: 0.07 σ: 0.14	[1.80, 7.65] μ: 4.66 σ: 1.16	[3, 35] μ: 13.03 σ: 5.33	[2, 63] μ: 9.79 σ: 15.90	[0.01, 0.78] μ: 0.52 σ: 0.17
Computer	[18, 10680] μ: 158.28 σ:2973.78	[0.0002, 0.50] μ: 0.05 σ: 0.11	[2.54, 39.1] μ: 6.95 σ: 8.67	[0.01, 0.50] μ: 0.12 σ: 0.14	[1.49, 18.98] μ: 4.31 σ: 3.48	[2, 46] μ: 11.65 σ: 8.71	[2, 352] μ: 38.13 σ: 86.11	[0.01, 0.88] μ: 0.43 σ: 0.26
Transportation	[75, 332] μ:174.40 σ: 107.60	[0.0327, 0.24] μ: 0.22 σ: 0.26	[4.23, 194,64] μ: 37.90 σ: 69.61	[0.01, 0.84] μ: 0.32 σ: 0.26	[1.21, 3.48] μ: 2.37 σ: 0.70	[3, 19] μ: 6.94 σ: 6.27	[2, 16] μ: 4.28 σ: 5.67	[0.01, 0.44] μ: 0.15 σ: 0.16

Density **avg. degree** **avg. CC** **avg. distance** **Diameter** Radius Modularity

[Kantarci et al. 2013]

CLASSIFYING NETWORKS

TABLE III. CORRELATION BETWEEN GLOBAL MEASURES

	$\boldsymbol{\delta}$	$\langle \boldsymbol{k} \rangle$	$\boldsymbol{C}$	$\langle \boldsymbol{d} \rangle$	$\boldsymbol{D}$	$\boldsymbol{R}$	$\boldsymbol{Q}$
$\boldsymbol{\delta}$	-	0.16	0.76	-0.45	0.02	-0.14	-0.71
$\langle \boldsymbol{k} \rangle$	-	-	0.12	-0.16	-0.01	0.00	-0.13
$\boldsymbol{C}$	-	-	-	-0.43	0.04	-0.09	-0.51
$\langle \boldsymbol{d} \rangle$	-	-	-	-	-0.09	0.59	0.60
$\boldsymbol{D}$	-	-	-	-	-	-0.03	-0.12
$\boldsymbol{R}$	-	-	-	-	-	-	0.16
$\boldsymbol{Q}$	-	-	-	-	-	-	-

TABLE VII. DISTRIBUTION OF DOMAINS OVER CLUSTERS

	Cluster 1	**Cluster 2**
Biomolecular	29	3
Citation	16	4
Computer	19	2
Ecology	1	19
Transportation	0	5
Social	5	20
Communication	5	23

[Kantarci et al. 2013]

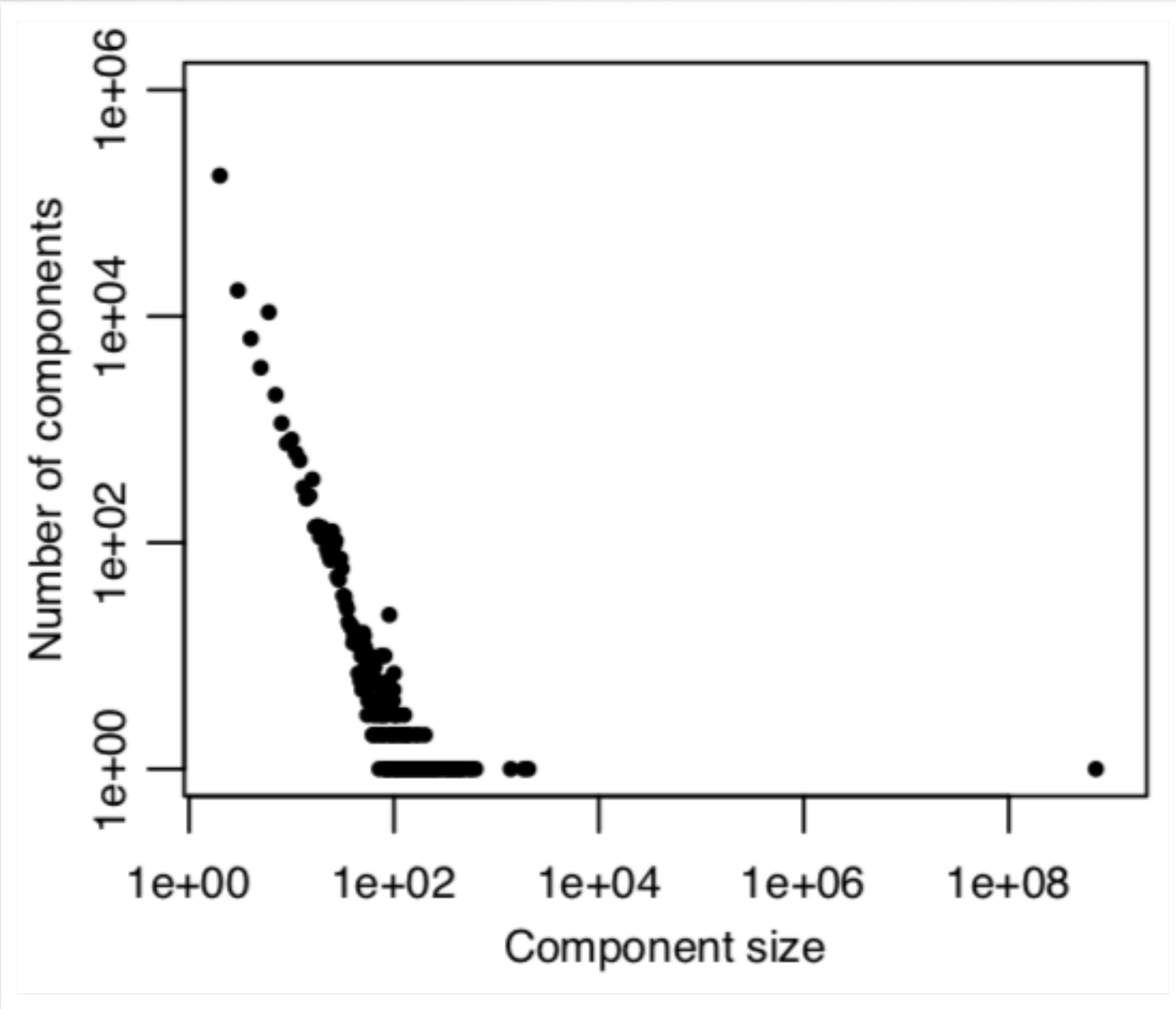
EXEMPLE OF GRAPH ANALYSIS

- Source: [The Anatomy of the Facebook Social Graph, Ugander et al. 2011]
- The Facebook friendship network in 2011

EXEMPLE OF GRAPH ANALYSIS

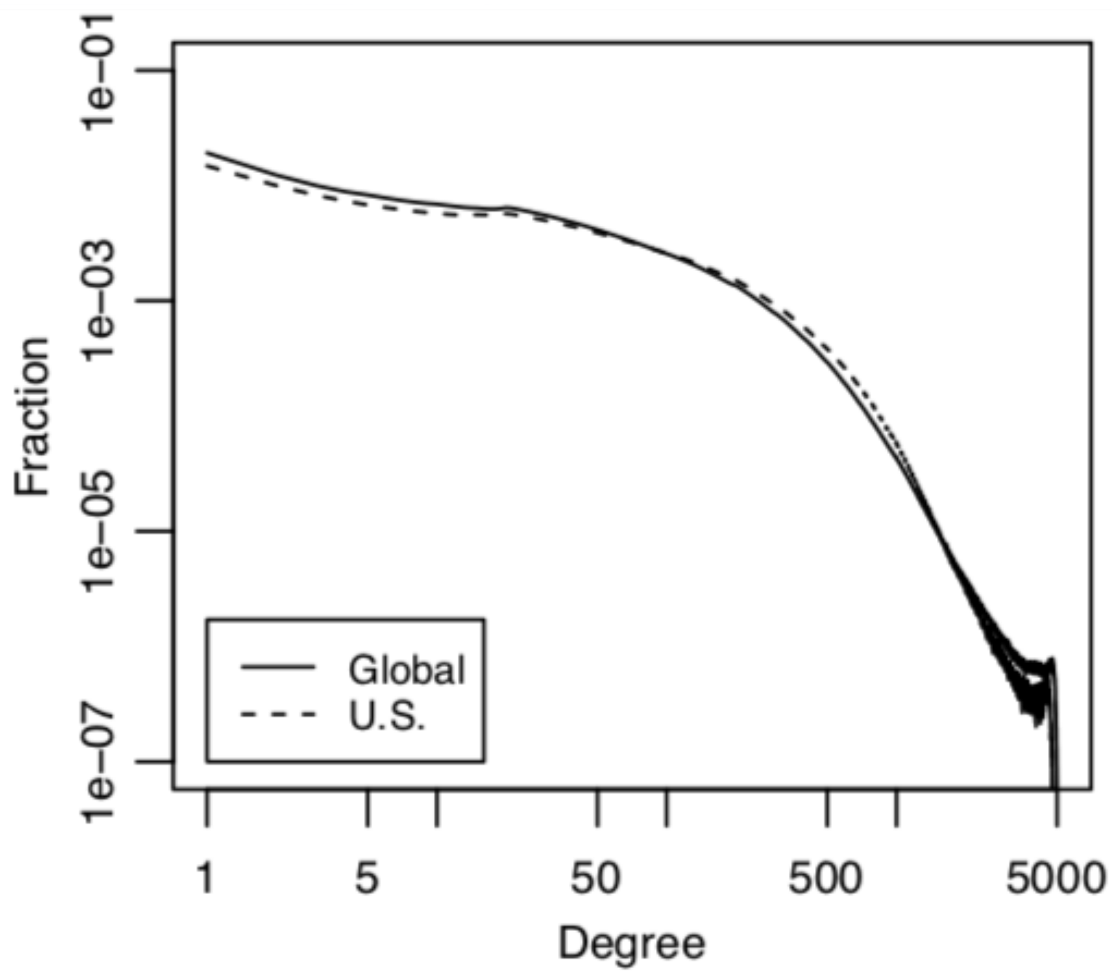
- 721M users (nodes) (active in the last 28 days)
- 68B edges
- Average degree: 190 (average # friends)
- Median degree: 99
- Connected component: 99.91%

EXEMPLE OF GRAPH ANALYSIS

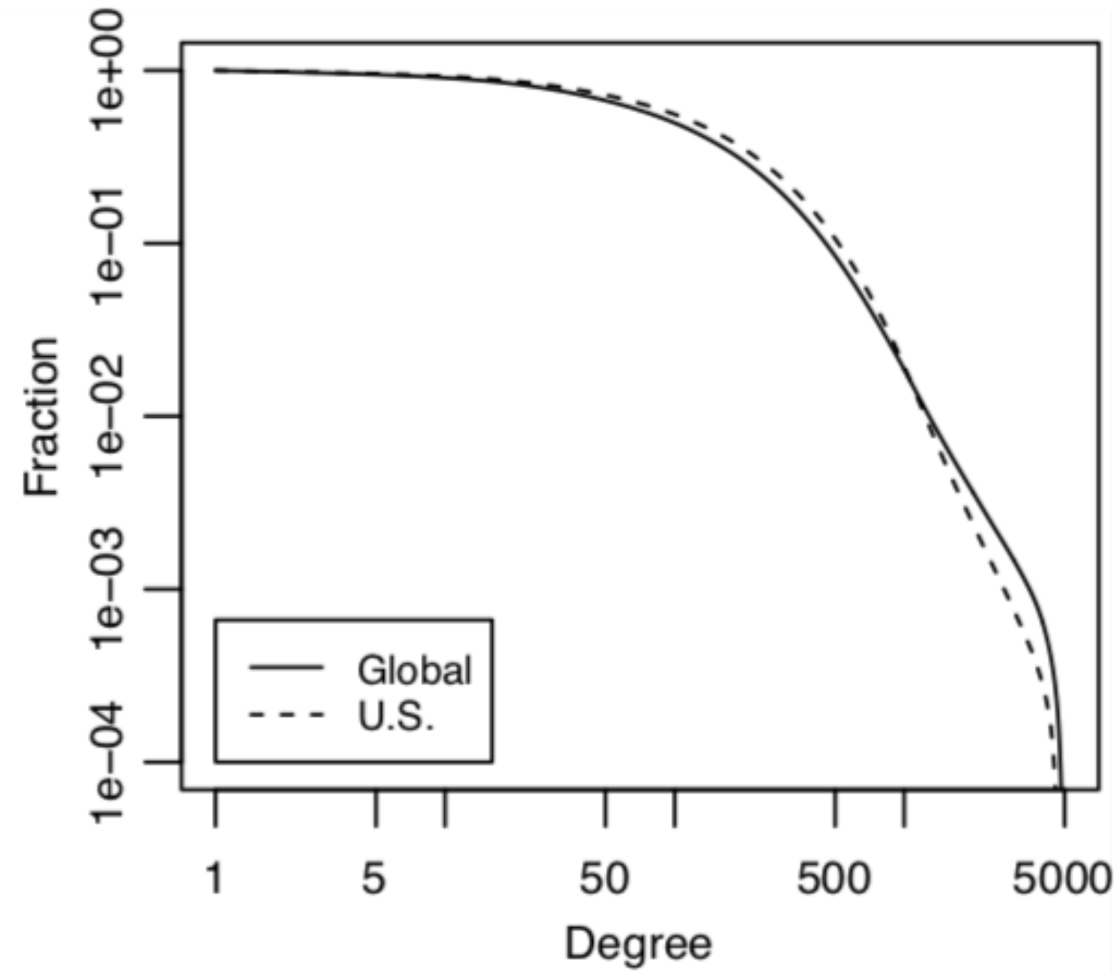


Component size
Distribution

EXEMPLE OF GRAPH ANALYSIS



(a)

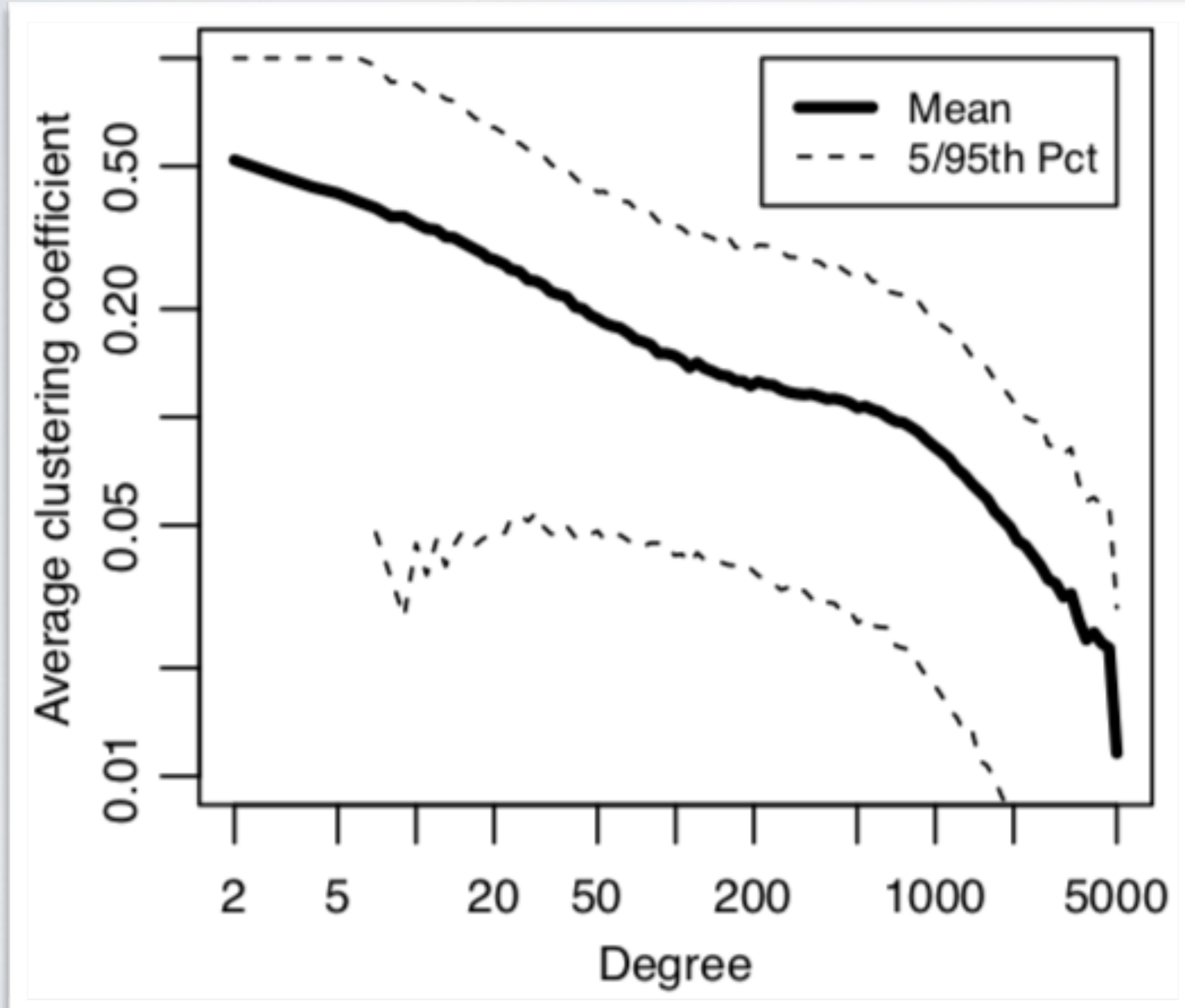


(b)

Cumulative

Degree distribution

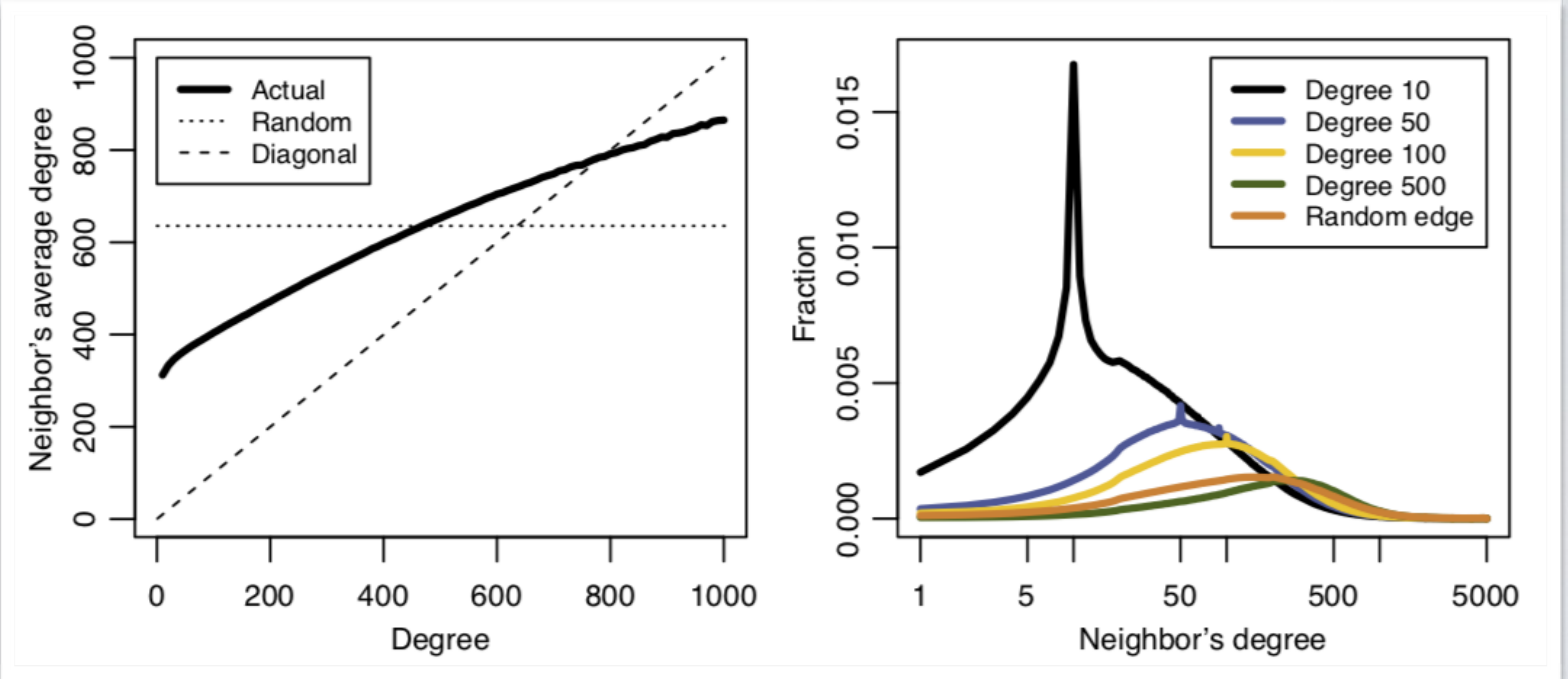
EXEMPLE OF GRAPH ANALYSIS



Clustering coefficient
By degree

Median user: 0.14:
14% of pair friends
Are actually friends

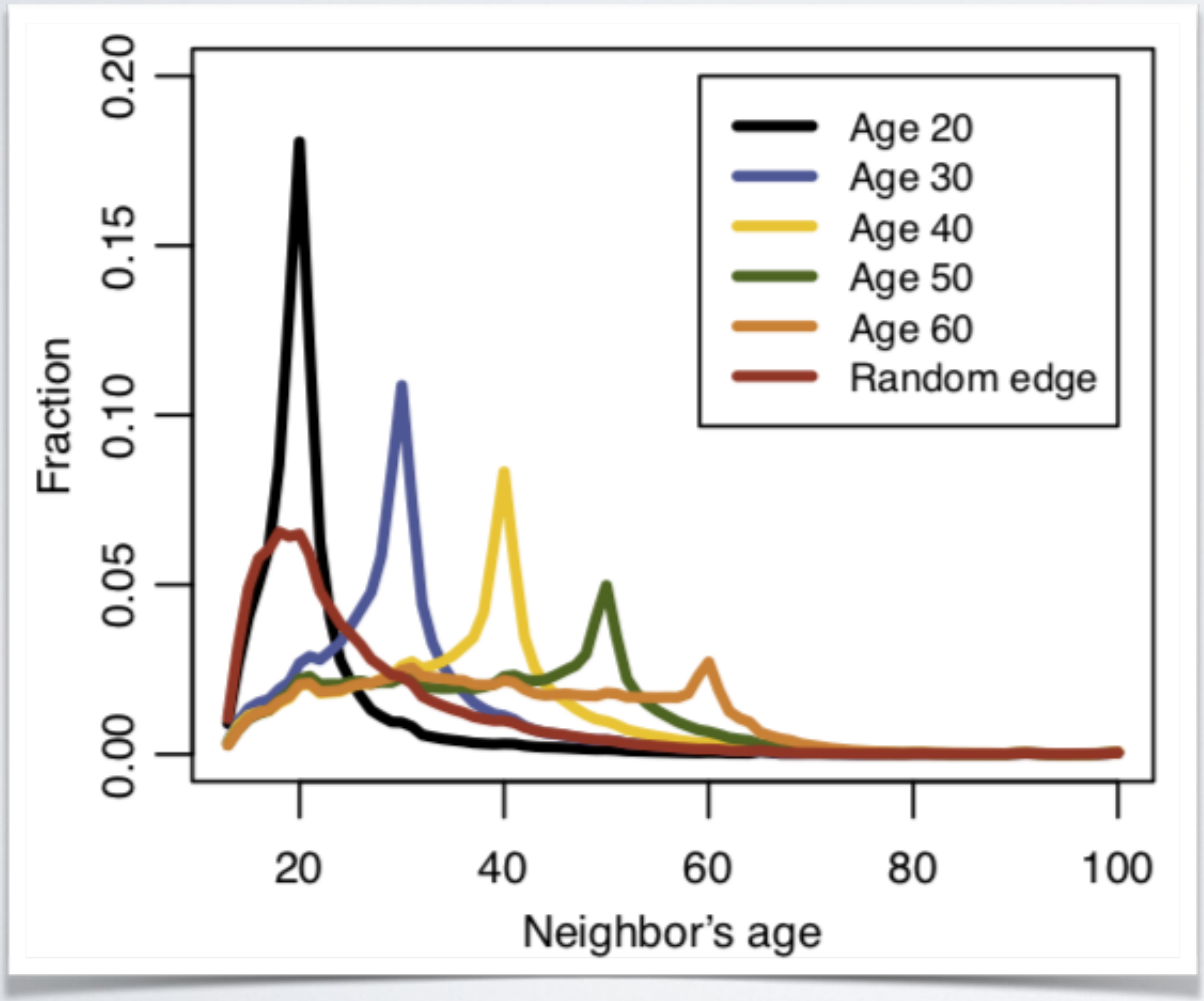
EXEMPLE OF GRAPH ANALYSIS



My friends have more
Friends than me!

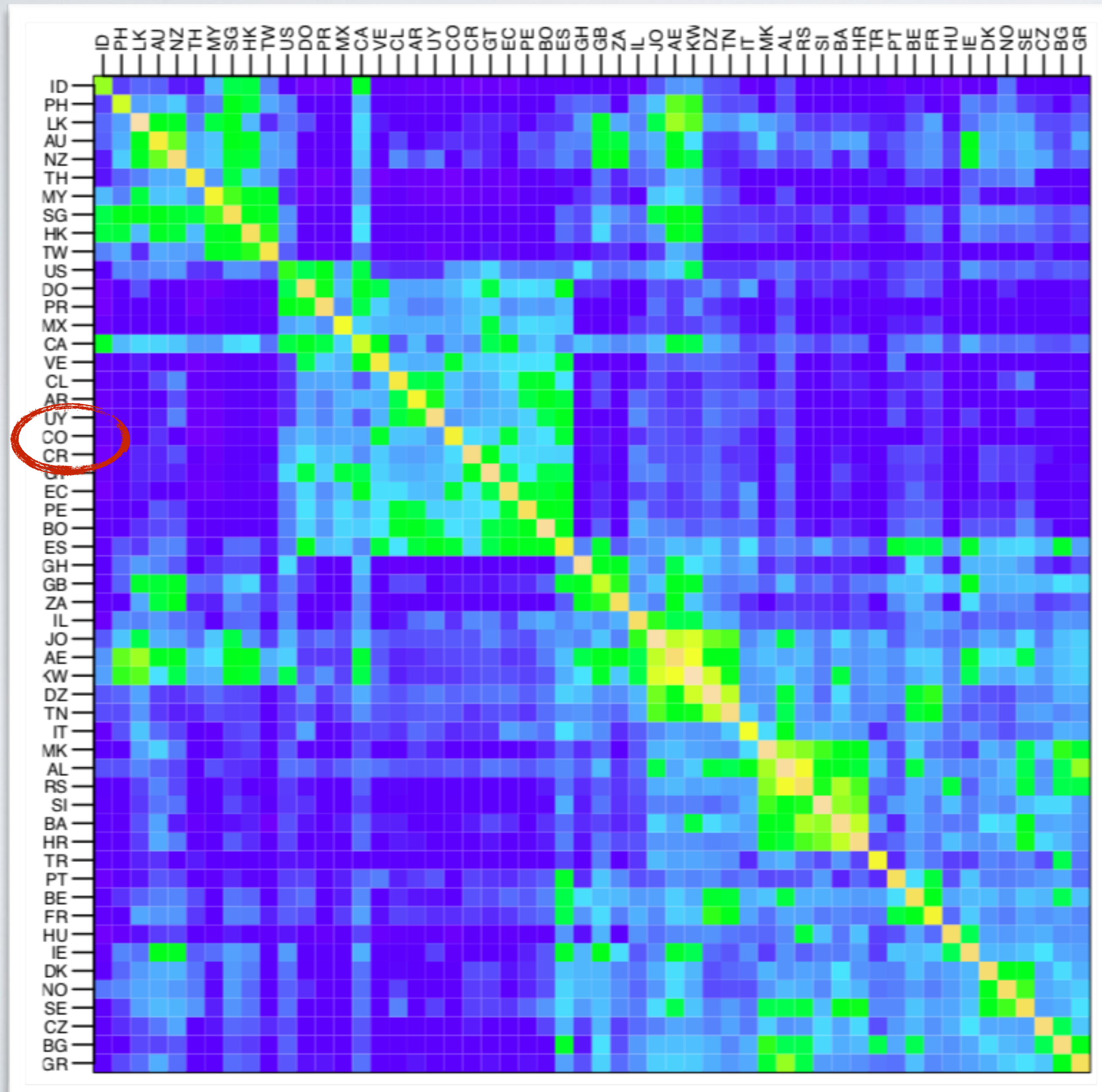
Many of my friends have the
Same # of friends than me!

EXEMPLE OF GRAPH ANALYSIS



Age homophily

EXEMPLE OF GRAPH ANALYSIS



Country similarity

84.2% percent of edges are within countries

(More in the community detection class)

MANIPULATING AND VISUALIZING GRAPHS

Using Gephi (Demo)

PRACTICAL

- Choose a network (I recommend to start with the soccer one ;))
  - http://cazabetremy.fr/Teaching/catedra.html
- Use Gephi to visualize it
  - Layout, node size and colors, edge size and colors, name…
- Choose a larger graph and try to visualize it
- Use filtering tools to clarify
- Export and interpret