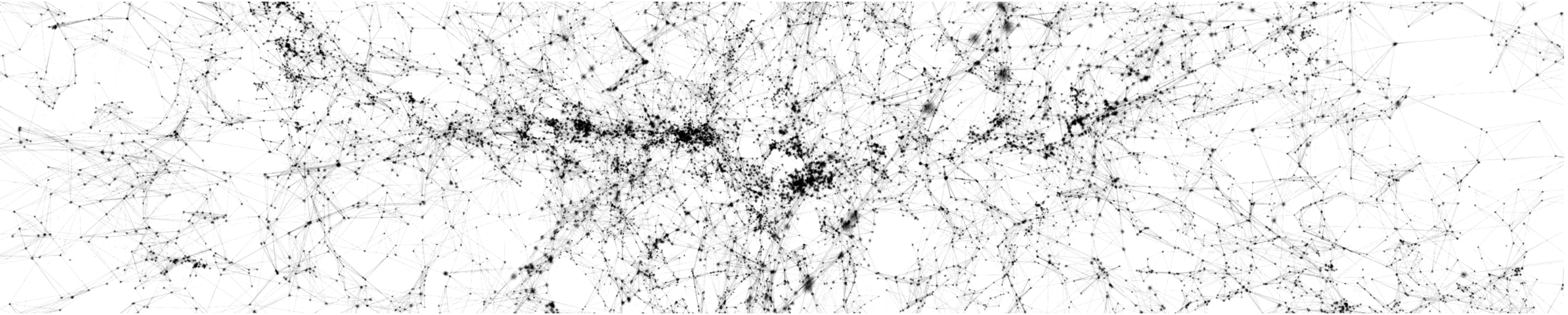
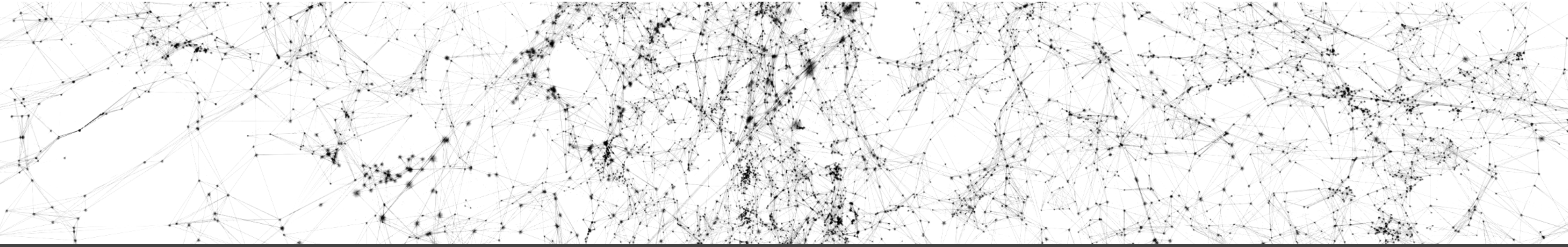


Patterns

Analysis of complex interconnected data



Quick Notes

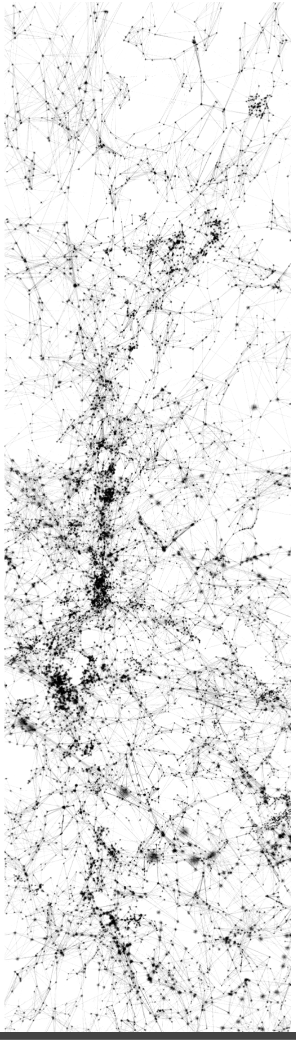
- First assignment is released
  - http://www.reirab.com/Teaching/NS22/Assignment_1.pdf
  - Join a Group in Mycourses & Submit the assignment through Mycourses
  - Late policy for assignments, 2^k% of the grade will be deducted per k days of delay.
- Use Ed discussion
  - Ask questions
  - Share tips & discuss the assignment

Deadlines

- assignment 1 due on Sep. 20th
- assignment 2 due on Oct. 4th
- assignment 3 due on Oct. 18th
- project proposal slides due on Oct. 26th
- Proposal presentations are scheduled for **Oct. 27th**
- project proposal due on Oct. 28th
- Reviews (first round) are due on Nov. 4th
- project proposal slides due on Nov. 9th
- Progress presentations are scheduled for **Oct. 10th**
- project progress report due on Nov. 11th
- Reviews (second round) are due on Nov. 19th [extended]
- project final report slides due on Nov. 28th
- Project presentations are scheduled for **Nov. 29th & Dec. 1st** [divided alphabetically]
- project final report due on Dec. 5th
- Reviews (third round) are due on Dec. 14th
- [Optional] project revised report and rebuttal are due on Dec. 20th
- Please note that these dates are tentative and subject to change.

Outline

- Sparsity Pattern
- Scale Free Pattern
  - Power-law degree distribution
  - Fitting a power-law
  - Preferential attachment and AB model
- Assortativity Pattern
- Transitivity Pattern
  - powers of A & counting triangles
- Small world Pattern
  - Shortest path
- How to pattern?



Adjacency Matrix: marginals

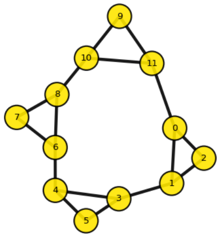
marginals of A => degree sequence

For undirected graphs: we have $A_{ij} = A_{ji} = 1$ if there is an edge between i and j, and degree of each node is:

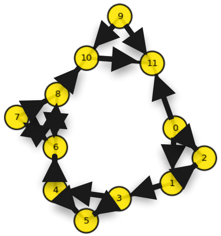
$$d_i = \sum_j A_{ij}$$

For directed graphs, $A_{ij} = 1$ if there is an edge from node j to i, and in/out degrees of each node are:

$$d_i^{in} = \sum_j A_{ij} \;, \quad d_i^{out} = \sum_j A_{ji}$$



	0	1	2	3	4	5	6	7	8	9	10	11	degrees
0	0	1	1	0	0	0	0	0	0	0	0	1	3
1	1	0	1	1	0	0	0	0	0	0	0	0	3
2	1	1	0	0	0	0	0	0	0	0	0	0	2
3	0	1	0	0	1	1	0	0	0	0	0	0	3
4	0	0	0	1	0	1	1	0	0	0	0	0	3
5	0	0	0	1	1	0	0	0	0	0	0	0	2
6	0	0	0	0	1	0	0	1	1	0	0	0	3
7	0	0	0	0	0	0	1	0	1	0	0	0	2
8	0	0	0	0	0	0	1	1	0	0	1	0	3
9	0	0	0	0	0	0	0	0	0	0	1	1	2
10	0	0	0	0	0	0	0	0	1	1	0	1	3
11	1	0	0	0	0	0	0	0	0	1	1	0	3
degrees	3	3	2	3	3	2	3	2	3	2	3	3	

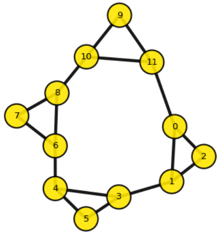


	0	1	2	3	4	5	6	7	8	9	10	11	in-degrees
0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0	0	0	0	1
2	1	1	0	0	0	0	0	0	0	0	0	0	2
3	0	1	0	0	0	0	0	0	0	0	0	0	1
4	0	0	0	1	0	0	0	0	0	0	0	0	1
5	0	0	0	1	1	0	0	0	0	0	0	0	2
6	0	0	0	0	1	0	0	1	1	0	0	0	3
7	0	0	0	0	0	0	1	0	0	0	0	0	1
8	0	0	0	0	0	0	1	1	0	0	0	0	2
9	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	1	1	0	0	2
11	1	0	0	0	0	0	0	0	0	1	1	0	3
out-degrees	3	2	0	2	2	0	2	2	2	2	1	0	

Adjacency Matrix: marginals

marginals of A => degree sequence

For undirected graphs: we have $A_{ij} = A_{ji} = 1$ if there is an edge between i and j, and degree of each node is:



	0	1	2	3	4	5	6	7	8	9	10	11	degrees
0	0	1	1	0	0	0	0	0	0	0	0	1	3
1	1	0	1	1	0	0	0	0	0	0	0	0	3
2	1	1	0	0	0	0	0	0	0	0	0	0	2
3	0	1	0	0	1	1	0	0	0	0	0	0	3
4	0	0	0	1	0	1	1	0	0	0	0	0	3
5	0	0	0	1	1	0	0	0	0	0	0	0	2
6	0	0	0	0	1	0	0	1	1	0	0	0	3
7	0	0	0	0	0	0	1	0	1	0	0	0	2
8	0	0	0	0	0	0	1	1	0	0	1	0	3
9	0	0	0	0	0	0	0	0	0	0	1	1	2
10	0	0	0	0	0	0	0	0	1	1	0	1	3
11	1	0	0	0	0	0	0	0	0	1	1	0	3
degrees	3	3	2	3	3	2	3	2	3	2	3	3	

$$d_i = \sum_j A_{ij}$$

What is $\sum_{ij} A_{ij}$? $\sum d_i = 2E$ twice the number of edges

$N = 12,\ E = 16$

Mean degree: $\bar{d} = \frac{1}{N} \sum_{ij} A_{ij} = \frac{1}{N} \sum_i d_i$

$\bar{d} = 2.6$

Density: $\rho = \frac{\sum_{ij} A_{ij}}{N(N-1)} = \frac{1}{N}\bar{d}$

$\rho = 0.24$

Real-world networks are **sparse**

mean degree << N-1
(or E << E_{max})

WWW (Stanford-Berkeley):	N=319,717	mean degree=9.65
Social networks (LinkedIn):	N=6,946,668	mean degree=8.87
Communication (MSNIM):	N=242,720,596	mean degree=11.1
Co-authorships (DBLP):	N=317,080	mean degree=6.62
Internet (AS-Skitter):	N=1,719,037	mean degree=14.91
Roads (California):	N=1,957,027	mean degree=2.82
Proteins (S. Cerevisiae):	N=1,870	mean degree=2.39

(Source: Leskovec et al., Internet Mathematics, 2009)

From Leskovec's slides

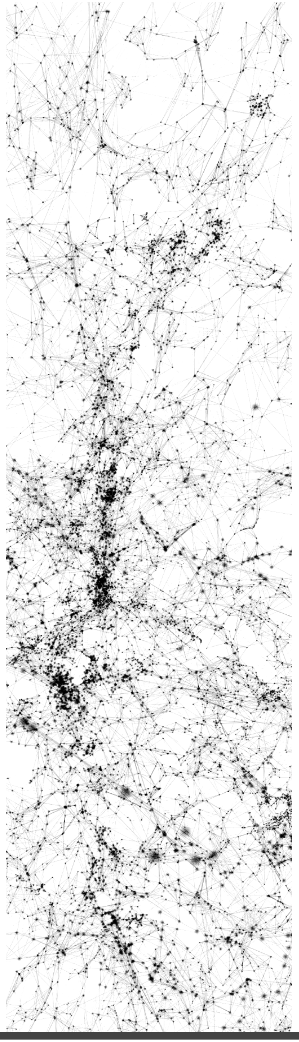
Adjacency matrix is filled with zeros!

(Density of the matrix: WWW=$1.51*10^{-5}$, MSNIM= $2.27*10^{-8}$)

Implications? Use sparse representations, density is not very informative!

Outline

- Sparsity Pattern
- **Scale Free Pattern**
  - Power-law degree distribution
  - Fitting a power-law
  - Preferential attachment and AB model
- Assortativity Pattern
- Transitivity Pattern
  - powers of A & counting triangles
- Small world Pattern
  - Shortest path
- How to pattern?

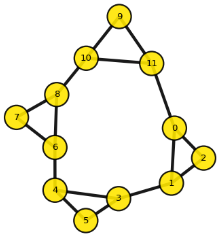


Adjacency Matrix: marginals

marginals of A => degree sequence

For undirected graphs: we have $A_{ij} = A_{ji} = 1$ if there is an edge between i and j, and degree of each node is:

$$d_i = \sum_j A_{ij}$$



	0	1	2	3	4	5	6	7	8	9	10	11	degrees
0	0	1	1	0	0	0	0	0	0	0	0	1	3
1	1	0	1	1	0	0	0	0	0	0	0	0	3
2	1	1	0	0	0	0	0	0	0	0	0	0	2
3	0	1	0	0	1	1	0	0	0	0	0	0	3
4	0	0	0	1	0	1	1	0	0	0	0	0	3
5	0	0	0	1	1	0	0	0	0	0	0	0	2
6	0	0	0	0	1	0	0	1	1	0	0	0	3
7	0	0	0	0	0	0	1	0	1	0	0	0	2
8	0	0	0	0	0	0	1	1	0	0	1	0	3
9	0	0	0	0	0	0	0	0	0	0	1	1	2
10	0	0	0	0	0	0	0	0	1	1	0	1	3
11	1	0	0	0	0	0	0	0	0	1	1	0	3
degrees	3	3	2	3	3	2	3	2	3	2	3	3	

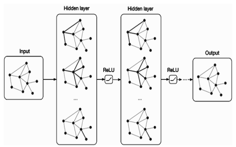
$N = 12, E = 16$

Degree distribution:

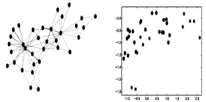
- shows how many nodes of degree d are in the graph
- degree sequence of all nodes $\Rightarrow$ count & get frequencies

  $[3, 3, 2, 3, 3, 2, 3, 2, 3, 2, 3, 3] \Rightarrow [0, 0, 4, 8]$

Recent Trend: Deep Learning for Graphs

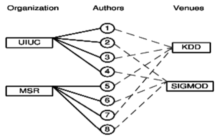


2015~2021
- **Graph Neural Networks**
- Deep Learning for Networks
- High-Order Networks [Benson et al.]

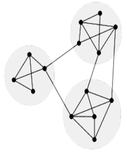


2010~2014
- Info. vs. Social Networks (Twitter) [Kwak et al.]
- **Signed** Networks [Leskovec et al.]
- Semantic Social Networks [Tang et al.]
- Four Deg. Of Separation [Backstrom et al.]
- Structural Diversity [Ugander et al.]
- Computational Social Science [Watts]
- **Network Embedding** [Perozzi et al.]

21st Century: More CS

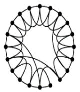


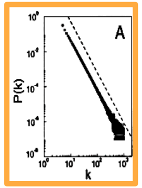
2005~2009
- Graph Evolution [Leskovec et al.]
- 3 Deg. Of Influence [Christakis & Fowler]
- Social **Influence** Analysis [Tang et al.]
- Six Deg. Of Separation [Leskovec & Horvitz]
- Network **Heterogeneity** [Sun & Han]
- Network **Embedding** [Tang & Liu]
- Computer Social Science [Lazer et al.]



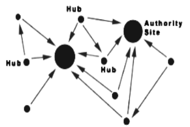
2000~2004
- Influence Max'n [Domingos & Kempe et al.]
- **Community Detection** [Girvan & Newman]
- Network Motifs [Milo et al.]
- Link Prediction [Liben-Nowell & Kleinberg]

Late 20th Century: CS & Physics

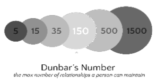




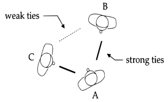
1998~1999
- **Small Worlds** [Watts & Strogatz]
- **Scale Free** [Barabasi & Albert]
- **Power Law** [Faloutsos x3]



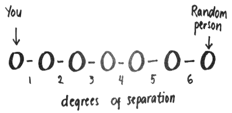
1997
- **HITS** [Kleinberg]
- **PageRank** [Page & Brin]
- Hyperlink Vector Voting [Li]



1992
- Structural Hole [Burt]
- **Dunbar's Number** [Dunbar]



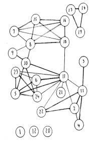
1970s
- The Strength Of **Weak Tie** [Granovetter]



1960s
- **Small Worlds** [Migram]

20th Century: Sociology

1950s
- **Random Graph** [Erdos, Renyi, Gilbert]
- Degree Sequence [Tuttle, Havel, Hakami]
- **Homophily** [Lazarsfeld & Merton]
- Balance Theory [Heider et al.]



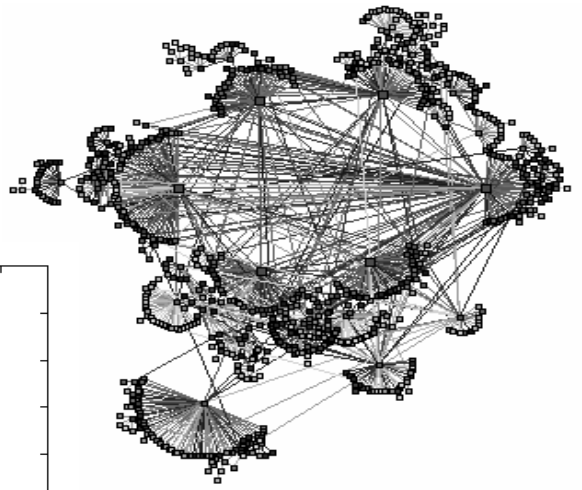
1930s
- **Sociogram** [Moreno]

Based on Slides from Jie Tang

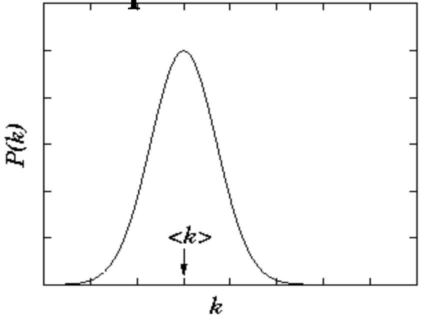
The first observations

Nodes: **WWW documents**
Links: **URL links**

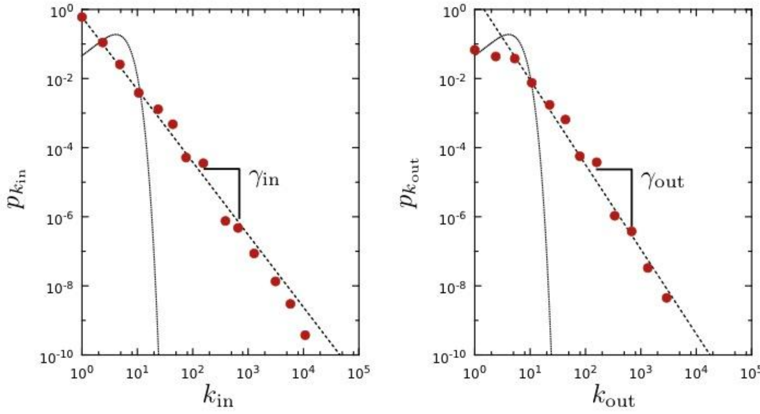
Over 3 billion documents
ROBOT: collects all URL's found in a document and follows them recursively



Expected



Observed

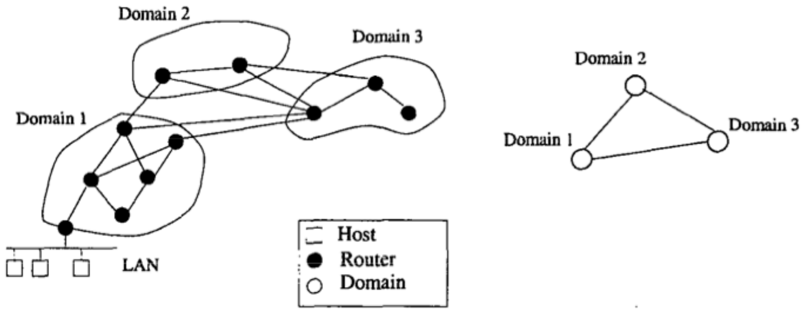


[HTML] **Diameter** of the **world-wide web**
R Albert, H Jeong, AL Barabási - nature, 1999 - nature.com
… the **diameter** of the **web**… **web** is a highly connected graph with an average **diameter** of only 19 links. The logarithmic dependence of <d> on N is important to the future potential of the **web**…
☆ Save 99 Cite Cited by 6292 Related articles All 42 versions

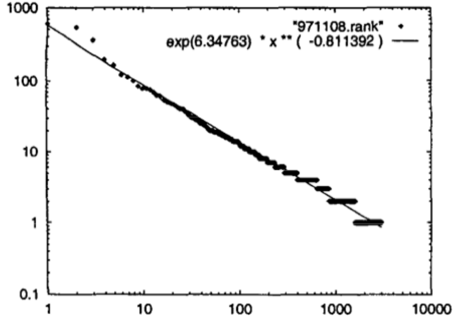
The first observations

Nodes: **Autonomous Systems (e.g. ISPs)**
Links: **Routing**

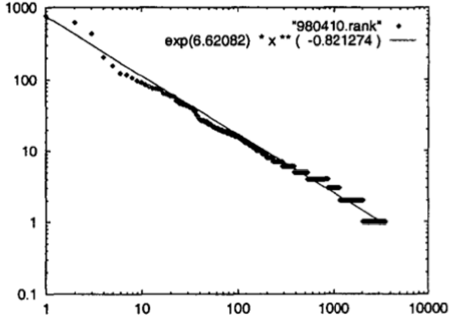
Around 4K nodes
Graphs from data in routing tables



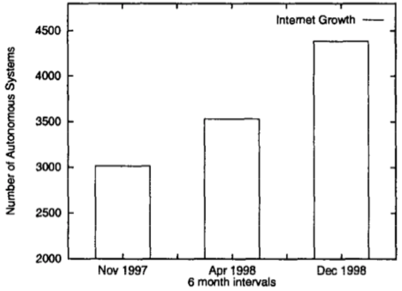
(a) (b)



(a) Int-11-97



(b) Int-04-98



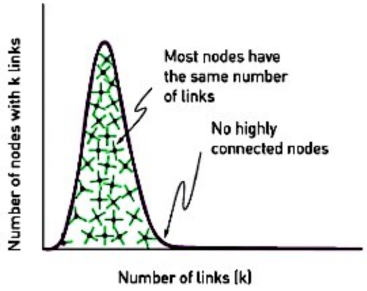
On power-law relationships of the internet topology
M Faloutsos, P Faloutsos, C Faloutsos - ACM SIGCOMM computer …, 1999 - dl.acm.org
Despite the apparent randomness of the Internet, we discover some surprisingly simple power-laws of the Internet topology. These power-laws hold for three snapshots of the Internet, between November 1997 and December 1998, despite a 45% growth of its size during that period. We show that our power-laws fit the real data very well resulting in correlation coefficients of 96% or higher. Our observations provide a novel perspective of the structure of the Internet. The power-laws describe concisely skewed distributions of graph …
☆ Save 99 Cite Cited by 7479 Related articles All 66 versions

Example

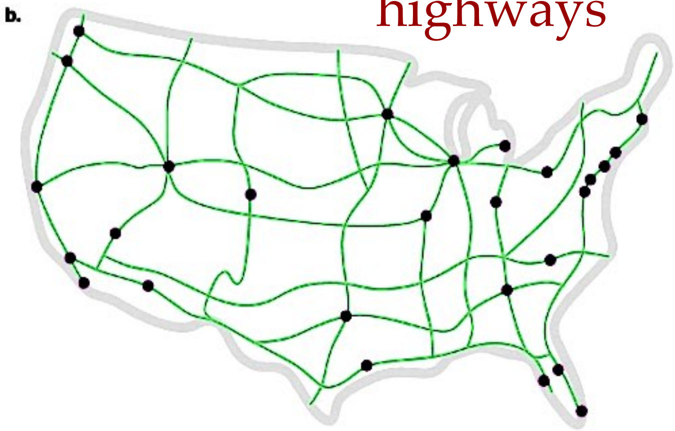
In highway networks, cities are of comparable connections, one has an expectation for it and each cities connections are usually close to this expectation: $\lambda = E(d) = \sigma^2(d)$

In air-traffic networks, we have major hubs and many smaller airports.

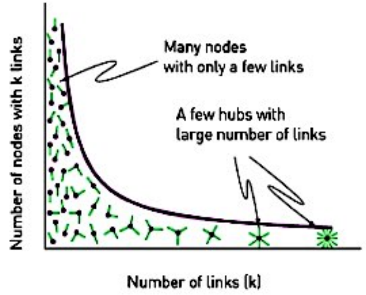
poisson



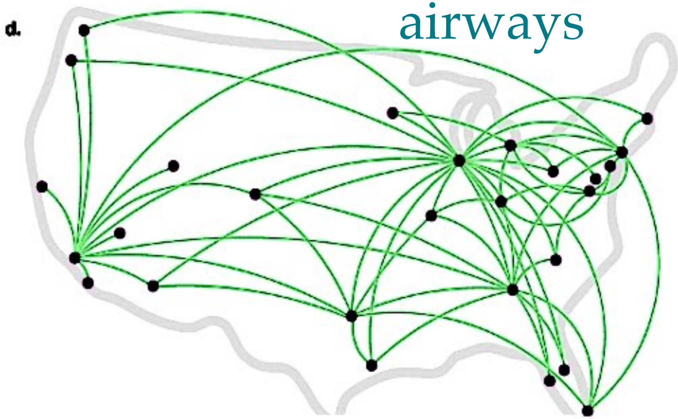
highways



powerlaw



airways



Power law distribution

Linear fit in log-log implies:

$$ln(p_d) = -\alpha\, ln(d) + \beta$$

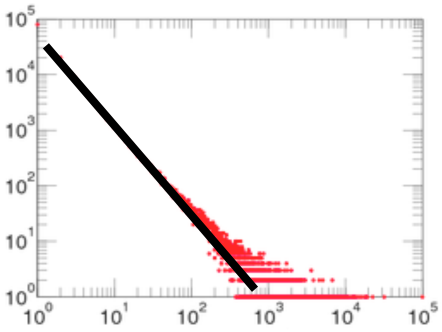
Which gives:

$$p_d = Cd^{-\alpha}$$

What is C? e^{β}

more info: [Power_law]

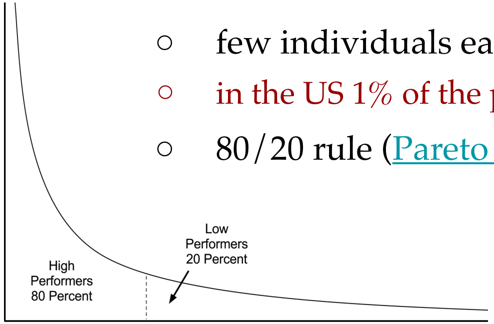
Provides a good fit to the linear pattern observed in log-log plots for degree distribution



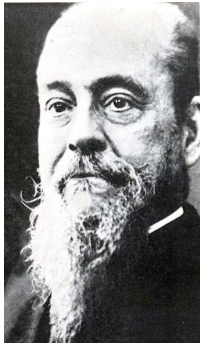
Even better fit when (logarithmically) bin the range

Powerlaws are common

- Income follows a Pareto distribution
  - few individuals earned most of the money & majority earned small amounts
  - in the US 1% of the population earns a disproportionate 15% of the total US income
  - 80/20 rule (Pareto principle): a general rule of thumb



e.g. 20 percent of the code has 80 percent of the errors



Vilfredo Federico Damaso Pareto (1848 – 1923)

- Zipf's law
  - distribution of words ranked by their frequency in a random text corpus is approximated by a power-law distribution
  - the second item occurs approximately 1/2 as often as the first, and the third item 1/3 as often as the first, and so on

George Kingsley Zipf (1902 – 1950)

Scale free networks

Networks with power-law degree distribution are coined as scale-free

Since power-law is scale invariance:

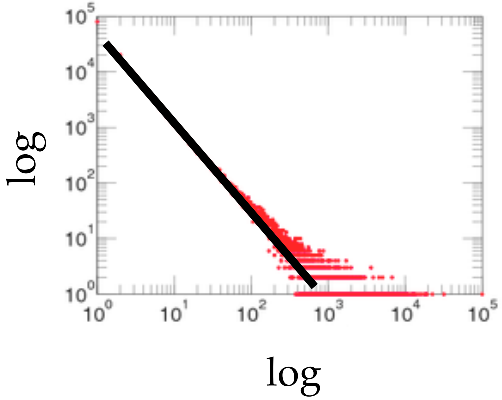
$$f(d) = p_d = Cd^{-\alpha}$$

$$f(\lambda d) = C(\lambda d)^{-\alpha} = \lambda^{-\alpha} f(d)$$

(invariant under all re-scalings)

Note: function f is scale invariance iff $f(\lambda x) = \lambda^a f(x)$ for some a & all λ

Provides a good fit to the linear pattern observed in log-log plots for degree distribution



Even better fit when (logarithmically) bin the range

Scale free networks are debated

Networks with power-law degree distribution are coined as scale-free

Commonly used but also debated

debate is around how test statistically

What we care about most is not the fit but the heavy-tail property

[HTML] **Scale-free networks are rare**
AD Broido, A Clauset - Nature communications, 2019 - nature.com
… **scale-free networks** 8,9 , and we find that 39% of **network** data sets have median estimated parameters in this range. We also find that 34% of **network** … the **scale-free network** literature. …
☆ Save 99 Cite Cited by 737 Related articles All 24 versions

[HTML] **Rare and everywhere: Perspectives on scale-free networks**
P Holme - Nature communications, 2019 - nature.com
… When "Scale-free networks are **rare**" appeared as a preprint in January 2018 it triggered a tremendous online activity, including articles, blog posts (by Barabási https://www.barabasilab.com/post/love-is-all-y need …
☆ Save 99 Cite Cited by 117 Related articles All 12 versions

Scale-free networks well done
I Voitalov, P van der Hoorn, R van der Hofstad… - Physical Review …, 2019 - APS
We bring rigor to the vibrant activity of detecting power laws in empirical degree distributions in real-world **networks**. We first provide a rigorous definition of power-law distributions, …
☆ Save 99 Cite Cited by 129 Related articles All 11 versions

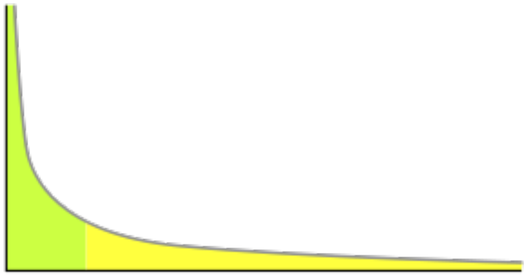
How **rare** are power-law **networks** really?
I Artico, I Smolyarenko… - Proceedings of the …, 2020 - royalsocietypublishing.org
… This means that it is impossible to detect **scale free networks**, whose power-law regime 'starts' at O(N). Every finite **network** degree distribution could potentially behave like a power-law …
☆ Save 99 Cite Cited by 12 Related articles All 12 versions

Heavy / fat / long Tailed Degree Distribution

Degree distribution is often **heavy tailed** in real world networks

There are many with very small degree & nodes with very high degree

This is the key point which is commonly referred to as powerlaw distribution and scale-free property. Powerlaw is a subtype of heavy tail and other subtypes might give a closer fit



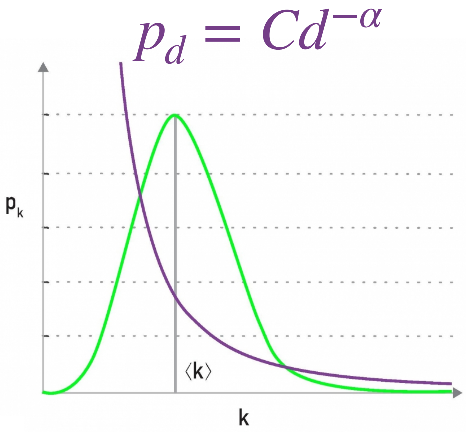
Read more on wiki if interested: Heavy-tailed_distribution, Fat-tailed_distribution, Power_law

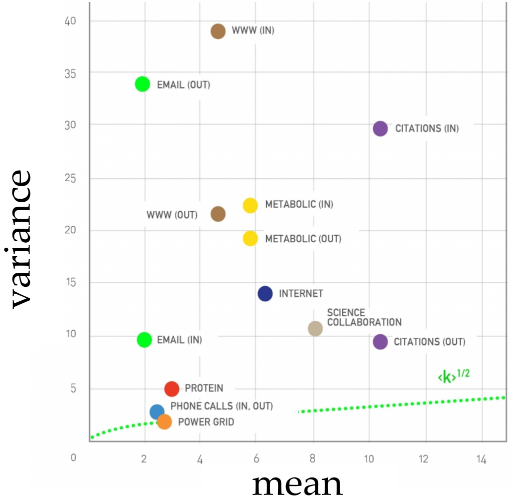
Implication? variance might not be finite, and even mean might not be well-defined

Mean & variance for a power-law

- Well-defined mean only if α>2
- No finite variance if α<3
  - the degree of a randomly chosen node can be significantly different from the mean degree

- Most real world networks are within this range
  - In the examples datasets of Barabasi book, we can see how variance deviates from expected variance of same mean random network with poisson distribution (dashed green line)





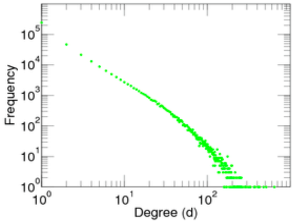
Heavy / fat / long Tailed Degree Distribution

Degree distribution is often heavy tailed in real world networks

There are many with very small degree & nodes with very high degree

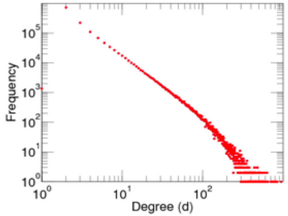
Degree distribution is almost always plotted in **log-log scale** (linear scale plots often show only a single point)

Actor-Movies



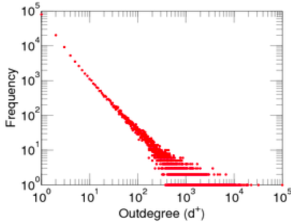
Actor degree distribution

Researcher-Publications



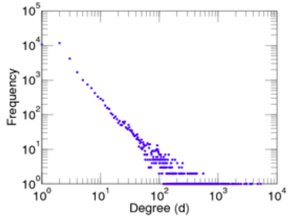
Author degree distribution

Wiki communications



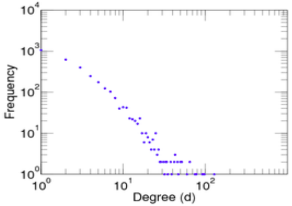
Outdegree distribution

Internet



Degree distribution

Protein Interactions

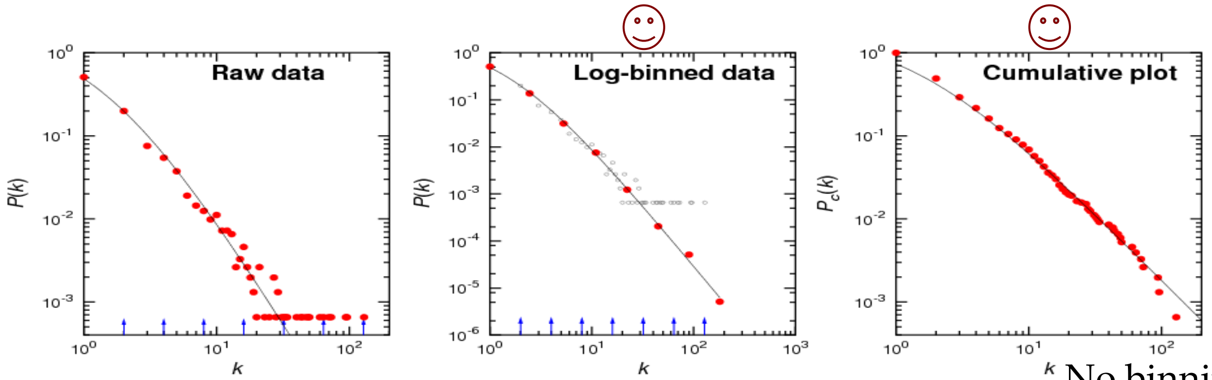


Degree distribution

Pro tip: it is better to (logarithmically) bin the range before plotting

Fitting a power law

- Use a log-log scale & fit a line
- Use logarithmic binning
- (C)CDF is preferred which is also powerlaw ⇒ more accurate exponent
  - $p(x = d) = Cd^{-\alpha} \Rightarrow p(x \leq d) \propto Cd^{1-\alpha}$

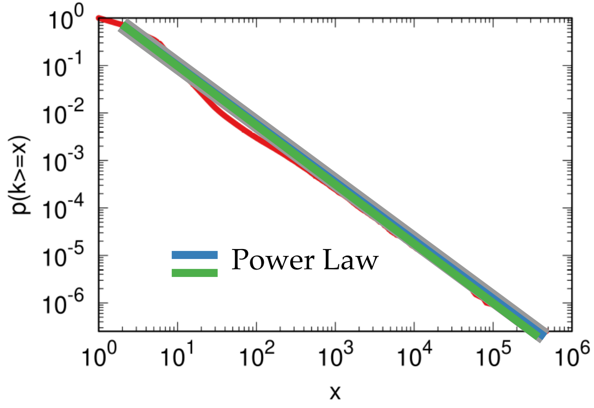


Complementary cumulative degree distribution, the fraction of nodes with degree greater than or equal to d

No binning needed

Fitting a power law

- Linear Fit in log-log
  - Common but debatable and might be misleading, e.g., here both distributions have a very good R2 and p-value because of log-log scale!
- Statistical Tests
  - For example, one tool based on log-likelihood, i.e., how likely is function f to fit the data? Allows p-value estimation between two alternatives: https://aaronclauset.github.io/powerlaws/



From Cosia's slides

What can create a powerlaw?

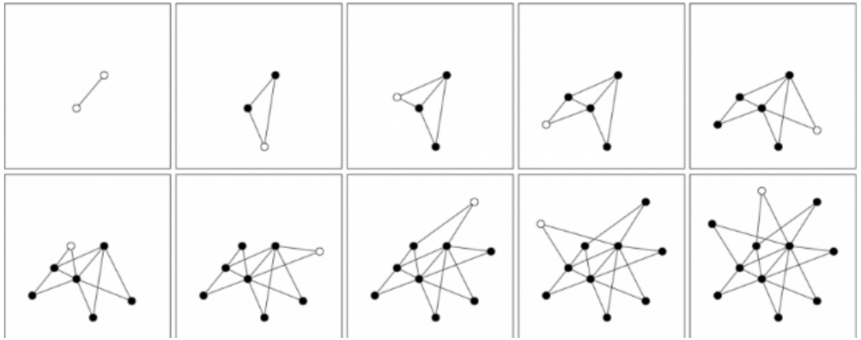
Preferential Attachment

a.k.a rich get richer, accumulative advantage, Yule process, Matthew effect

Albert Barabasi Model (AB)

- A simple graph generation process that adds one node at each iteration & connects it to m existing nodes, hence making m new connections
- the probability of forming a connection to an existing node is proportional to its degree

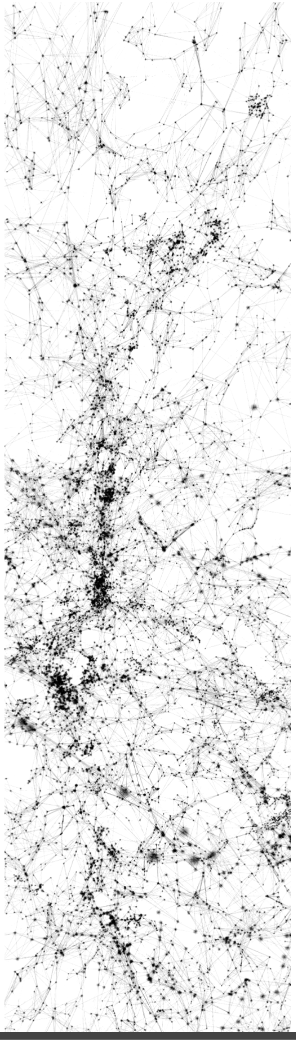
$$p(i) \propto d_i$$



What is m here? 2

Outline

- Sparsity Pattern
- Scale Free Pattern
  - Power-law degree distribution
  - Fitting a power-law
  - Preferential attachment and AB model
- **Assortativity Pattern**
- Transitivity Pattern
  - powers of A & counting triangles
- Small world Pattern
  - Shortest path
- How to pattern?



Degree Assortativity

marginals of A => degree sequence

For undirected graphs: $d_i = \sum_j A_{ij}$

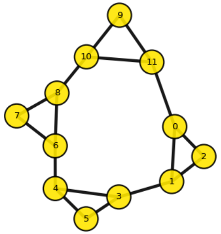
The degree sequence gives degrees of all nodes:

	0	1	2	3	4	5	6	7	8	9	10	11
$D =$	3	3	2	3	3	2	3	2	3	2	3	3

What are the patterns of how node connect?

Is there any relation between degree of neighbouring nodes?

Do popular people mingle together?



	0	1	2	3	4	5	6	7	8	9	10	11	
0	0	1	1	0	0	0	0	0	0	0	0	1	3
1	1	0	1	1	0	0	0	0	0	0	0	0	3
2	1	1	0	0	0	0	0	0	0	0	0	0	2
3	0	1	0	0	1	1	0	0	0	0	0	0	3
4	0	0	0	1	0	1	1	0	0	0	0	0	3
5	0	0	0	1	1	0	0	0	0	0	0	0	2
6	0	0	0	0	1	0	0	1	1	0	0	0	3
7	0	0	0	0	0	0	1	0	1	0	0	0	2
8	0	0	0	0	0	0	1	1	0	0	1	0	3
9	0	0	0	0	0	0	0	0	0	0	1	1	2
10	0	0	0	0	0	0	0	0	1	1	0	1	3
11	1	0	0	0	0	0	0	0	0	1	1	0	3
	3	3	2	3	3	2	3	2	3	2	3	3	

$(d_i, d_j) \forall (i, j) \in E$

{ (3, 3), (3, 2), (3, 3),
(3, 3), (3, 2), (3, 3),
(2, 3), (2, 3),
(3, 3), (3, 3), (3, 2),
(3, 3), (3, 2), (3, 3),
(2, 3), (2, 3),
(3, 3), (3, 2), (3, 3),
(2, 3), (2, 3),
(3, 3), (3, 2), (3, 3)
(2, 3), (2, 3),
(3, 3), (3, 2), (3, 3),
(3, 3), (3, 2), (3, 3) }

E

{ (0, 1), (0, 2), (0, 11),
(1, 0), (1, 2), (1, 3),
(2, 0), (2, 1),
(3, 1), (3, 4), (3, 5),
(4, 3), (4, 5), (4, 6),
(5, 3), (5, 4),
(6, 4), (6, 7), (6, 8),
(7, 8), (7, 6),
(8, 6), (8, 7), (8, 10)
(9, 10), (9, 11),
(10, 8), (10, 9), (10, 11),
(11, 0), (11, 9), (11, 10) }

Degree Assortativity

Strong correlation between degree of connecting nodes

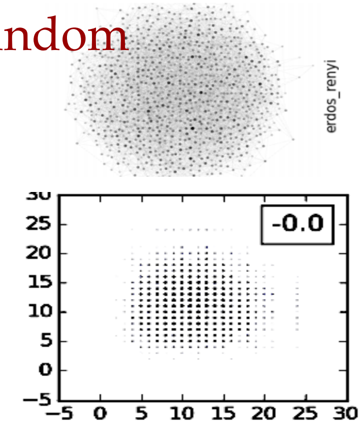
- For all edges, look at degrees of endpoints
  - Either nodes tend to connect to similar degree nodes or dissimilar

assortative mixing

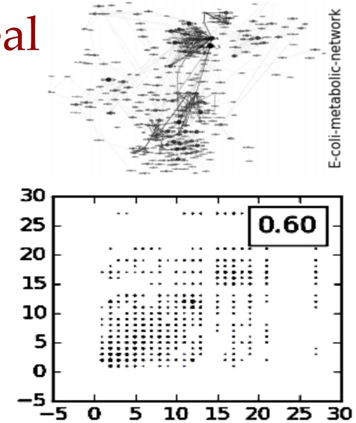
$M = (d_i, d_j) \forall (i, j) \in E$

{ (3, 3), (3, 2), (3, 3),
(3, 3), (3, 2), (3, 3),
(2, 3), (2, 3),
(3, 3), (3, 3), (3, 2),
(3, 3), (3, 2), (3, 3),
(2, 3), (2, 3),
(3, 3), (3, 2), (3, 3),
(2, 3), (2, 3),
(3, 3), (3, 2), (3, 3)
(2, 3), (2, 3),
(3, 3), (3, 2), (3, 3),
(3, 3), (3, 2), (3, 3) }

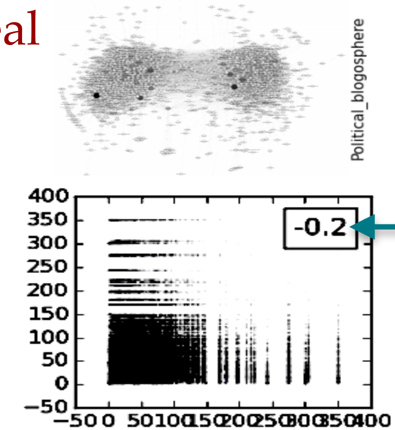
random



real



real



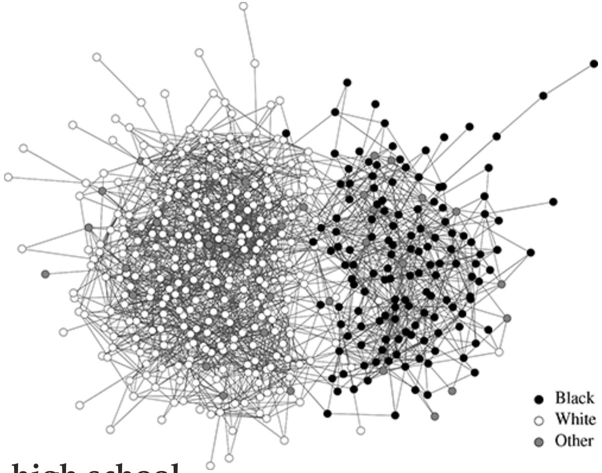
Pearson correlation

M[0, :] & M[1, :]

Assortativity & Mixing Patterns

Strong correlation between some properties of connecting nodes

- For all edges, look at property of endpoints
  - Either nodes tend to connect to similar nodes or dissimilar



A friendship network at a US high school.

We will discuss homophily later in the course

assortative mixing

$M = (f_i, f_j) \forall (i, j) \in E$

{ (3, 3), (3, 2), (3, 3),
(3, 3), (3, 2), (3, 3),
(2, 3), (2, 3),
(3, 3), (3, 3), (3, 2),
(3, 3), (3, 2), (3, 3),
(2, 3), (2, 3),
(3, 3), (3, 2), (3, 3),
(2, 3), (2, 3),
(3, 3), (3, 2), (3, 3)
(2, 3), (2, 3),
(3, 3), (3, 2), (3, 3),
(3, 3), (3, 2), (3, 3) }

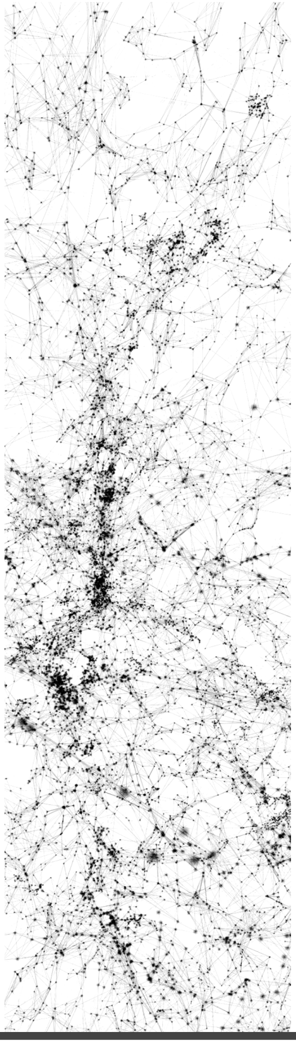
Pearson correlation

M[0, :] & M[1, :]

Valid when the property is ordered

Outline

- Sparsity Pattern
- Scale Free Pattern
  - Power-law degree distribution
  - Fitting a power-law
  - Preferential attachment and AB model
- Assortativity Pattern
- **Transitivity Pattern**
  - powers of A & counting triangles
- Small world Pattern
  - Shortest path
- How to pattern?



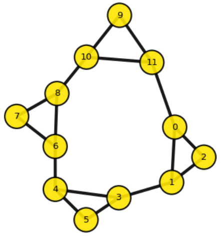
Adjacency Matrix: marginals

marginals of A => degree sequence

For undirected graphs: we have $A_{ij} = A_{ji} = 1$ if there is an edge between i and j, and degree of each node is:

$$d_i = \sum_j A_{ij}$$

What is $\sum_{ij} A_{ij}$? $\sum d_i = 2E$ twice the number of edges



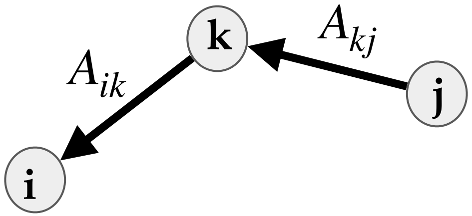
	0	1	2	3	4	5	6	7	8	9	10	11	degrees
0	0	1	1	0	0	0	0	0	0	0	0	1	3
1	1	0	1	1	0	0	0	0	0	0	0	0	3
2	1	1	0	0	0	0	0	0	0	0	0	0	2
3	0	1	0	0	1	1	0	0	0	0	0	0	3
4	0	0	0	1	0	1	1	0	0	0	0	0	3
5	0	0	0	1	1	0	0	0	0	0	0	0	2
6	0	0	0	0	1	0	0	1	1	0	0	0	3
7	0	0	0	0	0	0	1	0	1	0	0	0	2
8	0	0	0	0	0	0	1	1	0	0	1	0	3
9	0	0	0	0	0	0	0	0	0	0	1	1	2
10	0	0	0	0	0	0	0	0	1	1	0	1	3
11	1	0	0	0	0	0	0	0	0	1	1	0	3
degrees	3	3	2	3	3	2	3	2	3	2	3	3	

$N = 12$, $E = 16$

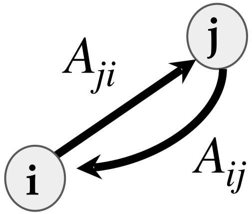
Powers of A

A^2 : number of walks with length two

$$A^2_{ij} = \sum_k A_{ik}A_{kj}$$



- **If undirected:**
  - **What is A^2_{ij} ?** number of common neighbours
  - **What is A^2_{ii} ?** number of neighbours = degree
- **What is A^2_{ii} in directed graph?** number of reciprocal neighbours



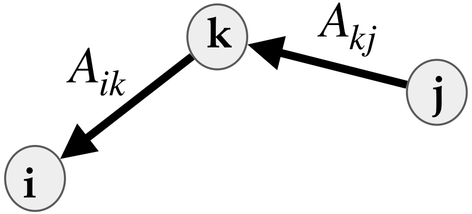
network's reciprocity

$$\frac{\sum_{ij} A_{ij}A_{ji}}{\sum_{ij} A_{ij}}$$

Powers of A

A^2 : number of walks with length two

$$A^2_{ij} = \sum_k A_{ik} A_{kj}$$

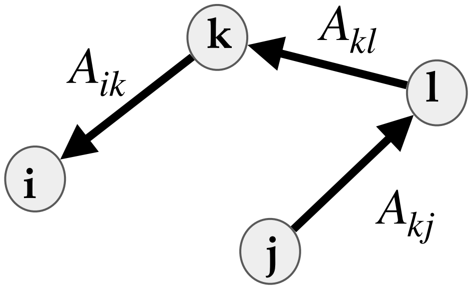


A^3 : number of walks with length three

Is it same as number of paths?

$$A^3_{ij} = \sum_{kl} A_{ik} A_{kl} A_{lj}$$

- A **walk** is a finite or infinite sequence of edges which joins a sequence of vertices.
- A **trail** is a walk in which all edges are distinct.
- A **path** is a trail in which all vertices are distinct.

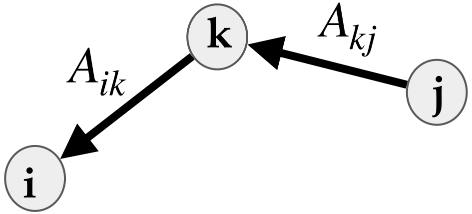


https://en.wikipedia.org/wiki/Path_(graph_theory)#Walk,_trail,_path

Powers of A

A^2 : number of walks with length two

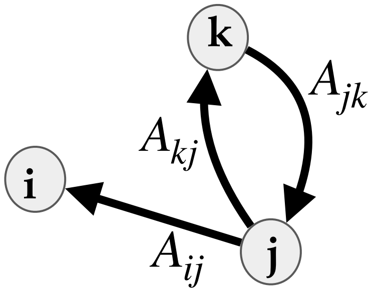
$$A_{ij}^2 = \sum_k A_{ik} A_{kj}$$



A^3 : number of walks with length three

Is it same as number of paths? No!

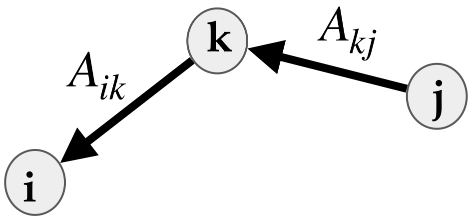
$$A_{ij}^3 = \sum_{kl} A_{ik} A_{kl} A_{lj}$$



Powers of A

A^2 : number of walks with length two

$$A_{ij}^2 = \sum_k A_{ik}A_{kj}$$

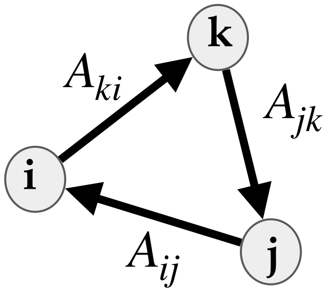


A^3 : number of walks with length three

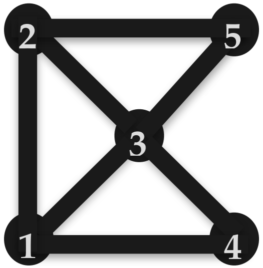
$$A_{ij}^3 = \sum_{kl} A_{ik}A_{kl}A_{lj}$$

What is A_{ii}^3 if graph is undirected?

Twice the Number of Triangles



Toy Example




```
import networkx as nx
G = nx.random_geometric_graph(5, 0.5)
A = nx.adjacency_matrix(G).todense()
print A
A2 = A*A
print A2
A3 = A2*A
print A3
```

A

```
[[0 1 1 1 0]
 [1 0 1 0 1]
 [1 1 0 1 1]
 [1 0 1 0 0]
 [0 1 1 0 0]]
```

A^2

```
[[3 1 2 1 2]
 [1 3 2 2 1]
 [2 2 4 1 1]
 [1 2 1 2 1]
 [2 1 1 1 2]]
```

common neighbours

degrees

A^3

```
[[4 7 7 5 3]
 [7 4 7 3 5]
 [7 7 6 6 6]
 [5 3 6 2 3]
 [3 5 6 3 2]]
```

walks of length 3

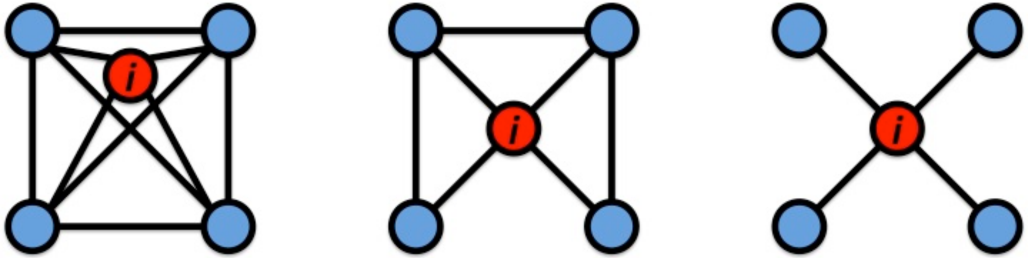
triangles x 2

Clustering Coefficient

Local clustering coefficient is defined per node:

$$c_i = \frac{A_{ii}^3}{d_i(d_i - 1)}$$

Shows how well connected the node's neighbourhood is:



$C_i = 1$ $C_i = 1/2$ $C_i = 0$

Clustering Coefficient

measures the density of triangles

Local clustering coefficient is defined per node:

$$c_i = \frac{A^3_{ii}}{d_i(d_i - 1)}$$

, then averaged over all nodes in the graph

Global clustering coefficient is defined for the whole graph:

$$c = \frac{triangles}{triangbles}$$

number of all triangles in the graph

number of all length two walks that can be a triangle if endpoints are connected

How can we measure total number of triangles in an undirected graph?

Clustering Coefficient

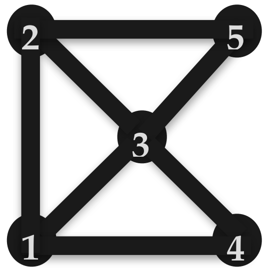
measures the density of triangles

Global clustering coefficient is defined for the whole graph:

$$c = \frac{triangles}{triangbles}$$

triangles: number of all triangles in the graph

triangbles: number of all length two walks that can be a triangle if endpoints are connected



How can we measure total number of triangles in an undirected graph?

$$Tr(A^3)/6$$

A^2

[[3 1 2 1 2]
[1 3 2 2 1]
[2 2 4 1 1]
[1 2 1 2 1]
[2 1 1 1 2]]

common neighbours

degrees

A^3

[[4 7 7 5 3]
[7 4 7 3 5]
[7 7 6 6 6]
[5 3 6 2 3]
[3 5 6 3 2]]

walks of length 3

triangles x 2

Clustering Coefficient

measures the density of triangles

Local clustering coefficient is defined per node:

$$c_i = \frac{A^3_{ii}}{d_i(d_i - 1)}$$

, then averaged over all nodes in the graph

Global clustering coefficient is defined for the whole graph:

$$c = \frac{Tr(A^3)}{Sum(A^2) - Tr(A^2)}$$

Do they give the same results?

Clustering Coefficient

measures the density of triangles

Local clustering coefficient is defined per node:

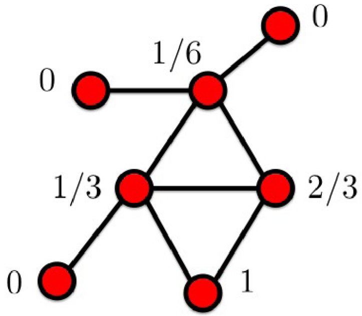
$$c_i = \frac{A^3_{ii}}{d_i(d_i - 1)}$$

, then averaged over all nodes in the graph

Global clustering coefficient is defined for the whole graph:

$$c = \frac{Tr(A^3)}{Sum(A^2) - Tr(A^2)}$$

Do they give the same results?



They differ:

$\langle C \rangle = \frac{13}{42} \approx 0.310$: Local average

$C = \frac{3}{8} = 0.375$: Global

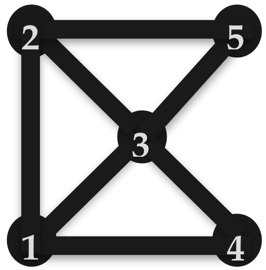
Clustering Coefficient measures the density of triangles

Global clustering coefficient is defined for the whole graph:

$$c = \frac{triangles}{triangbles}$$

triangles: number of all triangles in the graph

triangbles: number of all length two walks that can be a triangle if endpoints are connected



How can we measure total number of triangles in an undirected graph? $Tr(A^3)/6$

Can we compute number of triangles more efficiently?

since Tr(A)=$\sum_i \lambda_i$, and if λ is eigenvalue of A then λ^p is eigenvalue of A^p

Yes, from eigenvalues of A as $\frac{1}{6}\sum_i \lambda_i^3$

We can approximate this with using only top eigenvalues since this distribution is skewed

There are many works on approximating number of triangles in large graphs

A^2

[[**3** 1 2 1 2]
[1 **3** 2 2 1]
[2 2 **4** 1 1]
[1 2 1 **2** 1]
[2 1 1 1 **2**]]

common neighbours

degrees

A^3

[[**4** 7 7 5 3]
[7 **4** 7 3 5]
[7 7 **6** 6 6]
[5 3 6 **2** 3]
[3 5 6 3 **2**]]

walks of length 3

triangles x 2

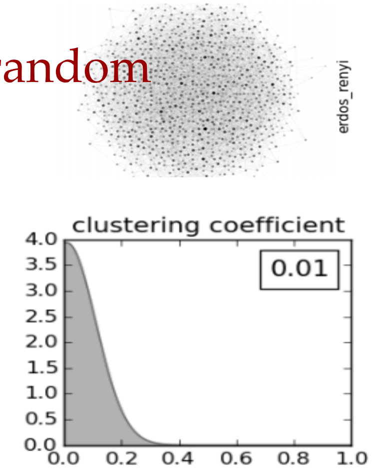
Transitivity Pattern

Real networks have a lot of triangles and strong transitivity

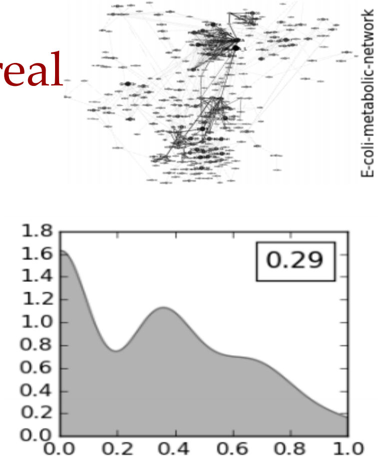
e.g. Friends of friends are friends

- High global clustering coefficient or high average local clustering coefficient
- Distribution of local clustering coefficient

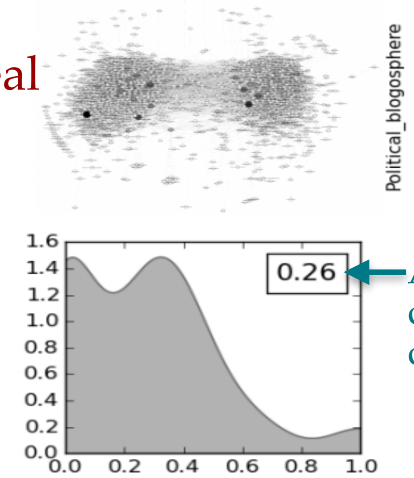
random



real



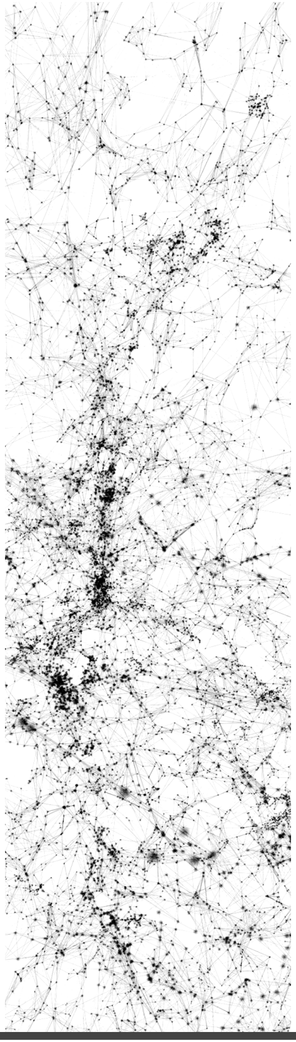
real



Average clustering coefficient

Outline

- Sparsity Pattern
- Scale Free Pattern
  - Power-law degree distribution
  - Fitting a power-law
  - Preferential attachment and AB model
- Assortativity Pattern
- Transitivity Pattern
  - powers of A & counting triangles
- **Small world Pattern**
  - Shortest path
- How to pattern?



Derived from the Adjacency matrix

network measure	scope	graph	definition	explanation
degree	L	U	$k_i = \sum_{j=1}^{n} A_{ij}$	number of edges attached to vertex i
in-degree	L	D	$k_i^{\text{in}} = \sum_{j=1}^{n} A_{ji}$	number of arcs terminating at vertex i
out-degree	L	D	$k_i^{\text{out}} = \sum_{j=1}^{n} A_{ij}$	number of arcs originating from vertex i
edge count	G	U	$m = \frac{1}{2} \sum_{ij} A_{ij}$	number of edges in the network
arc count	G	D	$m = \sum_{ij} A_{ij}$	number of arcs in the network
mean degree	G	U	$\langle k \rangle = 2m/n = \frac{1}{n} \sum_{i=1}^{n} k_i$	average number of connections per vertex
mean in- or out-degree	G	D	$\langle k^{\text{in}} \rangle = \langle k^{\text{out}} \rangle = 2m/n$	average number of in- or out-connections per vertex
reciprocity	G	D	$r = \frac{1}{m} \sum_{ij} A_{ij} A_{ji}$	fraction of directed edges that are reciprocated
reciprocity	L	D	$r_i = \frac{1}{k_i} \sum_{j} A_{ij} A_{ji}$	fraction of directed edges from i that are reciprocated
clustering coefficient	G	U	$c = \frac{\sum_{ijk} A_{ij} A_{jk} A_{ki}}{\sum_{ijk} A_{ij} A_{jk}}$	the network's triangle density
clustering coefficient	L	U	$c_i = \sum_{jk} A_{ij} A_{jk} A_{ki} \Big/ \binom{k_i}{2}$	fraction of pairs of neighbors of i that are also connected
diameter	G	U	$d = \max_{ij} \ell_{ij}$	length of longest geodesic path in an undirected network
mean geodesic distance	G	U or D	$\ell = \frac{1}{\binom{n}{2}} \sum_{ij} \ell_{ij}$	average length of a geodesic path
eccentricity	G	U or D	$\epsilon_i = \max_i \ell_{ij}$	length of longest geodesic path starting from i

From Clauset's slides

Shortest Path

Single-source shortest paths

- All shortest paths for a single node can be computed with BFS when graph is simple (unweighted, undirected), time complexity is linear in number of edges, i.e., $\mathcal{O}(E)$, assuming $E > V$
- There are alternatives that also work for weighted graphs: Dijkstra's algorithm($\mathcal{O}(E + VlogV)$), Bellman–Ford algorithm ($\mathcal{O}(VE)$)

All-pairs shortest paths

- Floyed-Warshall algorithm: $\mathcal{O}(V^3)$

https://en.wikipedia.org/wiki/Shortest_path_problem

In real world graph V and E are in the same order so there is not much difference between algorithms.

We often care about the longest & average shortest paths

Small average shortest path

Shortest path distribution is normal with small [shrinking] average in real world

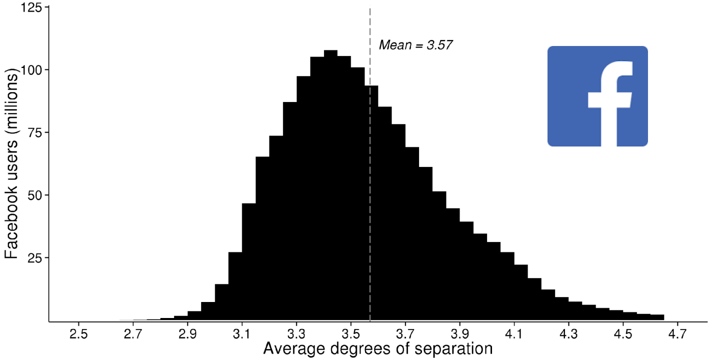
You can reach any node in a graph passing through few hubs

This is often referred to as small world

Diameter is also small {longest sp}

Stanley Milgram
(1933-1984)

Letter-passing experiment,
In 1967 discovered the
Six Degrees of Separation

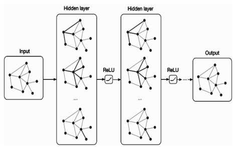


Four Degrees of Separation
You are 4 hops away from
anyone in the planet

Recent Trend: Deep Learning for Graphs

2015~2021

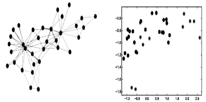
- **Graph Neural Networks**
- Deep Learning for Networks
- High-Order Networks [Benson et al.]



21st Century: More CS

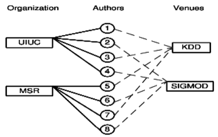
2010~2014

- Info. vs. Social Networks (Twitter) [Kwak et al.]
- **Signed** Networks [Leskovec et al.]
- Semantic Social Networks [Tang et al.]
- Four Deg. Of Separation [Backstrom et al.]
- Structural Diversity [Ugander et al.]
- Computational Social Science [Watts]
- **Network Embedding** [Perozzi et al.]



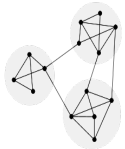
2005~2009

- Graph Evolution [Leskovec et al.]
- 3 Deg. Of Influence [Christakis & Fowler]
- Social **Influence** Analysis [Tang et al.]
- Six Deg. Of Separation [Leskovec & Horvitz]
- Network **Heterogeneity** [Sun & Han]
- Network **Embedding** [Tang & Liu]
- Computer Social Science [Lazer et al.]



2000~2004

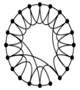
- Influence Max'n [Domingos & Kempe et al.]
- **Community Detection** [Girvan & Newman]
- Network Motifs [Milo et al.]
- Link Prediction [Liben-Nowell & Kleinberg]

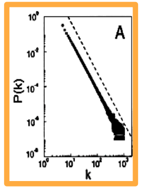


Late 20th Century: CS & Physics

1998~1999

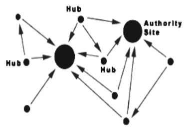
- **Small Worlds** [Watts & Strogatz]
- **Scale Free** [Barabasi & Albert]
- **Power Law** [Faloutsos x3]





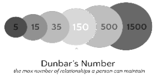
1997

- **HITS** [Kleinberg]
- **PageRank** [Page & Brin]
- Hyperlink Vector Voting [Li]



1992

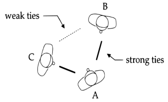
- Structural Hole [Burt]
- **Dunbar's Number** [Dunbar]



20th Century: Sociology

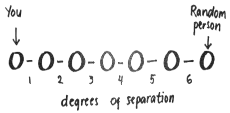
1970s

- The Strength Of **Weak Tie** [Granovetter]



1960s

- **Small Worlds** [Migram]

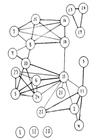


1950s

- **Random Graph** [Erdos, Renyi, Gilbert]
- Degree Sequence [Tuttle, Havel, Hakami]
- **Homophily** [Lazarsfeld & Merton]
- Balance Theory [Heider et al.]

1930s

- **Sociogram** [Moreno]



Based on Slides from Jie Tang

Pattern Detection

- WHY?
  - Understand the language of complex systems
  - Characterize different types of networks
  - Design {efficient} data structure & algorithms
  - Tangled with Measurements, Anomaly detection, Modelling
- HOW?
  - What do networks have in common?
  - How to measure or characterize (nodes, communities, whole) networks?
  - What are universal patterns observed in real world networks?
  - What is structure of real-world networks?

	Network	Type	n	m	c	S	ℓ	α	C	C_{WS}	r
Social	Film actors	Undirected	449913	25516482	113.43	0.980	3.48	2.3	0.20	0.78	0.208
	Company directors	Undirected	7 673	55392	14.44	0.876	4.60	–	0.59	0.88	0.276
	Math coauthorship	Undirected	253339	496489	3.92	0.822	7.57	–	0.15	0.34	0.120
	Physics coauthorship	Undirected	52909	245300	9.27	0.838	6.19	–	0.45	0.56	0.363
	Biology coauthorship	Undirected	1 520251	11803064	15.53	0.918	4.92	–	0.088	0.60	0.127
	Telephone call graph	Undirected	47000000	80000000	3.16			2.1			
	Email messages	Directed	59812	86300	1.44	0.952	4.95	1.5/2.0		0.16	
	Email address books	Directed	16881	57029	3.38	0.590	5.22	–	0.17	0.13	0.092
	Student dating	Undirected	573	477	1.66	0.503	16.01	–	0.005	0.001	−0.029
	Sexual contacts	Undirected	2 810					3.2			
Biological	Metabolic network	Undirected	765	3 686	9.64	0.996	2.56	2.2	0.090	0.67	−0.240
	Protein interactions	Undirected	2 115	2 240	2.12	0.689	6.80	2.4	0.072	0.071	−0.156
	Marine food web	Directed	134	598	4.46	1.000	2.05	–	0.16	0.23	−0.263
	Freshwater food web	Directed	92	997	10.84	1.000	1.90	–	0.20	0.087	−0.326
	Neural network	Directed	307	2 359	7.68	0.967	3.97	–	0.18	0.28	−0.226

Table 10.1
NS book

c: average degree
s: fraction of nodes in the largest component
l: average shortest path of connected nodes
α: powerlaw slope
C: global clustering coefficient
c_{WS}: average local clustering coefficient
r: degree correlation

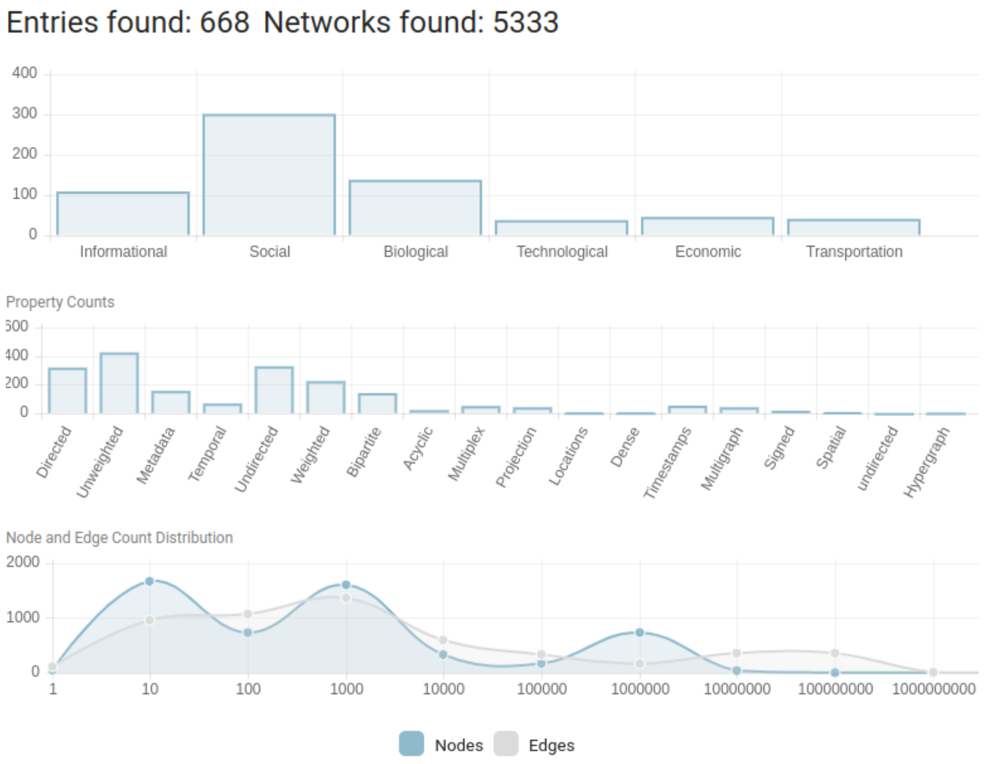
{common} Network Repositories

1. Newman's collection
2. Stanford Large Network Dataset Collection
3. The Colorado Index of Complex Networks (ICON)
4. The Koblenz Network Collection

From Clauset's slides

{common} Network Repositories

1. Newman's collection
2. Stanford Large Network Dataset Collection
3. **The Colorado Index of Complex Networks (ICON)**
4. The Koblenz Network Collection



{common} Network Repositories

1. Newman's collection
2. Stanford Large Network Dataset Collection
3. The Colorado Index of Complex Networks (ICON)
4. **The Koblenz Network Collection**

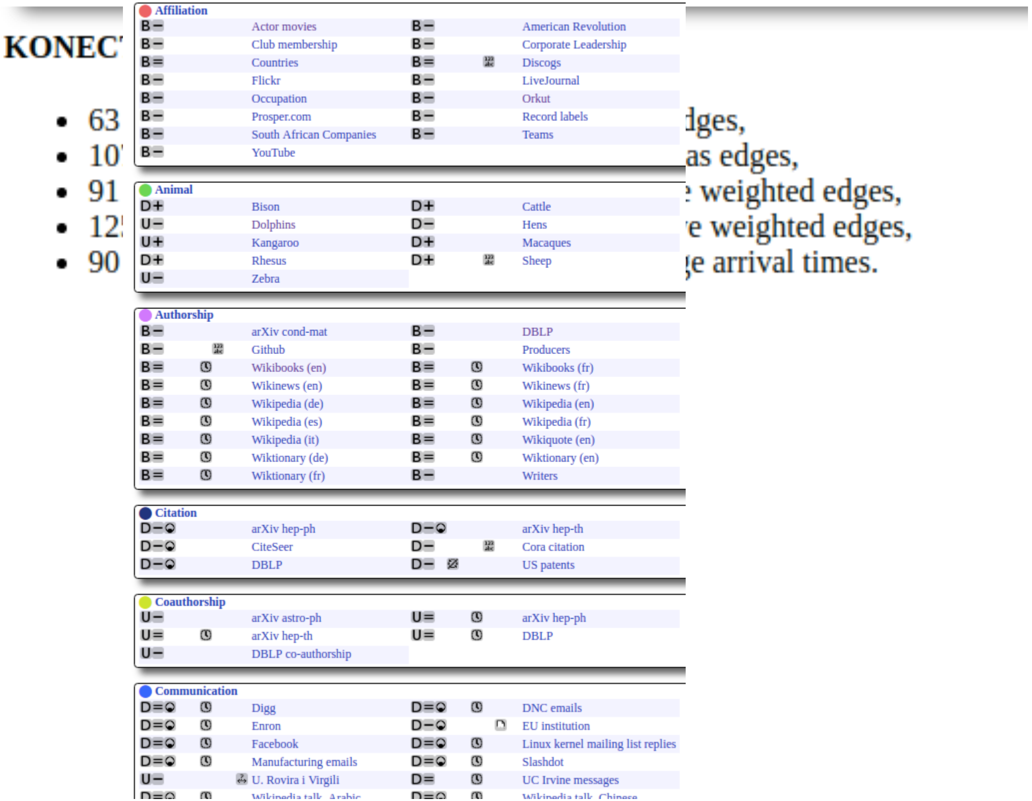
Let us know in slack if you come across other large repos

KONECT currently holds 261 networks, of which

- 63 are undirected,
- 107 are directed,
- 91 are bipartite,
- 125 are unweighted,
- 90 allow multiple edges,
- 6 have signed edges,
- 10 have ratings as edges,
- 3 allow multiple weighted edges,
- 18 allow positive weighted edges,
- and 89 have edge arrival times.

{common} Network Repositories

1. Newman's collection
2. Stanford Large Network Dataset Collection
3. The Colorado Index of Complex Networks (ICON)
4. **The Koblenz Network Collection**



Affiliation

B−	Actor movies	B−	American Revolution
B−	Club membership	B−	Corporate Leadership
B=	Countries	B=	Discogs
B−	Flickr	B−	LiveJournal
B=	Occupation	B=	Orkut
B−	Prosper.com	B−	Record labels
B=	South African Companies	B−	Teams
B−	YouTube		

Animal

D+	Bison	D+	Cattle
U−	Dolphins	D−	Hens
U+	Kangaroo	D+	Macaques
D+	Rhesus	D+	Sheep
U−	Zebra		

Authorship

B−	arXiv cond-mat	B−	DBLP
B−	Github	B−	Producers
B=	Wikibooks (en)	B=	Wikibooks (fr)
B=	Wikinews (en)	B=	Wikinews (fr)
B=	Wikipedia (de)	B=	Wikipedia (en)
B=	Wikipedia (es)	B=	Wikipedia (fr)
B=	Wikipedia (it)	B=	Wikiquote (en)
B=	Wiktionary (de)	B=	Wiktionary (en)
B=	Wiktionary (fr)	B−	Writers

Citation

D−	arXiv hep-ph	D−	arXiv hep-th
D−	CiteSeer	D−	Cora citation
D−	DBLP	D−	US patents

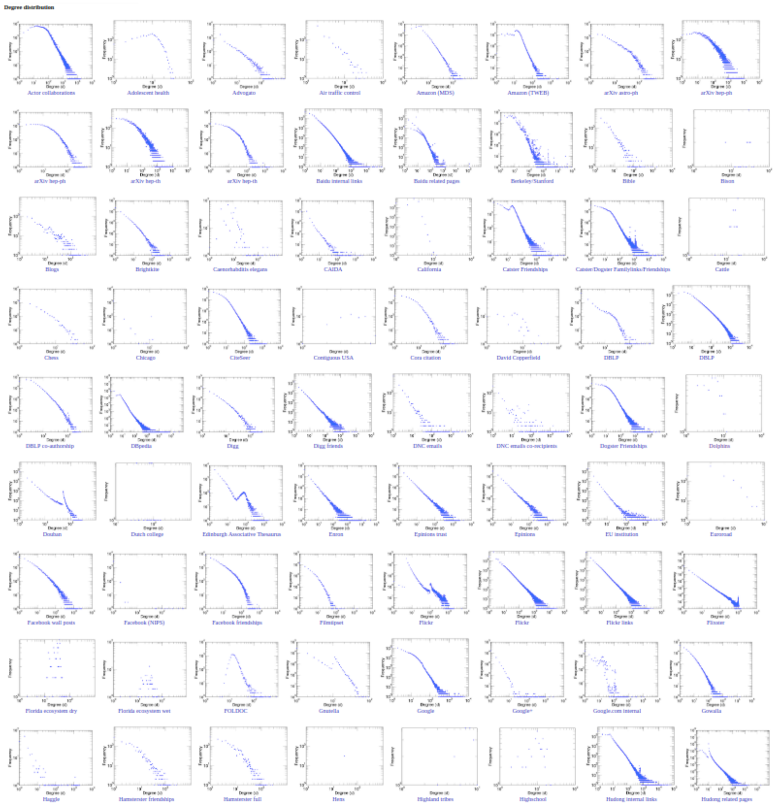
Coauthorship

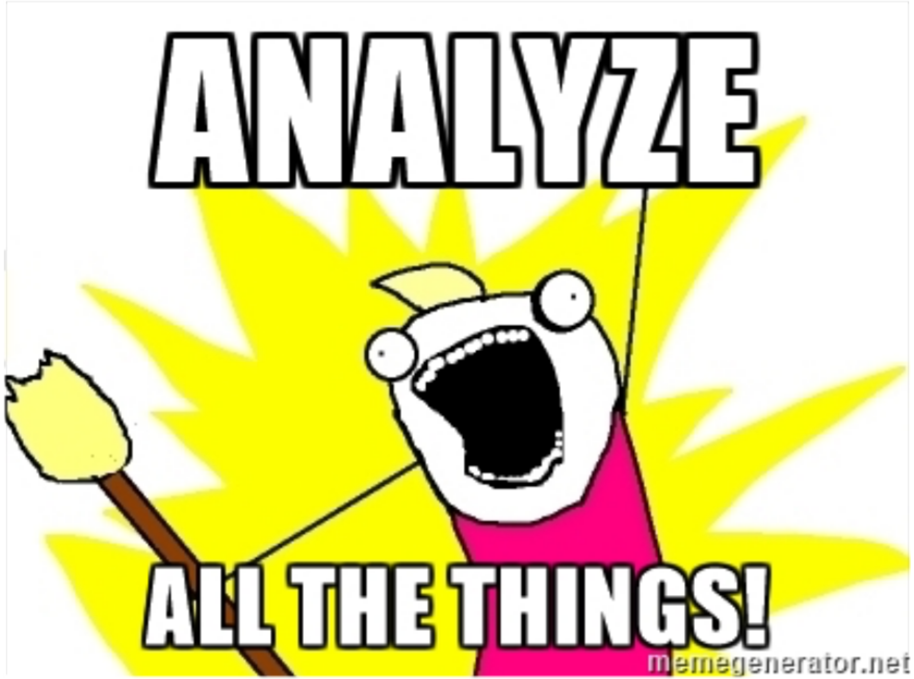
U−	arXiv astro-ph	U=	arXiv hep-ph
U=	arXiv hep-th	U=	DBLP
U−	DBLP co-authorship		

Communication

D=	Digg	D=	DNC emails
D=	Enron	D−	EU institution
D=	Facebook	D=	Linux kernel mailing list replies
D=	Manufacturing emails	D=	Slashdot
U−	U. Rovira i Virgili	D=	UC Irvine messages
D=	Wikipedia talk, Arabic	D=	Wikipedia talk, Chinese

Hypothesize, analyze & observe

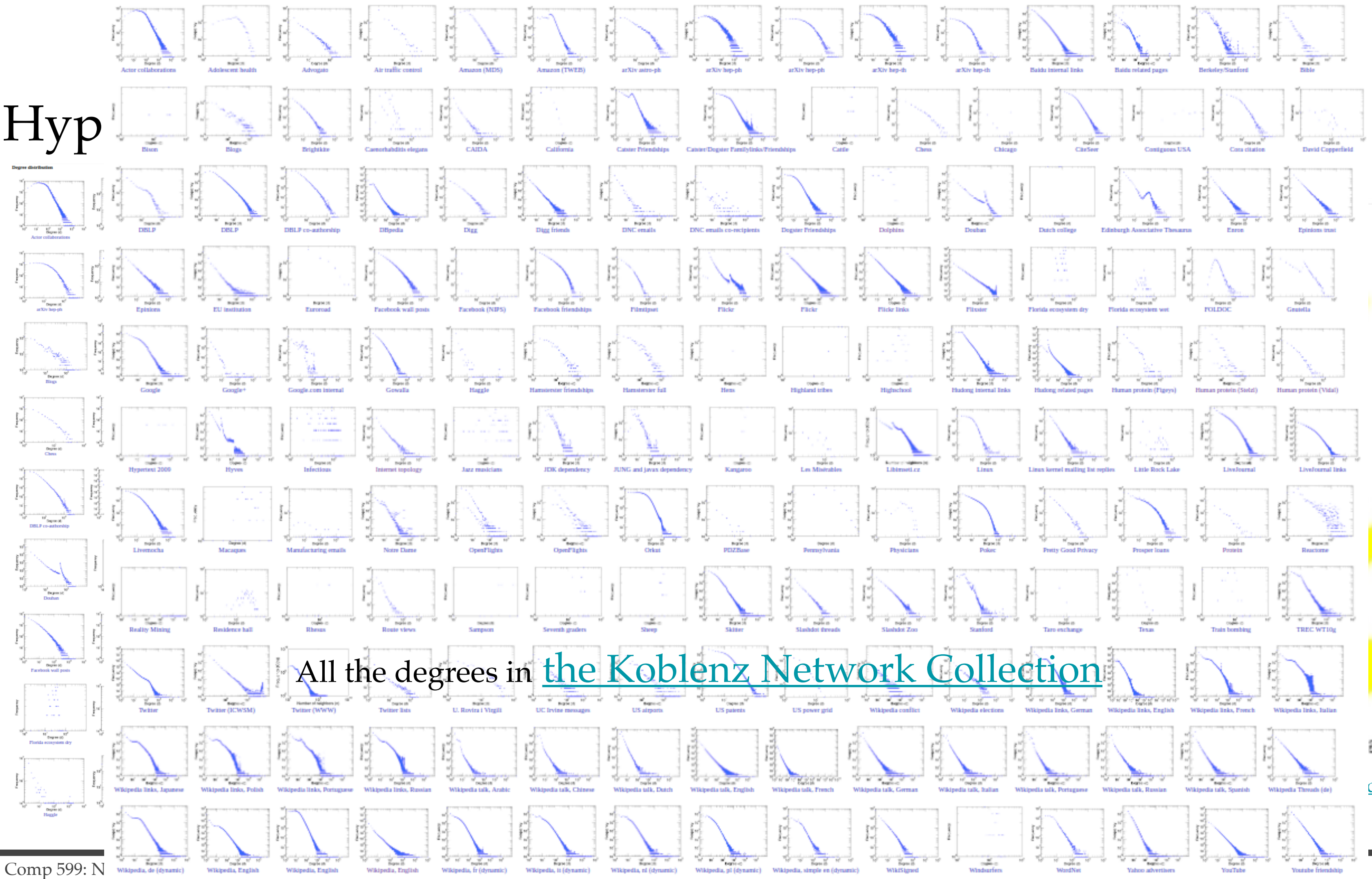




[From Clauset's slides]

[http://konect.cc/plots/degree_distribution]

Hyp



All the degrees in [the Koblenz Network Collection]

ator.net

des

Common benchmark repositories

- Stanford Large Network Dataset Collection (SNAP)
  - Social networks : online social networks, edges represent interactions between people
  - Networks with ground-truth communities : ground-truth network communities in social and information networks
  - Communication networks : email communication networks with edges representing communication
  - Citation networks : nodes represent papers, edges represent citations
  - Collaboration networks : nodes represent scientists, edges represent collaborations (co-authoring a paper)
  - Web graphs : nodes represent webpages and edges are hyperlinks
  - Amazon networks : nodes represent products and edges link commonly co-purchased products
  - Internet networks : nodes represent computers and edges communication
  - Road networks : nodes represent intersections and edges roads connecting the intersections

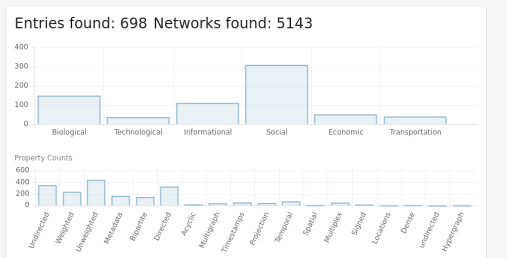
- Network Repository (networkrepository)

Data & Network Collections. Find and interactively VISUALIZE and EXPLORE hundreds of network data

ANIMAL SOCIAL NETWORKS	816	INTERACTION NETWORKS	29	SCIENTIFIC COMPUTING	11
BIOLOGICAL NETWORKS	37	INFRASTRUCTURE NETWORKS	8	SOCIAL NETWORKS	77
BRAIN NETWORKS	116	LABELED NETWORKS	105	FACEBOOK NETWORKS	114
COLLABORATION NETWORKS	20	MASSIVE NETWORK DATA	21	TECHNOLOGICAL NETWORKS	12
CHEMINFORMATICS	646	MISCELLANEOUS NETWORKS	2668	WEB GRAPHS	36
CITATION NETWORKS	4	POWER NETWORKS	8	DYNAMIC NETWORKS	115
ECOLOGY NETWORKS	6	PROXIMITY NETWORKS	13	TEMPORAL REACHABILITY	38
ECONOMIC NETWORKS	16	GENERATED GRAPHS	221	BHOSLIB	36
EMAIL NETWORKS	6	RECOMMENDATION NETWORKS	36	DIMACS	78
GRAPH 500	8	ROAD NETWORKS	15	DIMACS10	84
HETEROGENEOUS NETWORKS	15	RETWEET NETWORKS	34	NON-RELATIONAL ML DATA	211

Check the visualization demo here: https://networkrepository.com/graphvis.php

- The Colorado Index of Complex Networks (ICON)



- The KONECT Project (KONECT)

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...

Gephi, a notable visualization tool: https://gephi.org/users/tutorial-visualization/

More resources

- Listed on the course website

Resources

- Stanford Large Network Dataset Collection [Benchmark Datasets]
- Network Repository [Data + Interactive Visualization and Stats]
- The KONECT Project [Data + Basic Statistics]
- The Colorado Index of Complex Networks (ICON) [Varied Graph Data]
- Open Graph Benchmark [Large Graph Data]
- Networkx [Python Graph Library]
- Deep Graph Library [Benchmark Data + Graph ML Library]
- Pytorch Geometric [Benchmark Data + Graph ML Library]
- Papers with Code on Graph Related Tasks

Example benchmark datasets

NETWORK	NODES	LINKS	DIRECTED UNDIRECTED	N	L
Internet	Routers	Internet connections	Undirected	192,244	609,066
WWW	Webpages	Links	Directed	325,729	1,497,134
Power Grid	Power plants, transformers	Cables	Undirected	4,941	6,594
Mobile Phone Calls	Subscribers	Calls	Directed	36,595	91,826
Email	Email addresses	Emails	Directed	57,194	103,731
Science Collaboration	Scientists	Co-authorship	Undirected	23,133	93,439
Actor Network	Actors	Co-acting	Undirected	702,388	29,397,908
Citation Network	Paper	Citations	Directed	449,673	4,689,479
E. Coli Metabolism	Metabolites	Chemical reactions	Directed	1,039	5,802
Protein Interactions	Proteins	Binding interactions	Undirected	2,018	2,930

You can download these bundled from Barbasi's website, for the first assignment