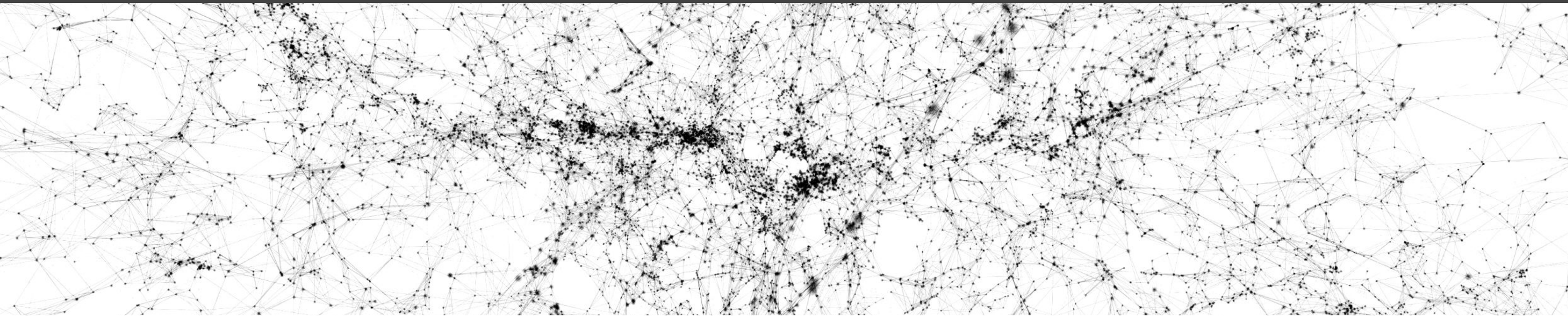
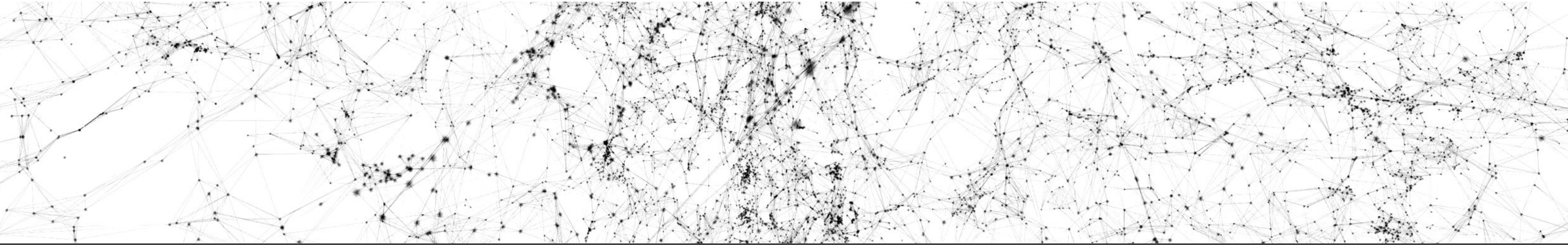


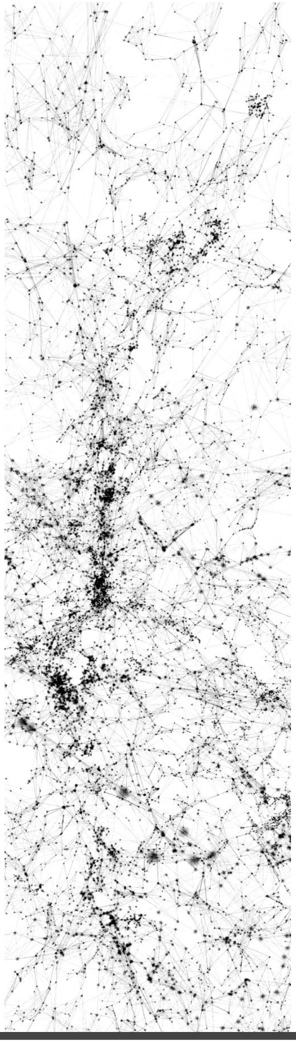
Patterns

Analysis of complex interconnected data



Outline

- **Quick Notes**
  - Assignment 1, slack
- Adjacency matrix and degree
- 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
- Connectivity & eigenvalues of Laplacian matrix
- How to pattern?



Quick Notes

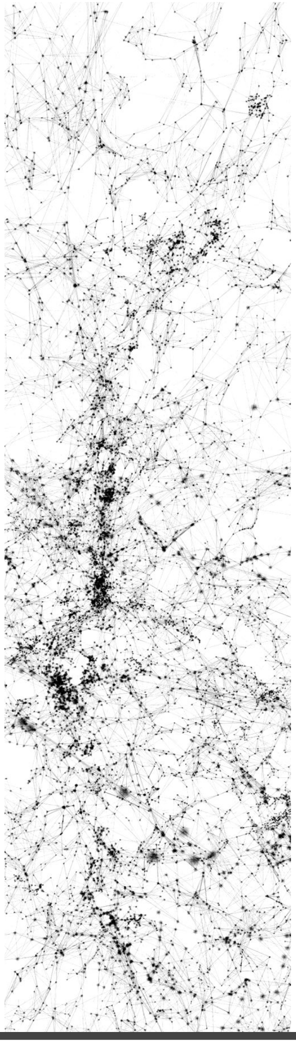
- Reminder, first assignment due in a week
  - http://www.reirab.com/Teaching/NS20/Assignment_1.pdf
  - Any questions from the description?
  - Join a Group in Mycourses
  - Submit the assignment in Mycourses
  - For assignments, 2^k% of the grade will be deducted per k days of delay.
- Use slack for easier communications
  - Let me know if you didn't get an invite
- Anyone new in the class?

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. 25th
- project proposal due on Nov. 1th
- Reviews (first round) due on Nov. 8th
- project progress report due on Nov. 22nd
- Reviews (second round) due on Nov. 29th
- project final report slides due on Dec. 1st
- project final report due on Dec. 6th
- Reviews (third round) due on Dec. 13th
- project revised report and rebuttal due on Dec. 20th
- note: dates are tentative, please check them for the updated deadlines

Outline

- Quick Notes
  - Assignment 1, slack
- **Adjacency matrix and degree**
- 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
- Connectivity & eigenvalues of Laplacian matrix
- How to pattern?



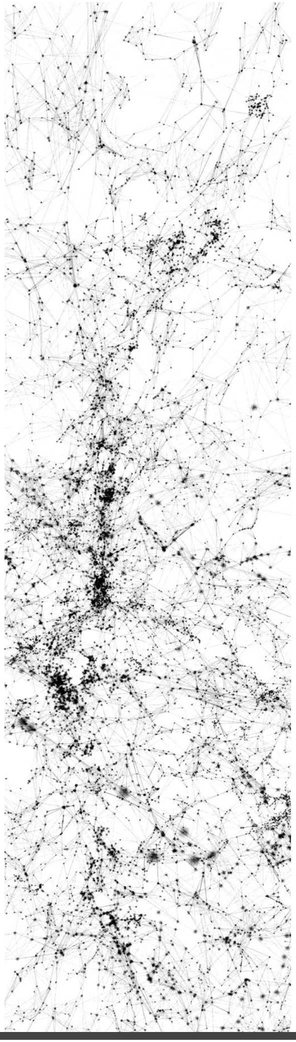
Marginals of Adjacency Matrix

- Marginals of **A** ⇒ Degrees
  - $\mathbf{d_i = \Sigma_j A_{ij}}$
- **Sum(A) = $\Sigma_i\Sigma_j A_{ij} = \Sigma_i d_i$ = ?**
- If directed indegree and outdegree
  - $\mathbf{d^{in}_i = \Sigma_j A_{ji}}$ and $\mathbf{d^{out}_i = \Sigma_j A_{ij}}$

$$\begin{array}{cc} & \begin{array}{cccccccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \end{array} \\ \begin{array}{c} 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \end{array} & \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix} \end{array}$$

Outline

- Quick Notes
  - Assignment 1, slack
- Adjacency matrix and degree
- **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
- Connectivity & eigenvalues of Laplacian matrix
- How to pattern?



Marginals of Adjacency Matrix

- Marginals of **A** ⇒ Degrees
  - $\mathbf{d_i = \Sigma_j A_{ij}}$
- **Sum(A) = $\Sigma_i\Sigma_j A_{ij} = \Sigma_i d_i = 2E$**

  If undirected

$$
\begin{array}{c|c}
 & \begin{array}{cccccccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \end{array} \\
\begin{array}{c} 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \end{array} &
\left[\begin{array}{cccccccccccc}
0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0
\end{array}\right]
\end{array}
$$

- mean degree: **$1/n\ \Sigma_i\Sigma_j A_{ij} = 1/n\ \Sigma_i d_i$**
- Density: **$\Sigma_i\Sigma_j A_{ij}$ / n(n-1)**

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 (MSN IM):	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=$1.51*10^{-5}$, MSN IM= $2.27*10^{-8}$)

Implications?

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 (MSN IM):	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]

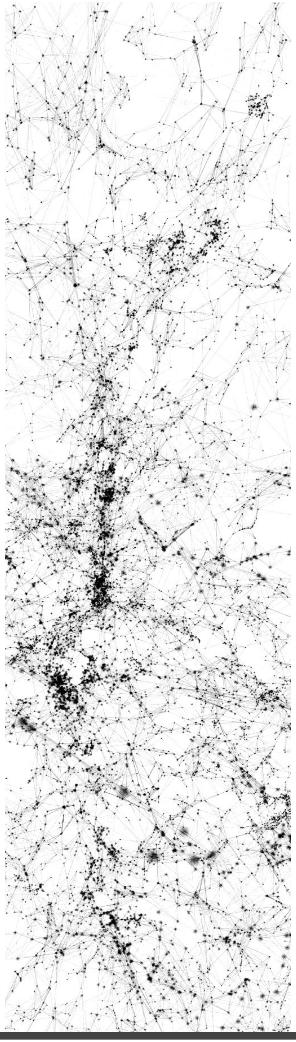
Adjacency matrix is filled with zeros!

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

Implications? Use sparse representations, density not so informative

Outline

- Quick Notes
  - Assignment 1, slack
- Adjacency matrix and degree
- 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
- Connectivity & eigenvalues of Laplacian matrix
- How to pattern?



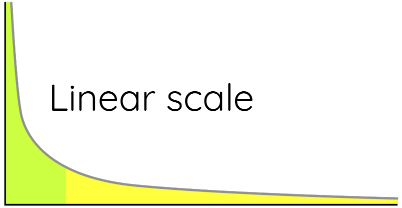
Marginals of Adjacency Matrix

- Marginals of **A** ⇒ Degrees
  - $\mathbf{d_i = \Sigma_j A_{ij}}$

$$
\begin{array}{cc}
 & \begin{array}{cccccccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \end{array} \\
\begin{array}{c} 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \end{array} &
\begin{bmatrix}
0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0
\end{bmatrix}
\end{array}
$$

- **Degree Distribution**
  - shows how many nodes of degree *d* are in the graph
  - degree sequence of all nodes ⇒ count frequencies

Heavy Tailed Degree Distribution

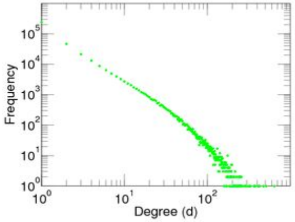


Degree distribution is often **heavy tailed** in real world networks

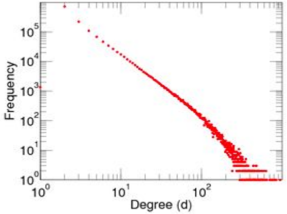
There are few nodes with very high degree & many with very small degree

This is often referred to as being **scale-free**

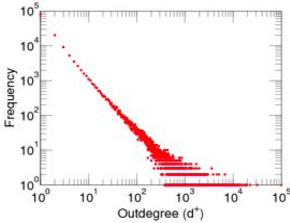
Degree distribution is almost always plotted in log-log scale



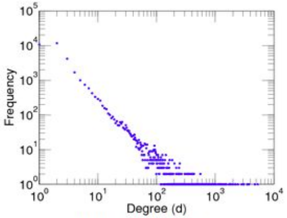
Actor degree distribution



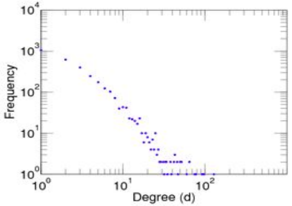
Author degree distribution



Outdegree distribution



Degree distribution



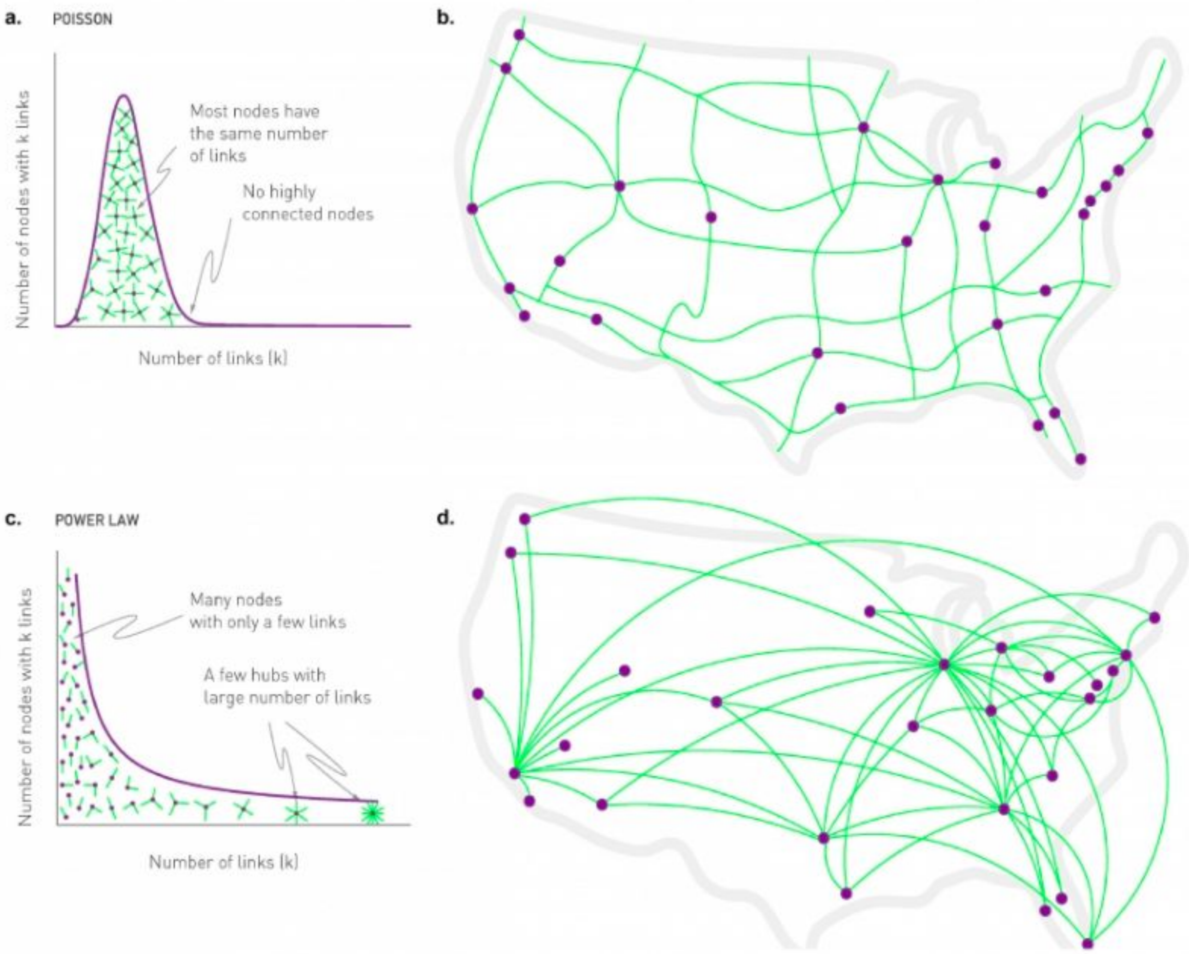
Degree distribution

[Actor-Movies] [Researcher-Publications] [Wiki communications] [Internet] [Protein Interactions]

Example

poisson vs powerlaw degree distribution
highways vs airways

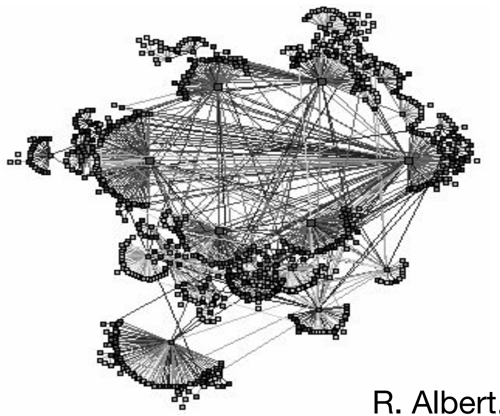
In air-traffic networks, we have major hubs and many smaller airports. In highway networks, cities are of comparable connections.

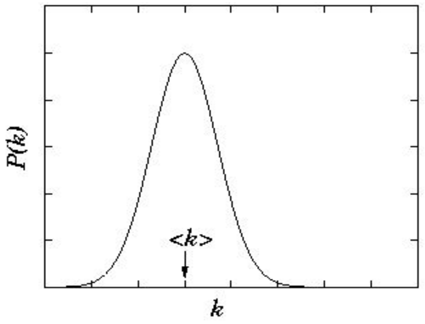


The first observations

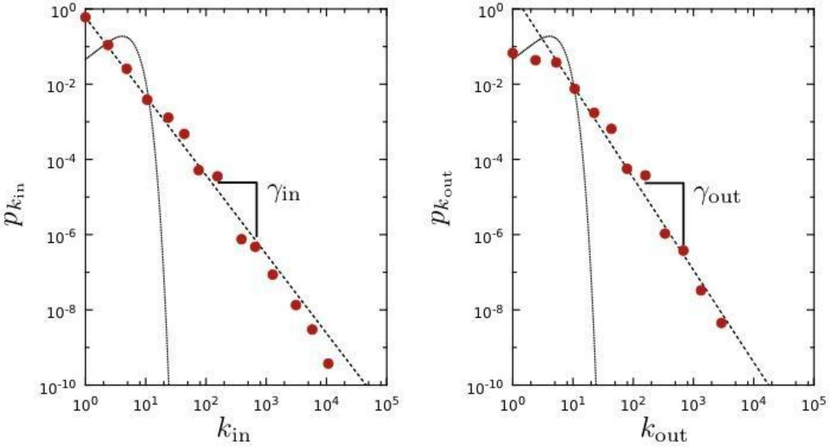
Nodes: **WWW documents**
Links: **URL links**

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





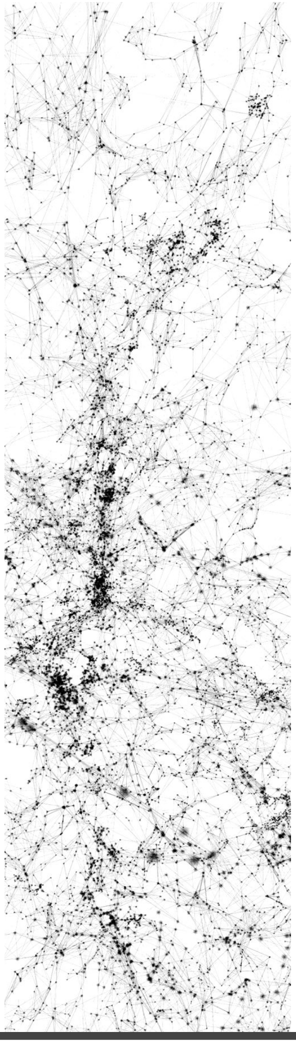
Expected



R. Albert, H. Jeong, A-L Barabasi, *Nature*, 401 130 (1999).

Outline

- Quick Notes
    - Assignment 1, slack
- Adjacency matrix and degree
- 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
- Connectivity & eigenvalues of Laplacian matrix
- How to pattern?



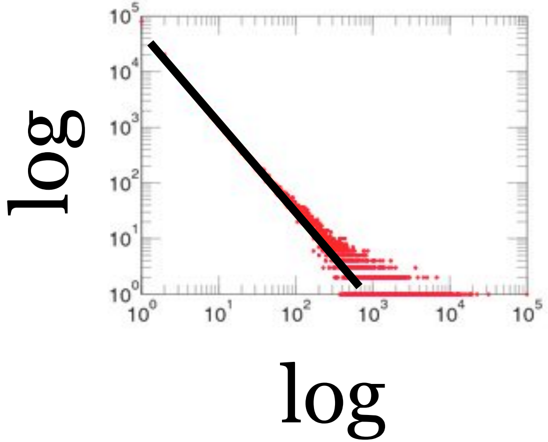
Power law distribution

$\ln p_d = -\boldsymbol{\alpha} \ln d + \boldsymbol{\beta}$

$p_d = C\, d^{-\alpha}$

What is C?

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



Power law distribution

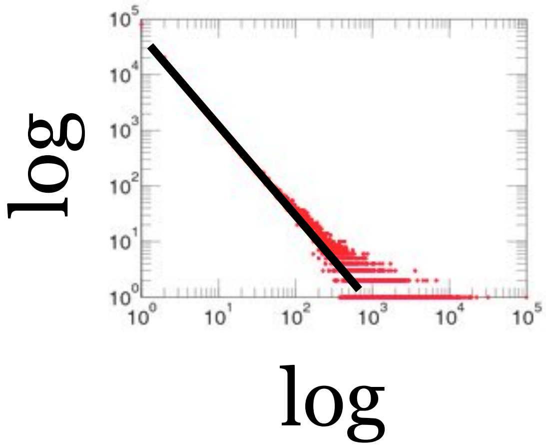
$\ln p_d = -\alpha \ln d + c$

$p_d = C\, d^{-\alpha}$

What is C?

$C = e^{\beta}$

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



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

Scale free networks

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

Since power-law is scale invariance:

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

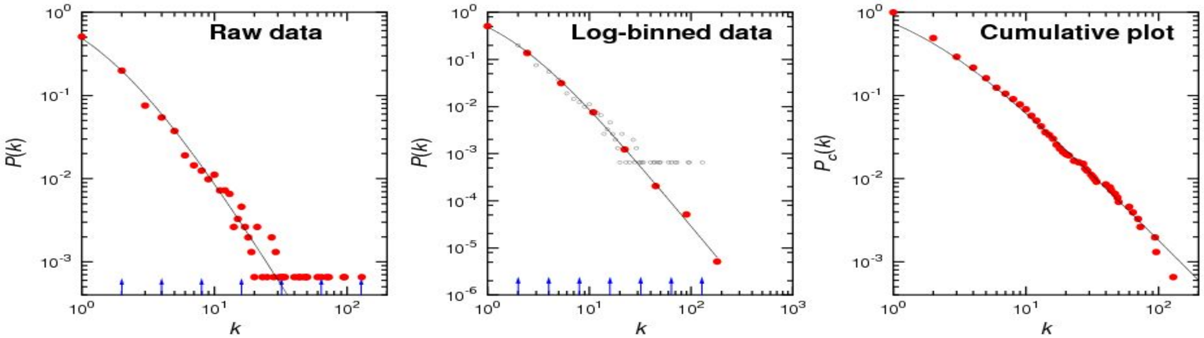
$$f(\lambda\, d) = C\, \lambda^{-\alpha} d^{-\alpha} = \lambda^{-\alpha} f(d)$$

function f is scale invariance iff

$$f(\lambda\, x) = \lambda^{a} f(x) \text{ for some a and all } \lambda$$

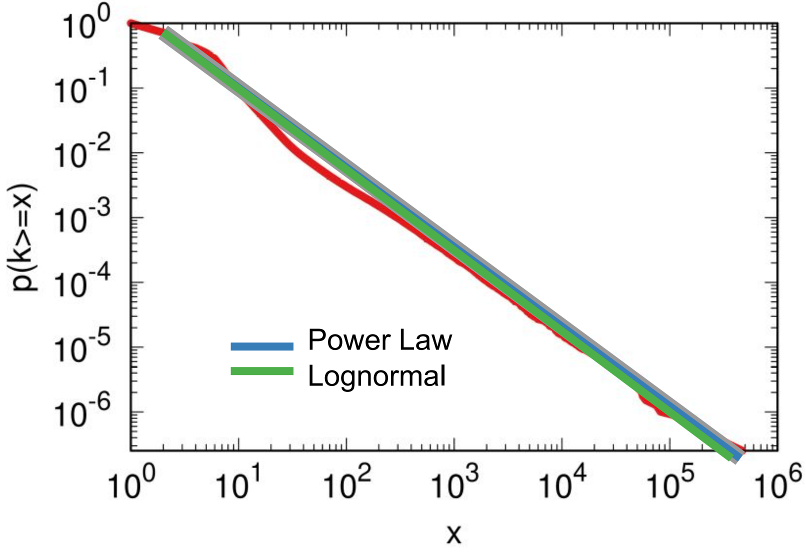
Fitting a power law

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



Fitting a power law

- Linear Fit in log-log space
    - Very good R2 and p-value because of log-log scale!
- Log-Likelihood
    - How likely is function f to fit the data? Allows p-value estimation between two alternatives, there is a tool for this: http://tuvalu.santafe.edu/~aaronc/powerlaws/
- Still an active research area



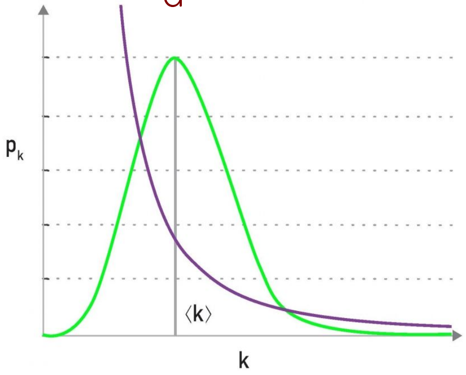
From Cosia's slides

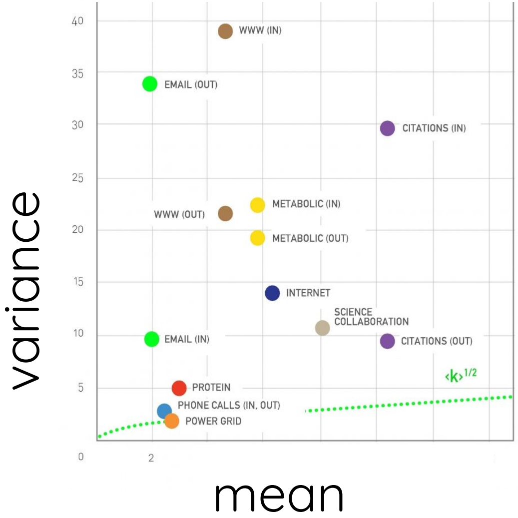
Mean & variance for a power-law

- 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 Barbasi book, we can see how variance deviates from expected variance of same mean random network with poisson distribution (dashed green line)

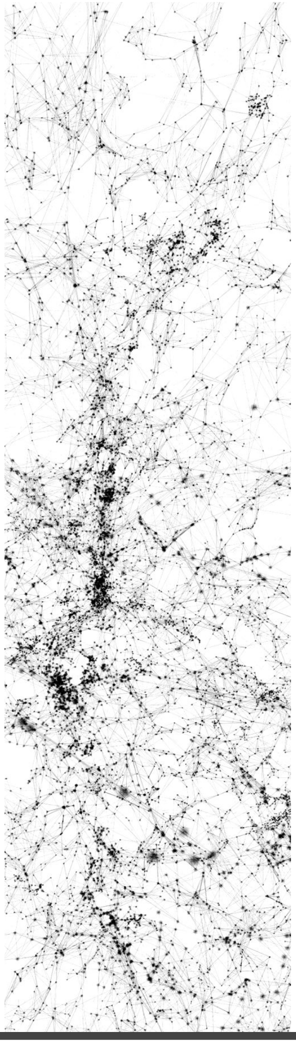
$p_d = C\,d^{-\alpha}$





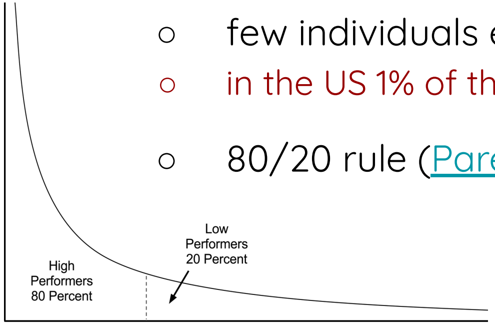
Outline

- Quick Notes
  - Assignment 1, slack
- Adjacency matrix and degree
- 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
- Connectivity & eigenvalues of Laplacian matrix
- How to pattern?

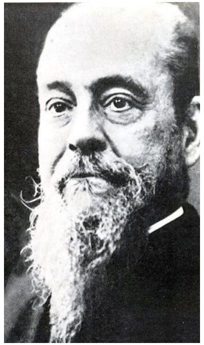


Powerlaws are normal?

- Income follow 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)

What creates a powerlaw?

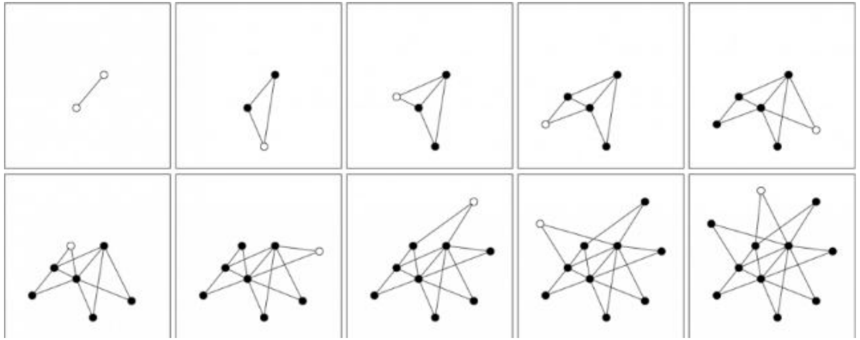
Preferential Attachment

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

Albert Barabasi Model (AB)

- Add one node at the time, add m connections per new node
- the probability of forming a connection to an existing node is **proportional to its degree**

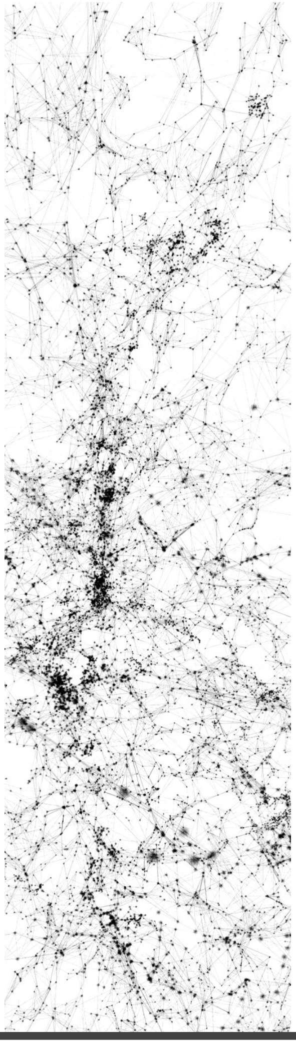
$p(i) \sim d_i$



m=?

Outline

- Quick Notes
  - Assignment 1, slack
- Adjacency matrix and degree
- 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
- Connectivity & eigenvalues of Laplacian matrix
- How to pattern?



Marginals of Adjacency Matrix

- Marginals of **A** ⇒ Degrees
  - $\mathbf{d_i = \Sigma_j A_{ij}}$

$$
\begin{array}{cc}
 & \begin{array}{cccccccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \end{array} \\
\begin{array}{c} 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \end{array} &
\begin{bmatrix}
0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0
\end{bmatrix}
\end{array}
$$

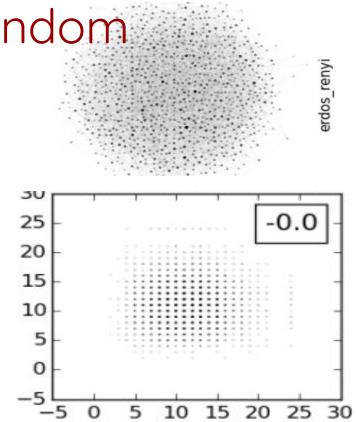
Degree Assortativity

Strong correlation between degree of connecting nodes

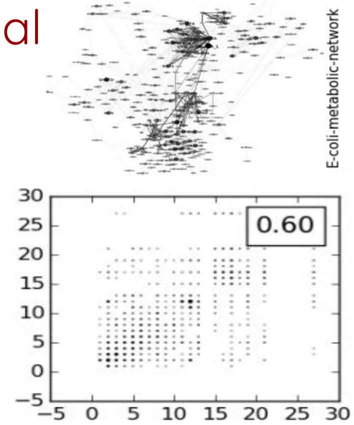
assortative mixing

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

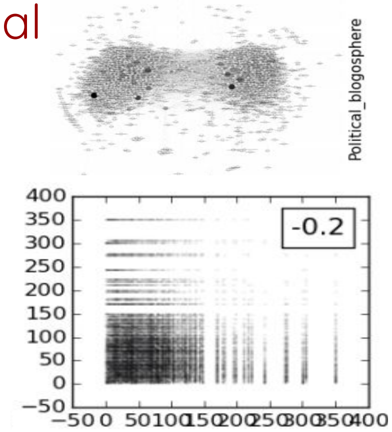
random



real

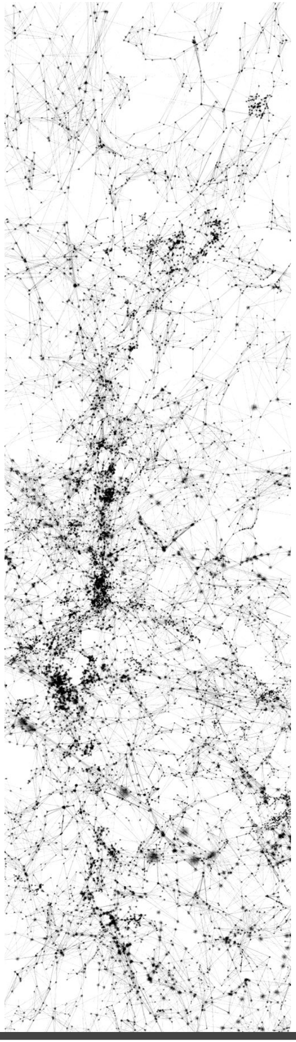


real



Outline

- Quick Notes
  - Assignment 1, slack
- Adjacency matrix and degree
- 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
- Connectivity & eigenvalues of Laplacian matrix
- How to pattern?



Marginals of Adjacency Matrix

- Marginals of **A** ⇒ Degrees
  - $\mathbf{d_i = \Sigma_j A_{ij}}$
- **Sum(A) = $\Sigma_i\Sigma_j A_{ij} = \Sigma_i d_i = 2E$**

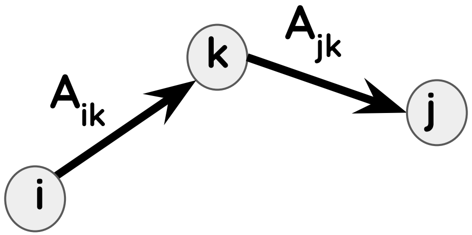
If undirected

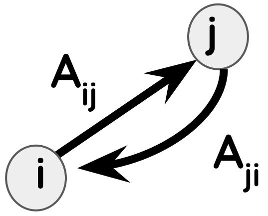
$$
\begin{array}{cc}
 & \begin{array}{cccccccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \end{array} \\
\begin{array}{c} 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \\ 11 \end{array} &
\begin{bmatrix}
0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0
\end{bmatrix}
\end{array}
$$

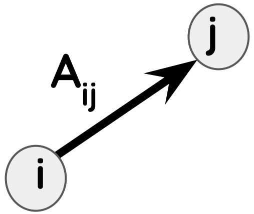
- If directed column-wise & row-wise marginals ⇒ indegree and outdegree
  - $\mathbf{d^{in}_i = \Sigma_j A_{ji}}$ and $\mathbf{d^{out}_i = \Sigma_j A_{ij}}$
  - **sum(A) = E**

Powers of A

- A^2 : # of walks with length two
  - $A^2_{ij} = \Sigma_k A_{ik} A_{kj}$
  - **If undirected:**
    - **A^2_{ij} : number of common neighbors**
    - **What is A^2_{ii} ?** number of neighbors = degree
  - **What is A^2_{ii} in directed graph?** number of reciprocal neighbors





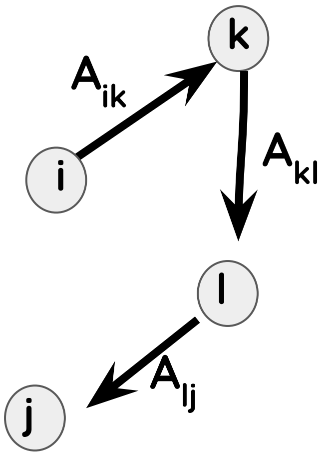


network's reciprocity

$$\Sigma_i \Sigma_j A_{ij} A_{ji} \,/\, \Sigma_i \Sigma_j A_{ij}$$

Powers of A

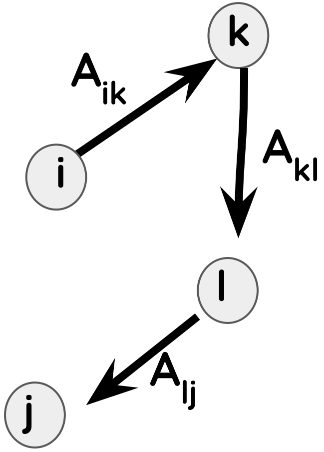
- A^2 : # of walks with length **two**
- A^3 : # of walks of length **three**
  - Is it same as number of paths?



Powers of A

- A^2 : # of walks with length **two**

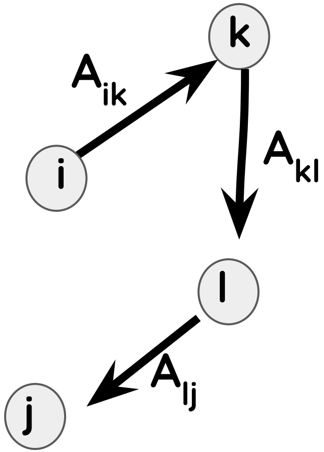
- A^3 : # of walks of length **three**
  - Is it same as number of paths?
    - A **walk** is a finite or infinite sequence of edges which joins a sequence of vertices
    - A **trail** is a walk in which all edges are distinct.
    - A **path** is a trail in which all vertices are distinct.

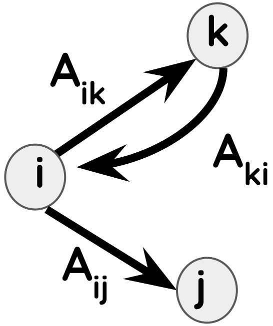


https://en.wikipedia.org/wiki/Path_(graph_theory)#Walk,_trail,_path

Powers of A

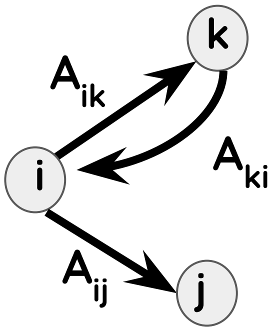
- A^2 : # of walks with length **two**
- A^3 : # of walks of length **three**
    - Is it same as number of paths? **No!**

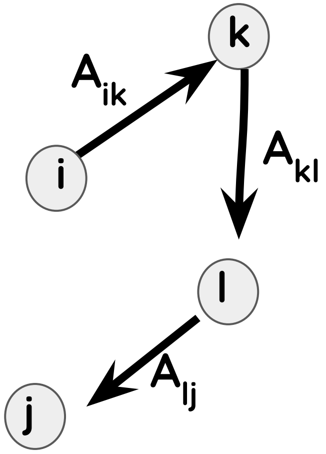




Powers of A

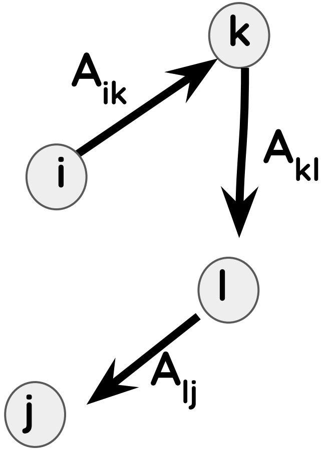
- A^2 : # of walks with length **two**
- A^3 : # of walks of length **three**
  - Is it same as number of paths? **No!**
  - **What is A^3_{ii} ?**



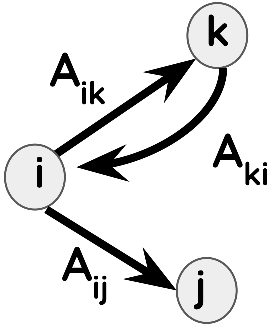


Powers of A

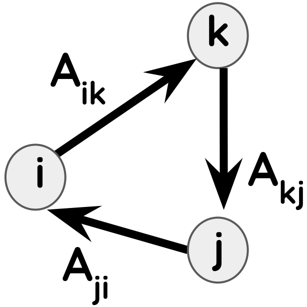
- A^2 : # of walks with length **two**



- A^3 : # of walks of length **three**
    - Is it same as number of paths? **No!**
    - **What is A^3_{ii} ?**



Number of Triangles?



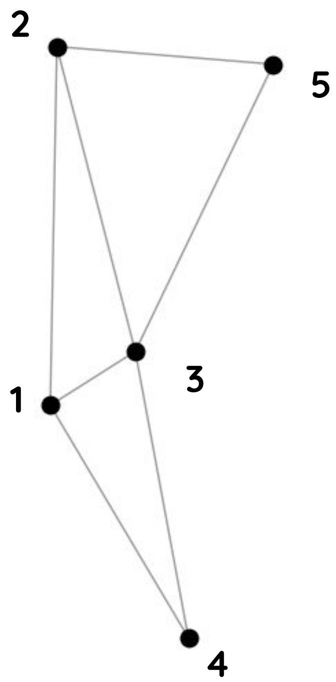
Example

```
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
```

[[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]]

[[**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**]]

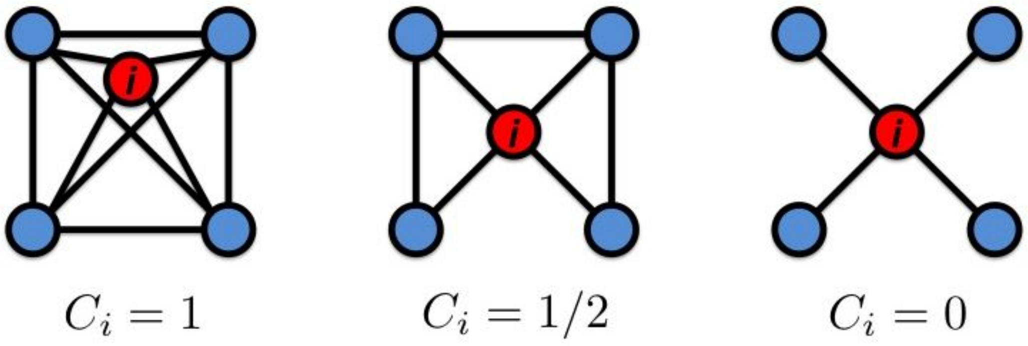
[[**4** 7 7 5 3]
 [7 **4** 7 3 5]
 [7 7 **6** 6 6]
 [5 3 6 **2** 3]
 [3 5 6 3 **2**]]

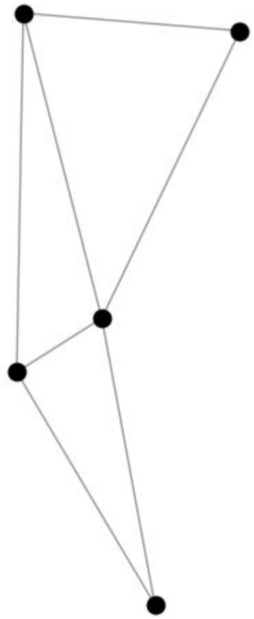


Clustering Coefficient

Local: $\mathbf{c_i = A^3_{ii} / d_i(d_i-1)}$

Total number of triangles?



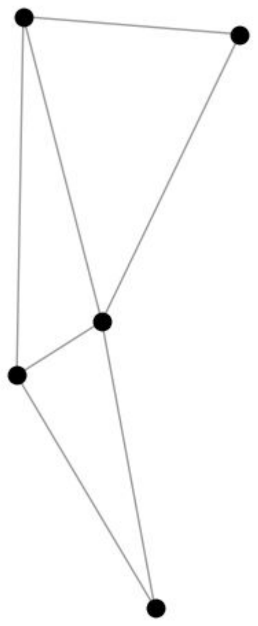


Clustering Coefficient

Local: $\mathbf{c_i = A^3_{ii} / d_i (d_i-1)}$

Total number of triangles?

$$\begin{bmatrix} \mathbf{4} & 7 & 7 & 5 & 3 \\ 7 & \mathbf{4} & 7 & 3 & 5 \\ 7 & 7 & \mathbf{6} & 6 & 6 \\ 5 & 3 & 6 & \mathbf{2} & 3 \\ 3 & 5 & 6 & 3 & \mathbf{2} \end{bmatrix}$$

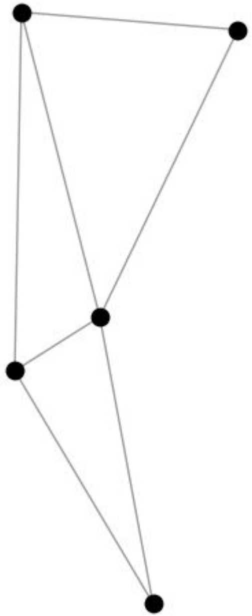


Clustering Coefficient

Local: $\mathbf{c_i = A^3_{ii} / d_i (d_i\text{-}1)}$

Total number of triangles? $\mathbf{Tr(A^3)/6}$

[[**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**]]



Real networks have a lot of triangles

Friends of friends are friends

Can we compute number of triangles more effectively?

Clustering Coefficient

Local: $\mathbf{c_i = A^3_{ii} / d_i (d_i\text{-}1)}$

Total number of triangles? $\mathbf{Tr(A^3)/6}$

Real networks have a lot of triangles

Friends of friends are friends

Can we compute number of triangles more effectively?

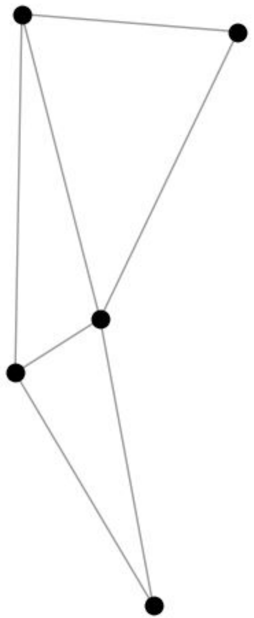
from eigenvalues of **A** as $\frac{1}{6} \Sigma_i \lambda_i^3$

Since if **λ** is eigenvalue of **A** then $\boldsymbol{\lambda^p}$ is an eigenvalue of $\mathbf{A^p}$

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

Many works on approximating number of triangles in large graphs

$$\begin{bmatrix} \mathbf{4} & 7 & 7 & 5 & 3 \\ 7 & \mathbf{4} & 7 & 3 & 5 \\ 7 & 7 & \mathbf{6} & 6 & 6 \\ 5 & 3 & 6 & \mathbf{2} & 3 \\ 3 & 5 & 6 & 3 & \mathbf{2} \end{bmatrix}$$



Clustering Coefficient

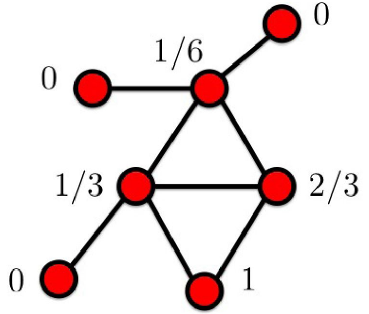
Local: $\mathbf{c_i = A^3_{ii} / d_i (d_i - 1)}$

Total number of triangles? $\mathbf{Tr(A^3)/6}$

Global: $\mathbf{Tr(A^3)/(Sum(A^2)-Tr(A^2))}$

measures the density of triangles

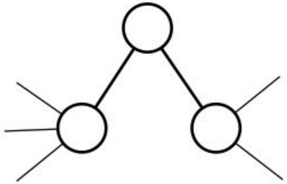
$$: \frac{\text{(number of closed paths of length 2)}}{\text{(number of paths of length 2)}}$$



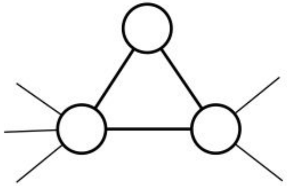
note the difference:

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

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



path of length 2

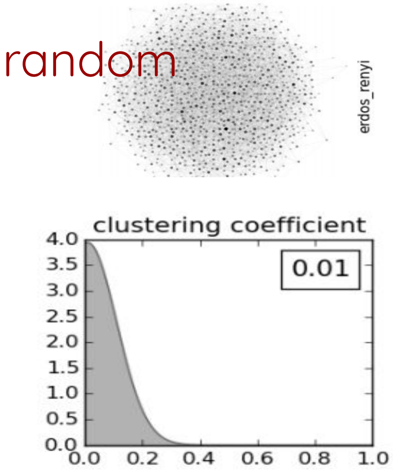


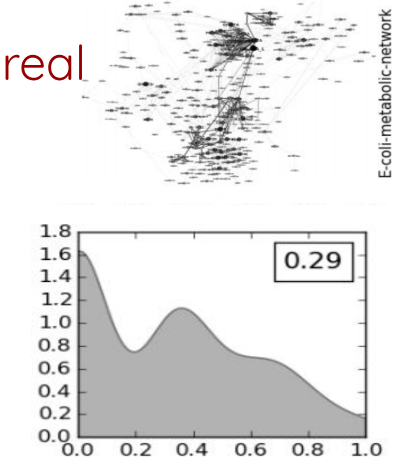
closed path of length 2

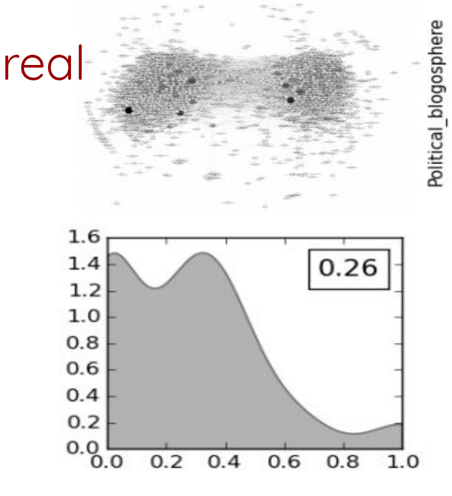
Transitivity & Assortativity

- High global clustering coefficient or high average local clustering coefficient

- Distribution of local clustering coefficient







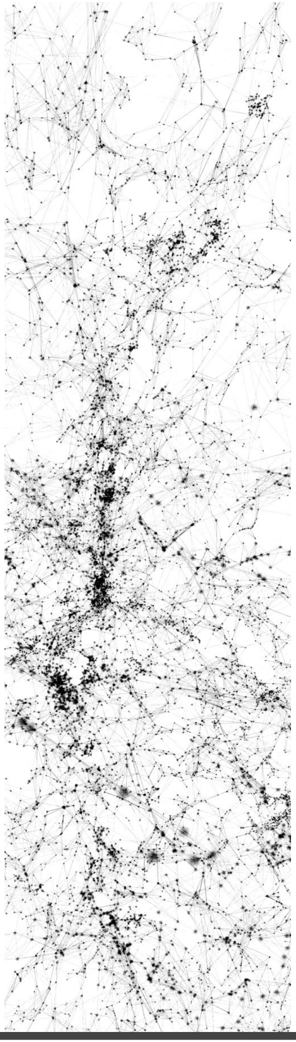
A

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

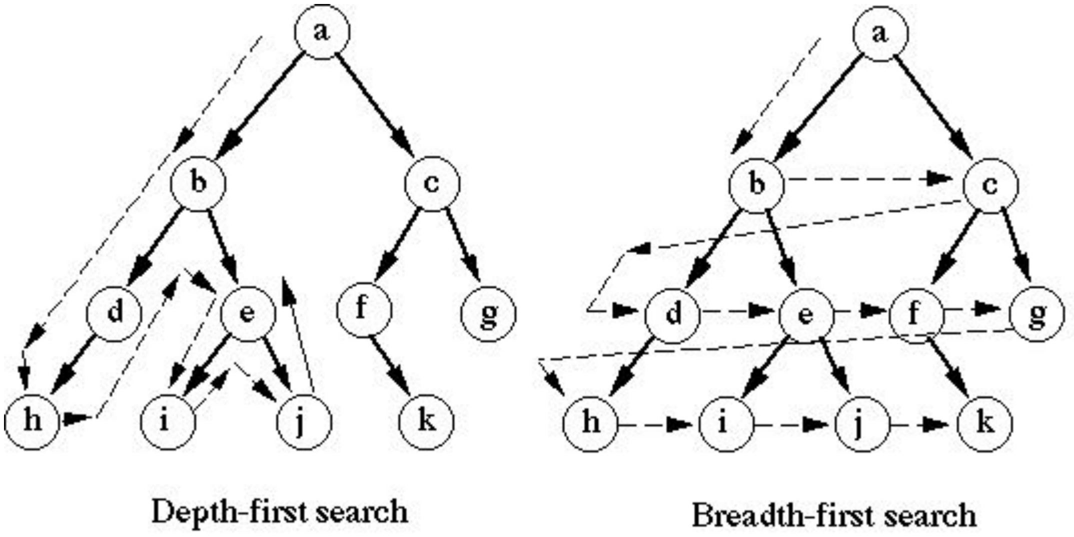
Outline

- Quick Notes
    - Assignment 1, slack
- Adjacency matrix and degree
- 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
- Connectivity & eigenvalues of Laplacian matrix
- How to pattern?



Shortest Path

https://en.wikipedia.org/wiki/Breadth-first_search



Longest & average shortest path

Small average shortest path

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

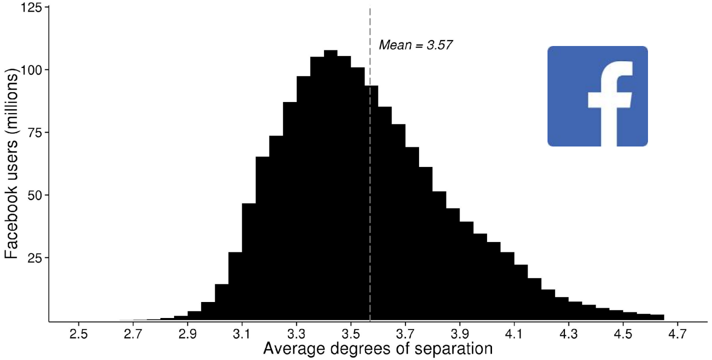
You can reach any node in a graph passing through few hubs

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

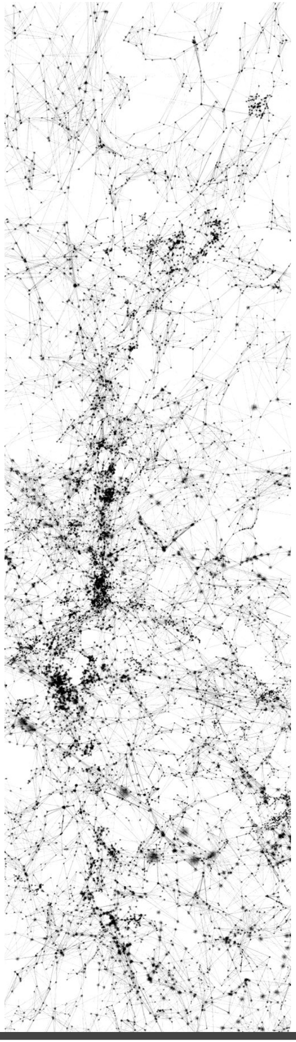


Four Degrees of Separation

You are 4 hops away from anyone in the planet

Outline

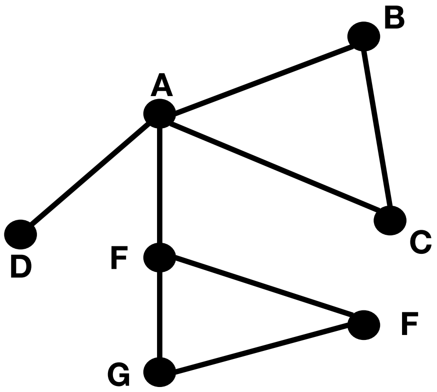
- Quick Notes
  - Assignment 1, slack
- Adjacency matrix and degree
- Sparsity Pattern
- Scale Free Pattern
  - Power-law degree distribution
  - Fitting a power-law
  - Preferential attachment and AB model
- Assortativity Pattern
- Transitivity Pattern
  - powers of A & counting triangles
- Small world Pattern
  - Shortest path
- **Connectivity & eigenvalues of Laplacian matrix**
- How to pattern?



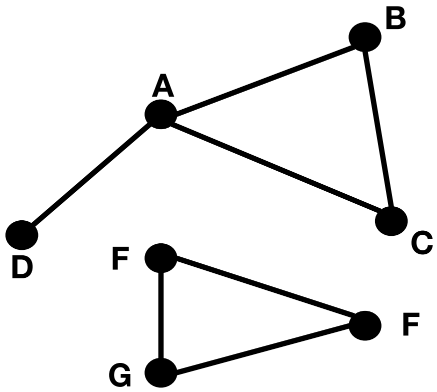
Connectivity

Connected (undirected) graph: any two vertices can be joined by a path

A **disconnected** graph is made up by two or more connected components



Connected



Not Connected

[From Barbasi's slides]

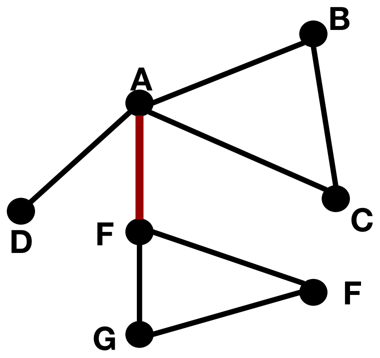
Connectivity: GCC & bridges

Connected (undirected) graph: any two vertices can be joined by a path

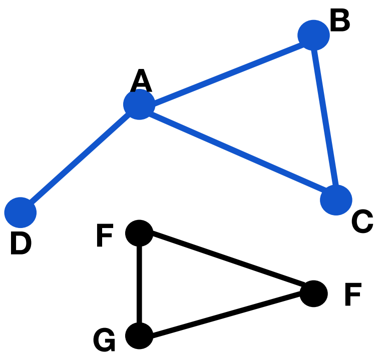
A disconnected graph is made up by two or more connected components

Largest Component is referred to as the **giant connected component (GCC)**

Bridge edges are those that if erased, the graph becomes disconnected



Connected



Not Connected

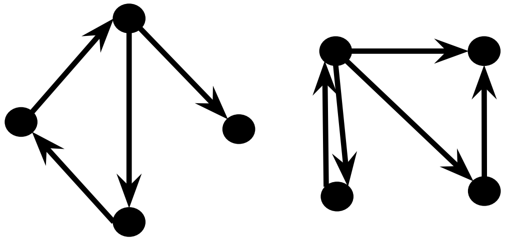
From Barbasi's slides

Connectivity in directed graphs

- **Strongly** connected component
  - has a path from each node to every other node and vice versa
    - e.g. A to B path and B to A path
- **Weakly** connected component
  - it is connected if we disregard the edge directions

How many scc do we have in this example graph?

How many wcc do we have in this example graph?



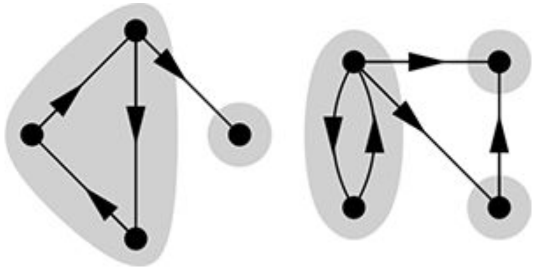
From Barbasi's slides &
From newman's book

Connectivity in directed graphs

- **Strongly** connected component
  - has a path from each node to every other node and vice versa
    - e.g. A to B path and B to A path
- **Weakly** connected component
  - it is connected if we disregard the edge directions

How many scc do we have in this example graph? 5

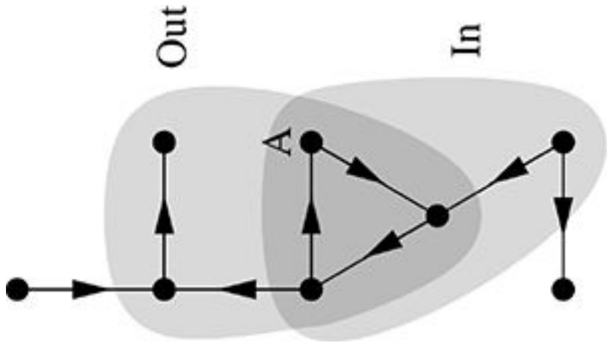
How many wcc do we have in this example graph? 2



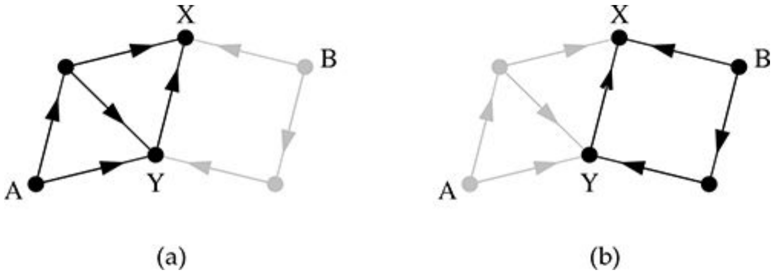
From Barbasi's slides &
From newman's book

In/Out components

In-component: nodes that can reach the scc
Out-component: nodes that can be reached from the scc



in/out-component of a specific node: set of nodes reachable by directed paths to/from that node

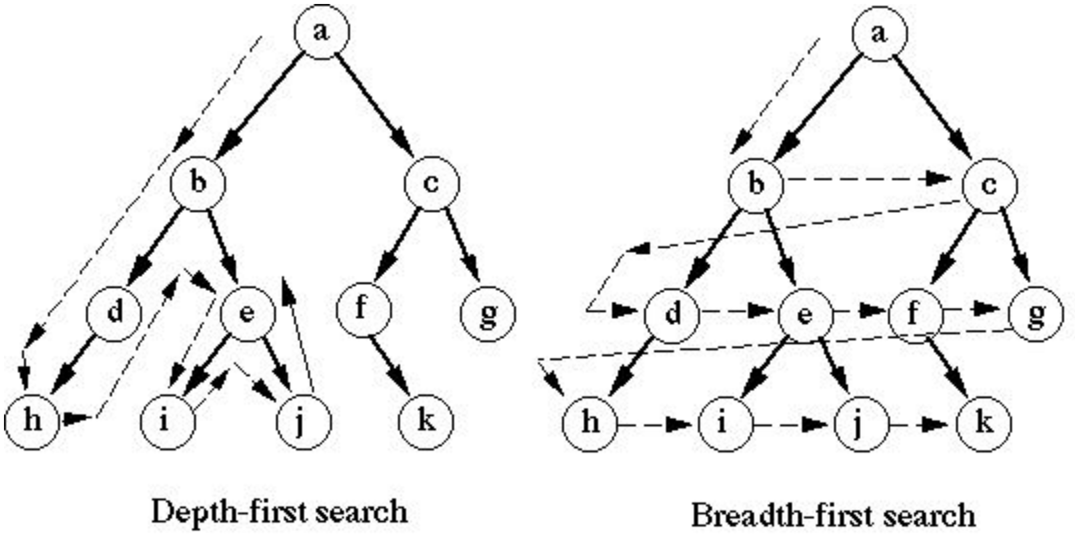


out-component of node A out-component of node B

From newman's book

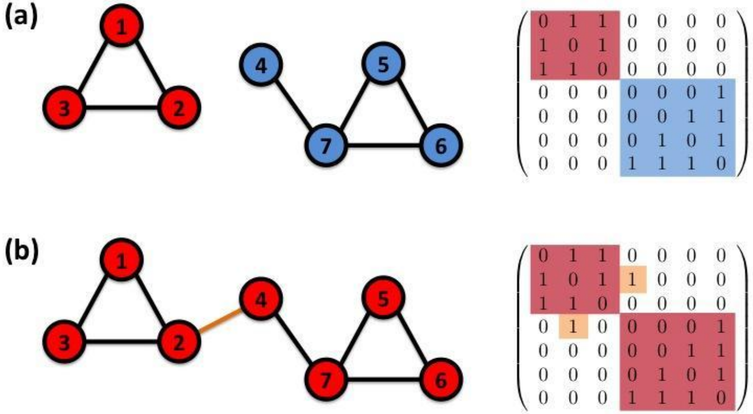
How to check connectivity?

Start from one node, traverse the graph and record the nodes you reach. If the size of this reached set of nodes is equal to all the nodes in the graph, then the graph is connected. If not, this is one component and continue until all nodes have been reached to get all the components.



Connectivity & Adjacency Matrix

The adjacency matrix of a network with several components can be written in a **block**-diagonal form, so that **nonzero elements are confined to squares**, with all other elements being zero:



From Barbasi's slides

How can we use this to see if the graph is connected based on A?

Connectivity & Laplacian Matrix

we need to consider a super useful matrix

comes into play in many many different contexts

D: diagonal matrix of degrees

Laplacian Matrix: **L = D - A**

```
[[ 3 -1 -1 -1  0]        [[3 0 0 0 0]        [[0 1 1 1 0]
 [-1  3 -1  0 -1]         [0 3 0 0 0]         [1 0 1 0 1]
 [-1 -1  4 -1 -1]         [0 0 4 0 0]         [1 1 0 1 1]
 [-1  0 -1  2  0]         [0 0 0 2 0]         [1 0 1 0 0]
 [0  -1 -1  0  2]]        [0 0 0 0 2]]        [0 1 1 0 0]]
```

L **D** **A**

Eigenvalues of Laplacian Matrix

- $\mathbf{Lu = \lambda u}$

wiki: Eigenvalues

- We have n eigenvalues which we call **Laplacian Spectrum**:

$$0 = \lambda_0 \leq \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$$

- λ_0 is always zero since we have **L(1,1…1) = 0**

why this holds?

- $E = \frac{1}{2} \Sigma d_i = \frac{1}{2} Tr(L) = \frac{1}{2} \Sigma \lambda_i$

- Laplacian Spectrum relates to graph connectivity & clustering

L= D - A

```
[[ 3 -1 -1 -1  0]
 [-1  3 -1  0 -1]
 [-1 -1  4 -1 -1]
 [-1  0 -1  2  0]
 [0  -1 -1  0  2]]
```

for undirected graphs, we can always do eigenvalue decomposition since L is symmetric

Connectivity & Laplacian Matrix

- smallest eigenvalue of **L** is always zero
- **second-smallest eigenvalue** of **L** is called Algebraic connectivity or Fiedler value and is nonzero only if graph is connected
- **number of zero eigenvalues** of **L** gives the **number of connected components**

Clustering & Laplacian Matrix

- **Signs of values in Fiedler eigenvector** (associated to Fiedler eigenvalue) tell us how to partition the graph into two components by breaking least edges, i.e. **minimum cut** solution

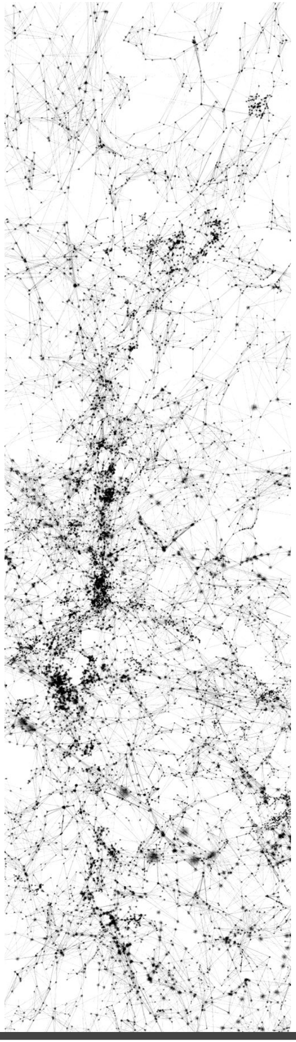
more on this and spectral clustering later

- eigengap is the difference between subsequent eigenvalues
  - first large eigengap is related to the number of clusters in data
  - first eigengap (=smallest nonzero eigenvalue) is called spectral gap which relates to how quickly the diffusion takes place on the network and density of the graph

See this: https://towardsdatascience.com/spectral-clustering-aba2640c0d5b

Outline

- Quick Notes
  - Assignment 1, slack
- Adjacency matrix and degree
- 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
- Connectivity & eigenvalues of Laplacian matrix
- **How to pattern?**



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?

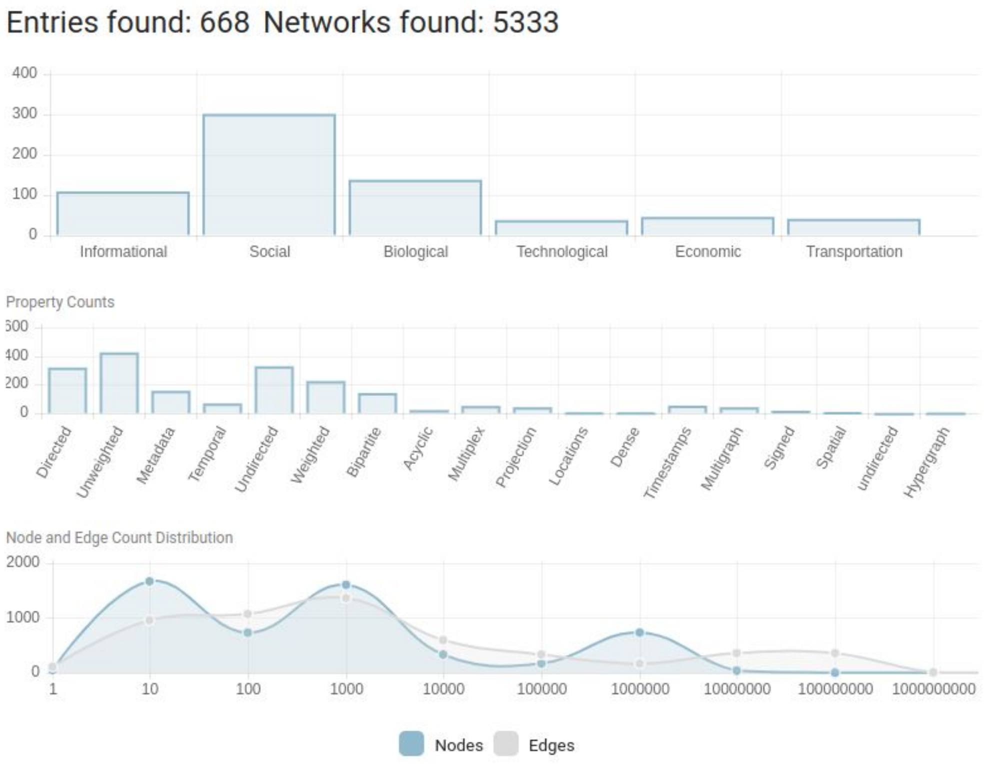
{common} Network Repositories

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**

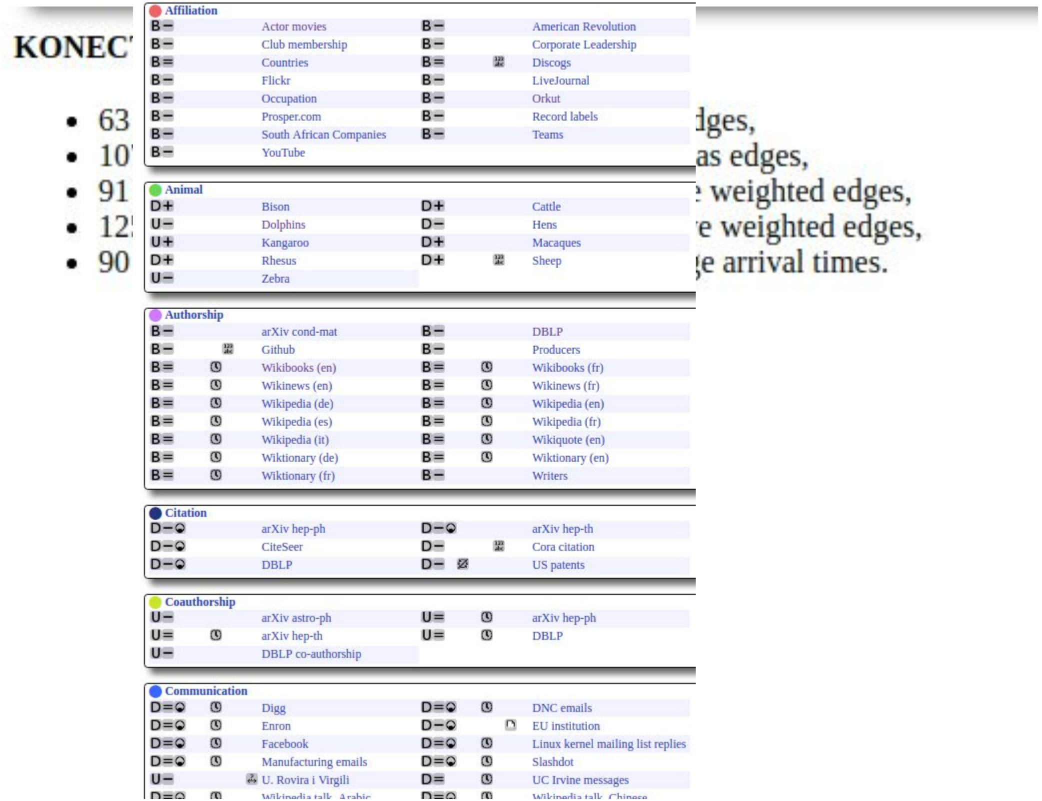
Let slack know if you come across other large repos

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

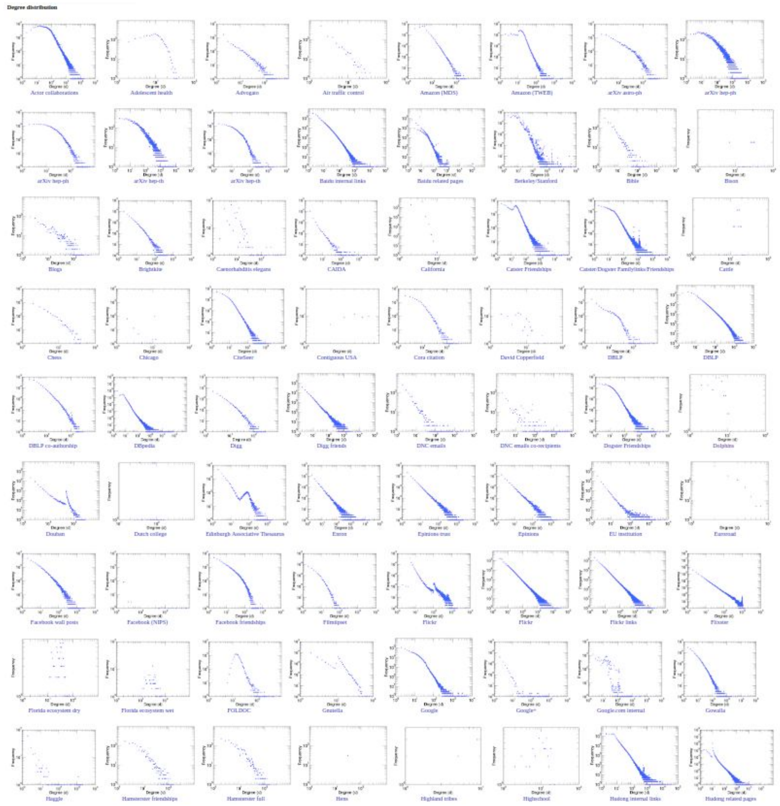
Coauthorship

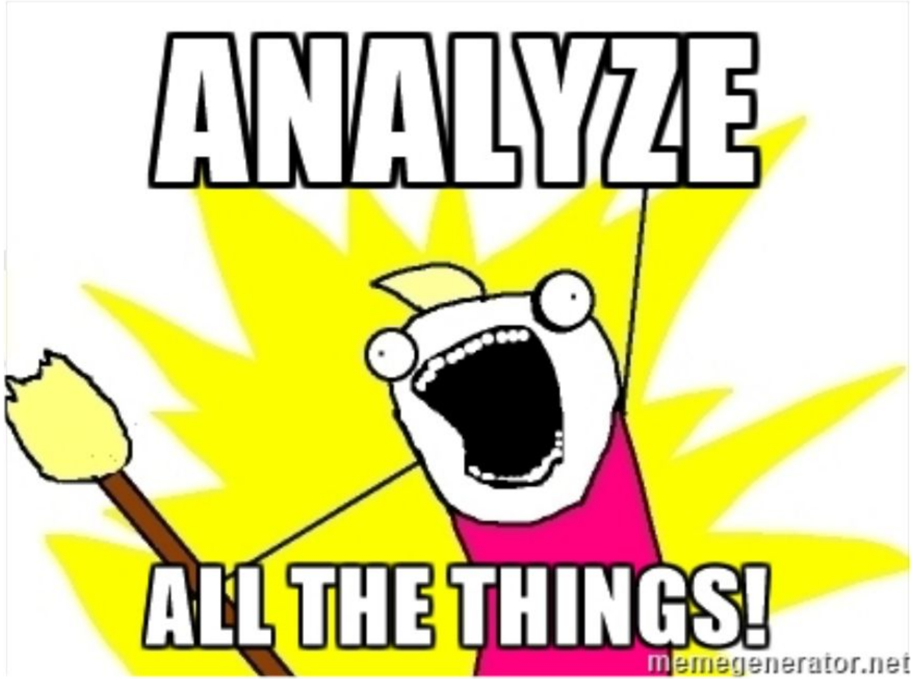
U−	arXiv astro-ph	U=	arXiv hep-ph
U=	arXiv hep-th	U=	DBLP
U−	DBLP co-authorship		

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

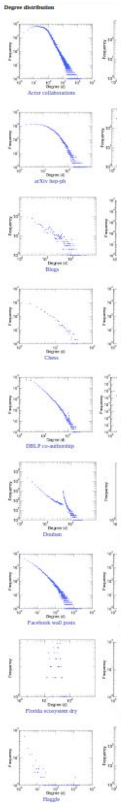


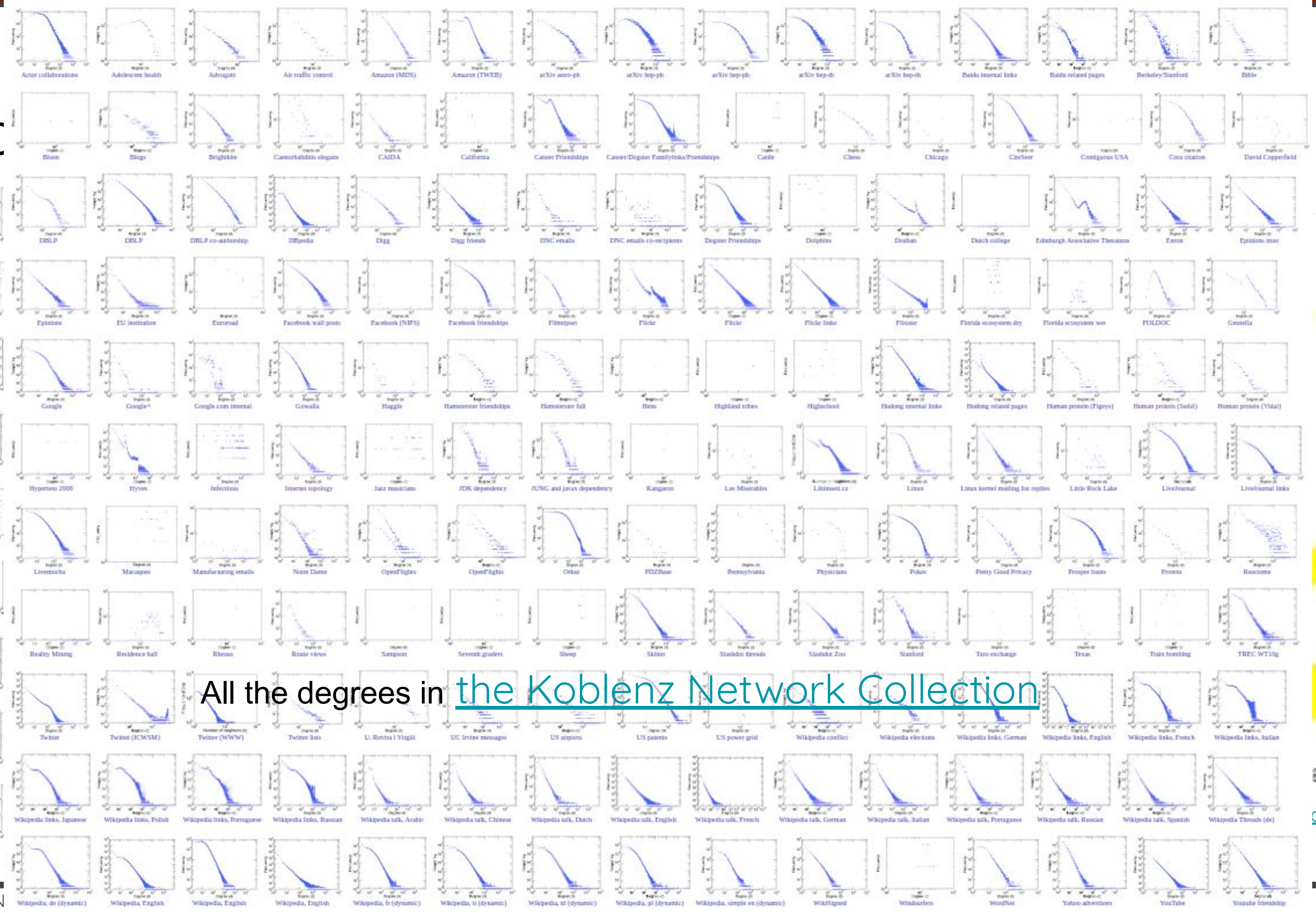


From Clauset's slides

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

Hyp





All the degrees in [the Koblenz Network Collection]

ator.net

des