# Анализ сетевых данных

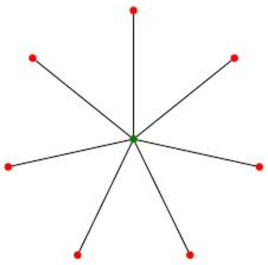
Лекция 2. Random Graph models.

12 Сентября, 2018

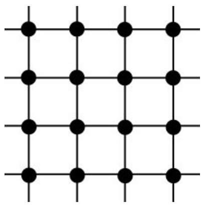
# MOOCs

- http://tuvalu.santafe.edu/~aaronc/courses/5352/ **Aaron Clauset**, lecture notes
- https://goo.gl/8CghUx **Leonid Zhukov**, http://www.leonidzhukov.net/, Full course videos + lecture notes.
- http://web.stanford.edu/class/cs224w/ **Jure Leskovec**, lecture notes, bunch of useful materials, *videos are unavailable.*

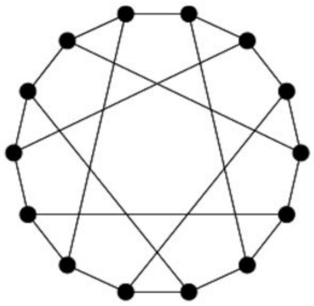
# Compare different networks



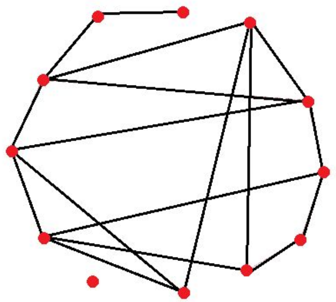
Star graph



Grid graph



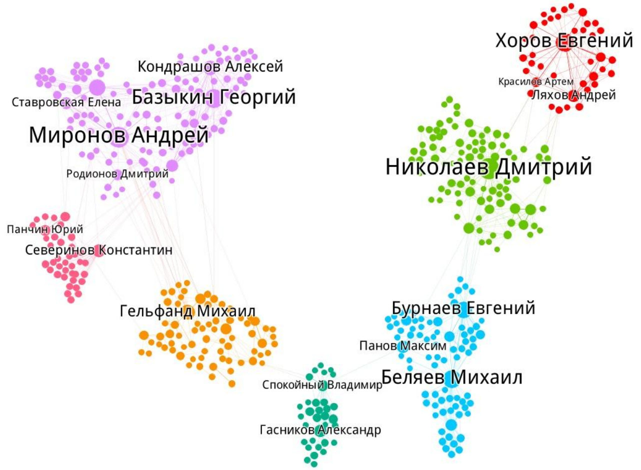
K-regular graph



Random graph

...?

VS



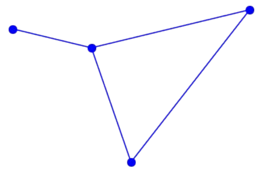

Real network

# How to compare different networks?

- # nodes, #edges, density
- (Average) node degree
- (Average) clustering coefficient
- (Average) path length
- Node degree distribution
- Centrality measures (next lecture)

# How to compare different networks?

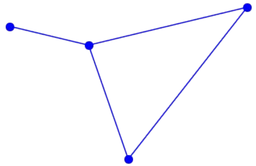
- # nodes, #edges, density
- (Average) node degree
- (Average) clustering coefficient
- (Average) path length
- Node degree distribution
- Centrality measures (next lecture)



# nodes = 4 (n)
# edges = 4 (m)
density = m / [n(n-1)/2] = 2/3

# How to compare different networks?

- # nodes, #edges, density
- (Average) node degree
- (Average) clustering coefficient
- (Average) path length
- Node degree distribution
- Centrality measures (next lecture)



$$k_i = \sum_{j=1}^{n} A_{ij}$$

k = (1 + 3 + 2 + 2) / 4 =

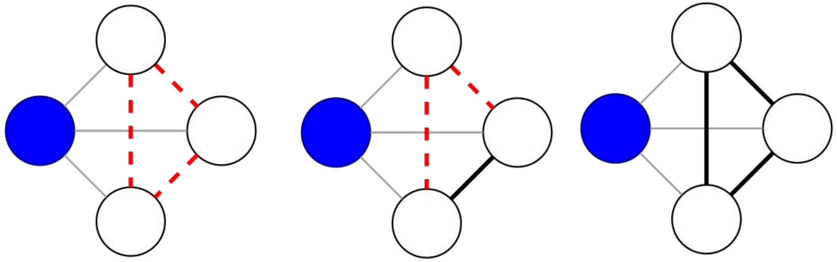
$$= \langle k \rangle = \frac{1}{n} \sum_{i=1}^{n} k_i = \frac{2m}{n}$$

# How to compare different networks?

- # nodes, #edges, density
- (Average) node degree
- (Average) clustering coefficient
- (Average) path length
- Node degree distribution
- Centrality measures (next lecture)

$$C_i = \frac{\text{(number of pairs of neighbors of } i \text{ that are connected)}}{\text{(number of pairs of neighbors of } i\text{)}} = \sum_{jk} A_{ij}A_{jk}A_{ki} \Big/ \binom{k_i}{2} = \frac{2\,n_i}{k_i\,(k_i-1)},$$

$n_i$ = # neighbours of node i

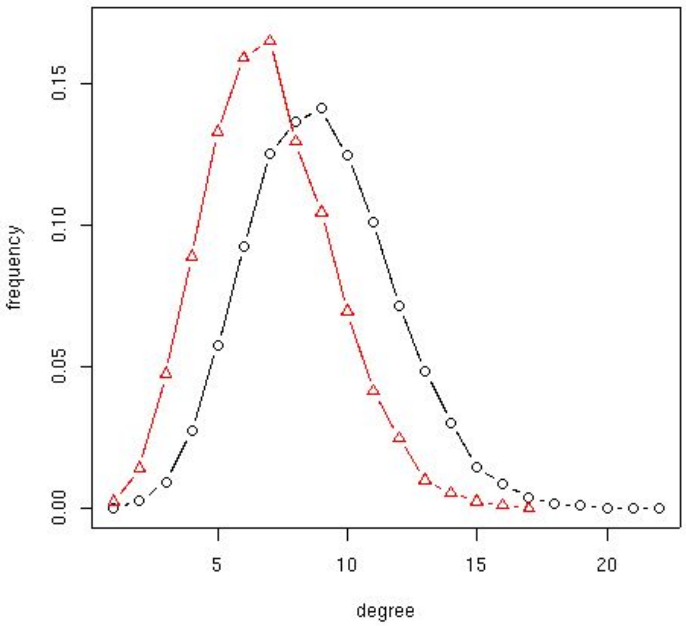


c = 0 c = 1/3 c = 1

$$C = \frac{\text{(number of triangles)} \times 3}{\text{(number of connected triples)}} = \sum_{ijk} A_{ij}A_{jk}A_{ki} \Big/ \sum_{ijk} A_{ij}A_{jk} \;,$$

# How to compare different networks?

- # nodes, #edges, density
- (Average) node degree
- (Average) clustering coefficient
- (Average) path length
- Node degree distribution
- Centrality measures (next lecture)



# Erdos-Renyi random graph

- $G_{n,p}$ model:

  Graph with $n$ nodes and for each pair of nodes the probability of an edge between them is equal to $p$.

- $G_{n,\ m}$ model:

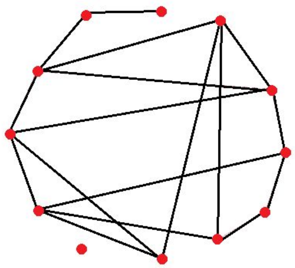
  A randomly selected graph from the set of $C_N^m$ graphs, with $N = n(n-1)/2$ , where $n$ = #nodes and $m$ = #edges

Random graph model (Erdos & Renyi, 1959)

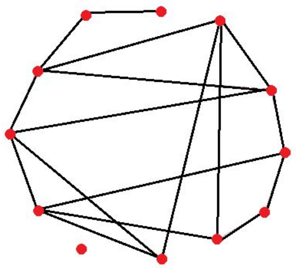
# Erdos-Renyi properties

$G_{n,p}$ model:

- $<m> = p*n*(n\text{-}1)/2$
- $<k> = (n\text{-}1)*p \approx n*p$



# Erdos-Renyi properties

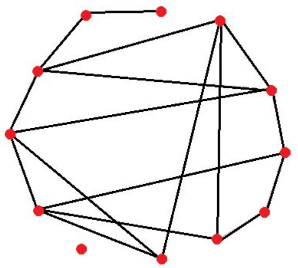


$G_{n,p}$ model:

- $<m> = p*n*(n\text{-}1)/2$
- $<k> = (n\text{-}1)*p \approx n*p$

What is node degree distribution?

# Erdos-Renyi properties



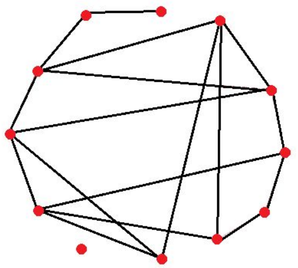
$G_{n,p}$ model:

- $\langle m \rangle = p * n * (n\text{-}1)/2$
- $\langle k \rangle = (n\text{-}1) * p \approx n * p$

What is node degree distribution?

**Probability that given node *i* has degree $k_i = k$**

# Erdos-Renyi properties



$G_{n,p}$ model:

- $\langle m \rangle = p*n*(n-1)/2$
- $\langle k \rangle = (n-1)*p \approx n*p$

What is node degree distribution?

**Probability that given node $i$ has degree $k_i = k$**

$$P(k_i = k) = P(k) = C_{n-1}^{k} p^k (1-p)^{n-1-k}$$

(Bernoulli distribution)

# Erdos-Renyi degree distribution

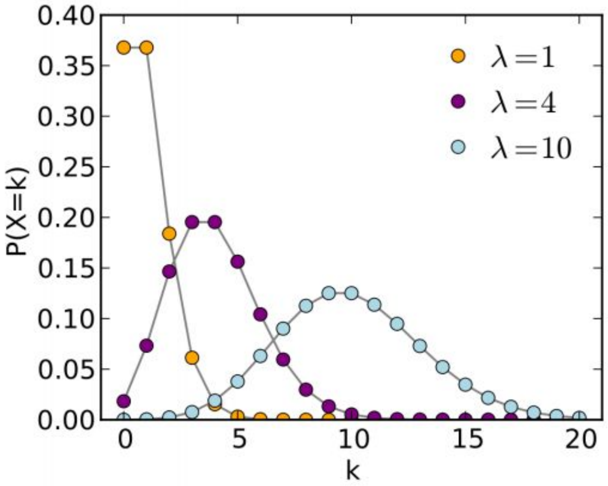
Limiting case of Bernoulli distribution (when *n* goes to infinity) - Poisson distribution (with parameter λ = <*k*> = *np*)

$$P(k) = \frac{\langle k \rangle^k e^{-\langle k \rangle}}{k!} = \frac{\lambda^k e^{-\lambda}}{k!}$$

# Erdos-Renyi degree distribution

Limiting case of Bernoulli distribution (when *n* goes to infinity) - Poisson distribution (with parameter λ = <*k*> = *np*)

$$P(k) = \frac{\langle k \rangle^k e^{-\langle k \rangle}}{k!} = \frac{\lambda^k e^{-\lambda}}{k!}$$

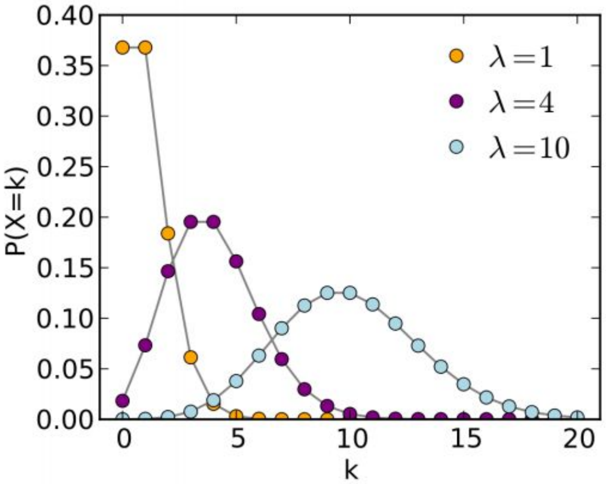



$$P(k_i = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad \lambda = pn$$

# Erdos-Renyi degree distribution

Limiting case of Bernoulli distribution (when *n* goes to infinity) - Poisson distribution (with parameter $\lambda = \langle k \rangle = (n-1)p$)

$$P(k) = \frac{\langle k \rangle^k e^{-\langle k \rangle}}{k!} = \frac{\lambda^k e^{-\lambda}}{k!}$$




Which in case of np → ∞, goes to Gaussian

$$P(k_i = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad \lambda = pn$$

# $G_{n,p}$ vs $G_{n,m}$

Bernoulli distribution

- Mean = $\langle k \rangle = (n-1)p$
- Variance = $\sigma^2 = p(1-p)(n-1)$

With fixed p and n $\rightarrow \infty$, distribution becomes *narrow*:

# $G_{n,p}$ vs $G_{n, m}$

Bernoulli distribution

- Mean = $\langle k \rangle = (n-1)p$
- Variance = $\sigma^2 = p(1-p)(n-1)$

With fixed p and $n \to \infty$, distribution becomes *narrow*:

$$\sigma / \langle k \rangle = [(1-p) / p(n-1)]^{1/2} \approx 1 / (n-1)^{1/2}$$

thus we are increasingly confident that the degree of a node is equal to $\langle k \rangle$

# $G_{n,p}$ vs $G_{n,m}$

Bernoulli distribution

- Mean = $\langle k \rangle = (n-1)p$
- Variance = $\sigma^2 = p(1-p)(n-1)$

With fixed p and $n \to \infty$, distribution becomes *narrow*:

$$\sigma / \langle k \rangle = [(1-p) / p(n-1)]^{1/2} \approx 1 / (n-1)^{1/2}$$

thus we are increasingly confident that the degree of a node is equal to $\langle k \rangle$

$G_{n,p}$ and $G_{n,m}$ are the same

# Erdos-Renyi clustering coefficient

$$C_i = \frac{2\, n_i}{k_i\,(k_i - 1)}$$

since edges appear i.i.d. with probability p:

$n_i = p * k_i(k_i - 1) / 2$

then $C_i \approx \langle k \rangle / n$

# Erdos-Renyi clustering coefficient

$$C_i = \frac{2\, n_i}{k_i\,(k_i - 1)}$$

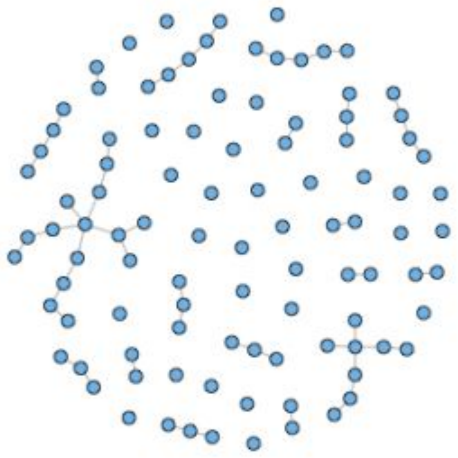
since edges appear i.i.d. with probability p:

$n_i = p * k_i(k_i - 1) / 2$

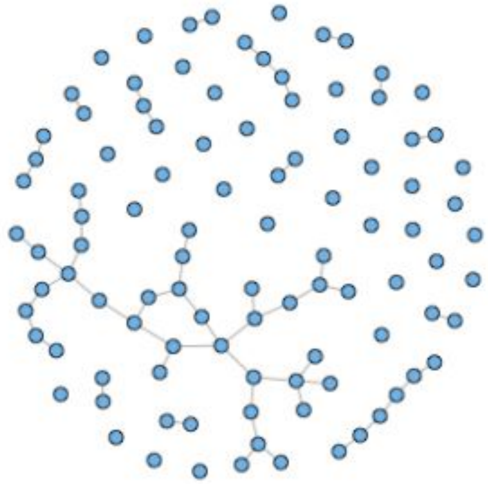
then $C_i \approx <k> / n$

This means that with n goes to infinity clustering coefficient of a random graph goes to 0

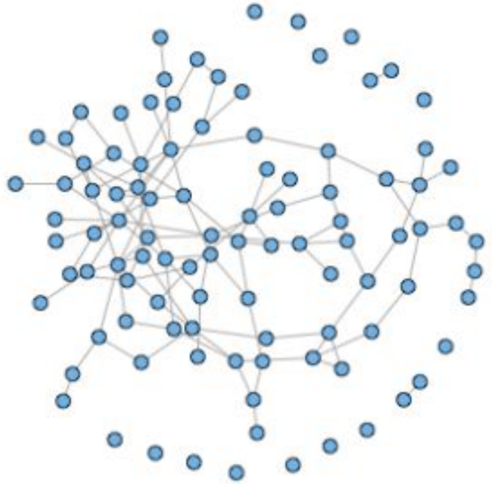
# What about connectivity?



$p < p_c$



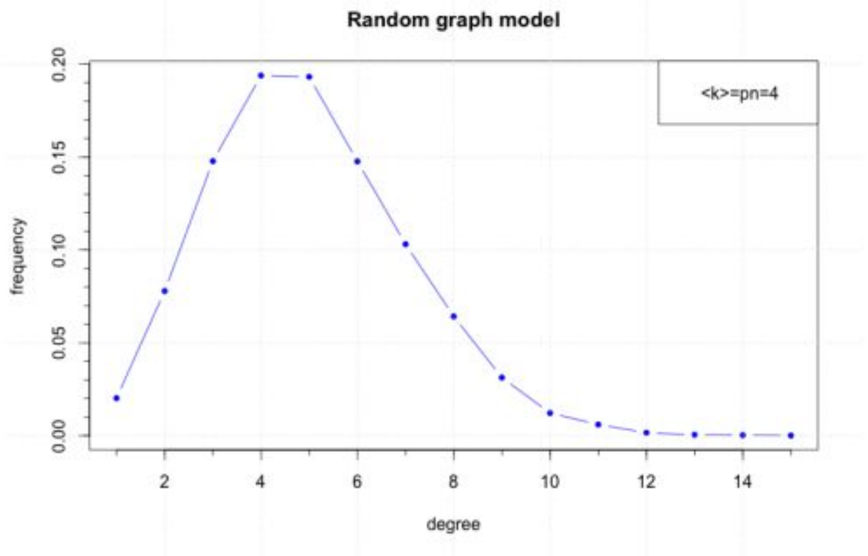
$p = p_c$

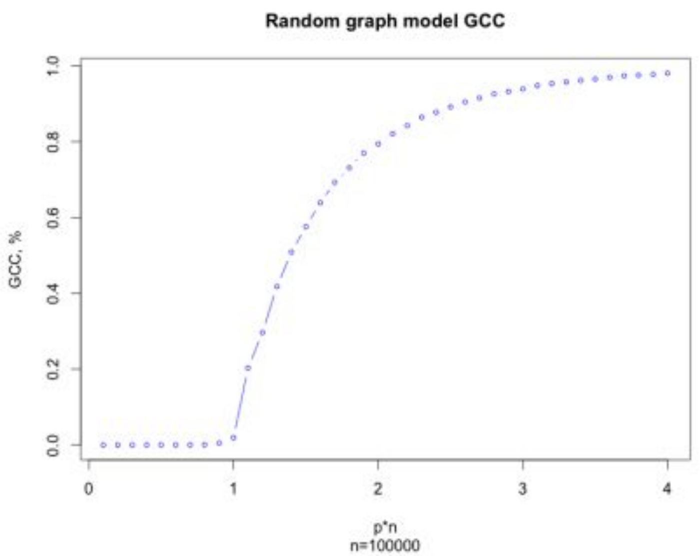


$p > p_c$

# What about connectivity?

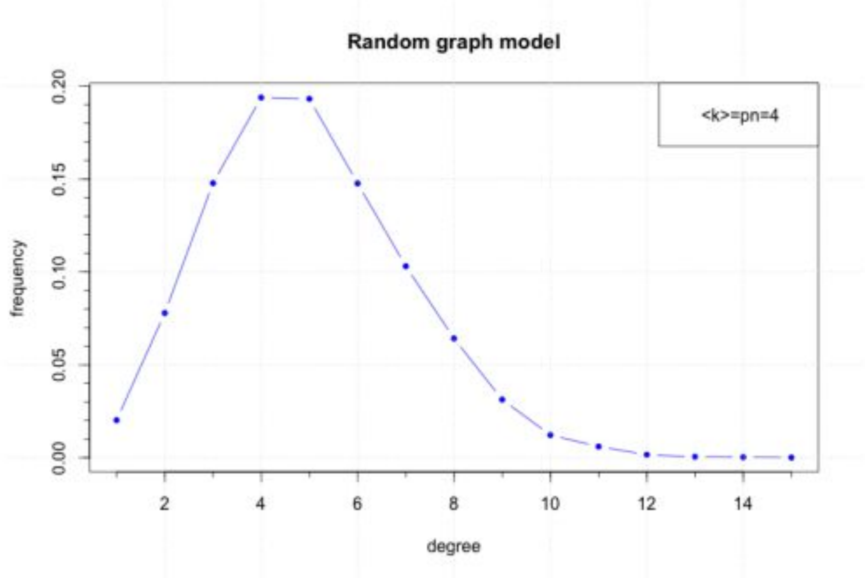
$$\langle k \rangle = pn$$

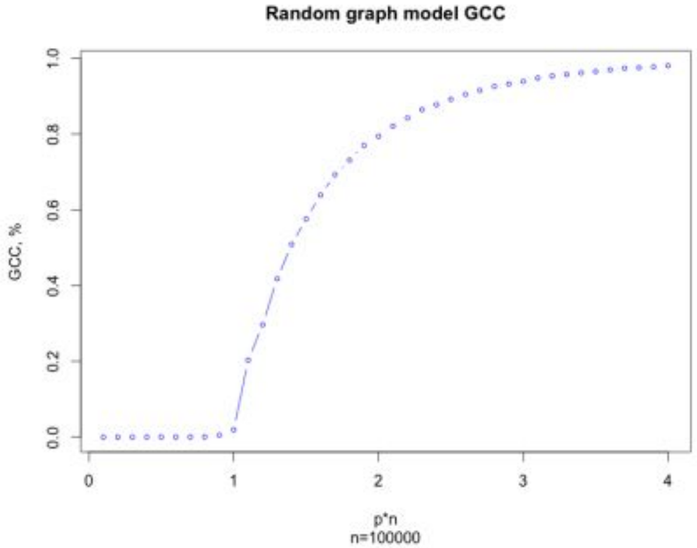


# What about connectivity?

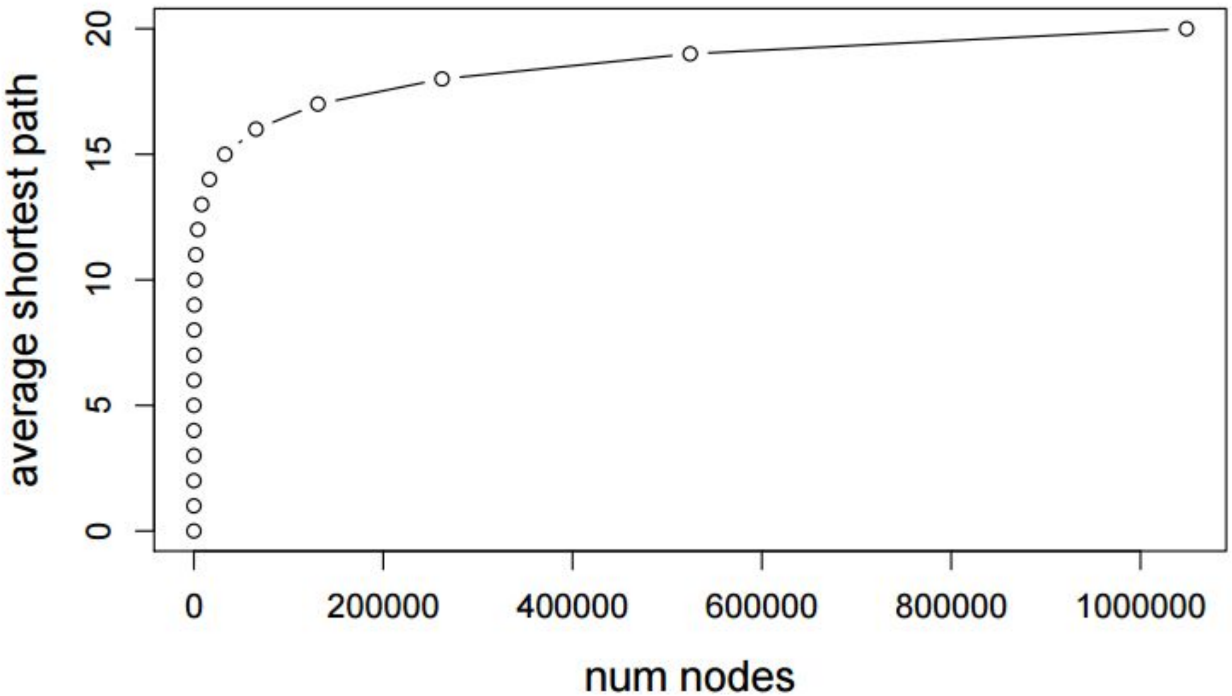
$$\langle k \rangle = pn$$




- It could be shown than with <k> = 1 the largest connected component contains $O(n^{2/3})$ nodes.
- With <k> > 1 it quickly has all the nodes.

# Average path length




For Erdos-Renyi graph average path length is of order *O(log n)*

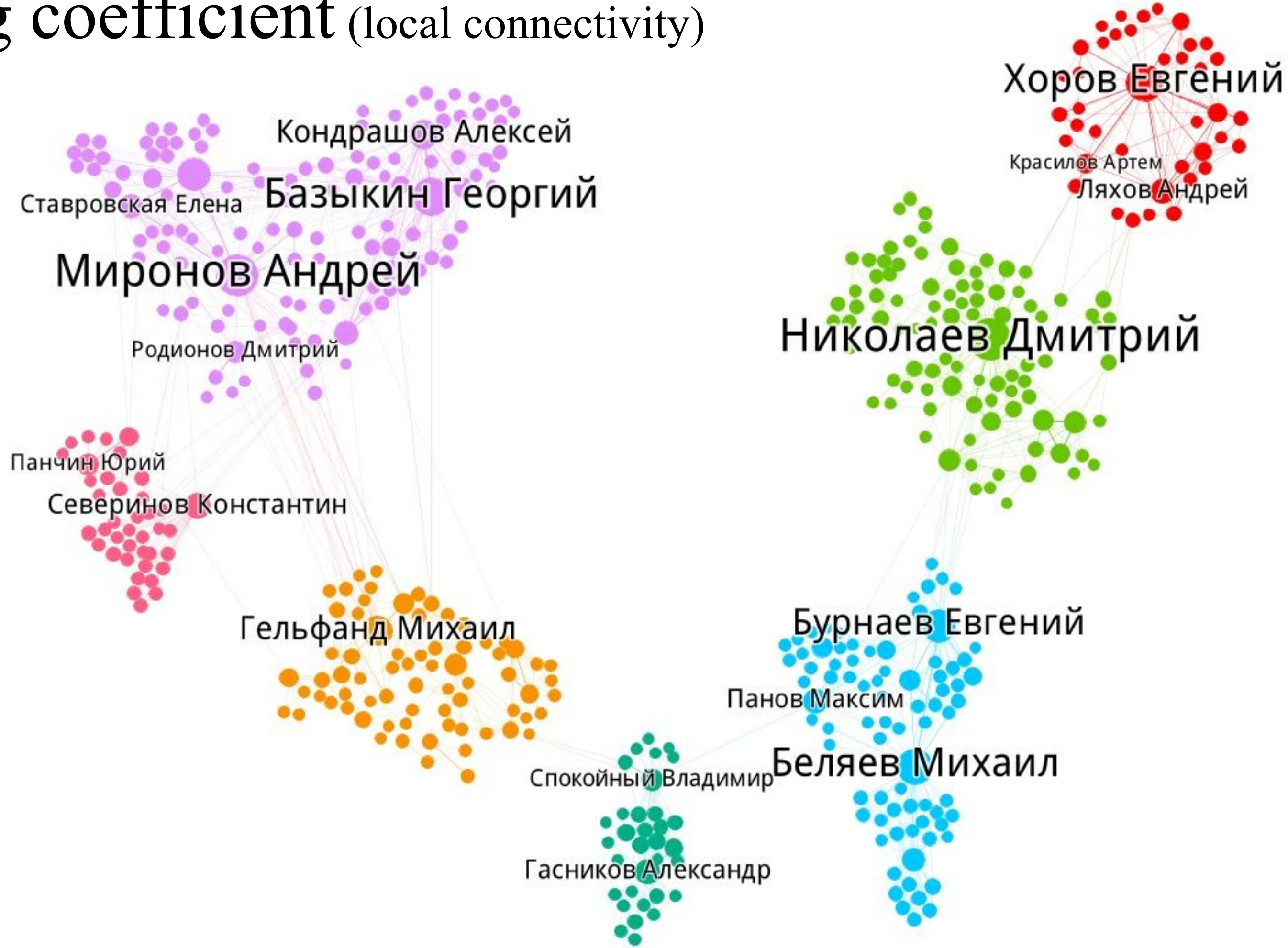
# How to compare different networks?

- # nodes, #edges, density
- (Average) node degree
- (Average) clustering coefficient
- (Average) path length
- **Node degree distribution**
- Centrality measures (next lecture)

# Key properties

- (Average) clustering coefficient
- (Average) path length
- Node degree distribution

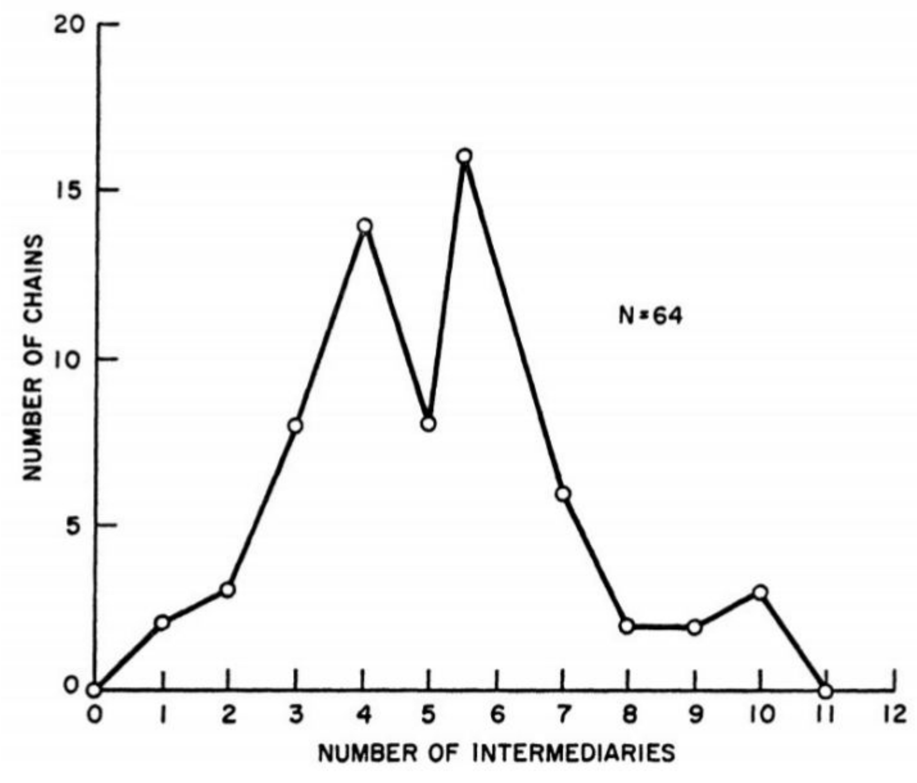
# Clustering coefficient (local connectivity)



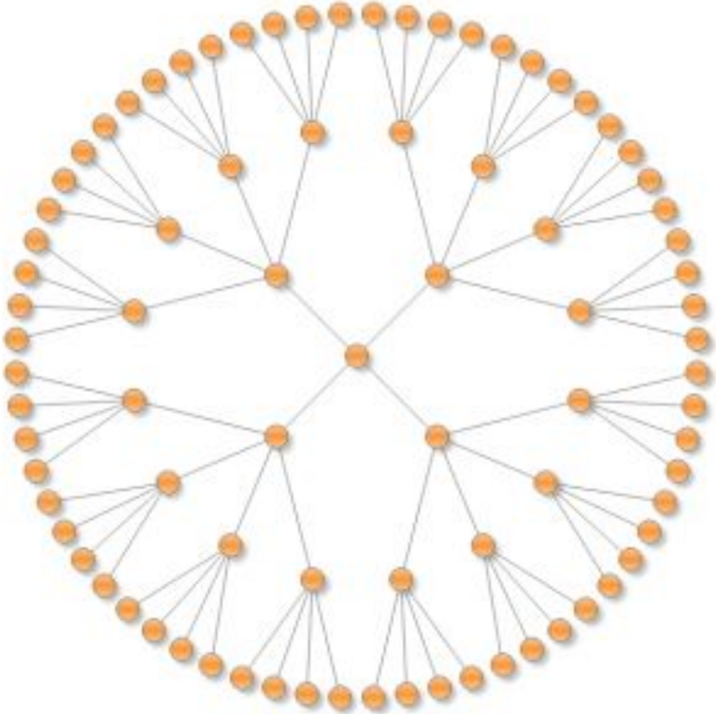
# Key properties

- (Average) clustering coefficient
- (Average) path length
- Node degree distribution

# Recall Milgram’s experiment



# Average path length (idea)



An estimate: $z^d = N$, $d = \log N / \log z$
$N \approx 6.7$ bln, $z = 50$ friends, $d \approx 5.8$.

Consider a simple model:

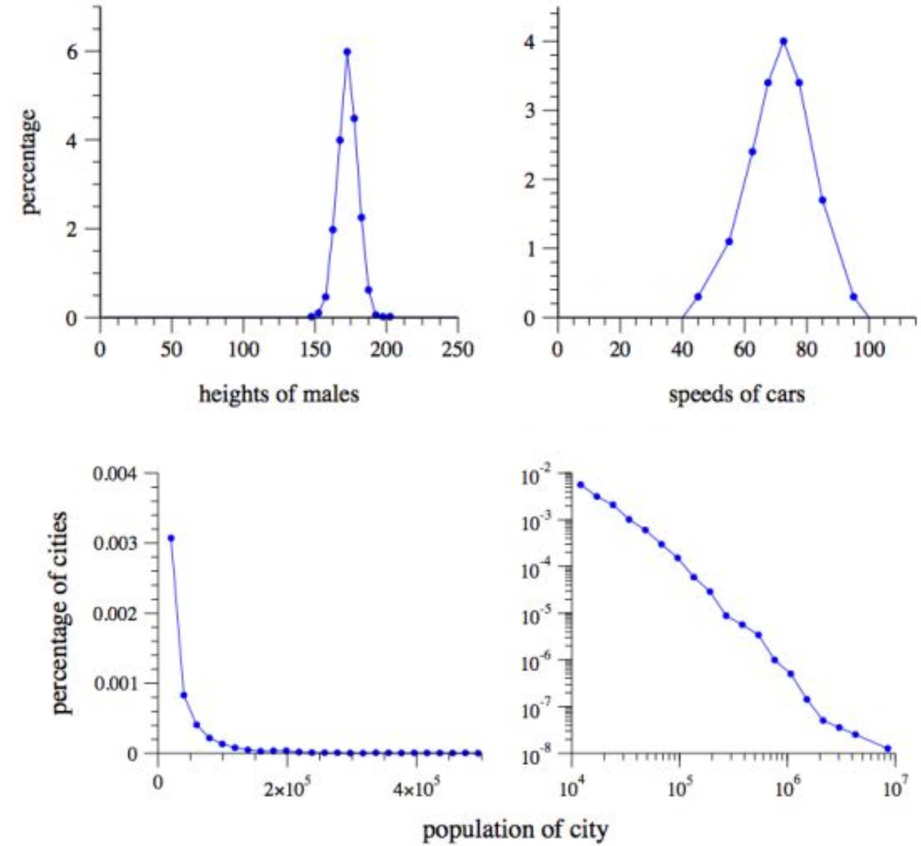
Each person has the same number of friends *z*, total # of people in the world is N, then what is a diameter?

<u>Diameter</u> = the longest of all the calculated shortest paths in a network

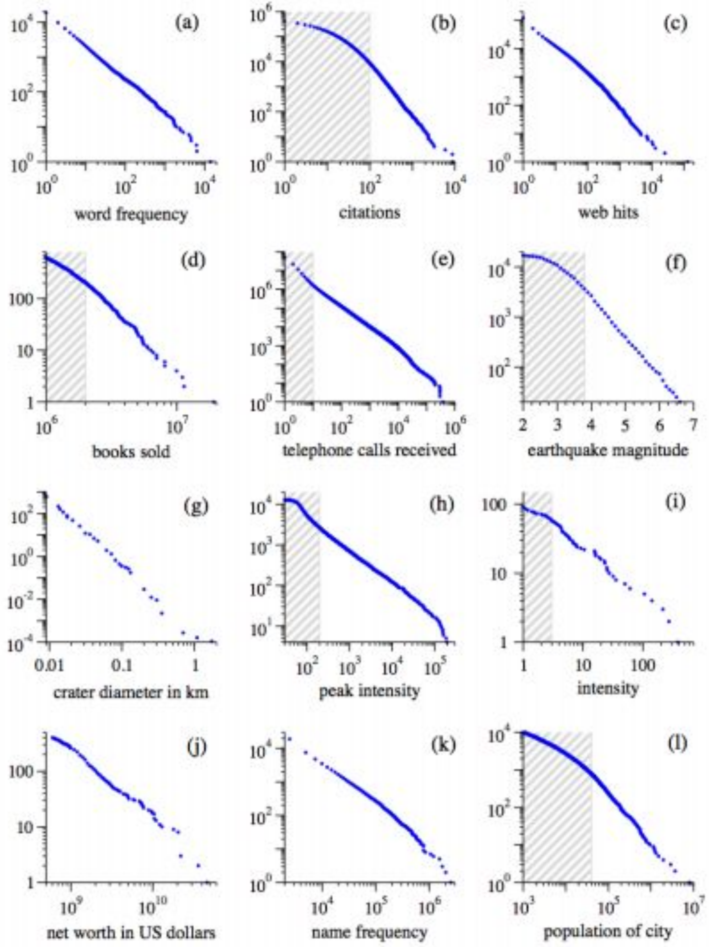
# Key properties

- (Average) clustering coefficient
- (Average) path length
- Node degree distribution

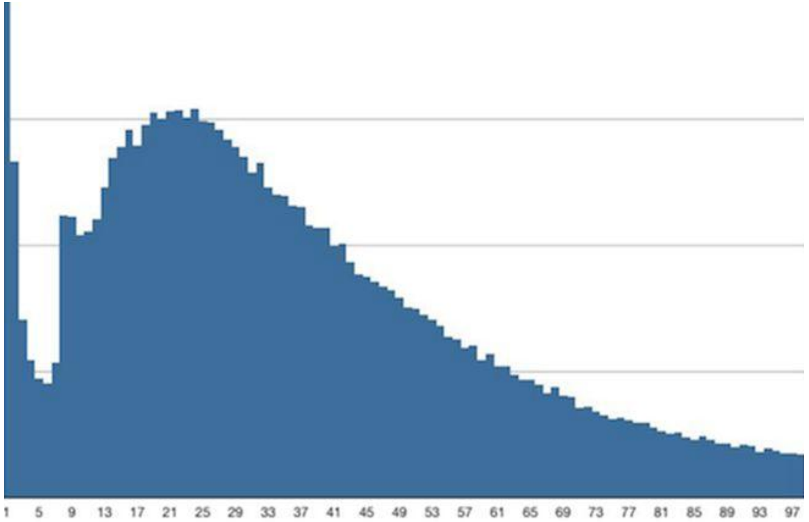
# Empirical distributions



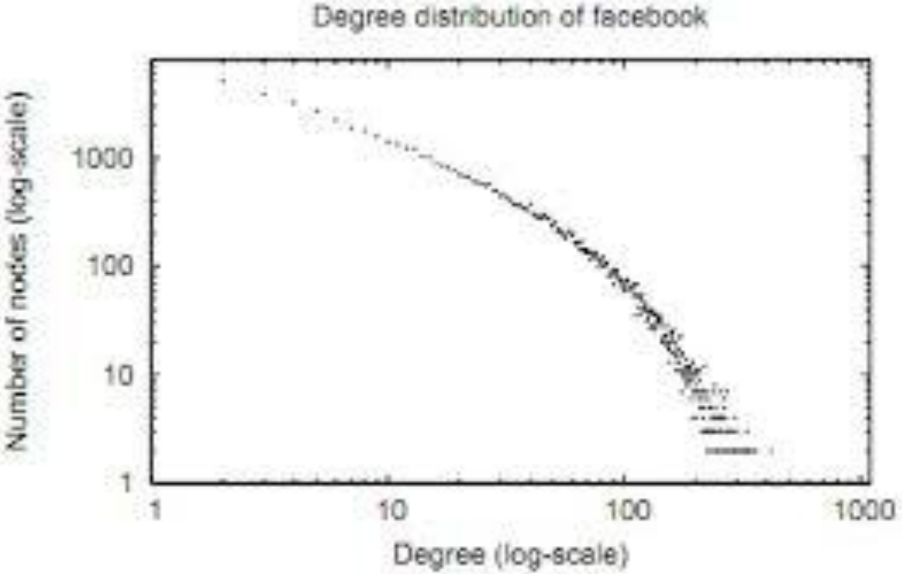

log-log scale



# Facebook degree distribution




y = number of people; x = number of friends for those people



# Empirical network features

- Power-law (heavy-tailed) degree distribution
- Small average distance (graph diameter)
- Large clustering coefficient (transitivity)
- Giant connected component, hierarchical structure,etc