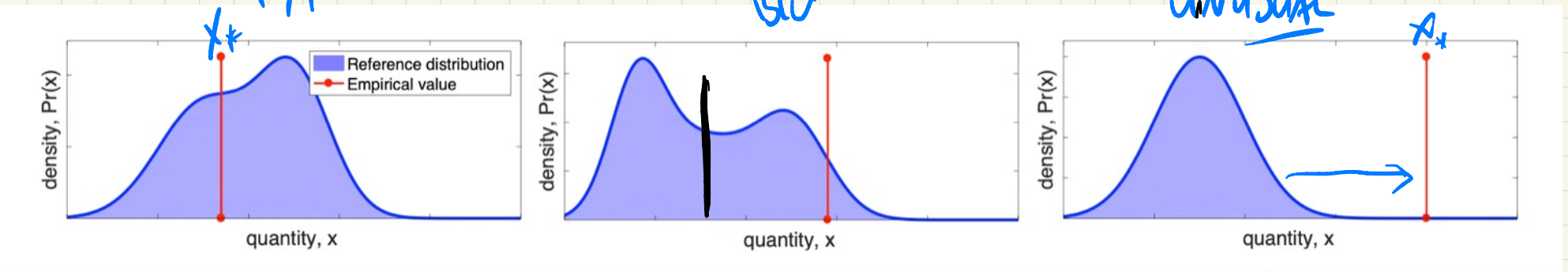
Lecture 3: random graph models

How do we interpret the value of $\langle k \rangle$, $\langle \ell \rangle$, $Pr(k)$, C, ...? Is it big or small? Is it typical or unusual?

reference distribution $Pr(x)$

empirical x_*



For networks 2a a random graph model defines the reference $Pr(x)$

2b many different models, with different assumptions

3 types of random graphs (for now...)

1) Erdös - Rényi (ER) random graph
   * edge density + iid

2) Configuration Model (and the Chung-Lu model)
   * degree structure (otherwise random)

3) Modular random graphs
   * groups of nodes

The Erdős-Rényi random graph

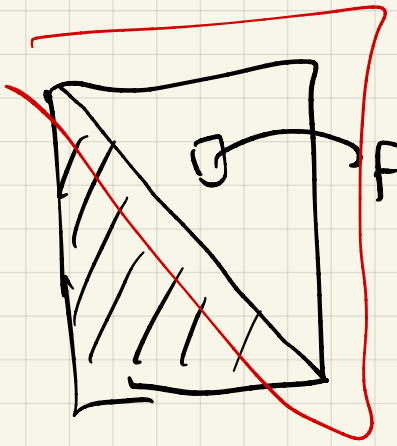
$2^{\binom{n}{2}} = O(2^{n^2})$

$G(n, p)$

n ↑ # nodes

p ↑ prob. $(i,j) \in E$

- defines an ensemble of graphs $Pr(G|p)$



unweighted

$\forall_{i>j} \quad \underbrace{A_{ij} = A_{ji}}_{\text{undirected}} = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{otherwise} \end{cases}$

expected degree c ← $\forall i$

$p = \frac{c}{n-1}$

$n-1$ ← # possible adjacencies

* $p = 1$ → complete graph on n nodes
* $p = 0$ → empty graph on n nodes

Properties of ER graphs

- simple graph
- connectivity is "homogeneous" $\Rightarrow$ edge "density" + iid
- deg. dist. $Pr(k) \cong Poisson(c)$ $\quad c = p(n-1)$
- diameter and MGD $\rightarrow O(\log n)$ $\Rightarrow$ "small-world"-like netw
- clustering $\Delta \rightarrow O(1/n)$
- largest connected component (LCC) $\Rightarrow O(n)$ when $c > 1$

Generating a $G(n,p)$ network

initial empty graph $G = (V, E)$ $\quad |V| = n$ $\quad E = \emptyset$

$\Theta(n^2)$ for each possible (i,j), draw $r \sim Uniform(0,1)$

if $r \leq p$, add (i,j) to E

can do it in $O(n+m)$ time if you're clever

Mean degree

$$\langle k \rangle = c = \sum_{j=1}^{n-1} p = p\left(\sum_{j=1}^{n-1} 1\right) = p(n-1)$$

$$\langle m \rangle = \sum_{i=1}^{n} \sum_{j>i} p = p\left(\frac{n}{2}\right) = p\left(\frac{n(n-1)}{2}\right)$$

recall $\langle k \rangle = \frac{2m}{n}$

$$\langle k \rangle = \frac{2}{n}\left(p\left(\frac{n(n-1)}{2}\right)\right) = p(n-1)$$

degree distribution

$$Pr(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}$$

$$= \frac{c^k}{k!} e^{-c}$$

c^k $\quad$ $k!$

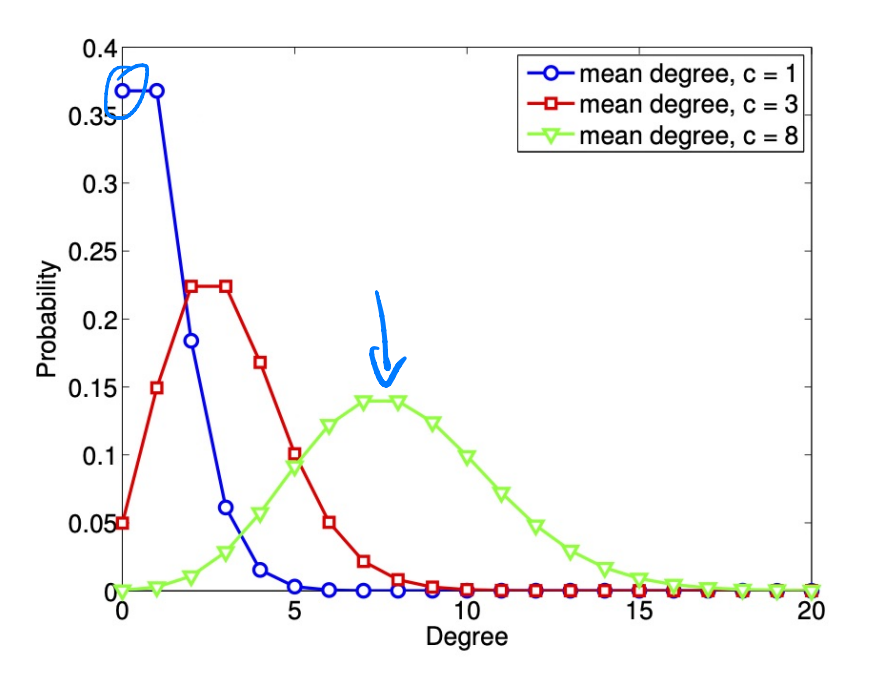
Most nets are sparse

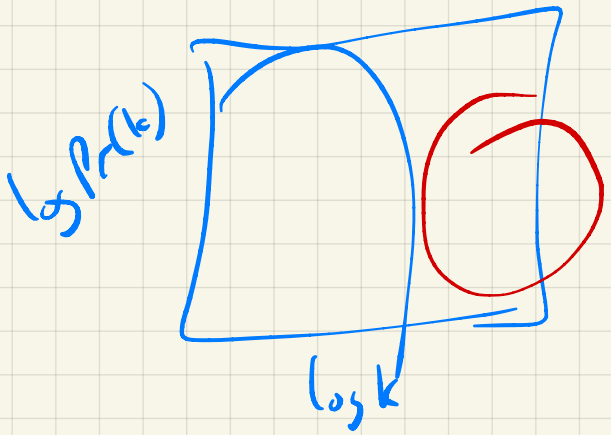
$\langle k \rangle = O(1)$

$= p(n-1)$

$p \approx 0$

$p = \frac{c}{n-1}$



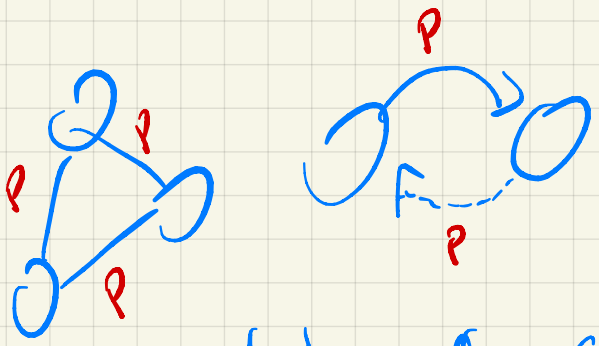


← these are not heavy-tailed distributions!

Motifs, reciprocity, clustering coefficient

$$\text{reciprocity } r = \frac{\#\text{ of recip. links}}{\#\text{ of links}} = \frac{(n^2-n)\,p^2}{(n^2-n)\,p} = p = O(1/n)$$

$$= \frac{c}{n-1}$$



$$\text{clust. coeff } C = \frac{\#\,\Delta}{\#\,\wedge} = \frac{\binom{n}{3}p^3}{\binom{n}{3}p^2} = p = O(1/n)$$

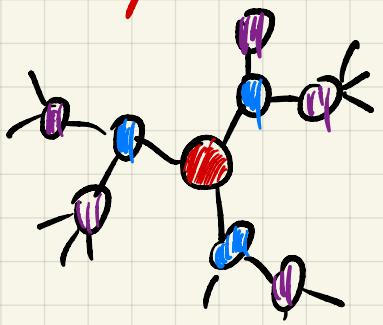
Diameter and MGD

$$(c-1)^{\ell_{max}} = n$$

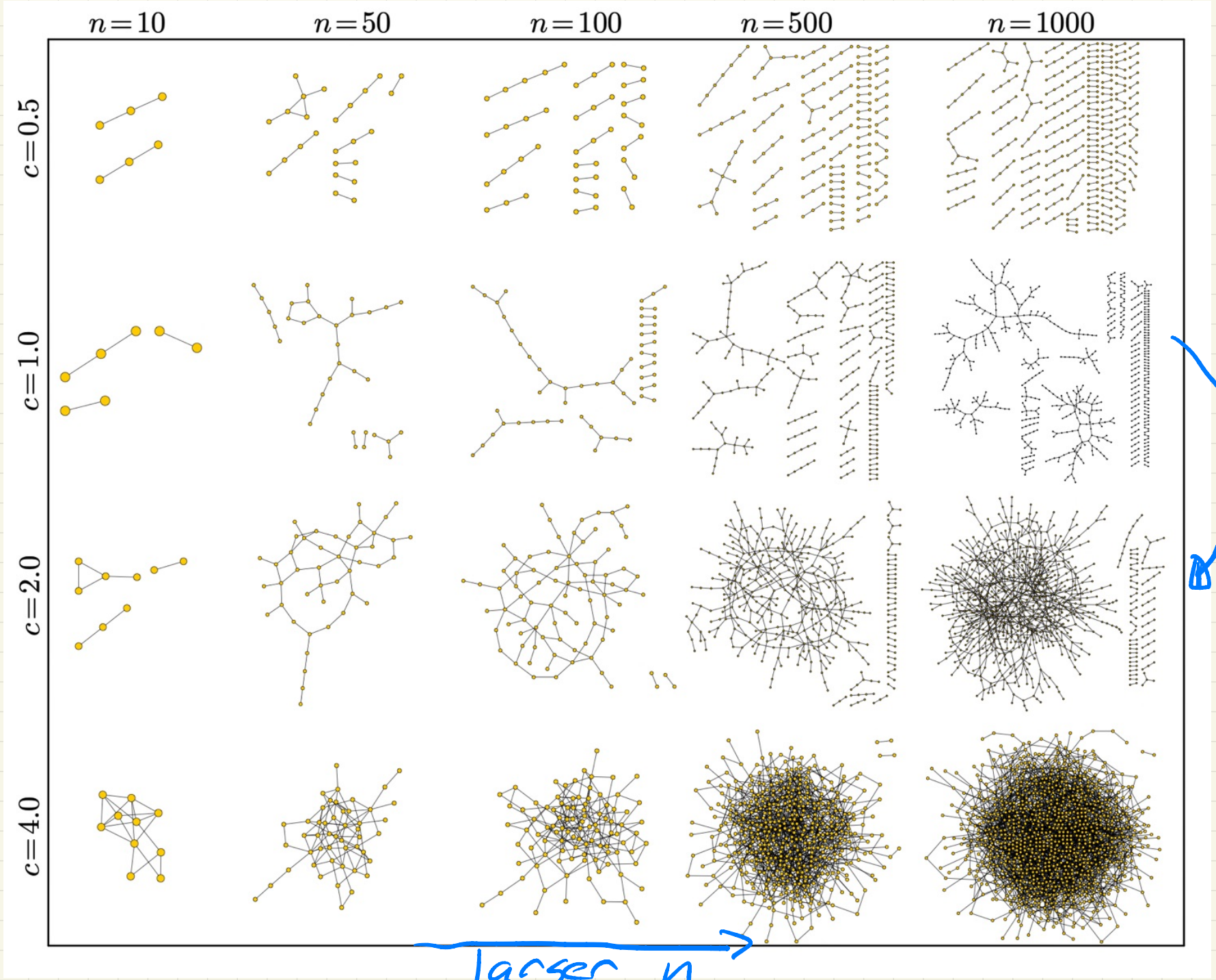
$$\ell_{max} = O(\log n)$$

ER graphs are

* locally tree-like *



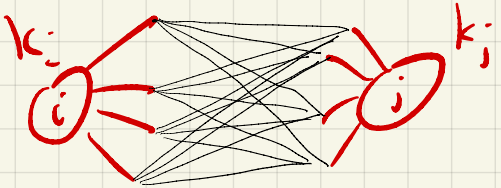
What do ER [G(n,p)] graphs look like?



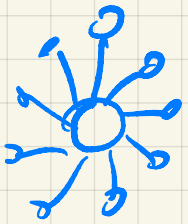
The configuration model

assume degree seq $\vec{k} = (k_1, k_2, k_3 \dots k_n)$

take from data



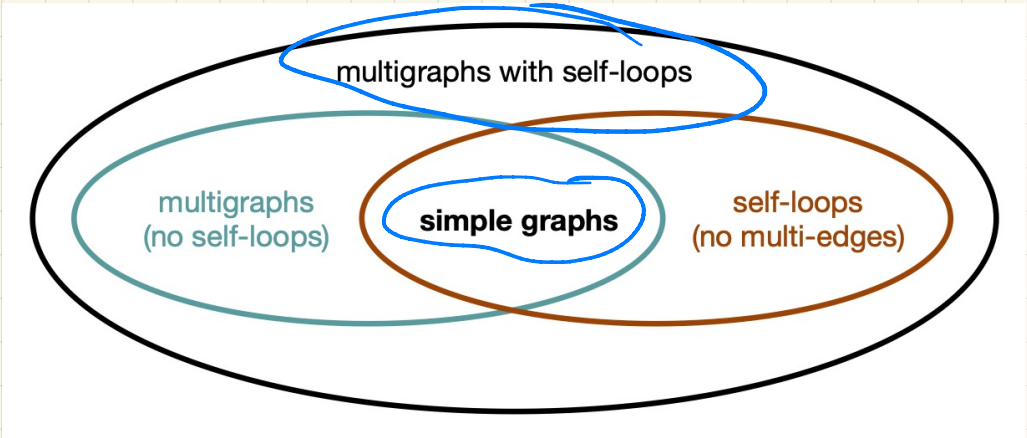
$k_i k_j$

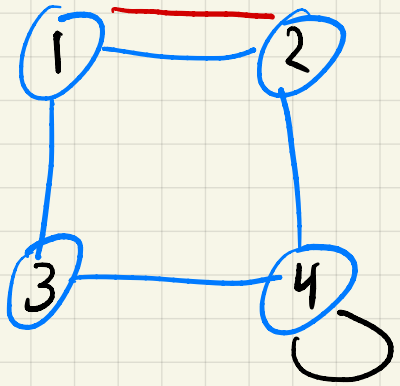


$$\forall_{i>j} \quad A_{ij} = A_{ji} = \begin{cases} 1 & \text{with probability } \propto k_i k_j \\ 0 & \text{otherwise} \end{cases}$$

($\propto$ means "is proportional to")

↳ contrast this with ER model of $p = \frac{2m}{n(n-1)}$ (a constant value)





Properties of the configuration model $G(n, \vec{k})$

- four flavors: simple or multigraph / w/ or w/o self-loops
- connectivity specified by $\vec{k}$
- diameter and cc $\sim O(\log n)$ ⇐ locally treelike
- motif freq $\triangle$ $\square$ $\sim O(1/n)$
- LCC $\sim \Theta(n)$ if G not too sparse

depend on mean $\langle k \rangle$ and variance $\langle k^2 \rangle$

generating a config. model

2 facts make it tricky:

1) not every $\vec{k}$ is graphical

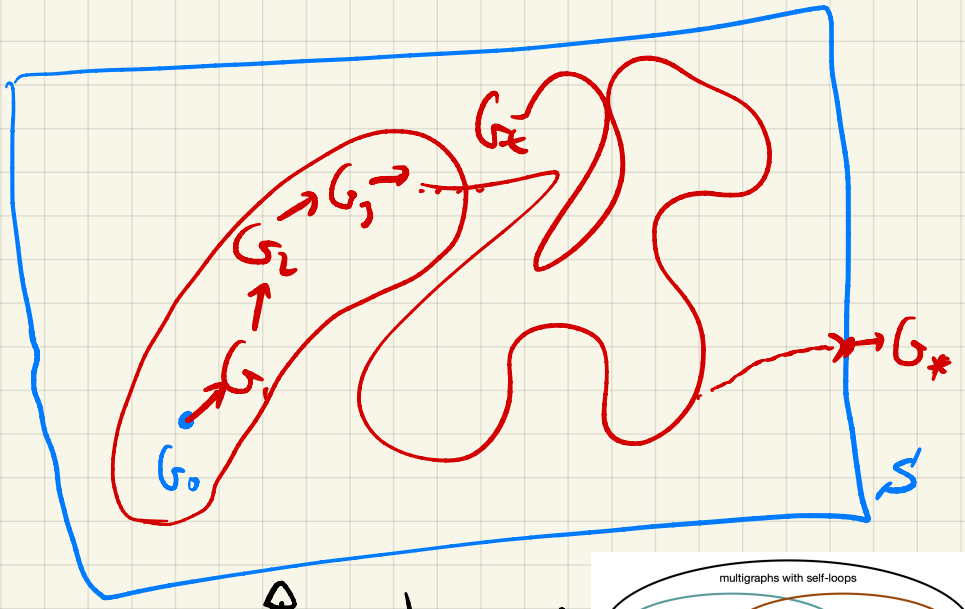
2) no simple procedure for $G \sim P(G)$

Solution: Markov chain Monte Carlo (MCMC)

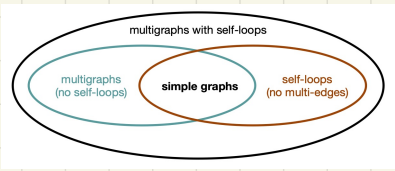
two parts:

1) generate a $G_0 \in S$ ← graph space

2) $g:(G_t) \to G_{t+1} \in S$



↑ graph space:

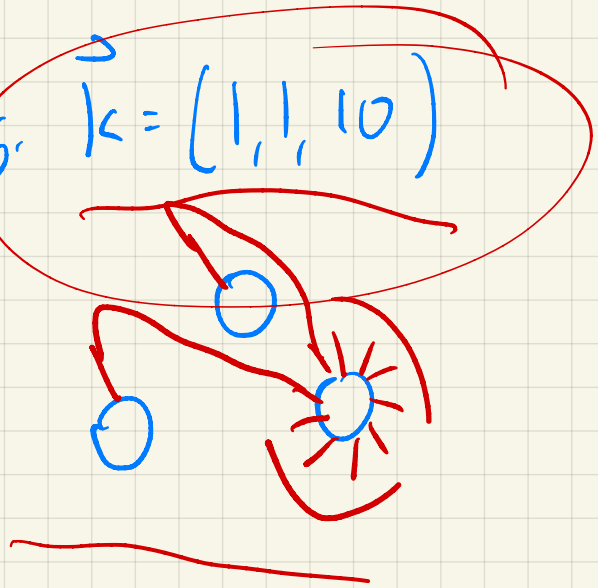


$Pr(G|\vec{x})$

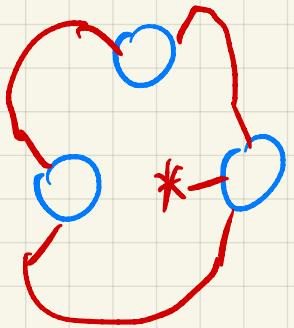
Is it graphical?

- Can we make a graph with $\vec{k} = \{k_1, k_2, k_3 \ldots k_n\}$?
  - So if so, then every "stub" can be paired up correctly

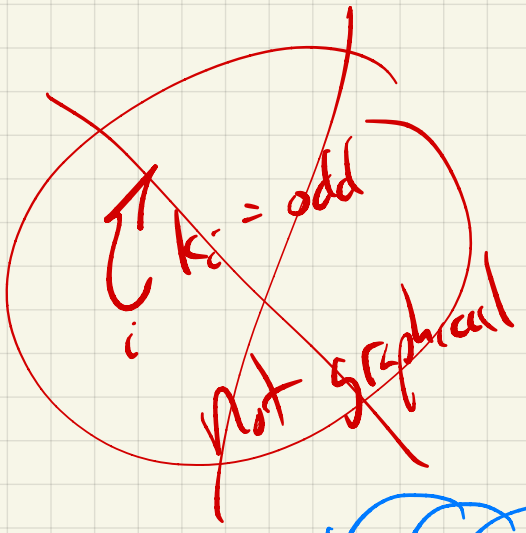
e.g. $\vec{k} = (1, 1, 10)$

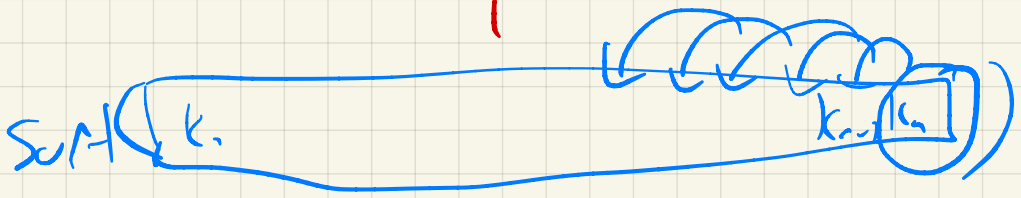


or $\vec{k} = (2, 2, 3)$

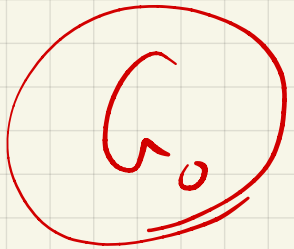


- Havel-Hakimi



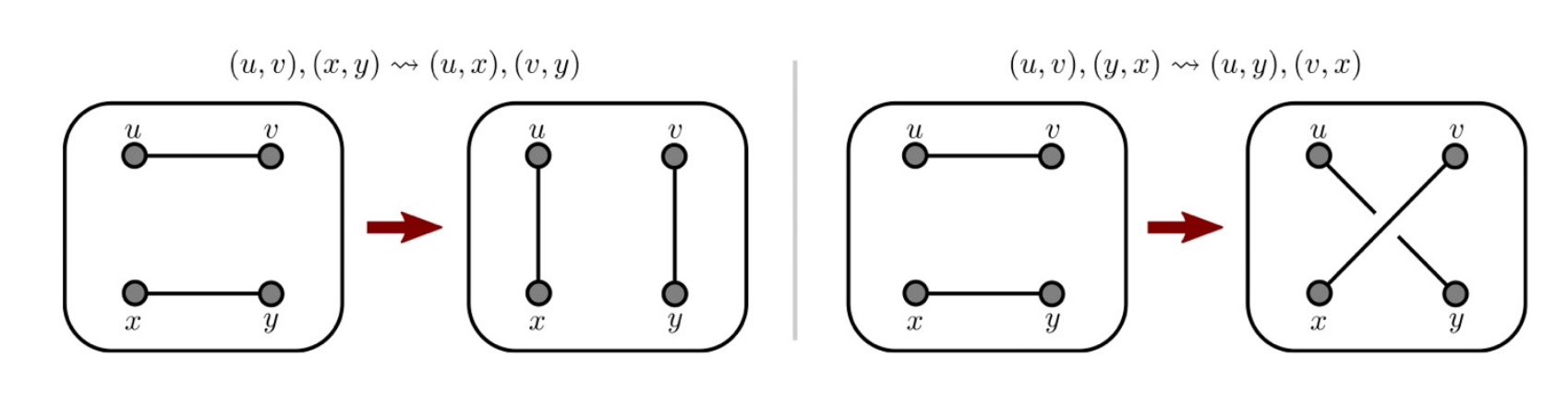


makes a synthetic seed for MCMC



g: the double edge swap : given $G_t \in \mathcal{S}$, apply $g(G_t) \to G_{t+1} \in \mathcal{S}$

- (select) choose uniformly at random
- (swap) choose output 1 or output 2 with equal prob
- (accept) if $G_{t+1} \in \mathcal{S}'$



degree preserving : k_u, k_v, k_x, k_y don't change

$G_0, G_1, G_2, \dots, G_t \equiv G_0, g(G_0), g(g(G_0)), \dots, g^t(G_0)$

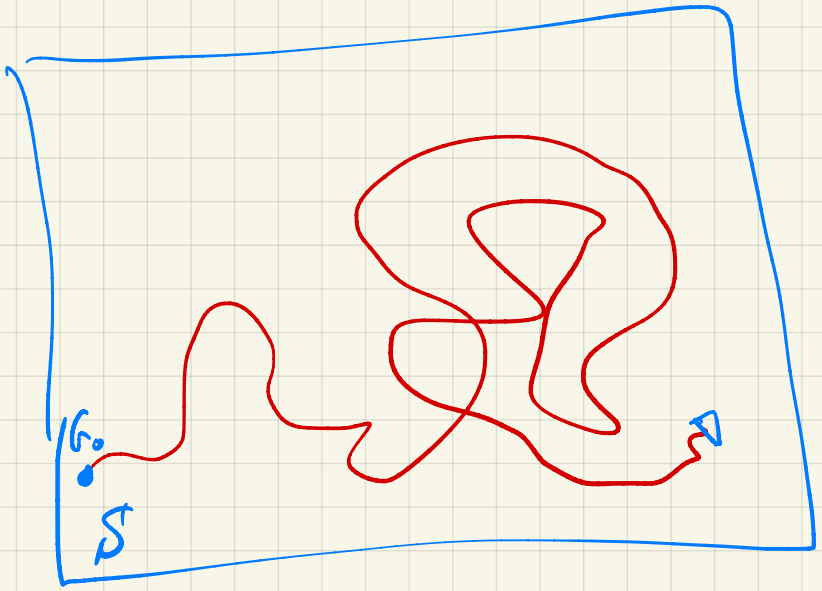
the chain

How long of a chain? (mixing times)

want independent samples:

G_t and $G_{t+\eta_0}$ (η_0: sampling gap)

heuristic: $\eta_0 = \frac{M}{3} + 300$

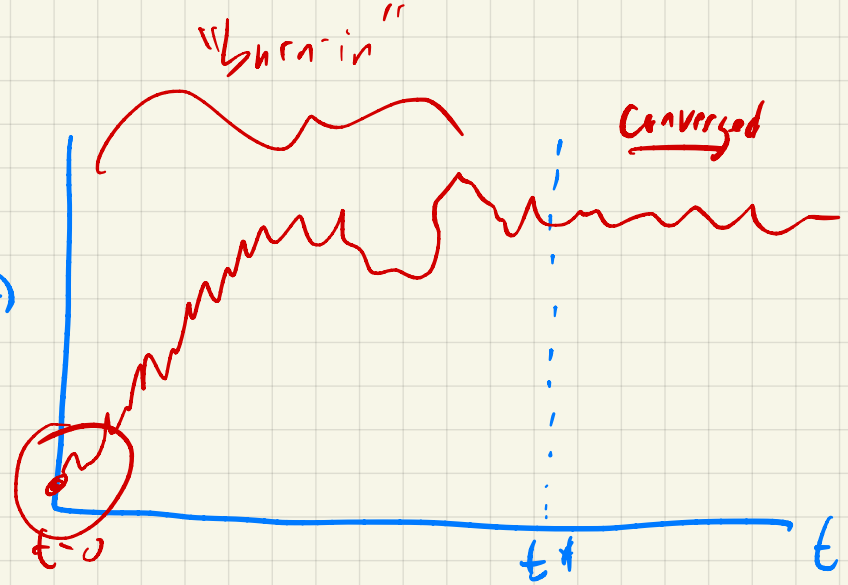


want $Pr(G)$ to be uniform over S

detect convergence

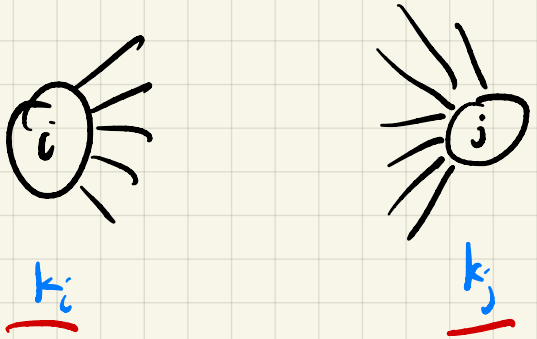
↳ test on two samples from MCMC

summary statistic



Chung-Lu model (simple graphs or directed graphs)

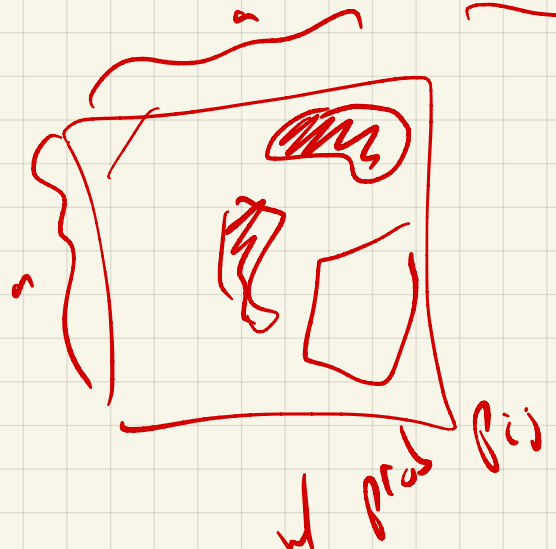
- like configuration model, but only $E[\vec{k}]$ rather than exactly $\vec{k}$



$$p_{ij} = k_i\left(\frac{k_j}{2m - k_i}\right) = \frac{k_i k_j}{(2m - k_i)}$$

if $\max_i k_i < 2m - k_i$

$$\forall_{i>j}\ A_{ij} = A_{ji} = \begin{cases} 1 & \text{with prob. } p_{ij} \\ 0 & \text{otherwise} \end{cases}$$



generating Chung-Lu:

for $i = 1$ to n
  for $j = i+1$ to n
    $A_{ij} = A_{ji} = 1$ ↓ prob

Clustering coefficient

$$C = \frac{1}{n}\frac{[\langle k^2\rangle - \langle k\rangle]^2}{\langle k\rangle^3} = O(1/n)$$ when $\langle k\rangle$, $\langle k^2\rangle$ are finite

Giant component [LCC = Θ(n)]

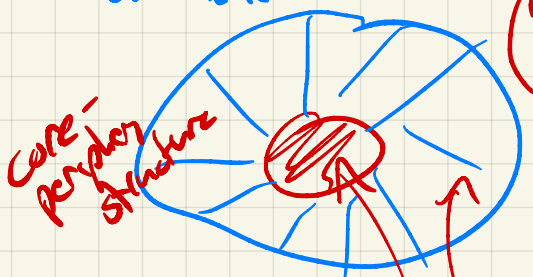
$$\langle k^2\rangle - 2\langle k\rangle > 0$$

diameter

$O(\lg n)$

sometimes $(O(\ln \ln n))$ if $\langle k^2\rangle$ very large

ultra small world

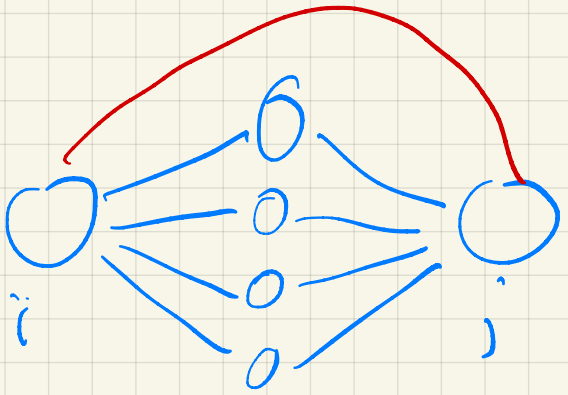


# of common neighbors

$$n_{ij} = P_{ij}\left(\frac{\langle k^2\rangle - \langle k\rangle}{\langle k\rangle}\right)$$

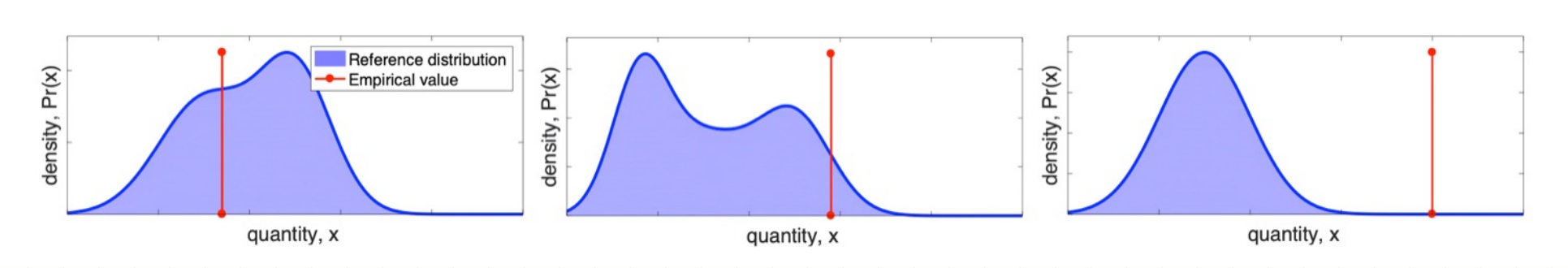
high degree nodes here

low degree nodes here



Random graphs as null models

recall: random graphs produce $Pr(G)$ as a reference distribution



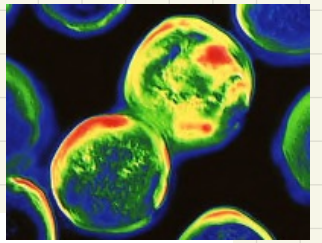
null model:

How much of an observed pattern is explained by {edge density or degrees} alone, under randomness?

need $Pr(G|\theta)$ ← est. from our data

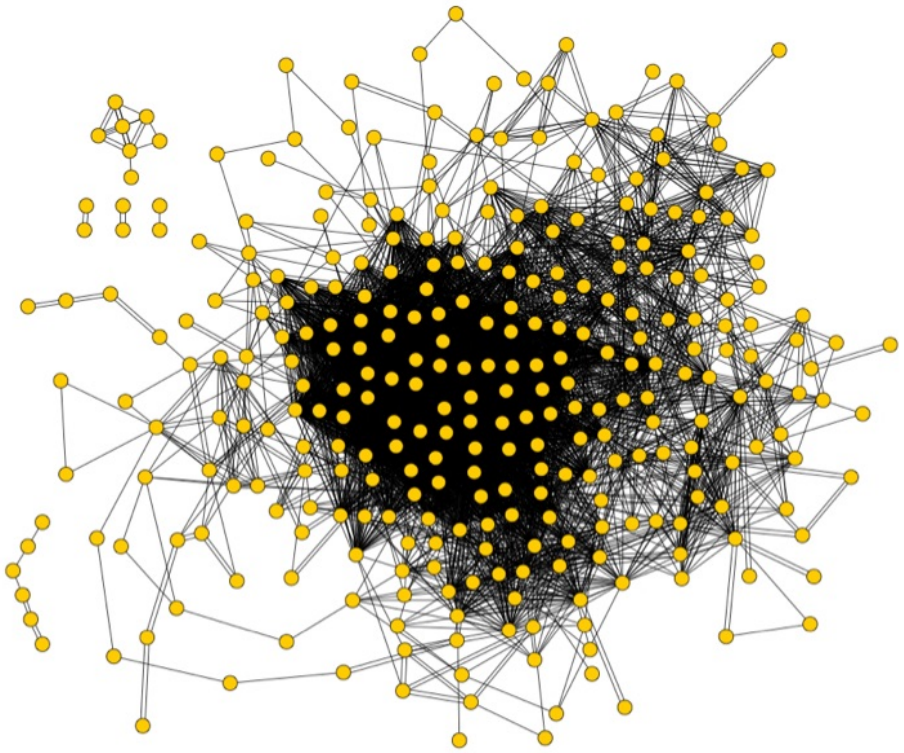
$\hat{p} = \frac{\langle k \rangle}{n-1}$

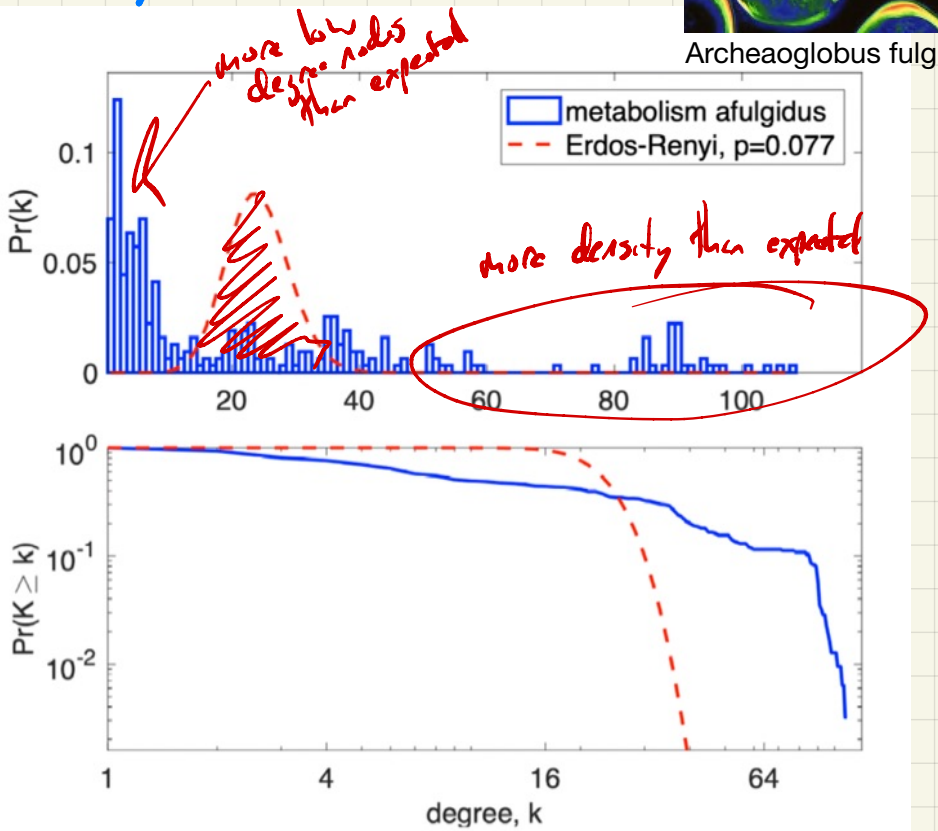
$\hat{\vec{k}} = \{k_1, k_2, \ldots, k_n\}$



Archeaoglobus fulgidus

$n = 315$ $m = 3793$ (undirected edges)





$\langle k \rangle = 24.08$ $\sigma_k = 28.17$ LCC: 296 nodes (94%)

ER model: $\hat{p} = \frac{\langle k \rangle}{n-1} = 0.077$

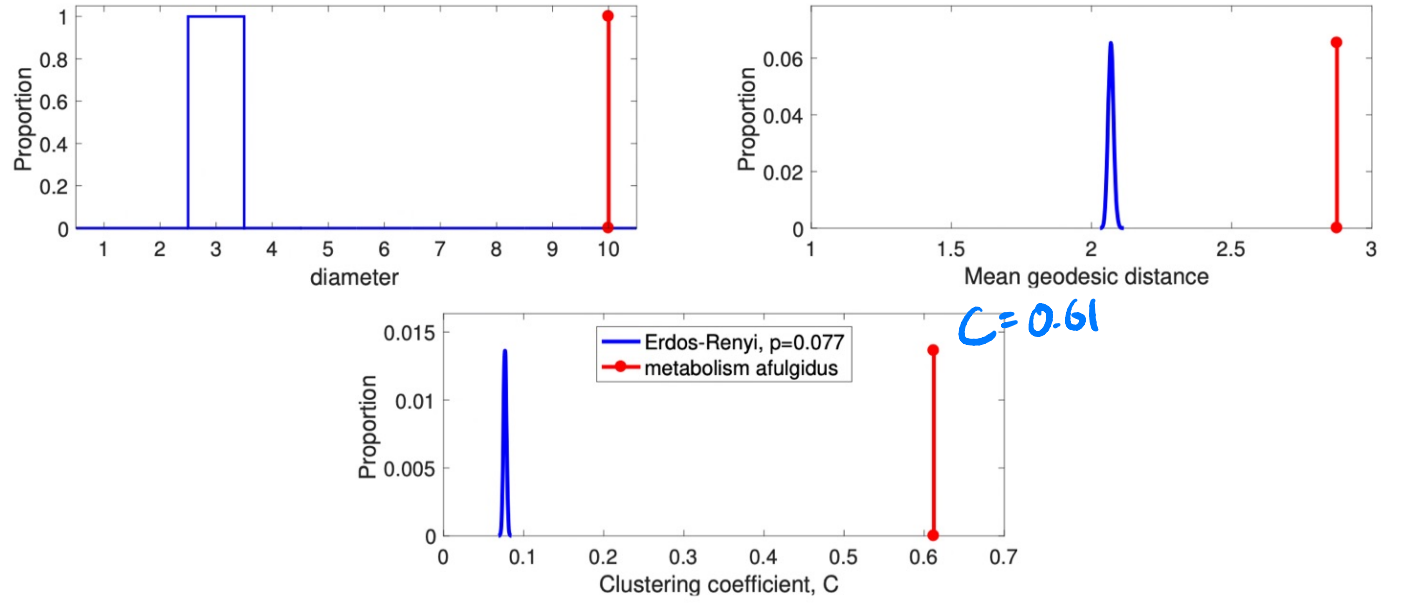
$k_{max} = 108$

$\Pr(k_{max} = 108 \mid ER, \hat{p}) = 4.3937 \times 10^{-36}$

$\ell_{max} = 10$

$\langle \ell \rangle = 2.88$

null ER Model (edge density alone)



$C = 0.61$

configuration null model (degree structure alone)