Lecture 5: modular networks, structure

"large-scale" structure

↓

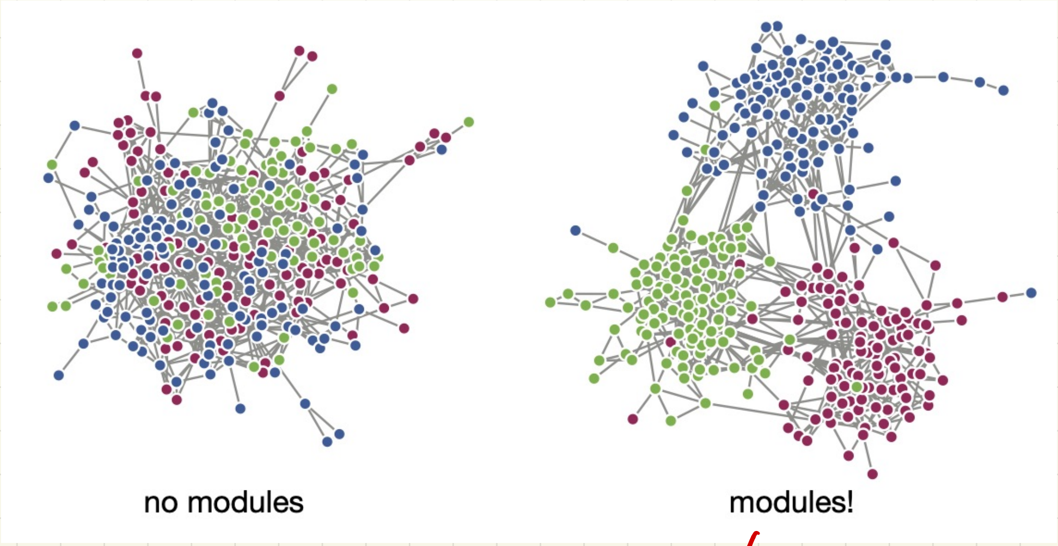
clustering node in groups

Module = group = community

group-level rep.

statistical notion

a **community** *is a group of nodes that connect to other groups in similar ways.*



"assortative" modules

finding the modules

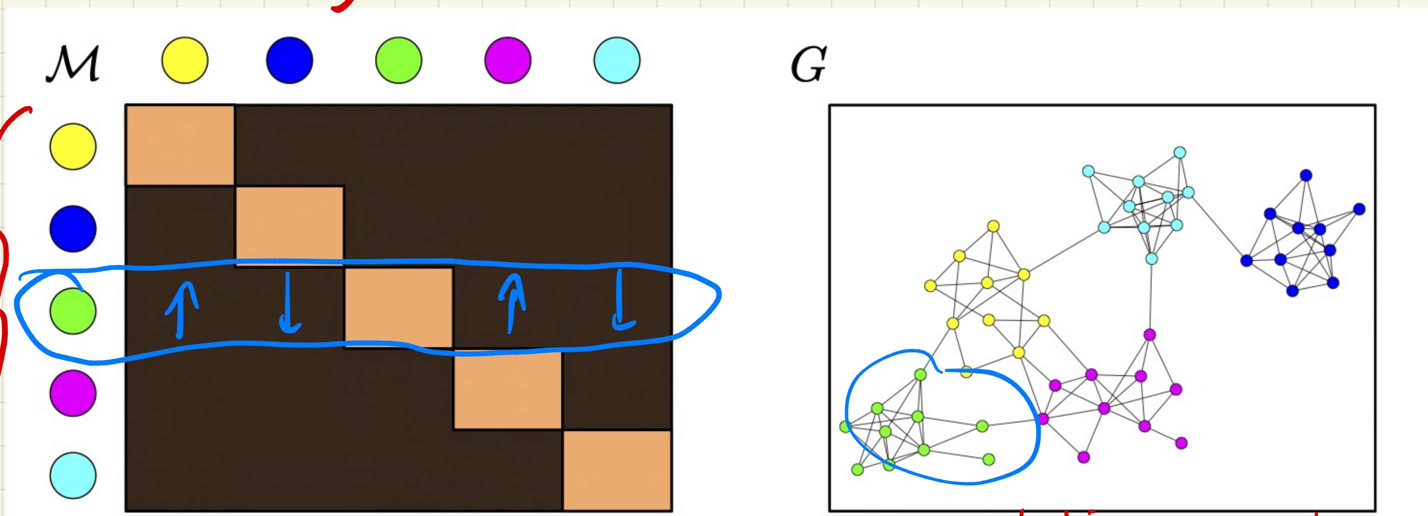
→ a "coarse graining" of the system

this is useful!

- functional clusters of genes/proteins
- computational groups of neurons
- groups of functionally similar species (e.g. predators)
- social units

Representing modular interactions

the mixing matrix $\mathcal{M}$ → inter and intra group interaction patterns



M_{rs} : density of connections from group r to group s

M_{rr} : within group density

c

c×c matrix

assortative communities → like links with like

modular structure generalizes Erdős-Rényi random graphs

$\forall_{rs} \; M_{rs} = p$

$$\begin{bmatrix} p & p & p \\ p & p & p \\ p & p & p \end{bmatrix}$$

$G(n,p)$

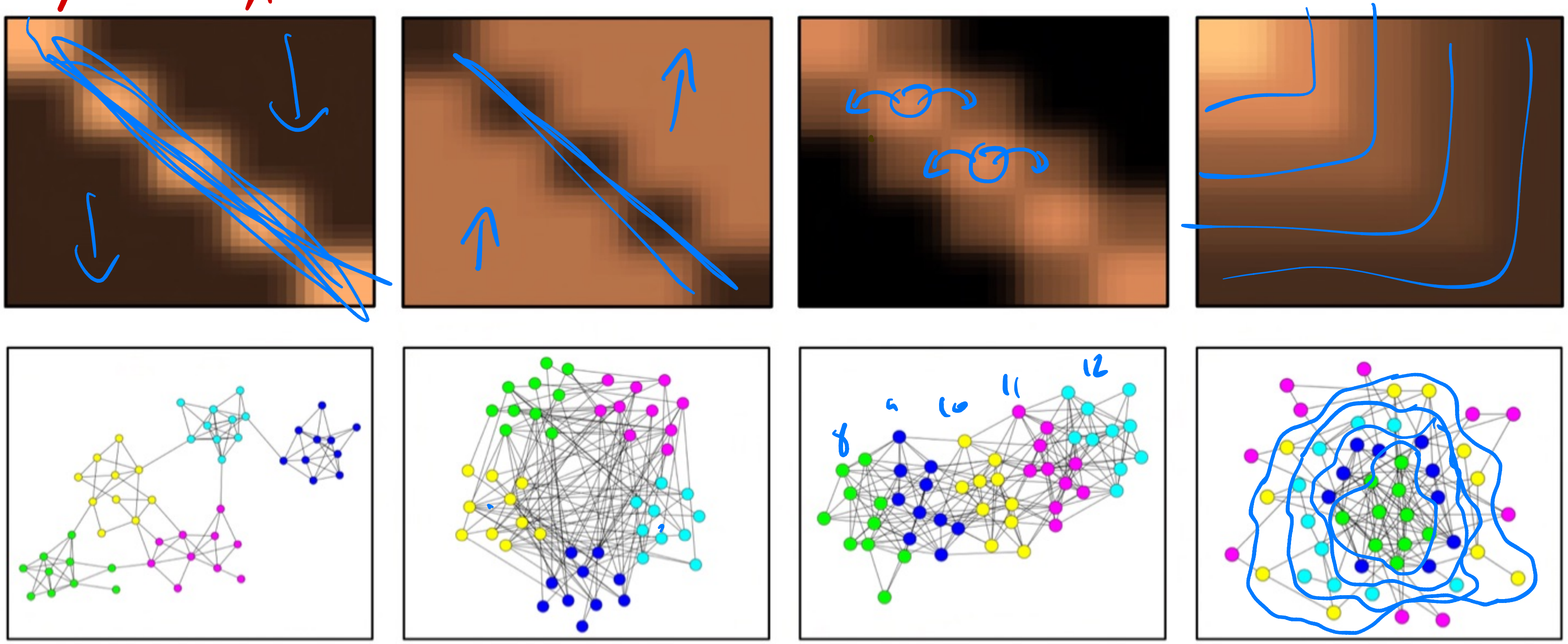
mixing matrices capture many patterns, depending on how M_{rs} values are organized.

c, the # of groups, sets the scale of the pattern

more groups = smaller scale patterns

→ more complex mixing patterns

4 stylized types of modular structure

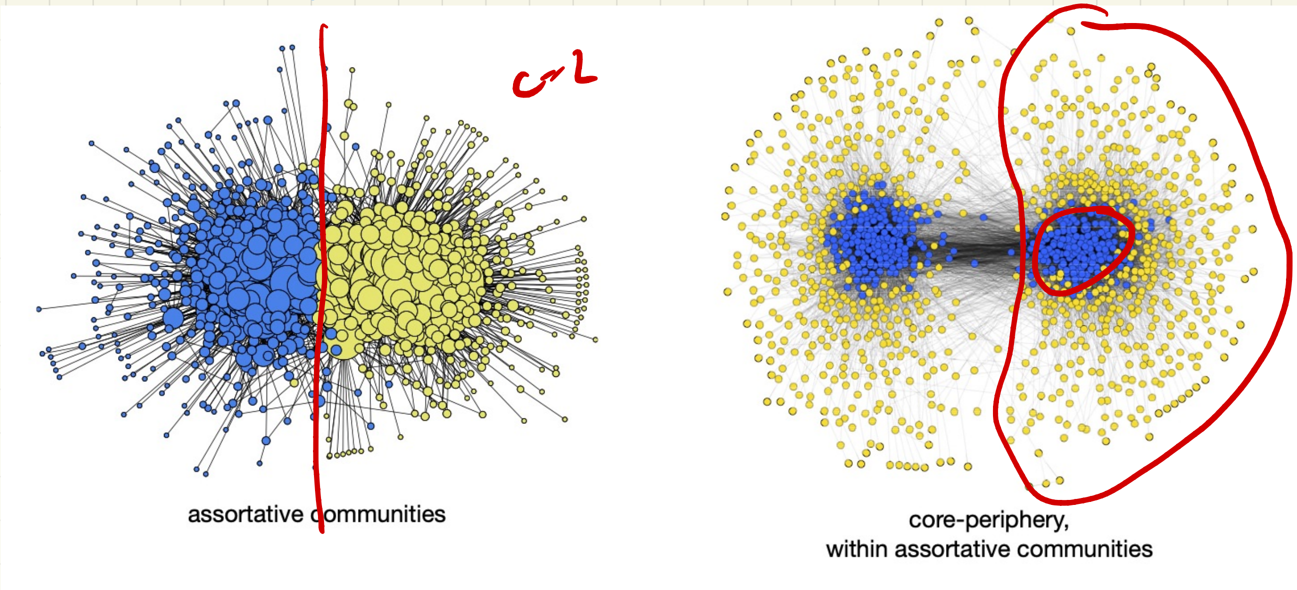


assortative	disassortative	ordered	core-periphery
edges inside communities	*edges between communities*	*linear group hierarchy*	*dense core, sparse periphery*
$M_{rr} > M_{r\neq s}$	$M_{rr} < M_{r\neq s}$	$M_{rr} > M_{r,r+\ell}$ for $\ell \geq 1$	$M_{11} > M_{21} = M_{22} = M_{21} > M_{31} \dots$
			core ... layer 1 ... layer 2

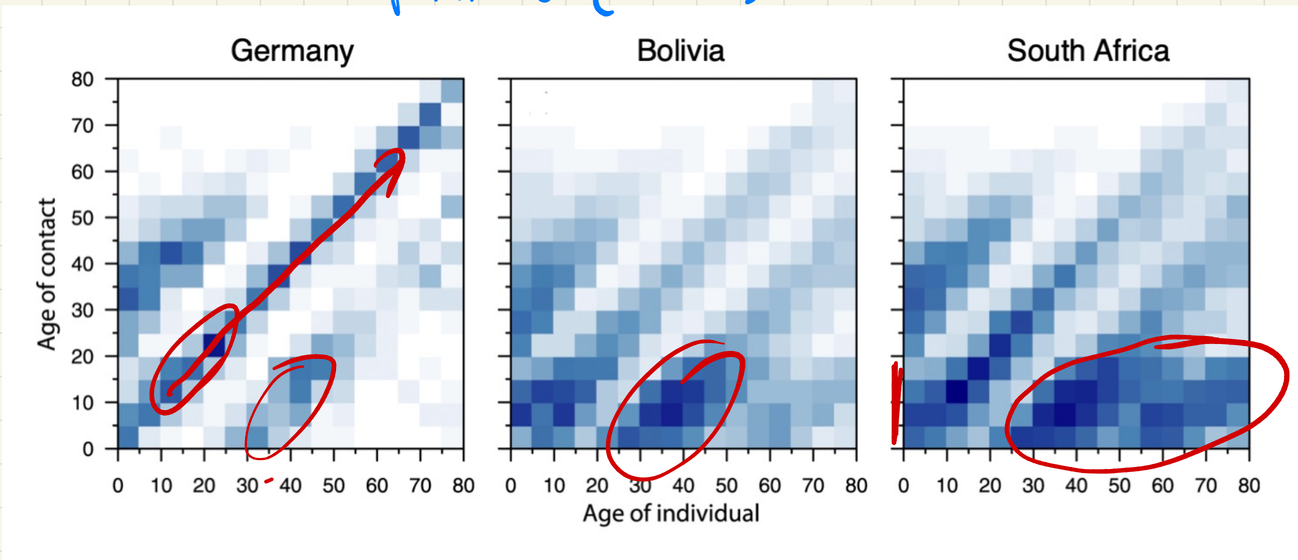
these examples illustrate undirected mixing matrices, but in directed networks $M_{rs} \neq M_{sr}$, which further enriches the representational breadth of modular structure

Some examples (with "clean" versions of our stylized patterns)

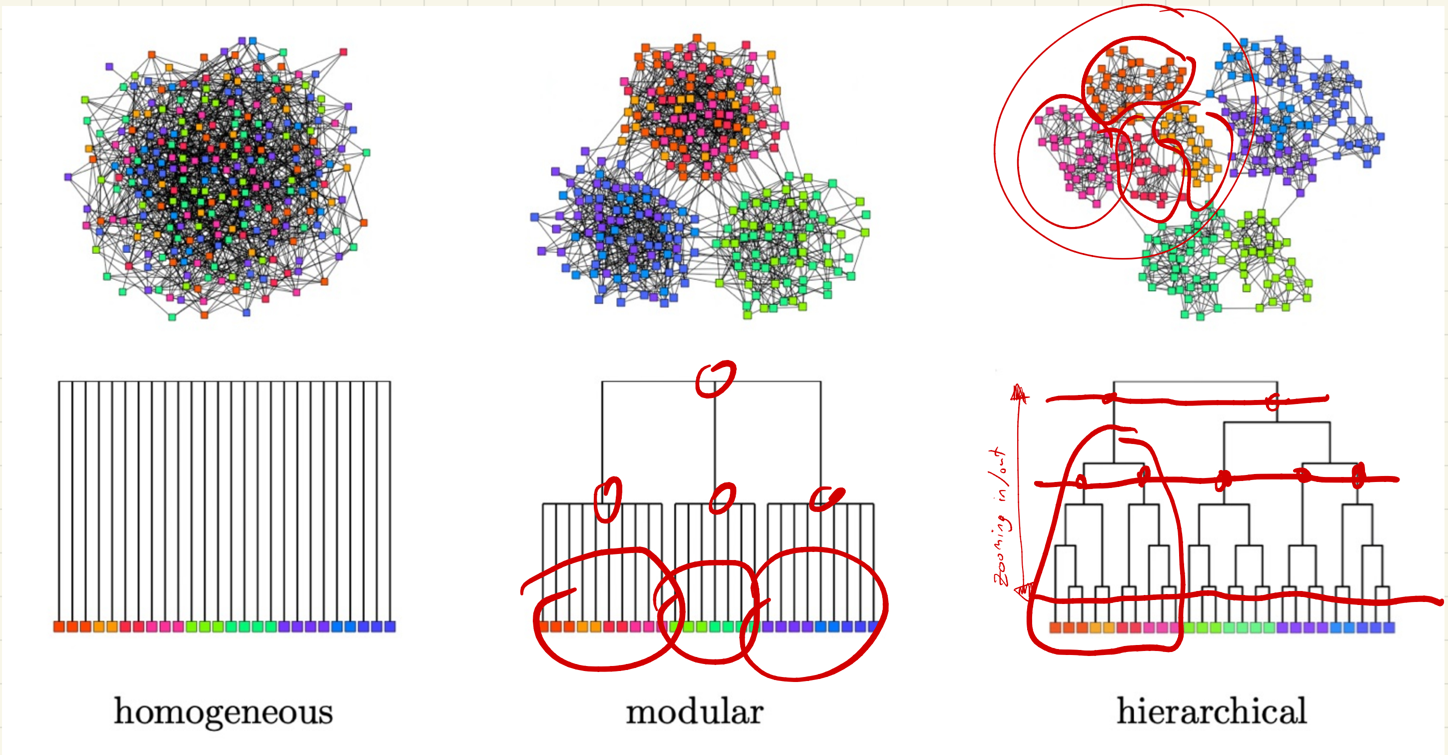
political blogs (2004)



social contact patterns (2017)



Why just one scale?



network

dendrogram

1 lvl of org

many lvls (nested)

Random graphs with modular structure

introduced by Holland, Laskey, & Leinhardt in 1983

introducing... the stochastic block model (SBM)

$$\delta_{ab} = \begin{cases} 1 & a=b \\ 0 & - \end{cases}$$

defined by $\underline{\theta} = (c, \vec{z}, M)$ and $n_r = \sum_{i=1}^{n} \underline{\delta_{z_i r}}$

c: # groups

$\vec{z}$: partition of nodes $z_i \in \{1, 2, \ldots c\}$

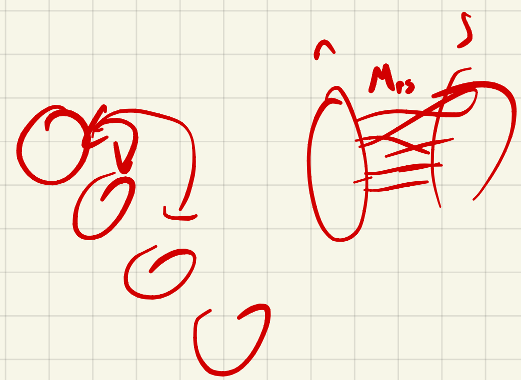
M: mixing matrix

n_r: # of nodes in group r

given $\underline{\theta}$, $\forall_{i>j}$ $\overbrace{A_{ij} = A_{ji}}^{\text{undirected}} = \begin{cases} 1 & \text{with probability } M_{\underline{z_i}\,\underline{z_j}} \\ 0 & \text{otherwise} \end{cases}$

* "block Erdős" network

- when $r=s$, we have $G(n_r, M_{rr})$
- when $r \neq s$, we have random bipartite graph with density M_{rs}



properties of SBM graphs

- simple graph
- connectivity is modular by design, M
- $\Pr(k) \approx$ Poisson (M_{rs}) mixture
- diameter & MGD → $O(\ln n)$ depends on M
- $C = O(1/n)$
- LCC = $O(n)$ if G is not too sparse

* key point *

edges are conditionally independent

↓

"stochastic equivalence"

the SBM is a transliteration of

a **community** *is a group of nodes that connect to other groups in similar ways.*

def of a "group" r is M_{r*}

generating a SBM network

given $\Theta = (c, \vec{z}, M)$

1) initialize $G = (V, \phi)$

2) flip a coin for each $i,j \in \binom{n}{2}$ pair

if Heads $(M_{z_i z_j})$

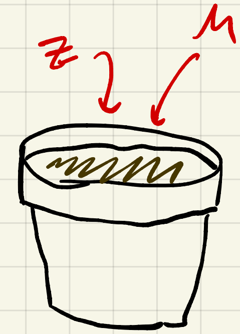
3) place an edge $(i,j) \to E$

$\binom{n}{2}$

$\Pr(i \to j \mid M, z) = M_{z_i, z_j}$

note: $\Theta(c^2)$ parameters = lots of flexibility

the planted partition (aka simplifying the SBM) ~~$++n+\binom{c}{2}$~~ . |



2. we specify the partition $\vec{z}$ and the mixing matrix M but simplified

planted-c: couple all off-diagonal $M_{r \neq s}$ AND equalize group sizes $\forall_i \; n_i = \frac{n}{c}$

$$M^{\text{planted-c}}_{\text{assort}} = \begin{pmatrix} p_1 & & q \\ & p_2 & \\ & \ddots & \\ & & p_c \end{pmatrix}$$

$\left\{ \begin{array}{l} c \text{ parameters for within-group densities} \\ 1 \text{ parameter for between-group densities} \end{array} \right.$

planted-2 : set $c=2$ AND couple within-group densities AND fix $\langle k \rangle = d$

$$M^{\text{planted-2}}_{\text{assort}} = \begin{pmatrix} p & q \\ & p \end{pmatrix}$$ and $n_1 = n_2 = n/2$

implies p and q not independent

→ we can reduce planted-2 to a one-parameter model (useful!)

define $p = d_{in}/n$
$q = d_{out}/n$

s.t. $d_{in}, d_{out} = 2\times$ within / between group degrees

algebra ↓

$2d = d_{in} + d_{out}$

define $\boxed{\varepsilon} = d_{in} - d_{out}$ ⇝ $d_{in} = d + \varepsilon/2$
$d_{out} = d - \varepsilon/2$ ⇝ $p = (d + \varepsilon/2)/n$
$q = (d - \varepsilon/2)/n$

note $\varepsilon \in [0, 2d]$

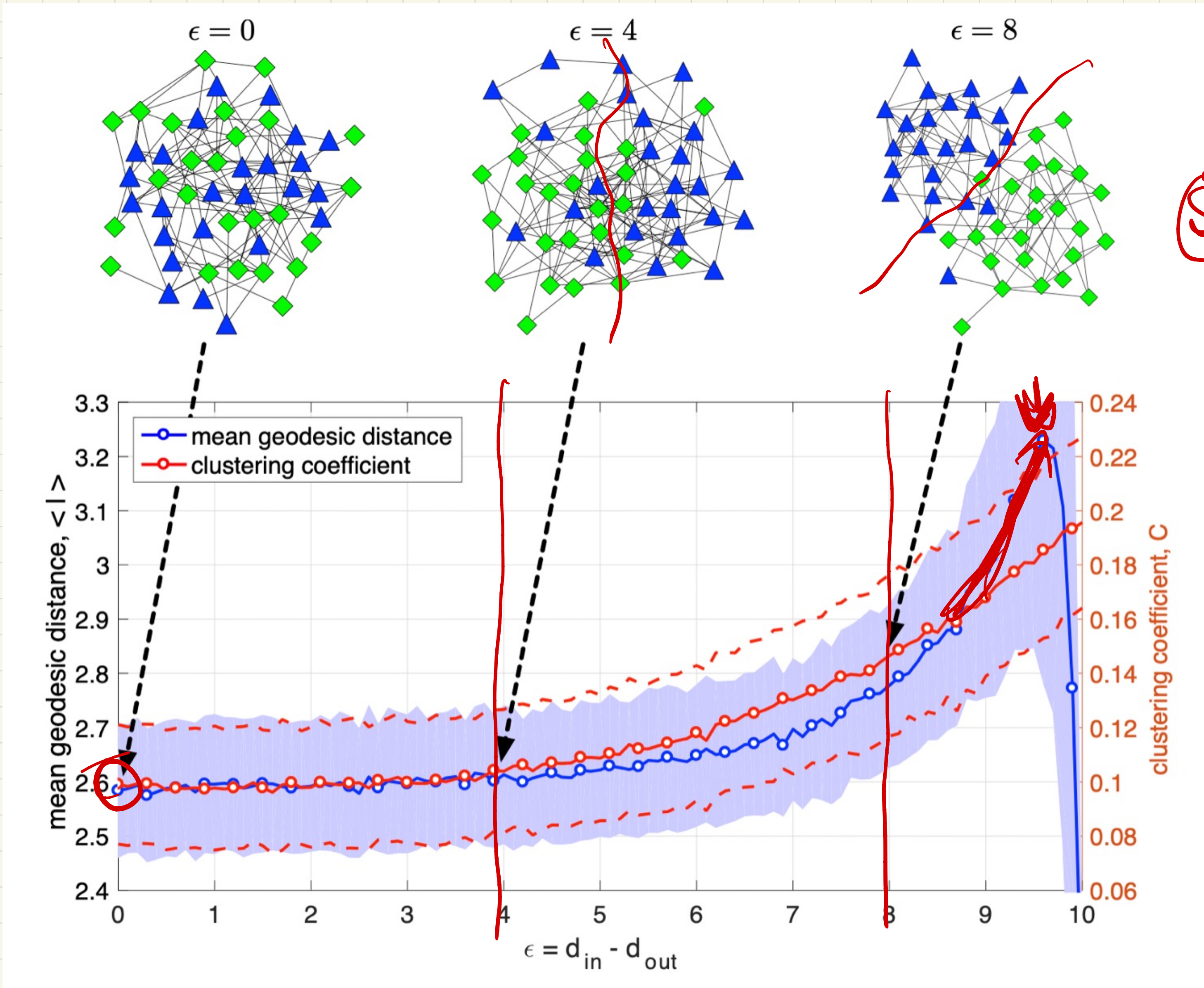
→ what does varying ε do to pattern of SBM? what happens if $\varepsilon < 0$

Cool! let's vary ε and measure G's shape

$n=50$

$\langle k \rangle = d = 5$

- $\varepsilon < 4$
- $4 < \varepsilon < 6$
- $\varepsilon > 8$



the degree-corrected SBM

introduced by Karrer and Newman in 2010

* the SBM struggles with heterogeneous Pr(k)
  → solution: the DC-SBM is like the Chung-Lu model + SBM

$\Theta = (c, \vec{z}, \vec{k}, M)$

SBM : ER ↴
DCSBM : Chung-Lu

- c: # of groups
- $\vec{z}$: assignment of nodes to groups

$$\begin{bmatrix} 1 \\ 1 \\ 2 \\ 2 \\ 3 \\ 1 \\ 4 \\ 1 \\ 2 \end{bmatrix}$$

- $\vec{k}$: degree structure
- M: Counts # of edges between groups r, s OR if r=s, 2x that number

$$M_{rs} = \sum_{ij} A_{ij} \underbrace{\delta_{z_i r} \delta_{z_j s}}_{=1 \text{ if } z_i = r \text{ AND if } z_j = s}$$

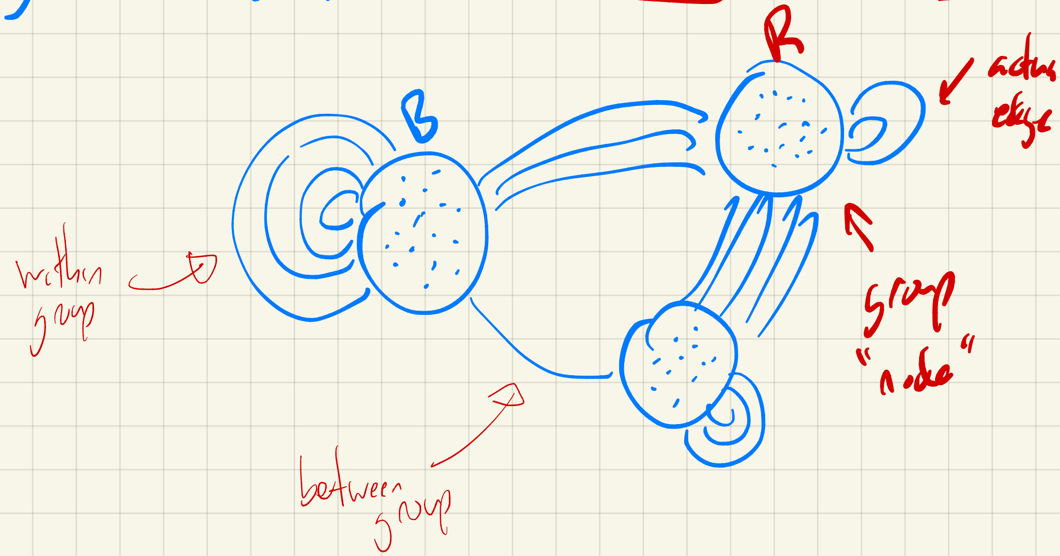
warning this def is diff. than SBM → DC-SBM: edge count; SBM: edge density

$\rightarrow \sum_{rs} M_{rs} = 2m$

* the DC-SBM is the most flexible random graph model in our toolbox

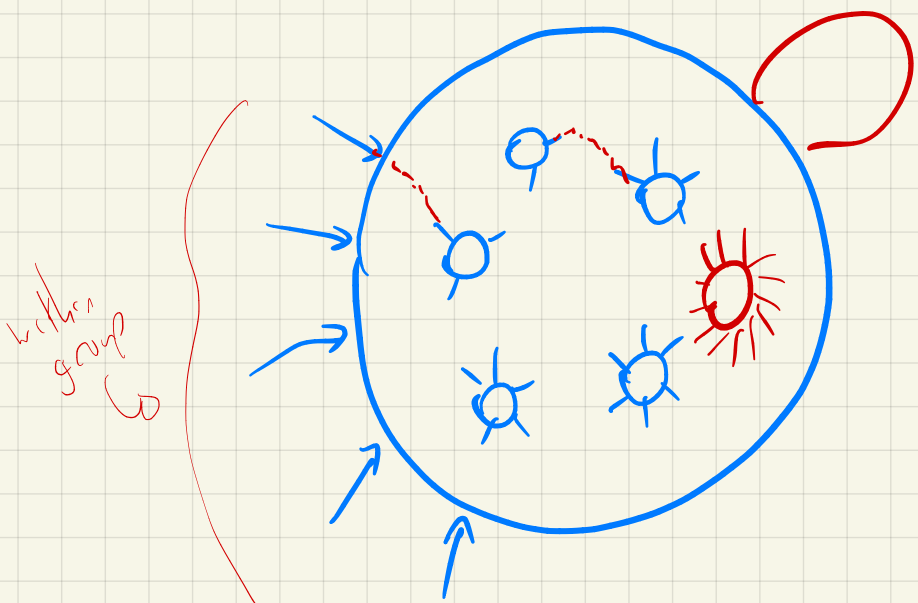
it's a 2-level model:

1) first, groups act like nodes in a group-level loopy multigraph model



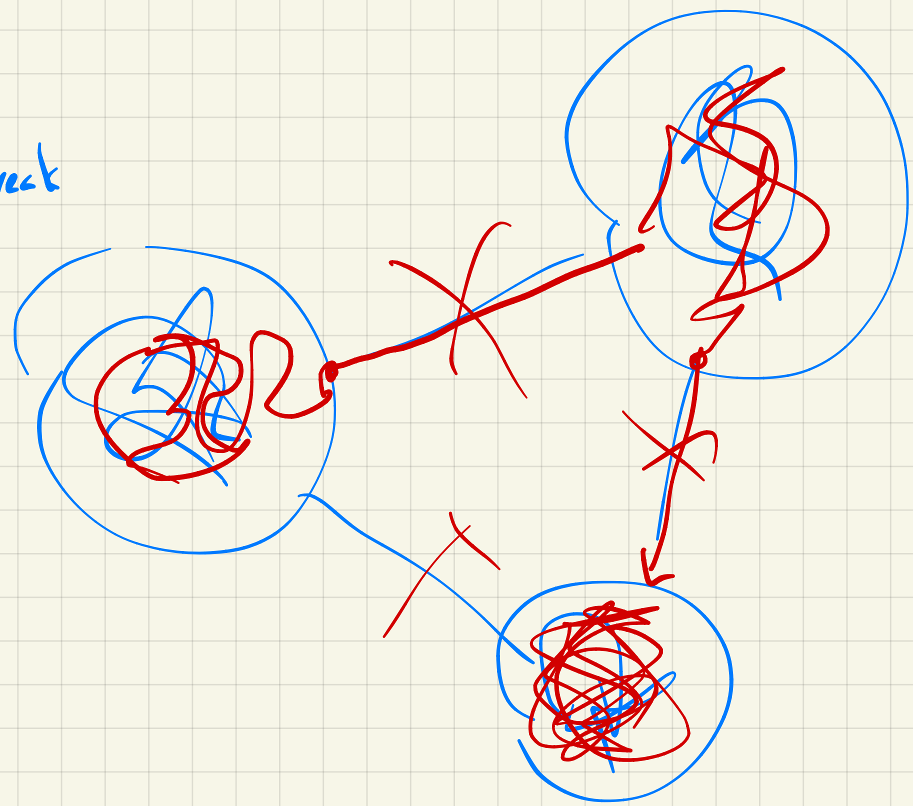
M sets between group connectivity

2) second, make connections from group to its members



Properties of the DC-SBM

- undirected multigraph
- modular by design, c groups, $\mathcal{M}$ mixing matrix
- $\Pr(k)$ $\vec{k}$ degree structure (specified)
- diameter & MGD $\to O(\ln n)$, unless $\mathcal{M}$ is weird
- $C \to O(1/n) \to$ locally tree-like
- LCC $\to O(n)$ unless $\mathcal{M}$ very weak



generating a DC-SBM network

level I

• a group level degree

$$\kappa_r = \sum_s m_{rs} = \sum_i k_i \overbrace{\delta_{z_i r}}$$

* sum of all the degrees of all nodes in group r

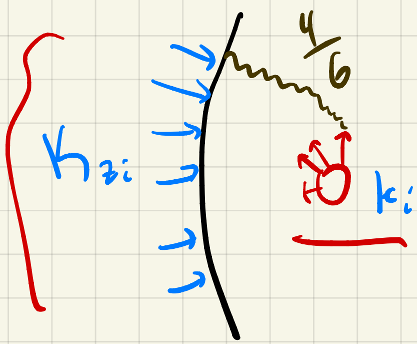
• $\vec{\kappa}$: the group level degree sequence

level 2

• node "propensity"

$$\gamma_i = \frac{k_i}{\kappa_{z_i}}$$

probability a connection to group r "lands" on node i



$$\gamma_i \in (0,1)$$

$$k_i \le \kappa_{z_i}$$

• $$\forall_{i>j} \quad A_{ij} = A_{ji} \sim \text{Poisson}\left(\overbrace{\gamma_i \gamma_j}^{k_i k_j} \overbrace{M_{z_i z_j}}\right)$$

generating a DC-SBM network

given $\theta = (c, \vec{z}, \vec{k}, M)$ $\quad M \rightarrow c \times c$

* how long does this take?

$\Theta(c^2)$

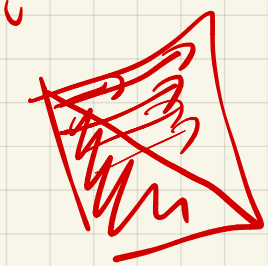
1) compute the $\kappa_r = \sum_s M_{rs}$ (group degrees)

$\Theta(n)$

2) compute the $\gamma_i = k_i / \kappa_{z_i}$ (node propensities)

3) initialize empty G with n nodes

$1 \le i < j \le n$



$\Theta(n^2)$

4) for each pair $i > j$, $r \sim \text{Poisson}(\gamma_i \gamma_j M_{z_i z_j})$

5) if $r > 0$, add (i,j) to G

enforces a simple graph

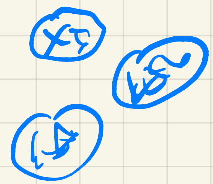
$A_{ij} = r$

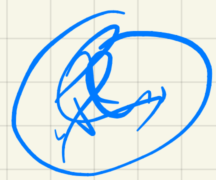
Planted partitions in the DCSBM

$$\overbrace{n_r}^{\text{equal sized groups}} = \frac{n}{c}$$

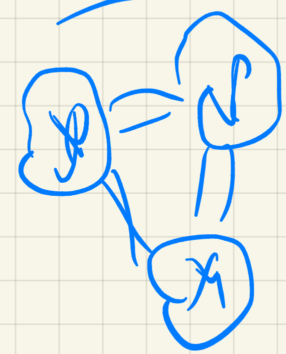
mixing parameter $\lambda \in [0,1]$

$$M_{rs} = \lambda \, \overbrace{M_{rs}^{assort}} + (1-\lambda)\, \overbrace{M_{rs}^{random}}$$

$\underline{\lambda = 1}$: $M_{rs} = M_{rs}^{assort}$ $\rightsquigarrow$ 

$\underline{\lambda = 0}$: $M_{rs} = M_{rs}^{random}$ $\rightsquigarrow$ 

modules!

$\lambda = \frac{1}{2}$: 

$$M_{rs}^{random} = k_r k_s / 2m \quad \forall rs$$

} same "count" in each M_{rs} cell

$$M_{rs}^{\text{assort}} = \begin{pmatrix} K_1 & & 0 \\ & K_2 & \\ & & \ddots \\ & & & K_c \end{pmatrix}$$

} c distinct ~~ER~~ Chung-Lu random graphs

all that's left is to specify the degree structure $\vec{k}$. ≥ 0 e.g. $\Pr(k) \sim \text{Exp}(\lambda)$

$\sim \text{Poisson}(\lambda)$

etc.

* one constraint: every $k_i \leq n_r$ WHY?

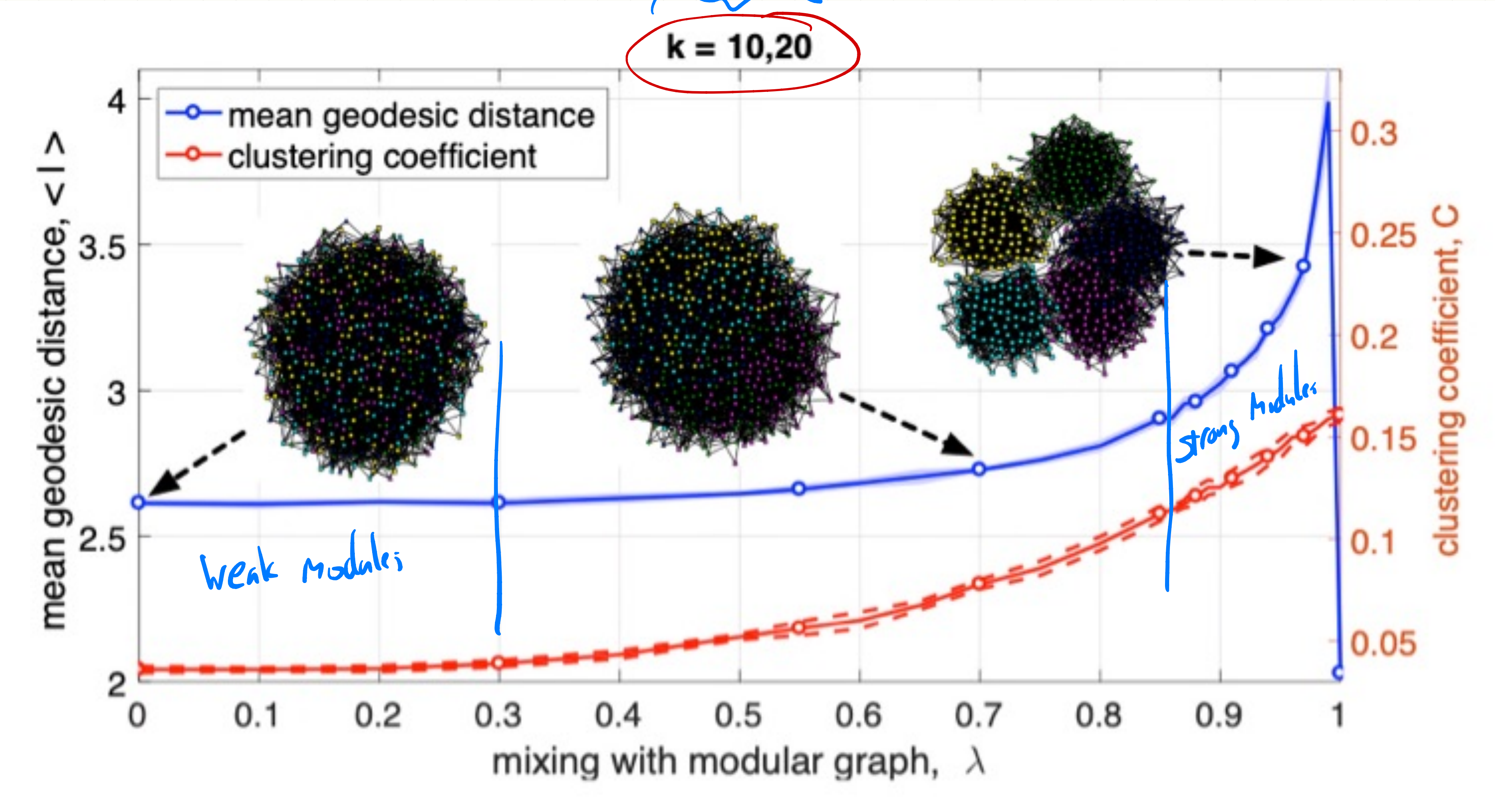
lets explore the shape of DC-SBM networks

$c=5$ groups $n=500$ nodes $n_r=100$

$\langle k \rangle = 15$

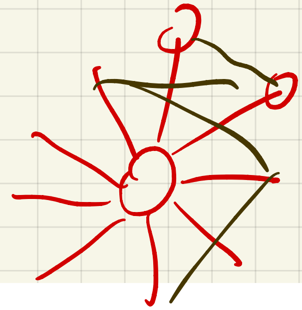
two values of k_i

low variance in $\vec{k}$

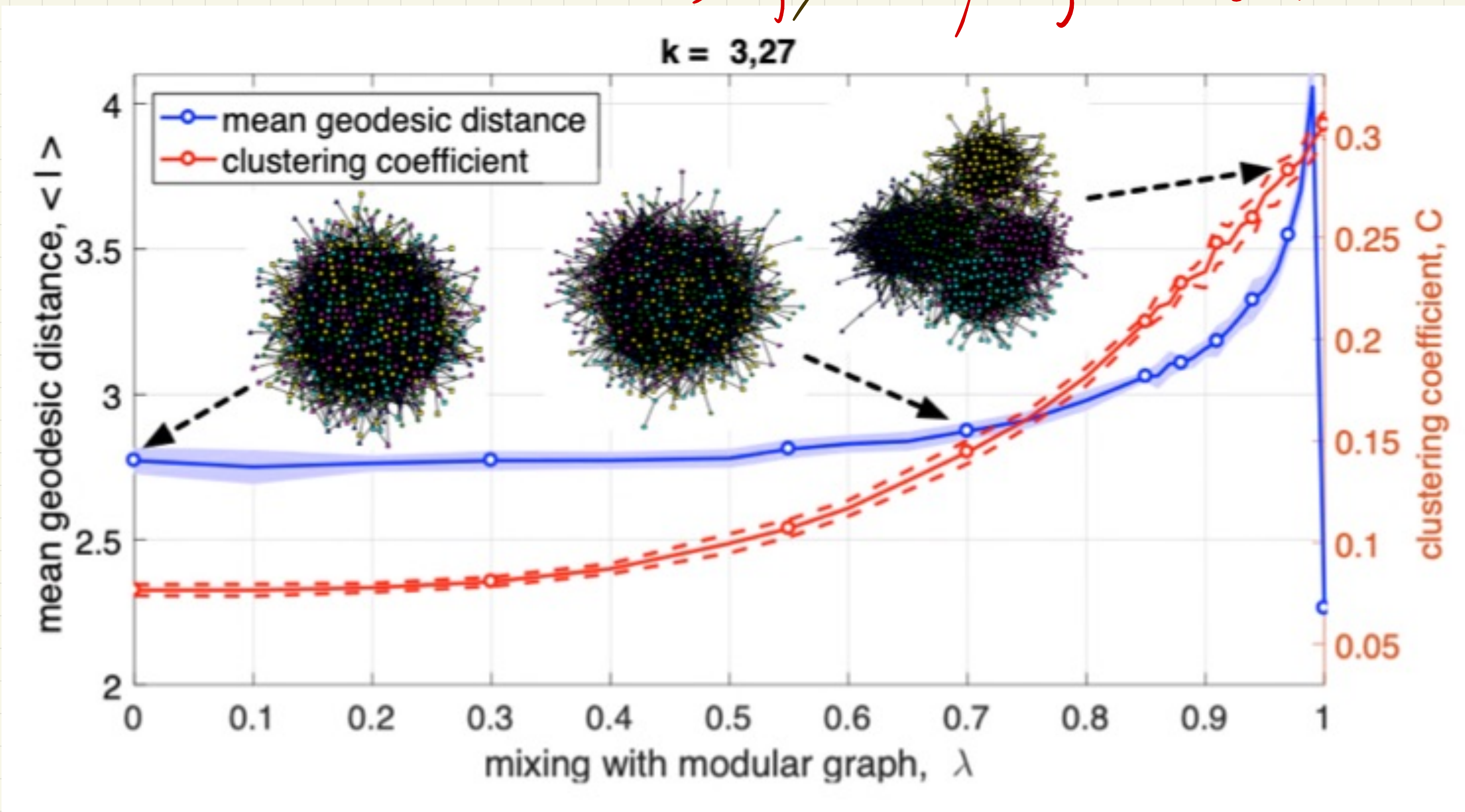


$c=5$

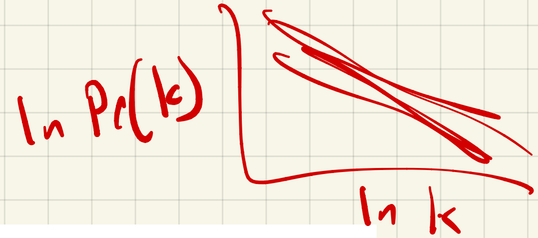
$n=500$ $\langle k \rangle = 15$

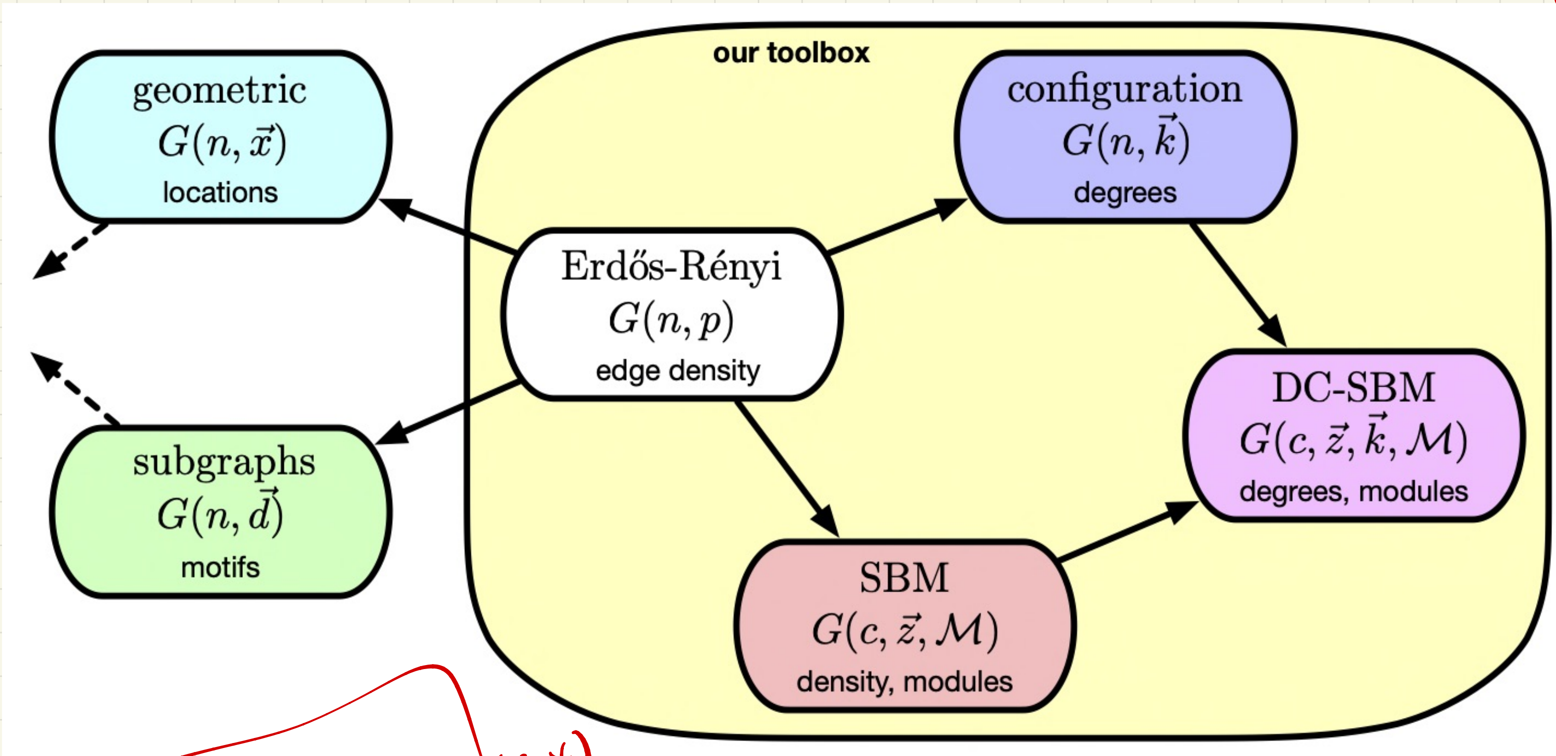


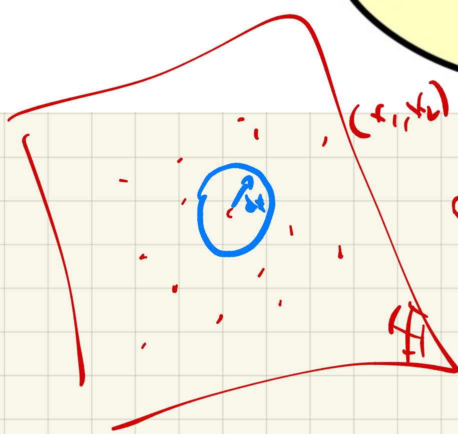
high variance in $\vec{k}$

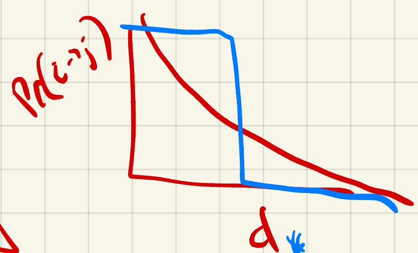


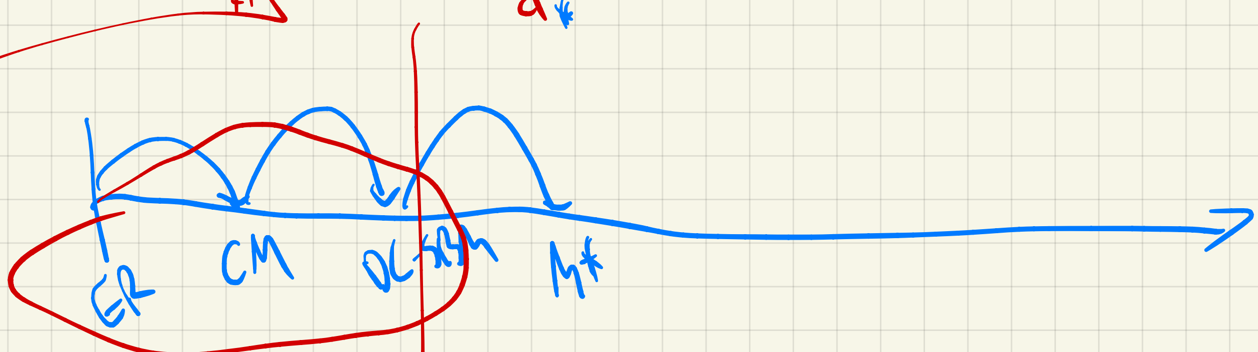
taking stock of random graph models











measure	pattern	edge density	degrees	modules, density	modules, degrees
	empirical	*Erdős-Rényi*	*configuration*	*SBM*	*DC-SBM*
edge density ρ	sparse	specified	specified ✓	specified ✓	specified ✓
degrees $\Pr(k)$	heavy tailed	~~Poisson~~	specified ✓	Poisson mixture ✗	specified ✓
diameter $\ell_{\max}$	$O(\log n)$	$O(\log n)$	$O(\log n)$ ✓	typically $O(\log n)$ ✓	typically $O(\log n)$ ✓
clustering C	social: many non-social: few	$O(1/n)$	$O(1/n)$ ✗	depends on $\mathcal{M}$ ✓	depends on $\mathcal{M}$ ✓
reciprocity r	high	$O(1/n)$	$O(1/n)$ ✗	depends on $\mathcal{M}$ ✓	depends on $\mathcal{M}$ ✓
large component	$\Theta(n)$	if $\langle k \rangle > 1$	if $\langle k^2 \rangle - 2\langle k \rangle > 0$	depends on $\mathcal{M}$	depends on $\mathcal{M}$ ✓