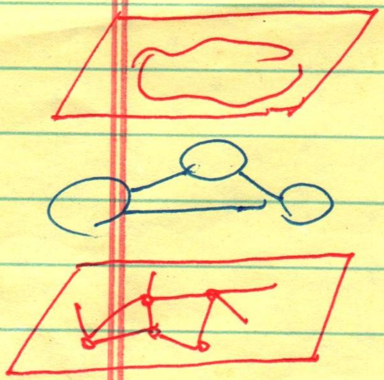
①

Lecture 7 a Feb 27, 2024

modular networks, structure.



global ↖ $\langle k \rangle$, LCC, clustering coefficient, etc.

← pattern of edges

local ← k_i, h_i (harmonic centrality), local clustering c_i, etc.

module = group = compartment = community

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

(statistical notion!)

Examples:

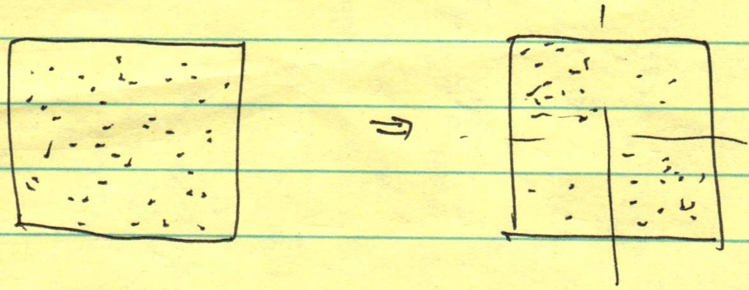
(1)

(2)

(3)

②

no modules   Modules

$A =$ 

assortative modules

⇒ finding the modules = finding a "coarse graining" of the system

⇒ empirically, # groups $= O(\sqrt{m})$

$= O(\sqrt{n})$ ⇒ bigger #'s, more communities.

How to represent modular interactions ?

» Use mixing matrices ! $[0,1]^{c \times c}$

$M = (M_{rs})$ $r \neq s$ density of connections from group $r \to$ group s.

$r = s$ density of communites inside group r.

» Idea: use M_{rs} to generate edges between groups r & s.

$$\Pr\left(i \to j \mid M_{z_i z_j} \right)$$

$$z_i \in \{1, 2, \ldots, c\}.$$

» What happens if $\forall\, rs,\ M_{rs} = p$?

⇒ Erdős-Rényi $G(n,p)$.

See handout for various mixing patterns.

Highlight :

- "block Erdős model"
- planted-partition model $\quad n_i = n_j \ \forall i,j \in \{1, 2, \dots, c\}$.
- disassortative ⇒ bipartite patterns.
- Structure of M tells the large-scale patterns
  - # groups set the scale of the pattern.
  - more groups = smaller pattern.
- when $r \neq s$, we are in directed #'s regime → richer patterns.

Examples :

- political blogs (2004) $\quad c=2 \Rightarrow c=2\times 2$ (seeing ~~core~~ phenipley)
- Social contact patterns (2017) "linear hierarchies" "within 5 years"
- hierarchies at multiple scales.

④

Assortative coefficient (r, Q)

network-level measure

>> how much often do attributes match accross edges than expected at random ?

⇒ we use the configuration model as the null model, i.e., $P_{ij} \propto k_i k_j$

1) categorical

2) scalar

1)
$$Q = \frac{1}{2m} \sum_{ij} \Big(\underbrace{A_{ij}}_{\substack{\text{observed} \\ \text{empirical network}}} - \underbrace{\frac{k_i k_j}{2m}}_{\substack{\text{expected} \\ \text{null model}}} \Big) \delta(x_i, x_j)$$

$$\delta(a,b) = \begin{cases} 1 & \text{if } a=b \\ 0 & \text{otherwise.} \end{cases}$$

(H) (n^2) cost

We can obtain "modularity" by taking the assortative coefficient at the group-level. (as summation over groups)

You will need :

$$e_{uv} = \frac{1}{2m} \sum_{ij} A_{ij}\, \delta(x_i, u)\, \delta(x_j, v).$$

fraction of stubs on edges from group u → group v

⇓

" mixing matrix "

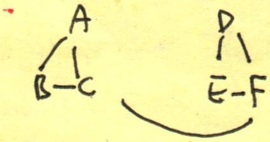
$$a_u = \frac{1}{2m} \sum_j k_j\, \delta(x_j, u)$$

"degree" of group u

5 ~ Half

-2

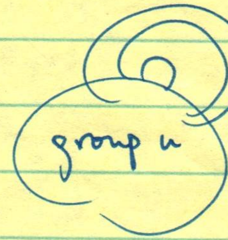
$\frac{1}{14}$ Lecture 7a / 7b.



$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \sum_u \delta(x_i, u)\, \delta(x_j, u)$$

[see lecture notes]

$$Q = \sum_u \left(e_{uu} - a_u^2 \right)$$



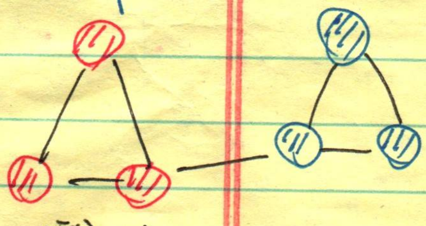
← self-loops in null model

↓

internal edge in a network.

$n=6$, $m=7$, $k=2$

Examples:



partition 1

$2m = 14$

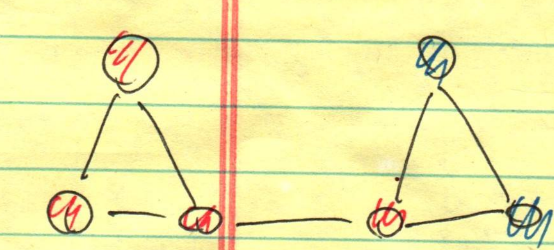
	red	blue
red	3/7	1/14
blue	1/14	3/7

sum counts

edge counts

$Q_1 = \frac{5}{14} = 0.357$

$$\cancel{Q = \underbrace{\frac{\#\text{ edges within modules}}{\#\text{ total edges}}}_{m} - \frac{\left(\sum e_{uu}\right)^2}{(2m)^2}}$$



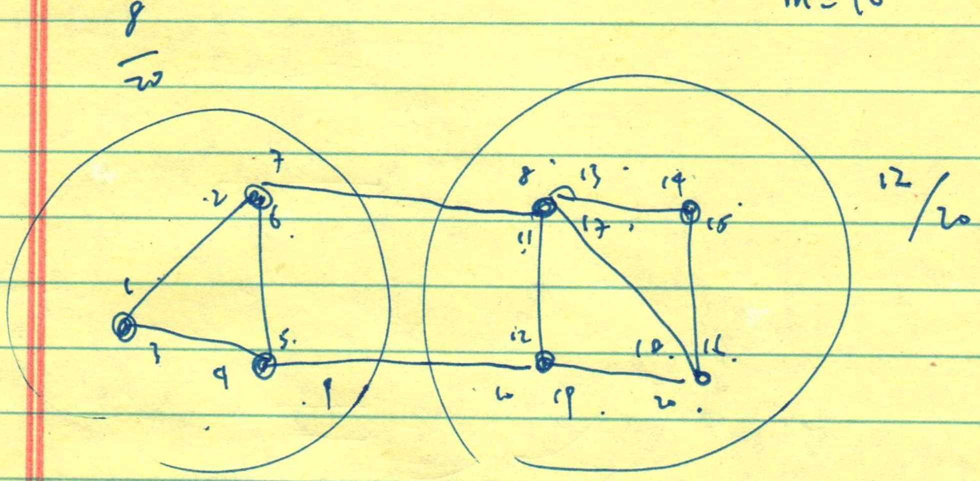
partition 2

	r	b
r	4/7	2/14
b	2/14	1/7

$Q = \frac{6}{49} = 0.122$.

(-1)

$n = 7$
$m = 10$



$$Q = \sum_u \left(e_{uu} - a_u^2 \right)$$

$$= \underbrace{\frac{3}{10}}_{e_{11}} + \underbrace{\frac{1}{2}}_{e_{22}} - \big(\underbrace{\tfrac{2}{5}}_{a_1}\big)^2 - \big(\underbrace{\tfrac{3}{5}}_{a_2}\big)^2 = \frac{7}{25}$$

~~a_u = frac edges join nodes of type u (stubs)~~

e_u : frac edges joining node of type u

a_u : stubs

① ①

Lecture 7 b

– Review the two examples for computing modularity.

Scalar

» How much more similar are attributes across edges, than expected at random?

» adapt covariance to network.

$$u = \frac{1}{2m} \sum_i k_i x_i$$

$$\mathrm{cov}(\{x\}, A) = \frac{\sum_{ij} A_{ij}(x_i - u)(x_j - u)}{\sum_{ij} A_{ij}}$$

(i.e., $2m$)

$$= \frac{1}{2m} \sum_{ij} \left(A_{ij} x_i x_j - \mu^2 \right)$$

$$= \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) x_i x_j$$

assortativity

$$r = \frac{\mathrm{cov}(\{x\}, A)}{\mathrm{var}(\{x\}, A)} = \frac{\sum_{ij} \left(A_{ij} - {}^{k_i k_j}/_{2m} \right) x_i x_j}{\sum_{ij} \left(k_i \delta(i,j) - {}^{k_i k_j}/_{2m} \right) x_i x_j}$$

or called ...

Pearson correlation

4 2

Holland, Laskey, Leinhardt (1983)

Random graphs w/ modular structure.

"SBM" $\Theta = (c, z, M)$ → mixing matrix.

groups, partition of nodes into groups.

and $z_i = \{1, 2, ..., c\}$

$n_r = \sum_{i=1} \delta_{z_i, r}$

→ # nodes in group r.

given Θ, $\forall i > j$ $A_{ij} = A_{ji} = \begin{cases} 1 & \text{w/ prob. } M_{z_i z_j} \\ 0 & \text{otherwise.} \end{cases}$

"block Erdős" random graph

③

Properties of SBM graphs

- simple / non-simple.
- connectivity is modular via M.
- degree distri $P(k)$ is a mixture of Poisson.
- diameter $\langle l \rangle \approx O(\log n)$
- $C = O(1/n)$, depends on M.
- LCC tend to $O(n)$ if G is large and $\langle k \rangle > 1$.

driven by locally-tree-like structure of ER graph.

* key point* edges are conditionally independent.
"Stochastic equivalence"
» assumes that the model is consistent under sampling
» see Shalizi & Rinaldo (2013) or ERGMs.

Generating a SBM network

given $(c, \{z\}, M)$

i) initialize an empty graph $G = (V, E = \phi)$
ii) for each $\binom{n}{2}$ pair, draw $r_{ij} \sim U(0,1)$
iii) if $r_{ij} \leq M_{z_i z_j}$, add (i,j) to E.

§ The planted partition (simplified version of the SBM)

1) "planted-c"

2) "planted-2" (when $c=2$)

1) $\Rightarrow$ specify $\{z\}$ and M, but simplified $\forall_r \; n_r = \frac{1}{c} n$.

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

$\begin{cases} c \text{ parameters for internal densities.} \\ 1 \text{ parameter for external densities.} \end{cases}$

2) $\Rightarrow$ $M^{\text{planted-}2}_{\text{assort}} = \begin{pmatrix} p & q \\ q & p \end{pmatrix}$, and fix $\langle k \rangle = d$

$n_1 = n_2 = n/2$

$\Rightarrow$ we can reduce planted-2 to a one-parameter model.
(total edges fixed)

$\Rightarrow$ $p \equiv d_{in}/n$, $q = d_{out}/n$

$\Rightarrow$ $2d = d_{in} + d_{out}$ $\qquad \Rightarrow \; p = (d + \epsilon/2)/n$

$\Rightarrow$ define $\epsilon = d_{in} - d_{out}$ $\qquad q = (d - \epsilon/2)/n$

$\epsilon \in [0, 2d]$ (note : $\epsilon \to 0 \Rightarrow$ ER)

5

See notes for one example (of $n=50$, $d=5$) varying ϵ.

* The degree-corrected SBM. Karrer & Newman (2011)

SBM struggles with heterogeneous $P(k)$
$\rightarrow$ dc-SBM $\Leftarrow$ Chung-Lu + SBM.

(Give an example phenomenon — for the problem.)

$\theta = (c, \{z\}, \{k\}, M)$ $\rightarrow$ conceptually:

SBM = ER + ~~graph~~ groups

DC-SBM = Chung-Lu + groups.

SBM $\leftarrow$ Bernoulli edges

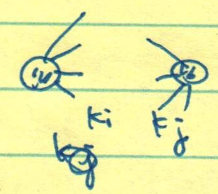
but DC-SBM $\leftarrow$ Poisson edges.

$$M_{rs} = \sum_{ij} A_{ij}\, \delta_{z_i r}\, \delta_{z_j s}.$$

$$A = \begin{pmatrix} & & \boxed{ij} \\ & & \\ & & \end{pmatrix}$$

SBM $\Rightarrow$ DC-SBM

$f(M_{rs})$ $\quad$ $f(M_{rs}, k_i, k_j)$

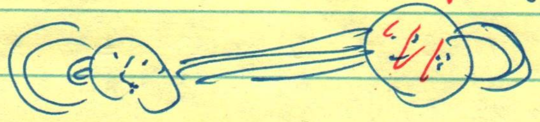


" most flexible random graph model " in our toolbox.

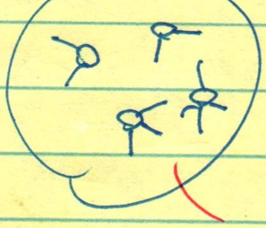
~~Abo~~ Remarks:

1) It's a two-level model

i) first, groups acts like nodes in a group-level loopy, multigraph model.



ii) second, make connections from group to it members.



total degree of groups

$= \sum_r M_{rs}$

may still generate ~~multiedges/self-loops~~

but very few if $n \to \infty$.

2) Properties.

- undirected loopy multigraph
- modular by design.
- degree structure is specified by $\{k\}$.
- diameter, mean geod. dist. $\langle l \rangle \sim O(\log n)$ ∴ b/c random graph.
- $C \approx O(1/n)$ ∴ M can change this
- LCC $\approx O(n)$ ∴ if G is not too sparse.

7

3) Generating DC-SBM network.

given $\theta = (c, z, k, M)$

i) compute $k_r = \sum_s M_{rs}$ (group degrees)

ii) $r_i = k_i / k_{z_i}$ (node propensities)

iii) Initialize empty G with n nodes.

iiii) for each pair $i > j$, $r \sim \text{Poisson}(r_i, r_j, M_{z_i z_j})$

v) if $r > \phi$, add edge (i, j) to G.

(?)

Final words: Peixoto (2013)

1. Highlight: canonical vs. microcanonical.

$z, k, M \rightarrow b, k, e.$

2. microcanonical:

$$\mathbb{P}(b, k, e) = \Omega(b, k, e)^{-1}$$

3. $\mathbb{P}(A, k, e, b) = \mathbb{P}(A \mid k, e, b)\underbrace{\mathbb{P}(k \mid e, b)\, \mathbb{P}(e \mid b)\, \mathbb{P}(b)}_{\text{priors}}$

(Bayesian hierarchy).

4. Next lectures: Inference!