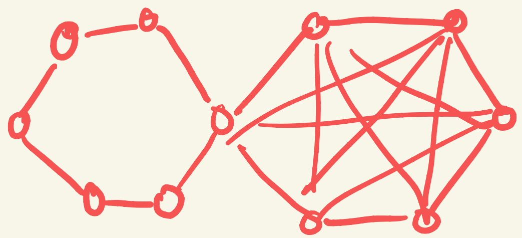
9/21/2020
Week 5

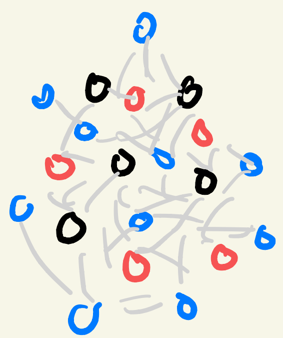
- HW2 grades: returned
- Q+A in office hrs!
- HW4 due Fri.



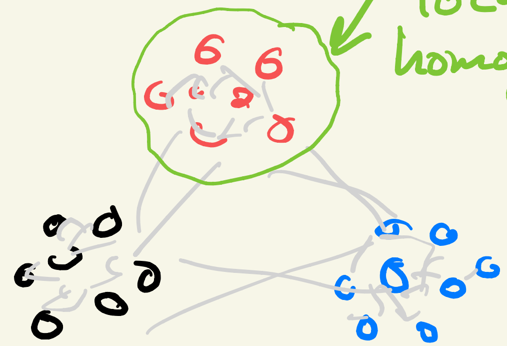
Can we go beyond pix?
Can we quantify heterogeneity?

Large-scale organizational patterns / structures.

"Mixing" patterns.



homogeneous
all regions net. equivalent



!!! locally homogeneous

heterogeneous
pattern or property depends on where you look!

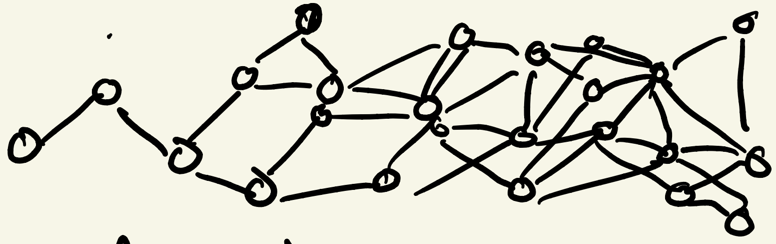
① Assortative
"like links w/ like" clustering.

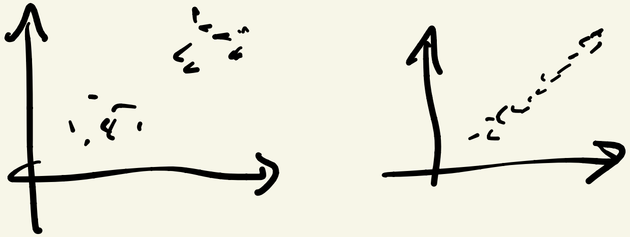
② Disassortative
"like links w/ unlike" bipartite – perfectly disassortative

Mixing on: degree, glasses, computer router switch,
plumage color, ..., unobserved attribute!

Global heterogeneity
↓ caused by
Local homogeneity,
w/ multiple communities.

"modular" or "community" structure.





All our models so far are "globally homogeneous." In the sense that all of our p_{ij} "coin flips" were independent — we never consider node group or anything like that!

Modular
Core Periphery
Ordered

3 kinds of ways to be locally homogeneous, but globally heterogeneous

Homophily + Assort. Mixing

pattern $\not\Rightarrow$ mechanism of pattern formation

Big Question: Radicalization?
Obesity ?

Mechanism Important!

Two types of attributes

① "Enumerative" i.e. "Categorical"
glasses vs. non-glasses wearing.

② Scalar height, degree

①. Let $x_i \in S$ denote the category of node i.

[How much more often do attributes match across edges than expected at random?]

model!
stochastic.

How should we model this?

Imagine:

fraction q vertices are blue,
$1-q$ are red.

If the network is E.R.,

How many blue-to-blue edges, in expectation?

E.R.: m edges, in expectation.

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

$\boxed{m q^2}$ blue to blue

$\boxed{m(1-q)^2}$ red to red

If empirical counts are higher than these $\Rightarrow$ assortative!

If some nodes in real network have high k $\longrightarrow$ our E.R. null model will be bad.

Instead, use the CM.

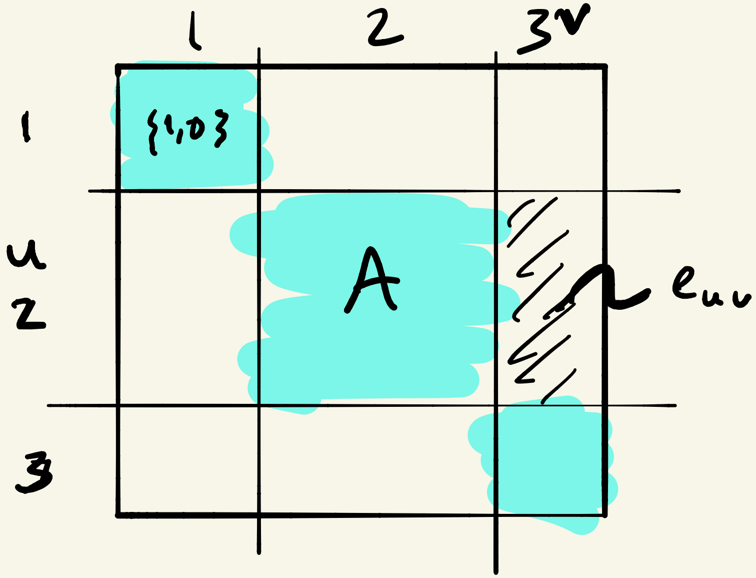
only add when i and j have same label!

$$Q = \frac{1}{2m} \sum_{ij} \Big(\underbrace{A_{ij}}_{\text{actual } i \leftrightarrow j} - \underbrace{\frac{k_i k_j}{2m}}_{\text{expected } i \leftrightarrow j} \Big) \delta(x_i, x_j)$$

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

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

only sum over vertices of same type!



Simplify?

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

= fraction of degrees between u, v.

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

= fraction of degrees attached to group u.

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

$$= \sum_u \left[\underbrace{\frac{1}{2m}\sum_{ij} A_{ij}\,\delta(x_i, u)\,\delta(x_j, u)}_{e_{uu}} - \underbrace{\frac{1}{2m}\sum_i k_i\,\delta(x_i, u)}_{a_u}\;\underbrace{\frac{1}{2m}\sum_j k_j\,\delta(x_j, u)}_{a_u}\right]$$

$$\boxed{= \sum_u e_{uu} - a_u^2}$$

Maj. Illusion:

Mean Neighbor Property:

$$\langle x_v \rangle = \frac{1}{2m} \sum_u \sum_v x_v A_{uv}$$

$$= \frac{1}{n\langle k \rangle} \sum_v k_v x_v$$

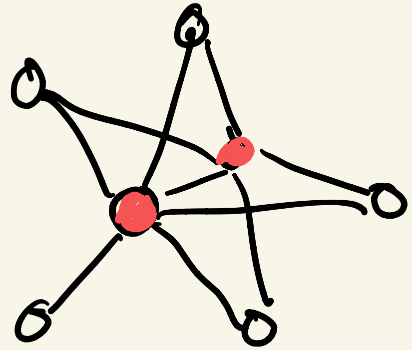
$$= \frac{\frac{1}{n} \sum_v k_v x_v}{\langle k \rangle}$$

If x_v uncorrelated w/ k_v

$$= \frac{E[x_v]\, \cancel{\frac{1}{n} \sum k_v}}{\cancel{\langle k \rangle}} = q$$

If x_v correlated w/ k_v

$\Rightarrow$ maj. illusion.



When do we see the <u>friendship paradox?</u>

$$\text{MND} > \langle k \rangle$$

$$\frac{\langle k^2 \rangle}{\langle k \rangle} > \langle k \rangle$$

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

$$\frac{1}{n}\left(k_1^2 + k_2^2 + k_3^2 + \ldots\right) > \frac{1}{n^2}\left(k_1 + k_2 + k_3 + \ldots\right)^2$$

LHS = RHS <u>only</u> when $k_i = \bar{k}$ $\forall i$

$\Rightarrow$ degree-regular network. ←

In general,

$$\langle x^2 \rangle \geq \langle x \rangle^2$$

with = when $\vec{x} = \bar{x}\,\vec{1}$

$$\frac{1}{n}\left[\left(k_1^2 + k_2^2 + k_3^2 + \ldots\right) - \frac{1}{n}\left(k_1 + k_2 + k_3 + \ldots\right)^2\right]$$

↓

variance

$$E[X^2] - E[X]^2$$

<u>uniform!</u>

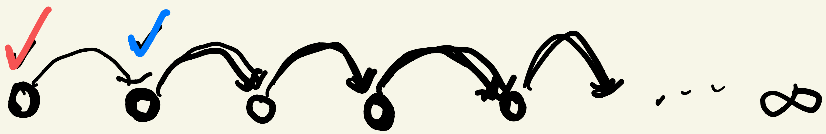
HW2

Induction.

Formulas, for natural numbers.

Show # nodes d steps from the center is $k(k-1)^{d-1}$ $\forall d \geq 1$

"n_d" natural numbers.

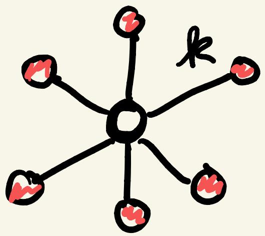


d=1 d=2

① d=1

② d=2

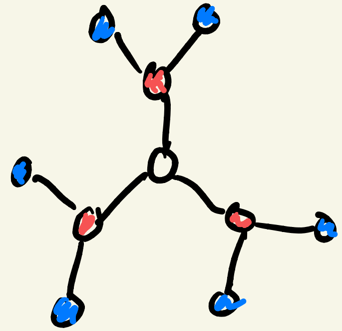
③ Assume d=n formula is true. ⟶ show d=n+1 formula true.



① $d=1 \rightarrow n_d = k$

formula: $k\cancel{(k-1)^{1-1}} = k$

② $d=2$



$k(k-1)$

formula: $k(k-1)^{2-1} = k(k-1)$

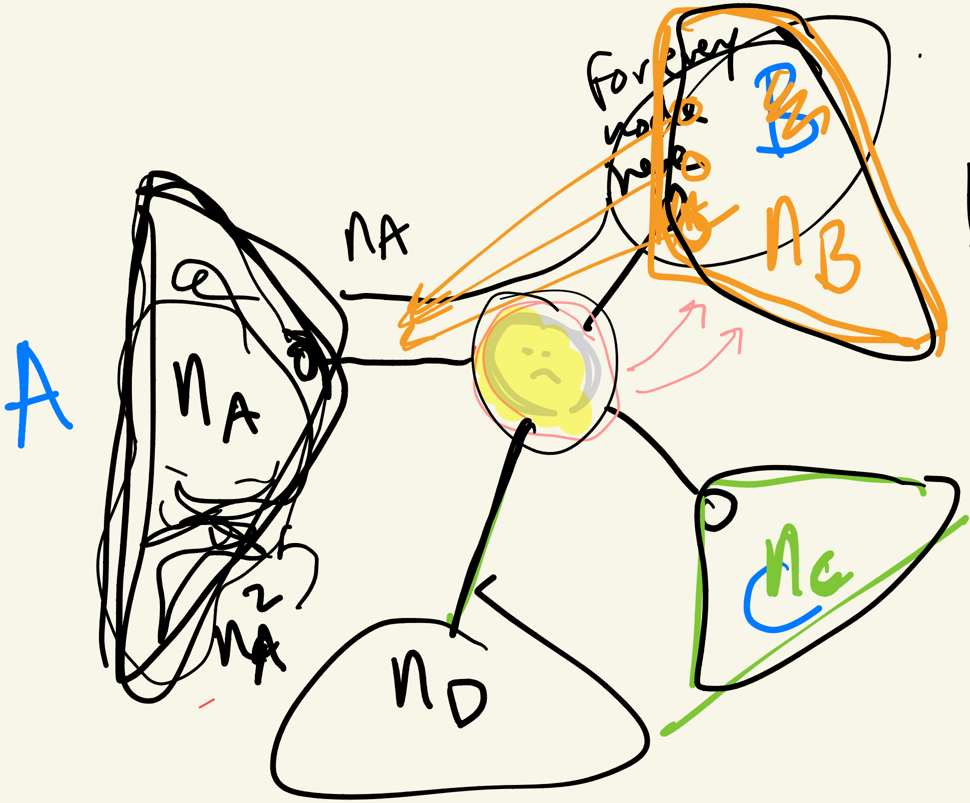
③ assume I have, at layer n, exactly $k(k-1)^{n-1}$ nodes.

For each node in layer n, we get $k-1$ nodes in layer n+1. (def'n of c.T.)

$$\Rightarrow n+1: \; k(k-1)^{n-1} \times (k-1)$$

$$= k(k-1)^n$$

formula: $k(k-1)^{n+1\cancel{-1}}$



$$b = 7^2 - 4^2 - 1^2 - 1^2$$
$$= 49 - 16 - 1 - 1$$
$$= 49 - 18$$
$$= 31.$$

$$n_1, n_2, \dots n_k$$

$$M_{ii} = 0$$

$$M_{ij} = n_i n_j = M_{ji}$$

$$\begin{array}{c|ccc|c} & A & B & C & \\ \hline A & 0 & n_A n_B & n_A n_C & n_A \\ B & n_A n_B & 0 & n_B n_C & n_B \\ C & n_A n_C & n_B n_C & 0 & n_C \\ \hline & n_A & n_B & n_C & 1 \end{array} \quad M$$

$$b = 2 \sum_{i<j}^{k,k} n_i n_j + 1 + 2 \sum_i^k n_i$$

$$\hat{b} = n^2 - \sum_{i=1}^{k} n_i^2$$

$$b = n^2 - \sum_{i=1}^{k} n_i^2$$

possible shortest, directed paths in the network.

n_i^2 = # possible shortest paths in each subtree.

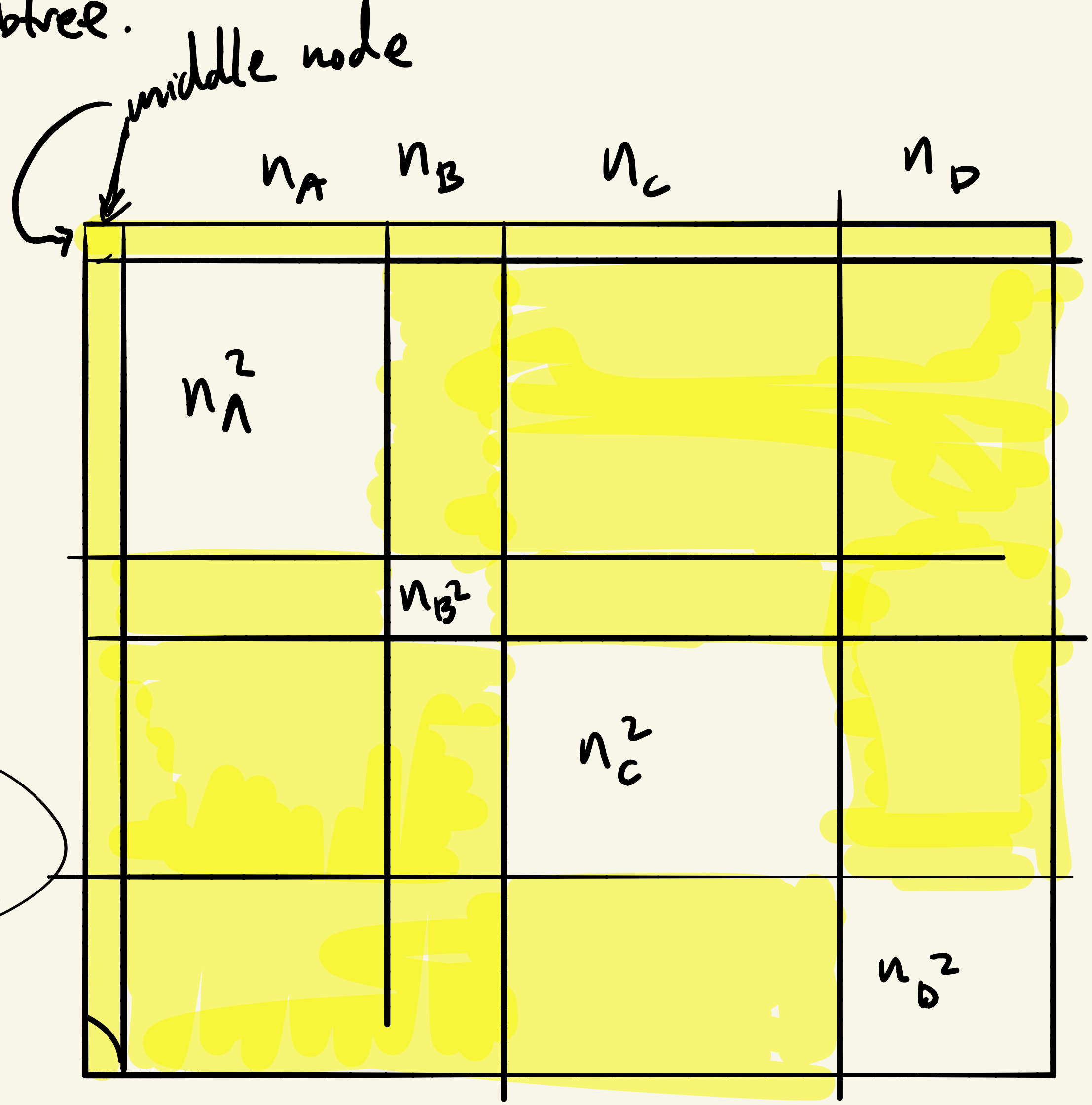
$$\sum_{i=1}^{k} \sum_{j \neq i} n_i n_j + 2\sum_i n_i + 1$$

$$+\sum_i n_i n_i - \sum_i n_i n_i$$

$$= \sum_i \sum_j n_i n_j - \sum_i n_i^2 + 2\sum_i n_i + 1$$

$$= \left(\sum_i n_i\right)\left(\sum_j n_j\right) - \sum_i n_i^2 + 2\sum_i n_i + 1$$

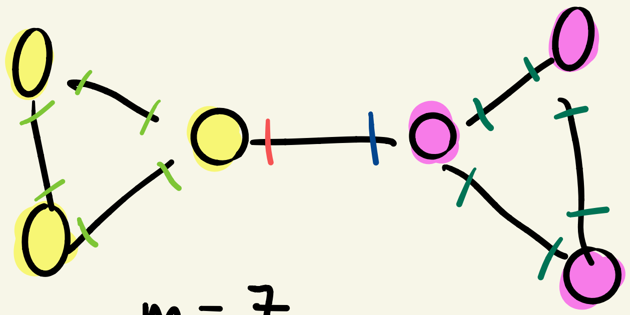
$$= n \cdot n - \sum_i n_i^2$$



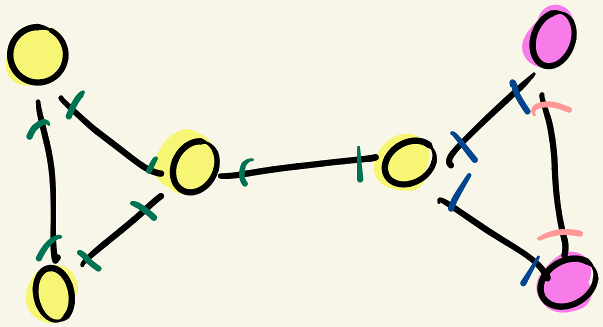
9/23/2020

$$\boxed{Q = \sum_{u} e_{uu} - a_u^2}$$

u are the groups.
Here, there are two: yellow + pink.



higher



lower

$m = 7$

$2m = \sum_i k_i = 14$

$$e \begin{bmatrix} \frac{6}{14} & \frac{1}{14} \\ \frac{1}{14} & \frac{6}{14} \end{bmatrix} \rightarrow \overset{a}{\begin{bmatrix} \frac{7}{14} \\ \frac{7}{14} \end{bmatrix}}$$

$$\frac{6}{14} - \left(\frac{7}{14}\right)^2 + \frac{6}{14} - \left(\frac{7}{14}\right)^2 = 0.357$$

$$e \begin{bmatrix} \frac{8}{14} & \frac{2}{14} \\ \frac{2}{14} & \frac{2}{14} \end{bmatrix} \rightarrow \overset{a}{\begin{bmatrix} \frac{10}{14} \\ \frac{4}{14} \end{bmatrix}}$$

$$\frac{8}{14} - \left(\frac{10}{14}\right)^2 + \frac{2}{14} - \left(\frac{4}{14}\right)^2 = 0.122$$

What about scalar correlations?

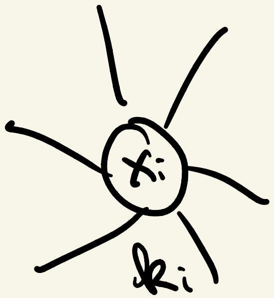
Let x_i be a scalar vertex attribute.

$x \in \mathbb{R}$. assortativity coefficient.

Network version of Pearson corr.

How much more similar are $\{x_i\}$ across edges than exp. at random?

mean: $\mu = \frac{1}{2m} \sum_i k_i x_i$



$\mathrm{cov}(\vec{x}, A)$

$$= \frac{\sum_{ij} A_{ij}(x_i - \mu)(x_j - \mu)}{\sum_{ij} A_{ij}}$$

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

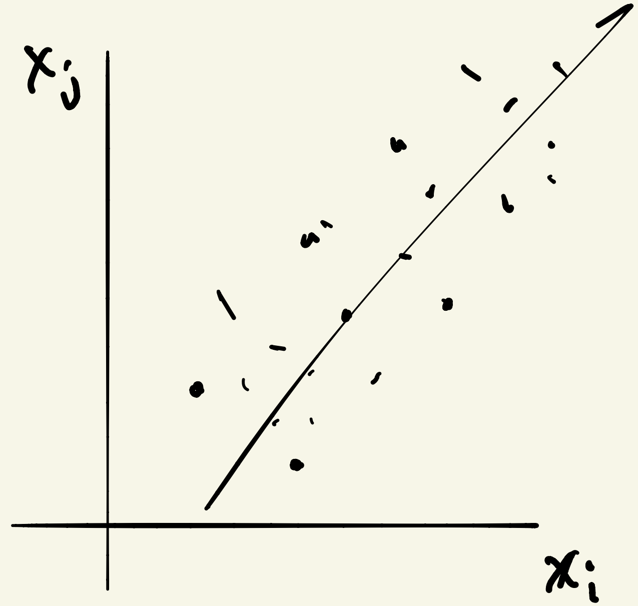
$$= \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m}\right) x_i x_j$$

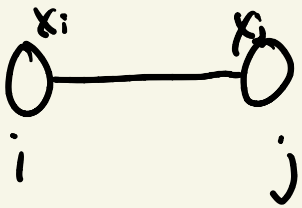
~~not normalized.~~

could be degrees!

Normalize:

$$\boxed{r = \frac{\sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m}\right) x_i x_j}{\sum_{ij} \left(k_i \delta(i,j) - \frac{k_i k_j}{2m}\right) x_i x_j}}$$





Why does work introduced Q have ~40K cites???!
(Kirvan, Newman)

Community Detection.

What labeling would max Q?

Goal: find labels to max Q.

assortative structure only!

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

doesn't depend on "partition" i.e. division into groups, or labels

$$\max_{\vec{x}} Q$$

[How many ways are there to divide n things into groups?]

Stirling Numbers.

n objects

divide into k nonempty groups.

$$S(n,k) = \frac{1}{k!}\sum_{j=0}^{k}(-1)^j\binom{k}{j}(k-j)^n$$

$S(4,2) = 7$

1/234 2/134 4/123 3/124

12/34 13/24 14/23

$$\sum_{k=1}^{n} S(n,k) = B_n$$

Bell number. $\uparrow$ super exponentially.

$n=1$	1
$n=2$	2
$n=3$	5
$n=4$	15
$n=5$	52
⋮	
$n=20$	5,832,742,205,057

$\left[\text{Max } Q \text{ is NP-hard.}\right]$

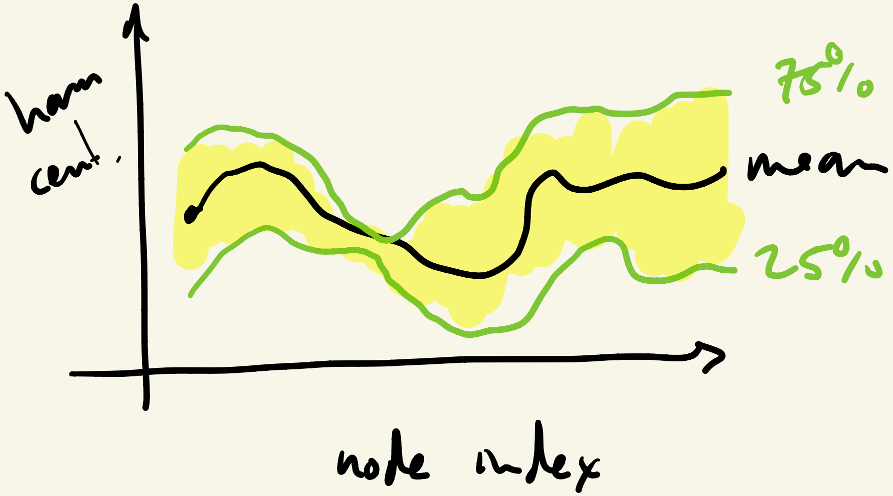
Search Problem

Algorithms. Heuristics.

Not unique to modularity Q.

Applies to all $\max_x f(A, x)$

adj mtx partition

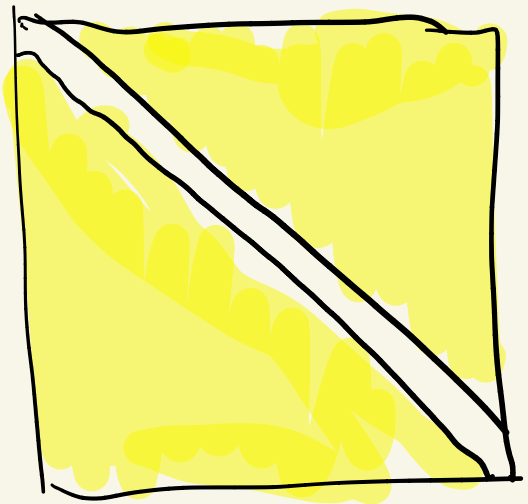


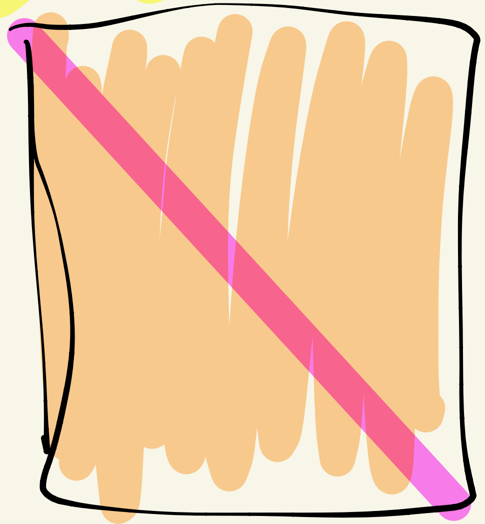
$$\left(\sum_i \sum_{j \neq i} n_i n_j + \sum_i n_i^2\right) - \sum_i n_i^2$$

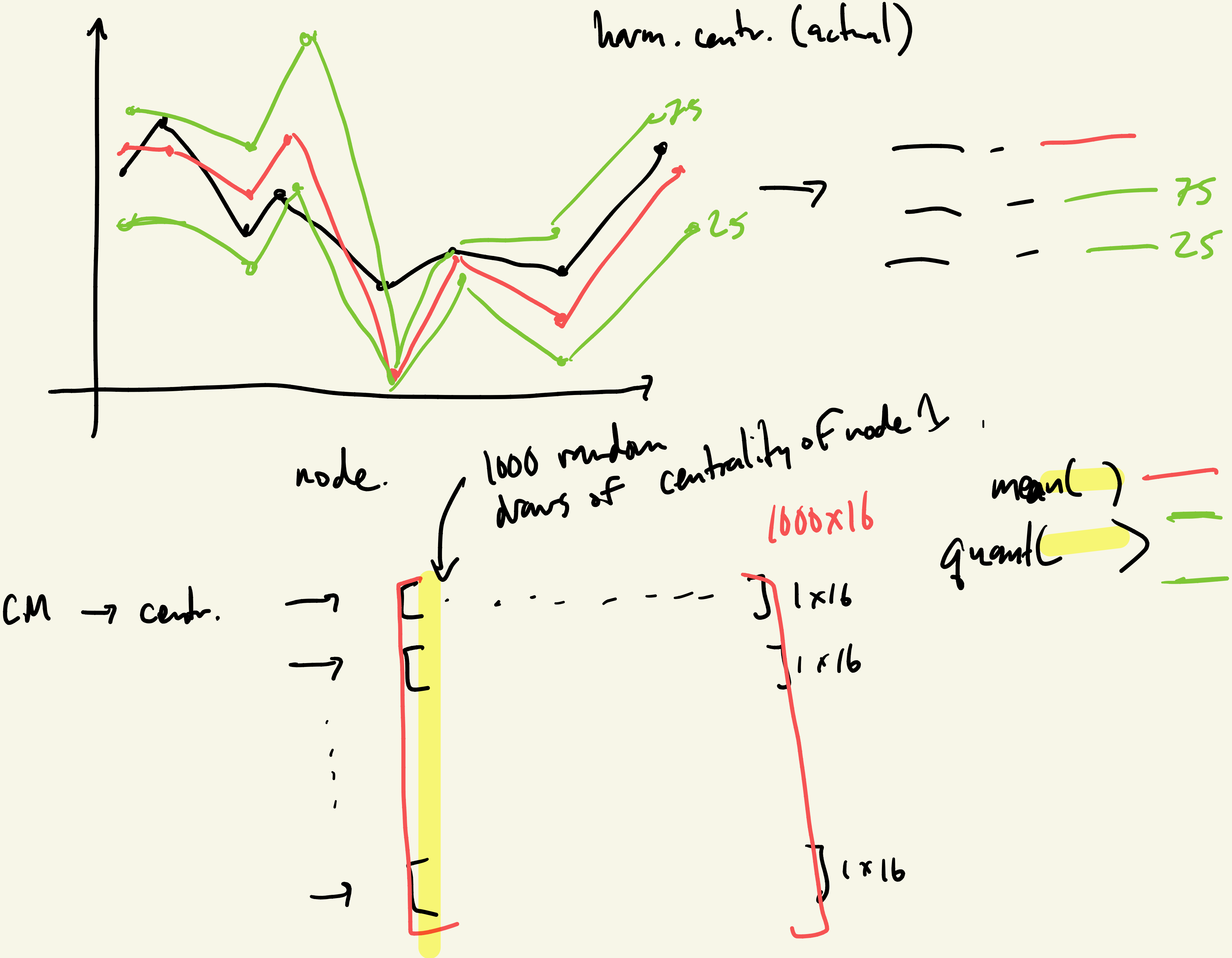
$$\sum_i \sum_j n_i n_j - \sum_i n_i^2$$

$$\sum_i \sum_{\substack{j \\ j \neq i}} n_i n_j = \sum_i \sum_j n_i n_j - \overbrace{\sum_i \sum_{\substack{j \\ j=i}} n_i n_j}^{\sum_i n_i^2}$$

$$\sum_i \sum_j n_i n_j - \sum_i n_i^2$$



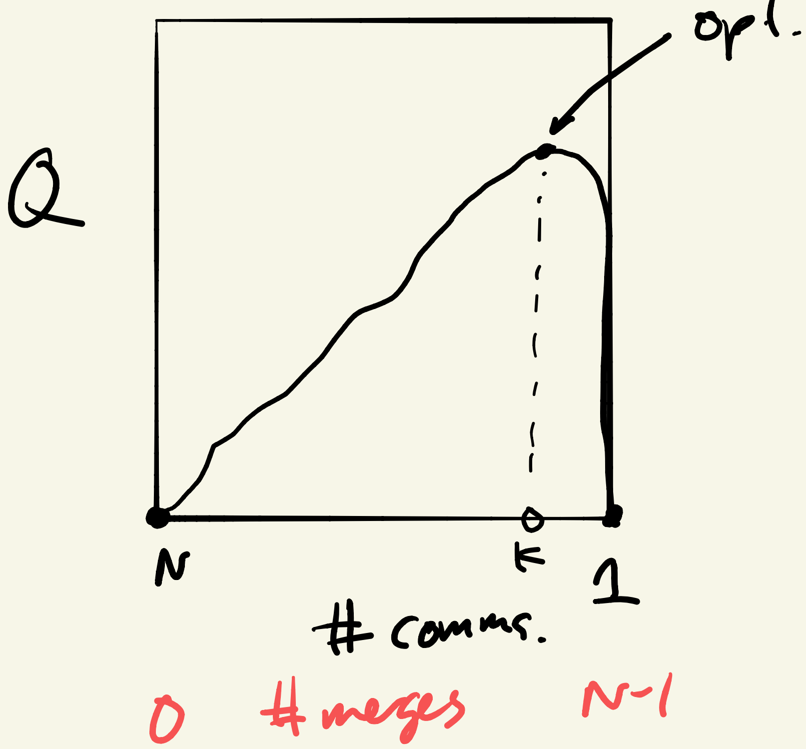



harm. centr. (actual)
75
25
75
25
node.
1000 random draws of centrality of node 1
1000x16
mean( )
quant( )
CM → centr.
1x16
1x16
1x16

9/25/2020

Searching space of partitions.

Goal: find division into groups to maximize $Q(A, \vec{x})$.



Greedy Agglomerative Approach

① Initialize: each node is its own group.

② Try all pairs of hypothetical merges, and choose the merge which most increases (or least decreases) Q. Do the merge.

compute Q for each hypothetical merge.

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

$$= \sum_u e_{uu} - a_u^2$$

③ Repeat ② until all nodes are in one group.

④ Choose the partition that max Q overall.

Speedup:

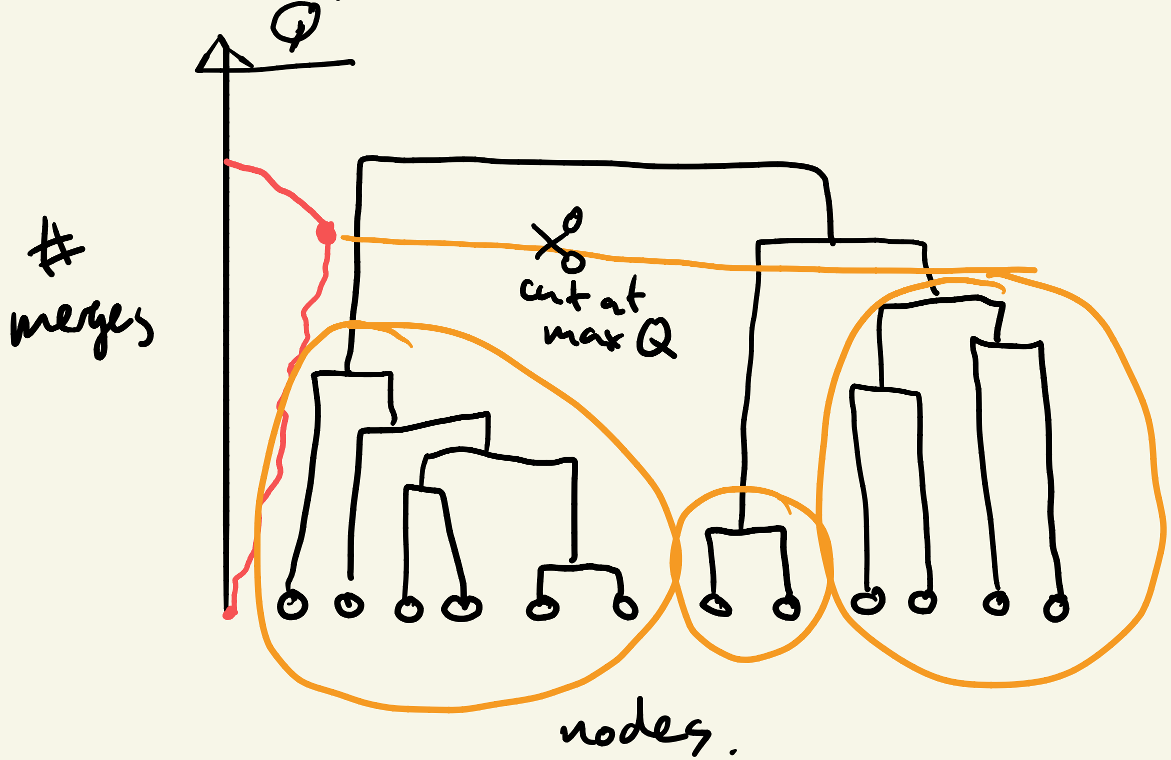
By how much would Q change, if we merged group u, and group v?

① Merging 2 comms that have no edges between $\Rightarrow$ impossible to increase Q.

② $$\Delta Q_{uv} = e_{uv} + e_{vu} - 2a_u a_v = 2(\underbrace{e_{uv}}_{\#\text{ edges } u\leftrightarrow v} - \underbrace{a_u a_v}_{\text{expected \# edges.}})$$

fun! derive this!

Greedy Approach Builds a tree.



Can you prove that greedy merge is concave in Q?
If so $\rightarrow$ +20 EC!