L25: Community Detection in Graphs

Trinity Consultants

We are seeking resumes of candidates for a position on our data management/software development team. The position will have a strong emphasis in meteorological and air quality data collection, interpretation, and validation. The ideal candidate will have a strong drive and interest in computer programming and software development as well as and understanding of meteorological and air quality instrumentation.

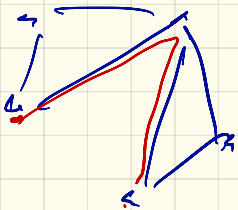
Resumes can be sent to either
Scott Adamson (sadamson@trinityconsultants.com) or
Casey Lenhart (clenhart@trinityconsultants.com).

Data $x \in X$

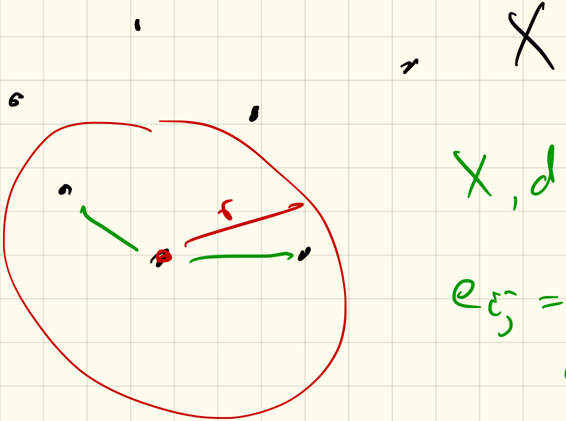
Graph $G = (V, E)$

distance : $X \times X \to \mathbb{R}$

similarity : $X \times X \to \mathbb{R}$



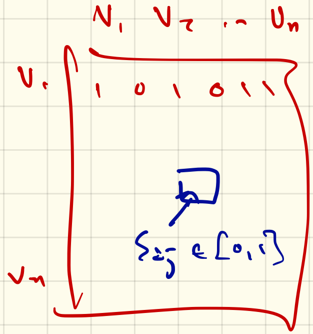
$V = X$



$X, d \to E$

$e_{ij} = 1$ iff $d(x_i, x_j) < r$

$M \in \mathbb{R}^{|V| \times |V|}$



Graphs arising in Social Networks

V = users, people

E: E_{ij} exists if interaction between v_i, v_j

follows on Twitter,
or mentions
or " in last 1 month

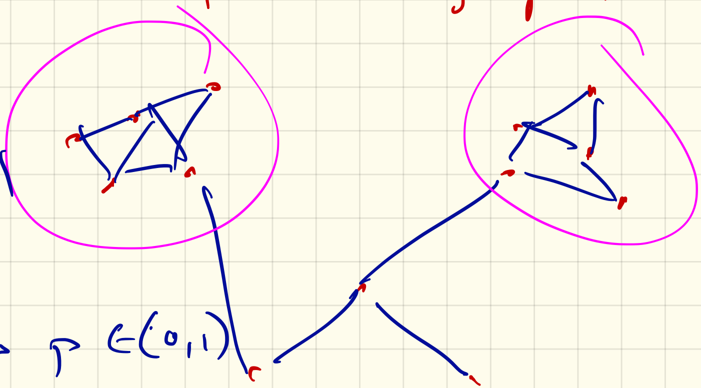
friends on Facebook

Enron emails

in 1990s : Sociology

Graphs $|V| = 100$ friends in high school

- usually draw graph "planar" graphs



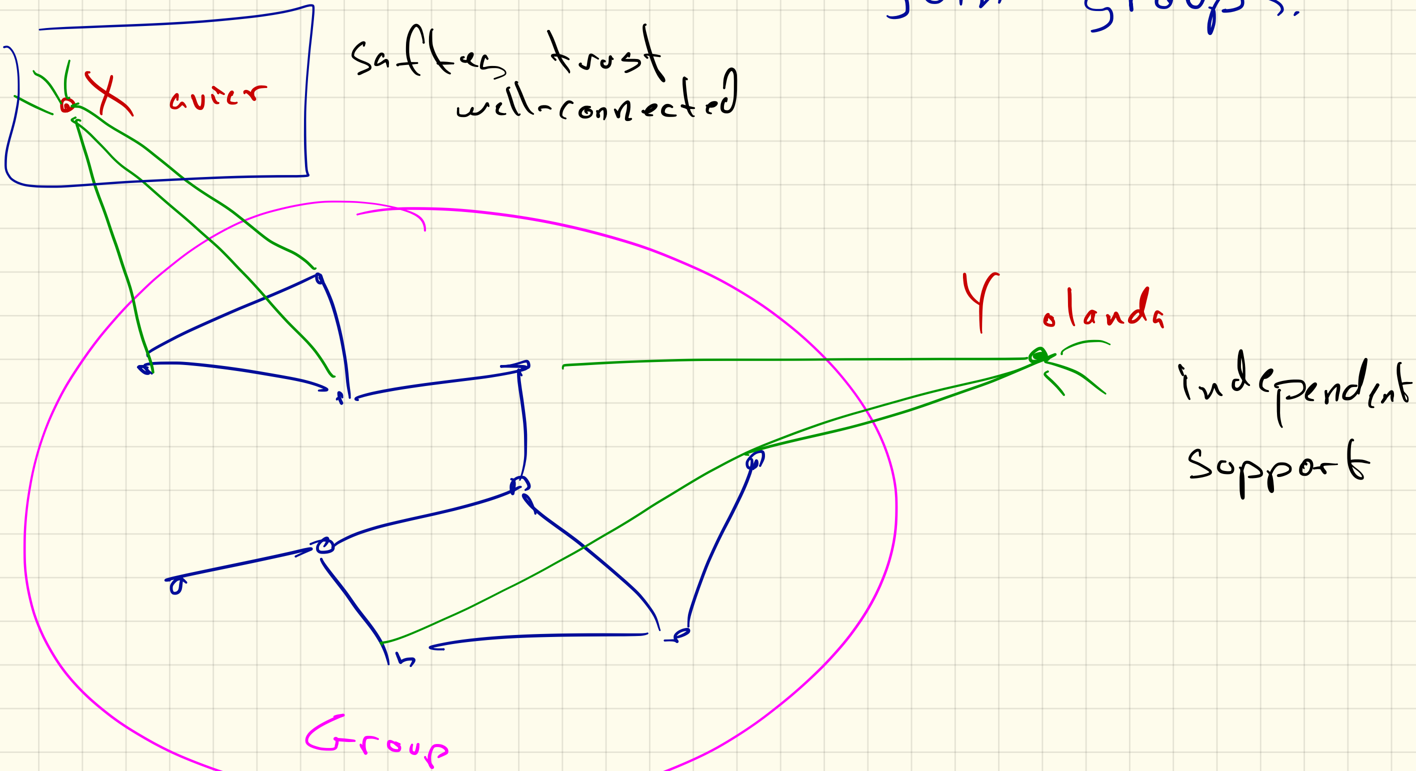
Mathematics

Erdös-Renyi Model

$|V| = n$

E_{ij} exist w/ prob $p \in (0,1)$ iid

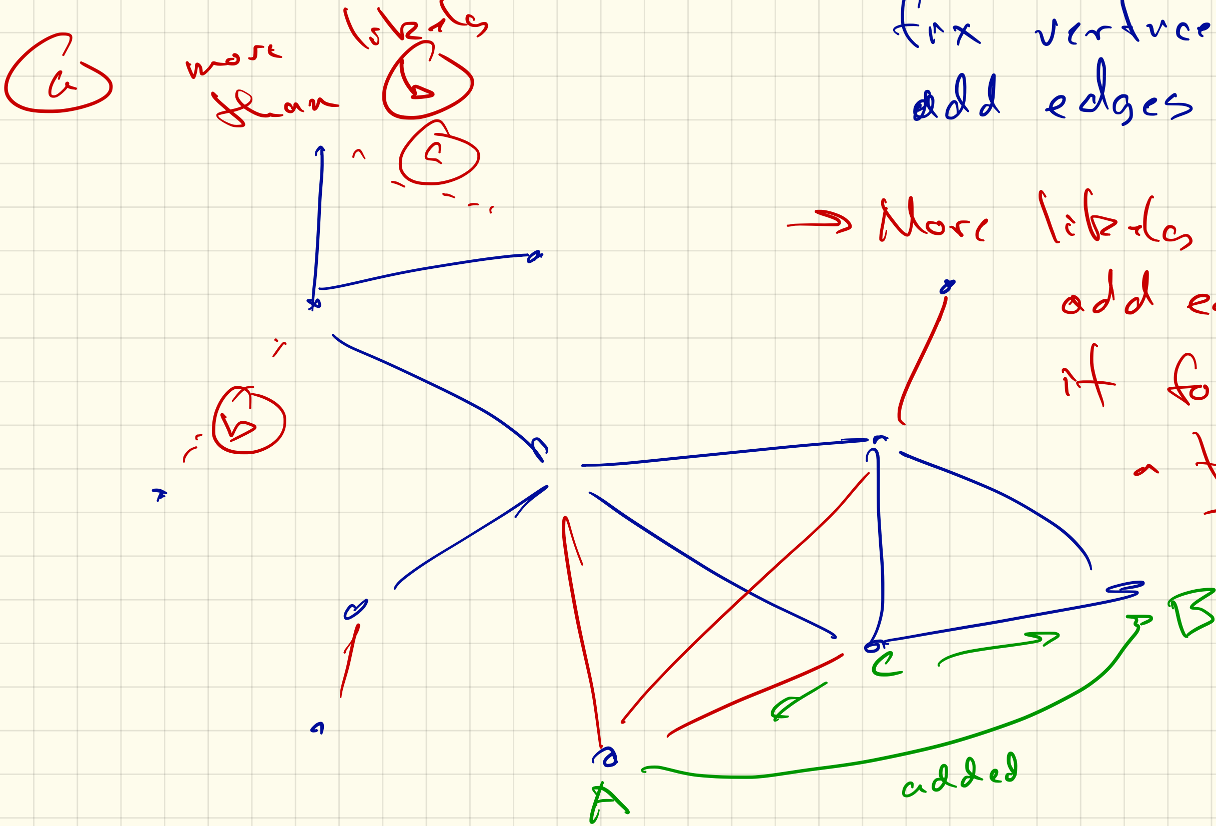
Example Question: Why do people join groups?



Preferential Attachment

fix vertices U
add edges over time

→ More likely to add edges if if forms a triangle



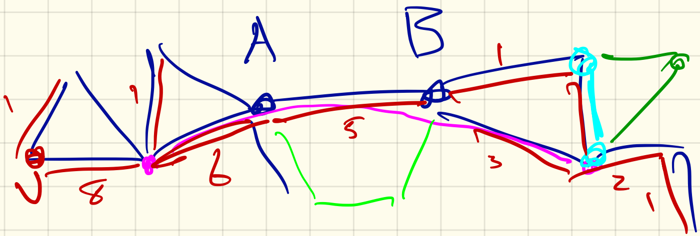
Centrality : importance of vertex (or edge)

- PageRank

Betweenness Centrality

vertices $A, B \in V$

btw(A,B) = fraction of shortest paths which use edge (A,B).



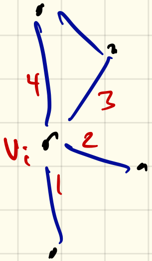
for all $v \in V$

| Dijkstra Alg
| walk back up

Modularity

finding dense community
↳ not partitions

Normalize by vertex degree



$d_i = 4$

Mod $C \subset V$ $\quad Q(C) = $ (fraction of edges in C) $-$ (expected # edges in C)

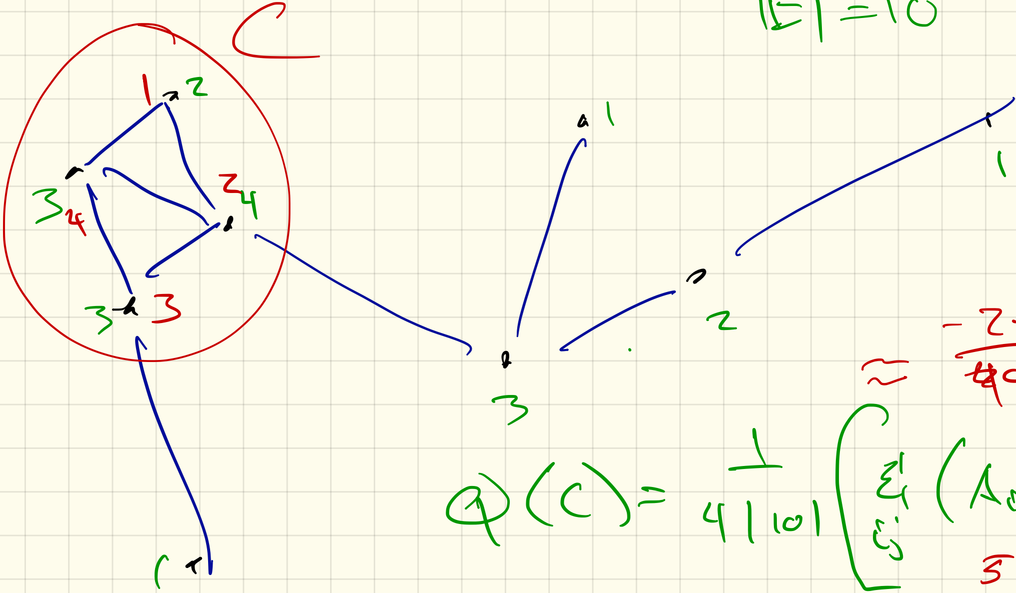
Adjacency matrix A : $A_{ij} = 1$ if edge, 0 o.w.

expected matrix E : $E_{ij} = \dfrac{d_i d_j}{2|E|}$ $\quad d_i = $ degree of v_i $= $ # edges

$$Q(C) = \frac{1}{4|E|}\left[\sum_{v_i, v_j \in C} (A_{ij} - E_{ij})\right] \in [-1/2, 1]$$

typical cluster $\in [0.3, 0.7]$

$$|E| = \sum_{i=1}^{n} d_i$$



$|E| = 10$

$$\approx \frac{-2.5}{40} \approx -0.05$$

$$Q(C) = \frac{1}{4|10|}\left[\sum_{ij}\left(A_{ij} - E_{ij}\right)\right]$$

$5 \qquad \frac{51}{20}$

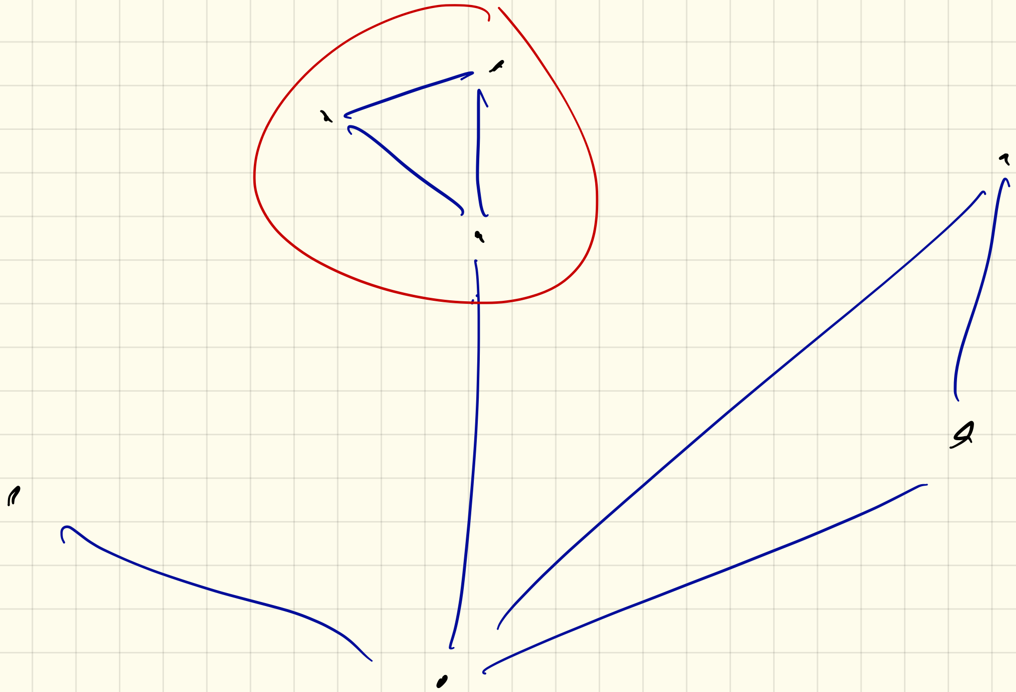
E_{1-4}

	1	2	3	4
1		2/5	3/10	3/10
2			3/5	3/5
3				9/20
4				

$$\frac{8}{20} + \frac{6}{20} + \frac{6}{20} + \frac{12}{20} + \frac{12}{20} + \frac{9}{20}$$

$$\frac{51}{20}$$

$$E_{ij} = \frac{d_i d_j}{2|10|}$$



Small Core Communities

via Cliques

$U' \subset U$ so all pairs $u_i, u_j \in U'$ have edge $\in E$