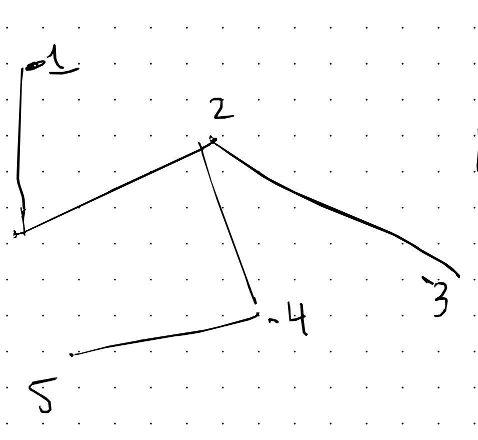
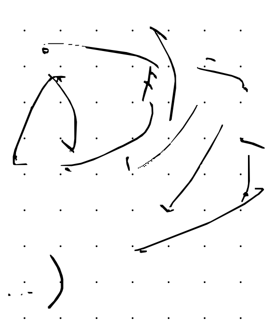
Network as data



Graph $\mathcal{G}$
Nodes $\mathcal{N}$
Edges $\mathcal{E}$

Financial



Nodes (Institutions)
Edges (Financial Exposure, Lending, buying stocks, ...)

During past weeks:
- Networks or graphs as deterministic $(G, \mathcal{N}, \mathcal{E})$
- Linear propagation using $\mathcal{G}$ (Adjacency A)
- Random approaches to networks (Erdos-Renyi graphs)
- Structured randomness (formation resembling real world)
  + Different degrees
  + High clustering
  + Communities (larger sub structures)
  + "Small worlds"

  } Configuration model or extensions

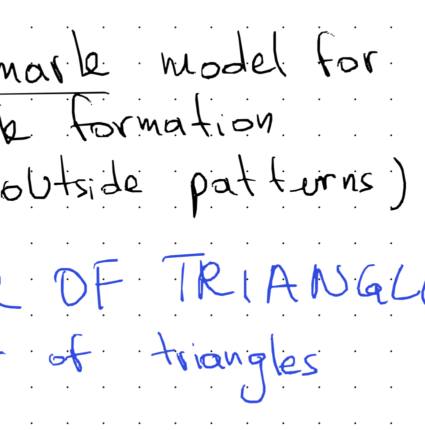
- Connect networks to data
  1. Properties of fixed $\mathcal{G}$
  2. " " random $\mathcal{G}$
  3. In the real world, which of these features matter most?

Main question: "How do observed networks form? What factors matter?

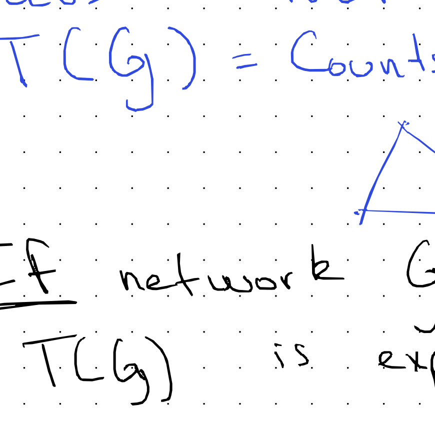
"How to distinguish randomness from true patterns in network data"

Observed graphs $\mathcal{G}^1_{obs}$ and $\mathcal{G}^2_{obs}$

$\mathcal{G}^1_{obs}$



$\mathcal{G}^2_{obs}$

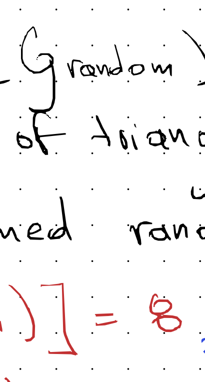


Are patterns forming or is this randomness?

//

"Null" models: Benchmark model for network formation
(In the absence of outside patterns)

Focus on NUMBER OF TRIANGLES
$T(\mathcal{G})$ = Counts # of triangles



If network $\mathcal{G}$ was random, how many $T(\mathcal{G})$ is expected?

If not random, what should $T(\mathcal{G})$ be?

$\rightsquigarrow$ If $T(\mathcal{G}_{obs})$ is very different from $T(\mathcal{G}_{random})$, then this property (# of triangles) is very unlikely to have formed randomly.

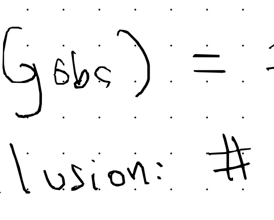
$E[T(\mathcal{G}_{random})] = 8$ (Size $n=50$) $\neq$ $T(\mathcal{G}_{obs}) = 50$ (Same size n)

Since Observed # $\neq$ Random # then, very likely triangles are important for formation

//

Null model: Erdos-Renyi (Random) graphs

Null: $\mathcal{G}(n, p)$ (n nodes, p probability any link is formed at random)



Pr (Triangle occurs)? $\rightsquigarrow$ $p^3 = p \times p \times p$

$$\boxed{T(\mathcal{G}(n,p)) = \binom{n}{3} p^3}$$

$n = 5$ $\rightsquigarrow$ $T(\mathcal{G}(5, 0.5)) = \binom{5}{3}(0.5)^3$
$p = 0.5$ ≈ 1

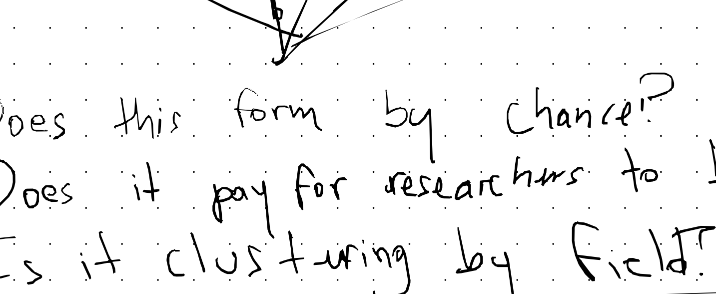
$T(\mathcal{G}_{obs}) = 15$

Conclusion: # of observed triangles is larger than expected by chance
$\Rightarrow$ Triangles are an important feature of the observed network

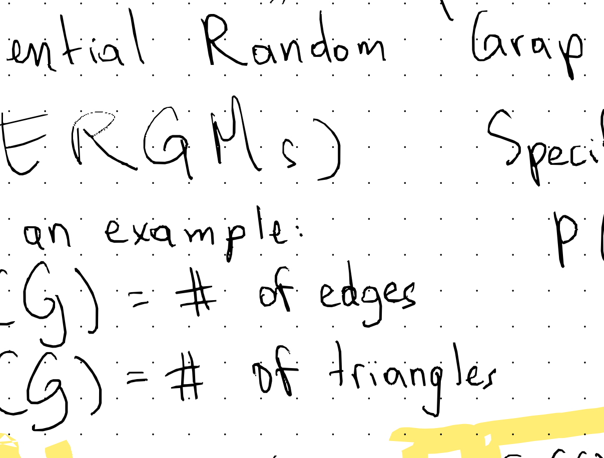
p-value: prob. of observing a feature in excess of just random chance

p-value ≈ 0.0001

Do the same w/ other properties:
1. Clustering coefficients
2. Number of edges
3. Centralities
4. Reciprocity: link to those who link to you (Directed)



Motivation: Collaboration for patents



1. Does this form by chance?
2. Does it pay for researchers to be friends of friends?
3. Is it clustering by field?

//

Exponential Random Graph Models (ERGMs)

Specify probabilities $P(G)$ for all possible G

First, an example:
1. $S_1(\mathcal{G})$ = # of edges
2. $S_2(\mathcal{G})$ = # of triangles

$$P(G) = \frac{\exp(\theta_0 + \underset{<0}{\theta_1} \cdot S_1(G) + \theta_2 \cdot S_2(G))}{Z(\theta)}$$

- $(\theta_0, \theta_1, \theta_2) \Rightarrow$ ERGM coefficients
- $Z(\theta) \rightsquigarrow$ Constant of normalization (ensure $P(G)$ is a probability) $[0,1]$
- $S_1(G)$, $S_2(G) \rightsquigarrow$ Network (graph) statistics

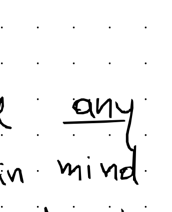
Intuitions:
1. If $\theta_1 > 0$ or $\theta_2 > 0$ then larger statistics $\Rightarrow$ higher probability
   "Encouraging structure"
   $\theta_1 > 0 \Rightarrow$ "More edges, higher prob."
   $\theta_2 > 0 \Rightarrow$ "More triangles, " " "
2. If $\theta_k < 0$, then $\uparrow S_k(\mathcal{G}) \Rightarrow \downarrow$ probability
   "Discouraging structure"
   $\theta_1 < 0 \Rightarrow$ "More edges, lower probability"
   $\theta_2 < 0 \Rightarrow$ "More triangles, " " "
3. Includes random graphs (Erdos-Renyi) as special case: $\theta_1 = \theta_2 = 0$
   With this restriction
   $$P(G) = \frac{\exp(\theta_0)}{Z(\theta)}$$
   $p := \frac{\exp(\theta_0)}{Z(\theta)} \in [0, 1]$ (Fixed number)
   $\rightsquigarrow$ Create an ER graph with

//

General ERGM:
$$P(G) = \frac{\exp(\theta_0 + \theta_1 \cdot S_1(G) + \dots + \theta_d \cdot S_d(G))}{Z(\theta)}$$

Use d network statistics $S_1(G), \dots, S_d(G)$ and YOU choose which ones are used.

$S_1(G)$ = Clustering coefficient ; $\theta_1 > 0$
$S_2(G)$ = # of communities ; $\theta_2 > 0$
$S_3(G)$ = # of squares ; $\theta_3 < 0$



$S_4(G)$ = # of triangles ; $\theta_4 < 0$
$S_5(G)$ = # of edges ; $\theta_5 > 0$
$\vdots$

ERGMs capture local and global incentives to connect (costs)

//

3 Takeaways

1. We can accurately describe any formation model we have in mind and translate to a statistical model
   # of possible networks (undirected) w/ n nodes: $2^{n(n-1)/2}$
2. You can either FIX coefficients $\boxed{(\theta_0, \theta_1, \dots, \theta_d)}$ or you can ESTIMATE them!
   We can now test: $H_0: \theta_1 = 0$
   $H_1: \theta_1 \neq 0$
   If reject, then $S_1(G)$ (Statistic 1) is relevant for network formation.
   $H_0: \theta_1 > 0$
   $H_1: \theta_1 \leq 0$
3. Be careful, ERGMs can have no solutions.
   For some configurations or values $(\theta_0, \theta_1, \dots, \theta_d)$
   $P(G) = 0$ for all relevant networks
   or $P(G) = 1$ for a single network