Session 8:

Social network formation

Andreas Bjerre-Nielsen

Agenda

Session on network formation

Common network patterns

What characterizes social networks?

-
-

Recap

What challenges were faced when measuring peer effects?

-

What are some of the potential ways to overcome these problems?

-

Introduction to network formation

- Leveraging random graphs

The fundamental patterns

Think about your top five friends:

- when, where did you meet them?
- how many friends do you have in common?
- how similar are you?

Mechanisms of social interaction

What characterizes networks sampled?

- Assortative (similarity in personal characteristics)
- Relational (e.g. shared friends)
- Proximity (shared space for meeting, e.g. same school, city)

Birds of a feather flock together

Old proverb - describes that people self select in friendships by ***similarity***. Confirmed by large meta-study McPherson et al. (2000) (https://www.annualreviews.org/doi/10.1146/annurev.soc.27.1.415) along the following dimensions:

- Socioeconomic status, education, ability
- Ethnicity, culture, age, gender
- Interests and hobbies
- Spatially, within country and across country

This pattern is called **homophily**, **sorting** and **assortative matching** / **mixing**.

Sorting in other "networks"

Research in economics and sociology has also focused sorting across:

Institutions

- Marriages (Becker (1973) (https://www.jstor.org/stable/1831130); Mare (1991) (https://www.jstor.org/stable/2095670?seq=1#page_scan_tab_contents))
- Firms, government departments as work places (Mendes et al. 2010 (https://www.sciencedirect.com/science/article/abs/pii/S0927537110000618))
- Schools (Reardon, Owens (2014) (https://www.annualreviews.org/doi/abs/10.1146/annurev-soc-071913-043152))

Space

- Residential areas (Massey, Denton (1988) (https://academic.oup.com/sf/article/67/2/281/2231999); Tiebout (1956) (https://www.journals.uchicago.edu/doi/pdfplus/10.1086/257839))

Policy and implications of sorting

Sorting in across institutions and in space has been subject to laws in recent history:

- United States:
  - School racial segregration was abolished in 1954 (Brown vs Board of Edu.)
  - Residential racial segregration was abolished in 1968 (Fair Housing Act)
- South Africa: Apartheid policies that segregated white and black

Implications of sorting

- increases inequality among households and across generations ([Kremer (1997) (https://academic.oup.com/qje/article-abstract/112/1/115/1870886)](https://academic.oup.com/qje/article-abstract/112/1/115/1870886); [Greenwood et al. (2014) (https://www.aeaweb.org/articles?id=10.1257/aer.104.5.348)](https://www.aeaweb.org/articles?id=10.1257/aer.104.5.348)).
- leads to lower spread of information across sub-populations ([Golub and Jackson (2012) (https://academic.oup.com/qje/article-abstract/127/3/1287/1923572)](https://academic.oup.com/qje/article-abstract/127/3/1287/1923572))

Measuring homophily

We define **homophily index** inspired by [Currarini et al. (2009) (https://doi.org/10.2139/ssrn.1021650)](https://doi.org/10.2139/ssrn.1021650):

- share of edges that are same type: $H = \frac{s}{s+d}$
- possible range [0,1]

If we observe that $H = 0.8$ on gender what should we conclude?

- Nothing!!! We want to know the potential to sort!

Potential homophily

We define **baseline homophily** as:

- We count fraction of potential edges in population of nodes which are same type:

$$B = \frac{\sum_t \#potential(n_t)}{\#potential(n)}, \qquad \#potential(k) = \frac{k \cdot (k-1)}{2}$$

- Interpretation: Expected homophily from random link formation.

Refined homophily measure

We define **inbreeding homophily** as:

$$IH = \frac{H - B}{1 - B}$$

Measures homophily in excess of potential - has upper bound of unity!

Revising conclusions

Resume example with $H = 0.8$. Suppose our comes from a college for nurses with 90 women and 10 men?

- $B = \frac{4050}{4950} \approx 0.82$ where
  - edges of same gender: 4050 ($= \frac{90 \cdot 89 + 10 \cdot 9}{2}$)
  - total edges: 4950 ($= \frac{100 \cdot 99}{2}$)
- $IH = \frac{0.8 - 0.82}{1 - 0.82} \approx -0.05$; we have ***inbreeding heterophily***!

Conditional edge independence

Basic model

Erdos-Renyi graph:

- Each link has a constant probability of forming

Implications:

- Edges follow are Bernoulli (i.i.d.):
  - Indendence
  - Identical distribution

Models of sorting

How can we estimate empirically whether people sort?

- Naive measure: correlation between node attributes.
- We can test for multiple attributes by using a random graph
    - This is basically a version of the Erdos-Renyi graph.
    - Estimation using logistic regression.
- Note: We need to cluster errors at person level (Aronow et al. (2015) (https://www.jstor.org/stable/24573193)).

Handling power laws

- Chatterjee, Diaconis, Sly (2011) (https://doi.org/10.1214/10-aap728) extends random graph to handle hetereogeneous degrees (e.g. power law distribution)
- Graham (2017) (https://doi.org/10.3982/ecta12679) extends the model to allow for unobserved homophily.
  - Assume there is a latent characteristic that people sort on.
  - This characteristic is a measure of "attractiveness" and also implies more connections.
  - Use econometric model to measure unobserved sorting.

New problems

What is problematic with the random graph models?

- We assume likelihood of links forming only depend on link characteristics.
- Implications:
    - Each link is formed *independently*.
    - Cannot recreate observed propensity *clustering*, *shortest paths*, *degree distributions*.

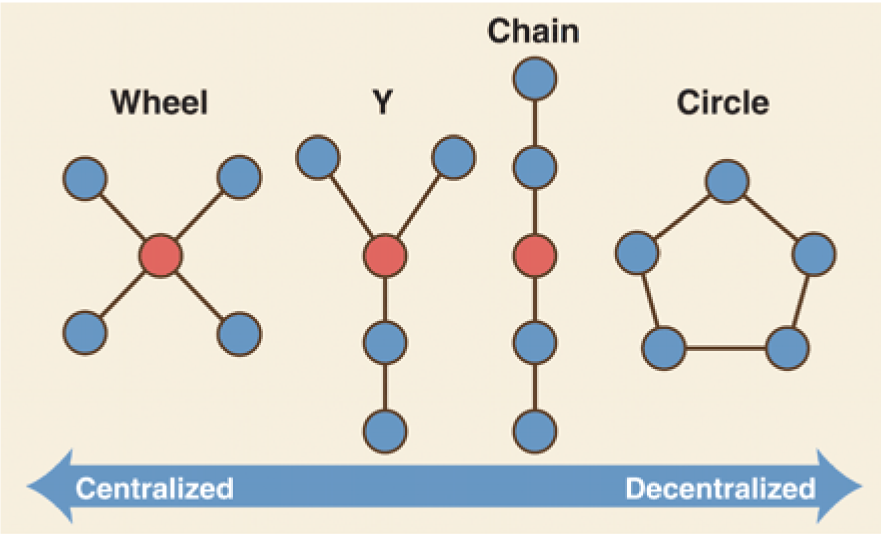
- What could go wrong?
    - People have preferences not captured by single edges:
        - for being in groups (clustering)
        - being more connected in network (centrality)
    - There are uobserved node/edge characteristics.

Dependent link formation

First attempt

[Frank, Strauss (1986) (https://doi.org/10.1080/01621459.1986.10478342)](https://doi.org/10.1080/01621459.1986.10478342) pioneered the use of Exponential Random Graph Models (**ERGM**)

- Handles non-independent formation by writing down likelihood over ***entire*** *graph.*
  - Can capture triangles and other motifs in networks:



ERGM estimation

[Robins, Snijders, Wang, Handcock, Pattison (2007) (https://doi.org/10.1016/j.socnet.2006.08.003)](https://doi.org/10.1016/j.socnet.2006.08.003) provides an overview of recent developments in ERGM

- Estimation is possible using Monte Carlo Markov Chain (MCMC).
- Idea: can express likelihood function $\mathbb{P}[Y = y|\theta] = \frac{\exp(\theta(y))}{c(\theta)}$

ERGM shortcomings

What could be problematic? Is convergence guaranteed?

Problem 1: slow convergence

- state space is huge: 1,000 nodes implies $2^{500,000} \approx 10^{30,000}$ configurations.
- the problem is computationally difficult

Problem 2: degeneracy

- Bhamidi, Bresler, Sly (2011) (https://doi.org/10.1214/10-aap740) that ERGM models may have different parameters consistent with network
- Chatterjee, Diaconis (2013) (https://doi.org/10.1214/13-aos1155) proved that ERGM is many cases indistinguashable from Erdos-Reny
    - We are back to where we started!!

Modelling social dynamics

Components of a structural model

- Who are the agents, their characteristics?
- How do people meet? (friends, institutions?)
- What decides if people become friends? (when meeting)

Applied structural models

[Mele (2017) (https://doi.org/10.3982/ecta10400)](https://doi.org/10.3982/ecta10400) develops a structural model and shows

- investigates potential games (everyone has same preferences over network)
    - EXTREMELY strong assumption
- microfounds ERGM as equilibrium outcome
- identification requires only non-positive externalities, otherwise degenerate

Graham (2016)
(http://bryangraham.github.io/econometrics/downloads/working_papers/DynamicNetwork
develops a structural model that can distinguish between unobserved homophily and preference for participating in groups.

Stochastic actor-based models by Snijders, van de Bunt, Steglich (2010) (https://doi.org/10.1016/j.socnet.2009.02.004) use dynamic evolution.

- See application by Lewis, Gonzalez, Kaufmann (2012) (https://doi.org/10.1073/pnas.1109739109) to measure peer effects.

A model of subgraphs

[Chandrasekhar, Jackson (2018) (https://web.stanford.edu/~arungc/CJ_sugm.pdf)](https://web.stanford.edu/~arungc/CJ_sugm.pdf) develops the idea of subgraph models

Focus on measuring fractions of realized subgraphs.

- The subsgraphs are motifs, i.e. fundamental network components. Examples include trinagles, bridges, stars, wheels etc.
- Fraction is computed as actual vs. potential

Subgraph advantages

Chandrasekhar, Jackson (2018) (https://web.stanford.edu/~arungc/CJ_sugm.pdf) show some advantages of using subgraph models:

- General approach which is applicable to many settings.
- Sophistacted counting to ensure that motifs are not incidentally generated.
- Ensure identification of parameters both when observing a single (large) network or multiple networks.

Subgraph example: counting triangles

We want to avoid incidental generation of motifs. E.g. was a triangle formed at random by three edges or on purpose? How can we avoid this? We count separately:

1. count actual and potential triangles
    - actual: paths of length 3 that starts and end at the same person
    - potential: binomial_coef(n,3)
2. count links that are not part of triangles
    - edge selection: edges that are not part of a triangle or if formed would be a triangle