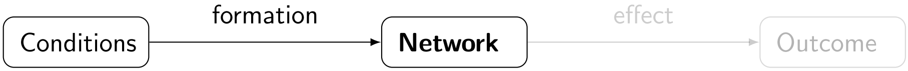
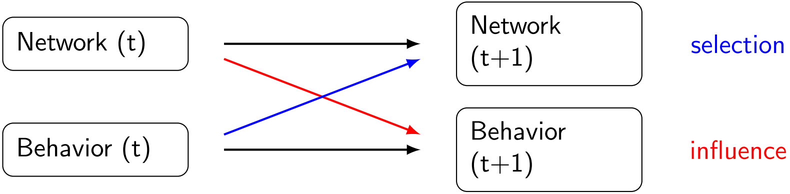
Network modeling

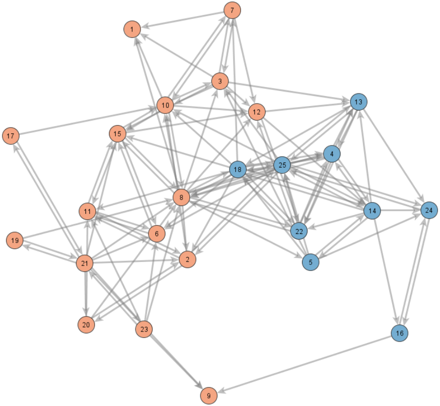
Network as a dependent variable



In longitudinal studies: co-evolution of networks and behaviors



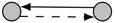
An example: friendship network



Why do ties occur?

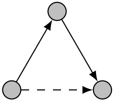
Multiple social/network processes (I) - dependence of ties on other ties and attributes

Explain the emergence of network structure (macro-level) by local (micro-level) processes



Reciprocity

Social exchange theory (?)
Game theory (?)
...



Transitivity (?)

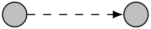
Balance theory (?)
Trust
Safety
...

Left: what we observed. Right: theoretical argument

Why do ties occur?

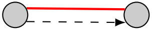
Multiple social/network processes (II) - dependence of ties on attributes

Explain the emergence of network structure (macro-level) by local (micro-level) processes



Homophily (**?**)

Meeting opportunity
Affinity/Attraction (**?**)
Organizational foci (**?**)
...



Dyadic attributes
Other relationships

Network proximity
Physical proximity (geography)
...

Left: what we observed. Right: theoretical argument

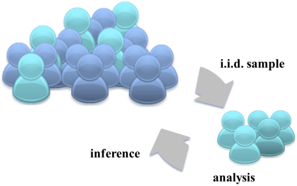
Why do we need statistical network models?

- Distinction between a dependent variable (network ties/individual outcome) and explanatory variables (endogenous: network ties, individual outcome, exogenous: indvidual and dyadic attributes)
- Combinations of multiple mechanisms of tie formation/attachment: test theories controlling for alternative explanations
- Assessment of uncertainties in inference
- Combine structure and attributes to explain the network formation (and possible evolution)

Why have network models been developed?

Standard statistical setting

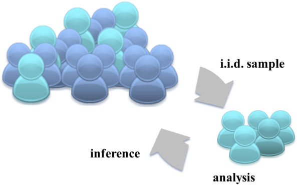
- Sampling from a population
- Set of variables associated with
  a set of entities
- Independence assumption



Why have network models been developed?

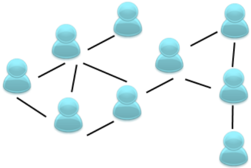
Standard statistical setting

- Sampling from a population
- Set of variables associated with
  a set of entities
- Independence assumption



Network analysis setting

- One group of entities and relations
- Set of variables associated with
  a set of pairs of entities that overlap
- Dependence assumption



Some notation

- $\mathcal{N} = \{1, \ldots, n\}$: set of nodes (also referred as to vertices, actors)
- $\mathcal{X}$ (sometimes I can also use Y): space of all the possible networks defined on $\mathcal{N}$
- x: observed adjacency matrix
- x_{ij}: cell of the adjacency matrix indicating if there is a relation between i and j

  focus on binary relations

$$x_{ij} = \begin{cases} 1 & \text{if there is a tie from } i \text{ to } j \\ 0 & \text{otherwise} \end{cases}$$

- $G(V, E)$ is a (random) graph with a node set V and an edge set E
- Capital letters to denote the corresponding random variables
  X random network and X_{ij} random tie variable
- Lower Greek letters denote parameters (e.g. θ) and a hat is used for their estimates (e.g. $\hat{\theta}$)
- Capital Greek letters to denote the parameter space (e.g. Θ)

Network models

Def.: *A network model is a probability distribution*

$$\{P(x;\theta),\ x \in \mathcal{X},\ \theta \in \Theta\}$$

indexed by the parameter θ and defined over the space of all possible networks $\mathcal{X}$

The richness of network models derives largely from how we choose to specify $P(x;\theta)$

- Uniform distribution on a set of graphs (random graph models)
- Specification based on mechanisms reproducing characteristics of observed networks (preferential attachment model, small-world model)
- Formulation to test the endogenous and exogenous generative mechanisms of a network

  (e.g., QAP regression, ERGMs, SAOMs)

The notion of Random Graphs

- Let $G = (V, E)$ be a graph. If E (and perhaps V) is a random set, then G is a random graph
  - Can consider G to be a random variable on some set G of possible graphs
  - we can write the graph probability mass function (pmf) as $P(G = g)$
- Let $\mathbf{X}$ be the adjacency matrix of the random graph G. Then $\mathbf{X}$ is a random matrix
  - W can write the graph pmf as $P(\mathbf{X} = \mathbf{x})$
  - $\mathbf{X}_{ij}$ is a binary random variable which indicates the state of the (random) i,j edge
  - $P(\mathbf{X}_{ij} = \mathbf{x}_{ij})$ is the probability of the $\mathbf{X}_{ij}$ edge state

Classical Random Graphs

- The n, m family (Erdos-Renyi, size/density conditional uniform graph (CUG)): a graph is chosen uniformly at random from the collection of all graphs which have n vertices and m edges.
  - let M be the maximum possible number of edges in G which is equal to $M = \frac{n(n-1)}{2}$, then:
  - $P(G = g|n, m) = \binom{M}{m}^{-1}$
  - the model assign non-null probability to networks with the same number of edges m (and null probability to all networks with a number of edges $\neq m$)
- The n, p family (homogeneous Bernoulli graphs again proposed by erdos and Renyi): a graph is constructed by connecting n nodes randomly.
- Each edge is included in the graph with probability $p = P(X_{ij} = 1)$, with the presence or absence of any two distinct edges in the graph being independent
  - $P(G = g|n, p) = p^m (1-p)^{n(n-1)/2-m}$
- the two models are equivalent for large n (e.g., asymptotic Poisson distribution of the degree)

Classical Random Graphs /2

- The parameter p in this model can be thought of as a weighting function; as p increases from 0 to 1, the model becomes more likely to include graphs with more edges
- In particular, the case $p = 0.5$ corresponds to the case where all $2^{\binom{n}{2}}$ graphs on n vertices are chosen with equal probability
- A graph $G(n, p)$ has on average $\binom{n}{2} p$ edges
- The distribution of the degree of any particular vertex v is binomial
    - $P(deg(v) = k) = \binom{n-1}{k} p^k (1-p)^{n-1-k}$
    - Since $P(deg(v) = k) \rightarrow \frac{np^k e^{-np}}{k!}$ as $n \rightarrow infinity$ and np= constant. This distribution is Poisson for large n and $np =$ const.
-

Small-world networks

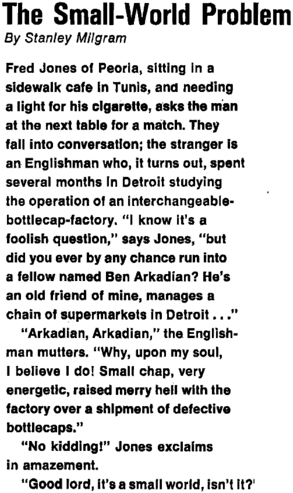
Two properties:

- Small distance between most nodes

  → average path length

- High level of clustering (two nodes with a common neighbor are more likely to be adjacent)

  → clustering coefficients



The Small-World Problem

By Stanley Milgram

Fred Jones of Peoria, sitting in a sidewalk cafe in Tunis, and needing a light for his cigarette, asks the man at the next table for a match. They fall into conversation; the stranger is an Englishman who, it turns out, spent several months in Detroit studying the operation of an interchangeable-bottlecap-factory. "I know it's a foolish question," says Jones, "but did you ever by any chance run into a fellow named Ben Arkadian? He's an old friend of mine, manages a chain of supermarkets in Detroit . . ."

"Arkadian, Arkadian," the Englishman mutters. "Why, upon my soul, I believe I do! Small chap, very energetic, raised merry hell with the factory over a shipment of defective bottlecaps."

"No kidding!" Jones exclaims in amazement.

"Good lord, it's a small world, isn't it?'

Small-world networks

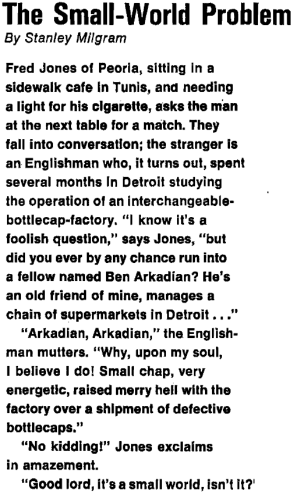
Two properties:

- Small distance between most nodes

  → average path length

- High level of clustering (two nodes with a common neighbor are more likely to be adjacent)

  → clustering coefficients



The Small-World Problem

By Stanley Milgram

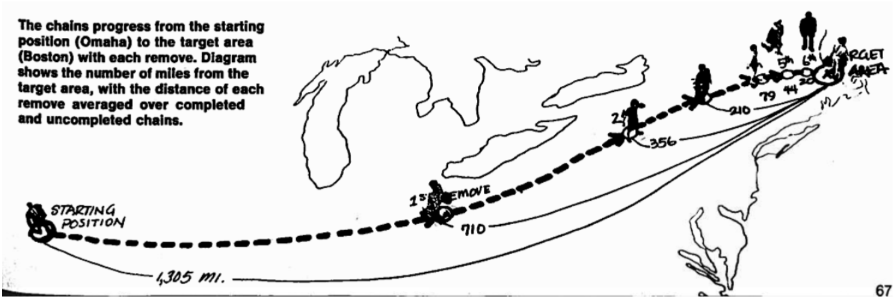
Fred Jones of Peoria, sitting in a sidewalk cafe in Tunis, and needing a light for his cigarette, asks the man at the next table for a match. They fall into conversation; the stranger is an Englishman who, it turns out, spent several months in Detroit studying the operation of an interchangeable-bottlecap-factory. "I know it's a foolish question," says Jones, "but did you ever by any chance run into a fellow named Ben Arkadian? He's an old friend of mine, manages a chain of supermarkets in Detroit . . ."

"Arkadian, Arkadian," the Englishman mutters. "Why, upon my soul, I believe I do! Small chap, very energetic, raised merry hell with the factory over a shipment of defective bottlecaps."

"No kidding!" Jones exclaims in amazement.

"Good lord, it's a small world, isn't it?'

Milgram's experiment

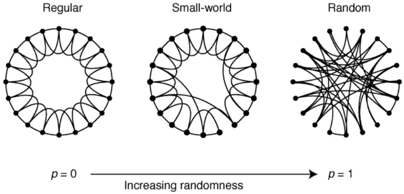


The chains progress from the starting position (Omaha) to the target area (Boston) with each remove. Diagram shows the number of miles from the target area, with the distance of each remove averaged over completed and uncompleted chains.

Chains: from 2 to 10 intermediaries, with a median of 5

[1] Milgram, S. (1967). The small world problem. Psychology today, 2(1), 60-67.

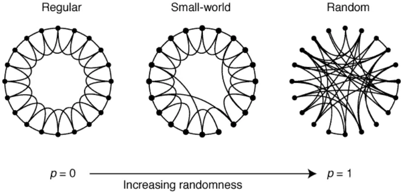
Small-world model



1Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks.
nature, 393(6684), 440-442.

- A ring lattice (high level of transitivity, large distances) is randomly rewired
- With only few rewires, the average distances drop dramatically while the level of cohesion stays relatively high
- One of the most cited "stylized" network models

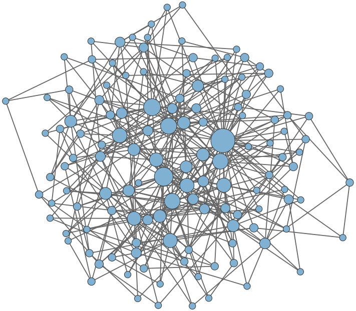
Rewiring

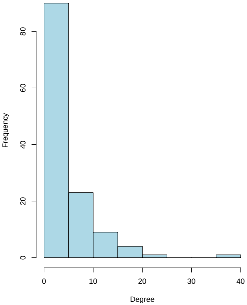


1Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of ‘small-world’networks.
nature, 393(6684), 440-442.

1. Start: ring of n vertices, each connected to its k nearest neighbors
2. Choose a vertex and the edge to its nearest neighbor in a clockwise sense
3. With probability p, reconnect this edge to a vertex chosen uniformly at random over the ring; otherwise leave the edge in place
4. Repeat this process by moving clockwise around the ring, considering each vertex in turn until one lap is completed
5. Circulate around the ring until each edge in the original lattice has been considered

Preferential attachment





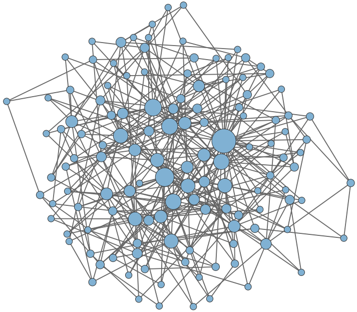
1Simon, H. A. (1955). On a class of skew distribution functions. Biometrika, 42(3/4), 425-440.

1Price, D. (1965). Networks of scientific papers. Science, 149(3683), 510-515.

Approximately a power law distribution

$$\{f_d\}_{d\geq 0} \approx \frac{1}{k^{\gamma}}, \quad \text{for some } \gamma \in \mathbb{R}$$

Preferential attachment model



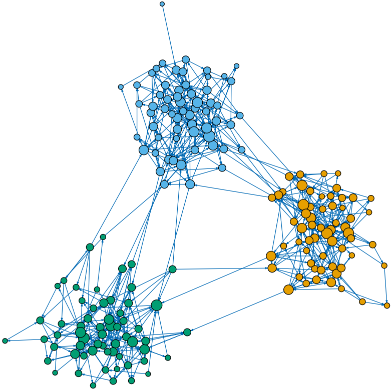
1Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. science, 286(5439), 509-512.

- Nodes are step-wise included to the graph
- Probability of a new node j to connect to an existing node i is proportional to their degree:

$$p_i = \frac{d_j}{\sum_{v' \in V} d_{v'}}$$

- Because this feature is irrespective of the network size, the resulting networks are labeled “scale-free”

Stochastic blockmodels



- The vertex set $V \in \{1..N\}$ is mapped to a set of blocks $B \in \{1..k\}$
- A $k \times k$ matrix P indicates the edge probability of ties within and between blocks (communities)
- Diagonal blocks the probability within-group ties (on the left: 5.1%), off-diagonal blocks that of between-groups ties (0.1%)

  $P = \begin{array}{|c|c|c|}\hline \blacksquare & \square & \square \\ \hline \square & \blacksquare & \square \\ \hline \square & \square & \blacksquare \\ \hline \end{array}$
- It can be used as a generative network model to generate test cases for clustering algorithms

What should a good network model do?

"A good model needs to be both estimable from data and a reasonable representation of that data, to be theoretically plausible about the type of effects that might have produced the network, and to be amenable to examining which competing effects might be the best explanation of the data."

Robins and Morris (2007)

What should a good network model do?

“A good model needs **to be both estimable from data and a reasonable representation of that data**, to be theoretically plausible about the type of effects that might have produced the network, and to be amenable to examining which competing effects might be the best explanation of the data.”

Robins and Morris (2007)

What should a good network model do?

“A good model needs to be both estimable from data and a reasonable representation of that data, **to be theoretically plausible about the type of effects that might have produced the network**, and to be amenable to examining which competing effects might be the best explanation of the data.”

Robins and Morris (2007)

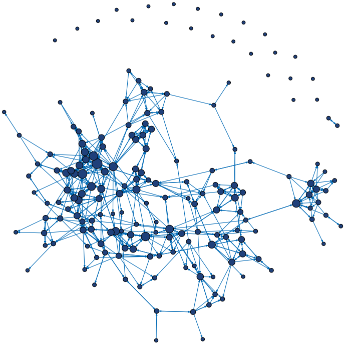
What should a good network model do?

“A good model needs to be both estimable from data and a reasonable representation of that data, to be theoretically plausible about the type of effects that might have produced the network, and **to be amenable to examining which competing effects might be the best explanation of the data**.”

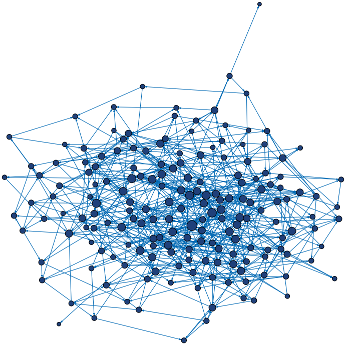
Robins and Morris (2007)

One empirical, one random network: How do they differ?

Empirical



G(183,590)

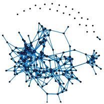


Ideas?

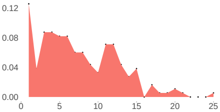
Comparison of two networks

- Density: ratio between the number of ties present and the possible number of ties
- Degree: number of ties incident to a node (for directed networks: indegree and outdegree)
- Isolates: a node with zero degree (indegree and outdegree)
- Degree centralization: a measure of the variance of the degree (range [0,1], 0 all the actors have the same degree, 1 one actor completely dominates the others)
- Geodesic distance: the length of the shortest path between two nodes
- Diameter: the largest geodesic distance between any two pair of nodes

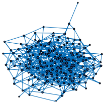
Comparison of two networks; N = 183, M = 590



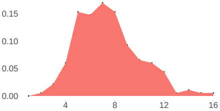
- Density: 0.0177
- Number of isolates: 23
- Reciprocal dyads: 166
- Number of triads: 332
- Degree Distribution:



- Degree centralization: $c_D(G) = 0.048$
- Diameter: 9



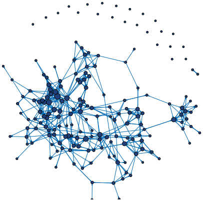
- Density: 0.0177
- Number of isolates: 0
- Reciprocal dyads: 3
- Number of triads: 33
- Degree Distribution:

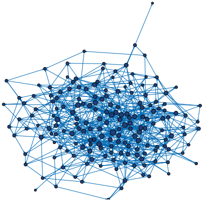


- Degree centralization: $c_D(G) = 0.026$
- Diameter: 5

Graphs generated from the ER model are different from the real network

- There are rarely isolated nodes
- There are too few reciprocal and triadic structures
- The degree distribution does not have a long tail
- The centralization is too low
- The distances between nodes is too small





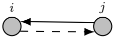
What about the other models?

Model	Criticism
Small world	Low degree variance, poor representation of groups
Preferential attachment	Disregard other network features (e.g., distances and clustering)
Blockmodel	Disregard other network features (e.g., degrees, distances)

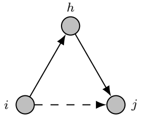
Testing network (tie formation) mechanisms

- The ER model is too simplistic to represent observed networks

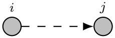
- Still it can be useful to test network mechanisms, e.g.



Reciprocity



Transitivity



Homophily

- If we observe a *large enough* number of a network feature/local configuration (e.g., reciprocal/mutual dyads, transitive triads or homophilous dyads), we have evidence for the corresponding mechanism

- What does “large enough” mean?

Conditional Uniform Graph (CUG) tests

- Is a certain network feature *more prevalent* than expected *by chance*?
  - Operationalization of a network feature ("more prevalent")
  - Definition of hypotheses (H_0: less prevalent or equally present, H_1: more prevalent)
  - Definition of a reference/null network model ("expected by chance")

- Steps:
  1. Calculate the number of local configurations on the observed network (statistic := number of local configurations)
  2. Generate networks from the reference model and compute the value of the statistic for each generated network
  3. Calculate a non-parametric p-value by comparing the empirical to the generated statistics

- The simplest reference distribution is the $G(n, m)$ model

Conditional Uniform Graph (CUG) tests

- Is a certain network feature *more prevalent* than expected *by chance*?
  - Operationalization of a network feature (“more prevalent”)
  - Definition of hypotheses (H_0: less prevalent or equally present, H_1: more prevalent)
  - Definition of a reference/null network model (“expected by chance”)

- Steps:
  1. Calculate the number of local configurations on the observed network (statistic $:=$ number of local configurations)
  2. Generate networks from the reference model and compute the value of the statistic for each generated network
  3. Calculate a non-parametric p-value by comparing the empirical to the generated statistics

- The simplest reference distribution is the $G(n, m)$ model

Conditional Uniform Graph (CUG) tests

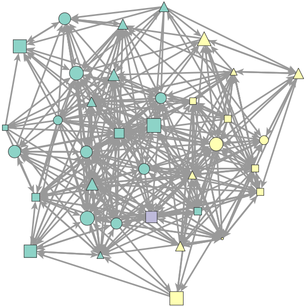
- Is a certain network feature *more prevalent* than expected *by chance*?
  - Operationalization of a network feature ("more prevalent")
  - Definition of hypotheses (H_0: less prevalent or equally present, H_1: more prevalent)
  - Definition of a reference/null network model ("expected by chance")

- Steps:
  1. Calculate the number of local configurations on the observed network (statistic := number of local configurations)
  2. Generate networks from the reference model and compute the value of the statistic for each generated network
  3. Calculate a non-parametric p-value by comparing the empirical to the generated statistics

- The simplest reference distribution is the $G(n, m)$ model

CUG test: example

Evidence for/against social mechanisms in an advice network? ?

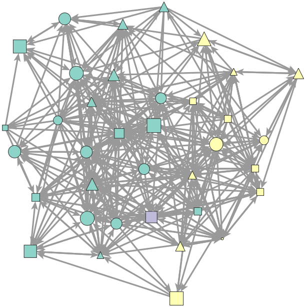


https://www.stats.ox.ac.uk/
~snijders/siena/

- 36 partners in a Northeastern US corporate law firm
- Advice relation:
  "Think back over the past year, consider all the lawyers in your Firm. To whom did you go for basic professional advice?"
- Friendship relation (not shown in the picture):
  "Would you go through this list, and check the names of those you socialize with outside work. You know their family, they know yours, for instance. I do not mean all the people you are simply on a friendly level with, or people you happen to meet at Firm functions."
- Vertex attributes:
  office (green=Boston; yellow=Hartford; violet=Providence)
  school (circle=Harvard, Yale; Triangle: Ucon; Square: other)
  years with the firm (node area)

CUG test: example

Evidence for/against social mechanisms in an advice network? ?



- Is there evidence for reciprocity?
- Is there evidence for transitivity?
- Is there evidence for school homophily?

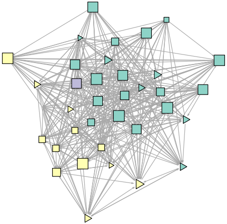
Reference/Null model: $G(n, m)$, $n = 36$, $m = 395$

Full description and data download:
`https://www.stats.ox.ac.uk/~snijders/siena/`

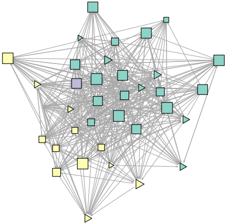
CUG test: example

Is there evidence for reciprocity?

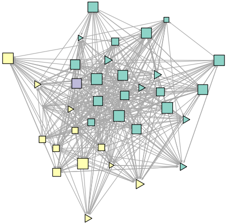
1. Number of observed mutual dyads: $M = 106$
2. Generate networks form $G(36, 395)$ and compute M



$M = 63$



$M = 54$



$M = 75$

...

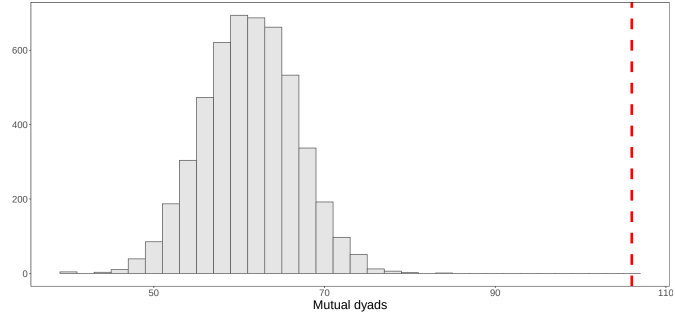
$M =$

...

CUG test: example

Is there evidence for reciprocity? Yes!

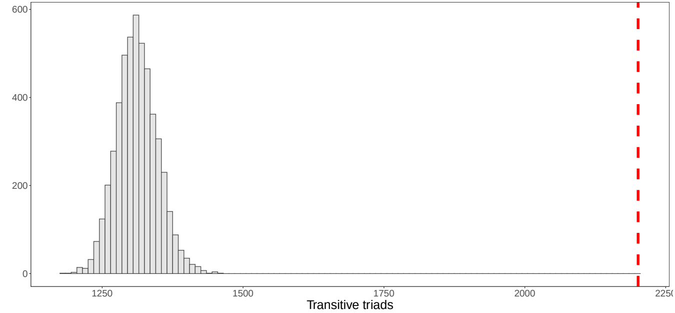
3. Calculate a non-parametric p-value



- None of the random networks has equal or more mutual dyads (p-value $=$ 0)
- Under a significance level of $\alpha = 0.05$ we consider this result significant

CUG test: example

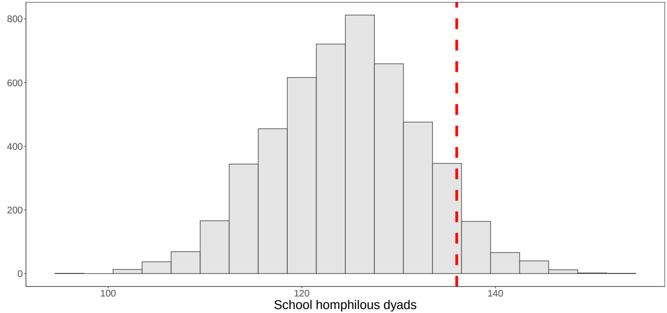
Is there evidence for transitivity? Yes!



- None of the random networks has equal or more transitive triads (p-value $= 0$)
- Under a significance level of $\alpha = 0.05$ we consider this result significant

CUG test: example

Is there evidence for school homophily? No!



- 7.8% of the random networks has equal or more school homophilous dyads (p-value = 0.078)
- Under a significance level of $\alpha = 0.05$ we consider this result not significant

Limitations of CUG tests

- CUGs allow testing hypotheses using a reference model conditioning on some network statistics
- CUGs have two limitations
  1. Given combinatorial complexity, CUGs are difficult to compute when large sets of conditioning statistics are considered
  2. We cannot "generalize" test results to the phenomenon we are analyzing. We can only claim that the observed value for the test feature is unlikely conditionally on some considered statistics

     E.g., we **can** claim:

     Reciprocal ties are **more** *likely to occur in the observed network* **than in random networks with m edges**

Limitations of CUG tests

- CUGs allow testing hypotheses using a reference model conditioning on some network statistics
- CUGs have two limitations

  1. Given combinatorial complexity, CUGs are difficult to compute when large sets of conditioning statistics are considered
  2. We cannot "generalize" results to the observed network. We can only claim that the observed value for the test feature is unlikely conditionally on some considered statistics

     E.g., we **cannot** claim:

     Reciprocal ties are likely to occur **in the observed network**

Good network models

‘A good model needs to be both estimable from data and a reasonable representation of that data, to be theoretically plausible about the type of effects that might have produced the network, and to be amenable to examining which competing effects might be the best explanation of the data.’ (Robins and Morris, 2007)

CUGs are not the best options to model network data given their limitations

Descriptives vs generative goals

- Descriptive: numerical summary measures
  - Nodal level: e.g., centrality
  - Configuration level: e.g., triad census
  - Network level: e.g., centralization, clustering (aka community detection)
- Generative: micro foundations for macro patterns
  - Global patterns are locally emergent
  - Recover underlying dynamic processes from cross-sectional data
  - Test hypotheses (on tie formation mechanisms)
  - Extrapolate and simulate from model

Some considerations: multiple mechanisms

Different (social) processes can lead to similar macro signatures

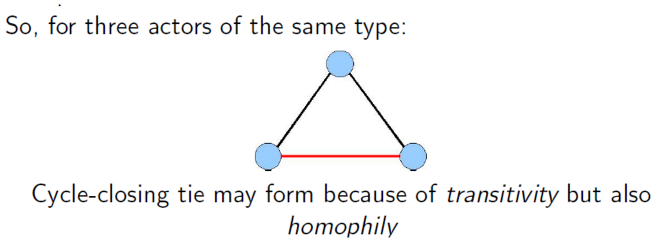
- For example: " typically observed in social nets can be a result of
  - Sociality - highly active persons create clusters
  - Homophily - assortative mixing by attribute creates clusters
  - Transitive triad closure - triangles create clusters

Want to be able to fit these terms simultaneously, and identify the independent effects of each process on the overall outcome.

Some considerations: multiple mechanisms

Example: Two theories about the process that generates 3-cycles in an undirected graph

1. Homophily: People tend to chose friends who are like them, in grade, race, etc. (birds of a feather), triad closure is a by-product
2. Transitivity: People who have friends in common tend to become friends (friend of a friend), closure is the key process



So, for three actors of the same type:

Cycle-closing tie may form because of *transitivity* but also *homophily*