Network Analysis

Dr. Shaina Race
Institute for Advanced Analytics

Hypothesis Testing for Networks

• • •

Individual Level Hypotheses

- Does the social capital of legislators predict success of the bills they sponsor?
- Does organizational connectedness predict speed of promotion?
- Does out-degree in an advice network predict learning?

Problems

1. Observations not independent!

    - If I have a high centrality, yours may go up by association
    - My connectedness/connections influence yours
    - Undirected networks force reciprocal links
    - Social connections have other limiting factors: time, money, freedom, happenstance.

Problems

2. Often not a random sample.
   - How useful would a random sample be if we're interested in relationships?
   - Snowball sampling
   - $\longrightarrow$ Population distribution of variables unknown

3. Because of #1 & #2, cannot compute significance through traditional tests.
   - Can compute correlations and statistics but cannot speak to their statistical significance.

Solution

Permutation Tests

- *Simulate* the null hypothesis: What would it look like if there were no association/difference?
- Take one column of data and shuffle (permute) it randomly
- Calculate the statistic of interest on the shuffled data.
- Repeat many times
- Get distribution of values you'd expect to find if there were no association/difference
- See where the value from the original observed data falls in that distribution

Example

- Trust network from an organization.
- Ask employees questions on team feedback "I trust this individual to operate effectively and efficiently with minimal guidance"
- Want to know if an individual's trustworthiness is related tenure with company.

- **Hypothesis**: In-Degree is correlated with tenure. (observed r=0.39)

Example

- **Hypothesis**: In-degree is correlated with tenure. (observed r=0.39)
- A p-value is the probability we got something as extreme as the observed result *if there is truly no relationship*.
- So simulate what it looks like when there is truly no relationship!

Example

Actual Data

In-Degree	Tenure
1	1
1	2
2	5
2	4
3	10
5	5
6	9

Shuffled Data

In-Degree	Tenure
1	10
1	2
2	4
2	9
3	5
5	1
6	5

Shuffled Data

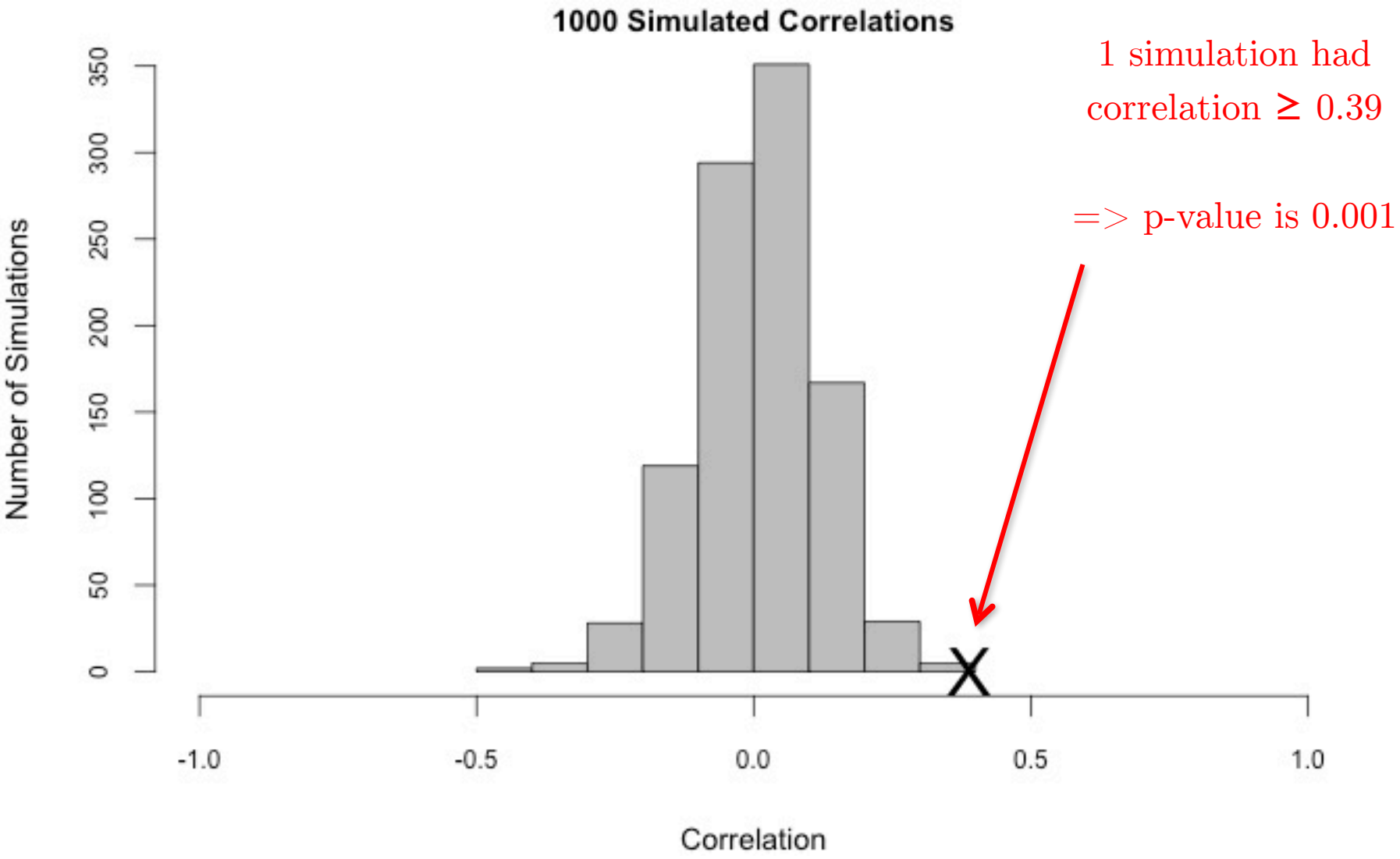
In-Degree	Tenure
1	1
1	9
2	5
2	5
3	10
5	4
6	2

Shuffled Data

In-Degree	Tenure
1	5
1	2
2	9
2	4
3	10
5	5
6	1

This is what it would look like if there were no relationship between In-Degree and Tenure (but univariate distributions in tact)

Example



1 simulation had correlation ≥ 0.39

$=>$ p-value is 0.001

Dyad Level Hypothesis

- Does homework group membership correlate with outside social interactions?
- Do people with strong social bonds tend to have many friends in common?

- **Challenge:** Correlating two adjacency matrices or comparing network statistics (which are based on two adjacency matrices)

Problems

- All the same problems listed previously.

- **Solution #1:** Turn the adjacency matrices into vectors. Simulate as before by randomly permuting one of the vectors.

- ***Problem* with Solution #1:** Randomly permuting one of the vectors does not handle the full range of dependencies between dyads (e.g. you could drastically change the degree distribution of the network).

Solution

QAP Approach

- Randomly permute rows AND columns of one matrix using the same permutation.
- In essence, this just re-labels the nodes in one matrix (but overall the degree structure is the same)
- Then compare the permuted matrix with second matrix, record statistic of interest, and repeat many times
- Compute proportion of trials that produced a result equal to or stronger than the one found. This is your p-value.

Network Level Hypothesis

- Is the density of a trust network in a practicum group associated with better performance?
- Is the clusterability of high-school social network associated with higher incidence of fights?

No Problem!

As long as the individual networks are selected randomly from the population of networks, we can use traditional statistical tests.

Network Autocorrelation (Ordinal/Continuous)

- Mixing dyadic and individual variables.
  - one variable is the network, another is for each node/individual

Example:

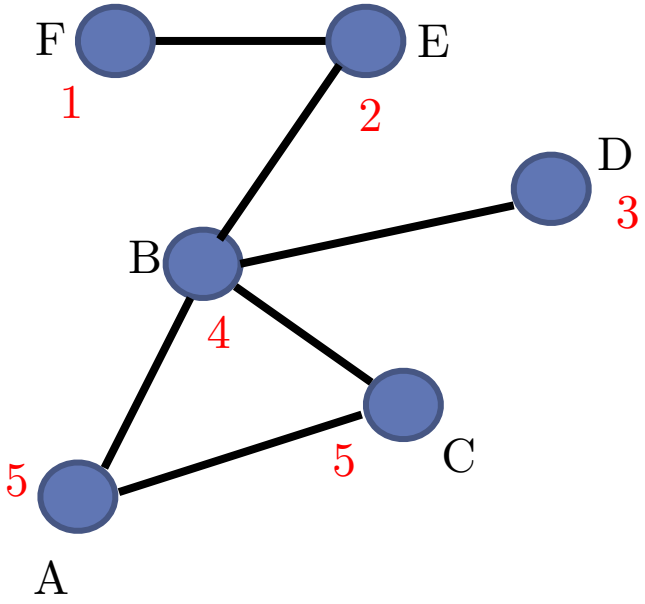
- Friendship network in organization. Individual variable is attitude toward the firm (ordinal). Is attitude contagious?
  - If so, friends should have similar attitudes
  - Nodes 2 links away should be more similar than those 5 links away.

Network Autocorrelation (Ordinal/Continuous)

Solution:

- Create a similarity measure for attitude
  (like the absolute difference $|attitude_i - attitude_j|$)
- See if it correlates with network distance.

Example: Autocorrelation



Numbers in red represent attitude metric

Difference in Attitude

	A	B	C	D	E	F
A	0	1	0	2	3	4
B	1	0	1	1	2	3
C	0	1	0	2	3	4
D	2	1	2	0	1	2
E	3	2	3	1	0	1
F	4	3	4	2	1	0

Network Distance

	A	B	C	D	E	F
A	0	1	1	2	2	3
B	1	0	1	1	1	2
C	1	1	0	2	2	3
D	2	1	2	0	2	3
E	2	1	2	2	0	1
F	3	2	3	3	1	0

If matrices are symmetric, only compare half of values

Network Autocorrelation (Categorical)

Examples:

- Are you more likely to respond to someone's post if that person is of the same gender?
- Do individuals in the same cohort communicate more/less frequently than individuals in different cohorts?

Network Autocorrelation (Categorical)

Connections within Blue Cohort

0	1	0	2	3	4
1	0	1	1	2	3
0	1	0	2	3	4
2	1	2	0	7	4
3	2	3	7	0	3
4	3	4	4	3	0

Connections within Orange Cohort

- Solution: Statistically, are the differences in means between matrix blocks greater than you would expect by chance? Shuffle the Categorical variable (QAP approach).

Fortune Interactive Consulting

- 71 Consultants
- Each **project has 1 or 2 leads**, managing a **team of 3 or 6** individuals respectively.
- Each **lead is responsible for creating a team of 3** other consultants to handle the project.
- Encouraged to work with many different consultants, not always draft same team
  - Newly hired consultants were not being utilized.
  - Some **accusations of preferential treatment** based on mutual interests

Fortune Interactive Consulting

Data:

- Edge variable: Reported **trust in ability** between consultants
- Node variables:
  - **Gender**
  - **Company Tenure**
  - **College Football Team Preference**
  - **College Basketball Team Preference**
  - **Number of Past Project Leads**

Fortune Interactive Consulting

Questions:

- Is there any **relationship between** whether consultants report **trust** in one another and whether they have the same **college football preference?**
- Is there any **relationship between** whether consultants will **invite one another to join** a project and whether they have the same **college football preference?**
- Can we determine factors that contribute to this network structure? **Are mutual interests in sports driving professional relationships?**