Computational Analysis of Network Data

Testing network-level statistics through simulation

Austin van Loon

SICSS 2019 Flash Talk

WARNING! Lots to Cover

Branch: master ▾ SICSS_2019_Flash_Talk / van Loon Flash Talk.ipynb Find file Copy path

AustinVL Add files via upload ff97dbd 1 minute ago

1 contributor

335 lines (334 sloc) | 65.8 KB Raw Blame History

Testing network-level statistics through simulation

The goal of this notebook is to walk you through how to test some basic properties of networks using two different null models: the **Erdos-Renyi model** and the **configuration model**. To begin, we'll import a series of libraries that you'll use for the analysis. Of note is the *networkx* library, which has a bunch of useful tools for analyzing networks in Python. I would also highly recommend the *snap* library which was created and is maintained by Jure Leskovec's lab at Stanford, especially if you're analyzing large networks. It requires a bit of work to install, however, and in my opinion has a slightly steeper learning curve than networkx. *seaborn* is a nice plotting tool that is built off of *pyplot*.

In [14]:

```
import networkx as nx
import csv
import seaborn as sns
import matplotlib.pyplot as plt
import numpy.random as rand
import random
from networkx.algorithms.assortativity import attribute_assortativity_coefficient as homophily
```

Setting up the network

Here we upload the network from a CSV file, which is available on the SICSS GitHub as well as my personal GitHub page. The network is drawn from the Enron email corpus, which is available in full at the Carnegie Mellon Computer Science website. A reduced form the dataset (just the network based off meta-data, no content) is available at the SNAP website.

Here we have a severely reduced form of the dataset. Specifically, this is a cumulative communication network of emails amongst the 150 custodians that were identified by a predictive algorithm to contain **expressive communication**, or communication that was perceived as intrinsically worthwhile (e.g. exchanging jokes, gossiping, etc.). This was done in the service of a larger project, but serves as a cool use case for us. I should mention that me and my co-author (Katariina Mueller-Gastelle) plan to build a stronger predictive algorithm in the future, so these results might not be reflected in future work.

Each row of the CSV we're using contains a "from" and "to", which denotes who sent at least one email in the observation period to whom, and each node is identified by their name. We call this a **directed edge list**, because the CSV file is a list of the connections

JUPYTER NOTEBOOK WITH ANNOTATED CODE AND EXPLANATION

Outline

- The why/when of networks
- Testing network-level statistics
  - Define a statistic
    - Homophily
    - Average clustering coefficient
  - Pick a null model
    - Erdös–Rényi model
    - Configuration model
- Code

Outline

- **The why/when of networks**
- Testing network-level statistics
  - Define a statistic
    - Homophily
    - Average clustering coefficient
  - Pick a null model
    - Erdös–Rényi model
    - Configuration model
- Node2Vec
  - Embedding space
  - Random walks
  - Embedding a node's position using random walks

When/Why Networks?

- Standard econometrics assumes i.i.d
- In many cases, we're interested in interdependence!
- Examples:
  - Relationships
  - Communication
  - Language
  - Computer networks
  - Protein interactions
  - Hyper-links
- Generally anything in which there's measurable interdependence!

Examining Gendered Exclusion and Multiplex Communication Networks

Austin van Loon

Stanford Sociology Department

Motivation

Motivation I: Embeddedness

- 'Embedded' ties [1] in organizations are important for organizational performance [2]
- To date, most research on the formal/informal divide in organizations have been qualitative [2-3]

Motivation II: Gender and Organizations

- Women face allocative discrimination in organizations [4]
- Internships seem to help mitigate this [5], but mechanisms are unknown

Theory

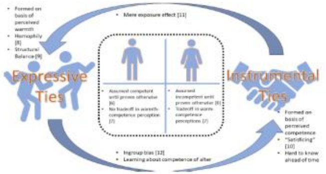
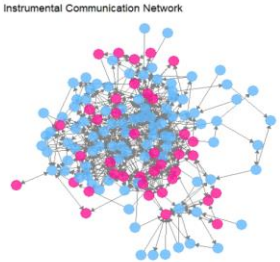
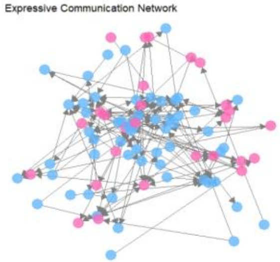


Figure 1: Theoretical Model of Gender and Multiplex Communication in Organizations

General Procedure

- Collect **Enron email data set** (0.5 million emails; in-box and out-box of 150 employees for 3 years); filter out a lot
- **Hand code** 1,500 emails as having or not having instrumental content and expressive content
- Train a **machine learning classifier** to identify both kinds of communication
- Create and analyze **networks** of expressive and instrumental communication





Conclusions

- Women's centrality in the instrumental network are marginally less correlated with their centrality in the expressive network than men ($p \approx 0.08$)
- Both instrumental and expressive communication are characterized by more clustering than we would expect by chance.
- Compared to the homophily we expect by chance, we see significantly more in the instrumental network ($p \approx 0.02$) and marginally more in the expressive network ($p \approx 0.06$)

Formulas

Eigenvector centrality is a measure of one's importance within a network and is defined as the following:

$$x_i = \frac{1}{\lambda}\sum_{j \in G} a_{i,j} x_j \qquad (1)$$

The **average clustering coefficient** for a graph describes how often there are 'triangles' in the graph:

$$\bar{C} = \frac{1}{n}\sum_{i \in G} \frac{2|\{e_{jk} : v_j, v_k \in N_i, e_{jk} \in E\}|}{k_i(k_i - 1)|} \qquad (2)$$

The **assortativity coefficient** of a graph is the degree to which 'birds of feather flock together'. Where e_{ij} is the fraction of edges in a graph that connect actors of type i and j:

$$r = \frac{Tr(\mathbf{e}) - ||\mathbf{e}^2||}{1 - ||\mathbf{e}^2||} \qquad (3)$$

Position Across Graphs

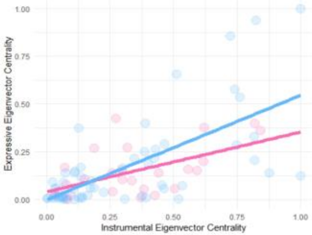


Figure 2: The relationship between network centralities by gender

Clustering

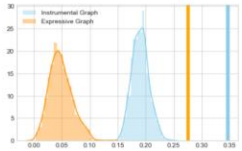


Figure 3: Histogram of clustering over 1,000 randomly re-wired networks and observed values for both networks.

Homophily

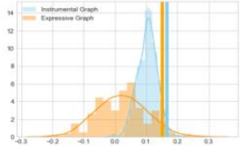


Figure 4: Histogram of homophily over 1,000 randomly re-wired networks and observed values for both networks.

References

[1] M Granovetter. Economic action and social structure. *AJS*, 91(3):481–510, 1985.

[2] M Dalton. *Men Who Manage.* John Wiley and Sons, 1959.

[3] P Blau. *Dynamics of Bureaucracy.* University of Chicago Press, 1955.

[4] T Peterson and I Saporta. The opportunity structure for discrimination. *AJS*, 109(4):852–901, 2004.

[5] A Sterling and R Fernandez. Once in the door. *Management Science*, 64(11):5444–60, 2018.

[6] BP Cohen J Berger and M Zelditch. Status characteristics and social interaction. *Stanford Technical Report.*

[7] P Glick ST Fiske, A Cuddy and J Xu. A model of (often mixed) stereotype content. *JPSP*, 82(6):878–902, 2002.

[8] L Smith-Lovin M McPherson and JM Cook. Birds of a feather. *ARS*, 27:415–44, 2001.

[9] D Cartwright and F Harary. Structural balance. *The Psychological Review*, 63(5):277–93, 1956.

[10] H Simon. Rational choice and the structure of the environment. *The Psychological Review*, 63(2):129–38, 1956.

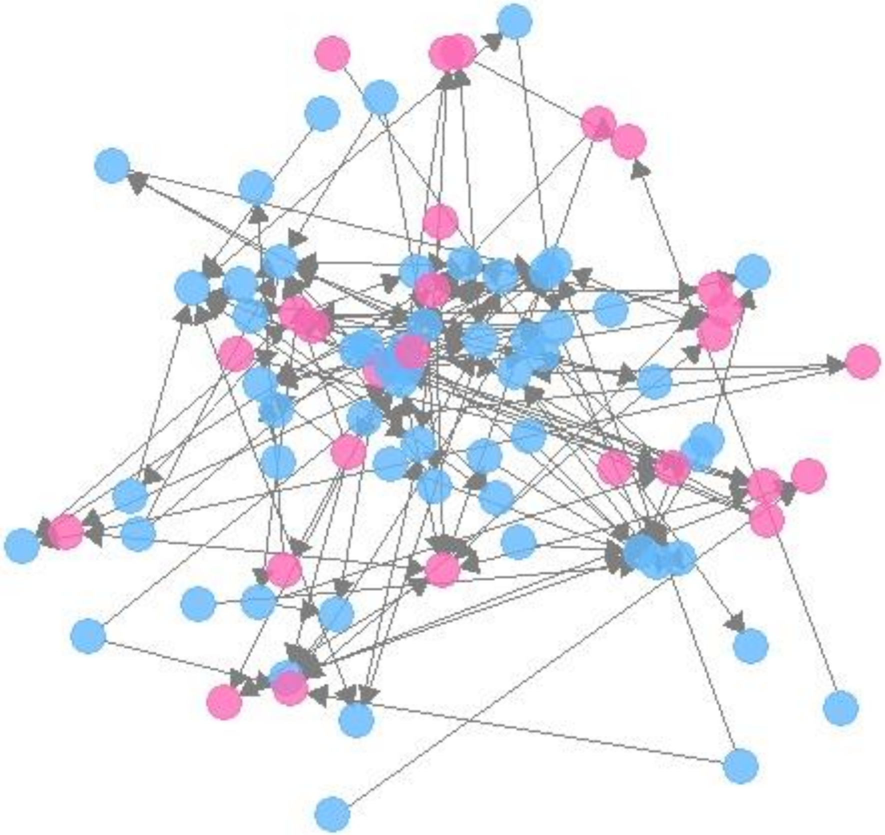
[11] RB Zajonc. Mere exposure: A gateway to the subliminal. *Current Directions in Psych Science*, 10(6):224–8, 2001.

[12] DM Taylor and JR Doria. Self-serving and group-serving bias in attribution. *JSP*, 113(2):201–11, 1981.

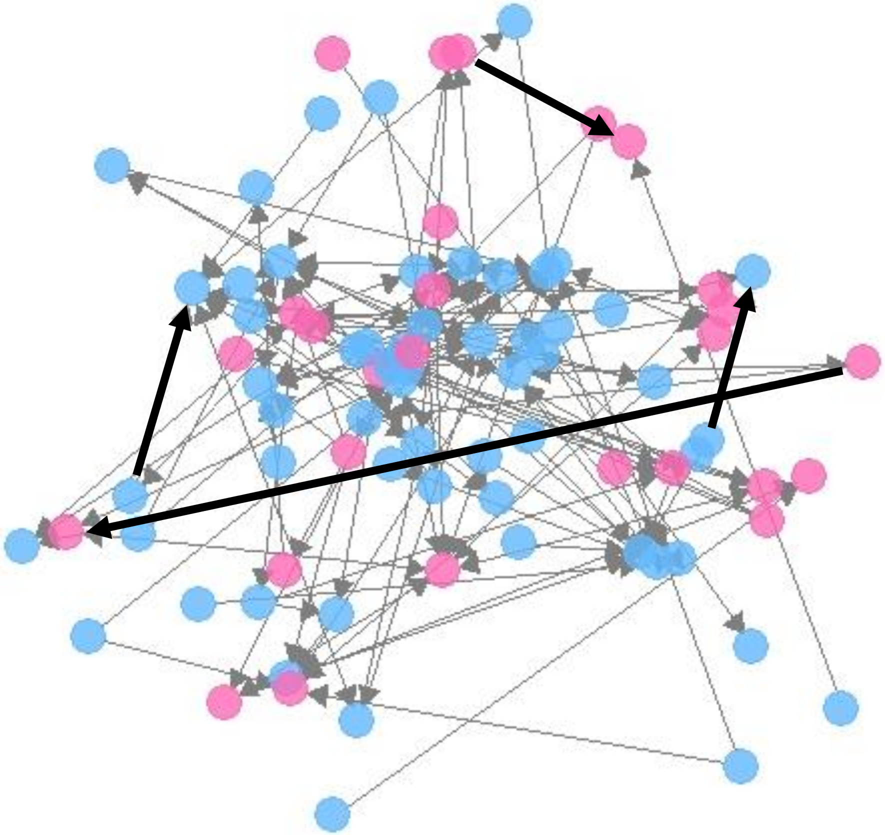
Outline

- The why/when of networks
- **Testing network-level statistics**
  - Define a statistic
    - Homophily
    - Average clustering coefficient
  - Pick a null model
    - Erdös–Rényi model
    - Configuration model
- Node2Vec
  - Embedding space
  - Random walks
  - Embedding a node’s position using random walks

Expressive Communication Network



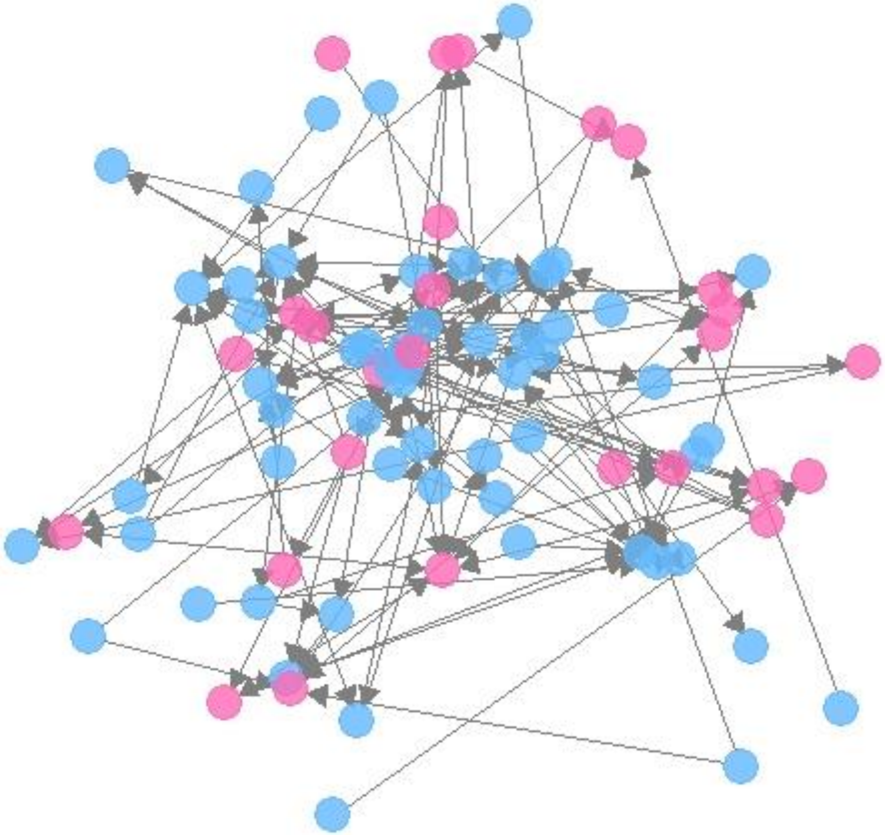
Expressive Communication Network



The **assortativity coefficient** of a graph is the degree to which **"birds of feather flock together"**. Where e is a matrix for which entry e_{ij} is the fraction of edges in a graph that connect actors of type i and j, it is defined as:

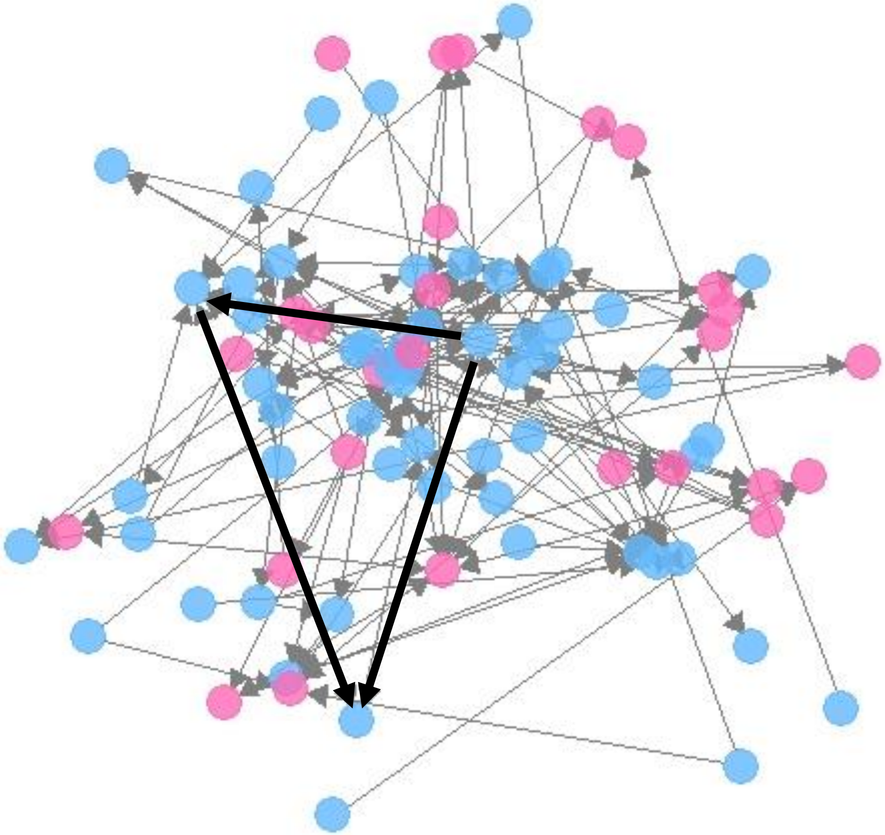
$$r = \frac{Tr(e) - ||e^2||}{1 - ||e^2||}$$

Expressive Communication Network



Stanford University

Expressive Communication Network



The **average clustering coefficient** for a graph describes **how often there are "triangles"** in the graph. Where A_i, is the set of nodes i is connected to and k_i is the size of that set, it is defined as:

$$\bar{C} = \frac{1}{n}\sum_{i \in V} \frac{2\left|\left\{e_{jk}:\ v_j, v_k \in\ A_i;\ e_{jk} \in E\right\}\right|}{k_i(k_i - 1)}$$

The **average clustering coefficient** for a graph describes **how often there are "triangles"** in the graph. Where A_i, is the set of nodes i is connected to and k_i is the size of that set, it is defined as:

$$\bar{C} = \frac{1}{n} \sum_{each\,node} \frac{2(number\ of\ \text{triangles})}{number\ of\ \text{possible triangles}}$$

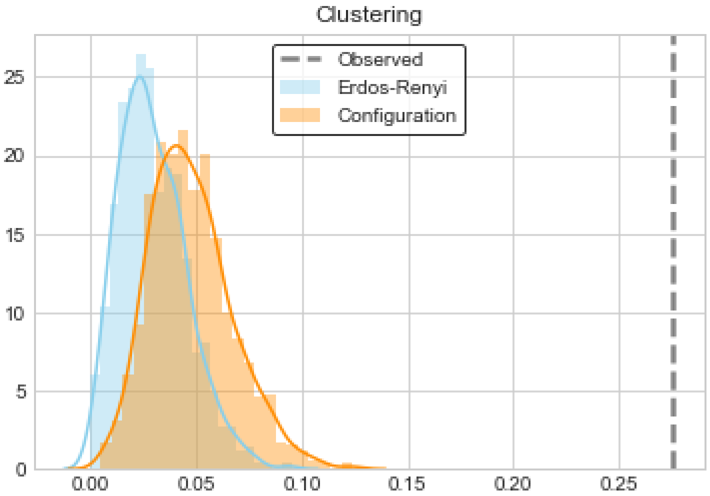
More than by chance?

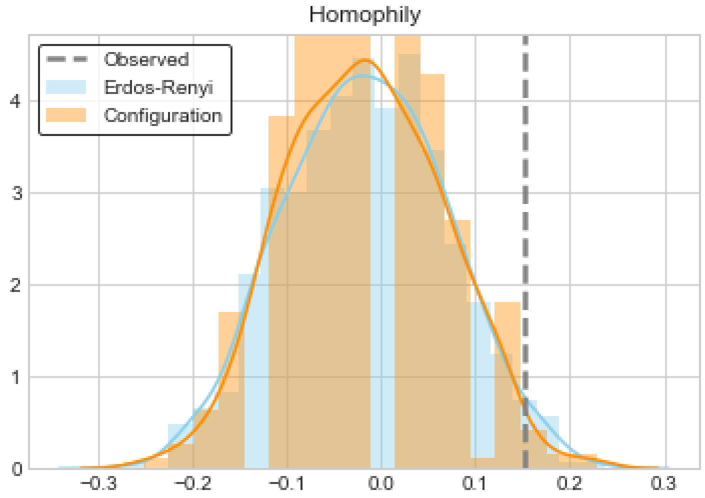
- Erdös–Rényi model holds constant
  - Number of nodes
  - Average degree
- Configuration model holds constant
  - Number of nodes
  - Exact degree distribution
  - Correlation between attributes and degree

Testing network-level statistics

- Simulate the network under the null model times
- For each simulation, measure the test statistic
- Compare your observed statistics to this distribution
- How likely of a draw is it? That’s your p-value!

What do we observe in our network?





Erdös–Rényi: $p \approx 0$

Configuration: $p \approx 0$

Erdös–Rényi: $p \approx 0.054$

Configuration: $p \approx 0.038$

Thank you!

PERSONAL WEBSITE:

HTTPS://ANKENYAV.WIXSITE.COM/AUSTINVANLOON

EMAIL:

AVANLOON@STANFORD.EDU

Austin van Loon