

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 *matplotlib*.

```
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 ([Katarina 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
 - Erdős–Rényi model
 - Configuration model



Outline

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



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!



Outline

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

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

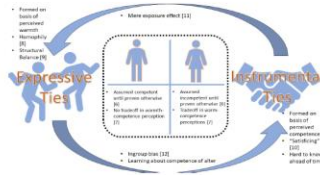
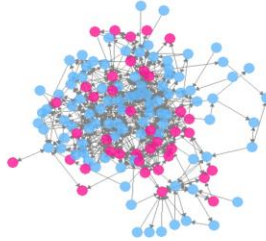


Figure 1: Theoretical Model of Gender and Multiplex Communication in Organizations

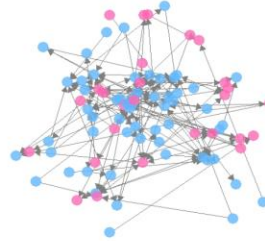
General Procedure

- 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

Instrumental Communication Network



Expressive Communication Network



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:

$$x_i = \frac{1}{\lambda} \sum_{j \in G} a_{ij} x_j \quad (1)$$

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

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

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 :

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

Position Across Graphs

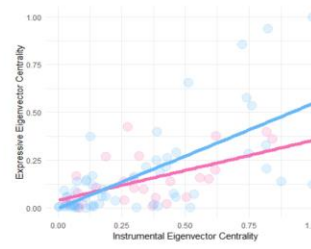


Figure 2: The relationship between network centralities by gender

Clustering

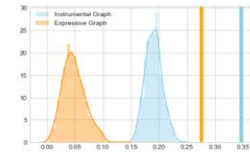


Figure 3: Histogram of clustering over 1,000 randomly re-wired networks and observed values for both networks.

Homophily

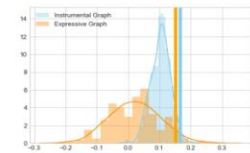


Figure 4: Histogram of homophily over 1,000 randomly re-wired networks and observed values for both networks.

References

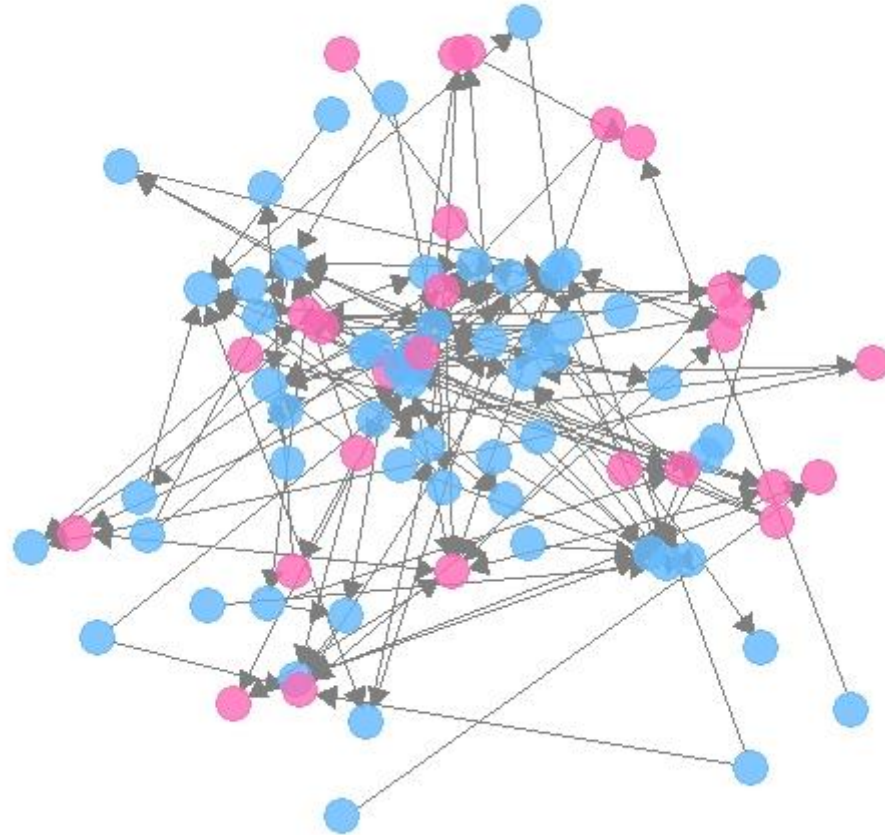
- [1] M. Granovetter, *Economic action and social structure: The problem of embeddedness*, *American Journal of Sociology*, 93(3):1361-1380, 1985.
- [2] A. Starling and R. Dierker, *Once in the field: Organizational success, failure, and the role of a leader*, *Academy of Management Review*, 34(1):114-131, 2009.
- [3] M. Dahab, *Men, Women, and the Gender Gap*, *John Wiley and Sons*, 2010.
- [4] P. Blau, *The Dynamics of Homophily*, *Journal of Sociology*, 10(1):1-11, 1974.
- [5] T. Freeman and J. Stawert, *The opportunity structure for discrimination*, *Journal of Sociology*, 10(1):1-11, 1974.
- [6] J. Smith-Lewis, *Men, Women, and the Gender Gap*, *Academy of Management Review*, 34(1):114-131, 2009.
- [7] R. Dierker, *Men, Women, and the Gender Gap*, *Academy of Management Review*, 34(1):114-131, 2009.
- [8] D. Cartwright and J. Harary, *Structural balance: A generalization of the Heider model*, *The Psychological Review*, 63(3):277-301, 1956.
- [9] H. Simon, *Structural balance and the structure of the network*, *Journal of Sociology*, 10(1):1-11, 1974.



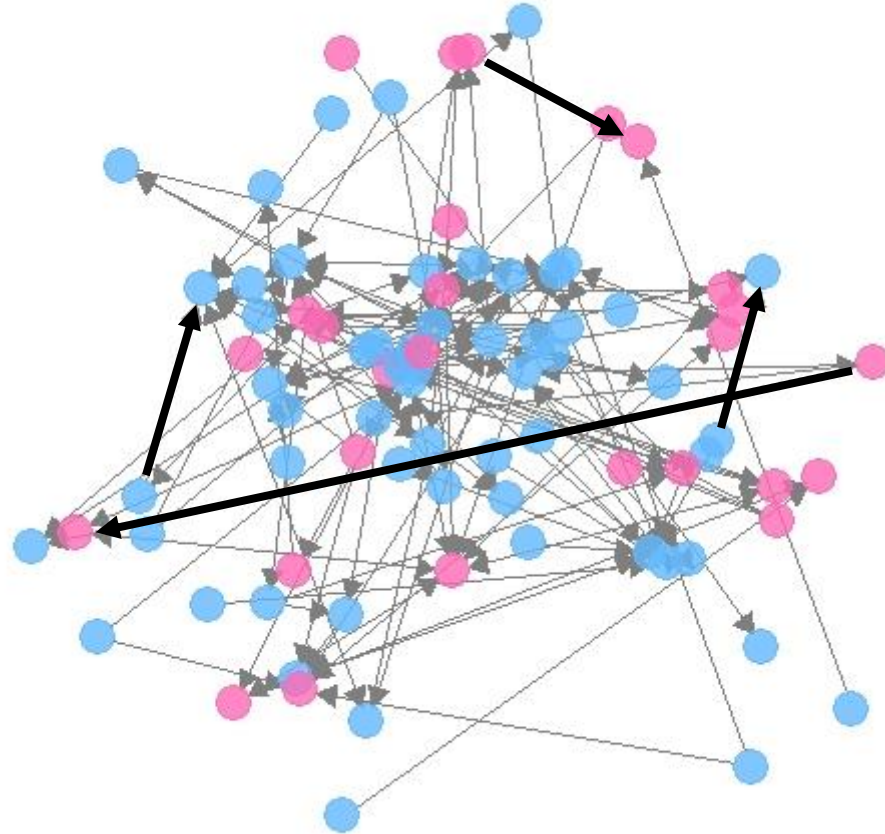
Outline

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

Expressive Communication Network



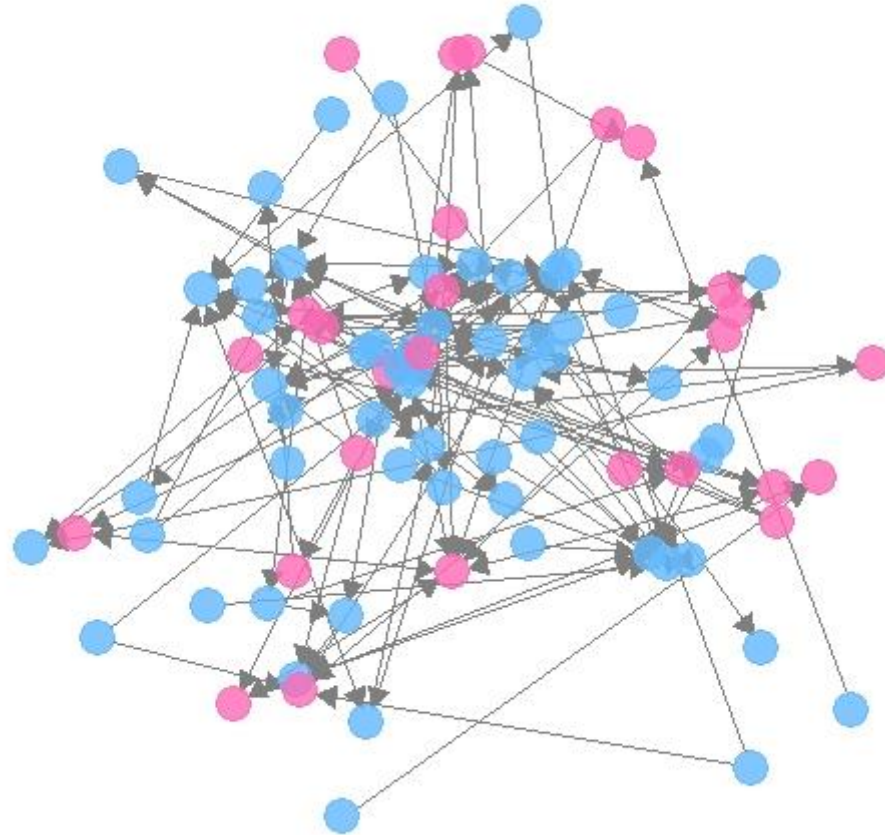
Expressive Communication Network



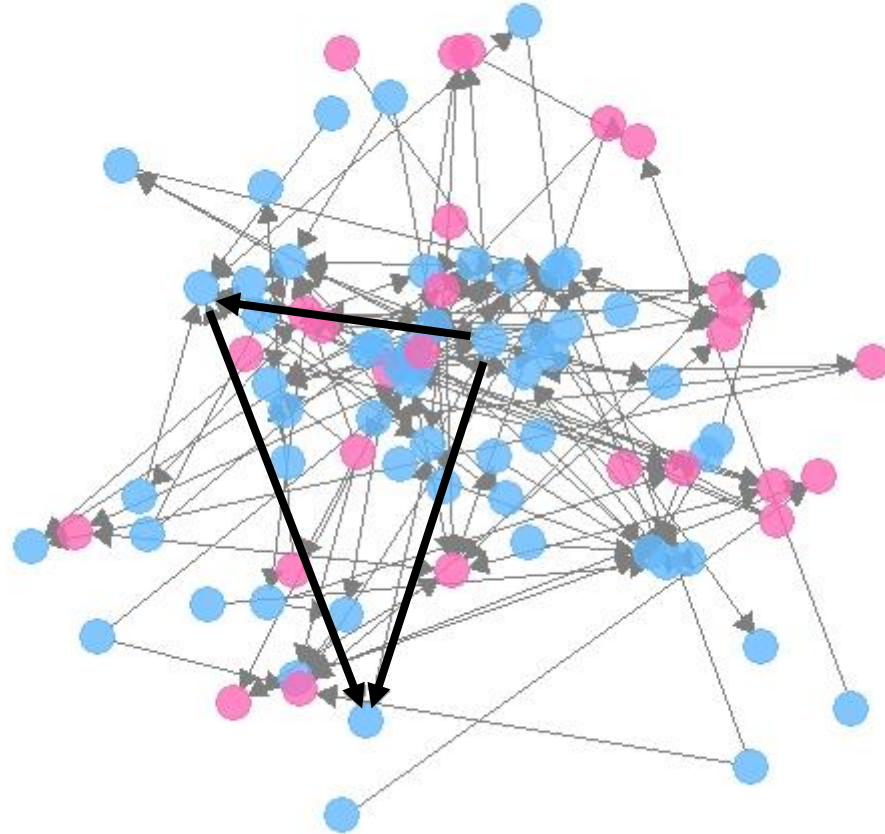
The **assortativity coefficient** of a graph is the degree to which "**birds of feather flock together**". Where e is a matrix for which entry e_{ij} is the fraction of edges in a graph that connect actors of type i and j , it is defined as:

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

Expressive Communication Network



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:

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

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:

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



Outline

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



More than by chance?

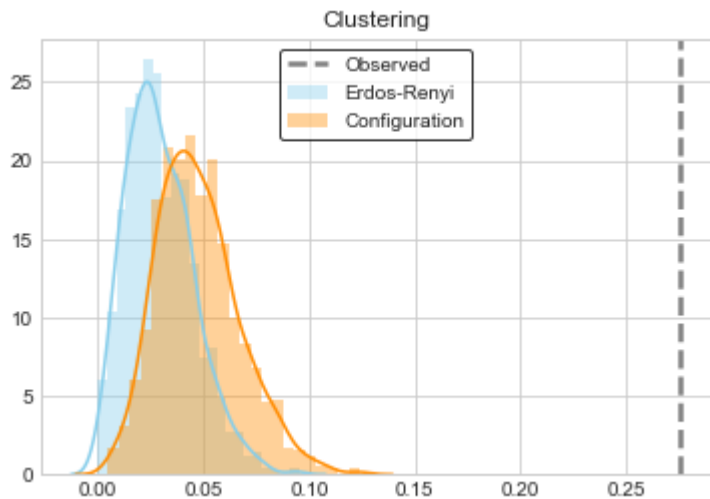
- Erdős–Rényi model holds constant
 - › Number of nodes
 - › Average degree
- Configuration model holds constant
 - › Number of nodes
 - › Exact degree distribution
 - › Correlation between attributes and degree



Testing network-level statistics

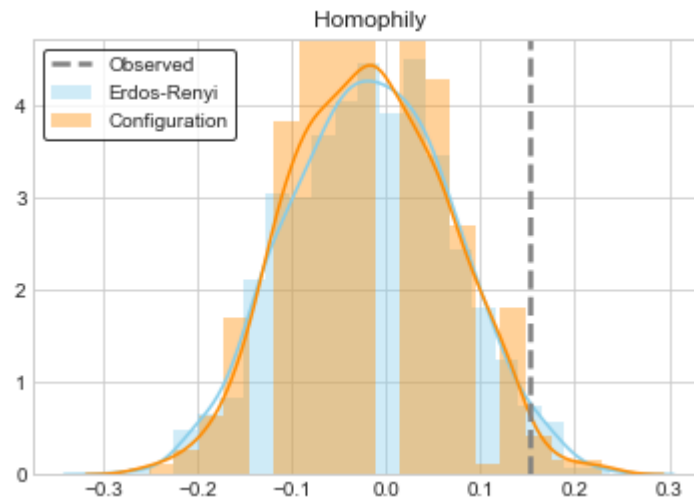
- Simulate the network under the null model many times
- For each simulation, measure the test statistic
- Compare your observed statistics to this distribution
- How likely of a draw is it? That's your p-value!

What do we observe in our network?



Erdős–Rényi: $p \approx 0$

Configuration: $p \approx 0$



Erdős–Rényi: $p \approx 0.054$

Configuration: $p \approx 0.038$

Thank you!

PERSONAL WEBSITE:

[HTTPS://ANKENYAV.WIXSITE.COM/AUSTINVANLOON](https://ankenya.v.wixsite.com/austinvanloon)

EMAIL:

AVANLOON@STANFORD.EDU

Austin van Loon

