


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](#) [Comment](#) [Delete](#)

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

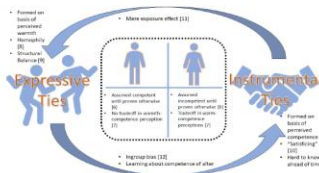
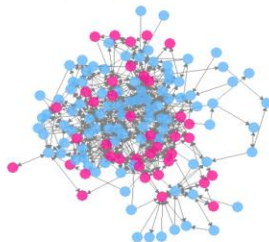


Figure 1: Theoretical Model of Gender and Multiplex Communication in Organizations

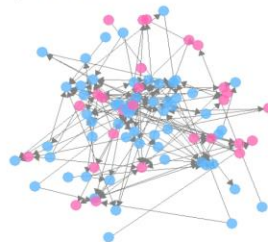
General Procedure

- 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:

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

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

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

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 :

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

Position Across Graphs

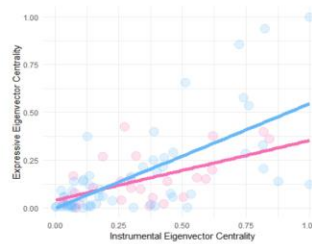


Figure 2: The relationship between network centralities by gender

Clustering

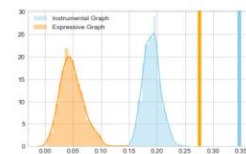


Figure 3: Histogram of clustering over 1,000 randomly re-wired networks and observed values for both networks.

Homophily

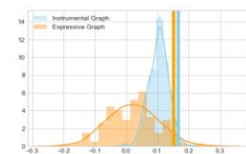


Figure 4: Histogram of homophily over 1,000 randomly re-wired networks and observed values for both networks.

References

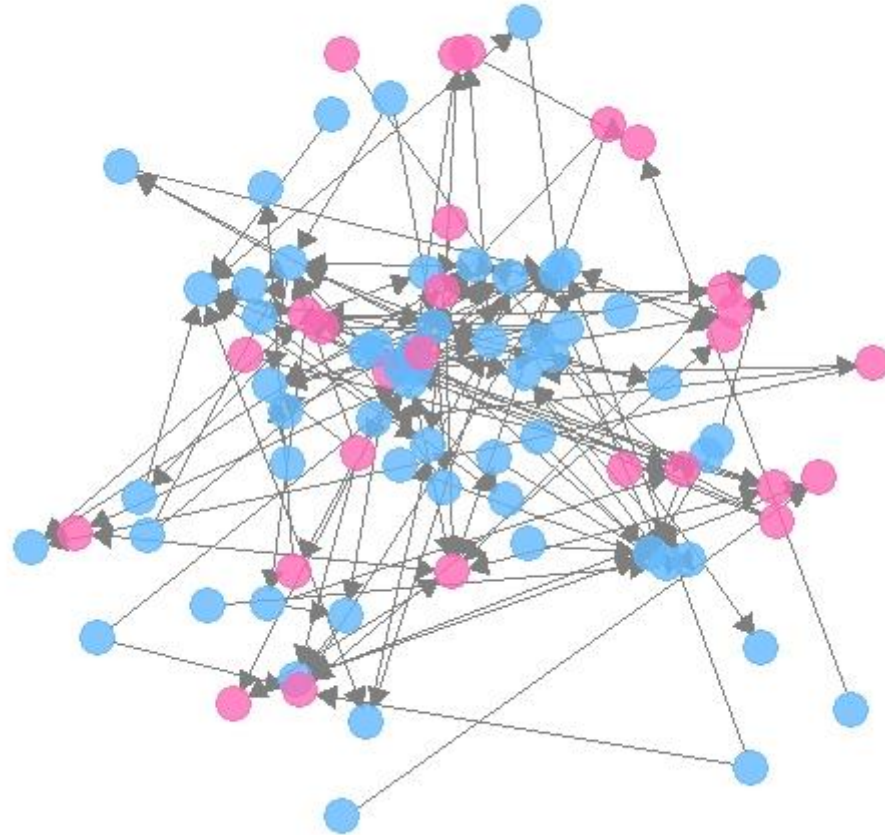
- | | | | | |
|--|---|--|---|--|
| [1] M. Economou et al., <i>Economic action and social interaction</i> , <i>Journal of Economic Psychology</i> , vol. 23, no. 1, pp. 105–118, 2002. | [12] J. S. Brehm and J. R. Burnstein, <i>On the face of the matter</i> , <i>Journal of Experimental Social Psychology</i> , vol. 31, pp. 101–114, 1995. | [13] L. S. Oishi, L. M. D. Brown, and M. P. Zanna, <i>Effects of a social desirability feedback</i> , <i>Journal of Experimental Social Psychology</i> , vol. 31, pp. 115–124, 1995. | [14] B. Zanna, <i>When we see others</i> , <i>Journal of Experimental Social Psychology</i> , vol. 31, pp. 125–134, 1995. | [15] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 529–534, 2002. |
| [16] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 535–540, 2002. | [17] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 541–546, 2002. | [18] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 547–552, 2002. | [19] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 553–558, 2002. | [20] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 559–564, 2002. |
| [21] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 565–570, 2002. | [22] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 571–576, 2002. | [23] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 577–582, 2002. | [24] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 583–588, 2002. | [25] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 589–594, 2002. |
| [26] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 595–600, 2002. | [27] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 601–606, 2002. | [28] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 607–612, 2002. | [29] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 613–618, 2002. | [30] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 619–624, 2002. |
| [31] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 625–630, 2002. | [32] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 631–636, 2002. | [33] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 637–642, 2002. | [34] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 643–648, 2002. | [35] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 649–654, 2002. |
| [36] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 655–660, 2002. | [37] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 661–666, 2002. | [38] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 667–672, 2002. | [39] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 673–678, 2002. | [40] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 679–684, 2002. |
| [41] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 685–690, 2002. | [42] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 691–696, 2002. | [43] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 697–702, 2002. | [44] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 703–708, 2002. | [45] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 709–714, 2002. |
| [46] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 715–720, 2002. | [47] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 721–726, 2002. | [48] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 727–732, 2002. | [49] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 733–738, 2002. | [50] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 739–744, 2002. |
| [51] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 745–750, 2002. | [52] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 751–756, 2002. | [53] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 757–762, 2002. | [54] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 763–768, 2002. | [55] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 769–774, 2002. |
| [56] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 775–780, 2002. | [57] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 781–786, 2002. | [58] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 787–792, 2002. | [59] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 793–798, 2002. | [60] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 799–804, 2002. |
| [61] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 805–810, 2002. | [62] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 811–816, 2002. | [63] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 817–822, 2002. | [64] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 823–828, 2002. | [65] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 829–834, 2002. |
| [66] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 835–840, 2002. | [67] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 841–846, 2002. | [68] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 847–852, 2002. | [69] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 853–858, 2002. | [70] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 859–864, 2002. |
| [71] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 865–870, 2002. | [72] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 871–876, 2002. | [73] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 877–882, 2002. | [74] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 883–888, 2002. | [75] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 889–894, 2002. |
| [76] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 895–900, 2002. | [77] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 901–906, 2002. | [78] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 907–912, 2002. | [79] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 913–918, 2002. | [80] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 919–924, 2002. |
| [81] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 925–930, 2002. | [82] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 931–936, 2002. | [83] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 937–942, 2002. | [84] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 943–948, 2002. | [85] J. P. Hain, <i>Psychological Science</i> , vol. 13, no. 6, pp. 949–954, 2002. |
| [86] J. P. Hain, <i>Psychological Science</i> , vol | | | | |



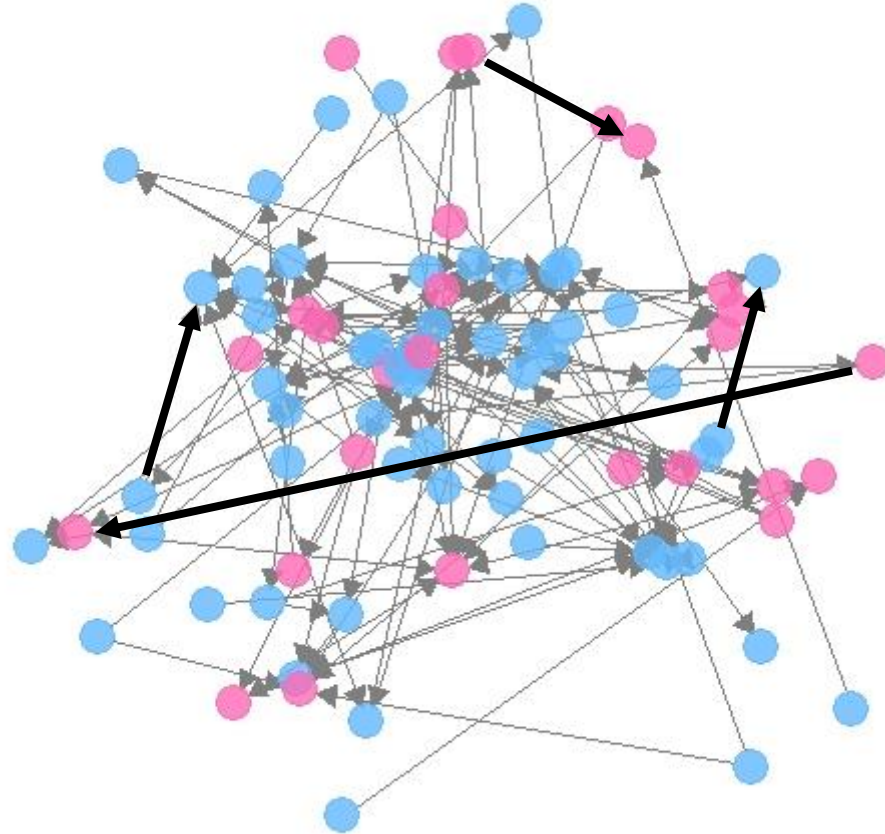
Outline

- 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



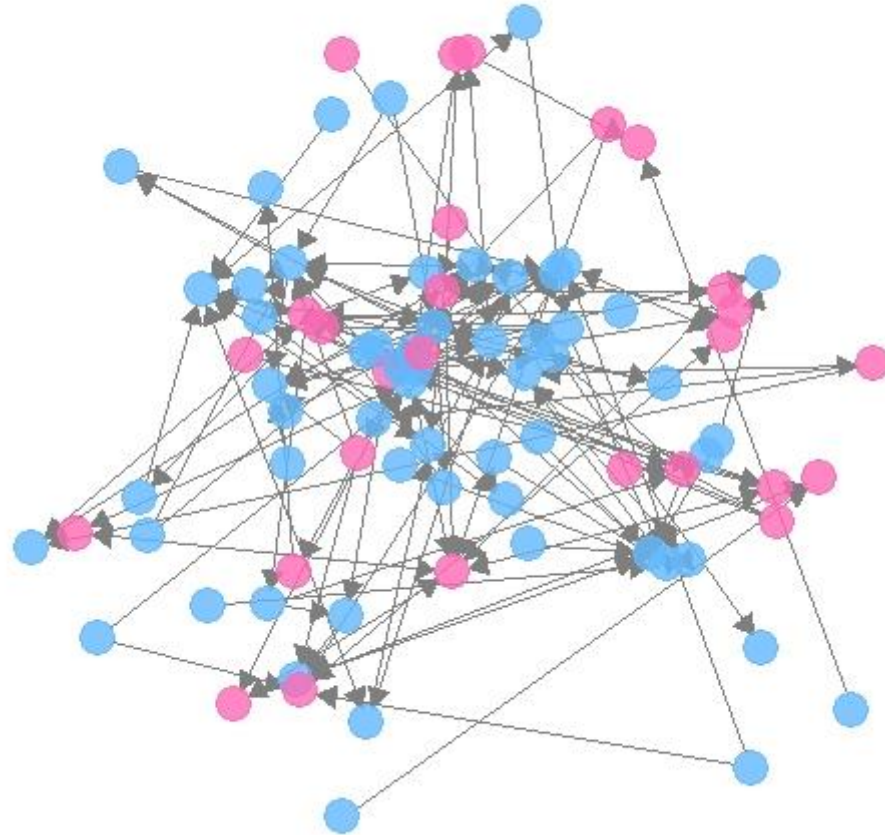
Expressive Communication Network



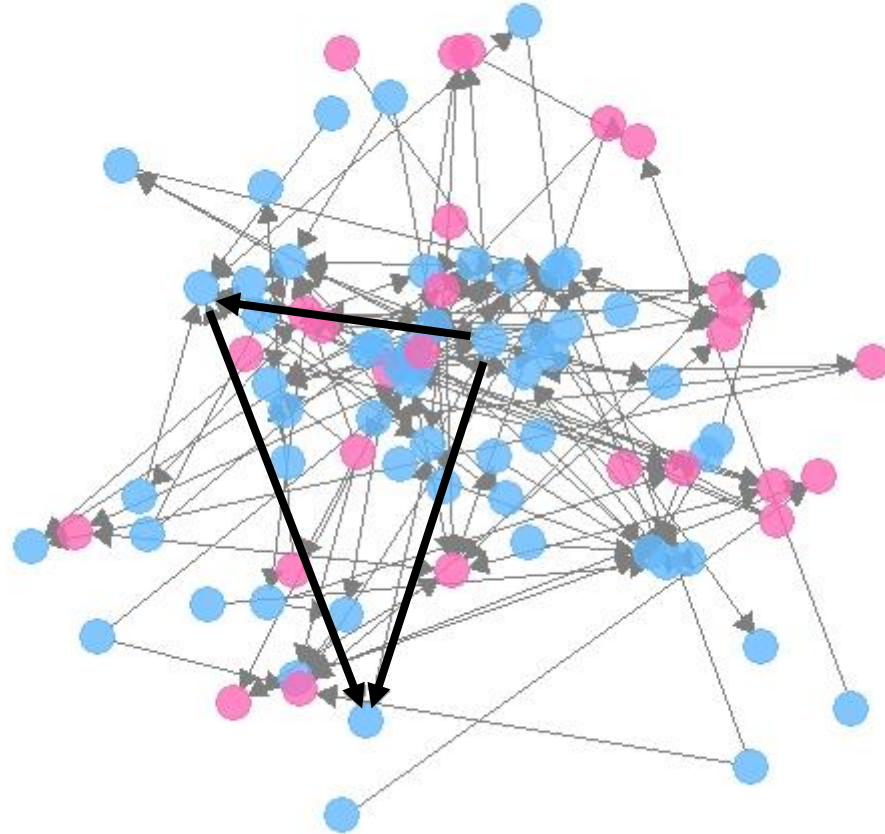
The **assortativity coefficient** of a graph is the degree to which "**birds of feather flock together**". Where e is a matrix for which entry e_{ij} is the fraction of edges in a graph that connect actors of type i and j , it is defined as:

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

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?

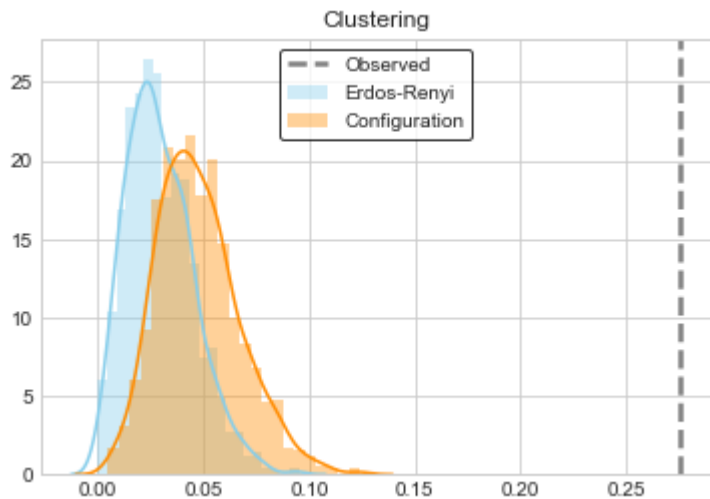
- Erdős–Rényi model holds constant
 - › Number of nodes
 - › Average degree
- Configuration model holds constant
 - › Number of nodes
 - › Exact degree distribution
 - › Correlation between attributes and degree



Testing network-level statistics

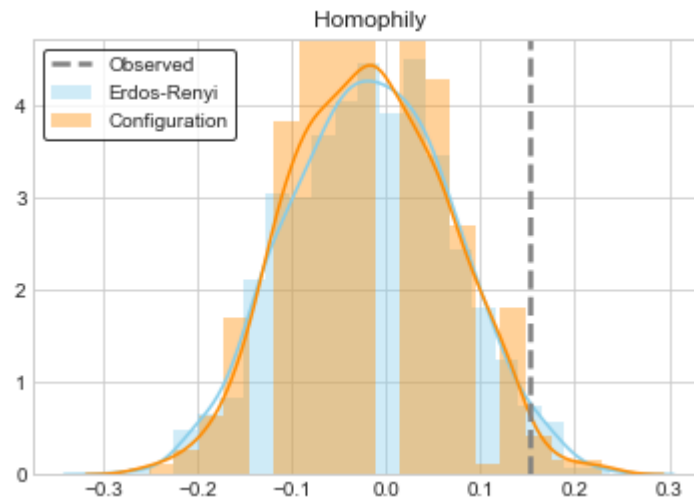
- Simulate the network under the null model many times
- For each simulation, measure the test statistic
- Compare your observed statistics to this distribution
- How likely of a draw is it? That's your p-value!

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.wixsite.com/austinvanloon)

EMAIL:

AVANLOON@STANFORD.EDU

Austin van Loon