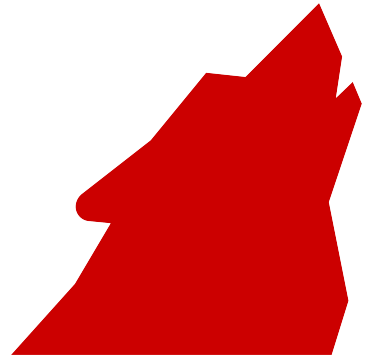
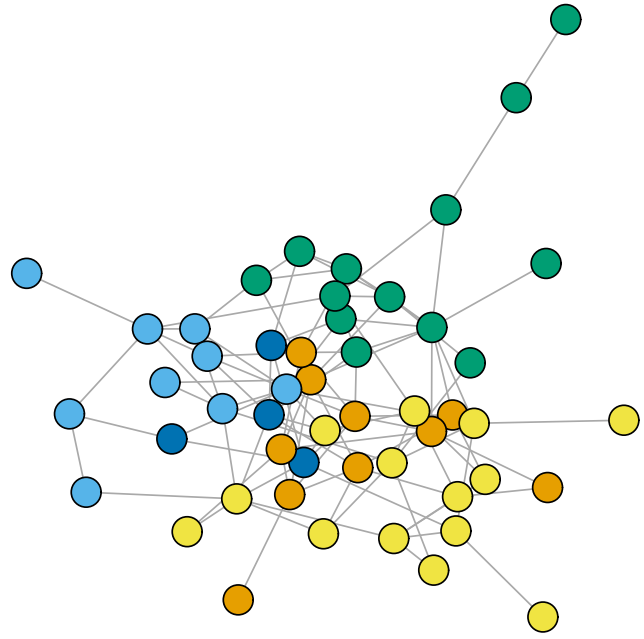


A model-agnostic hypothesis test for community structure and homophily in networks

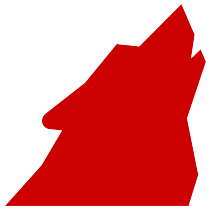
Eric Yanchenko and Srijan Sengupta
North Carolina State University
JSM 2021

Networks are everywhere!

- Networks are ubiquitous today
  - social media, infrastructure, biology, epidemiology...
- Community structure
- But are these communities meaningful?



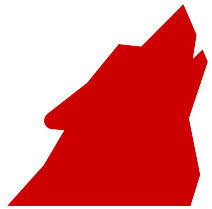
Community vs. Homophily



- Defining “community" is quite difficult
- Closely related is *homophily*: more likely to be an edge between nodes in the same community as compared to nodes in different communities
- Hypothesis test is:

$$H_0\text{: no homophily vs. } H_1\text{: homophily}$$

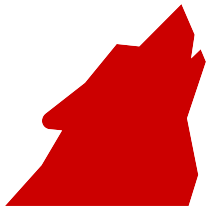
Outline



- Formal homophily definition
  - Nominal, collateral, intrinsic
- Test statistic and hypothesis test
- Simulations
- Real-data analysis

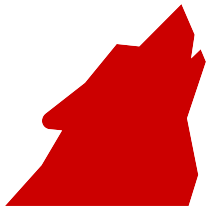
Homophily

Notation



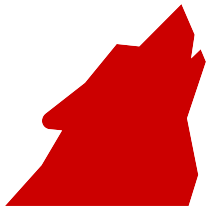
- Unweighted, undirected, no selfloops
- A is $n \times n$ adjacency matrix
  - $A_{ij} = 1$ if nodes i and j have an edge, and 0 otherwise
- $A \sim P$ means $A_{ij} \sim \text{Bernoulli}(P_{ij})$
- Community "assignment" vector $\boldsymbol{c}$
  - length n and $c_i = u \in \{1, \ldots, K\}$ means node i in community u

Homophily Parameter



- Formal, model-agnostic definition of homophily
- Probability of an edge *within* communities
  - $\bar{p}_{in}(\boldsymbol{c}) = \frac{1}{\sum_{k=1}^{K} \binom{n_k}{2}} \sum_{u=1}^{K} \sum_{i>j:c_i=c_j=u} P_{ij}$
- Probability of an edge *between* communities
  - $\bar{p}_{out}(\boldsymbol{c}) = \frac{1}{\sum_{k>l} n_k n_l} \sum_{u>v} \sum_{i>j,c_i=u,c_j=v} P_{ij}$
- Overall probability of an edge
  - $\bar{p} = \frac{1}{\binom{n}{2}} \sum_{i>j} P_{ij}$

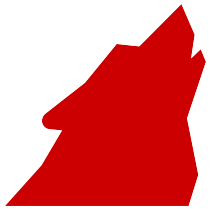
Homophily Parameter



$$\gamma(\boldsymbol{c}, P) := \frac{\bar{p}_{in}(\boldsymbol{c}) - \bar{p}_{out}(\boldsymbol{c})}{\bar{p}}$$

- Larger if intra-community edge is more likely than inter-community edge
- Scaled by overall sparsity of network
- Depends on community assignments $\boldsymbol{c}$ and generating matrix P
- Truly general (model agnostic)

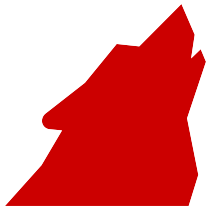
Hypothesis Test



- Labeled vs. unlabeled networks
- Unlabeled network has no other node information
  - co-appearances, friend groups, etc.
- Determine if *any* community assignment has homophily
- Define $\tilde{\gamma}(P) = \max_{\boldsymbol{c}} \gamma(\boldsymbol{c}, P)$ so the test is

$$H_0 : \tilde{\gamma}(P) \leq \gamma_0 \text{ vs. } H_1 : \tilde{\gamma}(P) > \gamma_0$$

Hypothesis Test



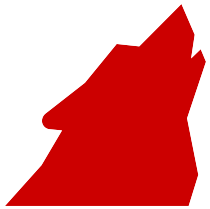
- How do we choose threshold γ_0?
- First principles implies $\gamma_0 = 0$
- Issue: this forces null hypothesis to be Erdös-Rényi (ER) model with $P_{ij} = p$ for all i, j

Lemma

$\gamma(\mathbf{c}, P) \leq 0$ *for all* $\mathbf{c}$ *if and only if* P *is from a homogenous Erdös-Rényi model.*

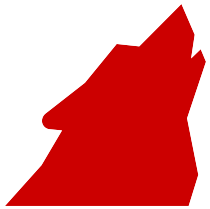
- $\gamma_0 = 0$ is equivalent to H_0: ER vs. H_1: not ER
- Implies only the ER model lacks homophily
- Problematic!

Nominal, Collateral and Intrinsic Homophily



- A network model has *nominal homophily* if $\gamma(\boldsymbol{c}, P) > 0$.
- A network model has *collateral homophily* if $\gamma(\boldsymbol{c}, P) > 0$ caused by other network feature (degree heterogeneity, transitivity, etc.).
- A network model has *intrinsic homophily* if homophily parameter is larger than what could be caused simply by some other network feature.
- Choose γ_0 as largest possible value arising from collateral homophily.

Collateral Homophily Toy Example



- Consder Chung-Lu model where $P_{ij} = \theta_i \theta_j$ for some parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$.
- Let $n = 4$ and $\boldsymbol{\theta} = (0.6, 0.7, 0.8, 0.9)$.

$$P = \begin{pmatrix} 0 & 0.42 & 0.48 & 0.54 \\ 0.42 & 0 & 0.56 & 0.63 \\ 0.48 & 0.56 & 0 & 0.72 \\ 0.54 & 0.63 & 0.72 & 0 \end{pmatrix}$$

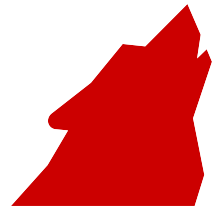
- If $\boldsymbol{c} = (1, 1, 2, 2)$, then $\gamma(\boldsymbol{c}, P) = 0.03 > 0$.
- Set $\gamma_0 = 0.03$.

Test Statistic

Test Statistic

- Formal, model-agnostic definition of homophily
- Estimated probability of an edge *within* communities
  - $\hat{p}_{in}(\boldsymbol{c}) = \frac{1}{\sum_{k=1}^{K} \binom{n_k}{2}} \sum_{u=1}^{K} \sum_{i>j:c_i=c_j=u} A_{ij}$
- Estimated probability of an edge *between* communities
  - $\hat{p}_{out}(\boldsymbol{c}) = \frac{1}{\sum_{k>l} n_k n_l} \sum_{u>v} \sum_{i>j,c_i=u,c_j=v} A_{ij}$
- Estimated overall probability of an edge
  - $\hat{p} = \frac{1}{\binom{n}{2}} \sum_{i>j} A_{ij}$

Test Statistic



- Similar to homophily parameter:

$$T(\boldsymbol{c}, A) = \frac{\hat{p}_{in}(\boldsymbol{c}) - \hat{p}_{out}(\boldsymbol{c})}{\hat{p}}.$$

- Reject H_0 for

$$T(A) = \max_{\boldsymbol{c}} T(\boldsymbol{c}, A) > C.$$

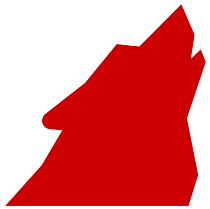
- C depends on network size and null model
- Analytical form of C is difficult!
  - Max of dependent random variables
- Instead use bootstrap

Hypothesis Testing Procedure

Result: p-value for testing nominal homophily
Input: $n \times n$ adjacency matrix A, number of iterations B;
Compute $\tilde{T}_{obs} = \max_{\boldsymbol{c}}\{T(\boldsymbol{c}, A)\}$ and $\hat{p} = \hat{p}_{obs} = \sum_{i,j} A_{ij}/(n(n-1))$;
for *B times* **do**
  $A_i^* \leftarrow$ ER network with $\hat{p}$;
  Compute $\tilde{T}_i^* = \max_{\boldsymbol{c}}\{T(\boldsymbol{c}, A_i^*)\}$;
end
p-val$= \#(\tilde{T}_i^* \geq \tilde{T}_{obs})/B$

Algorithm 1: Testing no homophily vs. nominal homophily

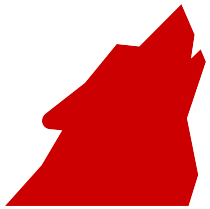
Method Advantages



- Rooted in formal, model-agnostic homophily definition
- Presents homophily as a continuum
- Interpretable test statistic
- Flexibility in null model
- Richer insights into networks

Theoretical Results

Consistency of Test



For some small $\epsilon > 0$, consider the threshold

$$C = \sqrt{\frac{2\{\log(2N_{n,K}) - \log \alpha\}}{n^2}} \cdot \frac{1+\varepsilon}{\hat{p}},$$

where $N_{n,K} = \binom{n-1}{K-1} \leq K^n$ and $K = \#$ of communities.
When the null hypothesis is true, for any $\eta > 0$ we have

$$\lim_{n\to\infty} P(\tilde{T}(A) > C) \leq \alpha + \eta.$$

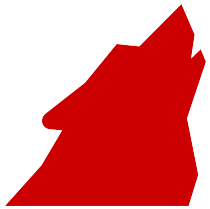
When the alternative hypothesis is true, i.e. intrinsic homophily, then for any $\eta > 0$ we have

$$\lim_{n\to\infty} P(\tilde{T}(A) > C) > 1 - \eta.$$

See paper for assumptions, details and proof.

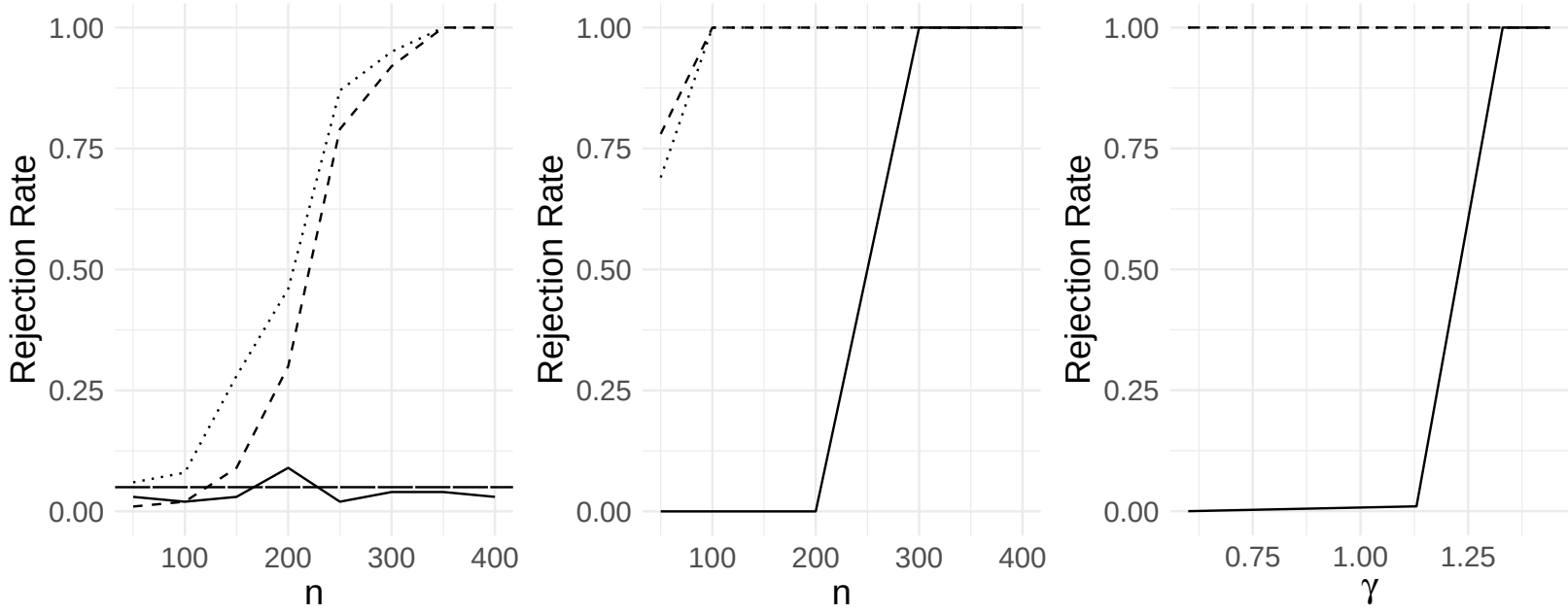
Simulations

Settings



- Compared with Spectral method of Bickel and Sarkar (2016)
- Considered rejection rate of test under:
  - Null model
  - Alternative model, fixed γ and increasing n
  - Alternative model, increasing γ and fixed n
- $B = 200$ bootstrap samples
- 100 Monte Carlo replications
- Walktrap method Pons and Latapy (2005) for communities

Collateral vs. intrinsic homophily

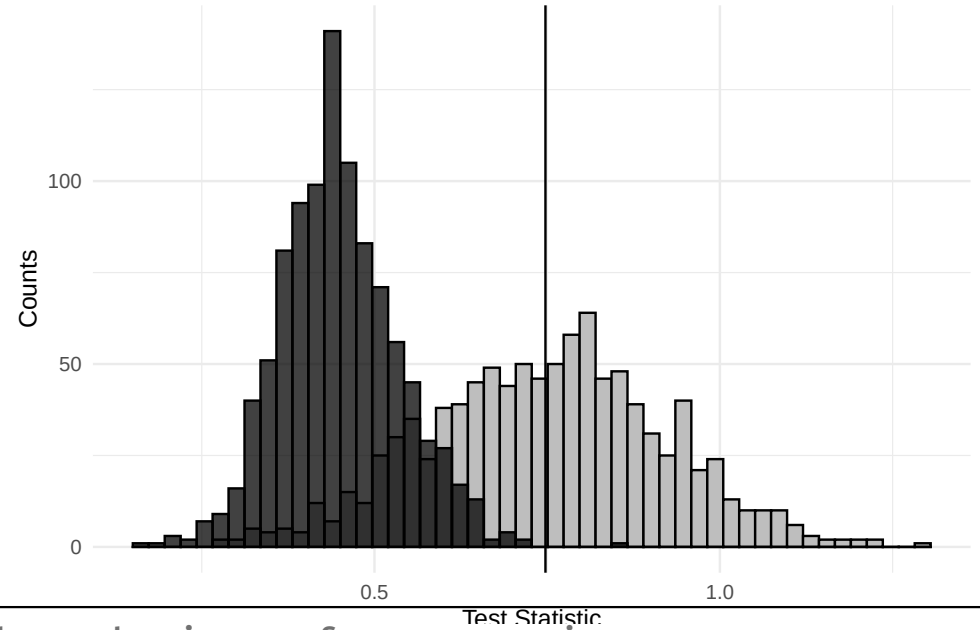


Chung-Lu (CL) vs. degree corrected stochastic block model (DCSBM). (- - - - -) Spectral, (·····) Spectral Adjusted, (——) Bootstrap and horizontal line at $\alpha = 0.05$ level of the test.

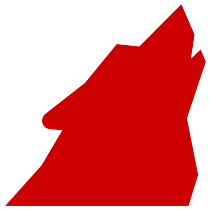
Real Data Analysis

Hospital Encounters

- Vanhems (2013) interactions between patients and workers in the geriatric unit of a French hospital
- $n = 75$ nodes representing people
- 2,278 edges representing their interactions
- p-values: < 0.001 Spectral adjusted; 0.001 bootstrap (ER); 0.512 bootstrap (CL)



Conclusions



- Novel and general method for testing homophily in networks
- Gives richer insights into real world data
- Only proved for ER and not with bootstrap
- Does not allow mixed membership
- Could be extended to other network features like small-world property

`https://arxiv.org/pdf/2107.06093.pdf`