A model-agnostic hypothesis test for community structure in networks

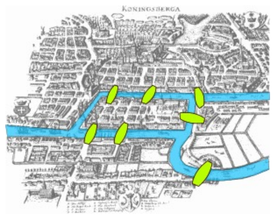
Eric Yanchenko and Srijan Sengupta

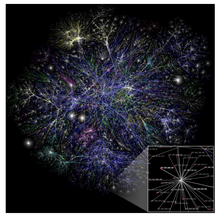
Department of Statistics, North Carolina State University

Motivation

Networks

- Networks are everywhere!
  - Collection of nodes and edges
  - Social, infrastructure, epidemiology,…

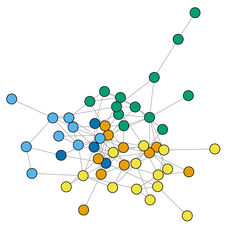




Euler's Sever Bridge of Konigsberg (1736) and Internet (2005, The Opte Project), two examples of networks

Network Properties

- Community structure
- Is this structure meaningful?



Network generated without community structure, but spectral clustering still returns communities.

Community vs. Homophily

- Defining "community" is quite difficult
- Homophily: more likely to be an edge between nodes in same community as compared to nodes in different communities
- Hypothesis test is:

$$H_0: no\ homophily\ vs.\ H_1: homophily$$

Methods

Notation

- Unweighted, undirected, no self-loops
- A is $n \times n$ adjacency matrix
  - $A_{ij} = 1$ if edge between nodes i and j and 0 otherwise
- $A \sim P$ means $A_{ij} \sim Bernoulli(P_{ij})$
- Community "assignment" vector c
  - Length n and $c_i = u$ means node i in community u for $u = 1, \dots, K$

Homophily Parameter

- Formal, model-agnostic homophily metric
- Probability of an edge *within* communities, $\overline{p}_{in}(c)$, *between* communities, $\overline{p}_{out}(c)$, and overall, $\overline{p}$

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

- Larger if intra-community edge more likely than inter
- Depends on community assignments and P

Hypothesis Test

- Labeled vs. unlabeled networks

$$H_0: max_c \gamma(c, p) \le \gamma_0\ vs.\ H_1: max_c \gamma(c, p) > \gamma_0$$

- How to chose γ_0? Should we chose 0?
- *Lemma: $\gamma(c, P) \le 0$ for all c if and only if P is from an Erdos-Renyi (ER) model ($P_{ij} = p$ for all i, j)*
- Implies only ER model lack homophily: problematic!

Nominal, Collateral and Intrinsic Homophily

- A network has *nominal homophily* if $\gamma(c, P) > 0$
- A network has *collateral homophily* $\gamma(c, P) > 0$ caused by some other network feature (degree heterogeneity)
- A network has *intrinsic homophily* if the homophily parameter is larger than could be caused by another network feature
- Choose γ_0 as largest value arising from collateral hom.
- Toy example: Consider Chung-Lu model ($P_{ij} = \theta_i \theta_j$)

$$P = \begin{pmatrix} - & .42 & .48 & .54 \\ - & - & .56 & .63 \\ - & - & - & .72 \\ - & - & - & - \end{pmatrix}$$

- If $c = (1, 1, 2, 2)$, $\gamma(c, P) = 0.03$.

Test Statistic

- *Estimated* probability of an edge within communities, $\hat{p}_{in}(c)$, between communities, $\hat{p}_{out}(c)$, and overall, $\hat{p}$

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

- Reject H_0 for

$$T(A) = max_c T(c, A) > C$$

- C depends on null model; analytic form is difficult!

Algorithm

Algorithm 1 Testing no homophily vs. nominal homophily

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$= \#(\tilde{T}_i^* \ge \tilde{T}_{obs})/B$

Theoretical Results

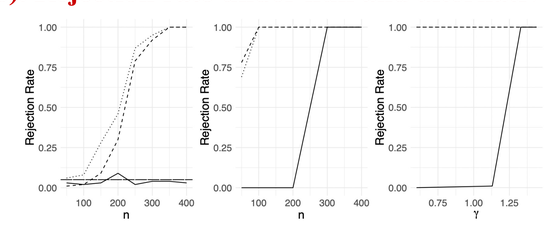
Hypothesis Test

- Convergence of test statistic: the test statistic $T(c, A)$ converges in probability to $\gamma(c, P)$ for all c
- Consistency of asymptotic test:
  - Under ER null, the rejection rate converges to Type I error rate
  - Under intrinsic homophily alternative, the rejection rate converges to 1

Data Analysis

Simulations

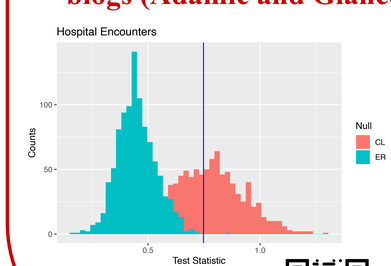
- Compare to Spectral method of Bickel and Sarkar (2016). Rejection rate under null and alternative hyps.

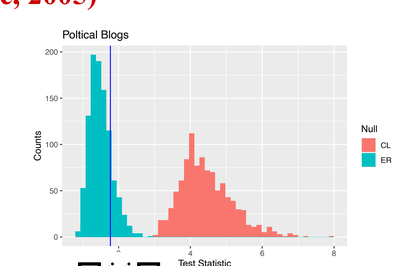


Chung-Lu (CL) vs. degree corrected stochastic block model (DCSBM). (- - - - -) Spectral, (·····) Spectral Adjusted, (———) Bootstrap and horizontal line at $\alpha = 0.05$ level of the test.

Real-World Data

- Hospital encounters (Vanhems et. al, 2013) and political blogs (Adamic and Glance, 2005)





Paper

Website