A generalized hypothesis test for community structure and homophily in networks

Eric Yanchenko and Srijan Sengupta
North Carolina State University

Abstract

Networks continue to be of great interest to statisticians, with an emphasis on community detection. Less work, however, has addressed this question: given some network, does it exhibit meaningful community structure? We propose to answer this question in a principled manner by framing it as a statistical hypothesis in terms of a formal and general homophily metric. Homophily is a well-studied network property where intra-community edges are more likely than between-community edges. We use the homophily metric to identify and distinguish between three concepts: nominal, collateral, and intrinsic homophily. We propose a simple and interpretable test statistic leveraging this homophily parameter and formulate both asymptotic and bootstrap-based rejection thresholds. We prove its asymptotic properties and demonstrate it outperforms benchmark methods on both simulated and real world data. Furthermore, the proposed method yields rich, provocative insights on four classic data sets; namely, that many well-studied networks do not actually have intrinsic homophily.

Keywords: bootstrap; community detection; community structure; homophily; hypothesis testing; networks; random graphs

1 Introduction

Networks are ubiquitous in the world we live in. From social media (Kane et al., 2014; Guo et al., 2020) to infrastructure (Pagani and Aiello, 2013) to biology (Mason and Verwoerd, 2007) to epidemiology (Leitch et al., 2019), many fields today are gathering and analyzing network data. Community structure is one of the most well-studied properties of such networks because nodes in the same community might share similar characteristics which can reveal details about the underlying data-generating process. A litany of different community detection methods exist, from spectral-based methods (Ng et al., 2002; Rohe et al., 2011; Jin, 2015; Sengupta and Chen, 2015), to greedy algorithms (Blondel et al., 2008) to random walk approaches (Pons and Latapy, 2005). While many approaches exist to find communities, there has been far less work studying whether a particular network even has meaningful community structure. In other words, are the community assignments returned by these algorithms informative about the underlying data-generating mechanism of the network, or are they simply some artifacts in the data?

As a motivating example, consider the network in Figure 1 which was generated from an Erdős-Rényi model (Erdős and Rényi, 1959) with \( n = 50 \) and \( p = 0.10 \). The color of the nodes represents the clusters returned by the Spectral clustering algorithm (Ng et al., 2002). Since an Erdős-Rényi model generated this data, there is not true community structure in the network. Thus, the community memberships returned by the algorithm are not meaningful.
This can be problematic if a practitioner then looked for relationships between nodes in the same “community”, for example.

Before continuing, we note that the concept of “community” in networks is vague and does not have a universal definition. Fundamentally connected to communities, however, is the concept of homophily (or assortative mixing (Newman, 2002)), which means that there is more likely to be an edge between nodes in the same community as compared to nodes in different communities (McPherson et al., 2001). In fact, most of the literature implicitly uses this meaning in their definition of communities, e.g., Fortunato (2010).

The goal of this work, then, is to develop a statistical test that determines whether a given network has true homophily or not, i.e. for a given network we want to test:

\[ H_0 : \text{no homophily} \quad \text{vs.} \quad H_1 : \text{homophily}. \]  \hspace{1cm} (1)

Some of the earliest work on this problem comes from Lancichinetti et al. (2010). Their method treats the configuration model as the null model and uses extreme and order statistics for the number of internal edges using the hypergeometric distribution. Another notable work, based upon the extension of the configuration model to weighted graphs, the so-called continuous configuration model, is by Palowitch et al. (2018). They show that the test statistic, the sum of all weights on edges connecting a node with a community, follows an asymptotic normal distribution under the continuous configuration model. Another method is that of He et al. (2018), which uses the number of internal edges in a given community to calculate a tight upper bound on the probability of finding a “better” community under the configuration model. He et al. (2020) uses both the number of internal edges and the number of external edges in a given community to calculate the same probability, under the ER model. Li and Qi (2020) determine the asymptotic distribution of the modularity to compute their \( p \)-values. This method assumes “random” labeling as their null model.

The main approach that we will be comparing the proposed method to is from Bickel and Sarkar (2016). Considering Erdös-Rényi as the null model, they find the distribution of the principal eigenvalue of an appropriately centered and scaled adjacency matrix and compare it to a Tracy-Widom distribution to compute their \( p \)-value. They present two methods in their paper, one based on asymptotic results and one based on a bootstrap small-sample correction.
For the remainder of the paper, these methods will be referred to as the “Spectral” method and the “Spectral Adjusted” method, respectively.

While the above methods present some nice solutions to this problem, there still remain some critical gaps in the literature. First, each existing method assumes a specific random graph model, like the Erdős-Rényi or configuration model, as the null model. The presence or absence of community structure, however, is a network property which is independent from any specific random graph model. In effect, these testing methods are testing against the null hypothesis that the network is generated from a specific null model, rather than testing against the null hypothesis that there is no homophily. Similarly, none of the current methods give a formal definition of homophily. While each method implicitly uses the concept of homophily (e.g., the Spectral method effectively treats “not Erdős-Rényi” as homophily), the methods do not define an explicit, statistical metric to quantify the strength of homophily. Finally, most methods do not yield test statistics that can be naturally interpreted in terms of community structure. For example, the Spectral method uses the largest eigenvalue of the centered matrix, which does not have a clear interpretation in terms of homophily. This test statistic is more related to the departure of the network from the Erdős-Rényi model than to homophily.

In light of these gaps, the main contributions of this paper are the following. First, we present a formal and general definition of homophily which forms the basis of our statistical inference framework. By general we mean that while a model must ultimately be selected, the method works for nearly any choice. We leverage this homophily metric to identify and distinguish between three key concepts: nominal homophily, collateral homophily, and intrinsic homophily. Second, we propose an intuitive and interpretable test statistic which is directly connected to the homophily parameter. We use this test statistic to develop our hypothesis test using both asymptotic and bootstrap-based rejection thresholds. Third, we establish theoretical properties of the proposed test and investigate its empirical performance via simulations. Finally, we apply the proposed method to a number of network datasets that are well-studied in the community structure literature. The results are simultaneously provocative and insightful, as we conclude that several of these well-known network datasets do not actually have intrinsic homophily, and in addition, our method yields rich, new insights about the underlying network structure. Source code for this work is available on GitHub: https://github.com/eyanchenko/NetHypTest.

The roadmap for the rest of this paper is as follows: in Section 2 we present the homophily metric and formulate the testing method. In Section 3 we prove certain asymptotic consistency properties of the proposed testing procedure. In Section 4 we perform a thorough simulation study of the proposed method and compare it a current leading approach. In Section 5, we apply the method to four real world datasets to validate its performance while also showing how it can elucidate new information about these networks.
2 Methodology

2.1 Notation

For this work, we will only consider simple, unweighted, undirected networks with no self-loops. Consider a network with $n$ nodes. Let $A$ denote the $n \times n$ adjacency matrix where $A_{ij} = 1$ if node $i$ and node $j$ have an edge, and 0 otherwise. We write $A \sim P$ as shorthand for $A_{ij} \sim \text{Bernoulli}(P_{ij})$ for $1 \leq i < j \leq n$. We define a “community assignment” to be an integer-valued vector $c$ of length $n$, such that $c$ takes $K$ distinct values $\{1, 2, \ldots, K\}$, and $c_i = u$ denotes node $i$ is assigned to community $u \in \{1, \ldots, K\}$. Here, $K$ is the number of communities, which is considered to be unknown in the rest of the paper.

2.2 A formal definition of homophily

To construct the hypothesis test, we must formulate a formal and general definition of homophily. The following approach resembles the intra- and inter-cluster density notions used to define communities in Fortunato (2010) and Mancoridis et al. (1998). The difference is that while these works consider the metric from the sample level (an observed network), we start at the population level (what generated the network). We choose a definition based on edge probabilities/densities as opposed to, say, connectedness because it is simpler to formulate statistically.

Consider a network model $P$ and let $c$ be a candidate community assignment. Define

$$\bar{p}_{\text{in}}(c) = \frac{1}{\sum_{k=1}^{K} \binom{n_k}{2}} \sum_{u=1}^{K} \sum_{i>j \in c_i = c_j = u} P_{ij} \quad \text{and} \quad \bar{p}_{\text{out}}(c) = \frac{1}{\sum_{k>l} n_k n_l} \sum_{u>v} \sum_{i>j \in c_i = u, c_j = u} P_{ij},$$

where $K$ is the number of communities, $n_u$ is the number of nodes in community $u$ for $u = 1, \ldots, K$ and $c_i$ is the community that node $i$ is in, $c_i \in \{1, 2, \ldots, K\}$. Here, $\bar{p}_{\text{in}}(c)$ and $\bar{p}_{\text{out}}(c)$ are the average intra- and inter-community edge probability, respectively. Intuitively, $\bar{p}_{\text{in}}(c) - \bar{p}_{\text{out}}(c)$ quantifies the strength of homophily in the network. This quantity, however, should be adjusted with respect to the overall sparsity of the network. So, we propose the homophily parameter $\gamma(c, P)$ as

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

where $\bar{p} = \sum_{i>j} P_{ij} / \binom{n}{2}$ is the overall probability of an edge between nodes in the network. The more pronounced the homophily is in the network, the larger $\bar{p}_{\text{in}}(c)$ is compared to $\bar{p}_{\text{out}}(c)$, thus making $\gamma$ large. Additionally, $\bar{p}$, the overall network sparsity, serves as a normalizing factor.

For example, if $\bar{p}_{\text{in}}(c) - \bar{p}_{\text{out}}(c) = 0.15$ and $\bar{p} = 0.10$, this should be considered as stronger homophily than if $\bar{p} = 0.70$ since the difference between the average intra- and inter-community edge probability is more striking in the former. Additionally, $\gamma$ depends on the community assignments and the data-generating probability matrix. While we eventually need to chose a model for $P$, $\gamma$ is a truly general definition for it is not rooted to any one model. This allows for hypothesis testing against multiple null models, yielding richer
information about the network. Furthermore, testing against multiple nulls can help ensure another network feature (e.g., degree heterogeneity) isn’t masquerading as homophily (Graham, 2017).

2.3 Hypothesis testing problem statement

In presenting the hypothesis test, we first acknowledge the difference between working with labeled and unlabeled networks. In a labeled network, we have information on certain categorical node covariates, e.g., the political party of a politician. Any such categorical node covariate \( \{X_i\}_{i=1}^n \) can be mapped to a community assignment \( \{c^*_i\}_{i=1}^n \) by mapping the distinct values of the categorical variable to the set \( \{1, 2, \ldots, K\} \) for some \( K \). We then want to know whether these labels correspond to meaningful community structure. In the terminology of the homophily parameter, there is some known \( c^* \) and the natural hypothesis test becomes:

\[
H_0 : \gamma(c^*, P) \leq \gamma_0 \text{ vs. } H_1 : \gamma(c^*, P) > \gamma_0,
\]

for some reasonable threshold value \( \gamma_0 \).

Conversely, in an unlabeled network, we don’t have any node covariate information and we must determine whether any community assignment has homophily. In this case, we define \( \tilde{\gamma}(P) = \max_c \gamma(c, P) \) to be the maximum homophily over the set of all possible community assignments. This is closely related to the idea of community detection algorithms that rely on maximizing a certain objective function, e.g., likelihood modularity, spectral clustering, etc. The hypothesis test is then

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

for some reasonable threshold value \( \gamma_0 \). For the remainder of this work, we will consider the unlabeled case because it is more realistic and more challenging. Our proposed method can be easily extended to the labeled network case.

The next question is, how to determine the “reasonable threshold” \( \gamma_0 \)? From first principles, a natural choice is to set \( \gamma_0 = 0 \). Doing so asserts that under the null hypothesis (no homophily), \( \gamma(c, P) \leq 0 \) for any community assignment \( c \) and that the alternative hypothesis is true (homophily) as long as there exists a community assignment \( c \) for which \( \gamma(c, P) > 0 \). This seems reasonable at first look, but there is a conceptual issue with these assertions. If \( P \) is a network model such that \( \gamma(c, P) \leq 0 \) for any community assignment \( c \), then \( P \) must be the homogeneous Erdős-Rényi (ER) model of Erdős and Rényi (1959) with \( P_{ij} = p \) for some \( p \in (0, 1) \) and for all \( i, j \). The formal lemma is below with proof in the Supplementary Materials.

**Lemma 2.1.** \( \gamma(c, P) \leq 0 \) for all \( c \) if and only if \( P \) is from a homogeneous Erdős-Rényi model.

Setting \( \gamma_0 = 0 \) is equivalent to testing \( H_0 : “\text{the network follows the ER model}” \) vs. \( H_1 : “\text{the network does not follow the ER model}” \). In other words, \( \gamma_0 = 0 \) implicates ER as the only model without homophily. To see why this might be a problem, consider the
following toy example of a network model with \( n = 4 \) nodes:

\[
P = \begin{pmatrix}
0 & 0.16 & 0.16 & 0.18 \\
0.16 & 0 & 0.23 & 0.18 \\
0.16 & 0.23 & 0 & 0.27 \\
0.18 & 0.18 & 0.27 & 0
\end{pmatrix}.
\]

For the community assignment \( c = (1, 1, 2, 2) \), i.e., nodes 1 and 2 are in one community and nodes 3 and 4 are in the second community, we have \( \gamma(c, P) = 0.14 > 0 \). A closer look at \( P \), however, shows that nodes 1 and 2 are more likely to connect to the nodes belonging to the other community than to each other. That is, for nodes in the first community, intra-community edges are less likely than inter-community edges, which goes against the concept of homophily. The positive value of the homophily parameter is only an artifact of non-uniformity.

We now define the first concept related to homophily, nominal homophily: a network model \( P \) has nominal homophily if there exists some community assignment \( c \) such that \( \gamma(c, P) > 0 \). As shown, this could occur merely as an artifact and in that case we have homophily in name only (hence, nominal). Setting \( \gamma_0 = 0 \) means we are testing for the presence of nominal homophily.

Now, suppose a network model \( P \) has nominal homophily. There could be two possibilities: the homophily arises as an artifact of some other network feature (e.g., non-uniformity, as observed in the toy example above) or the homophily is intrinsic to the network. To distinguish between these two cases, we introduce two more concepts: collateral and intrinsic homophily. Collateral homophily is indirect or secondary homophily that arises as a consequence of some network property which is itself not homophily. On the other hand, intrinsic homophily is true or inherent homophily which is greater than what could arise due to collateral homophily only. To fix ideas, consider the Chung-Lu model of Chung and Lu (2002) where \( P_{ij} = \theta_i \theta_j \) for a vector parameter \( \theta = (\theta_1, \ldots, \theta_n) \). Consider a toy example from this model with \( n = 4 \) and \( \theta = (0.6, 0.7, 0.8, 0.9) \). Then,

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

For \( c = (1, 1, 2, 2) \), \( \gamma(c, P) = 0.03 > 0 \), which means this network model has nominal homophily. Similar to the previous toy example, however, nodes 1 and 2 actually have higher inter-community edge probability than intra. This is an example of collateral homophily since \( \gamma(c, P) \) exceeds zero due to the collateral effect of degree heterogeneity, rather than due to true or intrinsic homophily. Similarly, other network features like transitivity might have an effect on \( \gamma(c, P) \) so selecting \( \gamma_0 \) in these cases should be done with equal care.

To distinguish between collateral and intrinsic homophily, we propose to select a “reasonable threshold” \( \gamma_0 \) by referring to the network itself in a property-specific manner. Here, we illustrate the idea at the population level in reference to the property of degree heterogeneity. Consider a network of \( n \) nodes with expected degrees \( \delta = \delta_1, \ldots, \delta_n \), and homophily parameter given by \( \tilde{\gamma} \). Note that, in practice, we will not have access to expected degrees.
or the true value of $\tilde{\gamma}$, but for now we are working under this supposition to theoretically illustrate the idea at a population level. We want to determine if the extent of homophily in the network can be explained by collateral homophily due to degree heterogeneity versus intrinsic homophily beyond what could be induced by degree heterogeneity. In this case, the “reasonable threshold” $\gamma_0$ can be determined as follows. Construct the “null” model $P^{(0)}$ by following the Chung-Lu formulation i.e., $P^{(0)}(\delta) = \delta_i \delta_j / \sum_j \delta_j$. Next, consider the set of all possible community assignments $c$, and define $\gamma_0 = \max_c \gamma(c, P^{(0)}(\delta))$. The threshold $\gamma_0$ represents the maximum extent of collateral homophily that could be induced by degree heterogeneity in the network. If $\tilde{\gamma} \leq \gamma_0$, we conclude that the network has collateral homophily, and if $\tilde{\gamma} > \gamma_0$, we conclude that the network has intrinsic homophily. The next sub-section builds this population-level idea to a sample-level statistical test.

To summarize, a network model $P$ has no homophily if $\gamma(c, P) \leq 0$ for all $c$, whereas it has nominal homophily if there exists some $c$ for which $\gamma(c, P) > 0$. The network has collateral homophily if it has nominal homophily, i.e., $\tilde{\gamma}(P) = \max_c \gamma(c, P) > 0$, but $\gamma(P)$ is lower than the threshold which is the maximum extent of homophily that could be induced by some network property, e.g., degree heterogeneity. The network has intrinsic homophily if $\tilde{\gamma}(P)$ is greater than the collateral threshold $\gamma_0$. Distinguishing collateral from intrinsic homophily is where we must finally settle on a model.

### 2.4 Test statistic and proposed test

We now introduce the test statistic and the test itself. Using the same notation as above and following closely the ideas from Section 2.2, define

$$
\hat{p}_{in}(c) = \frac{1}{\sum_{k=1}^{K} \binom{n_k}{2}} \sum_{u=1}^{K} \sum_{i>j, c_i = c_j = u} A_{ij} \\
\hat{p}_{out}(c) = \frac{1}{\sum_{k>l} \sum_{u,v} \sum_{i>j, c_i = u, c_j = v} A_{ij}}.
$$

Then our test statistic estimates $\gamma(c, P)$ from Eq. 3 as

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

where $\hat{p} = \sum_{i>j} A_{ij} / \binom{n}{2}$ and $\hat{p}_{in}(c), \hat{p}_{out}(c)$ and $\hat{p}$ are the sample versions of $p_{in}(c), p_{out}(c)$ and $p$ and estimate each parameter respectively. A network with significant homophily would result in a larger value of $T$ so we reject $H_0$ if

$$
\tilde{T}(A) = \max_c \{ T(c, A) \} > C
$$

for some cutoff $C$ that depends on the graph size and the null model. Computing the global maximum, $\tilde{T}(A)$, is difficult as it is a combinatorial optimization problem. The space of all possible community assignments is huge ($\sim K^n / K!$) so an exhaustive search is unfeasible. For theoretical purposes we assume that we can find the maximum, but in practice we find communities $\tilde{c}$ using any off-the-shelf optimization algorithm and then approximate $\tilde{T}(A)$ with $T(\tilde{c}, A)$. We suggest using the Walktrap algorithm (Pons and Latapy, 2005). This strategy of deriving theoretical results for the exact maximum but using an approximation
algorithm in applications is standard practice in the literature (Bickel and Chen, 2009; Zhao et al., 2012).

To obtain a test with level $\alpha$, we should set $C$ to be the $(1 - \alpha)$ quantile of the null distribution of $\tilde{T}(A)$. But this is a difficult task. The test statistic is a max taken over $\sim K^n/K!$ possible community assignments and these random variables are dependent since they are functions of the same adjacency matrix. We offer two options to sidestep this difficult theoretical problem. First, for the test of nominal homophily, we derive a threshold, $C$, which leads to an asymptotically valid test. We use an union bound approach to find a closed form expression for $C$ but this cutoff is too conservative since the union is over such a large set of dependent random variables. This derived expression for $C$ and further discussion can be found in Section 3.

The second option is a bootstrap based approach to determine an empirical $p$-value. We first compute the test statistic $\tilde{T}(A) = \max_c\{T(c, A)\}$ as in Eq. 6. Since the null distribution of $\tilde{T}(A)$ is unknown, we must simulate draws from this distribution in order to have a comparison with our observed test statistic. We illustrate the method for testing nominal homophily (ER null). First, compute $\hat{p}$ from $A$. Then generate $B$ ER networks $A_i^*$ based on this $\hat{p}$ and find $\tilde{T}_{i}^* = \max_c\{T(c, A_i^*)\}$ for $i = 1, \ldots, B$. The empirical distribution of $\{\tilde{T}_{i}^*\}_{i=1}^n$ serves as a proxy for the null distribution of $\tilde{T}(A)$ so we compute the $p$-value as

$$p-val = \#(\tilde{T}_{i}^* \geq \tilde{T}_{obs})/B$$

and we reject $H_0$ if the $p$-value is less than a pre-specified $\alpha$. Algorithm 1 outlines these steps.

**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 $\tilde{T}_{obs} = \max_c\{T(c, A)\}$ 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 $\tilde{T}_{i}^* = \max_c\{T(c, A_i^*)\}$

end

$p-val = \#(\tilde{T}_{i}^* \geq \tilde{T}_{obs})/B$

The bootstrap test can be easily extended to distinguishing between collateral homophily and intrinsic homophily. Consider testing whether the network has intrinsic homophily greater than what could appear as a consequence of degree heterogeneity. As before, we compute the test statistic $\tilde{T}(A) = \max_c\{T(c, A)\}$ but now we use the Chung and Lu (2002) model for generating bootstrap resamples. To do this, use the degree distribution of the observed network to fit model parameters. Then, generate networks $A_i^*$ from the fitted model and find $\tilde{T}_{i}^* = \max_c\{T(c, A_i^*)\}$ for $i = 1, \ldots, B$. Again compute the $p$-value by comparing the observed test statistic $\tilde{T}(A)$ to the empirical distribution of $\{\tilde{T}_{i}^*\}_{i=1}^n$. The steps are in Algorithm 2.
Algorithm 2 Test of collateral (degree heterogeneity) vs. intrinsic homophily

Result: \( p \)-value for testing collateral homophily

Input: \( n \times n \) adjacency matrix \( A \), number of iterations \( B \)

Compute \( \hat{T}_{\text{obs}} = \max_c \{ T(c, A) \} \) and \( \hat{\theta}_i = \sum_{j=1}^{n} A_{ij}/(2 \sum_{k>j} A_{kj})^{1/2} \) for \( i = 1, \ldots, n. \)

for \( B \) times do

\( A^*_i \leftarrow \) CL network with \( \hat{\theta} \)

Compute \( \hat{T}^*_i = \max_c \{ T(c, A^*_i) \} \)

end

\( p\text{-val}=\#(\hat{T}^*_i \geq \hat{T}_{\text{obs}})/B \)

We end this section by listing some advantages of the proposed method. First, the test is rooted in a formal and general definition of the homophily metric, \( \gamma \). This metric is simple, flexible, and well-connected to the conceptual notion of homophily which we argue makes it a more principled approach. Existing methods are based on specific random graph models, such as the ER model, which are implicitly presumed to be the only models that do not have homophily. As illustrated in this section, this approach can have unintended consequences, as unrelated properties, such as non-uniformity or degree heterogeneity, can masquerade as homophily. Second, existing methods frame the test as a dichotomy between the absence/presence of homophily. The proposed framework enables a broader and more realistic perspective towards homophily as a continuum, from none to nominal to collateral to intrinsic.

Next, existing methods typically use test statistics that do not have a clear interpretation with regards to homophily. For example, the spectral method of Bickel and Sarkar (2016) uses the eigenvalue of the centered adjacency matrix as the test statistic, which is statistically sound but does not have a natural interpretation as a measure of homophily. In contrast, the proposed test statistic has a clear and intuitive connection. In fact, it can even be used as a descriptive statistic to quantify the strength of homophily in a network. The proposed method also provides practitioners with greater flexibility in what kind or extent of homophily they wish to test for. In this section, we have described the test for collateral homophily induced from degree heterogeneity. In the Supplementary Materials, we describe the test for collateral homophily induced from transitivity. The proposed framework can be extended to other network properties because of the generalizability of the homophily metric and the flexibility of the bootstrap step.

Furthermore, the proposed method provides a richer set of practical insights compared to existing methods. Given a network, we recommend that practitioners first carry out the test of no homophily vs nominal homophily. If this test is rejected, further tests should be carried out to check whether it could be due to collateral homophily induced by properties such as degree heterogeneity. Thus, the method not only helps decide whether there is any homophily in the network, but also helps understand where this homophily is arising from. We illustrate this approach on a number of real-world networks in Section 5. Of independent interest is the simple but insightful visual aid for this analysis also from Section 5. The observed value of the test statistic is plotted next to a number of bootstrap histograms. This tool helps in visualizing how the strength of homophily observed in the network compares to benchmark network with the same density, or with the same degree distribution, and so on. Similar to
the test statistic, this plot is useful not only for inference but also as a useful descriptive tool. Finally, this method can be easily extended to labeled networks, which is not feasible for some existing methods such as the spectral method of Bickel and Sarkar (2016).

3 Theoretical Results

3.1 Uniform convergence of the estimator

Our first theoretical result relates to the performance of \( T(c, A) \) as an estimator. Note that, in our method, we maximize over all possible configurations \( c \). Therefore, it is not sufficient to fix a \( c \) and show that, given \( c \), \( T(c, A) \) converges to \( \gamma(c, P) \). Rather, we have to show that the estimator \( T(c, A) \) achieves uniform convergence, i.e., the convergence holds true uniformly over the range of \( c \). We make the following assumptions:

A1. The average expected node degree diverges to infinity, that is \( n\bar{p} \rightarrow \infty \).
A2. For any candidate community assignment, at least two community sizes must grow linearly with \( n \).
A3. For any candidate configuration, we must have \( n\bar{p}_{in}(c) \rightarrow \infty \) and \( n\bar{p}_{out}(c) \rightarrow \infty \).

The first assumption establishes a threshold for the sparsity level of the whole networks, whereas the third assumption establishes sparsity thresholds for the within-community and between-community links. The second assumption lays down a basic requirement for any legitimate candidate community assignment, since otherwise, one community will dominate the entire network. We now state the result establishing uniform convergence of the estimator, which does not depend on any specific random graph model, with proof in the Supplemental Materials.

**Lemma 3.1.** Under A1, A2, and A3, the estimator \( T(\cdot, A) \) uniformly converges to \( \gamma(\cdot, P) \) in probability, i.e.,

\[
\max_c |T(c, A) - \gamma(c, P)| \rightarrow 0.
\]

3.2 Level and power of the test statistic

We now prove asymptotic results about our testing procedure for the test of nominal homophily. Note that the theoretical results are for the asymptotic threshold test, not the bootstrap-based method. Consider the test statistic \( \tilde{T}(A) \) and rejection region \( R = \{ \tilde{T}(A) > C \} \) for a test at level \( \alpha \). The lemma is stated below.

**Lemma 3.2.** For some small \( \epsilon > 0 \), consider the threshold

\[
C = \left( \frac{2\{\log(2N_{n,K}) - \log n\}}{n^2} \right)^{1/2} \cdot \frac{1 + \epsilon}{\hat{p}}.
\]

where \( N_{n,K} \leq K^n \) is the number of possible community assignments. When the null hypothesis is true, i.e., \( \tilde{\gamma}(P) \leq 0 \), for any \( \eta > 0 \) we have

\[
\lim_{n \rightarrow \infty} P(\tilde{T}(A) > C) \leq \alpha + \eta.
\]
When the alternative hypothesis is true, i.e., \( \tilde{\gamma}(P) > 0 \) and there is intrinsic homophily such that \( \sqrt{n}(\bar{p}_{in} - \bar{p}_{out}) \to \infty \) and for any \( 1 \leq u < v \leq K \),

\[
\bar{p}_u := \frac{1}{\binom{n_u}{2}} \sum_{i,j: c_i = u, c_j = u} P_{ij} > \bar{p}_{uv} := \frac{1}{n_un_v} \sum_{i,j: c_i = u, c_j = v} P_{ij},
\]

then for any \( \eta > 0 \) we have

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

The proof proceeds by approximating the cutoff \( C \) under the null model using a union bound and then shows that the probability of not rejecting under the alternative goes to 0. Hoeffding’s inequality is leveraged both to find \( C \) and show that the probability goes to 0. All proofs are left to the Supplementary Materials. For the sake of simplicity, we assume that \( K \), the number of communities, is fixed. The theoretical properties can be generalized to the \( K \to \infty \) case by adding some more assumptions.

We also see that the alternative hypothesis is characterized in a general fashion. The classical SBM is a special case where \( \bar{p}_u^n = p_{in} \) for all \( u \) and \( \bar{p}_{uv}^n = p_{out} < p_{in} \) for all \( u \neq v \). Various extensions of the SBM, such as the degree corrected blockmodel or the popularity adjusted blockmodel, also satisfy these conditions under appropriate settings (Karrer and Newman, 2011; Sengupta and Chen, 2018) but, as desirable, the theorem is more general.

Moreover, this also shows that our method is consistent when the expected node degree increases sub-linearly whereas the proof for the Spectral method assumes that the node degree increases linearly.

While this asymptotic threshold is theoretically valid, the closed-form expression for \( C \) is not “sharp” in the sense that it yields a conservative test. The test statistic is found by maximizing \( T(c, A) \) over all valid community assignments, \( c \). Since \( T(c_1, A) \) and \( T(c_2, A) \) for two community assignments are dependent in a non-trivial manner, we have used the union bound to analyze the distribution of \( T(c, A) \). This makes the bound conservative. To obtain a sharper bound, we need more advanced analysis involving random field theory which is technically challenging. These difficulties motivate the use of the bootstrap approach in the following section.

4 Simulations

4.1 Settings

Our primary metric of interest is the rejection rate of the test under different settings. Under the null hypothesis, we look at the rejection rate as a function of \( n \), and we expect to see a rejection rate near the nominal significance level. Under the alternative hypothesis, we generate models with intrinsic homophily, and study the rejection rate in two ways: (1) fixing \( n \) and increasing \( \gamma(c, P) \), and (2) fixing \( \gamma(c, P) \) and increasing \( n \), and in both cases we expect the rejection rate to increase. For each case, we ran 100 Monte Carlo simulations and computed the mean rejection rate. Our bootstrap method used \( B = 200 \) bootstrap samples and we fix the level of the test at \( \alpha = 0.05 \). We chose these two Spectral methods proposed in Bickel and Sarkar (2016) as benchmarks since these are leading and
well-established methods with formal guarantees. Even though Bickel and Sarkar (2016)
suggest using only the adjusted method, we will still compare both. This is because, similar
to our proposed framework they propose a version of the test with an asymptotic threshold
and an adjusted version of the test with a bootstrap correction.

4.2 Test of no homophily vs. nominal homophily

First, we consider the simplest null hypothesis of no homophily, i.e., $\tilde{\gamma} = 0$. Following
Lemma 1, the null model must be the ER model. Following the theoretical result, we use
the rejection threshold $\frac{1}{p} \left[ \log(2\tilde{N}_{n,k}) - \log \alpha \right] / \left\{ n(n-2) \right\}^{1/2}$ for the asymptotic version of
our test where $\tilde{N}_{n,k} = n! / \left\{ (n/K)^K K! \right\}$. The results are in the top row of Figure 5. The
leftmost pane presents the rejection rate for the different methods when the networks were
generated from an ER with $p = 0.20$ and $n = 100, 200, \ldots, 500$. Since the null hypothesis is
true, these rejections are false positives. We observe that rejection rates are low, meaning
that each method won’t reject when the null hypothesis is true. In fact, all the methods
except the asymptotic threshold have a rejection rate that hovers around the level $\alpha$ which
is desirable for a test.

Next, we use the well-known stochastic blockmodel (SBM) to generate networks under
the alternative hypothesis (Holland et al., 1983). Under the stochastic blockmodel with $K$
communities, $P_{ij} = \omega_{c_i} \omega_{c_j}$ where $c_i, c_j$ represent the community membership of nodes $i$ and $j$
and $\omega$ is a $K \times K$ symmetric matrix of community-community interaction probabilities. We
used $K = 2$ communities. The top middle pane has networks generated with $p_{in} = 0.40$ and
$p_{out} = 0.20 (\tilde{\gamma} = 1.11)$ for $n = 40, 80, \ldots, 240$. The top right pane has networks generated
with $p_{in} = 0.25, 0.30, \ldots, 0.45, p_{out} = 0.20 (\tilde{\gamma} = 0.22, 0.40, 0.55, 0.66, 0.77)$ and $n = 200$. These two models are under the alternative hypothesis so these are true positives. The
rejection rate increases as either $n$ or $\gamma$ increase. The asymptotic cutoff value rejects less
often than the bootstrap approach. Both of the Spectral methods outperform the proposed
methods because the Spectral methods are designed for this exact scenario. The proposed
methods have low rejection rate under the null and increasing under the alternative which
is satisfactory. Overall, the Spectral and proposed methods are comparable in this setting.

4.3 Test of collateral homophily vs. intrinsic homophily

We now consider collateral homophily (from degree heterogeneity) as the null model and
intrinsic homophily as the alternative. Further simulations of networks with transitivity are
explored in the Supplementary Materials. We drop the asymptotic test from comparison as
do not have a theoretical expression for the asymptotic threshold for this case.

Under the null, we generate networks from the Chung-Lu (CL) model that was introduced in Section 2.2. Under the alternative, we generate networks from the degree-corrected
stochastic blockmodel (DCSBM) of Karrer and Newman (2011). Under the DCSBM with
$K$ communities, $P_{ij} = \theta_i \omega_{c_i} \omega_{c_j} \theta_j$, where $\omega$ is a $K$-by-$K$ symmetric matrix of community-
community interaction probabilities, and $\{c_i\}_{i=1}^n$ are the communities of the nodes, and
$\{\theta_i\}_{i=1}^n$ are degree parameters. This choice of alternative model is motivated by the fact
that the DCSBM has a natural interpretation as adding homophily to the CL model.

The results are in the bottom row of Figure 5. For the bottom left pane, we sample
\( \theta_i \sim \text{Uniform}(0.6, 0.8) \) and then generate CL graphs using these \( \theta \)'s for \( n = 50, 100, \ldots, 400 \) \( (\gamma(c, P) \approx 0.04) \). The bootstrap method has a rejection rate that hovers around the nominal level whereas the rejection rates for the Spectral methods increase as \( n \) increases. Since CL has only collateral and not intrinsic homophily, these are false positives. We generate DCSBM graphs by first sampling \( \theta_i \sim \text{Uniform}(0.6, 0.8) \) but now also adding block structure with \( K = 2 \) equal-sized blocks. For the bottom middle pane, fix \( p_{in} = 0.40, p_{out} = 0.10 \) and let \( n = 50, 100, 200, \ldots, 400 \) \( (\gamma(c^*, P) \approx 1.10) \). Also fix \( n = 200 \) and increase \( p_{in} = 0.30, 0.40, \ldots, 0.60 \) and \( p_{out} = 0.10 \) for the bottom right pane \( (\gamma(c^*, P) \approx 0.60, 1.13, 1.33, 1.44) \). All methods show an increase in rejection rate as either \( n \) or \( \gamma(c^*, P) \) increase. DCSBM has intrinsic homophily so these are true positives.

The main observation is that the Spectral methods have very high rejection rates even under the null. A network with large principal eigenvalue but small \( \gamma \) is rejected, even though there is no intrinsic homophily in the network. Since the Spectral method uses the ER model as the null model it therefore rejects when the network deviates from the ER model due to degree heterogeneity, rather than rejecting when the network has intrinsic homophily. The Spectral methods can only use the ER model as the null model so in a purely statistical sense, it is making the right decision by rejecting these networks. This is problematic from a practical perspective, however, since the rejection decision is interpreted from the perspective of community structure even though there is no intrinsic homophily present. On the other hand, the proposed Bootstrap method method performs well, with Type I error close to the nominal level and power increasing to one as \( \gamma \) increases.

5 Real Data Analysis

We now consider several real world data sets to study the proposed method. We choose well known networks where many are assumed to have homophily structure, but we are curious whether they actually do. For each data set, we compute the \( p \)-value and find the bootstrap histogram of the test statistic for the ER and CL null models in order to show how both null hypotheses can be used in conjunction to give a rich understanding of the network. We compare the proposed method to the \( p \)-value from the adjusted Spectral method.

For some of the data sets, we also have “ground-truth” community assignments \( c^* \) which could be compared to the community assignments \( \hat{c} \) returned by the Walktrap algorithm. However, it will always be the case that \( T(\hat{c}, A) \geq T(c^*, A) \) so we don’t consider these known labels. Lastly, for each example we used \( B = 1000 \) bootstrap samples.

See Table 1 for all \( p \)-values and description of the networks and Figure 3 for all histograms from the bootstrap method. While Table 1 provides a succinct summary with the \( p \)-values from each method, the plots in Figure 3 provide more details and insights. In these plots, the observed test statistic computed from the dataset is plotted as a vertical line along with histograms representing bootstrap distributions from various benchmark models. This simple but informative plots give practitioners a reference of how the observed homophily compares to the range of homophily in various benchmarks.

For the Jazz network, all methods have significant \( p \)-values and the test statistic for the proposed method is far in the right tail of the distribution under both nulls. Additionally, since both hypotheses are rejected, this implies that there is very likely intrinsic homophily.
Figure 2: Rejection rates from simulation study: Top row: Erdős-Rényi (ER) vs. stochastic block model (SBM) with $K = 2$ equal-sized communities. Bottom row: Chung-Lu (CL) vs. degree corrected stochastic block model (DCSBM). Left column: null model (rejection means false positive). Middle and right columns: alternative model (rejection means true positive). Method: (- - - - -) Spectral, (-----------) Spectral Adjusted, (----) Bootstrap, (-------) Asymptotic and horizontal line at $\alpha = 0.05$ level of the test.

Figure 3: Histograms of bootstrap samples from the proposed method for the four real data sets. The darker histogram is with Erdős-Rényi null and the lighter histogram is with Chung-Lu null.
Table 1: The number of nodes $n$ and edges $m$ for different networks along with their $p$-values for the adjusted Spectral method and Bootstrap method against different null hypotheses (ER=Erdős-Rényi, CL=Chung-Lu). In the Jazz network, nodes are bands from 1912-1940 and edges indicate that two bands have at least musician in common. The Karate network edges correspond to social connections between different members of the club represented as nodes. Nodes in the Hospital network are people and edges are interactions. The Political blogs network represent Blogs as nodes and there is an edge if one blog cites another, made undirected as per other papers.

| Network        | Reference             | $n$ | $m$ | Spectral Adj. | Bootstrap ER | Bootstrap CL |
|----------------|-----------------------|-----|-----|---------------|--------------|--------------|
| Jazz           | Gleiser and Danon (2003) | 198 | 2,742 | $< 0.001$    | 0.000        | 0.003        |
| Karate         | Zachary (1977)        | 34  | 78  | $< 0.001$    | 0.421        | 0.551        |
| Hospital       | Vanhems et al. (2013) | 75  | 2,278 | $< 0.001$   | 0.001        | 0.512        |
| Political Blogs| Adamic and Glance (2005)| 1,222 | 16,717 | $< 0.001$   | 0.168        | 1.000        |

in this network.

The proposed method yields a large $p$-value in each case for the Karate network but the adjusted Spectral method yields a significant $p$-value. We have seen how the adjusted Spectral method is much more sensitive to divergence from the ER model but this doesn’t necessarily mean that there is intrinsic homophily in the network. Thus, it’s reasonable to conclude that degree heterogeneity is causing the low $p$-value for the adjusted Spectral method, especially since four nodes have $> 10$ edges and twenty four nodes have $< 5$ edges.

All $p$-values are low in the Hospital network, except testing against CL. This indicates that perhaps the low $p$-values are detecting degree heterogeneity (collateral homophily) as opposed to true community structure (intrinsic homophily). Using the adjusted Spectral $p$-value alone would have led to the conclusion that there was true homophily in this network but, by using multiple nulls together with the proposed method, we can conclude that this is more likely degree heterogeneity and/or preferential attachment as opposed to intrinsic homophily.

For the Political Blogs network, while the Spectral method yields a significant $p$-value, the proposed method’s $p$-value is low for ER and large for CL. This seems to imply that both the adjusted Spectral method and the ER results are detecting divergence from ER towards CL, rather than true homophily. Also, the fact that the test statistic distribution under CL is so much farther right than the test statistic distribution under ER indicates strong heterogeneity in the network. Moreover, the proposed method $p$-value under the ER null is detecting changes in $\gamma$ from ER whereas the Spectral method is detecting any difference from ER. This could explain why the Spectral $p$-value is so much lower than the proposed ER $p$-value.

We plot the adjacency matrix in Figure 4 to further investigate. While there does appear to be some homophilic structure in the network, there is very clearly large degree heterogeneity. These two features together could cause the adjusted Spectral method to have a significant $p$-value. Regardless, looking at the adjacency matrix, concluding that there is true homophily or concluding that there is only degree heterogeneity seem to be legitimate.
6 Discussion

In this paper, we proposed a novel method for testing homophily and community structure in network. Some limitations of our method are that it was only proved with the ER null and with the asymptotic threshold. In particular, we do not have theoretical results for the bootstrap method. Furthermore, the derived bound is too conservative so an important avenue of future work is finding a sharper bound using more advanced probabilistic tools. Additionally, the method can be slow for large networks where the time is mostly determined by the speed of community detection algorithm. Future work could focus on using this method for larger networks while also considering different null hypotheses. This method could also be extended to different network features of interest, such as the small-world property. Also, we considered the framework where each node belongs to a single community, which means our framework does not include mixed membership models.

References

Adamic, L. A. and Glance, N. (2005). The political blogosphere and the 2004 us election: divided they blog. In Proceedings of the 3rd international workshop on Link discovery, pages 36–43.

Bickel, P. J. and Chen, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. Proceedings of the National Academy of Sciences, 106:21068–21073.

Bickel, P. J. and Sarkar, P. (2016). Hypothesis testing for automated community detection in networks. Journal of the Royal Statistical Society: Series B: Statistical Methodology, pages 253–273.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of
communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

Chung, F. and Lu, L. (2002). The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences*, 99(25):15879–15882.

Erdős, P. and Rényi, A. (1959). On random graphs. *Publicationes Mathematicae Debrecen*, 6:290–297.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486:75–174.

Gleiser, P. M. and Danon, L. (2003). Community structure in jazz. *Advances in Complex Systems*, 06(04):565–573.

Graham, B. S. (2017). An econometric model of network formation with degree heterogeneity. *Econometrica*, 85(4):1033–1063.

Guo, Z., Cho, J.-H., Chen, R., Sengupta, S., Hong, M., and Mitra, T. (2020). Online social deception and its countermeasures: A survey. *IEEE Access*, 9:1770–1806.

He, Z., Liang, H., Chen, Z., and Zhao, C. (2018). Detecting statistically significant communities. *arXiv*.

He, Z., Liang, H., Chen, Z., Zhao, C., and Liu, Y. (2020). Computing exact p-values for community detection. *Data Mining and Knowledge Discovery*, 34:833–869.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098.

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic block models: First steps. *Social Networks*, 5:109–137.

Jin, J. (2015). Fast community detection by SCORE. *The Annals of Statistics*, 43(1):57–89.

Kane, G. C., Alavi, M., Labianca, G., and Borgatti, S. P. (2014). What’s different about social media networks? a framework and research agenda. *MIS quarterly*, 38(1):275–304.

Karrer, B. and Newman, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Physical Review E*, 83:016107.

Krivitsky, P. N. and Handcock, M. S. (2008). Fitting position latent cluster models for social networks with latentnet. *Journal of Statistical Software*, 24.

Lancichinetti, A., Radicchi, F., and Ramasco, J. J. (2010). Statistical significance of communities in networks. *Physical Review E*, 81(4):046110.

Leitch, J., Alexander, K. A., and Sengupta, S. (2019). Toward epidemic thresholds on temporal networks: a review and open questions. *Applied Network Science*, 4(1):105.

Li, Y. and Qi, Y. (2020). Asymptotic distribution of modularity in networks. *Metrika*, 83(4):467–484.
Mancoridis, S., Mitchell, B. S., Rorres, C., Chen, Y., and Gansner, E. R. (1998). Using automatic clustering to produce high-level system organizations of source code. In Proceedings. 6th International Workshop on Program Comprehension. IWPC’98 (Cat. No. 98TB100242), pages 45–52. IEEE.

Mason, O. and Verwoerd, M. (2007). Graph theory and networks in biology. IET systems biology, 1(2):89–119.

McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. Annual Review of Sociology, 27(1):415–444.

Newman, M. E. (2002). Assortative mixing in networks. Physical review letters, 89(20):208701.

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In Advances in neural information processing systems, pages 849–856.

Pagani, G. A. and Aiello, M. (2013). The power grid as a complex network: a survey. Physica A: Statistical Mechanics and its Applications, 392(11):2688–2700.

Palowitch, J., Bhamidi, S., and Nobel, A. B. (2018). Significance-based community detection in weighted networks. Journal of Machine Learning Research, 18:1–48.

Pons, P. and Latapy, M. (2005). Computing communities in large networks using random walks. In Computer and Information Sciences-ISCI 2005, pages 284–293. Springer.

Rohe, K., Chatterjee, S., and Yu, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. The Annals of Statistics, 39(4):1878–1915.

Sengupta, S. and Chen, Y. (2015). Spectral clustering in heterogeneous networks. Statistica Sinica, 25:1081–1106.

Sengupta, S. and Chen, Y. (2018). A block model for node popularity in networks with community structure. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 80(2):365–386.

Vanhems, P., Barrat, A., Cattuto, C., Pinton, J.-F., Khanafer, N., Regis, C., Kim, B.-A., Comte, B., and Voirin, N. (2013). Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. PloS one, 8:e73970.

Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. Journal of anthropological research, 33(4):452–473.

Zhao, Y., Levina, E., and Zhu, J. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. The Annals of Statistics, 40:2266–2292.
Appendix

Simulations: Transitivity

We now test for collateral vs. intrinsic homophily against the network property of transitivity. Under the null, we generate networks with transitivity but no intrinsic homophily by using the latent space model (LSM) of Hoff et al. (2002). Under the LSM, each node is assumed to have a latent position in a $d$-dimensional space. Edge probabilities are determined by $L_2$ distances between latent positions of the nodes given by $\logit(P_{ij}) = \beta - |z_i - z_j|$ for $1 \leq i < j \leq n$, where $z_i$ is the latent position of the $i$th node and $\beta$ is a parameter that controls overall sparsity. If $d = 1$, then we have $z_i \sim N(0, \sigma^2)$. $P$ is estimated by using the maximum likelihood estimation strategy described in Hoff et al. (2002) as implemented in the R package latentnet (Krivitsky and Handcock, 2008). Under the alternative, we add intrinsic homophily to this model with $\logit(P_{ij}) = \beta_{in} - |z_i - z_j|$, when nodes $i$ and $j$ are in the same community and $\logit(P_{ij}) = \beta_{out} - |z_i - z_j|$, when nodes $i$ and $j$ are in different communities with $\beta_{out} < \beta_{in}$. We will denote this model as $LSM_{hom}$. The algorithm below outlines the steps involved for this test. The results are in Figure 5.

Algorithm 3 Testing collateral (transitivity) vs. intrinsic homophily

Result: $p$-value for testing collateral homophily

Input: $n \times n$ adjacency matrix $A$, number of iterations $B$

Compute $\hat{T}_{obs} = \max_c \{T(c, A)\}$ and $\hat{\beta}, \hat{\sigma}^2$ from latentnet package.

for $B$ times do

$A^*_i \leftarrow$ LSM network with $\hat{\beta}, \hat{\sigma}^2$ and $d = 1$

Compute $\bar{T}^*_i = \max_c \{T(c, A^*_i)\}$

end

$p\text{-val} = \#(\bar{T}^*_i \geq \hat{T}_{obs})/B$

For all networks, we have $d = 1$ latent space dimension and $\sigma^2 = 0.50$. For the left pane (collateral homophily), LSM networks were generated with $n = 40, 60, \ldots, 120$ and $\beta = 0$ ($\gamma(c, P) \approx 0.30$). For the middle pane (intrinsic homophily), LSMhom networks were generated with $\beta_{in} = 0.40$ and $\beta_{out} = -0.40$ and $n = 40, 80, \ldots, 400$ ($\gamma(c, P) \approx 0.52$). For the right pane (intrinsic homophily), LSMhom networks were generated with $\beta_{in} = 0, 0.5, 1, 1.5$ and $\beta_{out} = -0.50$ and $n = 100$ ($\gamma(c, P) \approx 0.37, 0.64, 0.84, 0.99$). For the LSM networks, both Spectral methods have a high rejection rate that increases with $n$ whereas the proposed method’s rejection rate is right around the $\alpha$ level. For both LSMhom networks, all methods have increasing rejection rate as $n$ or $\gamma(c^*, P)$ with the Spectral methods having a higher rejection rate than that of the proposed method. This is to be expected because the Spectral methods are more sensitive to any divergence from the ER null model.
Figure 5: Rejection rates from simulation study: latent space model (LSM) vs. LSM with homophily (LSMhom). Left column: null model (rejection means false positive). Middle and right columns: alternative model (rejection means true positive). Method: (-----) Spectral, (……..) Spectral Adjusted, (-----) Bootstrap, (……..) Asymptotic and horizontal line at \( \alpha = 0.05 \) level of the test.

Technical Proofs

Lemma 1

The only if direction of the claim is immediate. To prove the forward direction, we first show that \( \gamma(c, P) \leq 0 \) for all \( c \) implies that \( \gamma(c, P) = 0 \) for all \( c \). Then we show that if \( \gamma(c, P) = 0 \) for all \( c \), then \( P \) is from an ER model.

For the first part, this is equivalent to showing that if \( \gamma(c, P) < 0 \) for some \( c \) then there exists some \( c' \) such that \( \gamma(c', P) > 0 \). If there exists some \( c \) such that \( \gamma(c, P) < 0 \), then

\[
\frac{1}{\sum_{k=1}^{K} \binom{n_k}{2}} \sum_{u=1}^{K} \sum_{i>j; c_i = c_j = u} P_{ij} < \frac{1}{\sum_{k>l} n_k n_l} \sum_{u>v} \sum_{i>j; c_i = u, c_j = v} P_{ij}.
\]

But this means that there is some \( P_{ij} \) such that \( P_{ij} \geq P_{kl} \) for all \( i \neq k \) or \( j \neq l \) and is strictly greater for at least one \( P_{kl} \). Thus, if we consider the community assignment \( c' \) where nodes \( i \) and \( j \) are in one community and all other nodes are in the other, then \( \bar{p}_{in}(c') > \bar{p}_{out}(c') \) and thus \( \gamma(c', P) > 0 \).

We will prove the second part by induction. Let \( n = 3 \) and we are given that \( \gamma(c, P) = 0 \) for all \( c \). We start by writing out the probability matrix.

\[
P = \begin{pmatrix} - & P_{12} & P_{13} \\ - & - & P_{23} \\ - & - & - \end{pmatrix}.
\]

There are three possible community assignments: \( c_1 = \{1, 1, 2\}, c_2 = \{1, 2, 1\} \) and \( c_3 = \{2, 1, 1\} \). From each of these assignments, we have a corresponding statement relating the
probabilities:

\[ \bar{p}_{in} = P_{12} = \bar{p}_{out} = \frac{1}{2}(P_{13} + P_{23}) \]

\[ \bar{p}_{in} = P_{13} = \bar{p}_{out} = \frac{1}{2}(P_{12} + P_{23}) \]

\[ \bar{p}_{in} = P_{23} = \bar{p}_{out} = \frac{1}{2}(P_{12} + P_{13}) \]

Plugging the first equation into the second equation we find:

\[ P_{13} = \frac{1}{2}(\frac{1}{2} + P_{12} + P_{23}) \implies P_{13} = P_{23}. \]

Plugging this into the first equation we have \( P_{12} = P_{13} = P_{23} =: p \) which means that this must be an ER model.

Now assume that the claim holds for \( n - 1 \) and show it holds for \( n \). For convenience, assume \( n \) is even but the proof can easily be extended if \( n \) is odd. Consider a network with \( n \) nodes such that \( \gamma(c, P) = 0 \). Remove an arbitrary node such that we have a network with \( n - 1 \) nodes and apply the induction hypothesis, i.e. \( P_{ij} = p \) for all \( i, j \). We now add the removed node back to the network such that the node has probability \( P_{i,n} \) of an edge between itself and node \( i \) for \( i = 1, \ldots, n - 1 \). Thus, the probability matrix is:

\[
P = \begin{pmatrix}
-p & p & \cdots & p & P_{1n} \\
-p & -p & \cdots & p & P_{2n} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
& & & & \ddots \\
& & & & p_{n-1,n} \\
& & & & -p
\end{pmatrix}.
\]

Since \( \gamma(c, P) = 0 \) for all \( c \), then we want to show that \( P_{i,n} = p \) for \( i = 1, \ldots, n \). Assume for contradiction that \( P \) is not ER and we will show that \( \gamma(c, P) \neq 0 \) for some \( c \).

Without loss of generality, let \( \{P_{1,n}, \ldots, P_{n/2,n}\} \) be the smaller values of the last column and \( \{P_{n/2+1,n}, \ldots, P_{n-1,n}\} \) be the larger values and consider the community assignment where nodes \( \{1, \ldots, n/2\} \) are in one community and nodes \( \{n/2 + 1, \ldots, n\} \) are in the other community. Then

\[
\bar{p}_{in} = \frac{1}{2^{\binom{n/2}{2}}} \left( p \cdot \left( \frac{n}{2} - 1 \right)^2 + \sum_{i=n/2+1}^{n-1} P_{i,n} \right) > \bar{p}_{out} = \frac{1}{n^2/4} \left( p \cdot \left( \frac{n^2}{4} - \frac{n}{2} \right) + \sum_{i=1}^{n/2} P_{i,n} \right)
\]

since

\[
\sum_{i=n/2+1}^{n-1} P_{i,n} > \sum_{i=1}^{n/2} P_{i,n}.
\]

Thus \( \gamma(c, P) \neq 0 \) for this particular choice of \( c \) and we have completed the proof. \( \Box \)
Lemma 2

We have

\[ T(c, A) = \frac{\hat{p}_{\text{in}}(c) - \hat{p}_{\text{out}}(c)}{\bar{\hat{p}}} \]

and

\[ \gamma(c, P) = \frac{\bar{p}_{\text{in}}(c) - \bar{p}_{\text{out}}(c)}{\bar{\hat{p}}} \]

where \( \hat{p}_{\text{in}}(c) \) and \( \hat{p}_{\text{out}}(c) \) are mutually independent, but \( \hat{p} \) has dependence with both \( \hat{p}_{\text{in}}(c) \) and \( \hat{p}_{\text{out}}(c) \). For notational simplicity, we will write \( \hat{p}_{\text{in}}(c) \) and \( \hat{p}_{\text{out}}(c) \) as \( \hat{p}_{\text{in}} \) and \( \hat{p}_{\text{out}} \), respectively. Note that, for any \( \epsilon > 0 \),

\[
\{|T(c, A) - \gamma(c, P)| > \epsilon\} \subseteq \left\{|\hat{p}_{\text{in}} - \bar{p}_{\text{in}}| > \frac{\epsilon}{2} \bar{p}_{\text{in}}\right\} \cup \left\{|\hat{p}_{\text{out}} - \bar{p}_{\text{out}}| > \frac{\epsilon}{2} \bar{p}_{\text{out}}\right\} \cup \left\{|\hat{p} - \bar{p}| > \frac{\epsilon}{2} \bar{p}\right\},
\]

which means that

\[
P[|T(c, A) - \gamma(c, P)| > \epsilon] \leq P\left[|\hat{p}_{\text{in}} - \bar{p}_{\text{in}}| > \frac{\epsilon}{2} \bar{p}_{\text{in}}\right] + P\left[|\hat{p}_{\text{out}} - \bar{p}_{\text{out}}| > \frac{\epsilon}{2} \bar{p}_{\text{out}}\right] + P\left[|\hat{p} - \bar{p}| > \frac{\epsilon}{2} \bar{p}\right].
\]

(10)

Next, we will construct concentration bounds for the three probabilities on the right hand side of the above inequality. To do this, we will use the multiplicative form of Chernoff bound, which states the following. Let \( \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \) where \( X_i \sim \text{Bernoulli}(p_i) \) and \( X_i \) are independent. Let \( \mu = \frac{1}{n} \sum_{i=1}^{n} p_i \) and fix \( \epsilon \in (0, 1) \). Then,

\[
P[|\bar{X} - \mu| > \epsilon \mu] \leq 2 \exp\left(-\frac{n \mu \epsilon^2}{3}\right).
\]

It is easy to see that each of the random variables \( \hat{p}, \hat{p}_{\text{in}}, \) and \( \hat{p}_{\text{out}} \) is of the form \( \bar{X} \) in the above statement. Before applying the bound, we need to determine the number of Bernoulli trials involved in \( \hat{p}, \hat{p}_{\text{in}}, \) and \( \hat{p}_{\text{out}} \). For \( \hat{p} \), the number of trials is \( \binom{n}{2} \approx n^2/2 \). For \( \hat{p}_{\text{in}} \), the number of trials is \( O(n^2) \), since the largest community has \( O(n) \) nodes, by A2. For \( \hat{p}_{\text{out}} \), the number of trials is at least \( O(n^2) \), since the two largest communities have \( O(n) \) nodes, by A2. Therefore, ignoring constants, we get

\[
P\left[|\hat{p}_{\text{in}} - \bar{p}_{\text{in}}| > \frac{\epsilon}{2} \bar{p}_{\text{in}}\right] \leq 2 \exp\left(-\frac{n \mu_{\text{in}} \epsilon^2}{3}\right) \quad (11)
\]

\[
P\left[|\hat{p}_{\text{out}} - \bar{p}_{\text{out}}| > \frac{\epsilon}{2} \bar{p}_{\text{out}}\right] \leq 2 \exp\left(-\frac{n \mu_{\text{out}} \epsilon^2}{3}\right) \quad (12)
\]

\[
P\left[|\hat{p} - \bar{p}| > \frac{\epsilon}{2} \bar{p}\right] \leq 2 \exp\left(-\frac{n \mu \epsilon^2}{3}\right) \quad (13)
\]

The above results are true for a fixed \( c \). However, we want uniform convergence. Note that
the number of possible configurations is less than $K^n$. Therefore,

$$P[\max c |T(c, A) - \gamma(c, P)| > \epsilon]$$

$$= P[\cup c \{ |T(c, A) - \gamma(c, P)| > \epsilon \}]$$

$$\leq \sum c P[|T(c, A) - \gamma(c, P)| > \epsilon]$$

$$\leq K^n \left( P \left[ |\hat{p}_{in} - \bar{p}_{in}| > \frac{\epsilon}{2} \bar{p}_{in} \right] + P \left[ |\hat{p}_{out} - \bar{p}_{out}| > \frac{\epsilon}{2} \bar{p}_{out} \right] + P \left[ |\hat{p} - \bar{p}| > \frac{\epsilon}{2} \bar{p} \right] \right)$$

$$= K^n \left( 2 \exp(-n^2 \bar{p}_{in}) + 2 \exp(-n^2 \bar{p}_{out}) + 2 \exp(-n^2 \bar{p}) \right)$$

$$= \exp(n \log(K)) \times \left[ 2 \exp(-n^2 \bar{p}_{in}) + 2 \exp(-n^2 \bar{p}_{out}) + 2 \exp(-n^2 \bar{p}) \right]$$

$$= 2 \exp[n(\log(K) - \epsilon^2 n \bar{p}_{in})] + 2 \exp[n(\log(K) - \epsilon^2 n \bar{p}_{out})] + 2 \exp[n(\log(K) - \epsilon^2 \bar{p})]$$

$$\rightarrow 0.$$ (21)

The last line follows from assumption A3.

**Lemma 3**

Assume a rejection region of the form $R = \{T_*(n) > c(n)\}$ where $c(n) = \frac{k(n)}{\bar{p}(n)/(1 + \epsilon)}$ and

$$T_*(n) = \frac{U_*(n)}{S(n)}$$

where $U_*(n) = \max_i \{ \hat{p}_{in}(n) - \hat{p}_{i,out}(n) \}$ with max taken over all possible community assignments $i = 1, \ldots, N_{n,K}$; $S(n) = \bar{p}(n)$ and

$$\bar{p}(n) = \frac{1}{n \choose 2} \sum_{i>j} P_{ij}(n).$$

For the remainder of the proof we suppress all dependencies on $n$, i.e. $\bar{p}(n) = \bar{p}$. We can use DeMorgan’s Law to show

$$\{ S \geq \bar{p}/(1 + \epsilon) \} \cap \{ U_* \leq k \} \subseteq \{ T_* \leq c \}$$

$$\iff \{ S < \bar{p}/(1 + \epsilon) \} \cup \{ U_* > k \} \supseteq \{ T_* > c \}. \quad (22)$$

Thus,

$$P(T_* > c) \leq P(U_* > k) + P(S < \bar{p}/(1 + \epsilon)).$$

Thus, the general strategy of the proof is to show that under $H_0$, the first term on the right-hand side is at most $\alpha$ and the second term goes to 0.

Under $H_0$, $\bar{p} = p$. Now, we will need a more explicit form of $c$ for the second part of the proof so consider some fixed community assignment that yields $K$ communities for a network with $n = m_1 + \cdots + m_K$ nodes where $K$ is known and there are $m_j > 1$ nodes in community $j$ for $j = 1, \ldots, K$. By A2, at least two communities have size $O(n)$ which we will choose as $m_1, m_2$. There are $N_{n,K} \leq K^n$ such assignments and the algorithm will return the assignment that results in the maximum test statistic.
Now, consider the $U_*$ part of the test statistic. Consider a fixed community assignment $i$ for $i \in \{1, \ldots, N_{n,K}\}$ and let $M_{i,in}, M_{i,out}$ be the number of edges between nodes in the same community and between nodes in a different community, respectively, for this particular assignment. Then

$$M_{i,in} \sim \text{Binomial} \left( m_{i,in} = \sum_{j=1}^{K} \binom{m_j}{2}, p \right)$$

$$M_{i,out} \sim \text{Binomial} \left( m_{i,out} = \binom{n}{2} - \sum_{j=1}^{K} \binom{m_j}{2}, p \right)$$

where $M_{i,in} \perp M_{i,out}$ and $p \to 0$. Thus, we have

$$T_i = \frac{U_i}{S}$$

where

$$U_i = \frac{M_{i,in}}{m_{i,in}} - \frac{M_{i,out}}{m_{i,out}} := X_i - Y_i$$

where $X_i, Y_i$ are independent, scaled binomials and

$$S = \frac{1}{\binom{n}{2}} \sum_{i>j} A_{ij} \sim \frac{1}{\binom{n}{2}} \text{Binomial} \left( \binom{n}{2}, p \right).$$

Since the algorithm returns $T_* = \max_i T_i$, let’s bound the $(1 - \alpha/N_{n,K})$ %tile of the distribution of $U_i$ which we will use for $k$. By Hoeffding’s inequality, we can show

$$P \left( M_{i,in} \geq (p + \frac{\varepsilon}{2}) \cdot m_{i,in} \right) = P(X_i \geq p + \frac{\varepsilon}{2}) \leq e^{-\varepsilon^2 m_{i,in}/2}$$

and

$$P(Y_i \leq p - \frac{\varepsilon}{2}) \leq e^{-\varepsilon^2 m_{i,out}/2}$$

Then using DeMorgan’s Law, we have

$$\{X_i \leq p + \frac{\varepsilon}{2}\} \cap \{Y_i \geq p - \frac{\varepsilon}{2}\} \subseteq \{T_i \leq \varepsilon\} \iff \{X_i \geq p + \frac{\varepsilon}{2}\} \cup \{Y_i \leq p - \frac{\varepsilon}{2}\} \supseteq \{T_i \geq \varepsilon\}$$

We can now bound the necessary $(1 - \alpha/N_{n,K})$ %tile, $k$. To do so, we first must find the lower bounds of $m_{i,in}$ and $m_{i,out}$. For $m_{i,in}$, we want to minimize $m_{in} = \frac{1}{2} \sum_{j=1}^{K} m_j(m_j - 1)$ subject to $\sum_{j=1}^{K} m_j = n$. We can use Lagrange multipliers:

$$\mathcal{L}(m_1, \ldots, m_K, \lambda) = \frac{1}{2} \sum_{j=1}^{K} m_j(m_j - 1) - \lambda \left( \sum_{j=1}^{K} m_j - n \right)$$

Take the gradient:

$$\nabla \mathcal{L}(m_j, \lambda) = \left( m_1 - \frac{1}{2} - \lambda, \ldots, m_K - \frac{1}{2} - \lambda, n - \sum_{j=1}^{K} m_j \right)$$

24
Setting equal to 0 means that for all \( j \), \( m_j = \lambda + \frac{1}{2} \) so

\[
0 = n - \sum_{j=1}^{K} (\lambda + \frac{1}{2}) \implies \lambda = \frac{n}{K} - \frac{1}{2}.
\]  

(33)

Thus, \( m_{i,in} \) is minimized at \( m_1 = \cdots = m_k = \frac{n}{K} \) so

\[
m_{i,in} \geq \frac{1}{2} \sum_{j=1}^{K} \frac{n}{K}(\frac{n}{K} - 1) = \frac{n(n - K)}{2K}.
\]  

(34)

For \( m_{i,out} \),

\[
m_{i,out} = m_1(m_2 + \cdots + m_K) + m_2(m_3 + \cdots + m_K) + \cdots + m_{K-1}m_K
\]  

(35)

\[
\geq m_1m_2
\]  

(36)

\[
\geq n^2.
\]  

(37)

Lastly,

\[
\frac{\alpha}{N_{n,K}} = P(U_i \geq k)
\]  

(38)

\[
\leq P \left( \{X_i \geq p + \frac{k}{2}\} \cup \{Y_i \leq p - \frac{k}{2}\} \right)
\]  

(39)

\[
\leq P(X_i \geq p + \frac{k}{2}) + P(Y_i \leq p - \frac{k}{2})
\]  

(40)

\[
\leq e^{-k^2m_{in}^2/2} + e^{-k^2m_{out}^2/2}
\]  

(41)

\[
\leq 2e^{-k^2n^2/2}
\]  

(42)

\[
\implies k \leq \sqrt{\frac{2(\log(2N_{n,K}) - \log \alpha)}{n^2}} \leq \sqrt{\frac{2(\log(2K) - \frac{1}{n} \log \alpha)}{n}}.
\]  

(43)

Notice that \( k \to 0 \) as \( n \to \infty \) at a rate of \( O(n^{-1/2}) \) which will be important for the second part of the proof. Now, from the algorithm, \( U_* = \max_i U_i \) so using Bonferroni’s inequality we have:

\[
P_{H_0}(U_* \notin R) = P(U_* > k) = P \left( \bigcup_{i=1}^{N_{n,K}} U_i > k \right) \leq \sum_{i=1}^{N_{n,K}} P(U_i > k) = \sum_{i=1}^{N_{n,K}} \frac{\alpha}{N_{n,K}} = \alpha.
\]  

(44)

Next, consider \( S \). First, notice that

\[
\frac{p}{1 + \varepsilon} = p - \frac{\varepsilon}{1 + \varepsilon}p
\]  

(45)

Then, by Hoeffding’s inequality, we can show

\[
P(S < p/(1 + \varepsilon)) = P(S < p - \frac{\varepsilon}{1+\varepsilon}p) \leq e^{-\varepsilon^2p^2n(n-1)/(1+\varepsilon)^2} \to 0
\]  

(46)

(47)

(48)
since \(np \rightarrow \infty\) by A1. Thus, combining the previous work we have
\[
P(T_\ast > c) \leq P(U_\ast > k) + P(S < p/(1+\varepsilon))
\leq \alpha + \eta
\]
where \(\eta \rightarrow 0\) as we hoped to show.

Under \(H_1\), the homophily conditions implicitly assumes a block model with \(K\) communities. Let \(c_{\text{true}}\) be the true community structure, and let \(T_{\text{oracle}} = T(c_{\text{true}}, A)\) be the oracle version of the test statistic with the true community assignment. Since \(T_\ast = \max_c T(c, A)\), we have \(T_\ast \geq T_{\text{oracle}}\). We can also assume that \(T_{\text{oracle}} > 0\) by how the test statistic is constructed. We will use a similar approach to the proof under \(H_0\) noting that
\[
\{U_{\text{oracle}} > \frac{c_n}{p/(1+\varepsilon)}\} \cap \{S < \frac{p}{1-\varepsilon}\} \subseteq \{T_{\text{oracle}} > \frac{c_n}{p/(1+\varepsilon)}\}.
\]
so
\[
P(T_{\text{oracle}} > \frac{c_n}{p/(1+\varepsilon)}) \geq P(\{U_{\text{oracle}} > \frac{c_n}{p/(1+\varepsilon)}\} \cap \{S < \frac{p}{1-\varepsilon}\})
\geq P(U_{\text{oracle}} > c_{\frac{1+\varepsilon}{1-\varepsilon}}) + P(S < \frac{p}{1-\varepsilon}) - 1.
\]
Thus, we want to show that the first two terms on the right hand side of the inequality go to 1.

There is a different probability \(P_{ij}\) for every node pair \(i, j\) but we have the constraint that for two communities \(1 \leq u, v \leq K\)
\[
\bar{p}_u := \frac{1}{\binom{n_u}{2}} \sum_{i,j; c_i=u, c_j=u} P_{ij} \quad \bar{p}_{uv} := \frac{1}{n_u n_v} \sum_{i,j; c_i=u, c_j=v} P_{ij}
\]
Now, we let
\[
\Delta p = \frac{1}{K} \sum_{k=1}^{K} \bar{p}_k - \frac{2}{(K-1)(K-2)} \sum_{j \neq k} \bar{p}_{jk} := \bar{p}_{\text{in}} - \bar{p}_{\text{out}} > 0
\]
Then
\[
M_{\text{in}} \sim \sum_{u=1}^{K} \sum_{i,j; c_i=u, c_j=u} \text{Bernoulli}(P_{ij})
\]
\[
M_{\text{out}} \sim \sum_{u \neq v} \sum_{i,j; c_i=u, c_j=v} \text{Bernoulli}(P_{ij}).
\]
From our testing procedure, we let
\[
U_{\text{oracle}} = \hat{p}_{\text{in}} - \hat{p}_{\text{out}} = \frac{M_{\text{in}}}{m_{\text{in}}} - \frac{M_{\text{out}}}{m_{\text{out}}}
\]
where
\[
m_{\text{in}} = \sum_{j=1}^{K} \binom{m_j}{2}
\]
\[
m_{\text{out}} = \binom{n}{2} - \sum_{j=1}^{K} \binom{m_j}{2}.
\]
Then we can write
\[ P(U_{oracle} \leq k\frac{1+\varepsilon}{1-\varepsilon}) = P(|\hat{p}_{in} - \hat{p}_{out}| \leq k\frac{1+\varepsilon}{1-\varepsilon}) \]
\[ = P(|\hat{p}_{in} + \hat{p}_{in} - \hat{p}_{out} + \hat{p}_{out}| \leq k\frac{1+\varepsilon}{1-\varepsilon}) \]
\[ = P(|- (\hat{p}_{in} - \hat{p}_{in}) - (\hat{p}_{out} - \hat{p}_{out}) + (\hat{p}_{in} - \hat{p}_{out})| \leq k\frac{1+\varepsilon}{1-\varepsilon}) \]
\[ \leq P(|\hat{p}_{in} - \hat{p}_{out}| - |\hat{p}_{in} - \hat{p}_{in}| - |\hat{p}_{out} - \hat{p}_{out}| \leq k\frac{1+\varepsilon}{1-\varepsilon}) \]
\[ = P(|\hat{p}_{in} - \hat{p}_{in}| + |\hat{p}_{out} - \hat{p}_{out}| \geq |\hat{p}_{in} - \hat{p}_{out}| - k\frac{1+\varepsilon}{1-\varepsilon}) \]
\[ \leq P(|\hat{p}_{in} - \hat{p}_{in}| \geq \frac{1}{2}(|\hat{p}_{in} - \hat{p}_{out}| - k\frac{1+\varepsilon}{1-\varepsilon})) + P(|\hat{p}_{out} - \hat{p}_{out}| \geq \frac{1}{2}(|\hat{p}_{in} - \hat{p}_{out}| - k\frac{1+\varepsilon}{1-\varepsilon})) \]

Since \(\hat{p}_{in}, \hat{p}_{out}\) are the sample means of independent Bernoulli’s, we can use Hoeffding’s Second Inequality to show
\[ P(|\hat{p}_{in} - \hat{p}_{in}| \geq \varepsilon) \leq 2e^{-2m_{in}\varepsilon^2/m_{in}} = 2e^{-2m_{in}\varepsilon^2} \quad (61) \]
\[ P(|\hat{p}_{out} - \hat{p}_{out}| \geq \varepsilon) \leq 2e^{-2m_{out}\varepsilon^2/m_{out}} = 2e^{-2m_{out}\varepsilon^2} \quad (62) \]

Thus,
\[ P(U_{oracle} \leq k\frac{1+\varepsilon}{1-\varepsilon}) \leq 2e^{-m_{in}(\hat{p}_{in} - \hat{p}_{out} - k\frac{1+\varepsilon}{1-\varepsilon})^2/2} + 2e^{-m_{out}(\hat{p}_{in} - \hat{p}_{out} - k\frac{1+\varepsilon}{1-\varepsilon})^2/2} \quad (63) \]
\[ \leq 4e^{-n^2(\hat{p}_{in} - \hat{p}_{out} - k\frac{1+\varepsilon}{1-\varepsilon})^2/2}. \quad (64) \]

Now, \(\hat{p}_{in} - \hat{p}_{out} \to 0\) at a rate slower than \(k \to 0\) by A4 so we have \(\hat{p}_{in} - \hat{p}_{out} - k\frac{1+\varepsilon}{1-\varepsilon} > 0\) which we need for Hoeffding’s inequality. Also, \(n(\hat{p}_{in} - \hat{p}_{out}) \to \infty\) by A4. Thus,
\[ P(U_{oracle} \leq k\frac{1+\varepsilon}{1-\varepsilon}) \to 0 \quad (65) \]

or equivalently,
\[ P(U_{oracle} > k\frac{1+\varepsilon}{1-\varepsilon}) \to 1. \quad (66) \]

Next, consider \(S\). First, notice that
\[ \frac{\bar{p}}{1-\varepsilon} = \bar{p} + \frac{\varepsilon}{1-\varepsilon}\bar{p} \quad (67) \]

Then, by Hoeffding’s inequality, we can show
\[ P(S \geq \bar{p}/(1-\varepsilon)) = P(S \geq \bar{p} + \frac{\varepsilon}{1+\varepsilon}\bar{p}) \leq e^{-\varepsilon^2(\bar{p})^2/(1+\varepsilon)^2n(n-1)} \to 0 \quad (68) \]
\[ \geq P(S \geq \bar{p}/(1-\varepsilon)) \to 1. \quad (69) \]

since \(n\bar{p} \to \infty\) by A1. This is equivalent to
\[ P(S < \bar{p}/(1-\varepsilon)) \to 1. \quad (70) \]

Thus,
\[ P(T_{oracle} > \frac{k_n}{\bar{p}/(1+\varepsilon)}) \geq P(\{U_{oracle} > k\frac{1+\varepsilon}{1-\varepsilon}\} \cap \{S < \frac{\bar{p}}{1-\varepsilon}\}) \geq P(U_{oracle} > k\frac{1+\varepsilon}{1-\varepsilon}) + P(S < \frac{\bar{p}}{1-\varepsilon}) - 1 \to 1 + 1 - 1 = 1. \]

27