Detecting Structure in Graphical Data

Lawrence Tray
Ioannis Kontoyiannis

January 16, 2021

Abstract

So the agenda for today. I will take you through my exploratory work around the field in rough chronological order. We'll start off with the fundamentals of hypothesis testing. I will then introduce you to the most commonly used graphical model and talk about the ways of verifying structure in a graph with labelled nodes. However, that is not sufficient as we may want to detect structure in an unlabelled graph. Finally I'll talk about the future direction of the project.

Contents

1 Introduction **1**

2 The Stochastic Block Model **1**

3 Verifying Structure **2**

4 Early results **4**

5 Verifying Structure **4**
5.1 Single Sample against known mean 4
5.2 Two samples equality of means 6

6 Detecting Structure **7**
6.1 ABP 7

7 Composite Approaches **7**

8 Appendix **7**
8.1 Proving H.O.T can be neglected 7

1 Introduction

The title started off as "detecting causal structure in time series data" but the emphasis has pivoted to the more generic question: detecting structure in data and specifically graphical data.

There is a wealth of graphical data in the world and more is being produced each second; social networks, website hyperlinks and academic collaborations are just some examples of this data. There is a wealth of algorithms developed to analyse graphical data. Nevertheless, that same principled framework we have for querying classical data (hypothesis testing) is less developed for graphical data. Do my friends vote the same way I do or do researchers collaborate with those of the same gender? We want to answer these questions and not only that, we wish to report our confidence in the answers. To that end there is space to expand the hypothesis testing framework to graphs.

2 The Stochastic Block Model

The most popular graphical model in industry and indeed academia is called the Stochastic Block Model (SBM). We use a definition adapted from Abbe [1].

Definition 2.1 *Let $n \in \mathbb{Z}^+$ be the number of vertices and $k \in \mathbb{Z}^+$ be the number of communities in an SBM graph. We define a probability vector $\pi = [\pi_1, \pi_2 \dots \pi_K]^T$ to be the prior on the K-communities. Each vertex $v \in \mathcal{V} = \{1, 2 \dots N\}$ has a community label $X_v \in \{1, 2 \dots n\}$. Let W be a symmetric $k \times k$ matrix called the connectivity matrix. We say that the pair*

$(X, \mathcal{G}) \sim SBM(n, \pi, W)$ if X is an N-dimensional vector with each component independently distributed as the community prior $X_v \sim \pi$ and $\mathcal{G}$ is an N-vertex graph where each pair of vertices (i, j) is connected with probability $p(i \leftrightarrow j) = W_{X_i, X_j}$ independently of other pairs of vertices. Lastly, we define the community sets as $\Omega_i = \Omega_i(X) \coloneqq \{v \in \mathcal{V} : X_v = i\}$ which groups our partitions the vertex-set.

Obviously this definition imposes further constraints on the values that the connectivity matrix W namely: $0 \leq W_{ij} \leq 1$. Though the definition of the SBM is simple, it allows for very deep and rich analysis of graphical datasets.

3 Verifying Structure

Armed with this definition we tackle the simplest problem in structure verification. Given a graph $\mathcal{G}$ and vertex-labels X, we wish to determine whether the two communities a and b connect differently. Put formally this a hypothesis test on the parameters of W. There are three parameters we would wish to test: W_{aa}, W_{ab} and W_{bb} (note that the symmetry constraint requires $W_{ab} = W_{ba}$). To do this we can perform three-pairwise hypothesis tests. Here we test W_α against W_β where α and β are unique indices in $\{(a, a), (a, b), (b, b)\}$:

$$\begin{aligned} H_0 : &\quad W_\alpha = W_\beta \\ H_1 : &\quad W_\alpha \neq W_\beta \end{aligned} \tag{1}$$

We formulate this as a likelihood ratio test. Letting $\mathcal{L}(\mathcal{D}|H)$ denote the likelihood of observing the data $\mathcal{D} = (X, G)$ under hypothesis H. For ease of notation. Therefore, the test statistic is given by:

$$t_n \coloneqq \log \frac{\mathcal{L}(\mathcal{D}|H_1)}{\mathcal{L}(\mathcal{D}|H_0)} \tag{2}$$

At this point it helps to introduce some more notation. We define $n_i \coloneqq |\Omega_i(X)|$ leading to the result $n = \sum_i n_i$. Furthermore, we define $E_{ij} = E_{ij}(X, \mathcal{G})$ to denote the number of realised edges between communities i and j (in generality i may be equal to j). Furthermore, define $M_{ij} = M_{ij}(X)$ as the maximum number of possible edges between the communities. For an undirected graph this can be computed simply as follows:

$$M_{ij} = M_{ij}(X) = \begin{cases} n_i n_j & \text{for } i \neq j \\ \frac{1}{2} n_i (n_i - 1) & \text{for } i = j \end{cases} \tag{3}$$

With this new notation, likelihood function can be written explicitly:

$$\begin{aligned} \mathcal{L}(\mathcal{D}|H) &= p(X|\pi) \cdot p(\mathcal{G}|W, X) \\ &= p(X|\pi) \cdot \prod_{i=1}^{k} \prod_{j=i}^{k} p(E_{ij}|W, X) \\ &= p(X|\pi) \cdot \prod_{i=1}^{k} \prod_{j=i}^{k} W_{ij}^{E_{ij}} \cdot (1 - W_{ij})^{(M_{ij} - E_{ij})} \end{aligned} \tag{4}$$

The form of $p(\mathcal{G}|W, X)$ is simply a sequence of Bernoulli trials for each distinct community pair (i, j) (edge present with probability W_{ij} or edge absent with probability $1 - W_{ij}$ for every pair of vertices across those communities). A sequence of Bernoullis is the same as a Binomial distribution without the combinatoric term. By inspecting equation 4 we see that only terms involving W_α and W_β are going to differ under the two hypotheses; the rest of the terms will cancel in our calculation. Therefore, we can rewrite the likelihood as follows:

$$\mathcal{L}(\mathcal{D}|H) \propto f(W_\alpha, E_\alpha, M_\alpha) \cdot f(W_\beta, E_\beta, M_\beta) \tag{5}$$

$$\text{where } f(w, e, m) \coloneqq w^e \cdot (1 - w)^{(m-e)} \tag{6}$$

We note that $f(w, e, m)$ is simply the probability of observing a specific sequence of e successes in m independent Bernoulli trials with parameter w. Its maximiser with respect to the first argument is easily computed through partial differentiation giving:

$$\arg\max_w f(w, e, m) = \hat{w} = e/m \tag{7}$$

Furthermore, we spot the following property $f(w, e_1, m_1) \cdot f(w, e_2, m_2) = f(w, e_1 + e_2, m_1 + m_2)$ or in other words, the function f is linear in its second and third arguments given the same first argument. As such we can manipulate equation 2 greatly to give:

$$\begin{aligned} t_n &= \log \frac{\max_{W_\alpha \neq W_\beta}(f(W_\alpha, E_\alpha, M_\alpha) \cdot f(W_\beta, E_\beta, M_\beta))}{\max_{W_\alpha = W_\beta}(f(W_\alpha, E_\alpha, M_\alpha) \cdot f(W_\beta, E_\beta, M_\beta))} \\ &= \log \frac{\max_p f(p, E_\alpha, M_\alpha) \cdot \max_q f(q, E_\alpha, M_\alpha)}{\max_r f(r, E_\alpha + E_\beta, M_\alpha + M_\beta)} \\ &= \log \frac{f(\hat{p}, E_\alpha, M_\alpha) \cdot f(\hat{q}, E_\alpha, M_\alpha)}{f(\hat{r}, E_\alpha + E_\beta, M_\alpha + M_\beta)} \\ &= \log \frac{f(\hat{p}, E_\alpha, M_\alpha)}{f(\hat{r}, E_\alpha, M_\alpha)} + \log \frac{f(\hat{q}, E_\beta, M_\beta)}{f(\hat{r}, E_\beta, M_\beta)} \end{aligned} \tag{8}$$

Where $\hat{p} \coloneqq E_\alpha / M_\alpha$, $\hat{q} \coloneqq E_\beta / M_\beta$ and $\hat{r} \coloneqq (E_\alpha + E_\beta)/(M_\alpha + M_\beta)$. These symbols are introduced to make the notation more succinct.

Lemma 3.1 *With f defined as in equation 6, $0 \leq e \leq m$ and $r \in [0, 1]$ then it holds that $\log \frac{f(e/m, e, m)}{f(r, e, m)} = m \cdot \mathcal{D}\left(Bern(e/m) || Bern(r)\right)$ where $\mathcal{D}(g||h)$ is the Kullback-Leibler divergence between two probability mass functions $g, h : \mathcal{X} \mapsto [0, 1]$ defined in discrete space as $\mathcal{D}(g||h) \coloneqq \sum_{x \in \mathcal{X}} g(x) \log \frac{g(x)}{h(x)}$ and Bern(p) denotes the Bernoulli p.m.f with parameter p.*

Proving lemma 3.1 is simply a case of algebraic manipulation:

$$\begin{aligned} \log \frac{f(e/m, e, m)}{f(r, e, m)} &= e \cdot \log \frac{e/m}{r} + (m - e) \cdot \log \frac{1 - e/m}{1 - r} \\ &= m \cdot \left((e/m) \cdot \log \frac{e/m}{r} + (1 - e/m) \cdot \log \frac{1 - e/m}{1 - r} \right) \\ &= m \sum_{x \in \{0,1\}} \text{Bern}(x; e/m) \cdot \log \frac{\text{Bern}(x; e/m)}{\text{Bern}(x; r)} \\ &= m\mathcal{D}\left(\text{Bern}(e/m) || \text{Bern}(r)\right) \quad \therefore \text{QED} \end{aligned} \tag{9}$$

Thereby proving lemma 3.1. This allows us to simplify the test-statistic into a form that is more numerically stable:

$$t_n = M_\alpha \cdot \mathcal{D}\left(\text{Bern}(\hat{p}) || \text{Bern}(\hat{r})\right) + M_\beta \cdot \mathcal{D}\left(\text{Bern}(\hat{q}) || \text{Bern}(\hat{r})\right) \tag{10}$$

Lemma 3.2 *For $p \approx r$ then*

Theorem 3.3 *We posit that $(X, \mathcal{G}) \sim SBM(n, p, W)$. Given the realised graph and class labels $(X, \mathcal{G})$ we can perform a hypothesis test on parameters W_α and W_β of the connectivity matrix W.*

$$\begin{aligned} H_0 : &\quad W_\alpha = W_\beta \\ H_1 : &\quad W_\alpha \neq W_\beta \end{aligned}$$

If the log-likelihood ratio test statistic t_n is computed as in equation 10, repeated here:

$$t_n \coloneqq M_\alpha \cdot \mathcal{D}\left(Bern(\hat{p}) || Bern(\hat{r})\right) + M_\beta \cdot \mathcal{D}\left(Bern(\hat{q}) || Bern(\hat{r})\right)$$

Where $\hat{p} \coloneqq E_\alpha / M_\alpha$, $\hat{q} \coloneqq E_\beta / M_\beta$ and $\hat{r} \coloneqq (E_\alpha + E_\beta)/(M_\alpha + M_\beta)$. Then as the number of vertices $n \to \infty, t_n \sim \frac{1}{2}\chi^2_1$ under the null H_0. Therefore, we reject H_0 at the $100(\zeta)\%$ confidence level if and only if $2t_n \geq \psi^{-1}(\zeta)$, where ψ^{-1} is the χ^2_1 inverse cdf satisfying $Pr(Y \leq \psi^{-1}(\zeta)) = \zeta$ given $Y \sim \chi^2_1$.

Corollary 3.3.1 *We can also use a slightly simpler test-statistic z_n albeit with some loss of generality. If we define z_n to be:*

$$z_n \coloneqq \sqrt{\frac{M_\alpha M_\beta}{\hat{r}(1 - \hat{r})(M_\alpha + M_\beta)}} \, (\hat{p} - \hat{q})$$

With the symbols retaining their previous meanings. Then we show that under the null H_0, as $n \to \infty$ then $z_n \sim \mathcal{N}(0, 1)$. Meaning that, we can construct a similar test to reject H_0 at the $100(\zeta)\%$ confidence level if and only if $|z_n| \geq \phi^{-1}(\zeta)$, where ϕ^{-1} is the standard inverse Gaussian cdf on magnitude satisfying $Pr(|Y| \leq \phi^{-1}(\zeta)) = \zeta$ given $Y \sim \mathcal{N}(0, 1)$

4 Early results

We seek to apply theorem 3.3 to various real-world graphical datasets. We start by analysing social network graphs. The Stanford Network Analysis Project (SNAP for short) [5] offers a wealth of Facebook egonets. An egonet is simply a graph where all vertices (in this case each representing a Facebook user) are guaranteed to be connected to one central node (the ego-node). The data consists of the undirected set of edges $\mathcal{G}$ indicating whether any two vertices (Facebook users) are connected (friends on Facebook). We also have a set of binary labels for each vertex X. However, for the sake of privacy these features are anonymised. This is best explained through illustration:

```
75 first_name;anonymized feature 75
76 first_name;anonymized feature 76
77 gender;anonymized feature 77
78 gender;anonymized feature 78
79 hometown;id;anonymized feature 79
```

If we have a total of f features and n vertices, then the feature matrix X would be an $n \times f$ where each row is the feature vector for the corresponding vertex. The features are binary such that $X_{ij} \in \{0, 1\}$ indicating the feature of and on respectively. However, each feature is anonymised to avoid giving away personally identifiable information. I cannot tell whether vertex v is male or female but I can tell whether they are the same gender as another vertex w.

5 Verifying Structure

5.1 Single Sample against known mean

We start with the simple case of determining whether or not a coin is fair. Each coin flip can be represented by a random variable with a Bernoulli distribution $X_i \sim Bern(p)$. Each coin can result in either a Tails or a Heads denoted by $X_i \in \{0, 1\}$. We toss the coin n times leading us to a set $\{X_i\}_{i=1}^n$. We wish to test the null hypothesis $p = p_0$ against the alternative. We shall keep the p_0 notation for generality, though for a fair coin we require $p_0 = 1/2$.

$$\begin{aligned} H_0 : &\quad p = p_0 \\ H_1 : &\quad p \neq p_0 \end{aligned}$$

We denote the number of heads with the random variable $K \coloneqq \sum_{i=1}^n X_i \sim Bern(p, n)$. For a particular experiment we observe $K = k$. We employ a standard likelihood ratio test. The test statistic t_n is calculated as the log-likelihood ratio of observing $K = k$ under H_1 and H_0.

$$t_n \coloneqq \log \frac{\mathcal{L}(H_1)}{\mathcal{L}(H_0)} = \log \frac{\max_p P(K = k|p)}{P(K = k|p = p_0)} \tag{11}$$

We note that since K is distributed as a binomial, $P(K = k|p) = \binom{n}{k} p^k (1 - p)^{n-k}$. If we can vary p, this probability is maximised for $p = \hat{p} \coloneqq k/n$. Therefore, the test statistic is given by.

$$t_n = \log \frac{\binom{n}{k} \hat{p}^k (1 - \hat{p})^{n-k}}{\binom{n}{k} p_0^k (1 - p_0)^{n-k}} = \log \frac{\hat{p}^k (1 - \hat{p})^{n-k}}{p_0^k (1 - p_0)^{n-k}} \tag{12}$$

The combinatoric term $\binom{n}{k}$ cancels out. This implies that the order in which the heads land does not matter when determining the fairness of the coin. We can work the above expression into a more usable form:

$$\begin{aligned} t_n =& k \log \frac{\hat{p}}{p_0} + (n - k) \log \frac{1 - \hat{p}}{1 - p_0} \\ =& n \left(\hat{p} \log \frac{\hat{p}}{p_0} + (1 - \hat{p}) \log \frac{1 - \hat{p}}{1 - p_0} \right) \\ =& n \mathcal{D} \left(Bern(\hat{p}) || Bern(p_0) \right) \end{aligned} \tag{13}$$

Where $\mathcal{D}(f|g) \coloneqq \sum_{x \in \mathcal{X}} f(x) \log \frac{f(x)}{g(x)}$ is the Kullback-Leibler divergence between two arbitrary probability mass functions f and g. This is also called the relative entropy. The KL divergence has the property that:

$$\mathcal{D}(f||g) \geq 0 \quad \text{with equality iff} \quad f(x) = g(x) \quad \forall x \in \mathcal{X} \tag{14}$$

We can exploit this result for the case that $f \approx g$ to obtain a simplified expression for the KL divergence. We begin by defining $\delta(x) \coloneqq f(x) - g(x)$. We are interested in the region where δ is small. We start by substituting for $f = \delta + g$ and

then taking the Taylor expansion of $\log 1 + x$.

$$\begin{aligned}
\mathcal{D}(f||g) &= \sum_{x\in\mathcal{X}} (\delta + g) \log\left(1 + \frac{\delta}{g}\right) \\
&= \sum_{x\in\mathcal{X}} (\delta + g)\left(\frac{\delta}{g} - \frac{\delta^2}{2g^2} + O(\delta^3)\right) \\
&= \sum_{x\in\mathcal{X}} \delta + \frac{1}{2}\sum_{x\in\mathcal{X}} \frac{\delta^2}{g} + O(\delta^3) \\
&= \frac{1}{2}\sum_{x\in\mathcal{X}} \frac{\delta^2}{g} + O(\delta^3) \\
&= \frac{1}{2}\chi^2(f||g) + O(\delta^3)
\end{aligned}$$

Where the summation over δ evaluates to 0 because δ is the difference of two valid p.m.f's which each sum to 1 over $x \in \mathcal{X}$. We are able to neglect the $O(\delta^3)$ terms for f very close to g we shall see what this means later. $\chi^2(f||g)$ is known as the chi-squared distance between two distributions and is defined simply as $\chi^2(f||g) \coloneqq \sum_{x\in\mathcal{X}}(f-g)^2/g$. We now investigate the chi-squared divergence for $f = Bern(p)$ and $g = Bern(q)$.

$$\begin{aligned}
\chi^2(Bern(p)||Bern(q)) &= \frac{(p-q)^2}{q} + \frac{((1-p)-(1-q))^2}{1-q} \\
&= \frac{(p-q)^2}{q(1-q)} \\
&= \left(\frac{p-q}{\sqrt{q(1-q)}}\right)^2
\end{aligned}$$

Now we can exploit these results to get a workable expression for the test statistic t_n.

$$\begin{aligned}
t_n &= n\mathcal{D}\left(Bern(\hat{p})||Bern(p_0)\right) \\
&= \frac{n}{2}\chi^2\left(Bern(\hat{p})||Bern(p_0)\right) + nO(\delta^3) \\
&= \frac{1}{2}\left(\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}\right)^2 + \epsilon
\end{aligned}$$

So far we have been treating t_n as deterministic but it is merely an observation a random variable. To make this distinction clear we shall use upper-case to refer to random variables. Therefore, we have that the random variable T_n is a function of $\hat{P} \coloneqq K/n$. For large n, we can find the distribution of $\hat{P}$ by the Central Limit Theorem.

$$\begin{aligned}
\hat{P} = \frac{K}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i \\
H_0 : \mathbb{E}[X_i] = p_0, Var(X_i) = p_0(1-p_0) \\
\text{CLT, as } n\to\infty : \hat{P} \sim \mathcal{N}\left(\mu = p_0, \sigma^2 = p_0(1-p_0)/n\right) \\
\therefore \frac{\hat{P}-p_0}{\sqrt{p_0(1-p_0)/n}} = Z \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)
\end{aligned}$$

Therefore neglecting the error term ϵ[1], we have that under H_0, for sufficiently large n.

$$T_n = \frac{1}{2}Z^2 \sim \frac{1}{2}\chi_1^2 \tag{15}$$

By the definition of the chi-squared distribution with one degree of freedom. To reject the null hypothesis H_0 at the $100(1-\alpha)\%$ confidence level, we require that $P(T_n \geq t_n|H_0) < \alpha$. In other words, a low probability of observing this result under the null hypothesis.

[1]See the Appendix for proof of negligibility

5.2 Two samples equality of means

We now complicate things by introducing a second coin, with throws denoted by $\{Y_i\}_{i=1}^m$ where we assume each throw is i.i.d Bernoulli with parameter q ($Y_i \sim Bern(q)$). Note that the population sizes n and m may be different. We define $L \coloneqq \sum_{i=1}^m Y_i$ which is the analogue of K. We now set up our hypotheses to be:

$$\begin{aligned} H_0 : &\quad p = q \\ H_1 : &\quad p \neq q \end{aligned}$$

Proceeding as before we can derive a formula for the test statistic. This time we denote the test statistic by t_N where $N \coloneqq n + m$.

$$\begin{aligned} t_N &\coloneqq \log \frac{\mathcal{L}(H_1)}{\mathcal{L}(H_0)} = \log \frac{\max_{p,q} P(K = k|p)P(L = l|q)}{\max_p P(K = k|p)P(L = l|q = p)} \\ &= \log \frac{\binom{n}{k}\hat{p}^k(1-\hat{p})^{n-k}\binom{m}{l}\hat{q}^k(1-\hat{q})^{m-l}}{\binom{n}{k}\hat{r}^k(1-\hat{r})^{n-k}\binom{m}{l}\hat{r}^l(1-\hat{r})^{m-l}} \\ &= \log \frac{\hat{p}^k(1-\hat{p})^{n-k}}{\hat{r}^k(1-\hat{r})^{n-k}} + \log \frac{\hat{q}^k(1-\hat{q})^{m-l}}{\hat{r}^l(1-\hat{r})^{m-l}} \end{aligned}$$

Where, $\hat{p} \coloneqq k/n, \hat{q} \coloneqq l/m, \hat{r} \coloneqq (k+l)/(n+m) = (n\hat{p} + m\hat{q})/(n+m)$. Using the same tricks as before, we can express this in terms of the chi-squared distance between the various parameters:

$$\begin{aligned} t_N &= n\mathcal{D}\left(Bern(\hat{p})||Bern(\hat{r})\right) + m\mathcal{D}\left(Bern(\hat{q})||Bern(\hat{r})\right) \\ &\approx \frac{n}{2}\chi^2\left(Bern(\hat{p})||Bern(\hat{r})\right)) + \frac{m}{2}\chi^2\left(Bern(\hat{q})||Bern(\hat{r})\right)) \\ &= \frac{1}{2\hat{r}(1-\hat{r})}\left(n(\hat{p}-\hat{r})^2 + m(\hat{q}-\hat{r})^2\right) \\ &= \frac{1}{2\hat{r}(1-\hat{r})}\left(n\left(\frac{m(\hat{p}-\hat{q})}{n+m}\right)^2 + m\left(\frac{n(\hat{q}-\hat{p})}{n+m}\right)^2\right) \\ &= \frac{nm(\hat{p}-\hat{q})^2}{2\hat{r}(1-\hat{r})(n+m)} \\ &= \frac{1}{2}\left(\frac{(\hat{p}-\hat{q})}{\sqrt{\hat{r}(1-\hat{r})(1/n+1/m)}}\right)^2 \end{aligned}$$

Under the null hypothesis H_0, we require $p = q(= \mu)$; we introduce this third variable μ to refer to the true mean to avoid ambiguity. Applying the central limit theorem (for sufficiently large n and m) and combining Gaussians in the standard way, we have that:

$$\begin{aligned} \hat{P} &\sim \mathcal{N}\left(\mu, \frac{\mu(1-\mu)}{n}\right) \\ \hat{Q} &\sim \mathcal{N}\left(\mu, \frac{\mu(1-\mu)}{m}\right) \\ \hat{R} &\sim \mathcal{N}\left(\mu, \frac{\mu(1-\mu)}{n+m}\right) \\ \therefore \sqrt{\frac{nm}{\mu(1-\mu)(n+m)}}(\hat{P}-\hat{Q}) &= Z \sim \mathcal{N}(0,1) \\ \delta\hat{R} \coloneqq \hat{R} - \mu &\sim \mathcal{N}\left(0, \frac{\mu(1-\mu)}{n+m}\right) \end{aligned}$$

We almost have T_n we just need to demonstrate that $\hat{R}(1-\hat{R})$ is sufficiently close to $\mu(1-\mu)$ for our purposes.

$$\begin{aligned} \hat{R}(1-\hat{R}) &= (\mu + \delta\hat{R})(1-(\mu+\delta\hat{R})) \\ &= \mu(1-\mu) + O(\delta\hat{R}) \\ \therefore \frac{1}{\hat{R}(1-\hat{R})} &= \frac{1}{\mu(1-\mu)}\left(\frac{1}{1+O(\delta\hat{R})}\right) \\ &= \frac{1}{\mu(1-\mu)}(1+O(\delta\hat{R})) \\ &\approx \frac{1}{\mu(1-\mu)} \end{aligned}$$

We can neglect the terms of order $\delta\hat{R}$ and higher powers, as it is zero mean and for sufficiently large $n+m$ the variance approaches 0. Therefore, we have the desired expression for T_N.

$$T_N \approx \frac{1}{2}\left(\sqrt{\frac{nm}{\mu(1-\mu)(n+m)}}(\hat{P}-\hat{Q})\right)^2 = \frac{1}{2}Z^2 \sim \frac{1}{2}\chi_1^2 \tag{16}$$

For this test we can also use the z-statistic instead of the t-statistic. Since the former is distributed like a Gaussian it may be easier to deal with.

$$z_N = \frac{\hat{p}-\hat{q}}{\sqrt{\hat{r}(1-\hat{r})\,(1/n+1/m)}} \sim \mathcal{N}(0,1) \tag{17}$$

6 Detecting Structure

6.1 ABP

ABP(r, T) on a graph G with vertex set $V = V(G)$ and edge set $E = E(G)$

1. Initialise messages for (v, v'):

   $y^{(0)}_{v'\to v} \leftarrow \mathcal{N}(0,1)$

2. Iterate for $1 \le t \le T$ and for $(v, v') \in E$:

   compute average $s^{(t-1)} \leftarrow \frac{1}{2|E|}\sum_{(v,v')\in E} y^{(t-1)}_{v'\to v}$

   recentre messages $z^{(t-1)}_{v'\to v} \leftarrow y^{(t-1)}_{v'\to v} - s^{(t-1)}$

   sum incoming $y^{(t)}_{v'\to v} \leftarrow \sum_{(v',v'')\in E\setminus\{v\}} z^{(t-1)}_{v''\to v'}$

   if $(v''' \to v \to v')$ on cycle of length $r' \le r$ then correct:

   $y^{(t)}_{v'\to v} \leftarrow y^{(t)}_{v'\to v} - \sum_{(v,v'')\in E\setminus\{v',v'''\}} z^{(t-r')}_{v''\to v'}$

3. Assignment, for all $v \in V$:

   Sum incoming $y^{(T)}_v = \sum_{(v,v')\in E} y^{(t)}_{v'\to v}$

   Assign labels $\sigma_v = 1$ if $y^{(T)}_v > 0$ and 0 otherwise

7 Composite Approaches

8 Appendix

8.1 Proving H.O.T can be neglected

We have shown that under the null hypothesis, that for sufficiently large n, $(\hat{P}-p_0) \sim \mathcal{N}(0, \beta/n)$ for some positive finite constant β (in this case $\beta = p_0(1-p_0)$ but we just require finiteness for this proof). Therefore, $Q \coloneqq \sqrt{n}(\hat{P}-p_0) \sim \mathcal{N}(0,\beta)$. Manipulating our expression for the error term ϵ we can show that it can be expressed as a sum of

$$\begin{gathered}
\epsilon = nO(\delta^3) \\
\therefore |\epsilon| \le n\sum_{i=0}^{\infty}\alpha_i|\hat{P}-p_0|^{3+i} \quad \text{for some finite constants } \alpha_i \ge 0 \\
|\epsilon| \le \sum_{i=0}^{\infty}\alpha_i n^{-\frac{1+i}{2}}(\sqrt{n}|\hat{P}-p_0|)^{3+i} \\
|\epsilon| \le \sum_{i=0}^{\infty}\alpha_i n^{-\frac{1+i}{2}}|Q|^{3+i}
\end{gathered} \tag{18}$$

We know that Q is a Gaussian of zero mean and finite variance, therefore $|Q|$ will be a finite value. However, we see that it is scaled by a negative power of n, therefore for sufficiently large n we have that the error term asymptotes to 0. To be precise:

$$\lim_{n\to\infty} P(|\epsilon| < \eta) = 1 \text{ for arbitrarily small } \eta \ge 0 \tag{19}$$

References

[1] Emmanuel Abbe. "Community Detection and Stochastic Block Models: Recent Developments". In: *Journal of Machine Learning Research* 18.177 (2018), pp. 1–86. URL: http://jmlr.org/papers/v18/16-480.html.

[2] Emmanuel Abbe and Colin Sandon. "Achieving the KS threshold in the general stochastic block model with linearized acyclic belief propagation". In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee et al. Vol. 29. Curran Associates, Inc., 2016, pp. 1334–1342. URL: https://proceedings.neurips.cc/paper/2016/file/6c29793a140a811d0c45ce03c1c93a28-Paper.pdf.

[3] Emmanuel Abbe and Colin Sandon. "Proof of the Achievability Conjectures for the General Stochastic Block Model". In: *Communications on Pure and Applied Mathematics* 71.7 (2018), pp. 1334–1406. DOI: https://doi.org/10.1002/cpa.21719. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpa.21719. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.21719.

[4] Emmanuel Abbe and Colin Sandon. "Recovering Communities in the General Stochastic Block Model Without Knowing the Parameters". In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes et al. Vol. 28. Curran Associates, Inc., 2015, pp. 676–684. URL: https://proceedings.neurips.cc/paper/2015/file/cfee398643cbc3dc5eefc89334cacdPaper.pdf.

[5] Jure Leskovec and Andrej Krevl. *SNAP Datasets: Stanford Large Network Dataset Collection*. http://snap.stanford.edu/data. June 2014.

[6] Jure Leskovec and Rok Sosič. "SNAP: A General-Purpose Network Analysis and Graph-Mining Library". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 8.1 (2016), p. 1.