

Detecting Structure in Graphical Data

Lawrence Tray
Ioannis Kontoyiannis

January 12, 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	Verifying Structure	2
5	Verifying Structure	4
5.1	Single Sample against known mean	4
5.2	Two samples equality of means	5
6	Detecting Structure	6
6.1	ABP	6
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 $\Omega_i = \Omega_i(X) := \{v \in \mathcal{V} : X_v = i\}$ 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 $\mathcal{D} = (X, \mathcal{G})$ under hypothesis H . For ease of notation. Therefore, the test statistic is given by:

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

At this point it helps to introduce some more notation. We define $n_i := |\Omega_i(X)|$ 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 (note that i may be equal to j). Furthermore, define $M_{ij} = M_{ij}(X)$ as the maximum number of possible edges between the communities. The likelihood function can now be rewritten in more simple terms:

$$\begin{aligned} \mathcal{L}(\mathcal{D}|H) &= p(X|\pi) \cdot p(\mathcal{G}|W, X) \\ &= p(X|\pi) \cdot \prod_{i=1}^k \prod_{j=i}^k p(E_{ij}|W, X) \\ &= p(X|\pi) \cdot \prod_{i=1}^k \prod_{j=i}^k W_{ij}^{E_{ij}} \cdot (1 - W_{ij})^{(M_{ij} - E_{ij})} \end{aligned} \tag{3}$$

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 3 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:

$$\mathcal{L}(\mathcal{D}|H) \propto f(W_\alpha, E_\alpha, M_\alpha) \cdot f(W_\beta, E_\beta, M_\beta) \tag{4}$$

$$\text{where } f(w, e, m) := w^e \cdot (1 - w)^{(m-e)} \tag{5}$$

As such we can simplify equation 2 greatly to give:

$$\begin{aligned} t_n &= \log \frac{\max_{W_\alpha \neq W_\beta} (f(W_\alpha, E_\alpha, M_\alpha) \cdot f(W_\beta, E_\beta, M_\beta))}{\max_{W_\alpha = W_\beta} (f(W_\alpha, E_\alpha, M_\alpha) \cdot f(W_\beta, E_\beta, M_\beta))} \\ &= \log \frac{f(E_\alpha/M_\alpha, E_\alpha, M_\alpha) \cdot f(E_\beta/M_\beta, E_\beta, M_\beta)}{f((E_\alpha + E_\beta)/(M_\alpha + M_\beta), E_\alpha + E_\beta, M_\alpha + M_\beta)} \end{aligned} \tag{6}$$

4 Verifying Structure

So hypothesis testing. I'll start with an example that may seem trivial now but will be extremely important later for graphical analysis. This is the test for equality of means. Say we have two coins, X and Y. We assume each flip is an independent Bernoulli trial. Coin X has probability p of coming up heads, Y has parameter q . We flip X, n times and record k heads. Y is flipped m times and l heads.

Given this observed data we want to test the null hypothesis that $p=q$ against the alternative that $p \neq q$.

There is a very general theorem for hypothesis tests which states that the log-likelihood ratio test statistic (this is the likelihood of the observed data under H_1 divided by the likelihood under H_0 all logged) The theorem states that the log-likelihood ratio under the null is distributed like $\frac{1}{2}$ times a chi-squared distribution with p degrees of freedom. Where p is the number of additional degrees of freedom introduced by H_1 . This allows us to compute the test statistic and reject the null hypothesis if the statistic is large enough.

The first thing I did was formally prove that this theorem holds for the coin flipping example we just spoke about. For brevity I'll omit the details of the proof but it essentially manipulates the expression into such a form that we can apply the Central Limit Theorem to show that the term in the brackets is distributed as a standard gaussian. So we now have a robust tool for testing populations for equality of means. Put it in your toolbox for now because we will come back to it later.

So now on how to verify evidence of structure in a given graph. Suppose we have a graph G with known vertex labels. We assume it is drawn from a symmetric SBM with probability p of vertices within the same community being connected and probability q across communities. We say it is symmetric because these p and q are the same for all communities. We wish to test the hypothesis that these parameters are different against the null that they are equal. If the null is rejected then we can say that there is sufficient evidence to show the communities interact with one another differently.

And this is where we reap the dividends from the earlier hypothesis test. Two nodes within a community are connected with probability p (this is the equivalent of coming up heads) and they are disconnected with probability $1-p$. We can reduce this hypothesis test to exactly the one we had before by assigning the parameters as follows.

n is now the number of possible edges within communities m is those possible across k is the number that are realised within communities And l are those realised across

So we now try to apply this test to a real world dataset. The Stanford Network Analysis Project (SNAP) has a wealth of graphical datasets. In particular I analysed a couple of facebook ego-nets. So that is a graph of all my friends and the connections between them. Each vertex is also labelled by a feature vector which is entirely anonymised. I cannot even tell someone's gender only whether it is the same as somebody else's. The graph on the right is one such egonet with all gender-77 vertices coloured orange.

So we apply three hypothesis tests on this graph to determine whether gender impacts how people connect on facebook. In every case, we reject the null with a very high degree of confidence (in one case we even surpass the precision of the chi-squared cdf). We can be certain that gender impacts how people connect but this approach has several limitations.

First of all, we cannot say with certainty how large the impact is. Only that gender does impact graphical structure. Secondly we must be careful to avoid confounding variables. These are variables that we did not account for that could better explain the structure. To give an example, we might find evidence that students in Cambridge are more likely to be friends with those of the same gender. The confounding variable here could be college affiliation especially given the existence of all-female colleges such as Newnham. And lastly these approaches only work if the graph is labelled in advance.

To overcome some of these limitations we now look at the related issue of detecting structure in an unlabeled graph. The question is how well we can recover the communities from the graph. This is dependent on the parameters of the SBM. Each regime here is less strict than the one before. We will focus on weak recovery as that is the regime that most real-life graphs occupy and the same algorithms to solve weak recovery can be applied to more neat graphs.

Formally, weak recovery means that we can assign vertex labels with more accuracy than simply sampling from the community prior. A different way of putting this is that we simply need to partition the graph's vertices into two sets S and the complement of S such that this partition cuts across communities. Formally, the fraction of nodes from a certain community in our set S is different for different communities.

The most tractable weak recovery algorithm is something called Acyclic Belief Propagation. This is a form of message-passing that runs in $O(n \log n)$ on graphs with constant degree. Since we're pressed on time I'll skim over the details. We initialise the messages by drawing from a standard Gaussian. We propagate these messages along the edges of the graph correcting for cycles of length less than r and then take the output to be the sum of all incoming messages. If the output of a vertex is positive we assign it into set S .

So we run ABP on our same ego-net and find that weak recovery is satisfied. We partition the graph into two sets and find that the fraction of orange gender in set S is different to the fraction of purple gender in set S . 0.41 is not equal to 0.47. But these numbers are still quite close to one another so perhaps gender is not the most important feature by which we can partition the graph. And this is where I want to go to next.

So we can run ABP on an unlabelled graph to partition into the two most distinctive sets. But what if we also have the vertex labels. We could then train a binary classifier to map from feature vectors to classification under ABP. This way we will be able to include all feature variables automatically and rank order them by importance.

We would then start to ask more sophisticated questions of the dataset. Are friends made independent of politics given age? Is the gender of a research partner independent given the research topic? We can start to build a much more powerful tool.

Looking slightly longer term, I would like to formalise this approach and link the ABP partition to a robust hypothesis testing framework. I may look to bring in causality and seeing if we cannot start to predict graph evolution. Remember that almost no graph is static, new vertices and edges are being added all the time. I may also extend this approach to more sophisticated graphical models. Indeed, the standard SBM is not well suited to tightly clustered graphs. The geometric SBM introduces an additional parameter to account for the mutual friends phenomenon. You are much more likely to be friends with a friend of your friend than with a random vertex on the graph.

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 $X_i \sim \text{Bern}(p)$. 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 $K := \sum_{i=1}^n X_i \sim \text{Bern}(p, n)$. 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 .

$$t_n := \log \frac{\mathcal{L}(H_1)}{\mathcal{L}(H_0)} = \log \frac{\max_p P(K = k|p)}{P(K = k|p = p_0)} \quad (7)$$

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 $p = \hat{p} := k/n$. Therefore, the test statistic is given by.

$$t_n = \log \frac{\binom{n}{k} \hat{p}^k (1-\hat{p})^{n-k}}{\binom{n}{k} p_0^k (1-p_0)^{n-k}} = \log \frac{\hat{p}^k (1-\hat{p})^{n-k}}{p_0^k (1-p_0)^{n-k}} \quad (8)$$

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:

$$\begin{aligned} t_n &= k \log \frac{\hat{p}}{p_0} + (n-k) \log \frac{1-\hat{p}}{1-p_0} \\ &= n \left(\hat{p} \log \frac{\hat{p}}{p_0} + (1-\hat{p}) \log \frac{1-\hat{p}}{1-p_0} \right) \\ &= n \mathcal{D}(\text{Bern}(\hat{p}) || \text{Bern}(p_0)) \end{aligned} \quad (9)$$

Where $\mathcal{D}(f||g) := \sum_{x \in \mathcal{X}} f(x) \log \frac{f(x)}{g(x)}$ 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:

$$\mathcal{D}(f||g) \geq 0 \quad \text{with equality iff} \quad f(x) = g(x) \quad \forall x \in \mathcal{X} \quad (10)$$

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 $\delta(x) := f(x) - g(x)$. 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 $\chi^2(f||g) := \sum_{x \in \mathcal{X}} (f - g)^2 / g$. We now investigate the chi-squared divergence for $f = \text{Bern}(p)$ and $g = \text{Bern}(q)$.

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

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

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

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 $\hat{P} := K/n$. For large n , we can find the distribution of \hat{P} by the Central Limit Theorem.

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

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

$$T_n = \frac{1}{2}Z^2 \sim \frac{1}{2}\chi_1^2 \quad (11)$$

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.

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 ($Y_i \sim \text{Bern}(q)$). Note that the population sizes n and m may be different. We define $L := \sum_{i=1}^m Y_i$ 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 $N := n + m$.

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

¹See the Appendix for proof of negligibility

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

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

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:

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

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 .

$$T_N \approx \frac{1}{2} \left(\sqrt{\frac{nm}{\mu(1-\mu)(n+m)}}(\hat{P}-\hat{Q}) \right)^2 = \frac{1}{2}Z^2 \sim \frac{1}{2}\chi_1^2 \quad (12)$$

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.

$$z_N = \frac{\hat{p}-\hat{q}}{\sqrt{\hat{r}(1-\hat{r})(1/n+1/m)}} \sim \mathcal{N}(0,1) \quad (13)$$

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') :

$$y_{v' \rightarrow v}^{(0)} \leftarrow \mathcal{N}(0, 1)$$

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

$$\text{compute average } s^{(t-1)} \leftarrow \frac{1}{2|E|} \sum_{(v, v') \in E} y_{v' \rightarrow v}^{(t-1)}$$

$$\text{recentre messages } z_{v' \rightarrow v}^{(t-1)} \leftarrow y_{v' \rightarrow v}^{(t-1)} - s^{(t-1)}$$

$$\text{sum incoming } y_{v' \rightarrow v}^{(t)} \leftarrow \sum_{(v', v'') \in E \setminus \{v\}} z_{v'' \rightarrow v'}^{(t-1)}$$

if $(v''' \rightarrow v \rightarrow v')$ on cycle of length $r' \leq r$ then correct:

$$y_{v' \rightarrow v}^{(t)} \leftarrow y_{v' \rightarrow v}^{(t)} - \sum_{(v, v'') \in E \setminus \{v', v'''\}} z_{v'' \rightarrow v'}^{(t-r')}$$

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

$$\text{Sum incoming } y_v^{(T)} = \sum_{(v, v') \in E} y_{v' \rightarrow v}^{(T)}$$

Assign labels $\sigma_v = 1$ if $y_v^{(T)} > 0$ 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, $Q := \sqrt{n}(\hat{P} - p_0) \sim \mathcal{N}(0, \beta)$. Manipulating our expression for the error term ϵ we can show that it can be expressed as a sum of

$$\begin{aligned} \epsilon &= nO(\delta^3) \\ \therefore |\epsilon| &\leq n \sum_{i=0}^{\infty} \alpha_i |\hat{P} - p_0|^{3+i} \quad \text{for some finite constants } \alpha_i \geq 0 \\ |\epsilon| &\leq \sum_{i=0}^{\infty} \alpha_i n^{-\frac{1+i}{2}} (\sqrt{n} |\hat{P} - p_0|)^{3+i} \\ |\epsilon| &\leq \sum_{i=0}^{\infty} \alpha_i n^{-\frac{1+i}{2}} |Q|^{3+i} \end{aligned} \tag{14}$$

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:

$$\lim_{n \rightarrow \infty} P(|\epsilon| < \eta) = 1 \text{ for arbitrarily small } \eta \geq 0 \tag{15}$$

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/cfee398643cbc3dc5eefc89334cadd-Paper.pdf>.