TMR: Detecting Structure in Graphical Data

Lawrence Tray
Sup: Ioannis Kontoyiannis

January 18, 2021

Abstract

We produce vast quantities of graphical data every day and yet the techniques used to analyse their structure are still in their infancy. We start by defining the Stochastic Block Model (SBM) widely used in academia and develop a theorem to verify structure in a labelled graph through a hypothesis test formulation. We explore the results of applying this theorem to Facebook friendship egonets and find that gender does indeed influence how friendships form on Facebook but this binary yes-no result is unsatisfactory.

To rank order input features by importance we instead approach the problem from the opposite direction: given a graph we detect the partitions that are in some sense typical and then develop a crude classifier to map from vertex features to partition. The classifier can then rank order the input features by importance. Nevertheless, work must still be done to refine this classifier and tie it back to the rigorous hypothesis testing framework we have already developed.

1 Introduction

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. There are many algorithms developed to analyse graphical data. Nevertheless, that same principled hypothesis testing framework we have for querying classical data is less well-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 with entries in $[0, 1]$ 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 contains all vertices belonging to community i.*

Though the definition of the SBM is simple, it allows for very deep and rich analysis of graphical datasets. For certain problems it helps to define the symmetric SBM.

Definition 2.2 *The symmetric SBM is a special case denoted by $SSBM(n, k, q_{in}, q_{out}) \equiv SBM(n, p, W)$ if the community prior p is uniform ($p_i = 1/k$ for $i \in \{1, 2, \dots k\}$) and W_{ij} takes only two values, one on diagonal and another off diagonal such that $W_{ij} = q_{in}$ for $i = j$ and $W_{ij} = q_{out}$ for $i \neq j$.*

3 Verifying Structure

3.1 Theory

Armed with this definition we tackle the simplest problem in structure verification. Given an undirected graph $\mathcal{G}$ and vertex-labels X, we wish to determine whether the two communities a and b connect differently. Put formally, this is 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 for an

undirected graph $W = W^T$ necessarily so $W_{ab} = W_{ba}$). 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. 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 the number of vertices in community i by $n_i \coloneqq |\Omega_i(X)|$ leading to the result $n = \sum_i n_i$. Furthermore, we use $E_{ij} = E_{ij}(X, \mathcal{G})$ to denote the number of realised edges between communities i and j (in generality i may be equal to j) and similarly 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, the 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 of the test-statistic t_n. 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:

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

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 (KL divergence) *With f defined as in equation 6, $0 \leq e \leq m$ and $r \in [0,1]$ 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 $g, h : \mathcal{X} \mapsto [0,1]$. $\mathcal{D}$ is 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:

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

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

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

However, we must simplify further still.

Lemma 3.2 (χ^2 approximation) *For $g(x) \approx h(x)$ then $\mathcal{D}(g||h) \approx \frac{1}{2}\chi^2(g||h)$ where χ^2 is the chi-squared distance between two distributions defined as $\chi^2(g||h) \coloneqq \sum_{x \in \mathcal{X}} \frac{(g(x) - h(x))^2}{h(x)}$*

The proof requires defining $\delta(x) \coloneqq g(x) - h(x)$ and taking a Taylor expansion of $\log 1 + x$:

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

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 g very close to h. The chi-squared distance between two Bernoullis for $g = Bern(p)$ and $h = Bern(q)$ is simply:

$$\chi^2(\text{Bern}(p)||\text{Bern}(q)) = \left(\frac{p - q}{\sqrt{q(1 - q)}} \right)^2 \tag{11}$$

We can apply these results to equation 10 to obtain an approximate expression for t_n under the null for large n such that $\hat{p}$, $\hat{q}$ and $\hat{r}$ are all close meaning Lemma 3.2 is valid, giving:

$$\begin{aligned}
t_n &\approx \frac{1}{2\hat{r}(1 - \hat{r})} \left(M_\alpha \left(\hat{p} - \hat{r}\right)^2 + M_\beta \left(\hat{q} - \hat{r}\right)^2 \right) \\
&= \frac{1}{2\hat{r}(1 - \hat{r})} \left(M_\alpha \left(\frac{M_\alpha(\hat{p} - \hat{q})}{M_\alpha + M_\beta} \right)^2 + M_\beta \left(\frac{M_\beta(\hat{q} - \hat{p})}{M_\alpha + M_\beta} \right)^2 \right) \\
&= \frac{1}{2} \left(\frac{\hat{p} - \hat{q}}{\sqrt{\hat{r}(1 - \hat{r})(1/M_\alpha + 1/M_\beta)}} \right)^2 = \frac{1}{2} z_n^2
\end{aligned} \tag{12}$$

Through application of the Central Limit Theorem (CLT) and Slutsky's Theorem it can be shown that as $n \to \infty$ under the null H_0 the term in brackets on the final line is distributed like a standard Gaussian $z_n \sim \mathcal{N}(0,1)$. This is because under the same assumptions, assuming that $W_\alpha = W_\beta = \mu$, by the CLT $\hat{p} \sim (\mu, \mu(1 - \mu)/M_\alpha)$, $\hat{q} \sim (\mu, \mu(1 - \mu)/M_\beta)$ and $\hat{p} \sim (\mu, \mu(1 - \mu)/(M_\alpha + M_\beta))$. Therefore, the whole test-statistic is distributed as $t_n \sim \frac{1}{2}\chi_1^2$ by the definition of the χ_p^2 distribution: $\chi_p^2 \sim \sum_{i=1} pZ_i^2$ for $Z_i \sim \mathcal{N}(0,1)$. This leads to the following theorem:

Theorem 3.3 (χ^2 test) *For $(X, \mathcal{G}) \sim SBM(n, p, W)$, 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 (z-test) *We can also use a slightly simpler test-statistic z_n albeit with some loss of generality. If we define z_n to be:*

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

With the symbols retaining their previous meanings. 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 $Y \sim \mathcal{N}(0, 1)$

3.2 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) [8] 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 the example below:

Example anonymised feature flags

```
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
```

Each feature is anonymised to avoid disclosing 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 which suffices for our analysis. 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.

We perform a hypothesis test in the manner described by theorem 3.3 to determine whether gender influences how friends connect on Facebook. We choose to analyse SNAP egonet with id 0 (the id of the egonet is the id of the single egonode) though any choice is possible. The egonet is plotted on figure 1. We use the Python package NetworkX [7] for its visualisation tools. Indeed, we will discuss layout algorithms later as for now we focus on the simple hypothesis test.

We use an SBM model with $k = 2$ communities (1: gender-77 on and 2: gender-77 off) to model the egonet and perform a three-way hypothesis test on the parameters of the connectivity Matrix W.

$$W = \begin{bmatrix} p_1 & q \\ q & p_2 \end{bmatrix} \tag{13}$$

Therefore, p_1 is the probability that two vertices of gender-77 are connected, q is the crossover probability and p_2 is the connection probability within the gender-78 community. The results of the hypothesis tests are given in table 2 alongside p-values (we choose a 95% significance level to reject the null). However, in some cases the test statistics were so extreme that the p-value saturated to 0.

Test 2 being rejected gives evidence that one gender has on average more friends from their own gender. We see that gender-78 people (community 2) have on average more friends ($\hat{p_2} > \hat{p_1}$ in table 1). The rejection of the null in test 3 gives evidence that gender-78 treat gender-77 people differently to their own gender. Indeed, a gender-78 person is more likely to be friends

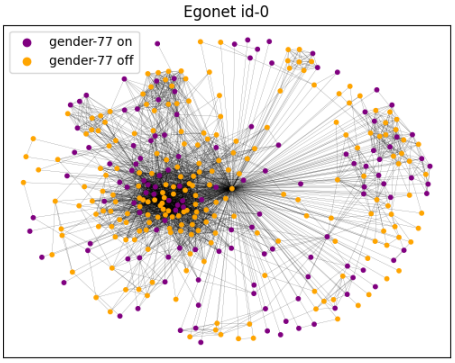


Figure 1: Egonet 0 with nodes coloured by gender

n_1	n_2	$\hat{p_1}$	$\hat{q}$	$\hat{p_2}$
130	218	8.13×10^{-2}	7.66×10^{-2}	10.6×10^{-2}

Table 1: Egonet-0 properties and parameter estimates

Test	H_0	H_1	p-value	H_0 rejected
1	$p_1 = q$	$p_1 \neq q$	0.158	No
2	$p_1 = p_2$	$p_1 \neq p_2$	0.000	Yes
3	$q = p_2$	$q \neq p_2$	0.000	Yes

Table 2: Egonet-0 hypothesis tests that gender influences friendship formation

with a fellow 78-er than with a 77-er. However, we do not reject the null in test 1 so there is insufficient evidence to claim the reverse (that 77-ers tend to stay away from 78-ers). Nevertheless, these are results on a single egonet so can hardly be generalised to society as a whole. However, the exercise has highlighted some interesting points:

1. The hypothesis test framework does not quantify the magnitude of the difference just whether or not a difference exists.
2. As the degree n increases, the p-values will necessarily become more extreme.
3. The method relies on the analyst to specify the features of interest so results may be misleading if third variables are missing from the analysis.

To address these shortcomings we instead approach the problem from a different angle. For now we have been given a graph with fully labelled vertices and ask if a given feature affect the structure. Instead, we could take a graph and partition into communities first (without using feature information) and then ask which features most reliably explain the partition. For that we need a way of detecting structure.

4 Detecting Structure

4.1 Weak recovery

This is the problem of recovering the communities from the graph $\mathcal{G}$. We use the notation $\hat{X}$ to denote the produced estimate of the community labels for the n vertices. Obviously, if the SBM is symmetric then we must allow for an arbitrary relabelling of the communities $r : \{1, 2 \dots k\} \mapsto \{1, 2 \dots k\}$ before we compare the agreement of the two vectors $\hat{X}$ and X. The agreement A and normalised agreement $\tilde{A}$ between two vectors x, y are computed as below:

$$A(x, y) = \max_r \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(x_i, r(y_i)) \tag{14}$$

$$\tilde{A}(x, y) = \max_r \frac{1}{k} \sum_{i=1}^{k} \frac{1}{n_i} \sum_{v \in \Omega_i(x)} \mathbb{1}(x_v, r(y_v)) \tag{15}$$

The normalised agreement is important for asymmetric SBMs as it takes the agreement averaged over communities rather than vertices. We define 4 recovery regimes (plus the trivial no recovery), such that the following holds (adapted from Abbe

[1]):

Definition 4.1 (Recovery regimes) *for $(X, \mathcal{G}) \sim SBM(m, p, W)$ the following recovery requirements are solved if there exists an algorithm which takes $\mathcal{G}$ and input and outputs a vector of classifications $\hat{X} = \hat{X}(\mathcal{G})$ such that as $n \to \infty$:*

- ***Exact recovery:*** *$Pr\{A(X, \hat{X}) = 1\} = 1 - o(1)$ correct classification recovered almost surely*
- ***Almost exact recovery:*** *$Pr\{A(X, \hat{X}) = 1 - o(1)\} = 1 - o(1)$ vanishing fraction misclassified almost surely*
- ***Partial recovery:*** *$Pr\{\tilde{A}(X, \hat{X}) \geq \alpha\} = 1 - o(1), 1 > \alpha > 1/k$ better than choosing from uniform prior*
- ***Weak recovery:*** *$Pr\{\tilde{A}(X, \hat{X}) \geq 1/k + \epsilon\} = 1 - o(1), \epsilon > 0$ marginally better than choosing from uniform prior*
- ***No recovery:*** *graph provides no information as to class labels*

Where $o(1)$ denotes a family of functions that tend to 0 as $n \to \infty$. The regime $\mathcal{G}$ occupies depends on the parameters of the SBM. Each recovery regime is weaker than the one above it so Exact recovery implies weak recovery is possible..

We choose to focus on weak recovery as it is the least strict of all regimes. An alternative statement of weak recovery is given below:

Definition 4.2 (Weak recovery) *for $(X, \mathcal{G}) \sim SBM(n, p, W)$ weak recovery is solved if there exists $\epsilon > 0$ and a certain choice of indices $i, j \in \{1, 2 \dots k\}$ and an algorithm that takes as input $\mathcal{G}$ and outputs a partition of the vertex set $\mathcal{V}$ into two distinct sets S and S^C (detected communities 1 and 2 respectively) such that:*

$$Pr\left\{\frac{|\Omega_i \cap S|}{|\Omega_i|} - \frac{|\Omega_j \cap S|}{|\Omega_j|} \geq \epsilon\right\} = 1 - o(1)$$

Bearing in mind the definition of $\Omega_i = \Omega_i(X) \coloneqq \{v \in \mathcal{V} : X_v = i\}$. In other words, we require the partition $S \subset \mathcal{V}$ to cut across communities such that the fraction of a particular community in partition S with respect to the whole community is different for a choice of two communities.

One of the most promising weak recovery algorithms is called Acyclic Belief Propagataion (ABP). This is a message-passing algorithm that solves weak recovery in the sense of definition 4.2. Algorithm 1 is heavily adapated from Abbe and Sandon [2]. The original paper goes into much more detail but it works on the premise that one can infer the community of a particular vertex by taking a majority vote among the of all its neighbours. The algorithm is non-determinsitic and requires the vertex set $\mathcal{V} = \mathcal{V}(\mathcal{G})$ and edge set $\mathcal{E} = \mathcal{E}(\mathcal{G})$ as well as two hyper-parameters r (the maximum cyclic length to correct for) and T (the number of iterations):

Algorithm 1 Linearised Acyclic Belief Propagation (ABP)

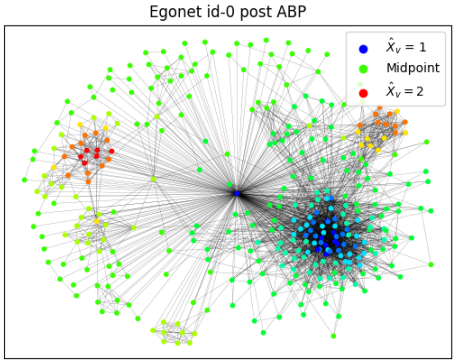
function ABP($\mathcal{E}$, $\mathcal{V}$, r, T)
  $y^{(0)}_{v' \to v} \leftarrow \mathcal{N}(0, 1)$ ▷ Initialise messages randomly
  for $t \in \{1, 2 \cdots T\}$ **do**
    for $(v, v') \in \mathcal{E}$ **do**
      $s^{(t-1)} \leftarrow \frac{1}{2|E|} \sum_{(v,v') \in \mathcal{E}} y^{(t-1)}_{v' \to v}$ ▷ Compute average
      $z^{(t-1)}_{v' \to v} \leftarrow y^{(t-1)}_{v' \to v} - s^{(t-1)}$ ▷ Recentre messages
      $y^{(t)}_{v' \to v} \leftarrow \sum_{(v', v'') \in \mathcal{E} \setminus \{v\}} z^{(t-1)}_{v'' \to v'}$ ▷ Sum incoming messages

      if $\exists v''' \in \mathcal{V}$ such that $(v''' \to v \to v')$ is on a cycle of length $r' \leq r$ **then**
        $y^{(t)}_{v' \to v} \leftarrow y^{(t)}_{v' \to v} - \sum_{(v, v'') \in \mathcal{E} \setminus \{v', v'''\}} z^{(t-r')}_{v'' \to v'}$ ▷ Cyclic correction
      end if
    end for
  end for
  $\sigma_v \leftarrow \sum_{(v,v') \in \mathcal{E}} y^{(T)}_{v' \to v}$ ▷ Sum incoming messages
  $S \leftarrow \{v \in \mathcal{V} : \sigma_v > 0\}$ ▷ Hard assignment
  $S^C \leftarrow \{v \in \mathcal{V} : \sigma_v \leq 0\}$
  return S, S^C
end function

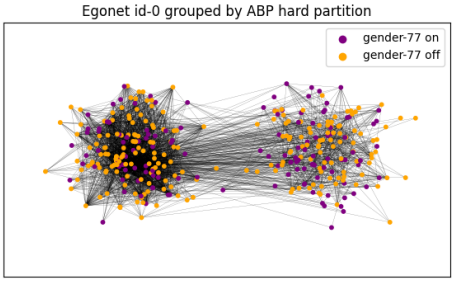
We implement this algorithm in Python and apply it to the Facebook Egonets described earlier. We found that the hyper-parameter settings $r = 3$ and $T = 5$ yielded good results. We define the normalised ABP output below:

$$\tilde{\sigma}_v \coloneqq \begin{cases} \frac{\sigma_v}{|\sigma_{max}|} \text{ for } \sigma_j > 0 \\ \frac{\sigma_v}{|\sigma_{min}|} \text{ for } \sigma_j \leq 0 \end{cases} \tag{16}$$

Where σ_{max} is the largest positive value of σ_v and σ_{min} is the most negative. The normalised quantity $\tilde{\sigma}_v \in [-1, +1]$ with 0 still being the midpoint that determines which detected community vertex v is assigned to. We colour each vertex according to this quantity $\tilde{\sigma}_v$ and plot the graph on figure 2a. We see intuitively that ABP has worked. The blue cluster in the lower right is clearly part of one community and a great many vertices do not seem to belong to any cluster at all so their classification is uncertain (the colour is close to green).



(a) Normalised ABP output $\tilde{\sigma}_v$ colouring vertices



(b) Vertices grouped by hard ABP assignments and coloured by gender. In assigned community 1 (S - left) we have {58, 122} gender-77 and gender-78 respectively and in detected community 2 (S^C - right) we have {72, 96}

Figure 2: Egonet-0 weak recovery results

We can evaluate formally whether weak recovery has succeeded by inspecting figure 2b and treating gender-77 and gender-78 as the two communities of interest:

$$\frac{|\Omega_{77} \cap S|}{|\Omega_{77}|} = \frac{58}{58+72} = 0.446 \quad \text{and} \quad \frac{|\Omega_{78} \cap S|}{|\Omega_{78}|} = \frac{122}{122+96} = 0.560 \tag{17}$$

Clearly the fraction of each community in the partition S differ substantially so we say that weak recovery is solved. The set S contains a higher proportion of the overall 78-population than of the 77-population.

4.2 From detection to inference

We now have an algorithm (ABP) that partitions our graph into two sets that are most readily explained by an SBM with $k = 2$ communities. We can now ask the question of which vertex labels most readily explain the separation. Therefore, we can determine which features have the largest impact on the graphical structure; this circumvents the binary resolution problem with the hypothesis testing approach.

This part of the project is still in its early stages so these are just preliminary results. For reference we plot the community fractions in the detected set S [1] on figure 3a. It is interesting to note that if two bars on this plot differ substantially then we have satisfied the requirment for weak recovery as defined in 4.2. We necessarily remove the constraint on the community sets Ω_i to be disjoint as one vertex can have multiple features turned on. This can be modelled more precisely through a Mixed Membership Model but for now an SBM with $k = 2$ communities suffices.

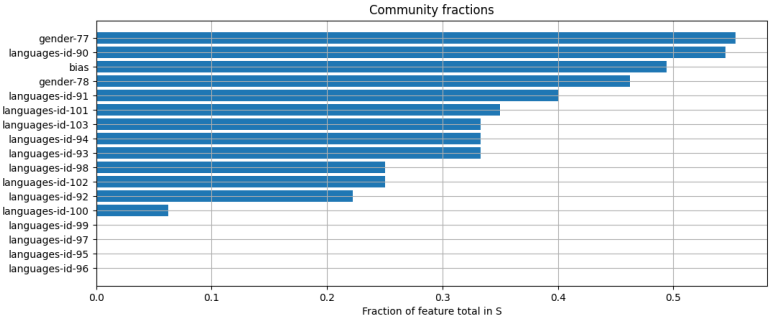
We perform a linear regression on the normalised ABP output vector $\tilde{\sigma}$ with the feature matrix $n \times f$ feature matrix X as the explanatory variables. $X_{ij} = 1$ if vertex i has feature j turned on and 0 if feature j is off. The equation we try to fit is given below:

$$\tilde{\sigma} = a + Xb + \epsilon \tag{18}$$

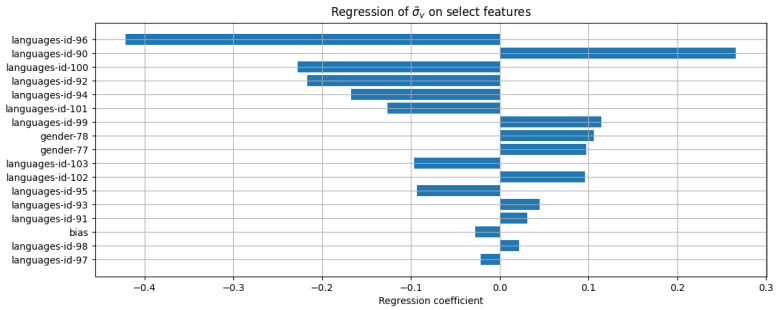
Linear regression finds the constant bias a and $f \times 1$ weight vector b that minimises the mean squared prediction error $\epsilon^T \epsilon / n$. The theory for linear regression is well known so omitted here (see [12] for a reference). We perform a linear regression using only gender and language features available in the dataset and plot the parameter coefficient in order of decreasing magnitude (figure 3b).

The results are rather interesting. Certain languages are a far stronger predictor of community assignment than language. After all, one would hope that a language barrier poses a more tangible obstacle to friendship than gender. The bias term has rather low magnitude which just implies that the partition S and S^C are of similar size. A keen reader would notice that

[1] ABP is non-deterministic so each subplot is the result of a different run of ABP. Subfigures use the same partition S but no exact comparison is possible between figures.



(a) Community fractions $|\Omega_i \cap S|/|\Omega_i|$



(b) Linear regression of $\tilde{\sigma}_v$ on select feature ($R^2 = 0.113$)

Figure 3: Egonet-0 linear classifier results

the bias term is included in figure 3a; this is because we implement the bias by augmenting X with an additional feature flag always to 1 such that the constant term a in equation 18 is subsumed into the b vector. Indeed, in figure 3a the fraction of vertices with bias turned on in partition S is very close to 0.5. Since all vertices have the bias feature turned on $\Omega_{bias} = \mathcal{V}$, this is equivalent to the number of vertices in S as a fraction of the total: $|S|/|\mathcal{V}|$.

This analysis is of course rather crude as we have failed to take into account multi-collinearity or even determined if a linear regression model is appropriate; indeed, an R^2 value of just 0.113 would suggest that it is not. Nevertheless, this serves as the platform for subsequent more rigorous analysis.

5 Future Direction

The immediate next steps for the project are to refine the classification approach outlined in section 4.2. Firstly, it should be tied back to the hypothesis testing framework to make sure that the results are proven to be meaningful. Secondly, the exact form of the classifier can be explored; a linear regression model is a good starting point but we can expand to a softmax neural network as that would allow us to generalise to more than just two partitions.

Indeed, ABP is severely limited by the fact that it only outputs 2 partitions S and S^C. Therefore, though weak recovery is solved the results from applying the classification approach may not be natural or easy to interpret. We instead would prefer to investigate various different algorithms. Force-directed approaches are the standard in industry. One of the most common is the Fruchterman-Reingold algorithm which treats vertices as repelling objects and edges as attractive forces (springs) and then performs iterative updates on vertex positions for a finite number of iterations. This algorithm is what is used by the NetworkX package [7] to determine graph layouts as seen in figures 1 and 2a. Song and Bressan [10] have had great success using the Fruchterman-Reingold algorithm followed by k-means clustering for community detection. Nevertheless, though they show the method produces accurate results there is no rigorous proof that it solves weak recovery.

Lastly, the analysis must of course be expanded to more datasets to benchmark efficacy against other methods. Work has already begun analysing the influence of gender academic collaboration graphs from Aminer [11]. Nevertheless, this dataset does not contain gender explicitly, only first-name. However, this is cross-referenced with a list of popular baby-names to get a good replacement (as suggested by West et al. [6]). Furthermore, the seminal dataset of political blogs from the 2004 US election [5] which started much of the analysis of the SBM will of course be analysed.

6 Conclusion

This concludes the Technical Milestone Report for the project thus far. We start by fully fleshing out the theoretical foundation of the Stochastic Block Model and hypothesis tests on connectivity parameters. Nevertheless, this approach is too narrow in scope to be a useful tool on its own. We therefore set about the problem from the opposite direction; rather than verify structure we are already given, we detect structure in the graph without labels and seek to identify which labels best recreate the detected partition. Early results are promising but this approach needs to be given the same rigorous treatment as the hypothesis testing framework. That will be the emphasis of future work.

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/cfee398643cbc3dc5eefc89334cacPaper.pdf.

[5] Lada A. Adamic and Natalie Glance. "The political blogospher and the 2004 US Election". In: *Proceedings of the WWW-2005 Workshop on the Weblogging Ecosystem*. 2005.

[6] Jevin West et al. "The role of gender in scholarly authorship". In: *PLoS One* (2013).

[7] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. "Exploring Network Structure, Dynamics, and Function using NetworkX". In: *Proceedings of the 7th Python in Science Conference*. Ed. by Gaël Varoquaux, Travis Vaught, and Jarrod Millman. Pasadena, CA USA, 2008, pp. 11 –15.

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

[9] 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.

[10] Yi Song and Stephane Bressan. "Force-directed Layout Community Detection". In: *School of Computing* (2013).

[11] Jie Tang et al. "ArnetMiner: Extraction and Mining of Academic Social Networks". In: *KDD'08*. 2008, pp. 990–998.

[12] Yale. *Linear Regression*. URL: http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm.