Spectral clustering in the weighted stochastic block model

Ian Gallagher*, Anna Bertiger**, Carey Priebe***, and Patrick Rubin-Delanchy*

*University of Bristol, U.K.
**Microsoft, U.S.A.
***Johns Hopkins University, U.S.A.

October 15, 2019

Abstract
This paper is concerned with the statistical analysis of a real-valued symmetric data matrix. We assume a weighted stochastic block model: the matrix indices, taken to represent nodes, can be partitioned into communities so that all entries corresponding to a given community pair are replicates of the same random variable. Extending results previously known only for unweighted graphs, we provide a limit theorem showing that the point cloud obtained from spectrally embedding the data matrix follows a Gaussian mixture model where each community is represented with an elliptical component. We can therefore formally evaluate how well the communities separate under different data transformations, for example, whether it is productive to ‘take logs’. We find that performance is invariant to affine transformation of the entries, but this expected and desirable feature hinges on adaptively selecting the eigenvectors according to eigenvalue magnitude and using Gaussian clustering. We present a network anomaly detection problem with cyber-security data where the matrix of log p-values, as opposed to p-values, has both theoretical and empirical advantages.

1 Introduction
Spectral clustering (Von Luxburg, 2007) refers to a number of different algorithms that have in common two main steps: first, of computing the spectral decomposition of a (possibly regularised) data matrix; and second, of applying a clustering algorithm to a point cloud extracted from the eigenvectors. When the matrix holds distances or affinities spectral clustering allows estimation of non-circular clusters in pointillist data (Ng et al., 2002). When the matrix represents a graph, it enables the discovery of communities (Von Luxburg, 2007).

In the case of graphs, one can talk quite precisely about the relative merits of different regularisation techniques (e.g. adjacency versus normalised Laplacian), which eigenvectors to select (e.g. corresponding to large eigenvalues versus large magnitude eigenvalues) and which clustering algorithm to use (e.g. K-means versus Gaussian mixture modelling). While the first decision is complicated (Tang and Priebe, 2019), asymptotic analysis now clearly favours the second option in each of the remaining (Rohe et al., 2011; Athreya et al., 2017; Rubin-Delanchy et al., 2018). These determinations are made under the assumption that the data follow a stochastic block model (Holland et al., 1983), where the probability of an edge is dependent only on the (unknown) community memberships of the corresponding nodes.

The natural extension to a real-valued matrix is to assume the $ij$th entry is a real random variable whose distribution depends only the communities of nodes $i$ and $j$ (Xu et al., 2017). (Under the ordinary stochastic block model this distribution would be Bernoulli.) While defining a normalised Laplacian is not entirely straightforward, since for example a node’s ‘degree’ could be negative and would need to be square-rooted, the second and third questions are still pertinent: which eigenvectors and which clustering algorithm should be used?

This paper presents a central limit theorem showing that asymptotically the point cloud obtained from spectrally embedding a real-valued matrix from a weighted stochastic block model follows...
a Gaussian mixture model with elliptical components whose centres and covariance matrices are explicitly calculable. This result implies that for statistical consistency, eigenvectors selected by eigenvalue magnitude must be used, and for optimality one should use Gaussian clustering, and not \( K \)-means.

Another application of this result is to allow a choice between data representations, for example, whether to embed the matrix of counts or log-counts. Since two data representations produce different mixture distributions, one can compare how well the components separate in each case. Following [Tang and Priebe (2019)], we do this using Chernoff information. In a relevant formalisation of the network anomaly detection problem, we are thus able to show that embedding the matrix of log \( p \)-values, rather than raw \( p \)-values, is statistically more efficient. This theoretical observation is validated in a cyber-security example.

Finally, affine transformation of a real-valued matrix’s entries does not change the Chernoff information of the associated asymptotic clustering problem. In other words, one need not worry about the origin and scale of the measurements in the data matrix, for example, whether temperature is measured in Celsius or Fahrenheit. Yet affine transformation can cause important eigenvalues to flip sign and Gaussian clusters to change shape, and so this invariance hinges on choosing eigenvectors from both sides of the spectrum and using Gaussian clustering; otherwise, performance will vary substantially.

2 Spectral clustering in the weighted stochastic block model

2.1 The weighted stochastic block model

**Definition 1** (Weighted stochastic block model). Given \( n \) nodes and \( K \) communities, an undirected weighted graph with symmetric adjacency matrix \( \mathbf{W} \in \mathbb{R}^{n \times n} \) follows a \( K \)-community stochastic block model if there is a partition of the nodes into \( K \) communities conditional upon which, for all \( i < j \),

\[
\mathbf{W}_{ij} \overset{\text{ind}}{\sim} \mathcal{F}_{z_i, z_j},
\]

where \( z_i \in \{1, \ldots, K\} \) is an index denoting the community of node \( i \), assigned independently according to a probability vector \((\pi_1, \ldots, \pi_K)\) where \( \sum_{k=1}^{K} \pi_k = 1 \).

Define matrices \( \mathbf{B}, \mathbf{C} \in \mathbb{R}^{K \times K} \) as the block means and variances respectively of the distributions \( \mathcal{F}_{k,l} \), for \( k, l \in \{1, \ldots, K\} \), where it is assumed the moments exist. For example, a 2-community unweighted stochastic block model with intra-community (respectively, inter-community) link probability \( p_1 \) (respectively, \( p_2 \)) has

\[
\mathbf{B} = \begin{pmatrix} p_1 & p_2 \\ p_2 & p_1 \end{pmatrix}, \quad \mathbf{C} = \begin{pmatrix} p_1(1-p_1) & p_2(1-p_2) \\ p_2(1-p_2) & p_1(1-p_1) \end{pmatrix}.
\]

The signature of a weighted stochastic block model, \((p, q)\), is defined as the number of strictly positive and strictly negative eigenvalues of \( \mathbf{B} \) respectively and let \( d = p + q \). We can choose \( v_1, \ldots, v_K \in \mathbb{R}^d \) such that \( v_k^T \mathbf{I}_{p,q} v_l = \mathbf{B}_{k,l} \), for \( k, l \in \{1, \ldots, K\} \), where \( \mathbf{I}_{p,q} = \text{diag}(1, \ldots, 1, -1, \ldots, -1) \), with \( p \) ones followed by \( q \) minus ones on its diagonal. One choice is to use the \( K \) rows of \( \mathbf{U} \Sigma \mathbf{U}^T \) over the diagonal matrix \( D \).

The vector \( v_q \) can be interpreted as a canonical latent position for node \( i \) in the weighted stochastic block model. Latent positions of a stochastic block model are only identifiable up to transformation by elements of the indefinite orthogonal group \( \mathcal{O}(p, q) = \{ \mathbf{M} \in \mathbb{R}^{d \times d} : \mathbf{M} \mathbf{I}_{p,q} \mathbf{M}^T = \mathbf{I}_{p,q} \} \). Attempts to infer the latent positions from the adjacency matrix of a weighted stochastic block model must take unidentifiability up to transformation from \( \mathcal{O}(p, q) \) into account.

2.2 Spectral clustering

**Definition 2** (Adjacency spectral embedding). Given an undirected weighted graph with symmetric adjacency matrix \( \mathbf{W} \in \mathbb{R}^{n \times n} \), consider the spectral decomposition \( \mathbf{W} = \mathbf{U} \Sigma \mathbf{U}^T + \mathbf{U} \mathbf{S} \mathbf{U}^T \), where \( \mathbf{S} \) is a \( d \times d \) diagonal matrix containing the \( d \) largest eigenvalues of \( \mathbf{W} \) in magnitude, and \( \mathbf{U} \in \mathbb{R}^{n \times d} \).
contains the corresponding orthonormal eigenvectors. Define the adjacency spectral embedding of the graph into $\mathbb{R}^d$ by

$$\hat{X} = [\hat{X}_1 | \ldots | \hat{X}_n]^T = \hat{U}|\hat{S}|^{1/2}. $$

We will interpret this spectral embedding procedure as providing an estimate $\hat{X}_i \in \mathbb{R}^d$ of the latent position for node $i$ in the network. Heuristically, nodes that are somehow ‘close’ in this space are likely to belong to the same community. Algorithm 1 (extending Algorithm 1 Rubin-Delanchy et al. (2018) to real-valued matrices) proposes an approach to recovering these communities.

**Algorithm 1** Spectral clustering for the weighted stochastic block model

**Input:** Weighted adjacency matrix $W$, dimension $d$, number of communities $K \geq d$

1: Compute spectral embedding $\hat{X}_1, \ldots, \hat{X}_n$ of the graph into $\mathbb{R}^d$ via Definition 2

2: Fit a Gaussian mixture model with full covariance matrices with $K$ components

**Output:** Cluster centres $\hat{v}_1, \ldots, \hat{v}_K$ and node memberships $\hat{z}_1, \ldots, \hat{z}_n$

There are two important features of Algorithm 1. Firstly, both sides of the spectral decomposition are used: in Definition 2 the largest eigenvalues by magnitude are retained (and the corresponding eigenvectors), not just the largest positive eigenvalues. This is needed for statistical consistency in general (Rohe et al., 2011). Large negative eigenvalues in computer network graphs can hold key information for node clustering and link prediction (Rubin-Delanchy et al., 2018). Secondly, the covariance matrices in the Gaussian mixture model are unconstrained, i.e. ellipsoidal with varying volume, shape, and orientation. This is a significant departure from the standard use of $K$-means (Von Luxburg, 2007).

Both of these algorithm features are well-justified by the theorem in the following section.

### 2.3 Central limit theorem

Rubin-Delanchy et al. (2018) derived a central limit theorem for adjacency spectral embedding under a ‘generalised random dot product graph’; a novel contribution of the present paper is to consider an extension of this theorem to the case of a weighted stochastic block model:

**Theorem 1** (Adjacency spectral embedding central limit theorem). Consider a sequence of adjacency matrices $W^{(n)}$ from a weighted stochastic block model with signature $(p,q)$. For any integer $m > 0$ and points $q_1, \ldots, q_m \in \mathbb{R}^d$, conditional on the community labels $z_1, \ldots, z_m$, there exists a sequence of random matrices $Q_n \in \mathcal{O}(p,q)$ such that

$$P \left\{ \prod_{i=1}^{m} n^{1/2} \left( Q_n \hat{X}_i^{(n)} - v_{z_i} \right) \leq q_i \right\} \rightarrow \prod_{i=1}^{m} \Phi(q_i; 0, \Sigma_{z_i}),$$

where

$$\Sigma_k = I_{p,q} \Delta^{-1} \left( \sum_{l=1}^{K} \pi_l C_{kl} v_l v_l^\top \right) \Delta^{-1} I_{p,q} \in \mathbb{R}^{d \times d},$$

the second moment matrix $\Delta$, assumed to be invertible, is

$$\Delta = \sum_{k=1}^{K} \pi_k v_k v_k^\top \in \mathbb{R}^{d \times d},$$

and $\Phi(\cdot; \mu, \Sigma)$ denotes the cumulative distribution function of a multivariate normal distribution with mean $\mu$ and covariance $\Sigma$.

The implication of Theorem 1 is that spectral embedding an adjacency matrix from a weighted stochastic block model produces a point cloud that is asymptotically a linear transformation (given by $Q_n^{-1} \in \mathcal{O}(p,q)$) of independent, identically distributed draws from a Gaussian mixture model. Each of its $K$ components corresponds to a community and has an explicitly calculable mean and covariance. A finite sample illustration of the theorem is given in Figure 1.

The result motivates the design of Algorithm 1 namely the importance of using both sides of the spectral decomposition and allowing full covariance matrices when fitting a Gaussian mixture model.

3
2.4 Example: Poisson counts versus Bernoulli presence events

Consider a 2-community weighted stochastic block model where weights represent event counts modelled by Poisson distributions with rate \( \lambda_{kl} \), \( k, l \in \{1, 2\} \), with block mean and variance matrices

\[
B = \begin{pmatrix} 0.5 & 0.7 \\ 0.7 & 0.6 \end{pmatrix}, \quad C = \begin{pmatrix} 0.5 & 0.7 \\ 0.7 & 0.6 \end{pmatrix}.
\]

We generate a weighted network from this model with \( n = 1000 \) nodes and probability of belonging to the first community \( \pi_1 = 0.2 = 1 - \pi_2 \), and apply Algorithm 1. Figure 1a) shows the 2-dimensional point cloud obtained from spectral embedding the graph (note, \( d = K = 2 \)), with colours indicating the true cluster assignment. The red and blue ellipses show the two 95% contours obtained by applying Gaussian clustering using the Python \textit{sklearn} library. In this example, the predicted community assignment is 98.5% accurate. Black ellipses show the 95% asymptotic contours of the components, calculated using Theorem 1 and approximately comparable.

Instead, suppose we simply report, for each pair of nodes, whether at least one event occurs. If \( X \sim \text{Poisson}(\lambda) \), then \( Y = \mathbb{1}(X \geq 1) \sim \text{Bernoulli}(1 - \exp(-\lambda)) \). The block mean and variance matrices for this unweighted stochastic block model are

\[
B' = \begin{pmatrix} 0.393 & 0.503 \\ 0.503 & 0.451 \end{pmatrix}, \quad C' = \begin{pmatrix} 0.239 & 0.250 \\ 0.250 & 0.248 \end{pmatrix}.
\]

We calculate this modified adjacency matrix directly from the original and Figure 1b) shows the resulting point cloud from spectral embedding, where the contours and true community labels are indicated as before. This time the predicted community assignment based on a Gaussian mixture model is only 96.3% accurate. This loss of accuracy is consistent with the theoretical contours appearing less well separated. We formally quantify cluster separation in Section 3.1 and find that the Poisson representation is indeed preferable in this example.

3 Choosing matrix data representation

3.1 Chernoff information

In order to define a measure of cluster separation we take inspiration from Tang and Priebe (2019), where the Chernoff information was proposed as a method to compare graph embedding based on the Laplacian versus the adjacency matrix. In a 2-cluster problem, the Chernoff information provides an upper bound on the probability of error of the Bayes decision rule that assigns each
data point to its most likely cluster a posteriori. If the clusters have distributions $F_1$ and $F_2$, the Chernoff information is (Chernoff [1952]):

$$C(F_1, F_2) = \sup_{t \in (0, 1)} C_t(F_1, F_2),$$

where $C_t$ is the Chernoff divergence

$$C_t(F_1, F_2) = -\log \int_{\mathbb{R}^d} f_1^t(x)f_2^{1-t}(x) \, dx,$$

and $f_1, f_2$ are the probability density functions corresponding to $F_1, F_2$ respectively. For $K > 2$, one reports instead the Chernoff information of the critical pair, $\min_{k,l \in \{1, \ldots, K\}, k \neq l} C(F_k, F_l)$.

The Chernoff information of the components in the limiting mixture distribution of Theorem 1 can be written in closed form. Suppose $F_1 = \text{Normal}(v_1, \Sigma_1)$ distribution and $F_2 = \text{Normal}(v_2, \Sigma_2)$ then, for $t \in (0, 1)$, denoting $\Sigma_t = (1-t)\Sigma_1 + t\Sigma_2$, we can compute (Pardo [2005]),

$$C(F_1, F_2) = \sup_{t \in (0, 1)} \left\{ \frac{t(1-t)}{2} (v_1 - v_2)^\top \Sigma_t^{-1} (v_1 - v_2) + \frac{1}{2} \log \left| \frac{\Sigma_1}{\left| \Sigma_1 \right|^{1/2} \left| \Sigma_2 \right|^{1/2}} \right| \right\}. \tag{2}$$

In their work motivating the use of Chernoff information to compare graph embeddings, Tang and Priebe (2019) make the point that a simpler criterion such as cluster variance is not satisfactory, since it is effectively measuring the performance of $K$-means clustering rather than clustering using a Gaussian mixture model.

### 3.1.1 Example: Poisson counts versus Bernoulli presence events

Returning to the Poisson versus Bernoulli example of Section 2.4, Figure 1c) shows the Chernoff divergence, and hence the Chernoff information, for the two representations. For the Bernoulli data, the Chernoff information is 0.002, achieved at $t = 0.497$; for the Poisson data, the Chernoff information is 0.012, achieved at $t = 0.481$, and this representation should therefore be preferred.

### 3.2 Invariance under affine transformation

As mentioned in the introduction, the choice of origin and scale for measurements in the data matrix is often arbitrary. The following lemma shows that cluster separation, as measured through Chernoff information, is not affected by this choice.

**Lemma 1** (Chernoff information invariance under affine transformation). Let $W$ be an adjacency matrix from a weighted stochastic block model and, for $a \neq 0$, define $W' = aW + b11^\top$, where $1$ is the all-one vector. For any $k, l \in \{1, \ldots, K\}$,

$$C(F_k, F_l) = C(F'_k, F'_l),$$

where $F_k = \text{Normal}(v_k, \Sigma_k)$ and $F'_k = \text{Normal}(v'_k, \Sigma'_k)$ denote the $k$th component from the limiting mixture distribution of Theorem 1 associated with $W$ and $W'$ respectively.

This lemma has some interesting consequences regarding common data transformations and their effect on spectral clustering.

**Remark 1.** Given an unweighted stochastic block model, rather than using 1 and 0 to respectively represent edges and missing edges, Chernoff information invariance suggests that any other two distinct values could be used.

**Remark 2.** Given a weighted stochastic block model where weights represent p-values, Chernoff information invariance suggests that there is no difference between analysing the matrix with entries $p_{ij}$ or with entries $1 - p_{ij}$.

Based on Lemma 1, it may appear that an affine transformation of the adjacency matrix entries will not affect the geometry of the point cloud. However, by transforming the entries we could potentially change the signature $(p, q)$ of the model and the underlying geometry of the invariant indefinite orthogonal group, $\mathbb{O}(p, q)$. 

---

5
Lemma 2 (Signature change under affine transformation). Let $B$ be a matrix with signature $(p, q)$. Then, depending on $a$ and $b$, the matrix $B' = aB + b11^\top$ has signature:

| $a$   | $b$   | Signature                      |
|-------|-------|--------------------------------|
| $a > 0$ | $b > 0$ | $(p, q)$ or $(p + 1, q - 1)$   |
| $a > 0$ | $b = 0$  | $(p, q)$                       |
| $a > 0$ | $b < 0$  | $(p, q)$ or $(p - 1, q + 1)$   |
| $a < 0$ | $b > 0$  | $(q, p)$                       |
| $a < 0$ | $b = 0$  | $(q, p)$                       |
| $a < 0$ | $b < 0$  | $(q, p)$ or $(q - 1, p + 1)$   |

3.2.1 Example: Beta distributions for p-values

Consider a 2-community weighted stochastic block model where weights represent p-values from a continuous test statistic. We model the p-values using Beta distributions, for $\alpha < 1$,

$$W_{ij} \sim \begin{cases} \text{Beta}(\alpha, 1) & \text{if } z_i = 1, z_j = 1, \\ \text{Uniform}[0, 1] & \text{otherwise}. \end{cases}$$ (3)

Following Corollary 2, there is no difference in Chernoff information between using the matrix $W$, with entries $p_{ij}$, or the matrix $W' = 11^\top - W$, with entries $1 - p_{ij}$, in Algorithm 1. Let $B$ and $B'$ be the corresponding block mean matrices,

$$B = \begin{pmatrix} \frac{\alpha}{\alpha + 1} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}, \quad B' = \begin{pmatrix} \frac{1}{\alpha + 1} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}. $$

Since $\alpha < 1$, the matrix $B$ has signature $(1, 1)$ while $B'$ has signature $(2, 0)$, changing the geometry of latent space. We investigate this model further in Section 4.

4 Application: network anomaly detection

Consider the problem of detecting a cluster of anomalous activity on a network. Assume that a p-value $p_{ij}$ for every unordered pair of nodes on a network can be obtained, quantifying our level of surprise in their activity. For example, a low p-value might occur if, relative to historical behaviour, a much smaller or larger volume of communication was observed [2], if a communication used an unusual channel [3] or took place at a rare time of day [4].

Assume that the network contains an unknown proportion $\pi$ of nodes of interest whose interactions tend to have low associated p-values. Interactions involving the remaining nodes generate p-values with no signal.

We model this using the 2-community stochastic block model specified in Section 3.2.1. One could hope to discover the anomalous cluster by spectrally embedding $W_{ij} = p_{ij}$ or, equivalently, $W'_{ij} = 1 - p_{ij}$. However, familiarity with statistical anomaly detection might suggest using instead $W''_{ij} = -\log p_{ij}$, since the most common method of combining p-values $p_1, \ldots, p_n$ is Fisher’s method [5],

$$-2 \sum_{k=1}^{n} \log p_k.$$  

This provides the uniformly most powerful approach if the p-values are Beta$(\alpha, 1)$ with $\alpha < 1$ under the alternative hypothesis [2]. Under a log transformation, these p-values have an Exp$(\alpha)$, whereas they have an Exp$(1)$ distribution under the null hypothesis.

Figure 2 shows the Chernoff information associated with these two matrix data representations, for $(\alpha, \pi) \in (0, 1)^2$. The log p-value representation appears to dominate over the full range $(\alpha, \pi) \in (0, 1)^2$ and this observation is confirmed in Lemma 3 below. Under this model, it is always preferable to use log p-values.
Lemma 3 (Log p-value dominance). Consider a 2-community stochastic block model with weights $W$ representing p-values modelled by the Beta distributions given in Equation 3 and define $W'_{ij} = -\log p_{ij}$. For all $(\alpha, \pi) \in (0, 1)^2$, $C(F_1, F_2) < C(F'_1, F'_2)$, where $F_k$ and $F'_k$ denote the $k$th component from the limiting mixture distribution of Theorem 1 associated with $W$ and $W'$ respectively.

4.1 Real data: detection of a cyber attack

In attacks on computer networks, attackers move between computers, leaving evidence in the form of anomalous connections between computers (Neil et al., 2013) (Turcotte et al., 2014). While individually these anomalous connections can sometimes be detected, they are often lost among the many unusual but nonetheless benign connections on the network. This calls for an approach that detects clusters of anomalous scores by exploiting network structure.

In this example we consider network log-in events between computers on a computer network. Further details on the data acquisition process can be provided by the authors upon request. For each log-in event, we use Poisson factorisation (Turcotte et al., 2016) to score the likelihood that a given computer logs in to another computer. In the case of log-in events in both directions, we combine the scores using Fisher’s method to produce a symmetric matrix of p-values.

In our first experiment, we insert three anomalous edges connected to three random vertices in the graph, drawing the value from a Beta$(0.2, 1)$ distribution. Figure 3a) shows the spectral embedding of the p-value matrix with entries $1 - p_{ij}$, while Figure 3b) shows the embedding corresponding to the matrix with entries $-\log(p_{ij})$. By Section 3.2.1 there is no advantage to using the matrix with entries $1 - p_{ij}$ over $p_{ij}$, the former is simply chosen so that large entries in this and the log representation indicate unusual events. The log representation results in an embedding which better separates the cluster of synthetic anomalous nodes, shown in red.

Our second experiment analyses data from a different computer network, now containing red-team log-in activity. The embeddings corresponding to the two rival representations are shown in Figure 4. Here, both embeddings seem to do well at separating the red team.

5 Conclusion

The performance of spectral clustering with real-valued matrices was investigated under a weighted stochastic block model assumption, extending recent statistical theory on graphs. Our theory recommends selecting eigenvectors by eigenvalue magnitude and using Gaussian clustering. This allows a choice between data representations using Chernoff information. We have identified cases where this choice is asymptotically immaterial (e.g. when the matrices are equal up to affine
transformation) and other cases where one representation always dominates (e.g. favouring the use of log p-values over p-values for network anomaly detection).

References

Athreya, A., Fishkind, D. E., Tang, M., Priebe, C. E., Park, Y., Vogelstein, J. T., Levin, K., Lyzinski, V., and Qin, Y. (2017). Statistical inference on random dot product graphs: a survey. The Journal of Machine Learning Research, 18(1):8393–8484.

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. The Annals of Mathematical Statistics, 23(4):493–507.

Fisher, R. A. (1934). Statistical methods for research workers. Edinburgh: Oliver & Boyd, 4th edition.

Heard, N. A. and Rubin-Delanchy, P. (2016). Network-wide anomaly detection via the Dirichlet process. In Proceedings of IEEE workshop on Big Data Analytics for Cyber-security Computing.

Heard, N. A. and Rubin-Delanchy, P. (2018). Choosing between methods of combining-values. Biometrika, 105(1):239–246.

Heard, N. A., Weston, D. J., Plataniotis, K., and Hand, D. J. (2010). Bayesian anomaly detection methods for social networks. The Annals of Applied Statistics, 4(2):645–662.
Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. Social networks, 5(2):109–137.

Horn, R. and Johnson, C. (2012). Matrix analysis, second edition. Cambridge University Press.

Neil, J., Hash, C., Brugh, A., Fisk, M., and Storlie, C. B. (2013). Scan statistics for the online detection of locally anomalous subgraphs. Technometrics, 55(4):403–414.

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.

Pardo, L. (2005). Statistical inference based on divergence measures. Chapman and Hall/CRC.

Price-Williams, M., Turcotte, M., and Heard, N. (2018). Time of day anomaly detection. In IEEE European Intelligence and Security Informatics Conference (EISIC2018). IEEE. to appear.

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

Rubin-Delanchy, P., Priebe, C., Tang, M., and Cape, J. (2018). A statistical interpretation of spectral embedding: the generalised random dot product graph. Arxiv preprint at https://arxiv.org/abs/1709.05506.

Tang, M. and Priebe, C. (2019). Limit theorems for eigenvectors of the normalized Laplacian for random graphs. The Annals of Statistics, To appear.

Turcotte, M., Heard, N., and Neil, J. (2014). Detecting localised anomalous behaviour in a computer network. In International Symposium on Intelligent Data Analysis, pages 321–332. Springer.

Turcotte, M., Moore, J., Heard, N., and McPhall, A. (2016). Poisson factorization for peer-based anomaly detection. In 2016 IEEE Conference on Intelligence and Security Informatics (ISI), pages 208–210. IEEE.

Von Luxburg, U. (2007). A tutorial on spectral clustering. Statistics and computing, 17(4):395–416.

Xu, M., Jog, V., and Loh, P.-L. (2017). Optimal rates for community estimation in the weighted stochastic block model. arXiv preprint arXiv:1706.01175.
6 Appendix

6.1 Proof of Lemma 1

Proof. Let \( B', C' \) be the block mean and variance matrices for the affine transformed weighted stochastic block model,

\[
B' = aB + b11^\top, \quad C' = a^2C.
\]

If \( B' = USU^\top \) is the spectral decomposition of \( B' \), then we consider latent positions given by

\[
V = [v_1 | \ldots | v_n]^\top = U[S]^{1/2},
\]

and \( B' = VI_{p\times q}V^\top \). Using this notation, the second moment matrix \( \Delta = V^\top \Pi V \), where \( \Pi = \text{diag}(\pi_1, \ldots, \pi_K) \). We can compute the covariance matrices of the asymptotic Gaussian mixture model distribution from Theorem 1,

\[
\Sigma_k = I_{p\times q}\Delta^{-1}\left(\sum_{i=1}^{K} \pi_i C_{k\ell} v_i v_i^\top\right) \Delta^{-1} I_{p\times q} = a^2I_{p\times q}\Delta^{-1}V^\top \Gamma_k V \Delta^{-1} I_{p\times q},
\]

where \( \Gamma_k = \text{diag}(\pi_1 C_{k1}, \ldots, \pi_K C_{kK}) \). For the Chernoff divergence at \( t \in (0,1) \), we require \( \Sigma_t = (1-t)\Sigma_k + t\Sigma_l \). This has the same form as the above equation, replacing \( k \) with \( t \), where we similarly define \( \Gamma_t = (1-t)\Gamma_k + t\Gamma_l \).

We individually analyse the two terms of the Chernoff divergence in Equation 1. For the first term, we can write \( v_k - v_l = V^\top (e_k - e_l) \), where \( e_i \) is the standard basis vector with 1 in position \( i \) and 0 elsewhere.

\[
(v_k - v_l)^\top \Sigma_t^{-1}(v_k - v_l) = a^{-2}(e_k - e_l)^\top VI_{p\times q} \Delta V^{-1} \Gamma_t^{-1} V^{-1\top} \Delta I_{p\times q} V^\top (e_k - e_l) = a^{-2}(e_k - e_l)^\top B^\top \Pi \Gamma_t^{-1} \Pi B (e_k - e_l) = (e_k - e_l)^\top B^\top \Pi \Gamma_t^{-1} \Pi B (e_k - e_l),
\]

where we have used \( B'(e_k - e_l) = aB(e_k - e_l) \) and the right hand side does not depend on \( a \) or \( b \).

Next, we consider the second term of the Chernoff divergence,

\[
\left|\frac{\Sigma_t}{\Sigma_k}\right|^{1-t} \left|\frac{\Sigma_l}{\Gamma_k}\right|^{t} = \frac{|\Gamma_t|}{|\Gamma_k|^{1-t} |\Gamma_l|^t}.
\]

Neither \( \Gamma_k \) nor \( \Gamma_t \) depend on \( a \) or \( b \). Therefore the Chernoff information is independent of \( a \) and \( b \) for all \( t \in (0,1) \), which implies that the Chernoff information is unaffected by affine transformation. \( \square \)

6.2 Proof of Lemma 2

Proof. The result follows from Corollary 4.3.9 from \textit{Horn and Johnson} (2012). If \( A \in \mathbb{Z}^{n \times n} \) and \( v \in \mathbb{Z}^n \) then, for \( i = 1, \ldots, n-1 \),

\[
\lambda_i(A) \leq \lambda_i(A + vv^\top) \leq \lambda_{i+1}(A),
\]

\[
\lambda_n(A) \leq \lambda_n(A + vv^\top).
\]

Firstly, we shall assume that \( a > 0, b > 0 \). Using the above result with \( v = b^{1/2}1 \) we have

\[
\lambda_i(B) = \lambda_i(aB) \leq \lambda_i(aB + b11^\top) \leq \lambda_{i+1}(aB) = \lambda_{i+1}(B).
\]

Since only the first \( q \) eigenvalues of \( B \) are negative, this means that either the first \( q \) or \( q-1 \) eigenvalues of \( B' \) are negative. Therefore, the signature for \( B' \) is either \( (p,q) \) or \( (p+1,q-1) \).

A version of Corollary 4.3.9 from \textit{Horn and Johnson} (2012) considers matrices of the form \( A - vv^\top \), which proves the lemma for \( a > 0, b < 0 \) using a similar argument. If \( a > 0, b = 0 \), then \( B \) and \( B' \) have the same eigenvalues and, therefore, the same signature.

If \( a < 0 \), then \( \lambda_{n-i}(P) = \lambda_i(aP) \), which swaps the role of \( p \) and \( q \) in the signature but the rest of the proof is unchanged. \( \square \)
6.3 Proof of Lemma 3

Proof. Consider block mean and variance matrices $B, C$ for edges between the two communities with the following form,

$$B = \begin{pmatrix} b_1 & b_2 \\ b_2 & b_2 \end{pmatrix}, \quad C = \begin{pmatrix} c_1 & c_2 \\ c_2 & c_2 \end{pmatrix}.$$

Using Equations 3 and 5 we can compute the Chernoff divergence directly,

$$C_t(F_1, F_2) \equiv g(t) + h(t),$$

where,

$$g(t) = \frac{(b_1 - b_2)^2}{2} \frac{t(1-t)}{(1-t)c_1 + tc_2},$$

$$h(t) = \frac{1}{2} \log \{(1-t)c_1 + tc_2\} - \frac{t}{2} \log c_1 - \frac{1-t}{2} \log c_2. \quad (6)$$

We consider the Chernoff information for the p-values and log p-values stochastic block models. Terms relating to the latter model are denoted using a dash. The block mean and variance matrices for the two models are

$$B = \begin{pmatrix} \frac{\alpha}{\alpha+1} & \frac{1}{\alpha+1} \\ \frac{1}{\alpha+1} & \frac{1}{\alpha+1} \end{pmatrix}, \quad C = \begin{pmatrix} \frac{\alpha}{\alpha+1} & \frac{1}{\alpha+1} \\ \frac{1}{\alpha+1} & \frac{1}{\alpha+1} \end{pmatrix}, \quad B' = \begin{pmatrix} \frac{1}{\alpha} & 1 \\ 1 & 1 \end{pmatrix}, \quad C' = \begin{pmatrix} \frac{1}{\alpha} & 1 \\ 1 & 1 \end{pmatrix}. \quad (7)$$

From the definition of Chernoff information, we have the following upper and lower bounds for the two different data representations,

$$C(F_1, F_2) = \sup_{t \in (0,1)} C_t(F_1, F_2) \leq \sup_{t \in (0,1)} g(t) + \sup_{t \in (0,1)} h(t),$$

$$C(F'_1, F'_2) = \sup_{t \in (0,1)} C_t'(F'_1, F'_2) \geq \sup_{t \in (0,1)} h'(t).$$

The maximum points of these functions can be found by differentiation,

$$\frac{dg}{dt} = 0 \Rightarrow t^* = \frac{\sqrt{c_1}}{\sqrt{c_1} + \sqrt{c_2}} \in (0, 1),$$

$$\frac{dh}{dt} = 0 \Rightarrow t^* = \frac{1}{\log c_2 - \log c_1} - \frac{c_1}{c_2 - c_1} \in (0, 1).$$

It is sufficient to prove that the Chernoff information of the log p-value model dominates the p-value model, if, for all $(\alpha, \pi) \in (0,1)^2$,

$$\sup_{t \in (0,1)} g(t) + \sup_{t \in (0,1)} h(t) \leq \sup_{t \in (0,1)} h'(t).$$

This inequality depends on $\pi$ only via $g(t)$ so we can assume the worst case scenario, $\pi = 1$. Substituting the maximum points into Equations 3 and 7 this inequality leads to the function,

$$f(\alpha) = \sup_{t \in (0,1)} h'(t) - \left( \sup_{t \in (0,1)} g(t) + \sup_{t \in (0,1)} h(t) \right)$$

$$= \frac{1}{2} \log \left( \frac{c_1' - 1}{\log c_1} \right) + \frac{\log c_1'}{2(c_1 - 1)} - \frac{(b_1 - 1/2)^2}{2(\sqrt{c_1} + \sqrt{1/12})^2} - \frac{1}{2} \log \left( \frac{c_1' - 1/12}{\log c_1 + \log 12} \right) - \frac{1}{2} \log c_1 + c_1 \log 12,$$

where $b_1, c_1, c_1'$ are the only parameters in the stochastic block models that depend on $\alpha$. Numerical analysis shows that $f(\alpha) > 0$ for all $\alpha \in (0,1)$, which completes the proof. \qed