Invariant Theory for Hypothesis Testing on Graphs

Priebe, Carey
Johns Hopkins University, Applied Mathematics and Statistics
3400 North Charles Street
Baltimore, Maryland 21218-2682, USA
E-mail: cep@jhu.edu

Rukhin, Andrey
Naval Surface Warfare Center, Sensor Fusion Department
18444 Frontage Road
Dahlgren, Virginia 22448, USA
E-mail: andrey.rukhin@navy.mil

1 Introduction

Following the setting outlined in Priebe et al. [2011] we aim to detect anomlies within attributed graphs. In particular, let $\mathcal{V} = \{1, \ldots, n\}$ be the fixed set of vertices and $\phi : \binom{\mathcal{V}}{2} \to \{0, 1, \ldots, K\}$ be an *edge-attribution* function. The graph on $\mathcal{V}$ is defined to be $G = (\mathcal{V}, \mathcal{E}_\phi)$ where

$$(u, v) \in \mathcal{E}_\phi \iff \phi(u, v) > 0.$$

We say that the edge (u, v) has attribute $c \in \{1, \ldots, K\}$ if $\phi(u, v) = c$. One can view the categorical edge attributes as some mode of the communication event between actors u and v (e.g., a topic label derived from the content of the communication).

The specific anomaly we aim to detect is the "chatter" alternative – a small (unspecified) subset of vertices with altered communication behavior in an otherwise homogeneous setting. Our inference task is to determine whether or not a graph $(\mathcal{V}, \mathcal{E}_\phi)$ includes a subset of vertices $\mathcal{M} = \{v_1, v_2, \ldots, v_m\}$ whose edge-connectivity within the subset exhibits a different behaviour than that found among the remaining vertices in the graph.

To this end we consider the problem of detecting chatter anomalies in a graph using hypothesis testing on a *fusion* of attributed graph invariants. In particular, the focus of this paper is analyzing and comparing the inferential power of the linear attribute fusion of the attributed q-clique invariant

$$T_q^W(G) = \sum_{c_1, \ldots, c_K \in P(\binom{q}{2}, K)} w_{c_1, \ldots, c_K} \sum_{(u_1, \ldots, u_q) \in \binom{\mathcal{V}}{q}} h\left(u_1, \ldots, u_q; c_1, \ldots, c_K\right),$$

where the sum is over the collection of partitions $P(\binom{q}{2}, K)$ of $\binom{q}{2}$ into K non-negative parts, $W = \{w_i\}_{i \in P(\binom{q}{2}, K)}$ are the fusion weights, and the summand $h\left(u_1, \ldots, u_q; c_1, \ldots, c_K\right)$ indicates the event that the vertices $u_1, \ldots, u_q$ are elements of a q-clique with c_r edges of color r. Specifically, we consider the cases $q = 2$ which yields the *size* fusion T_2^W and $q = 3$ which yields the *triangle* fusion T_3^W.

Our random graph model is motivated by the time series model found in Lee and Priebe [forthcoming]: for each vertex $v \in \mathcal{V}$ we assign a latent variable $\boldsymbol{X}^v = (X_1^v, \ldots, X_d^v)$ drawn independently of all other vertices from some d-dimensional distribution. The edge-attribution function will be a random variable where the probability of an edge (u, v) having attribute c is defined to be a some predetermined function of the inner product of the latent variables. We assume that the edge attributes, conditioned on the latent variables, are independent. In this paper, we will assume that $\boldsymbol{X}^v \sim Dirichlet\left(\lambda_0^v, \ldots, \lambda_K^v\right)$ and

$$\mathbb{P}\left\{\phi(u, v) = c\right\} = X_c^u X_c^v$$

for all $(u,v) \in \binom{\mathcal{V}}{2}$ and all $c \in \{1,\ldots,K\}$. This model choice is analogous to the first and second approximations found in Lee and Priebe [forthcoming]: if we write $\lambda^v = (\lambda_0^v,\ldots,\lambda_K^v) = (1+rx_0^v,\ldots,1+rx_K^v)$ for some fixed $(x_0^v,\ldots,x_K^v)$ in the unit simplex and non-negative real r, then $r \to \infty$ yields the first approximation model (i.e., the "independent edge model"). We mention that our approach herein differs from the second approximation in Lee and Priebe [forthcoming]; their second approximation yields a inner product model with truncated Gaussian latent variables.

Related work may be found in Bollobas et al. [2007] Section 16.4 and references therein. We also direct the interested reader to Priebe et al. [2011] in which the authors study other linear attribute fusion invariants; in particular, the authors consider

$$maxd^W(G) = \max_v \sum_{c=1}^{K} w_c \sum_{u \in N[v]} I\{\phi(u,v) = c\}$$

and

$$scan^W(G) = \max_v \sum_{c=1}^{K} w_c \sum_{u,x \in N[v]} I\{\phi(u,x) = c\},$$

where $N[v] = \{u \mid (u,v) \in \mathcal{E}\} \cup \{v\}$ is the closed neighborhood of vertex v in the graph.

Finally, we add that we will restrict ourselves to simple undirected graphs. We will not consider hyper-graphs (hyper-edges consisting of more than two vertices), multi-graphs (more than one edge between any two vertices), self-loops (an edge from a vertex to itself), or weighted edges.

2 Notation

For each positive integer l we use the notation $[l] = \{1,\ldots,l\}$.

For each $v \in \mathcal{V}$ we assign a *latent position* vector $\boldsymbol{X}^v = (X_0^v,\ldots,X_K^v) \sim Dirichlet(\lambda^v)$ for some fixed parameter vector $\lambda^v \in \mathbb{R}_+^{K+1}$. We also assume that the latent positions are independent.

Our null hypothesis assumes a version of homogeneity among the vertices; specifically,

$$\mathbb{H}_0 : \boldsymbol{X}^v = (X_0^v,\ldots,X_K^v) \sim Dirichlet(\lambda) \text{ for all } v \in \mathcal{V}$$

for some Dirichlet parameter vector $\lambda = (\lambda_0,\ldots,\lambda_K)$. Our alternative hypothesis incorporates the anomaly feature described in the preceding section as follows. Assume $m = m(n) < n$ satisfies the following two conditions: $\lim_{n\to\infty} m(n) = \infty$ and $\lim_{n\to\infty} \frac{m(n)}{n} = 0$. Our alternative hypothesis is defined to be

$$\mathbb{H}_1 : \boldsymbol{X}^v = \begin{cases} (Y_0^v,\ldots,Y_K^v) \overset{iid}{\sim} Dir(\eta) & i \in [m], \\ (X_0^v,\ldots,X_K^v) \overset{iid}{\sim} Dir(\lambda) & i \in [n]-[m]. \end{cases}$$

for some fixed Dirichlet parameter vector $\eta = (\eta_0,\ldots,\eta_K)$ and the same $\lambda = (\lambda_0,\ldots,\lambda_K)$ from the null hypothesis. For convenience, we also define $\Lambda = \sum_{0\le c\le K} \lambda_c$ and $H = \sum_{0\le c\le K} \eta_c$.

We define

$$\varepsilon_c = \sum_{(u,v)\in\binom{\mathcal{V}}{2}} \mathbb{I}\{\phi(u,v) = c\}$$

to be the *size* (i.e., number of 2-cliques) of attribute c in the graph. Similarly, for the number of *triangles* (i.e., 3-cliques) we write τ_c, $\tau_{b,c}$, and $\tau_{b,c,d}$ to denote the number of 3-cliques with three c-colored edges, two b-colored and one c-colored edge, and one edge of each of three edge-colors b,c,d, respectively.

Before proceeding, we highlight a relevant property of the mixed moments of the Dirichlet distribution (see Johnson and Kotz [1972]): if $r_1, \ldots, r_s$ are non-negative and $(X_1, \ldots, X_s) \sim Dirichlet(\theta_1, \ldots, \theta_s)$ then

$$(1) \quad E\left[\prod_{i=1}^{s} X^{r_i}\right] = \frac{\Gamma\left(\sum_{i=1}^{s} \theta_i\right) \prod_{j=1}^{s} \Gamma\left(\theta_j + r_j\right)}{\prod_{i=1}^{s} \Gamma\left(\theta_i\right) \Gamma\left(\sum_{j=1}^{s} \left(\theta_j + r_j\right)\right)}$$

where Γ denotes Euler's standard Gamma function. With this property one can compute the exact moments of the Hajek projection of T_2^W and T_3^W under either hypothesis. We write $\nu_{(c)}^{(i)}$ to denote the i-th moment of X_c in the null latent vector and $\nu_{(b,c)}^{(i,j)}$ to denote the joint (i,j)-moment of (X_b, X_c). Similarly, we'll write $\mu_{(c)}^{(i)}$ to denote the i-th moment of Y_c in the anomalous latent vector and $\mu_{(b,c)}^{(i,j)}$ to denote the joint (i,j)-moment of (Y_b, Y_c).

3 Analysis

We will appeal to Hajek's Projection method, detailed in Nowicki and Wierman [1988], in order to demonstrate the asymptotic normality of the fusion invariants in this article. This approach is outlined as follows: We define the *projection* of the fusion T to be the centered sum of independent random variables

$$T^* = \sum_{v \in \mathcal{V}} E\left[T \mid \boldsymbol{X}^v\right] - (n-1)E[T].$$

For both the size and triangle fusion we aim to show that

$$\frac{T - E\left[T\right]}{\sqrt{Var\left(T^*\right)}} = \frac{T - T^*}{\sqrt{Var\left(T^*\right)}} + \frac{T^* - E\left[T\right]}{\sqrt{Var\left(T^*\right)}} \xrightarrow{\mathcal{D}} N(0,1).$$

To this end, one appeals to Chebyshev's Inequality to show that $Var\left(T - T^*\right) = o(Var\left(T^*\right))$ (see Nowicki and Wierman [1988] for the detailed argument). Specifically, if

$$\mathbb{P}\left\{\frac{|T - T^*|}{\sqrt{Var\left(T^*\right)}} \geq \epsilon\right\} \leq \frac{Var\left(T - T^*\right)}{\epsilon^2 Var\left(T^*\right)} \to 0$$

for any positive ϵ, then

$$\frac{T - T^*}{\sqrt{Var\left(T^*\right)}} + \frac{T^* - E\left[T\right]}{\sqrt{Var\left(T^*\right)}} \xrightarrow{\mathcal{D}} N(0,1)$$

by applying the Central Limit Theorem to the normalized sum of independent random variables in the second term of the left-hand side.

3.1 The Attributed Size Fusion

For each $c \in K$ define

$$\varepsilon_c = \sum_{(u,v) \in \binom{\mathcal{V}}{2}} \mathbb{I}\left\{\phi\left(u,v\right) = c\right\}$$

to be the number of edges of color c in the graph. The linear attribute fusion with parameter $W = (w_1, \ldots, w_K)$ is defined to be

$$T_2^W = \sum_{c=1}^{K} w_c \varepsilon_c.$$

3.1.1 The Attributed Size Fusion under $\mathbb{H}_0$

We present all relevant terms within the Hajek Projection of the attributed size fusion under the null hypothesis.

The expectation of T_2^W under the null is given by

$$E_0\left[T_2^W\right] = \binom{n}{2}\sum_{c=1}^{K} w_c\left[\nu_{(c)}^{(1)}\right]^2.$$

$E_0\left[T_2^W \mid \boldsymbol{X}^a\right]$ for any fixed $a \in \mathcal{V}$ is given by

$$E_0\left[T_2^W \mid \boldsymbol{X}^a\right] = (n-1)\sum_{c=1}^{K} w_c\left[\nu_{(c)}^{(1)}\right]X_c^{(a)} + \binom{n-1}{2}\sum_{c=1}^{K} w_c\left[\nu_{(c)}^{(1)}\right]^2.$$

We can now evaluate T^* under the null:

$$\begin{aligned} T^* &= \sum_{a\in[n]} E\left[T_2^W \mid \boldsymbol{X}^a\right] - (n-1)E[T_2^W] \\ &= (n-1)\sum_{a\in[n]}\sum_{1\le c\le K} w_c\left[\nu_{(c)}^{(1)}\right]X_c^{(a)} - \binom{n}{2}\sum_{c=1}^{K} w_c\left[\nu_{(c)}^{(1)}\right]^2. \end{aligned}$$

The variance of this sum of independent and identically distributed random variables is

$$Var_0\left(T^*\right) = \Theta\left(n^3\right).$$

As $Var_0\left(T_2^W - T^*\right) \le \binom{n}{2}(2+1)^2 E\left[\mathbb{I}\left\{\phi\left(u,v\right) > 0\right\}\right] = o(Var_0\left(T^*\right))$ by the Cauchy-Schwarz Inequality (see Nowicki and Wierman [1988] for full details), we have the desired convergence to the standard normal distribution.

3.1.2 The Attributed Size Fusion under $\mathbb{H}_1$

For the alternative we write the edge attribution function as

$$\mathbb{P}\left\{\phi\left(u,v\right) = c\right\} = \begin{cases} Y_c^u Y_c^v & u,v \in [m], \\ Y_c^u X_c^v & u \in [m], v \in [n]-[m], \\ X_c^u X_c^v & u,v \in [n]-[m]. \end{cases}$$

We perform a similar but more involved analysis to deduce the limiting distribution of the attribued size fusion of the graph under these conditions, yielding

$$Var_1\left(T^*\right) = \Theta\left(n^3\right)$$

and

$$Var_1\left(T_2^W - T^*\right) = \Theta\left(n^2\right) = o(Var_1\left(T^*\right))$$

as desired.

3.1.3 Asymptotic Power Analysis of the Attributed Size Fusion

Returning to the context of hypothesis testing, assume we are interested in performing an α-level hypothesis test to determine whether or not the graph includes an anomalous set of m vertices whose underlying latent distribution differs from the the null component of the graph. We define $\beta_2^W = \lim_{n\to\infty}\mathbb{P}_1\left\{T_2^W > c_\alpha\right\}$ where $c_\alpha = c(\alpha, n)$ is the α-level critical value of the test.

Fix $c \in [K]$. The difference of the corresponding terms the hypotheses means can be written as

$$E_1[\varepsilon_c] - E_0[\varepsilon_c] = D_1^{(c)} + D_2^{(c)} = \binom{m}{1}\binom{n-m}{1}\nu_{(c)}^{(1)}\left(\mu_{(c)}^{(1)} - \nu_{(c)}^{(1)}\right) + \binom{m}{2}\left(\left[\mu_{(c)}^{(1)}\right]^2 - \left[\nu_{(c)}^{(1)}\right]^2\right)$$

(here $D_i^{(c)}$ corresponds to the edge-count that includes edges with exactly i anomalous vertices). The reader can verify that $\frac{Var_1(T^*)}{Var_0(T^*)} \to 1$. Moreover, given that the limiting distribution (under the null) is normal, we write

$$c_\alpha = z_\alpha\sqrt{Var_0(T^*)} + E_0\left[T_2^W\right]$$

and thus

$$\beta_2^W = \mathbb{P}\left\{Z > z_\alpha - \lim_{n\to\infty}\left(\frac{E_1\left[T_2^W\right] - E_0\left[T_2^W\right]}{\sqrt{Var_0(T^*)}}\right)\right\}.$$

Recall that $Var_0(T^*) = \Theta(n^3)$; thus, if

$$\sum_c w_c D_1^{(c)} \neq 0$$

(i.e. there is signal in the null-to-anomaly connectivity) then the limiting power $\beta_2^W > \alpha$ when $\frac{m(n-m)}{\sqrt{n^3}} \nrightarrow 0$ or, equivalently, when $m = \Omega(\sqrt{n})$ (similarly, if $m = \omega(\sqrt{n})$ then $\beta_2^W \to 1$). Furthermore, if

$$\sum_c w_c D_1^{(c)} = 0$$

(i.e. there is no signal in the null-to-anomaly connectivity) the limiting power $\beta_2^W > \alpha$ when $\sum_c w_c D_2^{(c)} \neq 0$ and $\frac{m^2}{\sqrt{n^3}} \nrightarrow 0$ (which is equivalent to $m = \Omega(\sqrt[4]{n^3})$). Moreover, if $m = \omega(\sqrt[4]{n^3})$ under these conditions then $\beta_2^W \to 1$.

It follows that the optimal choice of weights $(w_1, \ldots, w_K)$ is the one which maximizes the expression

$$\lim_{n\to\infty}\left(\frac{E_1\left[T_2^W\right] - E_0\left[T_2^W\right]}{\sqrt{Var_0(T^*)}}\right)$$

in either of the two above-mentioned cases.

3.2 The Attributed Number of Triangles Fusion

We begin by writing

$$\tau = \sum_{c\in[K]}\tau_c + \sum_{b\neq c}\tau_{b,c} + \sum_{d\neq b,c}\tau_{b,c,d}.$$

We denote the number-of-triangles fusion invariant to be

$$T_3^W = \sum_{c\in[K]} w_c\tau_c + \sum_{b\neq c} w_{b,c}\tau_{b,c} + \sum_{d\neq b,c} w_{b,c,d}\tau_{b,c,d}.$$

Similar to what was done in the previous section, we obtain

$$E_0[T^*] = \binom{n}{3}\left[\sum_{c\in[K]} w_c\left[\nu_{(c)}^{(2)}\right]^3 + \sum_{b\neq c} w_{b,c}3\nu_{(b)}^{(2)}\left[\nu_{(b,c)}^{(1,1)}\right]^2 + \sum_{d\neq b,c} w_{b,c,d}3\nu_{(b,c)}^{(1,1)}\nu_{(b,d)}^{(1,1)}\nu_{(c,d)}^{(1,1)}\right]$$

and

$$Var_0\left(T^*\right) = \Theta\left(n\binom{n-1}{2}^2\right)$$

under the null and

$$E_1\left[T^*\right] = \Theta\left(\sum_{i=0}^{3}\binom{m}{i}\binom{n-m}{3-i}\right)$$

and

$$Var_1\left(T^*\right) = \Theta\left(n\binom{n-1}{2}^2\right)$$

under the alternative. Again, $Var_0\left(T_3^W - T^*\right) = o(Var_0\left(T^*\right))$ and $Var_1\left(T_3^W - T^*\right) = o(Var_1\left(T^*\right))$.

As in the case with the attributed size fusion, we are interested in performing an α-level hypothesis test.

The terms within the difference in means can be expressed as

$$\begin{aligned} E_1\left[T^*\right] - E_0\left[T^*\right] &= \sum_{c\in[K]} D^{(c)} + \sum_{b\neq c} D^{(b,c)} + \sum_{d\neq b,c} D^{(b,c,d)} \\ &= \Theta\left(\sum_{i=1}^{3}\binom{m}{i}\binom{n-m}{3-i}\delta_i\left(H,\Lambda\right)\right) \end{aligned}$$

where $\delta_i(H,\Lambda)$ is the mixed-moments difference when there are i anomalous vertices in a 3-clique.

The reader can verify that $\frac{Var_1(T^*)}{Var_0(T^*)} \to 1$. Since $Var_0\left(T^*\right) = \Theta(n^5)$, we have that the limiting power $\beta_3^W > \alpha$ when when $m = \Omega(\sqrt[2i]{n^{2i-1}})$ and the corresponding mixed-moments expression $\delta_i(H,\Lambda)$ is non-zero.

4 Conclusion

We have presented preliminary results for linear attribute fusion for clique sizes $q = 2$ and 3 in terms of inferential power when detecting the prescribed anomaly within our model. In general, the most powerful choice of q depends on m as a function of n and on the Dirichlet parameter vectors λ and η through the mixed moments.

References

B. Bollobas, S. Janson, and O. Riordan. The phase transition in inhomogeneous random graphs. *Random Structures and Algorithm*, 31:3–122, 2007.

N. Johnson and S. Kotz. *Distributions in Statistics: Continuous Multivariate Distributions*. Wiley, New York, 1972.

N. Lee and C. E. Priebe. A Latent Process Model for Time Series of Attributed Random Graphs. *Statistical Inference for Stochastic Processes*, forthcoming.

K. Nowicki and J. Wierman. Subgraph counts in random graphs using incomplete u-statistics methods. *Discrete Mathematics*, 72:299–310, 1988.

C. E. Priebe, N. Lee, Y. Park, and M. Tang. Attribute fusion in a latent process model for time series of graphs. *The 2011 IEEE Workshop on Statistical Signal Processing (SSP2011)*, 2011.