Vertex nomination via attributed random dot product graphs

Marchette, David
Naval Surface Warfare Center
18444 Frontage Rd.
Dahlgren, VA 22448, USA
E-mail: david.marchette@navy.mil

Priebe, Carey
Johns Hopkins University, Applied Mathematics and Statistics
Baltimore, MD 21218, USA
E-mail: cep@jhu.edu

Coppersmith, Glen
Johns Hopkins University, Center of Excellence for Human Language Technology
Baltimore, MD 21218, USA
E-mail: coppersmith@gmail.com

Introduction

The vertex nomination problem addressed in this paper, introduced in Coppersmith and Priebe [2011] and illustrated in Figure 1, involves a (simple, undirected) graph in which vertices have associated attributes ("1" and "2", say; black and white in the figure). However, we observe the vertex attributes for only a (small) subset of the vertices. One of the vertex attributes identifies vertices of particular interest ("1", say; black in the figure), and we wish to nominate from the collection of vertices with unobserved attribute (the candidate set; gray in the figure) for further investigation. For example, we might nominate the candidate vertices which connect to the most known vertices of interest, or those with connectivity pattern most similar to that of the known vertices of interest.

Previous work in inferring a small region of inhomogeneity can be found in Priebe et al. [2005], Pao et al. [2010], Priebe et al. [2010] and Grothendieck et al. [2010]. These consider unattributed graphs, as well as edge-attributed graphs, and the inference involves determination of whether there exists a small collection of vertices connecting at a higher rate (and, in the case of edge-attributed graphs, with a distinguished edge-attribute distribution) than the majority of vertices. In effect, these manuscripts are concerned with the question of detection: is there a collection of anomalous vertices? In this paper we consider the related problem in which we have both vertex attributes and

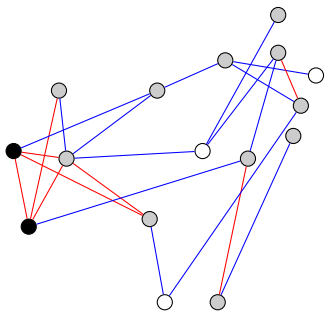

Figure 1: Illustration of the vertex nomination problem: given a graph with a few observed vertex attributes (black and white vertices), we wish to nominate from the candidate set (gray vertices) vertices which are likely to be of particular interest (those with unobserved vertex attribute being truly black, say).

edge attributes (e.g., topics extracted from the content of communication), and we observe the vertex attribute for only some of the vertices. We wish to use this edge and vertex attribute information to nominate other vertices for investigation.

Framework

A graph is a pair $G = (V, E)$ of vertices $V = \{1, \ldots, n\}$ and edges $E \subset V^{(2)}$, where $V^{(2)}$ denotes the set of unordered pairs of vertices. Let A denote the adjacency matrix for G, and let $deg(v)$ denote the degree of vertex v. We define an attributed graph $G^a = (V, E^a, \varphi_V, \varphi_E)$ as follows. Let $\Phi_V = \{1, \ldots, K_V\}$ and $\Phi_E = \{0, 1, \ldots, K_E\}$, and consider vertex attribution function $\varphi_V : V \to \Phi_V$ and edge attribution function $\varphi_E : V^{(2)} \to \Phi_E$. These functions attribute each vertex (resp. edge) of G^a with an element of Φ_V (resp. Φ_E). The edge set E^a for G^a is given by $E^a = \{vw : \varphi(vw) > 0\}$; we write $v \sim w$ to mean that $vw \in E$ or E^a. For simplicity, we will assume that $K_E = K_V = 2$, so that there are two distinguished attributes for each of the vertices and edges.

For the vertex nomination problem, we write $\mathcal{M} = \{v : \varphi_V(v) = 1\}$; this is the collection of vertices of particular interest. We observe not $G^a = (V, E^a, \varphi_V, \varphi_E)$ but $G^o = (V, E^a, \varphi'_V, \varphi_E)$, where $\varphi'_V : V \to \Phi_V \cup \{0\}$ and $\varphi'_V(v) = \varphi_V(v) = 1$ for $v \in \mathcal{M}' \subset \mathcal{M}$ and $\varphi'_V(v) = 0$ for $v \in V \setminus \mathcal{M}'$. Thus we observe a subset $\mathcal{M}'$ of cardinality $m' = |\mathcal{M}'|$ of the vertices of particular interest, and we wish to infer others – we wish to nominate vertices from $V \setminus \mathcal{M}'$ which are, we hope, in $\mathcal{M} \setminus \mathcal{M}'$.

Numerous generalization opportunities are apparent, but this formulation provides a simple framework for an initial investigation of the vertex nomination problem.

Methods

Random dot product graph (RDPG) models are discussed in Young and Scheinerman [2007], Marchette and Priebe [2008], Scheinerman and Tucker [2010] and are a special case of the latent position models of Hoff et al. [2002]. These latent position models posit a "social space" associated with the vertices, where the relative positions in this social space determine the relationship (edge) probabilities. We will first give the basic definitions, then discuss estimation in the model.

The basic idea is that each vertex v has associated with it a vector $x_v \in \mathbb{R}^d$, and these vectors determine the edge probabilities of the random graph via

$$P[v \sim w] = x_v^T x_w.$$

(Obviously, the vectors must satisfy $0 \leq x_v^T x_w \leq 1$.) We denote by X the $n \times d$ matrix whose v^{th} row corresponds to x_v.

Given a graph $G = (V, E)$ and a target dimensionality d for the latent vectors, we can fit an RDPG model using the iterative approach described in Scheinerman and Tucker [2010]. Alternatively, we can obtain an estimate $\widehat{X}$ in a single step as follows:

1. Let $\widetilde{A} = A + D$, where $D = deg(v)/(n-1)$ is the diagonal matrix with the normalized vertex degrees on the diagonal.
2. Compute the eigenvectors U and eigenvalues Λ of $\widetilde{A}$, and set all negative entries in Λ to zero.
3. $\widehat{X} = U_{1,\ldots,d}\sqrt{\Lambda_{1,\ldots,d}}$.

This yields $\widehat{X} = \arg\min_X ||\widetilde{A} - X^T X||_F$. Imputing the diagonal of $\widetilde{A}$ allows for a one-step solution, in contrast with the algorithm proposed in Scheinerman and Tucker [2010]. The justification for the choice $deg(v)/(n-1)$ is based on the observation that $E[deg(v)] = \sum_{w \neq v} x_v^T x_w$; if all the vectors are the same ($x_w = x_v \forall w$) then $E[deg(v)] = (n-1)x_v^T x_v$. Under the assumption that the latent vectors

are random variables and are independent and identically distributed, $E[deg(v)] \leq (n-1)E[X_v^T X_v]$ by Cauchy-Schwarz.

We extend the RDPG model to allow for edge attributes in a very natural way. The idea is to posit that the edge existence and edge attributes are fundamentally tied. Intuitively, we are modeling the edges as if they were communications, with the attributes corresponding to topics, and the vectors in the RDPG model encoding the interest level of each individual in each of the topics.

Let us consider the case where we allocate one dimension per edge attribute. Since $\Phi_E = \{0, 1, \cdots, K\}$, we define X to be an $n \times K$ matrix where each row x_v statisfies $x_{vk} \geq 0$ for all k and $\sum_k x_{vk} \leq 1$. (This is sufficient, but not necessary, to guarantee that the vectors satisfy $0 \leq x_v^T x_w \leq 1$.) Then the attributed RDPG model ARDPG(X) is given by

$$P[v \sim w] = x_v^T x_w \quad \text{and} \quad P[\varphi_E(vw) = k | v \sim w] = \frac{x_{vk} x_{wk}}{x_v^T x_w} \quad \text{for} \quad k = 1, \cdots, K.$$

Thus,

$$P[v \sim w \wedge \varphi_E(vw) = k] = x_{vk} x_{wk}.$$

That is, the random variable $\varphi_E(vw)$ is distributed according to the discrete distribution

$$\varphi_E(vw) \sim Discrete(\{0, 1, \cdots, K\}, [1 - \sum_{k=1}^{K} x_{vk} x_{wk}, x_{v1} x_{w1}, \cdots, x_{vK} x_{wK}]).$$

If the matrix X is itself random – for instance, if the individual vertex vectors (rows of X) are obtained by first drawing $X_v \sim Dirichlet(\alpha_v)$ where the $\alpha_v \in \mathcal{R}_+^{K+1}$ so that the vectors X_v satisfy the constraint that $X_v^T X_w \in [0, 1]$ – then the random variables $\varphi_E(vw)$ for ARDPG(X) are not independent but are *conditionally* independent, given X. (In a slight abuse of standard notation, we will consider $X_v \sim Dirichlet(\alpha_v)$ with $\alpha_v \in \mathcal{R}_+^{K+1}$ to be a length K random vector, with $X_{vk} \geq 0$ for all k and $\sum_k X_{vk} \leq 1$; that is, we drop the $(K+1)$st element of a standard Dirichlet in which $X_{v(K+1)} = 1 - \sum_{k=1}^{K} X_{vk}$.)

It is straightforward to extend this model to allow the vectors corresponding to each attribute to be multi-dimensional.

We modify the unattributed RDPG algorithm to produce an estimation of the vectors in the attributed case:

1. Let $L(G, d)$ denote the linear algebra approach to fit vectors of dimension d to the unattributed graph G. Thus $L(G, d)$ is an $n \times d$ matrix $\widehat{X}$ that is the estimate of the RDPG vectors (the parameters of the RDPG model) that produced G.

2. Given attributed graph G^a, for $k \in \{1, \ldots, K\}$ let $G_k^a = (V, E_k^a)$ where $E_k^a = \{vw : \varphi_E(vw) = k\}$ is the subset of edges in E^a with attribute k.

3. $\hat{x}_k = L(G_k^a, 1)$.

4. $\widehat{X}^o = (\hat{x}_1, \ldots, \hat{x}_K)$.

We can ensure that the resulting vectors are in the first orthant, although post-processing is necessary if we demand that the vectors be in the simplex (or the unit ball) so that $x_v^T x_w$ is guaranteed to be in $[0, 1]$. The simplicity of the algorithm is due to the decomposition in Step 2. It must be noted, however, that something is lost due to this decomposition – in the original attributed graph G^a each edge has exactly one attribute, while this dependency amongst the G_k is lost in our algorithm as we process each G_k independently. The justification for this approach (beyond "it simplifies" and "it works") is based on the observation that for large n the dependency is negligible.

Given an observed attributed graph G^o with only some of the vertex attributes observed, the vertex nomination procedure we propose is as follows. (1) Fit an attributed RDPG, obtaining $\widehat{X}^o$. Notice that this estimate does not use vertex attributes. (2) Using the subset $\mathcal{M}'$ of vertices with observed vertex attributes, rank the candidate vertices $V \setminus \mathcal{M}'$ according to their relationship to $\mathcal{M}'$. Having recovered latent space representations for the vertices in (1), the inferential step (2) may be performed in numerous ways; we present a few simple approaches in the experiments below.

Simulation Experiments

We define an attributed RDPG model $\kappa(n, \pi_{V\setminus\mathcal{M}}, m, \pi_{\mathcal{M}}, r)$ for the vertex nomination task as follows. Let $\pi_{V\setminus\mathcal{M}}$ and $\pi_{\mathcal{M}}$ be in the standard (unit) K-simplex $\Delta^K \subset \mathbb{R}^{K+1}$; that is, they are $(K+1)$-dimensional probability vectors. Let $n > m > 0$ and $r > 0$ be given. Consider

$$X_{V\setminus\mathcal{M}} \overset{iid}{\sim} Dirichlet(r\pi_{V\setminus\mathcal{M}} + \vec{1})$$

and

$$X_{\mathcal{M}} \overset{iid}{\sim} Dirichlet(r\pi_{\mathcal{M}} + \vec{1})$$

to be $(n-m) \times K$ and $m \times K$-dimensional matrices, respectively. Write $X = (X_{V\setminus\mathcal{M}}, X_{\mathcal{M}})$ as the $n \times K$ matrix of (random) vertex vectors. Then $\kappa(n, \pi_{V\setminus\mathcal{M}}, m, \pi_{\mathcal{M}}, r) = \text{ARDPG}(X)$ is our attributed random dot product graph model, with each dimension corresponding to an edge attribute as described above. Again, the extension to multi-dimensional vertex vectors for each attribute is straightforward.

$$G^a \sim \kappa(n, \pi_{V\setminus\mathcal{M}}, m, \pi_{\mathcal{M}}, r),$$

and for $0 < m' < m$ the observed attributed graph

$$G^o \sim \kappa(n, \pi_{V\setminus\mathcal{M}}, m, \pi_{\mathcal{M}}, r; m'),$$

involves selecting a random subset $\mathcal{M}' \subset \mathcal{M}$ – all $n-m$ vertex attributes for $V \setminus \mathcal{M}$ are missing and $m - m'$ of the vertex attributes for $\mathcal{M}$ are missing completely at random.

For our simulation experiments, we consider $\pi_{V\setminus\mathcal{M}} = [0.2, 0.2, 0.6]^T$ and $\pi_{\mathcal{M}} = [q, 0.2, 0.8 - q]^T$ with q ranging from 0.2 (no signal) to 0.8. The parameter r controls the variability of the Dirichlet random vectors; $r = 0$ gives a uniform distribution on the simplex (no signal) and $r \to \infty$ yields point mass (no variability); we use $r = 100$. We consider $m = 10$ and $m' = 5$, and consider two cases for n (100 and 250).

Table 1 presents our simulation results, based on 1000 Monte Carlo replicates. We obtain the estimate of the latent vertex vector matrix $\widehat{X}^o$ as described above, and then rank the candidate vertices $V \setminus \mathcal{M}'$ and evaluate based on (1) $p^* = P[v^* \in \mathcal{M} \setminus \mathcal{M}']$ where v^* is the top-ranked candidate vertex and (2) Normalized Sum of Reciprocal Ranks given by

$$\text{NSRR} = \left(\sum_{v \in \mathcal{M}\setminus\mathcal{M}'} \frac{1}{\text{rank}(v)} \right) \Big/ \left(\sum_{i=1}^{m-m'} \frac{1}{i} \right).$$

Our candidate vertex rankings, based on the latent space estimate $\widehat{X}^o$, are given by (1) the number $N(v)$ of observed signal vertices $\mathcal{M}'$ amongst the m' nearest neighbors of v and (2) the posteriors $\rho(v)$ from a linear discriminant analysis classifier.

As expected, the performance improves as q increases. Note that in this case the performance improves as n increases. While the latent vertex vector estimation improves as n increases, larger n (for fixed m, m') results in a harder vertex nomination problem. This trade-off deserves further study.

		q	0.2	0.3	0.4	0.5	0.6	0.7	0.8
$n = 100$	$N(v)$	p^*	0.05	0.10	0.26	0.50	0.73	0.92	0.98
		NSRR	0.09	0.12	0.23	0.40	0.60	0.79	0.91
	$\rho(v)$	p^*	0.07	0.20	0.49	0.78	0.93	0.98	0.99
		NSRR	0.13	0.24	0.46	0.68	0.83	0.92	0.96
$n = 250$	$N(v)$	p^*	0.02	0.07	0.30	0.70	0.93	0.99	1.00
		NSRR	0.04	0.07	0.23	0.53	0.79	0.92	0.97
	$\rho(v)$	p^*	0.01	0.20	0.62	0.92	0.98	1.00	1.00
		NSRR	0.05	0.22	0.54	0.82	0.93	0.98	0.99

Table 1: Vertex nomination results as a function of q, based on 1000 Monte Carlo replicates (see text).

Experiment on Enron Graphs

The Enron email corpus has been widely studied. Using the data described in Priebe et al. [2005], we construct an attributed graph $G^a_{[t_{min},t_{max}]}$ on the $n = 184$ vertices by including an edge $v \sim w$ if and only if there is at least one message between v and w during the time interval $[t_{min}, t_{max}]$ and attributing this edge based on classification of topics extracted from the content of the messages. For time interval T_1 (May 7, 2001 - Sep 23, 2001) we identify a partition $(\mathcal{M}, V \setminus \mathcal{M})$ for which $\mathcal{M}$ represents a small ($m = |\mathcal{M}| = 10$) collection of vertices connecting at a higher rate and with a distinguished edge-attribute distribution compared to $V \setminus \mathcal{M}$. Figure 2 presents the results of our latent vector extraction process on $G^a_{T_1}$ and on $G^a_{T_0}$ for time interval T_0 (Sep 24, 2001 - Feb 4, 2002) preceding T_1. A two-sample Wilcoxon rank sum test on the ranks r_v of the nearest element of $\mathcal{M}$ (not including v itself if $v \in \mathcal{M}$) to v (this test is closely related to the NSRR performance criterion for the vertex nomination task) yields p-values $p < 0.01$ for $G^a_{T_1}$ and $p \approx 0.7$ for $G^a_{T_0}$. That is, statistically significant signal ($\mathcal{M}$ vs. $V \setminus \mathcal{M}$) is identified based on these ranks for T_1 but not for T_0.

Discussion

The experiments presented above show that vertex vector estimation in the attributed RDPG model may be a viable approach to addressing the vertex nomination problem. Furthermore, sparse matrix methods make the approach suitable for large graphs. The methods presented are easily generalized to cases where there are more than two vertex attributes as well as observed vertex attributes for more than one attribute (as opposed to just the attribute-of-interest).

We have assumed in this work that the edges and their attributes, as well as whatever vertex attributes we observe, are perfectly observed. In real problems there may be some error in the attribution processes – some edges may be unobserved, and the observed vertex and edge attributes may have errors. Also, in applications such as the one considered in Priebe et al. [2005] and related papers, one may observe multiple random graphs on the vertices – a time series of graphs – and utilizing this extra information, if done properly, can improve performance. Furthermore, we have assumed that the edge-existence process and the edge-attribution process are both completely determined by the same vectors. It is straightforward to extend the model to allow some of the dimensions in the vectors to be associated with only the edge-existence, while others are associated only with the edge-attribution, and similar variations. Estimation procedures in these cases would be of great interest.

Finally, Bayesian methods are applicable to our vertex nomination problem. Prior distributions for the parameters of the attributed RDPG model $\kappa(n, \pi_{V\setminus\mathcal{M}}, m, \pi_{\mathcal{M}}, r)$ allow the potential for superior latent vector estimates and Bayesian model averaging.

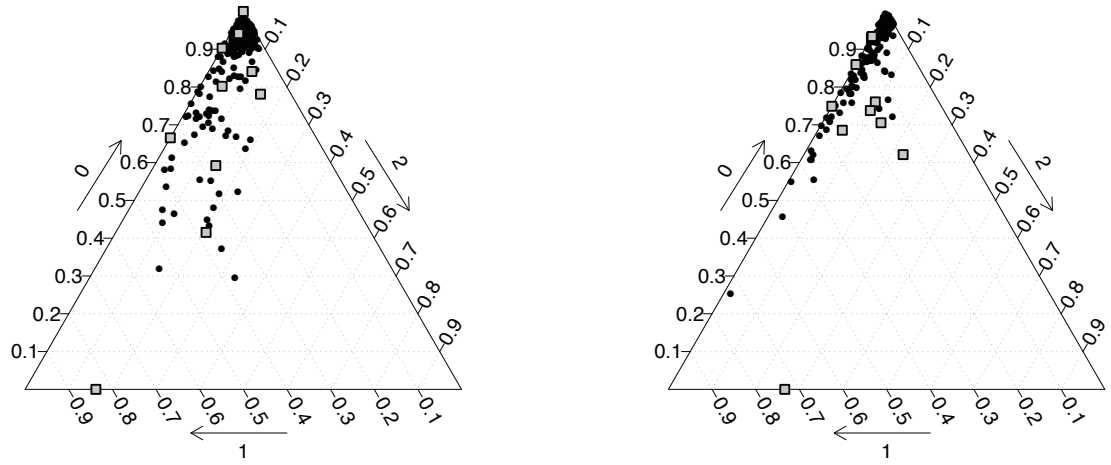

Figure 2: The results of our latent vector extraction process on two Enron graphs, $G^a_{T_0}$ (left) and $G^a_{T_1}$ (right). The symbols correspond to $\mathcal{M}$ (squares) and $V \setminus \mathcal{M}$ (circles). Inference suggests that vertex nomination via attributed random dot product graphs works for this real data (see text).

References

G. A. Coppersmith and C. E. Priebe. Vertex nomination via content and context. *Submitted for publication*, 2011.

J. Grothendieck, C. E. Priebe, and A. L. Gorin. Statistical inference on attributed random graphs: Fusion of graph features and content. *Computational Statistics and Data Analysis*, 54:1777–1790, 2010.

P. D. Hoff, A. E. Raftery, and M. S. Handcock. Latent space approaches to social network analysis. *JASA*, 97:1090–1098, 2002.

D. J. Marchette and C. E. Priebe. Predicting unobserved links in incompletely observed networks. *Computational Statistics and Data Analysis*, 52:1373–1386, 2008.

H. Pao, G. A. Coppersmith, and C. E. Priebe. Statistical inference on random graphs: Comparative power analyses via monte carlo. *Journal of Computational and Graph Statistics*, 2010.

C. E. Priebe, J. M. Conroy, D. J. Marchette, and Y. Park. Scan statistics on enron graphs. *Computational and Mathematical Organization Theory*, 11:229–247, 2005.

C. E. Priebe, Y. Park, D. J. Marchette, J. M. Conroy, J. Grothendieck, and A. L. Gorin. Statistical inference on attributed random graphs: Fusion of graph features and content: An experiment on time series of enron graphs. *Computational Statistics and Data Analysis*, 54:1766–1776, 2010.

E. R. Scheinerman and K. Tucker. Modeling graphs using dot product representations. *Computational Statistics*, 25:1–16, 2010.

S. J. Young and E. R. Scheinerman. Random dot product graph models for social networks. *Proc. 5th ICAM*, pages 138–149, 2007.