Detecting Time-dependent Structure in Network Data via a new Class of Latent Process Models

Lucy F. Robinson¹ and Carey E. Priebe²

¹. Department Epidemiology and Biostatistics, Drexel University ². Department of Applied Mathematics and Statistics, Johns Hopkins University
Abstract

We introduce a new class of latent process models for dynamic network data with the goal of detecting time-dependent structure. Network data are often observed over time, and static network models for such data may fail to capture relevant dynamic features. We introduce a new technique for finding distinct subpopulations of vertices. A network is observed over time, with attributed edges appearing at random times. At unknown time points, subgroups of vertices may exhibit a change in behavior. Such changes may take the form of a change in the overall probability of connection within or between subgroups, or a change in the distribution of edge attributes. A mixture distribution for latent vertex positions is used to detect heterogeneities in connectivity behavior over time and over vertices. The probability of edges with various attributes at a given time is modeled using a latent-space stochastic process associated with each vertex. A random dot product model is used to describe the dependency structure of the graph. As an application we analyze the Enron email corpus.

1 Introduction

Network data are often observed over time, with evolving connections between nodes. Our goal is to detect time-dependent structure in network data, finding change points at which the characteristics of the network are altered. We consider "streaming" network data, a system in which connections between $n$ vertices are observed over a time period $(0, T)$. At random times $(t_1, t_2, \ldots, t_N)$, undirected connections between edges are created, each of which may be associated with a categorical attribute contained in $\{1, 2, \ldots, K\}$. At unknown time points, subgroups of vertices may exhibit a change in behavior. Such changes may take the form of a change in the overall probability of connection within or between subgroups, a change in the distribution of edge attributes, or both. Examples of emerging subpopulations with distinct behavior may include bursts of "chatter" activity among a group of individuals, or the creation of new communities of vertices such as parents of members of a little league team at the beginning of a new season.

Real world networks, particularly social networks, generally exhibit structure beyond that which can be explained by simple stochastic models. Common features of network structure include heterogeneity across vertices with respect to number and types of connections observed, transitivity, and clusters of vertices which are more likely to be connected to one another. One approach to modeling network structure is to partition the graph, grouping together similar vertices. Similarity can be defined in various ways. Many analyses of network data focus on discovering community structure, in which groups of vertices that have a high degree of connection between them belong to the same community or block (Girvan and Newman, 2002; Airoldi et al., 2008). Vertices can also be classified according to structural equivalence. A group of structurally equivalent
vertices have the same “role” in the network, exhibiting similar patterns of connectivity with other groups of vertices. Stochastic block models (Wang and Wong, 1987; Snijders and Nowicki, 1997) parameterize block structure by assigning within and between-block probabilities of connection, and weighted membership vectors specifying to which blocks a vertex is likely to belong.

A variety of approaches have been applied to modeling individual vertices’ connectivity properties. A detailed review of statistical network modeling is available in Goldenberg et al. (2010). We use a latent space model in which vertex behavior can be characterized by a projection onto a low-dimensional space. This approach is fruitful both in constructing analytically tractable models, and in creating a framework for comparison across vertices. Positions in a low-dimensional latent space can be used to parameterize the probability of connection between two vertices, and to define a relative measure of vertex similarity. Hoff et al. (2002) propose a class of latent space models in which the probability of connection between two vertices depends on their respective latent positions through a Euclidian distance function. Vertices with neighboring latent positions are more likely to be connected than those with distant positions. After conditioning on latent vertex positions, the edges are independent of one another.

We use the inner product between latent positions to define the probability of connection. This “random dot product model” for undirected connections is described by Scheinerman and Tucker (2010). The properties of this model with respect to random graph properties such as clustering, diameter and degree distributions are explored in Young and Scheinerman (2007). In a related model, Hoff (2009) describes the probability of directed connections using a multiplicative latent factor model in which each vertex is endowed with unobserved sender and receiver characteristics. Variability in connectivity which cannot be explained by observed attributes is modeled using random effects that are products of vector-valued sender and receiver characteristics.

In latent position models, the similarity measure induced by distances in latent space can be used in graph partitioning. Handcock and Raftery (2007) apply a model-based clustering algorithm to latent positions in a Euclidean space to infer a latent cluster structure over vertices. Latent positions are assumed to be drawn from a Gaussian mixture, and the probability of connection varies with the pairwise distance between latent positions. The latent mixture model induces clustering in the observed network.

We are interested in estimating partitions that identify dissimilarity in connectivity behavior across time as well as across vertices. While most network analysis assumes a static topology, dynamic network models have been proposed in contexts where data on the temporal location of connections is available. Dynamic mixed-membership stochastic blockmodels (Xing et al., 2010) are an extension of the mixed-membership stochastic blockmodel (Airoldi et al., 2008) in which the vectors identifying block membership (for each
node) can change over time, such that each actor may take on various roles in a network at different times. Latent space models have also been generalized to include dynamic positions, (Sarkar and Moore, 2005; Westveld and Hoff, 2010; Xu and Zheng, 2009). We are particularly interested in dynamic position models which can be analyzed to detect change points at which there is a shift in network behavior.

Also of interest are networks in which observed edges contain additional information on the type of connection between two vertices. We develop a model for the “attributed edge” case in which each edge has a categorical attribute contained in a finite set, as well as for the standard unattributed edge case. These attributes may be observed directly or, as in the case of the Enron email data, assigned by a classification procedure (Priebe et al., 2010). For data with attributed edges, we use the additional information to construct a more complex description of vertex behavior, in which vertex behavior varies with respect to the types as of connections they are associated with as well as probabilities of connection.

Lee and Priebe (2011) propose a dynamic dot product model for attributed graphs in which each vertex is associated with a latent stochastic process. Using a mathematically tractable approximate model, they develop a test statistic for detecting changes in attributed multigraphs observed at discrete times. The changes of interest are shifts in the behavior of an unknown anomalous subgroup of vertices. We approach a similar problem using a related latent position model, and extend the detection problem to multiple change points and groups of vertices. Costa et al. (2007) investigate a related detection task in a different context, developing a scan statistic to detect spatio-temporal disease clusters. A spatial graph structure is used to define cylinders in space-time which are the bases for the scan statistic.

We propose a model in which each vertex occupies a position in a latent space, and these positions may shift over time. The presence and attributes of observable edges depend probabilistically on the latent positions of the associated vertices through the random dot product model. Through the inferred latent positions, we can identify dissimilarity in behavior across groups of vertices or through time. We introduce a mixture model in which latent positions are drawn from either a homogeneous or heterogeneous distribution with respect to vertices and time. The goal of the analysis is to detect change points at which there is an identifiable shift between heterogeneity and homogeneity (in either direction) in the network. We approach the possibility of multiple change points through an iterative partitioning procedure.

Network data are complex. We aim to construct a simple model which can be used to detect time-varying structure in a variety of datasets. If more subtle features of the network are of interest, an initial evaluation of time-dependent block structure may be useful in segmenting the data before fitting a more elaborate model. Practicality of implementation is an important issue in network modeling, and we have found that this model can be fit relatively quickly for moderately large datasets. As an application we analyze the Enron email
2 Model

We introduce a generative model for streaming network data in which a population of \( n \) vertices are observed continuously in time, with edges appearing at random times, possibly with categorical attributes. The density of edges over time is of interest, as is the density of edges across the population of vertices. The distribution of edges over time is assumed to follow a doubly stochastic Poisson process model. The generative model is as follows:

1. Generate a point process \((t_1, t_2, \ldots, t_{N+})\) on the interval \((0, T)\) according to a Poisson process with rate \( \lambda \).

2. For each \( j = 1, \ldots, N^+ \):

   (i) Randomly choose a pair \((u_j, v_j)\) such that \( u_j \neq v_j \) from the population of \( n \) vertices.

   (ii) For vertices \( u_j \) and \( v_j \), generate latent positions \( X_{u_j}(t_j) \) and \( X_{v_j}(t_j) \) from distributions \( F(\theta_{u_j}(t_j)) \) and \( F(\theta_{v_j}(t_j)) \) independently. The vertex-specific parameter processes \( \theta_v(t) \) are described in section 2.1.

   (iii) Unattributed edge case: generate a Bernoulli random variable \( z_j \) such that

   \[
P(z_j = 1) = X_{u_j}(t_j) \cdot X_{v_j}(t_j) = \sum_{k=1}^{K} x_{u_j}^{(k)}(t_j)x_{v_j}^{(k)}(t_j).
   \] (1)

   If \( z_j = 1 \), draw an edge between \( u_j \) and \( v_j \).

   Attributed edge case: Draw an element \( k_j \) from the set \( 0, 1, 2, \ldots, K \) such that

   \[
P(k_j = k) = \begin{cases} x_{u_j}^{(k)}(t_j)x_{v_j}^{(k)}(t_j) & \text{if } k > 0, \\ 1 - \sum_{k=1}^{K} x_{u_j}^{(k)}(t_j)x_{v_j}^{(k)}(t_j) & \text{if } k = 0, \end{cases}
   \] (2)

   where \( x_{u_j}^{(k)}(t_j) \) and \( x_{v_j}^{(k)}(t_j) \) are the \( k \)th elements of vectors \( X_{u_j}(t_j) \) and \( X_{v_j}(t_j) \), respectively. If \( k_j > 0 \), draw an edge of attribute \( k_j \) between \( u_j \) and \( v_j \).

The underlying poisson process \((t_1, t_2, \ldots, t_{N^+}) \in (0, T)\) creates edge opportunities uniformly over the graph with exponentially distributed inter-arrival times. For simplicity, we consider data observed over an fixed interval \((0, T)\), but a related model for ongoing data acquisition could be constructed in which each new edge
opportunity occurs at an exponentially distributed interval after the last. Each pair of vertices has the same expected number of edge opportunities, and the expected number of edges opportunities does not change over time. The edge opportunity process is unobserved; what we observe are the realized edges, a filtered version of the edge opportunity process. Through this filtering, we can model inhomogeneities in number and attributes of edges over time and over the graph. A schematic for this two-level process is given in Figure 1.

Figure 1: An underlying unobserved Poisson process (top) generates opportunities for edges between pairs of vertices. The observed edge process (bottom) is a filtered version of the edge opportunity process in which edges may have categorical attributes.

The probability an edge opportunity at time $t_j$ will be realized depends on the latent processes associated with vertices $u_j$ and $v_j$. At each time $t$, $u_j$ and $v_j$ occupy positions $X_{u_j}(t)$ and $X_{v_j}(t)$ in a latent space. We define the latent space $S$ to be the subset of $\mathbb{R}^K$ bounded by the $K$-dimensional simplex:

$$S = \{ x \in \mathbb{R}^K_+ : \sum_{k=1}^K x_k \leq 1 \}. \quad (3)$$

The position of vertex $v$ in the latent space at time $t$, $X_v(t)$, is generated from the distribution $F(\theta_v(t))$. In what follows, we take $F$ to be the $(K + 1)$-dimensional Dirichlet distribution. To allow for flexible modeling of edge probabilities, we take $X_v(t)$ to be the first $K$ components of a $(K + 1)$-dimensional Dirichlet random variable, i.e. if $Y = (y_1, \ldots, y_{K+1}) \sim \text{Dir}(\theta_v(t))$, $X_v(t) = (y_1, \ldots, y_K)$. Note that since Dirichlet random variables must sum to one, the $(K + 1)^{th}$ component is fixed given the first $K$.

The relative locations of the two vertices in latent space determine the probability that edges between them will be realized. The probability an edge between vertices $u$ and $v$ at time will be realized at time $t$
depends on $X_u(t)$ and $X_v(t)$ through the dot product:

$$P(z_j = 1|X_{u_j}(t_j), X_{v_j}(t_j)) = P(k_j > 0|X_{u_j}(t_j), X_{v_j}(t_j)) = X_{u_j}(t_j) \cdot X_{v_j}(t_j) = \sum_{k=1}^{K} x^{(k)}_{u_j}(t_j) x^{(k)}_{v_j}(t_j).$$  \hspace{1cm} (4)$$

The dot product captures inhomogeneities in the probability of connection across different vertex pairs. In particular, two types of inhomogeneities can be easily expressed through the dot product: differences in overall connectivity for an individual vertex, and clustering, or the tendency for some groups of vertices to be more likely to be connected with one another. The former can be expressed through variations in the magnitude of the latent position vector. As $||X_v(t)||$ approaches 0 (i.e. the latent position approaches the origin), the dot product $X_v(t) \cdot X_u(t)$ will approach 0 for any other vertex $u$. Similarly, as $||X_v(t)||$ approaches 1, the probability of realized edges between $v$ and all other vertices will increase. A low probability of connection can also be expressed through a latent position vector which is nearly orthonormal.

The second type of inhomogeneity, clustering, can be expressed through the angles between vertices. For vertex pairs with a large angle between them (different directions in the latent space), the dot product will be small, even if the overall magnitude of each vector is large. When vertex pairs point in the same direction in the latent space, they will be more likely to communicate.

Using the dot product model, clusters in the latent space will give rise to clusters in the graph, as nearby vertices, with respect to angle from the origin, are more likely to communicate with one another. We are therefore interested in analyzing the structure of the graph in terms of distributions of latent positions.

The generative process above produces a collection of events denoted $e^+$ in the unattributed edge case and $a^+$ in the attributed edge case. Each event has a time stamp $t_j$, a vertex pair $(u_j, v_j)$, and either an edge indicator variable $z_j$ or an attribute $k_j$:

$$e^+ = \{(t_j, u_j, v_j, z_j); j = 1 \ldots N^+\} = (T^+, U^+, V^+, Z^+)$$ \hspace{1cm} (5)

and

$$a^+ = \{(t_j, u_j, v_j, k_j); j = 1 \ldots N^+\} = (T^+, U^+, V^+, K^+)$$ \hspace{1cm} (6)

where $T^+ = \{t_j, j = 1 \ldots N^+\}$, $U^+ = \{u_j, j = 1 \ldots N^+\}$, $V^+ = \{v_j, j = 1 \ldots N^+\}$, $Z^+ = \{z_j, j = 1 \ldots N^+\}$, and $K^+ = \{k_j, j = 1 \ldots N^+\}$. The collections $e^+$ and $a^+$ contain events for which $z_j = 0$ or $k_j = 0$. Since no edges are created in these events, we do not observe them. We observe $e$ or $a$, subsets of $e^+$ and $a^+$ for
which edges are realized:

\[ e = \{ (t_j, u_j, v_j) : z_j = 1 \} = \{ (t_i, u_i, v_i) : i = 1 \ldots N \} = (T, U, V) \]  

\[ a = \{ (t_j, u_j, v_j, k_j) : k_j > 0 \} = \{ (t_i, u_i, v_i, k_i) : i = 1 \ldots N \} = (T, U, V, K) . \]  

As an example, consider \( e \) to be a record of email messages, each of which contains a sender/receiver pair and time label, and \( a \) to be a similar collection which also contains a categorical email topic attribute.

Using the latent positions model, we are interested in doing inference on \( \theta_1, \ldots, \theta_n \), the parameters describing the distributions of the latent positions. If \( \theta_v = \theta \) for \( v = 1 \ldots n \), each vertex pair will have the same expected number of edges. In the general case, if \( \theta_u = \theta_v \) for a particular pair \( (u, v) \) vertices \( u \) and \( v \) may be more likely to communicate, depending on their overall probability of edge creation. Given the latent positions, the probability of the complete collection of events can be factored as

\[
P(e^+ | X, \lambda) = P(U^+, V^+ | N^+) P(Z^+ | X, N^+) P(T^+, N^+ | \lambda) \]

\[
= C_{u,v,t} \prod_{j: z_j = 1} X_{u_j} (\tau_j) \cdot X_{v_j} (\tau_j) \prod_{j: z_j = 0} (1 - X_{u_j} (\tau_j) \cdot X_{v_j} (\tau_j)) \frac{(\lambda T)^{N^+} e^{-\lambda T}}{N^+!} ,
\]

where \( C_{u,v,t} \) is a constant which does not depend on \( X \) or \( \theta \). The first product term describes the \( N \) elements of \( e \), ie the observed edges. The second term describes the subset of elements of \( e^+ \) which are unobserved, and therefore not in \( e \). The number of edge opportunities \( N^+ \), and the latent vertex positions at unobserved events appear in the likelihood, but are unobserved. If we assume the Poisson process parameter \( \lambda \) is given, as will be discussed in section 3, we can work with the expectation of the likelihood \( l(\theta, e^+) \) given the observed data \( e \):

\[
E_{e^+}(l(\theta, e^+) | e) = C_{u,v,t} \sum_{N^+ = 0}^{\infty} \prod_{i=1}^{N} X_{u_i} (t_i) \cdot X_{v_i} (t_i) (1 - E(X_u \cdot X_v | \theta))^{(N^+ - N)} \frac{(\lambda T)^{N^+} e^{-\lambda T}}{N^+!}
\]

\[
= C_{u,v,t} \prod_{i=1}^{N} X_{u_i} (\tau_i) \cdot X_{v_i} (\tau_i) (1 - E(X_u \cdot X_v | \theta))^{-N} e^{-(\lambda T) E(X_u \cdot X_v | \theta)} ,
\]

where the second inequality is simply an identity of the exponential function. This allows us to fit the model based on the observed data \( e \) or \( a \).
2.1 Change Points

We are interested in detecting changes in the behavior of the network, particularly changes in which a subset of vertices alter their behavior as a group. Where changes exist, we seek to partition the network over vertices and over time, detecting time-dependent inhomogeneities. Through an iterative procedure, it will ultimately be possible to detect multiple time points at which the network changes behavior, and multiple subpopulations of vertices exhibiting distinct behavior. As an initial step, we define a partition to identify unusual behavior in one subgroup of vertices over a single unknown time interval. A schematic is given in Figure 2.

Recall that $X_v(t)$, the latent position of vertex $v$ at time $t$ is drawn from a distribution $F(\theta_v(t))$. Changes in behavior for vertex $v$ through time are modeled through changes in $\theta_v$. For the Dirichlet distribution, this is $\alpha = (\alpha_1, \ldots, \alpha_{K+1})$, which specifies both the center and spread of the latent positions. Our initial hypotheses are

\begin{align}
H_0 : \theta_v(t) &= \theta_0, v \in \{1, \ldots, n\}, t \in (0, T) \\
H_1 : \theta_v(t) &= \theta_0, v \in \{1, \ldots, n\}, t \in (0, \tau_1) \cup (\tau_2, T), \\
\theta_v(t) &= \theta_0, v \in \{1, \ldots, n\} \setminus \{v_1, \ldots, v_m\}, t \in (\tau_1, \tau_2) \\
\theta_v(t) &= \theta_1, v \in \{v_1, \ldots, v_m\}, t \in (\tau_1, \tau_2)
\end{align}

for some unknown $\theta_0, \theta_1, m, \{v_1, \ldots, v_m\}, \tau_1$ and $\tau_2$. We approach the detection of multiple change points and multiple sub-populations of vertices in a hierarchical manner. An initial partition produces estimates of $(v_1, \ldots, v_m)$ and $(\tau_1, \tau_2)$ from the alternative model in (13) using the EM algorithm, as described in the next section. If the initial partition is shown to describe the data significantly better than the null model, as measured by the test statistic $M_n$ defined below, the model is fit for further partitions, which are then tested for significance.

We seek to identify the number of partitions which optimally describe the data by defining a stopping rule to decide when further partitions are unnecessary. The hierarchical approach is used rather than simultaneously estimating multiple partitions of over vertices and time because the latter is computationally impractical.

At each partitioning stage, the significance of the current partition can be phrased in terms of the number of components in a mixture distribution. If there are no distinct subgroups in a given set of vertices, and behavior does not change over a given time interval, the latent positions can be modeled using one component distribution, i.e. $X_v(t) \sim F(\theta), v \in \{1, \ldots, n\}, t \in (0, T)$. If the distribution of the $X_v(t)$ differs over time
or over vertices, then a mixture model with two or more components describes the distribution of the latent positions.

Determining the correct number of components in a mixture distribution is a difficult and well-studied problem. In this context, the likelihood ratio has been shown (Chernoff and Lander, 1995; Sen and Ghosh, 1985) to be badly behaved, without a simple limiting distribution. Chen et al. (2004) have developed a modified version of the likelihood ratio test, employing penalty functions to control the behavior of the test statistic near the boundary of the parameter space. Their test statistic is

\[
M_n = 2\{l(\hat{\theta}_0, \hat{\theta}_1) - l(\hat{\theta}) + C \log(4\gamma(1 - \gamma))\},
\]  

where \(\gamma\) is a mixing probability and \(C\) is a tuning parameter. We use \(M_n\) to assess the significance of each new partition.

Under certain regularity conditions, this test statistic has been shown asymptotically to have a mixture
of $\chi^2$ as its limiting null distribution:

$$\begin{align*}
M_n & \sim \frac{1}{2} \chi^2_0 + \frac{1}{2} \chi^2_1. 
\end{align*}$$

(15)

We compare this asymptotic approximation to a simulation-based threshold in section 4.

3 Model Fitting

In order to fit the model under heterogeneity and to compute the value of the test statistic $M_n$, we must compute maximum likelihood estimates of the model parameters $\theta$, $\theta_1$ and $\theta_0$. We do so using a stochastic conditional EM algorithm (Meng and Rubin, 1993).

The E-step at the $(j+1)$th iteration of the algorithm consists of computing a stochastic approximation to the conditional expectation of $(\tau_1, \tau_2)$ given the current parameter estimates $\theta_0^j$, $\theta_1^j$, and $\{v_1, \ldots, v_m\}^j$, the current estimates of $\{v_1, \ldots, v_m\}$. The approximate expectation is computed by simulating a large number (5,000, in our case) of candidate intervals $\{(t_{1,z}, t_{2,z}) : z = 1, \ldots, 5000\}$ and computing $p((t_{1,z}, t_{2,z})|\{v_1, \ldots, v_m\}^j, \theta_0^j, \theta_1^j)$ for each interval. The new estimate $\{v_1, \ldots, v_m\}^{j+1}$ then computed using the updated estimate of the interval $(\tau_1, \tau_2)$.

The M-step requires finding the updated maximum likelihood estimates $\theta_0^{j+1}$ and $\theta_1^{j+1}$ given the current estimates of $t_1, t_2$ and $\{v_1, \ldots, v_m\}$. As in other latent position models (Handcock and Raftery, 2007; Gormley and Murphy, 2010), the likelihood in the unattributed edge dot product model is invariant to reflections and rotations of the latent positions (although it is not invariant to translations.) The likelihood is convex with respect to the dot product of the latent positions, but not necessarily with respect to the latent positions themselves. In the attributed edge case, the likelihood is not invariant to rotations and reflections, because the directions in the latent space have meaning with respect to the probabilities of various edge attributes. Handcock and Raftery (2007) resolve the non-identifiability in the latent distance model by post-processing the MCMC output to find the configuration of latent positions which is optimal in terms of Bayes risk.

In order to resolve the non-identifiability for the unattributed random dot product model, we define an additional condition based on minimizing the squared difference between a modified adjacency matrix and a matrix consisting of the expected number of edges between each pair.

Let $\bar{X}_v$ be the average latent position of vertex $v$ over a given time interval $(t_1, t_2)$. The collection of latent positions over over all vertices is the $K \times n$ matrix $\bar{X} = (\bar{X}_1, \ldots, \bar{X}_n)$. The expected number of edges over $(t_1, t_2)$ between vertices $u$ and $v$ is the $(u,v)^{th}$ element of the matrix $B\bar{X}^T\bar{X}$, where

$$B = \frac{\lambda(t_2 - t_1)}{\binom{n}{2}}.$$  

(16)
Let $A$ be the multiadjacency matrix consisting of the number of edges between each vertex pair over the interval $(t_1, t_2)$. We would like to compute an initial estimate of the latent positions by minimizing the difference between the observed and expected number of edges for each pair, given the latent positions. We cannot directly compare $A$ and $B\mathbf{X}^T\mathbf{X}$, however, because the diagonal entries in the two matrices are not comparable. We instead use $\tilde{A}$, a modification of $A$ with an augmented diagonal as described in Scheinerman and Tucker (2010). Using singular value decomposition, we can easily find the latent positions $\mathbf{X}$ which minimize

$$\|B\mathbf{X}^T\mathbf{X} - \tilde{A}\|^2,$$

the Frobenious norm of the difference between expected and observed edges. This least squares estimate of the latent positions is used as a starting point to a hill-climbing procedure to find the local maximum likelihood estimates of $\theta, \theta_0$ and $\theta_1$ which are closest to the minimizer of the least-squares condition. Note that in the attributed edge case, this additional condition is unnecessary.

Based on simulation results, the sensitivity of the algorithm to initial conditions seems to vary with the number of change points and of distinct subgroups of vertices, with more components increasing the sensitivity. A procedure to define initial conditions is described below.

Initial values of $(\tau_1, \tau_2)$ and $\{v_1, \ldots, v_m\}$ are produced by dividing up the entire observed time interval $(0, T)$ into segments, finding candidate values of $\{v_1, \ldots, v_m\}$ for each segment, and then assessing which combination of time interval and candidate $v_1 \ldots v_m$ produces the highest likelihood under the 2-component mixture model.

The time interval $(0, T)$ is divided into $r$ equally sized segments, producing non-overlapping time intervals $(0, T/r), (T/r, 2T/r), \ldots, ((r-1)T/r, T)$. In the unattributed edge case, the least squares estimates of the latent positions over each interval $\mathbf{X}^1, \ldots, \mathbf{X}^r$ are computed from the data in each time interval using the SVD procedure described above. This group of $n$ latent positions are then clustered into two groups using k-means, producing an estimate of $\{v_1, \ldots, v_m\}$ for each interval. In the attributed edge case, using the data from each interval, a vector $(p_1, \ldots, p_K)$ is created for each vertex, where $p_k$ is the proportion of edges with attribute $k$. These vectors are then clustered using k-means, producing estimates of $\{v_1, \ldots, v_m\}$ for each interval.

In our model, there is non-identifiability between the Poisson process rate $\lambda$, and the parameters of the latent position distribution, $\theta$ which can be easily resolved. The expected number edges created in a given time interval increases with both $\lambda$ and the expected value of $\mathbf{X}_u \cdot \mathbf{X}_v$; increasing $\lambda$ produces more edge opportunities, and increasing $E(\mathbf{X}_u \cdot \mathbf{X}_v)$ increases the probability of an edge at each opportunity. We
approach this non-identifiability by fixing $\lambda$ to be 1.5 times the maximum number of edges observed in any time unit (weeks, in our application). Thus the expected number of edge opportunities (not necessarily edges) in a unit time interval is 50% greater than the maximum observed number.

4 Simulations

We simulate network datasets $a = \{a_1, \ldots, a_N\}$ using the generative model in section 2 to test the performance of the partitioning algorithm. At random times $t_1, \ldots, t_{N^+}$, pairs of latent positions

$\{(X_{u_1}, X_{v_1}), \ldots, (X_{u_{N^+}}, X_{v_{N^+}})\}$

are drawn from the $(K + 1)$ - dimensional Dirichlet distribution. Change points are inserted at times $\tau_1$ and $\tau_2$, between which a subset of vertices $\{v_1, \ldots v_m\} \in \{1, \ldots, n\}$ has latent positions drawn from a Dirichlet distribution with parameter $\alpha_2$ rather than $\alpha_1$.

The ability to detect time periods with heterogeneous subgroups increases with $\phi$, the angle between $\alpha_1$ and $\alpha_2$, $N = N/(n(n - 1)/2)$, the average number of edges per pair, $\tau_2 - \tau_1$, the length of the anomalous interval, and with $m$, the size of the anomalous subgroup (up to $m = n/2$). The average number of edges per pair is a function of the length of the total time interval, the number of vertices $n$, the Poisson process rate $\lambda$ and the values of $\alpha_1$ and $\alpha_2$. Performance can be measured by the correct identification of vertices as anomalous and non-anomalous, as represented by sensitivity (correct identification of anomalous vertices) and specificity (correct identification of non-anomalous vertices), and by $|(\tau_1 - \hat{\tau}_1) + (\tau_2 - \hat{\tau}_2)|/2$, the average error in estimates of the change points $\tau_1$ and $\tau_2$, when they are detected. Because of the invariance discussed in section 3, recovery of $\alpha_1$ and $\alpha_2$ is not of interest.

Simulated data from networks of $n = 50$ vertices with anomalous subsets of size $m = 10$, $K = 4$ and various values of $\lambda$ and $\phi$ are shown in Figure 3. A total time interval of length 100 is used with Poisson process rates 500, 350, 250, 150 and 75. The length of the anomalous interval is 40% of the total observed interval. In simulated data with moderate angles between the centroids of the latent position distribution, $(\phi = 76^\circ)$ find good performance in datasets with $N \geq 2$. For larger angles between cluster centroids, good performance can be achieved for $N = 1$.

5 Application to Enron Data

As an application, we analyze the Enron email corpus (Cohen, 2009). Following its investigation of Enron, the Federal Energy Regulatory Commission publicly released emails sent between 1998 and 2002 among 184 email addresses belonging to roughly 150 Enron employees, including high-level executives. We use version of the dataset which has been processed to correct some integrity problems (Priebe et al., 2005). Previous
analyses (Fu et al., 2009; Diesner et al., 2005; Priebe et al., 2005) have found these data have been exhibit heterogeneity over time, and to exhibit structure across vertices. We use Michael Berry’s topic classifications for the 2001 data (Berry et al., 2007), which assigns one of 32 topics to most messages. We restrict our attention to data on emails which have been assigned one of the 4 most popular email topics. The average number of connections between pairs is 1.1. Although this is a subset of the entire email corpus, and likely contains some residual integrity problems, we find it to be an informative application for the partitioning algorithm.

Fu et al. (2009) have analyzed these data using the dynamic mixed-membership stochastic block model. Using information on the direction of the connections, the blockmodel provides a useful categorization of email behavior. Fu et al. identify, for example, groups of employees who tend to send and receive emails among themselves as cliques, groups that receive emails but rarely send them, etc. Their analysis of the changing membership vectors associated with individual email addresses reveals changes in the frequency of emails in late 2001 as the company was going bankrupt, and a decrease in emails sent by high-level executives over the same period.

Analysis of the Enron data using the attributed edge dot product model reveals multiple change points. In the attributed model, the dominant change point detected (that which is identified in the first partition) is in mid-August 2001, after which the detected heterogeneity persists until mid-December 2001. This time interval is significant with respect to major events related to the collapse of the company. CEO Jeffrey Skilling resigned on August 15, 2001 and over the following weeks the Enron stock price decline sharply and concerns over the company’s accounting practices became public. Enron declared bankruptcy in December 2001. The overall message rate increased somewhat in this period with respect to the January - August period (423 vs 297 messages per week, on average.) However, the graph partitioning algorithm identifies
a subgroup of 36 email addresses for which communication decreases significantly. There is also a shift in
distribution of message subjects (edge attributes). A second anomalous time period is detected between mid
April and late July. Here, a group of 88 email addresses show a greatly increased communication rate and
a change in email attributes. These results are consistent with the “chatter anomalies” detected in Priebe
et al. (2005) using scan statistics on an unattributed version of the same data.

6 Discussion

We have presented a novel algorithm and associated model for time-dependent network partitioning capable
of discovering dynamic network structure. The filtered Poisson process model is a flexible framework for
describing inhomogeneities in the rate of edge creation over time and over the graph. The detection of mean-
ingful partitions is approached by fitting a mixture model and comparing it to a model with a homogeneous
distribution of latent positions.

The model fitting algorithm simultaneously estimates latent positions and cluster structure, which has
been shown in related contexts (Handcock and Raftery, 2007) to give better results than a two-stage procedure
(position estimation followed by clustering.) We approach the detection of multiple change points and multiple
vertex subpopulations using by iterating the partitioning algorithm. Iterative partitioning may not perform
as well as a simultaneous estimation of multiple components, but we find the latter to be computationally
infeasible. In order to test each partition for significance, we use the modified likelihood ratio test statistic
proposed by Chen et al. (2004) Fitting the random dot product mixture model presents some computational
challenges. Like the latent-distance model (Hoff et al., 2002) , the likelihood surface described by the random
dot product model is non-convex, and the latent positions are non-unique. We impose an additional condition
based on the least-squares optimal latent positions in order to define a unique maximum likelihood solution.
We find this to be a fast and reliable solution to the non-uniqueness problem, although the precise relationship
between the least-squares optimal solution estimates and the maxima of the likelihood surface is unclear.

Joseph and Wolfson (1993) discuss the consistency of maximum likelihood estimates in the “multi-path”
change point problem, in which a collection of time series are observed, each potentially containing a unique
change point. The locations of change points in individual time series are assumed to be independent of
one another. They find that under certain regularity conditions, the maximum likelihood estimates of the
change point locations and of the pre-and post-change parameters are consistent estimators. In our context,
each vertex has an associated time course with a possible change point. In contrast to the case described by
Joseph and Wolfson (1993), the change points of individual vertex time courses are not independent of one
another; this non-independence makes analytic results on the asymptotic behavior of the parameter estimates difficult.

The proposed model for edge creation has a simple functional form, and while it is able to capture features frequently observed in real network data, it may not incorporate all available information. Several extensions of the model are possible. In particular, we have allowed for the parameters distributions of latent positions to change in time, but have not incorporated temporal dependence in the latent processes associated with each vertex. Lee and Priebe (2011) have analyzed the performance of inferences based on first and second order approximations to time-dependent latent dot product models. It may be possible to (for example) fit an autoregressive component to the proposed latent process model by updating latent position estimates with each new edge event, conditional on the estimated positions at the last edge event. For data in which the overall rate of edges is known to change over time, the model could be modified in a simple manner by specifying a non-constant Poisson process rate $\lambda$. Under such conditions, the partitioning algorithm would detect changes in network behavior with respect to the overall dynamic message rate. In some applications, for example the Enron email data, vertex-level covariates and directional information are also available. A model which includes covariate information could be used in place of the simple dot product model. In general, the filtered Poisson model and change-point framework described above may be useful in combination with other latent positions models for network data.

References


Diesner, J., Frantz, T., and Carley, K. (2005), “Communication Networks from the Enron Email Corpus It’s Always About the People. Enron is no Different,” *Computational and Mathematical Organization Theory*, 11, 201–228, 10.1007/s10588-005-5377-0.


