Discovering Long Range Properties of Social Networks with Multi-Valued Time-Inhomogeneous Models

Danny Wyatt
Dept. of Computer Science and Engineering
University of Washington
danny@cs.washington.edu

Tanzeem Choudhury
Dept. of Computer Science
Dartmouth College
tanzeem.choudhury@dartmouth.edu

Jeff Bilmes
Dept. of Electrical Engineering
University of Washington
bilmes@ee.washington.edu

Abstract

The current methods used to mine and analyze temporal social network data make two assumptions: all edges have the same strength, and all parameters are time-homogeneous. We show that those assumptions may not hold for social networks and propose an alternative model with two novel aspects: (1) the modeling of edges as multi-valued variables that can change in intensity, and (2) the use of a curved exponential family framework to capture time-inhomogeneous properties while retaining a parsimonious and interpretable model. We show that our model outperforms traditional models on two real-world social network data sets.

Introduction

It is becoming increasingly easy to collect data that captures the real-world social interactions of entire groups of people (Wren et al. 2007; Wyatt, Choudhury, and Kautz 2007; Eagle, Pentland, and Lazer 2009). These new data sets provide opportunities to study the social networks of people as they are observed "in the wild," instead of as they are reported in surveys. And while it is tempting to turn to traditional methods of social network analysis (SNA), those methods are often inadequate for behavioral data. Most existing SNA techniques apply only to static, binary data (Wasserman and Faust 1994). Social networks derived from behavioral data will almost always be observed through time (not in one static snapshot) and will often have finer grained observations about interactions than simple binary indicators. New techniques are needed that can take into account multiple tie intensities and the dynamics of a network as it evolves in time.

Additionally, the few existing works on temporal network models (Robins and Pattison 2001; Guo et al. 2007; Hanneke, Fu, and Xing 2009) apply a traditional time-homogeneous approach that assumes that the underlying properties of a network—e.g. density, transitivity—remain constant throughout time. We hypothesize that many networks will exhibit significant time-*inhomogeneity* and that one of the key uses of a temporal network model is to discover the pattern of that inhomogeneity. For example, networks may change their densities or path length distributions over time. If we want to maximize the spread of information or minimize the spread of contagion it would be useful to know when there will periods of high or low connectivity. A time-inhomogeneous model can capture such phenomena.

The two key contributions of this paper are (1) a new method for time-inhomogeneous modeling of dynamic social networks that allows the underlying network properties to vary over time while still requiring only a small, fixed number of parameters; and (2) the extension of existing exponential random graph models (ERGMs) to account for edges of multiple intensities. We believe both of these techniques are of central importance to the modeling, mining, and prediction of automatically collected real-world social network data. We demonstrate that our model produces better fits on two real-world social networks. We also show that the model discovers interpretable sociological properties of the populations being modeled. To the best of our knowledge, this is the first implementation of dynamic, multi-valued ERGMs as well as the first application of them to behavioral data.

Background and Related Work

In recent decades, a new class of methods for SNA known as exponential random graph models has been developed (Frank and Strauss 1986; Wasserman and Pattison 1996; Robins et al. 2007). ERGMs depart from traditional descriptive models by considering a social network as a realization of a set of random variables. By considering a distribution over networks and network statistics (instead of considering just a single observed value), ERGMs can exploit and understand any underlying uncertainty in the data.

Given an observed network, an ERGM estimates the parameters of an exponential family model that describes the joint distribution of variables used to model that network—typically one variable per potential edge. The probability distribution takes the form

$$p(\mathcal{Y} = \mathbf{y}) = \frac{1}{Z_{\boldsymbol{\eta}}} e^{\langle \boldsymbol{\eta},\, \mathbf{f}(\mathbf{y}) \rangle} \tag{1}$$

$\mathcal{Y}$ are random variables representing edges in the graph, with a specific realization $\mathbf{y}$ in which edge (i, j) takes value y_{ij}. $\mathbf{f}$ are feature functions defined on $\mathbf{y}$, and $\boldsymbol{\eta}$ is a vector of weights to be learned. $\langle \boldsymbol{\eta},\, \mathbf{f}(\mathbf{y}) \rangle$ denotes the inner product $\boldsymbol{\eta}^{\mathsf{T}} \mathbf{f}(\mathbf{y}) = \sum_i \eta_i f_i(\mathbf{y})$ and the model thus has the standard

form of a log-linear combination of parameters and features. $Z_{\boldsymbol{\eta}} = \sum_{\mathbf{y}} e^{\langle \boldsymbol{\eta}, \mathbf{f}(\mathbf{y}) \rangle}$ is the usual normalizing constant.

The features are deterministic functions (or statistics) of the network. Typical features are counts of subgraph occurrences, such as the number of triangles or even simply the number of edges. The strength of these models lies in their ability to capture structural dependencies in a probabilistic manner. Individual properties of the network can then be interpreted in terms of how changes to them affects the network's probability.

When represented as an undirected graphical model, an ERGM has one node in the graphical model for each dyad in the social network. Edges in the graphical model exist between variables that occur together in the same feature function. ERGMs are typically specified by the set of feature functions employed, and not by an explicit structure for the graphical model.

Advances in the ability to fit models in the form of (1) using MCMC (Geyer and Thompson 1992) have also revealed that older ERGM specifications tended to be *degenerate* (Handcock 2003; Snijders 2002). Models are considered degenerate if only a small set of parameter values lead to plausible networks. Slight changes in the parameter values of a degenerate model can cause it to put all of its probability on almost entire empty or entirely complete networks.

To ameliorate degeneracy, Hunter and Handcock (2006) proposed using a curved exponential family model. A curved exponential family places constraints on $\boldsymbol{\eta}$ that restrict its possible values to lie on a non-linear manifold. In that case, $\boldsymbol{\eta}$ is redefined as a non-linear function mapping a point $\boldsymbol{\theta}$ in q-dimensional space to a point $\boldsymbol{\eta}(\boldsymbol{\theta})$ in p-dimensional space, where $q < p$. The points $\boldsymbol{\theta} \in \Theta$ then define a q-dimensional curved manifold in p-dimensional space and thus models defined in a such a way are called curved exponential families (Efron 1978). The model thus takes the form

$$p(\mathbf{Y} = \mathbf{y}) = \frac{1}{Z_{\boldsymbol{\theta}}} e^{\langle \boldsymbol{\eta}(\boldsymbol{\theta}), \mathbf{f}(\mathbf{y}) \rangle} \tag{2}$$

And the gradient of the log-likelihood is

$$\frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}|\mathbf{y}) = \nabla \boldsymbol{\eta}(\boldsymbol{\theta})^{\mathsf{T}} \left(\mathbf{f}(\mathbf{y}) - \mathop{\mathrm{E}}_{\mathbf{y}} \left[\mathbf{f}(\mathbf{y}) \,\middle|\, \boldsymbol{\theta} \right] \right) \tag{3}$$

where $\nabla \boldsymbol{\eta}(\boldsymbol{\theta})$ is the $p \times q$ Jacobian of $\boldsymbol{\eta}(\boldsymbol{\theta})$. The presence of the Jacobian shows that the log-likelihood is no longer, in general, convex.

This new formulation, known as curved ERGMs (CERGMs) has led to better model fits than linear ERGMs (Hunter and Handcock 2006; Hunter, Goodreau, and Handcock 2008).

Features for CERGMs

The features used for CERGMs can lead to better model fits, but they also make possible more nuanced sociological interpretations. The simple subgraph counts used as features in linear ERGMs can lead to model degeneracy, but they often also do not fully capture the intuitions that motivated the features.

The models of (Hunter and Handcock 2006) allow entire histograms of statistics to be used as features while still requiring only a small number of parameters. For example, the traditional ERGM feature for capturing transitivity is the count of all triangles that appear in the network. A CERGM replaces that count with the edgewise shared partner (ESP) histogram of the network: a vector where component i counts the numbers of edges whose endpoints have exactly i shared partners.

Formally, that means that a span of entries in $\mathbf{f}$, call it $\mathbf{f}^g$, are the network's edgewise shared partner histogram. There is a corresponding portion of $\boldsymbol{\eta}$, $\boldsymbol{\eta}^g$, that assigns weights to the bins of $\mathbf{f}^g$. The form that $\boldsymbol{\eta}^g$ has taken in all existing CERGMs is

$$\eta^g_i(m, r) = m \left(e^r (1 - (1 - e^{-r})^i) \right) \tag{4}$$

Thus, while an n-node network has $n-2$ bins in the ESP histogram, there are only two free parameters controlling their weights: m, the usual multiplicative weight, and r, the rate at which the growth of m in i diminishes. Since that rate of diminishing increase is geometric, the above combination of features and constrained parameters is known as geometrically weighted edgewise shared partners (GWESP). GWESP models a richer notion of transitivity through its diminishing increase in weights. A simple feature that counts triangles (with a single weight) implies that adding a shared partner always increases log-probability by a constant amount no matter how many shared partners a pair already has.

Previous Multi-valued and Temporal ERGMs

Robins, Pattison, and Wasserman (1999) first proposed extending ERGMs to handle discrete multi-valued relations. Where edges can take one of v values, v new binary indicator variables are introduced per edge. New features are then defined for this expanded all-binary model. This transformation is motivated by the fact that a pseudo-likelihood fit (Strauss and Ikeda 1990) can be then be found using existing logistic regression software. Our model avoids the need for such a conversion by simply defining features over the multi-valued edge variables directly.

Robins and Pattison (2001) proposed the first temporal ERGM by defining the probability of a sequence of just two observed networks, $\mathbf{y}^1$ and $\mathbf{y}^2$, as

$$p(\mathbf{y}^1, \mathbf{y}^2) = p(\mathbf{y}^2) p(\mathbf{y}^2 | \mathbf{y}^1) \tag{5}$$

$$= \frac{1}{Z_{\boldsymbol{\eta}^1}} e^{\langle \boldsymbol{\eta}^1, \mathbf{f}^1(\mathbf{y}^1) \rangle} \times \frac{1}{Z_{\boldsymbol{\eta}^2, \mathbf{y}^1}} e^{\langle \boldsymbol{\eta}^2, \mathbf{f}^2(\mathbf{y}^2, \mathbf{y}^1) \rangle} \tag{6}$$

Robins and Pattison test this model on a two-step data set, fitting it using pseudo-likelihood. They explain that their model could be extended to longer sequences $\mathbf{y}^1, \ldots, \mathbf{y}^T$ if a time-homogeneity assumption is made.

Recently, Hanneke, Fu, and Xing (2009) applied the model of Robins and Pattison to longer network sequences

by making exactly such an assumption. In the model proposed by Hanneke, Fu, and Xing the probability of a sequence of T networks is

$$p(\mathbf{y}^1, \ldots, \mathbf{y}^T) = p(\mathbf{y}^1) \prod_{t=2}^{T} p(\mathbf{y}^t | \mathbf{y}^{t-1}) \tag{7}$$

$$= \frac{1}{Z_{\boldsymbol{\eta}_s}} e^{\langle \boldsymbol{\eta}_s, \mathbf{f}_s(\mathbf{y}^1) \rangle} \times \prod_{t=2}^{T} \frac{1}{Z_{\boldsymbol{\eta}, \mathbf{y}^{t-1}}} e^{\langle \boldsymbol{\eta}_s, \mathbf{f}_s(\mathbf{y}^t) \rangle + \langle \boldsymbol{\eta}_d, \mathbf{f}_d(\mathbf{y}^t, \mathbf{y}^{t-1}) \rangle} \tag{8}$$

In this model the features and their associated parameters have been divided into two sets: *static* features $\mathbf{f}_s$ that only consider variables within one timestep and *dynamic* features $\mathbf{f}_d$ that consider variables across two timesteps. The time-homogeneity assumption is expressed in the use of the same $\boldsymbol{\eta}_s$ and $\boldsymbol{\eta}_d$ for all timesteps.

Hanneke, Fu, and Xing also restrict their model to use only dyad-independent static features and dynamic features that can be decomposed as

$$\mathbf{f}_d(\mathbf{y}^t, \mathbf{y}^{t-1}) = \sum_{ij} f_{ij}(y_{ij}^t, \mathbf{y}^{t-1}) \tag{9}$$

They show that the restriction in (9) helps the model avoid degeneracy. Additionally, they show that using only dynamic features of the form in (9) in a distribution that conditions on the first observation (i.e. removing the $p(\mathbf{y}^1)$ factor in (7)) ensures that the (conditional) log-likelihood gradient can be computed exactly. Such conditioning may be problematic, though, since the initial state may not represent the network at equilibrium. Indeed, Hanneke, Fu, and Xing explain that they have to discard the first few observations from their data since they seem to be outliers when compared to later observations.

Multi-valued Time-Inhomogeneous Dynamic Exponential Random Graph Models

The fact that observations from one part of a sequence seem very different from observations from a later part of the same sequence motivates our model. We hypothesize that most social networks will display this sort of time-inhomogeneity[1] at the scales we would like to model them at. We use the curved exponential framework to define a model capable of capturing changes in the underlying properties of the network—not just changes in the values of variables—while still requiring only a fixed set of parameters. In addition to that, we redefine existing ERGM features so that they apply to multi-valued edges.

[1]Note that time-inhomogeneity is different from non-stationarity. A homogeneous process is one whose parameters do not change over time and it always has at least one stationary distribution. An inhomogeneous process can ensure that there is no stationary distribution and may therefore be more suitable for temporal processes whose properties evolve over time.

Multi-Valued CERGMs

Our model allows edges to take one of v discrete, ordinal values. These values represent the observed intensity of a social tie. Larger values indicate a stronger tie. To permit comparisons with binary-valued models, the values are scaled so that the smallest is 0 and the largest is 1. Many simple network statistics can be redefined for this model in a straightforward manner: the density of a network is the sum of its edge values; a node's degree is the sum of the values of the edges incident to that node.

More complicated features that involve subgraphs require defining the intensity of a subgraph. For that, we use the geometric mean of the edge values composing the subgraph. For example, a shared partner k for nodes i and j is defined to be a partner of intensity $(y_{ik}y_{jk})^{\frac{1}{2}}$, where y_{ij} represents the multi-valued edge between nodes i and j. The count of shared partners for a pair, SP_{ij} is the sum of these intensities:

$$SP_{ij} \triangleq \sum_k (y_{ik}y_{jk})^{\frac{1}{2}} \tag{10}$$

To model edgewise shared partners we take the product of an edge's value with its shared partner sum:

$$ESP_{ij} \triangleq y_{ij} SP_{ij} \tag{11}$$

Note that if $v = 2$ and all edge values are either 0 or 1, then our features are equivalent to the traditional CERGM features.

Time-Inhomogeneous Dynamic CERGMs

For a time-homogeneous model such as (8) with s static features and d dynamic features, a feature vector of length $s+d$ can be computed for each pair of adjacent timesteps. Time-homogeneity allows all of these vectors to be summed into a single vector (clearly also of length $s+d$) that summarizes the entire sequence. By doing that, (8) can be rewritten as

$$p(\mathbf{y}^1, \ldots, \mathbf{y}^T) = \frac{1}{Z_{\boldsymbol{\eta}_s}} e^{\langle \boldsymbol{\eta}_s, \mathbf{f}_s(\mathbf{y}^1) \rangle} \times \frac{1}{Z_T} \times e^{\langle \boldsymbol{\eta}_s, \sum_{t=1}^{T} \mathbf{f}_s(\mathbf{y}^t) \rangle + \langle \boldsymbol{\eta}_d, \sum_{t=2}^{T} \mathbf{f}_d(\mathbf{y}^t, \mathbf{y}^{t-1}) \rangle} \tag{12}$$

where $Z_T = \prod_{t=2}^{T} Z_{\boldsymbol{\eta}, \mathbf{y}^{t-1}}$.

For our time-inhomogeneous model we compute the same set of features for each timestep but keep their values separate and allow each timestep to have its own set of parameters:

$$p(\mathbf{y}^1, \ldots, \mathbf{y}^T) = \frac{1}{Z_{\boldsymbol{\eta}_s^1}} e^{\langle \boldsymbol{\eta}_s^1, \mathbf{f}_s(\mathbf{y}^1) \rangle} \times \prod_{t=2}^{T} \frac{1}{Z_{\boldsymbol{\eta}_s^t, \mathbf{y}^{t-1}}} e^{\langle \boldsymbol{\eta}_s^t, \mathbf{f}_s(\mathbf{y}^t) \rangle + \langle \boldsymbol{\eta}_d^t, \mathbf{f}_d(\mathbf{y}^t, \mathbf{y}^{t-1}) \rangle} \tag{13}$$

Thus the feature vector for the entire sequence is of length $T(s+d)$ and the feature output for time t begins at index $[(t-1)(s+d)+1]$ in $\mathbf{f}$. For example, consider a model that includes a single feature: network density. The density

of each $\mathbf{y}^t$ is computed and placed at index t in the feature vector. The resulting vector is the sequence of densities as the network evolves through time. Clearly, the longer the sequence gets, the longer its feature vector gets.

However, by leveraging the functional form of $\boldsymbol{\eta}$ in a curved exponential family we can keep the number of parameters fixed. And by choosing a flexible form for $\boldsymbol{\eta}$ we can smooth away short term variations in the data to discover long range patterns of change over time.

Note that not only does having separated features per timestep allow for time-inhomogeneity, it also allows—with properly defined transition features—for irregularly spaced observations. When observing a real-world social network it is likely that observations may not appear regularly.

Additionally, it is possible to model time-inhomogeneity through time-varying features (e.g. by defining separate $\mathbf{f}^t(\mathbf{y}^t)$ for each timestep). However, by using the parameters to model time-inhomogeneity it is possible to learn the form of that inhomogeneity from the data instead of having to specify it in advance through a fixed set of feature functions.

Features and Parameter Constraints

The models we employ in this paper use different combinations of three features: (i) the edge value histogram, (ii) network anti-stability, and (iii) GWESP.

The edge value histogram is the simple vector of counts of how many edges take each of the v discrete values. One value (the highest) is excluded to avoid having a linear dependency among the features.

Network anti-stability, $a(\mathbf{y}^{t'}, \mathbf{y}^t)$, is the amount that each edge changes its value between observations:

$$a(\mathbf{y}^{t'}, \mathbf{y}^t) \triangleq \sum_{ij} \frac{(y_{ij}^{t'} - y_{ij}^t)^2}{t' - t} \tag{14}$$

where $t' > t$ and there is no other observed timestep between t' and t (thus implying a Markov property). Note that t' need not be $t + 1$ (and frequently is not in our evaluations) and this feature is still capable of modeling irregularly spaced observations. Dividing by $t' - t$ makes (14) equivalent to modeling the change in an edge's value (when all other features are held constant) as a discrete time random walk with step sizes drawn from a Gaussian. The mean of that Gaussian is zero, and its variance will be inversely proportional to the negative of the weight learned for this feature.

GWESP is as it is defined in (11) with the parameter constraint defined in (4). It models the network's tendency towards transitivity.

All three of these features are strictly "local" in nature. This captures the intuition that social ties are formed through local decisions without access to global network properties.

The parameter constraints for the edge value histograms, the anti-stability sums, and the multiplicative weight m for GWESP are all constrained to follow a sigmoid with offset:

$$\eta_{f_k}^t(w_k, a_k, b_k, s_k) \triangleq w_k \left(\frac{1}{1 + e^{-(a_k + b_k t)}} + s_k \right) \tag{15}$$

This equation is analogous to (4): $\eta_{f_k}^t$ are the weights for some feature f_k and the weights change in a constrained way over time. Specifically, w_k is the ordinary multiplicative weight for feature k. That weight is scaled by the logistic with parameters a_k and b_k. Since the logistic will only take values between 0 and 1, the offset parameter s_k shifts it up or down, allowing it to cross zero. Features with a positive weight make the data more likely as they increase in value and those with a negative weight make the data less likely as they increase in value. If the learned sigmoid crosses zero at some time, it means that the model has found a point at which a feature has shifted between helpful and harmful for the network.

Note that what previously would have been one parameter, w_k, in a time-homogeneous model is now 4 parameters in our time-inhomogeneous model. That is the cost of the increased flexibility provided, but it is fixed: the number of parameters stays the same no matter how long the data sequence is.

Any number of functions could have been chosen to model time-inhomogeneity. We chose the sigmoid for 3 reasons. First, the networks we consider are observed within bounded "episodes" for their respective populations (one academic year, one senate session). We want to see if there is a shift from one underlying regime to another, e.g. from low transitivity to high. Second, the logistic has an asymptotic bound. With 4 parameters we could have used a degree 3 polynomial, but that would grow infinitely as time increased. An asymptotic function is more plausibly extended into the future. Third, while the logistic is defined for all real values of t, in our specification t will always be positive and will be effectively bounded by some maximum T. The a and b parameters allow the sigmoid to be shifted left and right, so it is free to only decrease or only increase. It can also stay constant if there is no time-inhomogeneity present in the data.

Evaluation

We test our model on two real-world social network data sets. First, a simple model applied to data from the U.S. Senate illustrates the basic advantages of a time-inhomogeneous approach. Then we apply a more complex model to a corpus of face-to-face conversations and use the model to discover basic properties of the conversational network.

In both data sets we quantize continuous edge values to v discrete values. All zero values are left at zero and all non-zero values are quantized to $v - 1$ discrete points using k-means. The quantized values are then normalized so that the maximum value is 1. We also experimented with equally-spaced and equally-weighted binning schemes but found that the non-uniform binning provided by k-means produced the best model fits. For the senate data, $v = 5$ and for the conversation data $v = 10$. (Initial experiments showed that the model was robust across larger values of v (Wyatt, Choudhury, and Bilmes 2009).)

To learn the parameters we first use pseudo-likelihood to find a starting point and then use Gibbs sampling to approximate the expectation in (3). Despite their non-convexity,

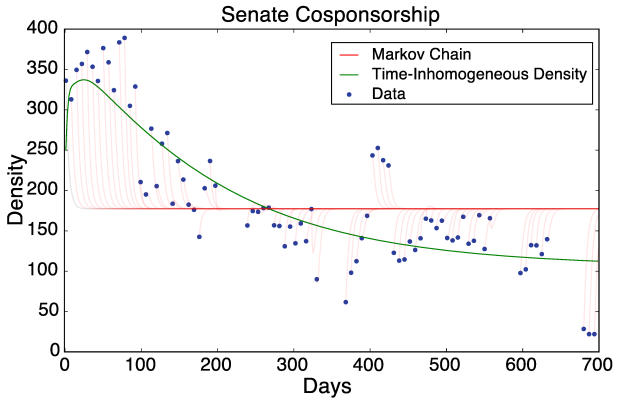

Figure 1: Density of senate networks with best fit values from time-homogeneous Markov chains and a time-inhomogeneous density model.

BFGS has been successfully used for learning CERGMs (Hunter et al. 2008) and we use it as well.

Senate Data

The senate data comes from (Fowler 2006) and is the same population from which Hanneke, Fu, and Xing (2009) discard observations that they considered outliers. The data captures the cosponsorship network of senators in the 108th United States Senate. When a bill (or resolution or amendment, all referred to as "bills" here) is proposed in the U.S. Senate it must be sponsored by one senator. Additional senators may sign on as co-sponsors of the bill any time before the senate votes on the bill.

We divide the senate data into sliding windows that are 28 calendar days long with 7 calendar day offsets. If the senate is not in session for more than 28 calendar days in a row we include no measurement for that period. After such a gap, the next window starts at the soonest date the senate is in session. From each window we build an undirected cosponsorship network by adding edges between two senators if one cosponsored the other's bill during that window. The strength of the edge is the number of bills cosponsored during the window, normalized by the number of days in session within the window (which adjusts for small variations due to e.g. 3 day weekends).

Figure 1 shows that the network's density is clearly time-varying. We fit the simplest of our models to this data: one that includes only the edge value histogram feature. Since the edges histogram feature assumes all edges are independent, the gradient for this model can be computed exactly as can its predicted networks. The green line in Figure 1 shows the expected density of the network as predicted by our model over time. With $v = 5$ this model has 16 parameters.

The red lines in Figure 1 represent prediction from a time-homogeneous Markov chain that learns a complete $v \times v$ transition matrix (and thus has 20 parameters). Each red line shows the expected density of a separate chain run forward from every observed data point. As is to be expected, the chains quickly converge to their stationary distribution. But that distribution is not near the data: the root mean square error for the Markov chain is 87, but for the time-inhomogeneous model it is only 45.

Conversation Data

The second dataset we use is one that captures the face-to-face conversations between a cohort of incoming graduate students. We recruited a group of 24 (out of 27) first-year graduate students from the same department at a large university (Wyatt, Choudhury, and Kautz 2007). For one week per month, over the 9 months of an academic year, the students wore microphones connected to a wearable computer. No raw audio was ever saved. Instead, the computer extracted a set of privacy-sensitive features in real-time. After recording is complete, the resulting streams of features are combined and we can automatically find face-to-face conversations in them with accuracies ranging from 96% to 99% (Wyatt, Choudhury, and Bilmes 2007).

We divide this data into 2 day long windows with a sliding offset of 1 day. Due to academic calendar fluctuations (and a technical issue after the 3rd week) the recording weeks do not all start at evenly spaced intervals. A network is built from each window by putting an edge between two students if they spent time in conversation during the window. The edge weight is set to the proportion of time in conversation: the amount of time that the pair spent in conversation divided by the amount of time that both members of the pair simultaneously recorded.

The model we apply to this data includes all three features described above: edge value histograms, anti-stability, and GWESP. We fit both a time-inhomogeneous model that uses sigmoid constraints on the weights and a time-homogeneous model that learns the same weights for all timesteps. To test the models we simulate entire sequences of networks from them.

For the time-inhomogeneous model we provide only the time indexes at which it should generate networks. For the time-homogeneous model we provide both the time indexes for which it should generate networks, as well as the true first network observation—thus giving it potentially more information about the network series.

We use Gibbs sampling to generate sample sequences, with a burn in of 1000 sweeps over all variables and subsequent samples saved every 100 sweeps. We compare the simulated sequences to the data using more features than just those in the model. Such comparisons will easily show whether a model is degenerate (Hunter, Goodreau, and Handcock 2008).

Figure 2 shows the density of the conversation networks along with the densities of networks sampled from the two models. On the top, in red, the time-homogeneous model shows a very poor fit to the data. The extreme samples that extend beyond the plot's limits show that the model is exhibiting degeneracy and assigning significant probability to completely connected graphs. In fact, if we sample from this model without conditioning it on the first observation it only returns sequences of graphs with all edges set to or near their maximum value. The time-inhomogeneous model on the bottom, in green, shows a much better fit to the data.

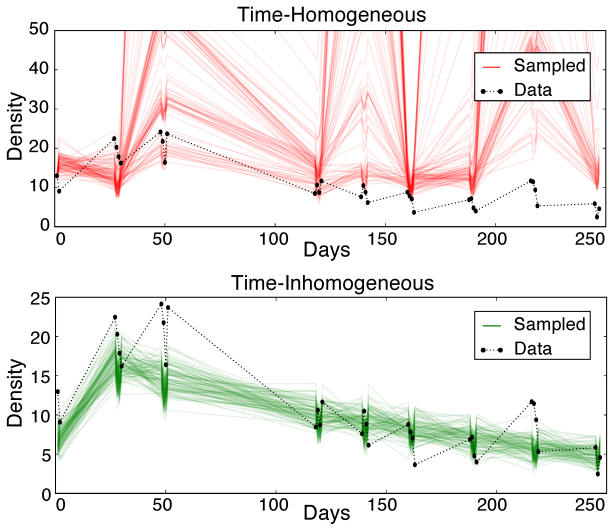

Figure 2: Density of sampled networks compared to data.

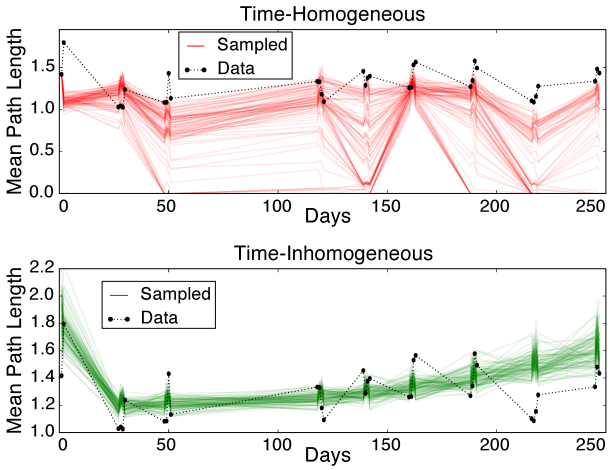

Figure 3: Mean path length of sampled networks compared to data.

Figure 3 shows mean path lengths. To compute path lengths, every non-zero edge y_{ij} has its value replaced with $1 - y_{ij}$ so that shorter paths travel across stronger edges. Path length is a global property of the network and thus is not directly modeled by our strictly local set of features. Good reproduction of global properties is evidence of good model fit (Hunter, Goodreau, and Handcock 2008). Again, the time-homogeneous model (top) shows a poor fit to the data and exhibits degeneracy by generating many maximally connected networks with all paths at length zero. The time-inhomogeneous model (bottom) provides a much better fit.

What Didn't Work Before arriving at the above features we also tried two others. Simple density—the sum of all edge values—yielded networks with very low total densities and was replaced with the edge value histograms. A raw triangles count defined as $D \triangleq \sum_{ijk}(x_{ij}x_{ik}x_{jk})^{\frac{1}{3}}$ and a "poor man's GWESP" of $\log(1 + D)$ both lead to degeneracy.

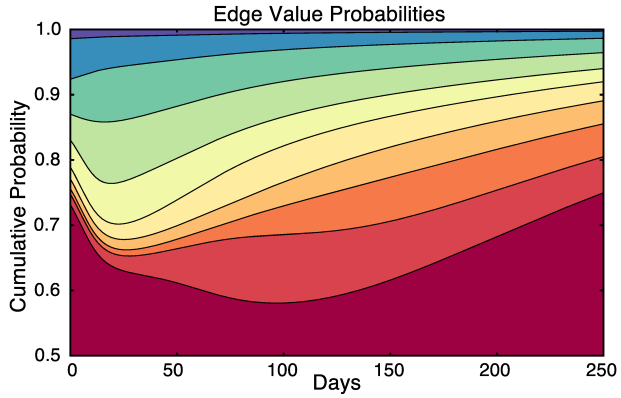

Figure 4: Edge value probabilities over time, with all other features kept equal. Values increase from 0 at bottom to 1 at top. Note that y axis starts at 0.5.

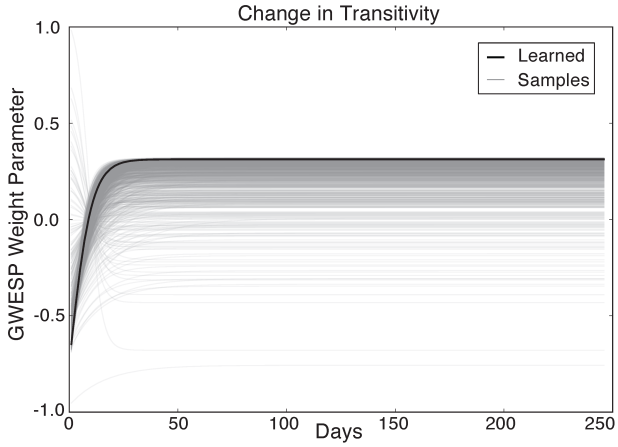

Figure 5: Estimated GWESP weight, with additional weights drawn from its nominal sampling distribution.

Interpreting the Model Once an ERGM has been fit to data, its parameters can be interpreted to provide insight into the process that generated the network. A linear ERGM is easiest to interpret: a parameter's value is the log-odds of a unit increase in that feature. Features like GWESP are also fairly interpretable. The curve defined in (4) shows the log-odds of an edge having i shared partners and the smooth curve shows the rate at which the log-odds diminishes in i. The same is true of our time-inhomogeneous parameters. Since the edge histograms are simple multinomials, we can convert their parameters from canonical to mean-value form and view them as the probability that an edge takes a given value at a given time, with all other features held equal.

Figure 4 shows such edge value probabilities for the conversation networks. In the beginning, strong edges (top) have larger probability than others. In the middle, weak edges become more probable, and eventually zero-valued edges (bottom) increase their dominance.

Of course, there is the usual uncertainty associated with the single set of parameters learned by the model. Fortunately, curved exponential families still allow for the Fisher information to be used to estimate standard errors around

Table 1: Sigmoid parameter values learned for GWESP weight.

Parameter	Value	s.e
w	-2.24	1.17
a	-0.09	0.95
b	-0.19	0.05
s	-0.14	0.07

learned parameter values (Hunter and Handcock 2006). Table 1 shows the parameter values (from (15)) learned for the GWESP weight in the conversation data, along with their nominal standard errors. Unfortunately, as $\boldsymbol{\eta}$ becomes more complex, standard errors around a learned $\hat{\boldsymbol{\theta}}$ become harder to reason about. We can get a coarse feel for the uncertainty, though, by sampling new parameter values from the nominal asymptotic normal sampling distribution of $\hat{\boldsymbol{\theta}}$. We can then feed those sampled parameters through $\boldsymbol{\eta}$ to see how different $\hat{\boldsymbol{\theta}}$ values might effect our interpretation of the model. Figure 5 shows such samples for the learned GWESP weight. The MLE (the output of (15) for the values in Table 1) is in solid black and the gray lines are weights computed from sampled values of $\hat{\boldsymbol{\theta}}$. The sampled values follow the general form of the point estimate and all suggest that in this network transitivity quickly increases in importance and then stays important.

Conclusion and Future Work

We have shown that traditional time-homogeneous models may not be best for modeling sequences of social networks and that time-inhomogeneous variants of them can perform better. We have also shown a way that existing CERGM features can be successfully extended to multi-valued networks that arise naturally in temporal network data.

Temporal network modeling is in its infancy and there are many additional avenues to explore. We can add dynamic structural features, like transitivity over time; we can try different parameter constraint functions; we can incorporate the other streams of observations that come with the new kinds of network data like conversational styles, physical activity, and location. The future work is immense.

Acknowledgements

This work was supported by NSF grants IIS-0433637, IIS-0535100, and IIS-0845683.

References

Eagle, N.; Pentland, A. S.; and Lazer, D. 2009. Inferring friendship network structure by using mobile phone data. *PNAS* 106(36):15274–15278.

Efron, B. 1978. The geometry of exponential families. *The Annals of Statistics* 6(2).

Fowler, J. H. 2006. Legislative cosponsorship networks in the U.S. house and senate. *Social Networks* 28(4):454–465.

Frank, O., and Strauss, D. 1986. Markov graphs. *J. Am. Stat. Assoc.* 81(395):832–842.

Geyer, C. J., and Thompson, E. 1992. Constrained monte carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society* 54(3):657–659.

Guo, F.; Hanneke, S.; Fu, W.; and Xing, E. C. 2007. Recovering temporally rewiring networks: A model-based approach. In *Proc. of ICML*.

Handcock, M. 2003. Assessing degeneracy in statistical models of social networks. Technical Report 39, UW CSSS.

Hanneke, S.; Fu, W.; and Xing, E. 2009. Discrete temporal models of social networks. arXiv.

Hunter, D. R., and Handcock, M. 2006. Inference in curved exponential family models for networks. *J. Computational and Graphical Statistics* 15(3).

Hunter, D. R.; Handcock, M. S.; Butts, C. T.; Goodreau, S. M.; and Morris, M. 2008. ergm: A package to fit, simulate and diagnose exponential-family models for networks. *J. Stat. Software* 24(3).

Hunter, D.; Goodreau, S.; and Handcock, M. 2008. Goodness of fit of social network models. *J. Am. Stat. Assoc.* 103(481):248–258.

Robins, G., and Pattison, P. 2001. Random graph models for temporal processes in social networks. *J. Mathematical Sociology* 25(1):5–41.

Robins, G.; Snijders, T.; Wang, P.; Handcock, M.; and Pattison, P. 2007. Recent developments in exponential random graph (p*) models for social networks. *Social Networks* 29(2).

Robins, G.; Pattison, P.; and Wasserman, S. 1999. Logit models and logistic regressions for social networks: III. valued relations. *Psychometrika* 64(3):371–394.

Snijders, T. 2002. Markov chain monte carlo estimation of exponential random graph models. *J. Social Structure* 3(2).

Strauss, D., and Ikeda, M. 1990. Pseudolikelihood estimation for social networks. *J. Am. Stat. Assoc.* 85:204–212.

Wasserman, S., and Faust, K. 1994. *Social Network Analysis*. Cambridge UP.

Wasserman, S., and Pattison, P. 1996. Logit models and logistic regression for social networks: 1. An introduction to markov graphs and (p*). *Psychometrika* 61.

Wren, C.; Ivanov, Y.; Leigh, D.; and Westhues, J. 2007. The MERL motion detector dataset. In *MD '07: Proc. of the 2007 Workshop on Massive Datasets*.

Wyatt, D.; Choudhury, T.; and Bilmes, J. 2007. Conversation detection and speaker segmentation in privacy sensitive situated speech data. In *Proc. of Interspeech*.

Wyatt, D.; Choudhury, T.; and Bilmes, J. 2009. Dynamic multi-valued network models for predicting face-to-face conversations. In *NIPS workshop on Analyzing Networks and Learning with Graphs*.

Wyatt, D.; Choudhury, T.; and Kautz, H. 2007. Capturing spontaneous conversation and social dynamics: A privacy-sensitive data collection effort. In *Proc. of ICASSP*.