Learning Hidden Curved Exponential Family Models to Infer Face-to-Face Interaction Networks from Situated Speech Data

Danny Wyatt
Dept. of Computer Science and Engineering
University of Washington
Seattle, WA 98195
danny@cs.washington.edu

Tanzeem Choudhury
Dept. of Computer Science
Dartmouth College
Hanover, NH 03755
tanzeem.choudhury@dartmouth.edu

Jeff Bilmes
Dept. of Electrical Engineering
University of Washington
Seattle, WA 98195
bilmes@ee.washington.edu

Abstract

In this paper, we present a novel probabilistic framework for recovering global, latent social network structure from local, noisy observations. We extend curved exponential random graph models to include two types of variables: hidden variables that capture the structure of the network and observational variables that capture the behavior between actors in the network. We develop a novel combination of informative and intuitive conversational (local) and structural (global) features to specify our model. The model learns, in an unsupervised manner, the relationship between observable behavior and hidden social structure while simultaneously learning properties of the latent structure itself. We present empirical results on both synthetic data and a real world dataset of face-to-face conversations collected from 24 individuals using wearable sensors over the course of 6 months.

Introduction

The structure of social networks and the nature of communication among people are important in trying to understand a variety of social phenomena, such as diffusion of information, coalition formation, decision-making, and the spread of disease. Complex macro-social phenomena can arise from simple micro-level behavior without global coordination. For example, racial segregation in neighborhoods can occur simply from individuals wanting to avoid being in the minority even in a population that prefers diversity (Schelling 1971). Similarly, micro-level behavior can reveal information about macro-social structure. For example, people's speaking style can be predictive of their position in a social network (Choudhury and Basu 2004).

These correlations between local behavior and global structure are difficult to study because they require behavioral data gathered at a resolution that is not possible using traditional manual techniques. The advent of the Internet has made it easier to study virtual social behavior (Adamic and Adar 2003; Kossinets and Watts 2006), but face-to-face interactions are still people's predominant method of communication (Baym, Zhang, and Lin 2004). Recent advances in wearable and ubiquitous computing have made it easier to collect face-to-face interaction data which has led to new efforts to study real-world social behavior by automatically collecting empirical data about people's interactions, activities, and locations (Choudhury 2003; Eagle and Pentland 2003).

A fundamental challenge to all such efforts is inferring a global, abstract social structure from the local, concrete behavioral observations. All of these new studies require, in the words of Marsden (1990), "some means of abstracting from these empirical acts to relationships or ties."

In the language of machine learning, the abstract structure can be considered a hidden state and the empirical behavioral data are indirect, noisy observations conditioned on that hidden state. Given a set of observations, the challenge is to infer the hidden state that generated the observations. And since the state is truly hidden—it can never be observed—any uncertainty contained in the observations should be acknowledged and utilized.

Typically, researchers define simple thresholds or heuristics to discard observations that are believed *a priori* to be noise. The remaining observations are then considered a direct observation of the "true" network (e.g. (Palla, Barabási, and Vicsek 2006; Kossinets and Watts 2006)). This surviving network is used as data for subsequent social network analysis. Such methods are unsatisfying because the definition of noise is ad hoc and it does not propagate any uncertainty about the latent structure into later analyses.

In this paper we propose a model that can automatically learn, in an unsupervised manner, the relationship between observable behavior and hidden social structure while simultaneously learning properties of the latent structure itself. We extend traditional statistical network analysis methods to handle data from automatically collected face-to-face conversations. We evaluate our method on a 24 person real-world social network comprising over 3,000 hours of data.

Exponential Random Graph Models

There exists a rich body of work on the statistical analysis of social networks. Traditionally, such analysis has focused on finding descriptive statistics—path lengths, degree distributions, clustering coefficients—that describe global features of the network (Wasserman and Faust 1994). In recent decades, a new class of relational models known as exponential random graph models (ERGMs, also sometimes called

Copyright © 2008, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

p-star 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, one variable for each potential edge in the network.

Given an observed network, ERGMs estimate the parameters of an exponential family model that describes the joint distribution of the edge variables. The probability distribution takes the form (typical for exponential families) of a log-linear combination of features and weights:

$$p(\mathbf{Y} = \mathbf{y}) = \frac{1}{Z_{\boldsymbol{\eta}}} e^{\boldsymbol{\eta}^{\mathsf{T}} \boldsymbol{\phi}(\mathbf{y})} \tag{1}$$

$\mathbf{Y}$ are the variables representing edges in the graph, $\boldsymbol{\phi}$ are feature functions defined on $\mathbf{y}$, $\boldsymbol{\eta}$ is a vector of weights to be learned, and $Z_{\boldsymbol{\eta}}$ is a normalizing constant. The features are deterministic functions (or statistics) of the variables and they define conditional independence assumptions between variables.

As with all exponential families, these models have the property that the gradient of their log-likelihood is equal to the difference between the observed features and the expected features:

$$\frac{\partial}{\partial \eta_i} \log p(\mathbf{y}) = \phi_i(\mathbf{y}) - E_{\boldsymbol{\eta}} \left[\phi_i(\mathbf{y}) \right] \tag{2}$$

which is clearly maximized when the expectation equals the observed data.

For an example, consider a basic model with just two features: (i) the total number of edges in the network and (ii) the number of triangles. The edges feature models network density and the triangles term models an intuitive notion of transitivity: people form ties with friends of their friends. In this model, two ties are conditionally independent (given all other ties) if they have no person in common. This independence assumption corresponds to an intuition about the processes that create social networks: a social network is formed by people making local decisions about their ties to others.

The strength of these models lies in their ability to capture the structural dependencies in a probabilistic manner. Properties of the network can then be interpreted in terms of how they affect the network's probability. In the example above, assume that a negative weight is learned for density and a positive weight for transitivity. That would mean that, everything else being equal, it is unlikely for an edge to be added to the network unless that edge completes a triangle.

ERGMs can be represented by undirected graphical models that contain a node for each variable and a clique or set of cliques for each feature. If a model has p features, then $\boldsymbol{\phi}(\mathbf{y})$ and $\boldsymbol{\eta}$ are points in p-dimensional spaces.

In the simple example this dimensionality is clearly two, but—as the conditional independence assumption would suggest—there are more than two cliques in the graphical model. For this example, the two features can be represented by an expanded set of separate indicator features that have their weights constrained to be equal. If there are n people in the social network, the expanded feature set would have $\binom{n}{2}$ indicator variables for each edge and $\binom{n}{3}$ indicator features for each possible triangle ($p = \binom{n}{2} + \binom{n}{3}$). The weights for the edge indicators would be tied and the weights for the triangle indicators would be tied.

Tying weights so that they must be equal is a special case of putting a linear constraint on the components of $\boldsymbol{\eta}$. Any set of q linear constraints, $q < p$, on $\boldsymbol{\eta}$ can be represented by a $p \times q$ matrix $\mathbf{A}$, so that $\boldsymbol{\eta} = \mathbf{A}\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ is a q-dimensional vector. This effectively reduces the number of parameters from p to q. The number of features can similarly be reduced by redefining the feature function to be $\boldsymbol{\phi}'(\mathbf{y}) = \mathbf{A}^{\mathsf{T}} \boldsymbol{\phi}(\mathbf{y})$. The likelihood then becomes $p(\mathbf{y}) = \frac{1}{Z_{\boldsymbol{\theta}}} e^{\boldsymbol{\theta}^{\mathsf{T}} \boldsymbol{\phi}'(\mathbf{y})}$. (Clearly, this has the same form as Equation 1 and thus its gradient takes the same form as Equation 2.)

In this regard, ERGMs belong to a long tradition of graphical models that employ repeated features with tied parameters: from time-homogeneous Markov chains (Shannon and Weaver 1949) to HMMs (Baum and Petrie 1966) and DBNs (Dean and Kanazawa 1988), to CRFs (Lafferty, McCallum, and Pereira 2001), RMNs (Taskar, Abbeel, and Koller 2002), and Markov logic (Domingos et al. 2008).

In the example model, $q = 2$ and $\mathbf{A}$ is an $\left(\binom{n}{2} + \binom{n}{3}\right) \times 2$ matrix. The first column of $\mathbf{A}$ contains $\binom{n}{2}$ ones and $\binom{n}{3}$ zeros, and the second column contains $\binom{n}{2}$ zeros and $\binom{n}{3}$ ones. $\mathbf{A}^{\mathsf{T}} \boldsymbol{\phi}(\mathbf{y})$ results in a 2×1 vector whose first component is the number of edges in the graph and whose second component is the number of triangles. And thus the model has been reduced back to its original two features.

Despite the rich theory behind ERGMs, parameter learning has proven to be difficult due to a problem known as model degeneracy. 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. This can result in inferences that render the observed data extremely unlikely and can cause parameter estimation using Markov Chain Monte Carlo procedures not to converge. A full discussion of the potential pitfalls are beyond the scope of this paper, but for a detailed analysis see (Handcock 2003) and (Snijders 2002).

Curved Exponential Random Graph Models

Recently, (Snijders et al. 2006) proposed a new set of features for ERGMs that avoids degeneracy but at the price of using more complicated features. These new features involve "buried" parameters that must be set through cross-validation outside of the normal parameter learning process. An example of a such a feature is the geometrically weighted sum of all actors' degrees. To model (just) that feature, $\boldsymbol{\phi}(\mathbf{y})$ would be the complete degree histogram for the network. $\mathbf{A}$ would contain a sequence of geometrically diminishing coefficients (instead of only ones and zeros). The geometrically weighted sum would then be $\mathbf{A}^{\mathsf{T}} \boldsymbol{\phi}(\mathbf{y})$. The buried parameter in this example is the rate at which the coefficients in $\mathbf{A}$ diminish. That rate must be fixed in advance, and can only be learned through external cross validation.

To make those buried parameters part of ordinary learning, Hunter and Handcock (2006) have proposed using a curved exponential family model.

A curved exponential family allows the constraints on $\boldsymbol{\eta}$ to be non-linear. 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. 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. The likelihood for a curved exponential family is written as

$$p(\mathbf{Y} = \mathbf{y}) = \frac{1}{Z} e^{\boldsymbol{\eta}(\boldsymbol{\theta})^{\mathsf{T}} \boldsymbol{\phi}(\mathbf{y})} \tag{3}$$

and the gradient of the log likelihood is

$$\frac{\partial}{\partial \theta_i} \log p(\mathbf{y}) = \nabla_{\boldsymbol{\eta}}{}^{\mathsf{T}} \left(\phi_i(\mathbf{y}) - E\left[\phi_i(\mathbf{y})\right] \right) \tag{4}$$

where $\nabla_{\boldsymbol{\eta}}$ is the Jacobian of $\boldsymbol{\eta}$: the $p \times q$ matrix of partial derivatives of $\boldsymbol{\eta}$ with respect to $\boldsymbol{\theta}$. For more on curved exponential families see (Efron 1978).

This new formulation, known as curved ERGMs (CERGMs) has led to better performance during learning than linear ERGMs. Additionally, CERGMs have the benefit of continuing to use intuitive features while also learning interesting aspects of those features. Continuing the example, the geometrically weighted degree feature captures the intuition that a person experiences a diminishing rate of return in adding more and more ties. By learning the geometric rate parameter, we can discover exactly how that rate of return diminishes for the population being studied.

Features for CERGMs

The simple counts used in ERGMs to model social ties can lead to model degeneracy (Handcock 2003), but they often also do not fully capture the intuitions that motivated the features. For example, one expects social networks to exhibit transitivity, but only up to a point. Networks do not eventually become their complete transitive closures.

To model more nuanced notions of social processes, (Snijders et al. 2006) replaced simple counts with geometrically weighted sums. We take three of those features for our model: the geometrically weighted degree distribution (GWD), the geometrically weighted edgewise shared partner distribution (GWESP) and the geometrically weighted dyadwise shared partner distribution (GWDSP).

For an n actor network, these three features (as defined by (Hunter 2007)) all take a similar form:

$$f\left(\mathbf{H}(\mathbf{y}), \theta_r^{\mathbf{H}}\right) = e^{\theta_r^{\mathbf{H}}} \sum_{i=1}^{n-1} \left[1 - \left(1 - e^{-\theta_r^{\mathbf{H}}} \right)^i \right] H_i(\mathbf{y}) \tag{5}$$

where $\mathbf{H}(\mathbf{y})$ is a histogram describing the empirical distribution of some statistic of the network $\mathbf{y}$. Each feature of this form seen as a sum of geometrically diminishing weights—reflecting an intuition about diminishing rates of return—multiplied by a fundamental statistic. A separate multiplicative weight $\theta_w^{\mathbf{H}}$—which can be positive or negative—is learned for each feature so that the complete model is a linear weighted combination of these features.

For the GWD, $\theta_r^{\mathbf{H}} = \theta_r^{\mathbf{D}}$ and $\mathbf{H}(\mathbf{y}) = \mathbf{D}(\mathbf{y})$, the degree histogram of the network, where $D_i(\mathbf{y})$ is the count of actors in the network with degree i. Intuitively, this feature models a geometrically diminishing increase in the probability that any node will increase its degree.

For the GWESP, $\theta_r^{\mathbf{H}} = \theta_r^{\mathbf{T}}$ and $\mathbf{H}(\mathbf{y}) = \mathbf{T}(\mathbf{y})$, the edgewise shared partners histogram, where $T_i(\mathbf{y})$ is the number of ties in the network where the actors involved in that tie have mutual ties to i other actors. (In other words, the number of ties that form the base of exactly i triangles.) Similar to the GWD, this models the intuition that two friends will have the same mutual friends but that the value of adding a new mutual friend will gradually diminish.

For the GWDSP, $\theta_r^{\mathbf{H}} = \theta_r^{\mathbf{S}}$ and $\mathbf{H}(\mathbf{y}) = \mathbf{S}(\mathbf{y})$, the dyadwise shared partner histogram, where $S_i(\mathbf{y})$ is the number of unconnected pairs of actors in the network with mutual ties to i other actors. It models the probability that two individuals with mutual friends are likely to not be friends.

In (Snijders et al. 2006) each of these three weighted sums defined a single feature to be used in the model. So a model that used just these three features would have $\boldsymbol{\phi}(\mathbf{y}) = \left[f(\mathbf{D}(\mathbf{y}), \theta_r^{\mathbf{D}}), f(\mathbf{S}(\mathbf{y}), \theta_r^{\mathbf{S}}), f(\mathbf{T}(\mathbf{y}), \theta_r^{\mathbf{T}}) \right]$. Each feature has a multiplicative weight associated with it, but the geometric rate parameters are buried.

(Hunter and Handcock 2006) redefine these features using the CERGM framework. If the three sums above are replaced by the larger set of fundamental statistics (the histograms, so $\boldsymbol{\phi}(\mathbf{y}) = [\mathbf{D}(\mathbf{y}), \mathbf{S}(\mathbf{y}), \mathbf{T}(\mathbf{y})]$) and the weights learned for this larger set of features are constrained to be geometrically diminishing, then the model becomes a curved exponential family where p is the number of statistics and q is the number of θ's. The reformulation allows the model to learn the geometric rate parameter directly while maintaining the parsimonious description of the likelihood.

The weights that are learned for these features retain their socially intuitive interpretations. Specifically, for any feature parameterized as above, if $\theta_w^{\mathbf{H}}$ is its multiplicative weight and $\theta_r^{\mathbf{H}}$ is its geometric rate parameter, then $\theta_w^{\mathbf{H}}(1 - e^{-\theta_r^{\mathbf{H}}})^k$ is the log odds ratio for moving an observation from bin k to bin $k+1$. This means that the learned parameters can still be used to state useful properties of the social process that formed the network. Additionally, other features that are functions of these statistics can be incorporated into the model by incorporating their weights into the function $\boldsymbol{\eta}$ without increasing the number of statistics (p). For example, network density can be computed from the degree distribution as $\sum_i \frac{1}{2} D_i(\mathbf{y})$. (Hunter 2007) provides a thorough history and derivation of these features.

Hidden CERGMs to Infer Latent Networks

All work to date with CERGMs assumes that the network is fully observed during learning (Hunter 2007). Even for ERGMs, we are aware of only one other paper that learns the hidden structure from noisy observations (Guo et al. 2007).

Due to their probabilistic nature, it is straightforward to extend CERGMs to handle noisy observations. By treating the network itself as hidden, we can marginalize it out and consider only the marginal likelihood of the observations. Let $\mathbf{x}$ be the observed interaction behavior for all nodes n and let $\mathbf{y}$ be realizations of graphs. The marginal likelihood of $\mathbf{x}$ can be written as:

$$p(\mathbf{X}=\mathbf{x}) = \frac{1}{Z_{\boldsymbol{\theta}}} \sum_{\mathbf{y}} e^{\boldsymbol{\eta}(\boldsymbol{\theta})^{\mathsf{T}} \boldsymbol{\phi}(\mathbf{x},\mathbf{y})} \tag{6}$$

Capturing Social Activity via Conversational Features

In our model the $\mathbf{x}$ represent data about face-to-face conversations. Specifically, for any pair (i, j) we observe two variables: c_{ij}, the amount of time that i and j spend in conversation, and o_{ij} the total amount of time observed for i and j. $t_{ij} = c_{ij}/o_{ij}$ is the proportion of time that i and j spend in conversation.

We define two features that relate the proportion of time two people spend in conversation to the probability of a tie between them existing in the latent network. Similar to the structural features above, we model the intuition that spending more time in conversation increases the probability of a tie, but with ever diminishing returns. This intuition is captured using geometrically weighted features similar to GWD, GWESP, and GWDSP. We use two such features:

$$c(\mathbf{C}(\mathbf{x},\mathbf{y}), \theta_r^{\mathbf{C}}) = e^{\theta_r^{\mathbf{C}}} \sum_{i=1}^{v} \left[1 - \left(1 - e^{-\theta_r^{\mathbf{C}}}\right)^i\right] C_i(\mathbf{x},\mathbf{y}) \tag{7}$$

$$n(\mathbf{N}(\mathbf{x},\mathbf{y}), \theta_r^{\mathbf{N}}) = e^{\theta_r^{\mathbf{N}}} \sum_{i=1}^{v} \left[1 - \left(1 - e^{-\theta_r^{\mathbf{N}}}\right)^i\right] N_i(\mathbf{x},\mathbf{y}) \tag{8}$$

The statistics $\mathbf{C}(\mathbf{x},\mathbf{y})$ and $\mathbf{N}(\mathbf{x},\mathbf{y})$ are histograms of pairs where the i-th bin counts how many pairs spent approximately $i \times z\%$ of their observed time in conversation and v is an arbitrary upper limit that specifies the maximum proportion of the total observed time that any pair could spend in conversation. For our data, $z = 0.14\% \approx 3$ minutes for the pair with the largest o_{ij} and $v = 180 \approx 25\%/0.14\%$. Pairs are counted in the $\mathbf{C}(\mathbf{x},\mathbf{y})$ histogram if a tie exists between them and in the $\mathbf{N}(\mathbf{x},\mathbf{y})$ histogram if there is no tie between them. One expects the weights for $\mathbf{C}(\mathbf{x},\mathbf{y})$ to increase and then level off reflecting the reduced social utility in spending more and more time in conversation.

Our model In the model considered in this paper we use the three structural and two conversational features defined above. In addition, we include features for the density of the network and the global propensity of any pair to spend time in conversation. These two additional features can be incorporated into the weight function $\boldsymbol{\eta}$ without increasing the total number of statistics that need to be computed for the curved model in the p-dimensional space. The density can be computed from the degree histogram $\mathbf{D}$, and the global propensity for conversation can be computed from both conversation histograms.

Concretely, if we let the $n-1$ bins of $\mathbf{D}(\mathbf{y})$ be our first $1 \ldots n-1$ features, the portion of $\boldsymbol{\eta}(\boldsymbol{\theta})$ that defines the mapping for the weights on $D_i(\mathbf{y})$ is

$$\eta_i(\boldsymbol{\theta}) = [\theta_d] + \left[\theta_w^{\mathbf{D}} e^{\theta_r^{\mathbf{D}}} [(1 - (1 - e^{-\theta_r^{\mathbf{D}}})^i]\right] \tag{9}$$

where θ_d is the multiplicative weight for the density feature, $\theta_w^{\mathbf{D}}$ and $\theta_r^{\mathbf{D}}$ are the multiplicative weight and geometric rate, respectively, for the GWD feature. The weights for the $T_i(\mathbf{y})$ and $S_i(\mathbf{y})$ statistics are defined similarly, but without the additional density weight.

If the k-th feature is the first bin of $\mathbf{C}(\mathbf{x},\mathbf{y})$, then the portion of $\boldsymbol{\eta}(\boldsymbol{\theta})$ that constrains the weights of the v "tie exists" conversation histogram bins is

$$\eta_k(\boldsymbol{\theta}) = [\theta_c] + \left[\theta_w^{\mathbf{C}} e^{\theta_r^{\mathbf{C}}} [(1 - (1 - e^{-\theta_r^{\mathbf{C}}})^i]\right] \tag{10}$$

where $\theta_w^{\mathbf{C}}$ is the multiplicative weight for the "tie exists" conversation feature and $\theta_r^{\mathbf{C}}$ is its geometric rate parameter. The weights for the "no tie" histogram are defined similarly.

In addition to time spent in conversation we include one feature that attempts to capture preferential attachment between highly social people. Let $t_i = \sum_j c_{ij} / \sum_j o_{ij}$ be the total proportion of time that i spends in conversation. In other words, t_i captures how prone to conversation—or chatty—i is. This is a more behavioral observation of sociability than simple degree since it measures not how many people i interacts with but how much time i spends interacting.

Let $t_{i\setminus j} = \sum_{k \neq j} c_{ik} / \sum_{k \neq j} o_{ik}$. $t_{i\setminus j}$ represents i's baseline sociability for all partners other than j. Let m_t be the median sociability for all t_i. Each pair can be put in one of three pair-sociability categories:
(i) both are highly social: $t_{i\setminus j} > m_t$ and $t_{j\setminus i} > m_t$,
(ii) one is highly social, the other is not $t_{i\setminus j} > m_t$ and $t_{j\setminus i} \leq m_t$, or
(iii) neither is highly social $t_{i\setminus j} \leq m_t$ and $t_{j\setminus i} \leq m_t$.
From these categories we define six new features: the three counts of pair-sociability categories for pairs with a tie between them, and the three counts of pair-sociability categories for pairs with no tie between them.

Overall, the 431 dimensional $\boldsymbol{\eta}$ is represented using an 18 dimensional $\boldsymbol{\theta}$ in the curved model.

Learning and inference The maximum likelihood estimate (MLE) for a curved exponential model is the point $\hat{\boldsymbol{\theta}}$ that satisfies $\nabla_{\boldsymbol{\eta}(\hat{\boldsymbol{\theta}})}{}^{\mathsf{T}} \left(\boldsymbol{\phi}_i(\mathbf{y}) - E_{\hat{\boldsymbol{\theta}}}\left[\boldsymbol{\phi}_i(\mathbf{y})\right]\right) = 0$. This makes learning possible using simple gradient ascent methods that move in the direction of greatest increase in likelihood.

The addition of hidden variables to our model means that computing the gradient requires computing two expectations:

$$\frac{\partial}{\partial \eta_i} = E_{\boldsymbol{\theta}}\left[\boldsymbol{\phi}_i(\mathbf{y})|\mathbf{x}\right] - E_{\boldsymbol{\theta}}\left[\boldsymbol{\phi}_i(\mathbf{x},\mathbf{y})\right] \tag{11}$$

We compute these expectations using MCMC, sampling $\mathbf{y}$ with $\mathbf{x}$ fixed to find the left term, and sampling both $\mathbf{y}$ and $\mathbf{x}$ to find the right term.

Since the log likelihood is not convex for models with hidden variables we use stochastic gradient ascent to find the MLE. We take one gradient step for each training network. Even with the CERGM features we found that learning could diverge if gradient steps moved too far beyond a realistic range of the parameters (Handcock 2003). To avoid this, we normalize the gradient by its norm so that no step will increase any weight by more than 1. We use a Gaussian prior to aid with regularization. All multiplicative weights have a mean of zero and unit variance. All rate weights have

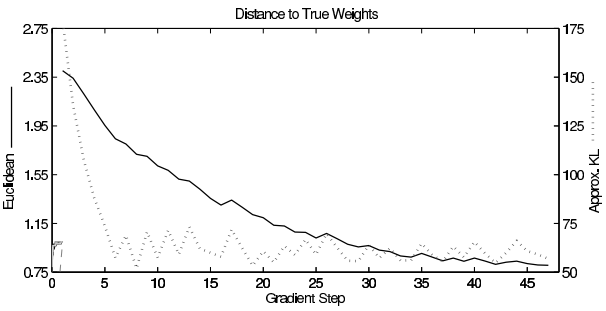


Figure 1: Distances from true to learned weights during learning.

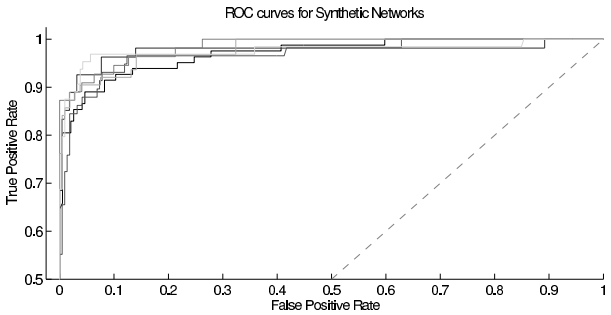


Figure 2: Solid lines are ROC curves for 6 synthetic data experiments. Dashed line indicates equal TPR and FPR.

a mean of 1 and variance of 0.5. The "edge on" conversation weight has a mean of 1, and the "edge off" weight has a mean of -1. Both have unit variance. The six pair-sociability features have means of 0 and unit variances.

Once the parameters have been learned, they can be used to infer the posterior distribution of the latent structure given the observations: $p(\mathbf{y}|\mathbf{x})$. Any properties of the global social structure can be computed from that posterior. We believe that using the posterior will yield more plausible analyses of global properties since it carries through any uncertainty inherent in the observations. Furthermore, as mentioned above, the parameters themselves encode interpretable global properties—density, transitivity—of the latent network. By learning these parameters using the marginal likelihood of the observations, the parameter estimates will include any observational uncertainty.

Experimental Results

We performed two sets of evaluations of our model: one on synthetic data and one on data gathered from a real social network.

Both evaluations take the same form. Each must learn the parameters for a 24 node network given 6 separate sets of noisy observations for that network. Each set of observations contains information about all of the $\binom{24}{2} = 276$ pairs. We assume that all 6 observations were generated by a singled, fixed distribution and thus use all six observation sets to learn one set of parameters for the model specified above. Note that learning one set of parameters is not the same as learning a single latent structure. Each of the 6 observations may be generated by different latent networks. We only assume that all of the latent networks have the same global properties (density, transitivity, etc.).

After the parameters have been learned, they are used to sample from the posterior distribution for the latent social network of each of the 6 examples separately. The mean of the samples for each latent edge variable is interpreted as the posterior probability of that edge. A concrete realization of the posterior network can be had by fixing a threshold and assembling the network of all edges with posterior probability greater than that threshold.

Synthetic Data

To test that our model is capable of recovering latent structure we ran it on a synthetic data set designed to simulate our real data. We used weights that had been fit to actual data to generate synthetic tie and time in conversation variables. In the synthetic data we know both the true parameters and the true latent structure, so we can evaluate our technique in terms of how well it recovers the original parameters and how well it can infer the latent structure.

Table 1: Mean performance on synthetic data at varying thresholds.

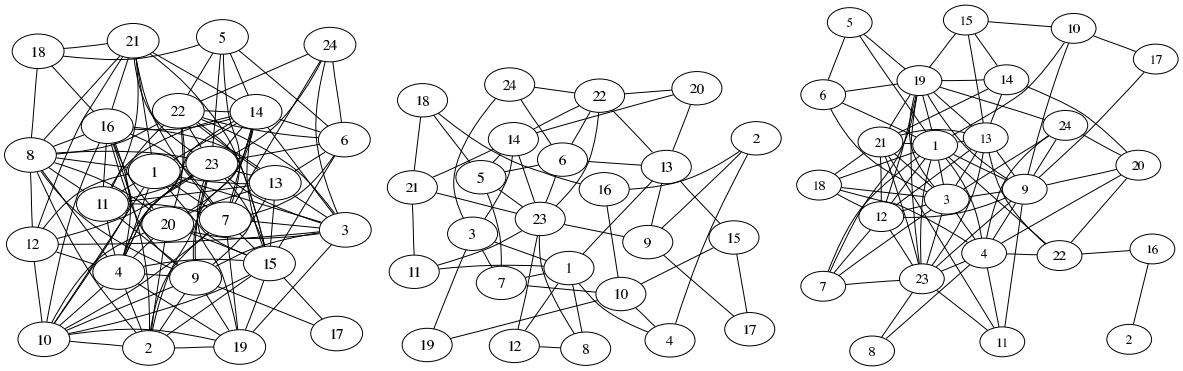
threshold	Acc.	True Pos. Rate	False Pos. Rate
0.50	87.4%	96.3%	15.2%
0.75	93.0%	92.6%	6.9%
0.90	94.8%	89.1%	3.6%
0.95	95.5%	85.9%	1.7%

Figure 1 shows, for each gradient step in the learning procedure, both the Euclidean distance between the current learned weights and the true weights and an approximation of the Kullback-Leibler divergence between the two sets of weights. The expectation required for the KL divergence is estimated using only the 6 training examples, so the KL approximation is extremely coarse.

Figure 2 shows the ROC curves for all 6 examples. For a threshold t, the true positive rate is the number of edges with posterior mean greater than t that are in the true latent network, divided by the total number of edges in the true latent network. The false positive rate is the number of edges with mean greater than t that are *not* in the true network, divided by the total number of non-edges (unconnected pairs) in the true network. As t is raised, the true positive rate increases sharply while the false negative rate remains small. Table 1 shows specific values for aggregate accuracy, true positive rate, and false positive rate at 4 different thresholds. For example, at a threshold of 0.75, we can recover the latent structure with 93% accuracy while only suffering a 7% false positive rate. In the synthetic data, the model is able to recover the latent structure quite successfully.

The Social Network Dataset

We have collected a social network corpus that contains sensor-derived measurements of conversations occurring within a cohort of 24 subjects. All of the subjects were members of the incoming graduate class of a single department at a large research university. To collect data each subject wore a wearable sensing device containing 8 different sensors useful for detecting conversations, activities, and environmental



(a) Raw conversation network. (b) Inferred network of edges with mean posterior > 0.5. (c) Survey network.

Figure 3: Conversation, inferred, and survey networks for Week 1.

Table 2: Agreements between survey networks and random networks, raw conversation networks, and inferred conversation networks.

				Inferred - Random		Inferred - Raw		Raw - Random	
Week	Random	Raw	Inferred	diff.	p-value	diff.	p-value	diff.	p-value
1	59.5%	61.6%	71.7%	12.3%	.0012	10.1%	.0058	2.1%	.3035
2	54.4%	62.9%	69.6%	15.1%	1.2E-4	6.7%	.0493	8.5%	.0214
3	51.1%	63.3%	63.8%	12.7%	.0013	0.4%	.4578	12.3%	.0018
4	66.4%	80.3%	82.3%	15.9%	1.0E-5	1.9%	.2788	13.9%	1.1E-4
5	76.8%	79.4%	83.0%	6.2%	.0350	3.6%	.1400	2.6%	.2313
6	76.8%	82.2%	84.4%	7.6%	.0117	2.3%	.2383	5.4%	.0592
Agg.	64.1%	71.3%	75.8%	11.6%	1.4E-13	4.5%	.0016	7.1%	6.0E-6

context. Data was collected during working hours for one week each month over the 9 month course of an academic year. To collect data in an ethical (and legal) manner, no raw audio was ever recorded. Only a set of privacy-sensitive features necessary for conversation extraction were saved. A complete description of this data can be found in (Wyatt, Choudhury, and Kautz 2007).

For the experiments in this paper, we used only the 6 weeks (from consecutive months) with the most data. Using the technique described in (Wyatt, Choudhury, and Bilmes 2007), we automatically extract multi-person conversations from the sensor data and segment the speaker turns within those conversations. For this analysis, we only consider conversations with two participants. As described above, we learn a single set of parameters for all six weeks and use those parameters to computer posterior distributions for the latent networks.

Evaluation of those inferred networks is difficult since the hidden structure that we are trying to recover is genuinely hidden—there is no ground truth for us to compare it to. However, at the end of each week the subjects were asked in a survey who they worked with on coursework and research, who they visited outside of school, and who they talked to on the phone. Since they only recorded sensor data during school, we can use the responses to the research and coursework questions to build a separate observation of the same latent network. We stress that these surveys should not be considered ground truth, but rather a second noisy observation of the same latent social structure. Nevertheless, one would expect there to be some agreement between our inferred structures and those expressed in the surveys.

Indeed, that is what we found. Using a threshold of 0.5, we compared the inferred networks to the survey networks and computed the number of edges for which they agreed. We compute the same agreements for random networks with the same expected density as the survey network, and for the network formed by the raw conversation data (that is, a network with edges for any pair who ever spent time in conversation). Results for these comparisons are in Table 2. Samples of the raw, inferred, and survey networks for Week 1 are in Figure 3.

For all weeks, the inferred networks have better agreement with the surveys than random networks, and the improved agreements are statistically significant (one-tailed t-test). The inferred networks also have better agreement than the raw networks in all weeks. While those improvements are not statistically significant in all weeks, they are significant in aggregate. It is well known that people have poor recall of their own social behavior and that there can be a divergence between perceived and behavioral social networks (Bernard, Killworth, and Kronenfeld 1984). An earlier study similar to this one found that survey responses had only 54% agreement with automatically extracted networks (Choudhury 2003). In light of that, these agreements are very promising.

Conclusion and Discussion

We present a hidden curved exponential family model as an intuitive approach for inferring latent network structure

from noisy behavioral observations. We also propose socially relevant conversational features derived from observations that, in combination with the structural features, are informative for recovering the hidden network graph. Using a synthetic dataset, we demonstrated the predictive capabilities of our model. On a real-world interaction dataset the network structure inferred showed better agreement with survey data than both random networks and the raw conversation networks.

An appealing property of both ERGMs and CERGMs is that they provide a framework for formalizing the modeling assumptions about the network and encoding intuitions about social processes. One of the challenges we faced in this work was evaluating the accuracy of our inferences for the real-world network data because the latent network structure is not available. One possible future evaluation strategy is to try to predict the observations for one edge given observations for other edges. For the current model trying to predict the time a pair spends in conversation pair is challenging. However, that may improve if the dynamics of the network from day-to-day or week-to-week are modeled. By using the past information about a given pair's interaction as well as observations from other pairs, a dynamic model may be able to predict held out observations for the pair. In our future work, we plan to develop a temporal extension to our current model.

References

Adamic, L., and Adar, E. 2003. Friends and neighbors on the web. *Social Networks* 25(3):211–230.

Baum, L., and Petrie, T. 1966. Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics* 37(6):1554—1563.

Baym, N.; Zhang, Y. B.; and Lin, M. C. 2004. Social interactions across media: Interpersonal communication on the internet, telephone and face-to-face. *New Media and Society* 6:299–318.

Bernard, H.; Killworth, P.; and Kronenfeld, D. 1984. The problem of informant accuracy: The validity of retrospective data. *Annual Review of Anthropology* 13:495–517.

Choudhury, T., and Basu, S. 2004. Modeling conversational dynamics as a mixed-memory markov process. In *Proceedings of NIPS*.

Choudhury, T. 2003. *Sensing and Modeling Social Networks*. Ph.D. Dissertation, Massachusetts Institute of Technology.

Dean, T., and Kanazawa, K. 1988. Probabilistic temporal reasoning. In *Proceedings of AAAI*.

Domingos, P.; Kok, S.; Lowd, D.; Poon, H.; Richardson, M.; and Singla, P. 2008. Markov logic. In Raedt, L. D.; Frasconi, P.; Kersting, K.; and Muggleton, S., eds., *Probabilistic Inductive Logic Programming*. Springer.

Eagle, N., and Pentland, A. S. 2003. Social network computing. In *Proceedings of the 5th International Conference on Ubiquitous Computing*, 289–296.

Efron, B. 1978. The geometry of exponential families. *The Annals of Statistics* 6(2):362–376.

Frank, O., and Strauss, D. 1986. Markov graphs. *Journal of American Statistical Association* 81:832–842.

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

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

Hunter, D. R., and Handcock, M. 2006. Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics* 15(3):565–583.

Hunter, D. 2007. Curved exponential family models for social networks. *Social Networks* 29:216–230.

Kossinets, G., and Watts, D. J. 2006. Empirical analysis of an evolving social network. *Science* 311:88–90.

Lafferty, J.; McCallum, A.; and Pereira, F. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML*.

Marsden, P. V. 1990. Network and data measurement. *Annual Review of Sociology* 16:435–463.

Palla, G.; Barabási, A.-L.; and Vicsek, T. 2006. Quantifying social group evolution. *Nature* 446:664–667.

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):192–215.

Schelling, T. 1971. Dynamic models of segregation. *Journal of Mathematical Sociology* 1:143–186.

Shannon, C. E., and Weaver, W. 1949. *The Mathematical Theory of Communication*. Urbana, Illinois: University of Illinois Press.

Snijders, T. A.; Pattison, P.; Robins, G. L.; and Handcock, M. S. 2006. New specifications for exponential random graph models. *Sociological Methodology* 36:99–153.

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

Taskar, B.; Abbeel, P.; and Koller, D. 2002. Discriminative models for relational data. In *Proceedings of UAI*.

Wasserman, S., and Faust, K. 1994. *Social Network Analysis*. Cambridge: Cambridge University Press.

Wasserman, S., and Pattison, P. 1996. Logit models and logistic regression for social networks: 1. an introduction to markov graphs and (p*). *Psychometrika* 61:401–425.

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

Wyatt, D.; Choudhury, T.; and Kautz, H. 2007. Capturing spontaneous conversation and social dynamics: A privacy-sensitive data collection effort. In *Proceedings of the IEEE International Conference on Acoustic and Speech Signal Processing*.