Latent Set Models for Two-Mode Network Data

Christopher DuBois
Department of Statistics
University of California, Irvine

James Foulds, Padhraic Smyth
Department of Computer Science
University of California, Irvine

Abstract

Two-mode networks are a natural representation for many kinds of relational data. These networks are bipartite graphs consisting of two distinct sets ("*modes*") of entities. For example, one can model multiple recipient email data as a two-mode network of (a) individuals and (b) the emails that they send or receive. In this work we present a statistical model for two-mode network data which posits that individuals belong to latent sets and that the members of a particular set tend to co-appear. We show how to infer these latent sets from observed data using a Markov chain Monte Carlo inference algorithm. We apply the model to the Enron email corpus, using it to discover interpretable latent structure as well as evaluating its predictive accuracy on a missing data task. Extensions to the model are discussed that incorporate additional side information such as the email's sender or text content, further improving the accuracy of the model.

Introduction

Event participation data can be studied as a two-mode network where a bipartite graph associates individuals with the events that they participate in. By explicitly representing events as a second class of node (a "*mode*"), two-mode networks capture more relational structure than the standard one-mode representation. Two-mode networks have a rich tradition in the social network literature (Breiger 1974; Borgatti and Everett 1997) and are a natural representation for relational data involving co-occurrences, membership, or affiliation between sets of entities.

Two-mode data arises in email and other digital social media, where events often have multiple participants. Previous studies tend to treat multi-recipient emails as a collection of dyadic interactions (e.g. Eckmann, Moses, and Sergi 2004) even though such emails can account for more than 40% of emails exchanged within organizations (Roth et al. 2010). Two-mode networks also arise in social networking sites, such as Facebook™, where ties can be formed between individuals and non-person entities such as event invitations and fan pages. With the increasing amount of online social interaction of all forms, there is a growing need for statistical models that can make predictions from such data.

For example, suggesting groups of email contacts can help users interact separately with family, friends, and coworkers. Users find it tedious and time-consuming to create these groups manually, motivating automatic alternatives (Roth et al. 2010; MacLean et al. 2011). To address this, recent work has demonstrated the practical utility of tools that can suggest possibly forgotten or incorrect recipients in email (Carvalho and Cohen 2008; Roth et al. 2010).

When studying network data, the goal is often to make predictions about missing or future data, as well to explore scientific hypotheses—and ideally to achieve both of these goals within a principled framework. Complicating factors can include the presence of missing data (perhaps due to privacy issues) or sparse data (e.g. the amount of information per individual is highly skewed). When additional information is available, such as attributes or covariates that capture individual-level or event-level characteristics, we would like to be able to incorporate such information into our model so as to improve our understanding of the data.

Statistical network modeling provides a general framework for handling such issues, including prediction, hypothesis testing, sparsity, attribute dependence, and so forth. Specifically, in this paper, we propose a statistical latent variable model for two-mode data based on the intuition that co-appearance patterns among individuals are driven in part by their shared group memberships.

For instance, people in the same research group are likely to be senders or recipients of the same emails. Similarly, a set of individuals that form a social clique are likely to attend the same social events. This is a relatively old idea in sociology. Simmel (1955) postulated that people's social identities are defined by their membership in various groups such as family, occupation, neighborhood, and other organizations. Feld (1981) called these shared activities and interests *foci*, and argued they help explain dyadic interaction.

In this paper we "operationalize" these concepts by proposing a statistical model that automatically infers groups or foci from event-based network data. Our proposed model explains co-appearance relationships in event data by positing that there exist latent sets of individuals that tend to appear together at the same events. We provide an algorithm for inferring the latent sets and other model parameters from observed two-mode data. The inferred latent sets impose a clustering on the individuals, with the property that individ-

uals may belong to more than one set. This can give insight into the social structure of the network, making our method useful for exploratory data analysis. Specifically, the model allows us to analyze the social identities of the individuals in terms of the sets they belong to, and the nature of the events by the sets of individuals that attend them.

Latent variable models such as this offer a sensible trade-off between modeling flexibility and computational tractability (Hoff 2009). A variety of related latent variable network models have been described in the literature. For example, two-mode blockmodels (Borgatti and Everett 1997) assign events and individuals to latent equivalence classes.

Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) is a popular latent variable model for text that has also been applied to network data (Zhang et al. 2007). A key aspect of latent variable modeling for networks is the assumption of a hypothetical latent space that characterizes individual behavior, where network processes (such as events relating sets of individuals) are conditionally independent given the individuals' representations in that space (Hoff 2009). Many latent variable models, including LDA, can alternately be seen as matrix factorization methods, where a $T \times N$ co-appearance matrix can be decomposed into a product of a $T \times K$ matrix with a row for each event and a $K \times N$ matrix with a column for each individual (Buntine and Jakulin 2006). Our proposed model can also be understood as a matrix factorization method that, unlike methods such as LDA, explictly decomposes the co-occurrence matrix using a sparse latent set representation of the individuals.

The remainder of the paper develops as follows. We then describe our model as well as the statistical methods used to learn the model from data. We illustrate the ability of the model to extract social structure on a small, two-mode dataset. Following an exploratory analysis and a set of prediction experiments on multi-recipient email data, we conclude with a discussion of related work and future directions.

Model

Our proposed model for two-mode networks assumes that co-appearance patterns in individual-event data can be explained by a relatively small number of latent sets of individuals.[1] The goal of the model is to find these latent sets, resulting in a sparse representation of individuals and events that cannot be recovered existing matrix factorization techniques. Intuitively, these sets may correspond to shared foci such as social cliques, club memberships, or workplace collaboration. According to our model, events are produced via a two-step process. Firstly, a number of sets are activated for the event. In the second step, conditioned on the list of activated sets, the list of individuals participating in the event is generated as a noisy realization of the union of individuals from the active sets. A key property of the proposed model is that performing inference to recover the latent sets from the observed event data produces a (possibly overlapping) clustering of the individuals. By encoding individuals by their inferred latent set memberships and events by the sets that are active, this framework provides a dimensionality reduction for the observed data which can aid in exploratory data analysis as well as prediction. Before describing the model formally, we first introduce some notation:

T	Number of events
N	Number of individuals
K	Number of latent sets
y_{ij}	Binary value indicating whether individual j is present at event i
z_{jk}	Binary value indicating whether individual j is a member of set k
w_{ik}	Binary value indicating whether set k is active for event i
p_{jk}	Probability that individual j appears in an event, given that they are a member of set k and only set k is active
θ_{jk}	A positive real-valued transformation of p_{jk}
q_{ijk}	Binary value indicating whether a Bernoulli trial with probability p_{jk} succeeded for event i
ρ_k	Prior probability that each individual is a member of set k
τ_i	Prior probability that each set is active in event i.

Capitalized versions of the variables above correspond to matrices, e.g. Y is the co-occurrence matrix. We now describe the model formally. The model corresponds to assuming that the data are generated by the following process:

For each individual j and set k, sample j's membership:
  $z_{jk} \sim \text{Bernoulli}(\rho_j)$
For each event i and set k, sample whether k is active:
  $w_{ik} \sim \text{Bernoulli}(\tau_i)$
For each event i and individual j:
  For each active set k that j belongs to ($w_{ik} = z_{jk} = 1$):
    Flip a weighted coin giving j the chance to attend i:
    $q_{ijk} \sim \text{Bernoulli}(p_{jk})$
  $y_{ij} = 1$ iff $\exists k : q_{ijk} = 1$.

Note that the event data y_{ij} are generated via a noisy-OR over i's active sets that j belongs to, with every such set giving individual j an opportunity to attend event i. Reparametrizing p_{jk} via $\theta_{jk} = -\log(1 - p_{jk})$, the probability of the ith individual appearing in the jth event is therefore conditionally independent of the other entries of Y given the latent variables and model parameters, and is given by the noisy-OR likelihood:

$$p(y_{ij} = 1|W, Z, \Theta) = 1 - \exp\ (-\sum_k w_{ik} z_{jk} \theta_{jk})\ . \quad (1)$$

The model can also be understood as a factor model in which the data matrix Y is probabilistically factorized as

$$p(Y|W, Z, \Theta) \sim f(W(Z \cdot \Theta)^T)\ , \quad (2)$$

where $f(x)$ is a Bernoulli distribution on the noisy-OR of each of the entries of the matrix x, and $Z \cdot \Theta$ is the Hadamard

[1] While two-mode data pertains to any bipartite network between two distinct classes of entities, we focus on the case where the classes of entities are individuals and events. However, our model is applicable to more general classes of two-mode data.

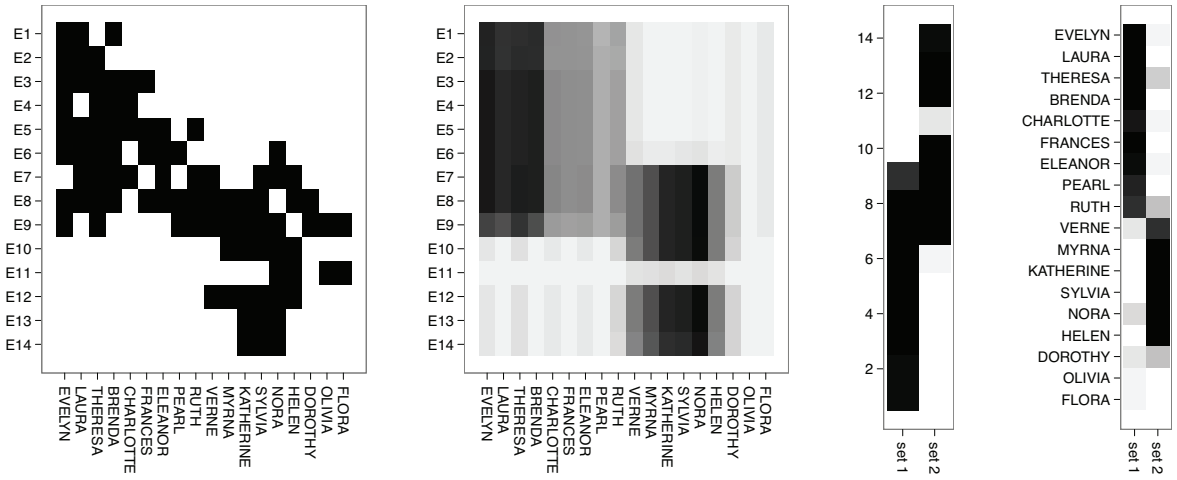

Figure 1: Left: Davis' Southern women data, as presented in Davis, Gardner, and Gardner (1941). Next, posterior estimates of the predictive distribution, W, and Z (ordered left to right) after fitting the model with $K = 2$.

(entry-wise) product of Z and Θ. In practice, we also subtract an intercept term ϵ inside the exp in Equation 1. This corresponds to including an extra coin flip in the noisy-OR that determines the probability that an individual appears at an event, even if she is not a member of any active sets.

We take a Bayesian approach to building the model, allowing us to incorporate prior knowledge about the latent variables and parameters. We place an informative $\text{Beta}(\epsilon_\alpha, \epsilon_\beta)$ prior on ϵ with $\epsilon_\alpha = 1$ and $\epsilon_\beta = 20$ so that the probability of appearing without being a member of an active set is small. We also place a prior on the set memberships, $\rho_j \sim \text{Beta}(Z_\alpha, Z_\beta) = \text{Beta}(1, K)$, implying individuals belong to only one set on average. For each event, we assume the probability of an active set $\tau_i \sim \text{Beta}(W_\alpha, W_\beta) = \text{Beta}(1, K)$. We let $\theta_{jk} \sim N(\mu, \sigma)$.

A useful aspect of the statistical framework is that it allows us to leverage side information about a particular event by placing additional structure for determining which sets are active. For example, we can instead model whether a set is active for a given event with a logistic regression

$$\text{logit}\left(p(w_{ik} = 1)\right) = X_i \beta_k \tag{3}$$

where X_i is a p-dimensional vector of covariates concerning event i, $\beta_k \sim \text{Normal}_p(0, \lambda I)$ are the respective parameters specific to set k, and $\text{logit}(x) = \log(x/(1-x))$. Later we use this extension in an email context to incorporate information about each email's sender and textual content.

In this formulation of the model one must choose the number of latent sets a priori. In our experiments below, we show the predictive performance for a particular dataset varies with different values of K. If overfitting is a concern, the value of K can be chosen via a model selection procedure using cross-validation, for example. Alternatively, one could consider a nonparametric prior on the binary matrices W and Z (Wood and Griffiths 2007; Doshi-Velez and Ghahramani 2009).

Inference

We show how to perform inference on the unknown parameters and hidden variables of our model, given the observed Y. Maximum likelihood optimization for this model via gradient-based methods is intractable as the likelihood function is not concave due to the presence of the hidden variables. An expectation-maximization algorithm is also difficult for this model: the dependence between W and Z makes it impractical to compute the expectation step.

We instead use a Markov chain Monte Carlo (MCMC) method to perform inference. MCMC algorithms are a technique for sampling from complex distributions by traversing the state space via an appropriately chosen Markov chain. The posterior distribution of our unobserved variables can be simulated using Gibbs sampling: we iteratively sample each unknown variable from its full conditional distribution given all available data and other variables. We can improve the mixing time by integrating out the set membership prior probabilities ρ and τ rather than sampling them.

Sampling Z

Given the model, we have the following sampling equation for Z:

$$\begin{aligned}
&p(z_{jk} = 1 | z_{\neg(j,k)}, W, Y, \Theta, \epsilon, Z_\alpha, Z_\beta) \\
&\propto p(Y | W, Z^*, \Theta, \epsilon) p(z_{jk} = 1 | z_{j,\neg k}, Z_\alpha, Z_\beta) \\
&= p(z_{jk} = 1 | z_{j,\neg k}, Z_\alpha, Z_\beta) \prod_{i=1}^{T} p(y_{ij} | W, Z^*, \Theta, \epsilon) \\
&= \frac{\sum_{k' \neq k} z_{jk'} + Z_\alpha}{K + Z_\alpha + Z_\beta} \prod_{i=1}^{T} p(y_{ij} | W, Z^*, \Theta, \epsilon)
\end{aligned}$$

where Z^* is the current Z but with $z_{jk} = 1$.

Sampling Θ

It is not possible to sample directly from the posterior of each θ_{jk}, so we sample these parameters using Metropolis-Hastings steps. The Metropolis-Hastings algorithm is an MCMC method where in each iteration a candidate for the next sample is generated from a "proposal" distribution, such as a Gaussian centered at the current location of the chain. The proposed sample is accepted with some probability, otherwise the previous sample is duplicated. In our case, in each Gibbs iteration we propose each θ_{jk}'s new value from $\theta_{jk} \sim \text{Beta}(\theta_\alpha, \theta_\beta)$ and accept this new value with probability

$$\min\left(1, r_{jk} = \frac{p(Y|W, Z, \Theta^{(t)}, \epsilon)}{p(Y|W, Z, \Theta^{(t-1)}, \epsilon)} \frac{p(\Theta^{(t)}|\theta_\alpha, \theta_\beta)}{p(\Theta^{(t-1)}|\theta_\alpha, \theta_\beta)}\right).$$

If the new value is rejected, the previous value is retained. We only need to sample θ_{jk} as above where $z_{jk} = 1$, rather than all $N \times K$ elements. If $z_{jk} = 0$, we can sample θ_{jk} from its prior $p(\theta_{jk}|\theta_\alpha, \theta_\beta)$. The intercept term ϵ is also sampled using similar Metropolis-Hastings updates.

Optimizing W and β

From Bayes rule we can also derive the posterior distributions of W and β:

$$p(w_{ik}|\ldots) \propto p(Y|W, Z, \Theta)p(w_{ik}|X, \beta) \text{ , and}$$
$$p(\beta_{kp}|\ldots) \propto p(W|X, \beta)p(\beta_{kp})$$

It is straightforward to sample the Ws and βs using Metropolis-Hastings updates using a procedure similar to that for Θ. However, we obtained better mixing performance by instead maximizing their posterior probabilities via a logistic regression conditioned on the other variables and the observed covariates X. This strategy is reminiscent of iterated conditional modes (ICM) (Besag 1986). While there is little theory on the long-run convergence properties of MCMC sampling used in conjunction with ICM, such a strategy can work well in practice.

Making Predictions

To make predictions we use S samples of the posterior distribution obtained via MCMC to compute a Monte Carlo estimate $\hat{y}_{ij}$ of the posterior predictive probability for individual j occurring at a given event i:

$$p(y_{ij} = 1) \approx \hat{y}_{ij} = \frac{1}{S}\sum_{s=1}^{S} p(y_{ij} = 1|W^{(s)}, Z^{(s)}, \Theta^{(s)}, \epsilon^{(s)}) \,.$$

If entries of the Y matrix are unavailable, under a missing completely at random (MCAR) assumption we can sample them in each iteration of the MCMC procedure using Equation 1, and then use the equation above at prediction time. We use this method in our prediction experiments on held-out data in the Experiments section.

Algorithmic Considerations

The sampling equations described above require us to frequently compute the current log-likelihood, so it is important to make this computation efficient. Due to the form of the log-likelihood, a sparse implementation is straightforward. Note that we have a $\log(\epsilon)$ term in the log-likelihood for each element in the set $\{(i, j) : y_{ij} = 0, \sum_k w_{ik} z_{jk} = 0 \; \forall k\}$ and an additional $\log(1-\epsilon)$ term in the log-likelihood for each element in the set $\{(i, j) : y_{ij} = 1, \sum_k w_{ik} z_{jk} = 0\}$. Only for the other entries do we need to compute $p(y_{ij} = 1|W, Z, \Theta, \epsilon)$ explicitly. This makes the log-likelihood scale with the number of edges instead of the number of entries in Y, allowing us to exploit the sparsity found in many real world networks.

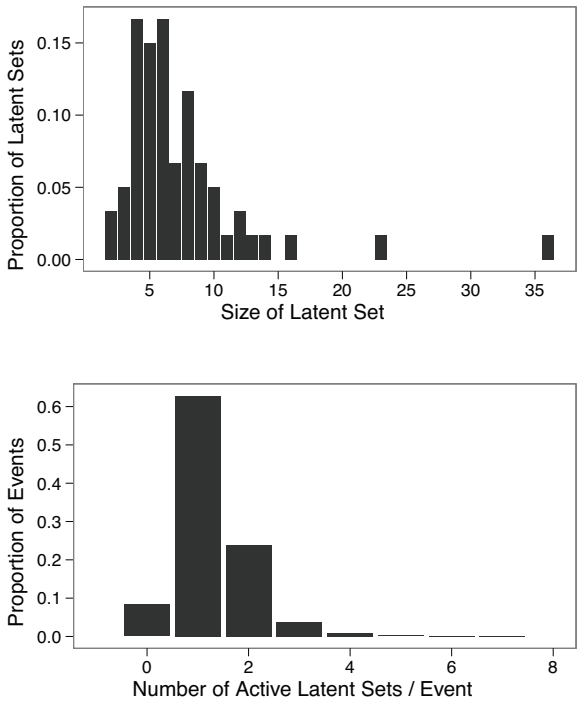

Figure 2: Latent set statistics after fitting the model with $K = 20$ to the Enron dataset. Top: Proportion of a given latent set size. Bottom: Proportion of events having a given number of active sets.

As with all MCMC algorithms, it is important to consider the mixing behavior and convergence properties of the procedure. We typically find the log-likelihood of the data converges within 30-50 Gibbs sampling iterations. As expected with a Metropolis-Hastings step the Θ and ϵ parameters tend to change slowly. However, this may not be an issue in practice as small changes in these parameters do not have a great effect on our parameters of interest (e.g. estimates of the posterior predictive density or the set memberships). In preliminary experiments we found that the algorithm benefits greatly from a sensible initialization strategy. We initialize each row of the Z matrix by computing the normalized counts of the observed sets and sampling from the resulting multinomial distribution. We also increase the robustness of the algorithm to local modes by simulating from several MCMC chains.

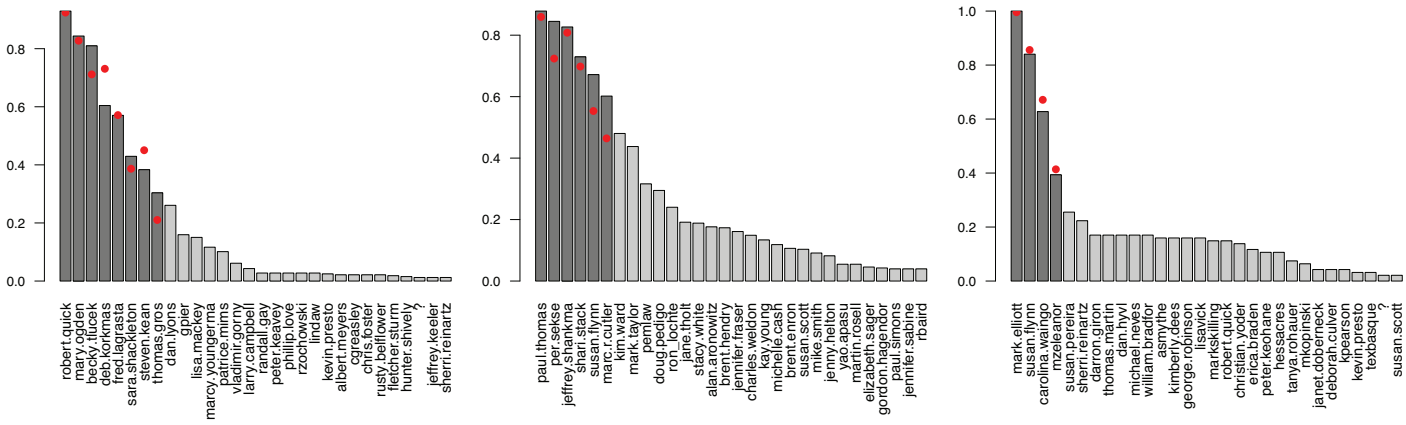

Figure 3: Proportion of events for which each individual is present and the given latent set is active, sorted from largest to smallest. Dark bars correspond to members of the set. Red dots indicate estimates of p_{jk} for members of the sets.

Illustrative Example

We illustrate the use of the model on a small social network dataset collected by Davis concerning the attendance of women at a set of social events (Davis, Gardner, and Gardner 1941). The data have a long tradition in the social network analysis literature as a means to investigate methods for two-mode data (Freeman 2003).

According to Freeman's meta-analysis of previous studies using this dataset, the general consensus is that the women are partitioned into two groups[2] whose events occasionally include members of both groups. We explore this by fitting our model with $K = 2$ latent sets. Figure 1 shows the mean of 50 samples from the posterior of W and Z after 50 iterations as burn-in. In Figure 1, the posterior predictive distribution shows general agreement with the true Y (e.g. events 7, 8, and 9 draw from both groups of women). The success of the model that fits the data satisfactorily with two latent groups both validates the appropriateness of the model for such data and supports the hypothesis that there are in fact two groups of women. The analysis here shows the least certainty regarding the set membership of Dorothy, Olivia, and Flora. It is sensible for the model to be uncertain regarding the memberships of these individuals, as they only appeared at events attended by both groups.

Evaluation

In this section we use the Enron dataset (Klimt and Yang 2004) to demonstrate the utility of the model qualitatively via an exploratory data analysis (EDA) task, and quantitatively with missing data prediction experiments. We employ a two-mode network representation consisting of individuals and the emails that they participated in as senders or receivers. For visualization purposes we consider a subset of 7319 multirecipient emails among the 191 individuals who participated in more than 150 emails between the years of 1999 and 2002.

When fitting our model to the training data using MCMC, we simulate 3 chains with 50 iterations each using the Gibbs sampler described previously on the training data. Hyperparameter settings can be found in the Model section.

Exploratory Analysis of Email Data

A primary strength of the model is the dimensionality reduction that results from the clustering of individuals into overlapping latent sets. The recovered sets help us understand the structure of the network, making the model a useful tool for exploratory data analysis.

After fitting the model with a chosen number of latent sets K, we can plot the distribution of set sizes (i.e. sums of the columns of Z) and the number of active sets per event (ie. sums of the rows of W) as seen in Figure 2. Distributions such as these are typical of fits from the model. The typical latent set size is around 5; occasionally we see sets of larger sizes of 20 to 35.

The activity of members for a few latent sets is shown in Figure 3. Each bar denotes the proportion of times a given individual was present for an event for which that set was active (ie. $\sum_i w_{ik} z_{jk}$ for a given individual j and set k). Members of the set have darker bars, and the red dots indicate model estimates for their probability of appearing, p_{jk}. Though a variety of people may be present when a given set is active (only the top 30 are shown in the Figure 3), the model finds a smaller set of individuals that are typical of the set. The model determines these sparse sets simultaneously with individuals' membership strength, as shown by the red dots. In Table 1 we list the members of these sets and include the covariates with the top five largest parameters β_{kp}. Note that some senders strongly indicate a particular latent set is active, even when that sender is not a member of the latent set. This illustrates the model can capture asymmetries between senders and sets, an effect which can be present in real networks (e.g. one person sends out emails with humorous pictures and a subset of recipients rarely gives a response).

[2]Group 1 includes Evelyn, Laura, Theresa, Brenda, Charlotte, Frances, Eleanor, Pearl, and Ruth. Group 2 includes Verne, Myrna, Katherine, Sylvia, Nora, Helen, Dorothy, Olivia, and Flora.

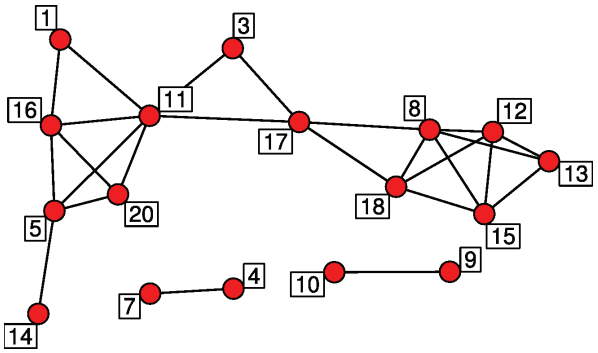

Figure 4: Network of greater-than-average co-activity among the latent sets after fitting the presented model on Enron email data with with $K = 20$.

The ability to model overlapping sets provides a tool for exploring the co-activity of the sets themselves. In Figure 4 we plot a network denoting the sets which co-appear more often than average. Three sets were isolates and were omitted from the figure. Some interesting structure becomes apparent, where a clique of sets 5, 11, 16, and 20 tend to co-appear with each other but not with sets 9 and 10, for example. This sort of decomposition is difficult to achieve with traditional blockmodels or clustering techniques.

Prediction Experiments

The model can also be used for prediction tasks such as predicting the presence of individuals given information about an event. As shown in Equation 2, the model can be understood as a constrained matrix factorization method, where W is constrained to be binary and $Z \cdot \Theta$ is sparse. While these constraints are essential to the latent set interpretation, the model pays a penalty in representational power relative to unconstrained factorization methods. As might be expected, unconstrained matrix factorization approaches such as LDA appear to provide slightly better predictive performance, and hence may be better suited if prediction is the primary goal.

We therefore focus our experimental comparison on methods that produce an interpretable latent set representation of the data. The purpose of these experiments is not to demonstrate the model's predictive power, but instead to validate its ability to infer sets that capture the latent structure of the data. Specifically, we compare our model to K-means clustering, a method for assigning data points to K non-overlapping latent clusters. The clusters that K-means extracts are a similar notion to our latent sets, making it a natural competitor to our model.

In our experiments, we also consider several extensions to the model where covariates specific to email data are incorporated via extra observed variables X and corresponding regression coefficients β in the logistic regression on the Ws in Equation 3. The covariates we considered were topic variables extracted from the text of the emails, as well as the identity of senders and the day of the week. The topics, intended to correspond to semantic themes such as "sports" or "politics", were extracted using the `lda` package[3] with the number of topics set to 20. We introduced binary variables X_{im} for each topic m indicating whether email i contained a word from that topic.

Experimental Setup

For the experiment we fit our models to a training set consisting of 80% of email events. Observations from the remaining test set were chosen to be missing completely at random with probability $p = .5$.

Predictions from the model can be used to rank the most likely individuals to appear. For missing observations in the test set we consider the top M-ranked individuals predicted by the model and ask if any of these individuals are present in the event. Models are compared using their average performance on this task across events for $M = 5$, 10, and 20. These values correspond to the far left side of the receiver-operating characteristic curve. We also consider the area under the receiver-operating characteristic curve (AUC) for the task of filling in all missing entries.

After estimating Z and Θ from the training set with K sets, for each event we must decide which sets are active while conditioning on observed information (e.g. individuals who are known to be present/absent, available covariates X, and so on). For each chain we sampled the w_{ik} values for all test events, using the values after 15 iterations for computing predictions.

For comparison, we applied K-means for each value of K, clustering the rows (events) of Y. Missing entries of the matrix were set to zero at training time. To perform prediction we assigned each test data point to its closest cluster and used the cluster means to rank individuals.

Results

In Table 2 we provide the results of our missing data experiment. Without any covariates our model outperforms the K-means method for $K = 20, 40$ and 60, and for each prediction task. We conjecture that this is due to the ability of our model to handle noise via a probabilistic formulation.

Unsurprisingly, the model containing information about the email's sender outperforms the model without any covariates. The difference in performance when we add in topic model covariates (labeled "Both") is not statistically significant. This indicates that individuals tend to send emails to the same sets of people, and that in a relative sense the content of emails is not particularly informative.

We note that the difference in performance of our model versus the baseline (K-means) is most pronounced at the highest precision.

Discussion

Our proposed model considers two-mode network data to be represented by latent sets of individuals that may be active for each observed event. Connections can be drawn between the proposed model and other models and algorithms in the literature. The model presented here can be viewed

[3] Available at www.cran.r-project.org/web/packages/lda/ .

Set Members	p_{jk}	Set Members	p_{jk}	Set Members	p_{jk}
robert.quick	0.89	jeffrey.shankma	0.73	mark.elliott	0.99
mary.ogden	0.84	shari.stack	0.68	susan.flynn	0.85
becky.tlucek	0.71	paul.thomas	0.66	carolina.waingo	0.66
deb.korkmas	0.64	per.sekse	0.62	mzeleanor	0.46
fred.lagrasta	0.54	susan.flynn	0.49		
sara.shackleton	0.33	marc.r.cutler	0.25		
steven.kean	0.28				
thomas.gros	0.24				

Top Covariates	β_{kp}	Top Covariates	β_{kp}	Top Covariates	β_{kp}
eric.bass	12.61	kaye.ellis	7.2	kay.chapman	7.44
martin.cuilla	9.19	carol.clair	6.6	beverly.stephen	4.66
daren.farmer	8.76	sara.shackleton	6.51	communications enron	1.74
monique.sanchez	8.27	mark.taylor	6.34	customers think get	1.27
jim.schwieger	8.21	janette.elberts	5.73	kay lynn meeting	1.27

Table 1: Top: Estimates of p_{jk} for members of selected latent sets after fitting the model to the Enron data with $K = 20$. Bottom: Five covariates with the largest effect for predicting that the given set is active.

as Bayesian matrix factorization, as per Equation 2. The chief difference between our model and other factor models is the sparsity constraints that are used to impose the latent set interpretation for the model. Wood and Griffiths (2007) propose a similar model, but they use an Indian Buffet Process prior on Z, do not use a Θ matrix, and the noisy-OR likelihood that they use assumes each additional active set contributes equally to the probability of an appearance. Factor modeling with a noisy-OR likelihood has been applied to link analysis before (Singliar and Hauskrecht 2006). The model of Singliar and Hauskrecht is similar to our proposed model, but a Z matrix is not used so the representation is not sparse, and thus, it cannot be interpreted as a latent set model. The work here is most similar to the single-mode network model of Mørup, et al (2010) (Mørup and Schmidt 2010) who use an IBP prior on Z, though their parameterization of the noisy-OR likelihood resembles a stochastic block model rather than allowing for varying membership strengths as in the proposed model. Another model leveraging the IBP for network modeling is due to Miller, Griffiths and Jordan (2009) , who propose a sparse latent feature model for single-mode network data. Single-mode matrix factorization models such as this typically have a bilinear form $Y \sim f(ZWZ^T)$. K-means clustering can be viewed as a latent variable method, where objects are assumed to each belong to a latent cluster. The latent sets in our model can be viewed as cluster memberships that are allowed to overlap. Although K-means is a non-probabilistic method, it is closely related to probabilistic mixture models, and has a matrix factorization interpretation similar to the factorization that our model performs.

Relative to deterministic models like K-means, the noisy-OR likelihood essentially puts a noise model on top of the factorized representation of the data matrix. This is reminiscent of Gelman et al. (2010), where a stochastic component extends Guttman scaling to a probabilistic framework. From another view, our model can be understood as a spike-and-slab distribution, where we place a large portion of probability on a small set of individuals and place a small probability, ϵ, on all other occurrences.

Co-clustering (also known as bi-clustering) is a related approach to finding latent structure in matrix data, where a compressed approximation of matrix data is obtained by simultaneously clustering the rows and columns of the matrix. A general formulation of co-clustering can be found in Banerjee et al. (2004).

Two prominent approaches to statistical modeling of network data are exponential family random graph models (ERGMs) (Wasserman and Pattison 1996; Robins et al. 2007) and latent variable methods (Hoff 2009). ERGMs have been applied in the past to two-mode data (Wang et al. 2009), though currently these methods do not scale well to large networks, as are common in social media.

Methods for ranking the relevance of items to a small set of query objects have been explored previously (Ghahramani and Heller 2005; Roth et al. 2010), however these methods do not extract latent set representations from the data. Alternative inference methods which have also been applied to Bayesian factorization models could potentially be applied here, such as the particle filter of Wood and Griffiths (2007) or variational inference (Doshi-Velez and Ghahramani 2009; Jaakkola and Jordan 1999).

Conclusion

We presented a generative model for two-mode network data that assumes there exist latent sets of individuals that tend to co-appear, and showed how to infer these sets using a Markov chain Monte Carlo algorithm. The latent sets impose a clustering on the individuals, providing insight into the social structure of the network. We apply the model to an exploratory analysis of multiple recipient email communication data. Extensions for the model were given, showing how to model the influence of covariates such as topics in the email text and sender-specific effects. Predictive experiments on a missing data task demonstrated that the model performs well relative to K-means clustering, and showed the benefits of incorporating covariates. Although we focused on email data, the methods presented here can readily be applied to other types of two-mode data.

	Top 5			Top 10			Top 20			AUC		
Number of Sets	20	40	60	20	40	60	20	40	60	20	40	60
K-Means	0.486	0.493	0.538	0.559	0.603	0.618	0.786	0.846	0.862	0.887	0.898	0.902
No Covariates	0.609	0.694	0.71	0.698	0.773	0.798	0.855	0.907	0.931	0.904	0.931	0.94
Topics	0.616	0.651	0.663	0.709	0.738	0.753	0.868	0.879	0.894	0.901	0.918	0.935
Senders	0.667	**0.742**	**0.76**	0.768	**0.825**	**0.843**	**0.919**	**0.936**	**0.957**	**0.929**	**0.953**	**0.963**
Both	**0.689**	0.726	0.749	**0.794**	0.819	0.83	**0.919**	0.934	0.946	0.921	0.943	0.952

Table 2: Experimental results on a missing data task.

Acknowledgments This work was supported in part by an NDSEG Graduate Fellowship (CD) and by ONR/MURI under grant number N00014-08-1-1015 (CD, JF, PS). PS was also supported by a Google Research Award.

References

Banerjee, A.; Dhillon, I.; Ghosh, J.; Merugu, S.; and Modha, D. S. 2004. A generalized maximum entropy approach to bregman co-clustering and matrix approximation. In *Proceedings of the 10th ACM SIGKDD*, 509–514. ACM.

Besag, J. 1986. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society, Series B* 48(3):259–302.

Blei, D.; Ng, A.; and Jordan, M. 2003. Latent Dirichlet allocation. *JMLR* 3:993–1022.

Borgatti, S., and Everett, M. 1997. Network analysis of 2-mode data. *Social Networks* 19(3):243–269.

Breiger, R. 1974. The duality of persons and groups. *Social Forces* 53(2):181–190.

Buntine, W., and Jakulin, A. 2006. Discrete component analysis. In *Statistical and Optimization Perspectives Workshop on Subspace, Latent Structure and Feature Selection, LNCS 3940*, 1–33. Springer.

Carvalho, V. R., and Cohen, W. W. 2008. Ranking users for intelligent message addressing. In *Proc. European Conference on Advances in Information Retrieval*, 321–333.

Davis, A.; Gardner, B.; and Gardner, M. 1941. *Deep South*. University of Chicago Press.

Doshi-Velez, F., and Ghahramani, Z. 2009. Correlated nonparametric latent feature models. In *Proc. UAI*, 143–150.

Feld, S. L. 1981. The focused organization of social ties. *American Journal of Sociology* 86(5):1015–1035.

Freeman, L. C. 2003. Finding social groups: A meta-analysis of the southern women data. In *Dynamic Social Network Modeling and Analysis*, 39–97. The National Academies Press.

Gelman, A.; Leenen, I.; Mechelen, I. V.; Boeck, P. D.; and Poblome, J. 2010. Bridges between deterministic and probabilistic models for binary data. *Statistical Methodology* 7(3):187–209.

Ghahramani, Z., and Heller, K. 2005. Bayesian sets. *Advances in NIPS* 18:435–442.

Hoff, P. 2009. Multiplicative latent factor models for description and prediction of social networks. *Computational & Mathematical Organization Theory* 15(4):261–272.

Jaakkola, T. S., and Jordan, M. I. 1999. Variational probabilistic inference and the QMR-DT network. *Journal of Artificial Intelligence Research* 10:291–322.

Klimt, B., and Yang, Y. 2004. The Enron corpus: A new dataset for email classification research. *European Conference on Machine Learning* 217–226.

MacLean, D.; Hangal, S.; Teh, S. K.; Lam, M. S.; and Heer, J. 2011. Groups without tears: mining social topologies from email. In *Proceedings of the 16th international conference on intelligent user interfaces*, 83–92. ACM.

Miller, K.; Griffiths, T.; and Jordan, M. 2009. Nonparametric latent feature models for link prediction. *Advances in NIPS* 22:1276–1284.

Mørup, M., and Schmidt, M. 2010. Infinite multiple membership relational modeling for complex networks. *Networks Across Disciplines: Theory and Applications Workshop at NIPS*.

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.

Roth, M.; Ben-David, A.; Deutscher, D.; Flysher, G.; Horn, I.; Leichtberg, A.; Leiser, N.; Matias, Y.; and Merom, R. 2010. Suggesting friends using the implicit social graph. In *Proceedings of the 16th ACM SIGKDD*, 233–242. ACM.

Simmel, G. 1955. *Conflict and the Web of Group-Affiliations*. Free Press: New York.

Singliar, T., and Hauskrecht, M. 2006. Noisy-OR component analysis and its application to link analysis. *JMLR* 7:2189–2213.

Wang, P.; Sharpe, K.; Robins, G. L.; and Pattison, P. E. 2009. Exponential random graph (p^*) models for affiliation networks. *Social Networks* 31:12–25.

Wasserman, S., and Pattison, P. 1996. Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p^*. *Psychometrika* 61(3):401–425.

Wood, F., and Griffiths, T. 2007. Particle filtering for nonparametric Bayesian matrix factorization. *Advances in NIPS* 19:1513–1520.

Zhang, H.; Qiu, B.; Giles, C.; Foley, H.; and Yen, J. 2007. An LDA-based community structure discovery approach for large-scale social networks. In *Intelligence and Security Informatics*, 200–207. IEEE.