Modeling Relational Events via Latent Classes

Christopher DuBois

Department of Statistics
University of California, Irvine

May 25, 2010

This material is based on research supported by the Office of Naval Research under award N00014-08-1-1015.

Social networks and relational events

- Aim: study how massive networks of social entities interact
- Often such data is a sequence of *relational events*, a timestamped event with a sender, receiver, and action type
- Examples
  - Online social networks: sharing of media
  - One-to-one communication: email, phone, etc
  - International political events

Goal: Prediction

- What is the probability the next event is sent by individual s to receipient r?

Goal: Prediction

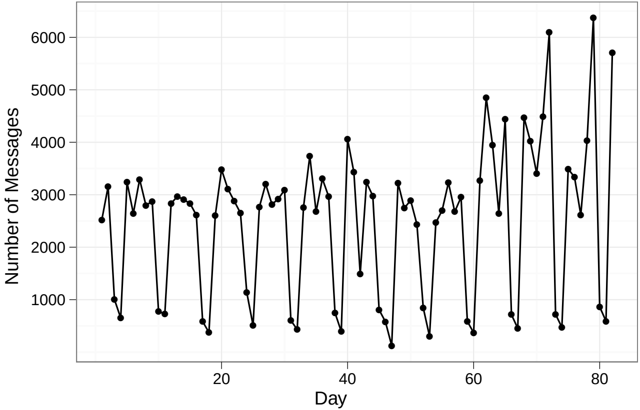
- What is the probability the next event is sent by individual s to receipient r?
- Want models that are:
  - scalable
  - interpretable
  - easily extended
  - robust to missing data
  - work when few covariates are available
  - able to share statistical strength over similar individuals/events

Real World Data: Eckmann Email Data

- 200,000 messages among 2997 individuals over 82 days

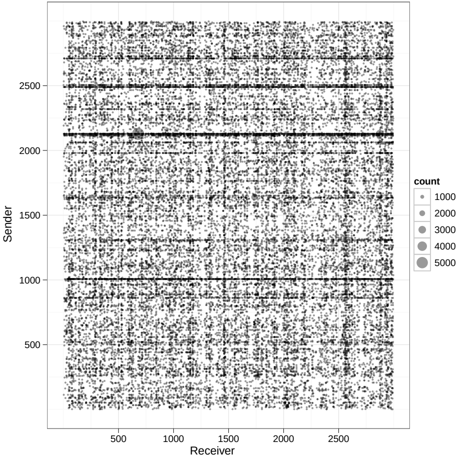
Real World Data: Eckmann Email Data

- 200,000 messages among 2997 individuals over 82 days

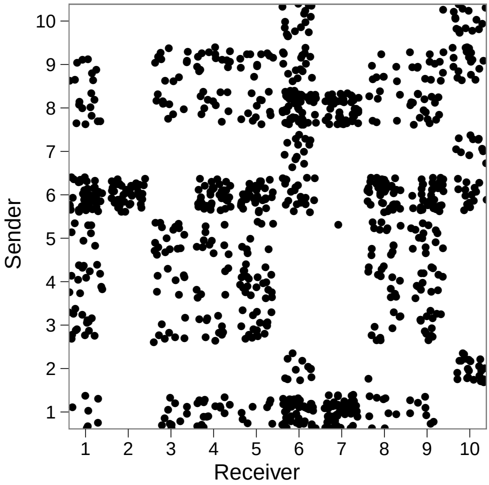


Real World Data: Eckmann Email Data

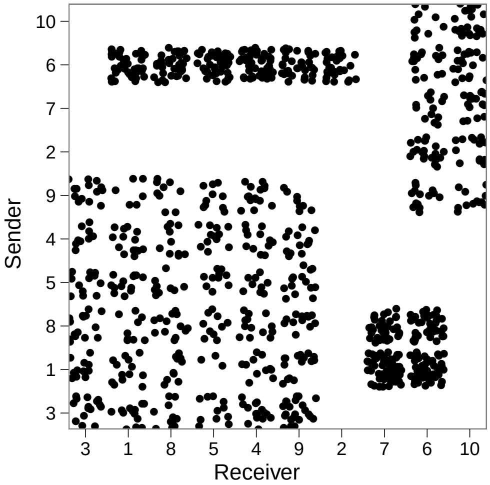
- 200,000 messages among 2997 individuals over 82 days



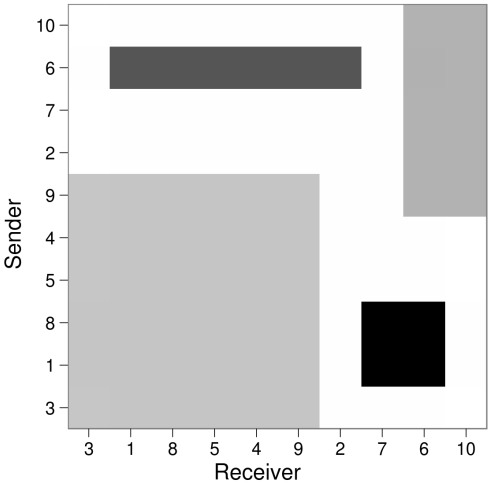
Data



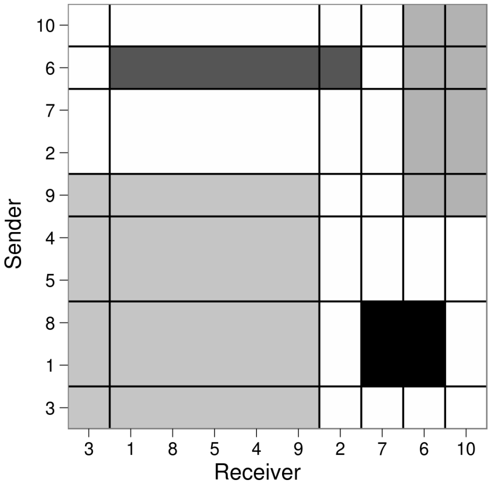
Data



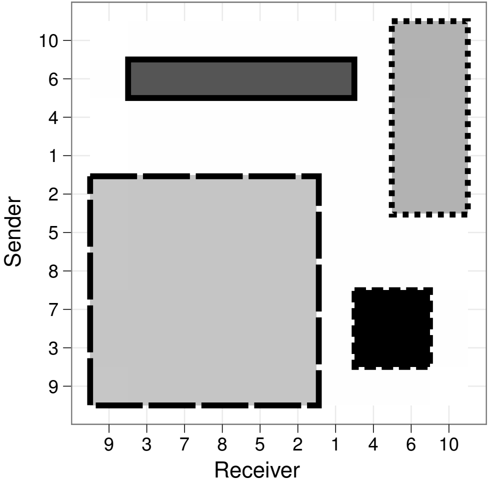
Model



Other approaches: Block models



A different approach



Marginal Product Mixture Model

Sender, receiver, action type cond. ind. given a latent class

Marginal Product Mixture Model

Sender, receiver, action type cond. ind. given a latent class

- For each event
  - Draw $c \sim \text{Multinomial}(\pi)$, the event's class
  - Draw $s|c \sim \text{Multinomial}(\theta_c)$, the event's sender
  - Draw $r|c \sim \text{Multinomial}(\phi_c)$, the event's receiver
  - Draw $a|c \sim \text{Multinomial}(\psi_c)$, the event's type

Marginal Product Mixture Model

Sender, receiver, action type cond. ind. given a latent class

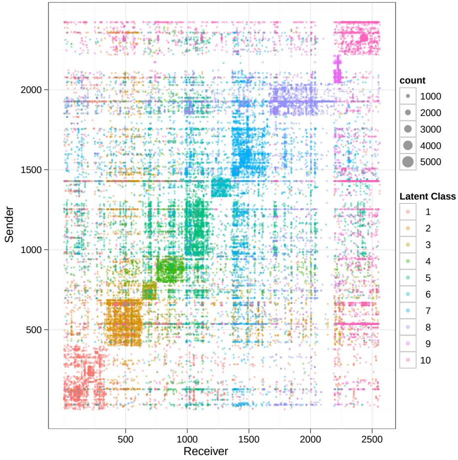
- For each event
  - Draw $c \sim \text{Multinomial}(\pi)$, the event's class
  - Draw $s|c \sim \text{Multinomial}(\theta_c)$, the event's sender
  - Draw $r|c \sim \text{Multinomial}(\phi_c)$, the event's receiver
  - Draw $a|c \sim \text{Multinomial}(\psi_c)$, the event's type
- Likelihood:

$$
\begin{aligned}
P(D|\Phi) &= \prod_{i=1}^{T} \sum_{c=1}^{C} P(s_i|\theta, c) P(t_i|\phi, c) P(a_i|\psi, c) P(c|\pi) \\
&= \prod_{i=1}^{T} \sum_{c=1}^{C} \theta_{c,s_i} \phi_{c,r_i} \psi_{c,a_i} \pi_c
\end{aligned}
$$

Inference: Leverage advances for similar models

- Data Augmentation - latent variable which represents a class assignment
- Conjugate Dirichlet priors make deriving the posterior easy
- E-step and M-step derivations are straightforward
- Integrate out θ, ϕ, ψ to derive collapsed Gibbs sampling equations for the latent assignments *c* (minimal bookkeeping required)

Exploratory Analysis with MPMM



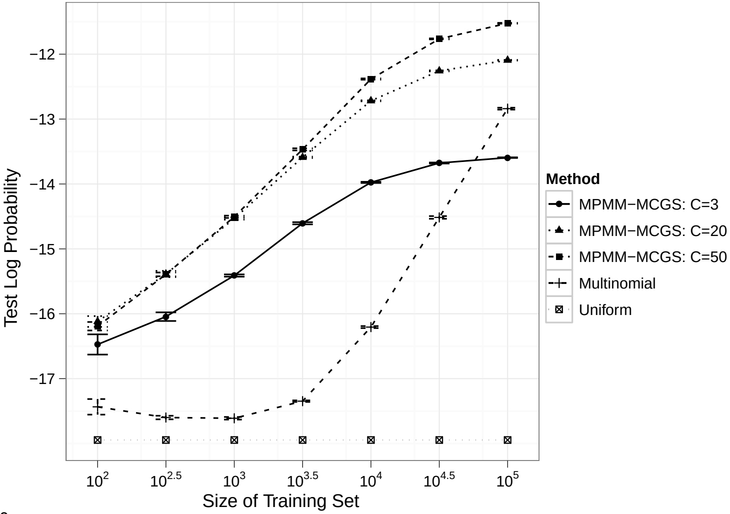
Experiments - Evaluating predictive accuracy

- Split data in training set and test set
- Evaluate log probability of test events under model:

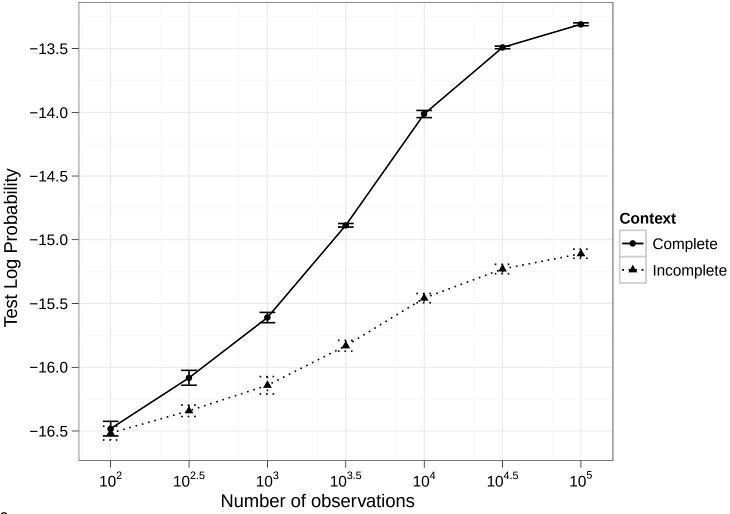
$$L_{\text{test}} = \frac{1}{T}\sum_{i=1}^{T}\log(f(Y_i|Y_{train})) = \frac{1}{T}\sum_{i=1}^{T}\log(\hat{p}_{s_i,r_i,a_i})$$

- Larger values indicate the model assigns higher probability to observed events

Experiments



Experiments



Data: International Political Events

- Automatically-coded Reuters news articles
- Subset with only US-foreign interactions:
  - 40031 events from 81 entities associated with the United States to 2695 foreign entities over 5 years
  - 178 action types (e.g. criticize, host a meeting, military occupation)

Exploratory Analysis with MPMM

Class A

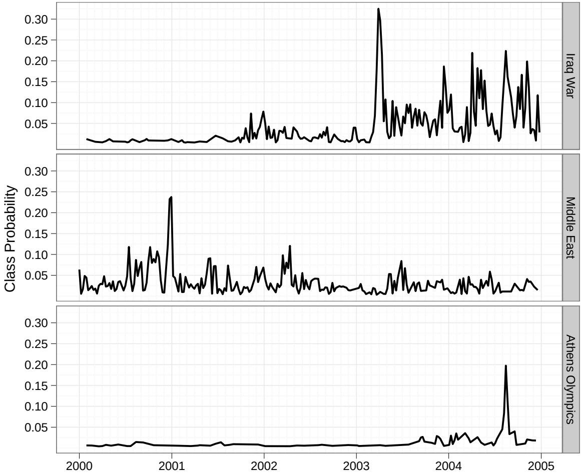
Top Senders	Pr.	Top Receivers	Pr.	Top Actions	Pr.
U.S. : Government agents	0.47	Greece : NA	0.05	Sports contest	0.59
U.S. : Athletes	0.29	Australia : Government agents	0.02	Agree or accept	0.14
U.S. : Nominal agents	0.04	United Kingdom : NA	0.02	Optimistic comment	0.04
U.S. : Police	0.04	Canada : Government agents	0.02	Comment	0.03
U.S. : Occupations	0.04	France : NA	0.01	Control crowds	0.03
U.S. : Ethnic agents	0.03	Belgium : Government agents	0.01	Improve relations	0.01

Class B

Top Senders	Pr.	Top Receivers	Pr.	Top Actions	Pr.
U.S. : Military	0.88	Iraq : Government agents	0.17	Comment	0.19
U.S. : Government agents	0.08	Iraq : National executive	0.07	Military raid	0.14
U.S. : Military hardware	0.01	Iraq : Military	0.05	Military clash	0.10
U.S. : Officials	0.00	Iraq : Ethnic agents	0.05	Military occupation	0.10
U.S. : Police	0.00	Iraq : Intangible things	0.04	Shooting	0.10
U.S. : Motor vehicles	0.00	NA : Insurgents	0.04	Political arrests and detentions	0.04

Exploratory Analysis with MPMM

International political events



Future Directions for the MPMM

- Time dependence: HMM at the class level is a simple extension
- Nonparametric: Dirichlet Process instead of a Dirichlet prior on the class distribution
- Non-symmetric priors
- Smoothing that is more specific to social networks (e.g. friend-of-a-friend effects)

Thank you!

Collapsed Gibbs Sampling Equations

$$
\begin{aligned}
P(c_i = c | z^{\neg i}, \mathcal{C}, \Phi) \quad & \propto \left(M_c^{\neg i} + \alpha_c\right) \left(\frac{U_{c,s_i}^{\neg i} + \beta}{\sum_{s=1}^{ns} U_{c,s}^{\neg i} + n_s \beta}\right) \\
& \left(\frac{V_{c,r_i}^{\neg i} + \gamma}{\sum_{r=1}^{nr} V_{c,r}^{\neg i} + n_r \gamma}\right) \left(\frac{W_{c,a_i}^{\neg i} + \delta}{\sum_{a=1}^{na} W_{c,a}^{\neg i} + n_a \delta}\right)
\end{aligned}
$$

MAP Estimates

$$
\begin{aligned}
\hat{\pi}_c &= \frac{M_c}{\sum_c M_c} \\
\hat{\theta}_{c,r} &= \frac{N_{c,s} + \beta}{\sum_{s=1}^{n_s} N_{c,s} + n_s\beta} \\
\hat{\phi}_{c,r} &= \frac{U_{c,r} + \gamma}{\sum_{r=1}^{n_r} U_{c,r} + n_r\gamma} \\
\hat{\psi}_{c,a} &= \frac{W_{c,a} + \delta}{\sum_{a=1}^{n_a} W_{c,a} + n_a\delta}
\end{aligned}
$$

Expectation-Maximization Equations

E-step:

$$P(c_i = c | s_i r_i, a_i, \Phi) \propto \theta_{c,s_i} \phi_{c,r_i} \psi_{c,a_i}$$

M-step:

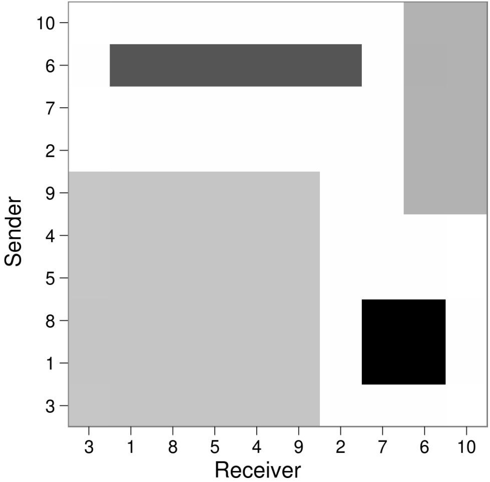
$$\begin{aligned}
\hat{\theta}_{c,s} &= \frac{\sum_{i=1}^{T} I(s_i = c) P(c_i = c)}{\sum_{i=1}^{T} P(c_i = c)} \\
\hat{\phi}_{c,r} &= \frac{\sum_{i=1}^{T} I(r_i = c) P(c_i = c)}{\sum_{i=1}^{T} P(c_i = c)} \\
\hat{\psi}_{c,r} &= \frac{\sum_{i=1}^{T} I(a_i = c) P(c_i = c)}{\sum_{i=1}^{T} P(c_i = c)}
\end{aligned}$$

Marginal Product Mixture Model

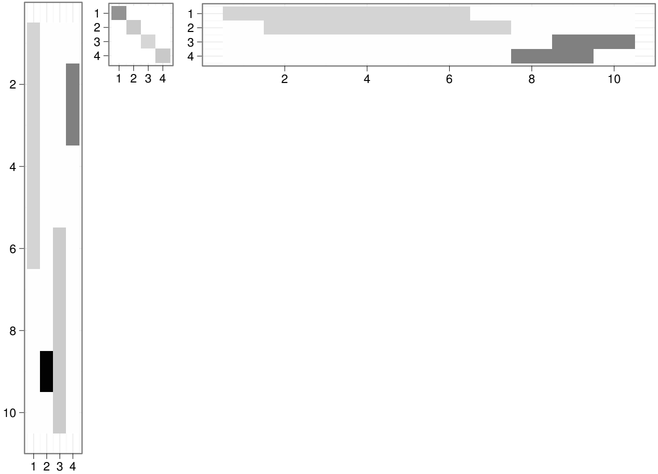
Sender, receiver, action type cond. ind. given a latent class

- Baseline: $n_s \times n_r \times n_a$ parameters
- MPMM: $C(n_s + n_r + n_a)$ parameters

Discussion: Nonnegative Matrix Factorization



Discussion: Nonnegative Matrix Factorization



Inference

- Uninformative hyperparameters for both baseline and model so that $\Pr(p) \propto 1$ and $\Pr(\Phi) \propto 1$
- Choosing C: Can use predictive accuracy on validation set (or other model selection approaches, e.g. BIC or DIC)