Topic and Role Discovery in Social Networks

Andrew McCallum, Andres Corrada-Emmanuel, Xuerui Wang

Group and Topic Discovery from Relations and Their Attributes

Xuerui Wang, Natasha Mohanty, Andrew McCallum

Presented by Steven Damer

Outline

- Background
- Problem
- Related Work
- Model
- Algorithm
- Results
- Future Work
- Conclusion

Topic and Role Discovery In Social Networks

- International Joint Conference on Artificial Intelligence 2005
- All authors from University of Massachusetts Amherst

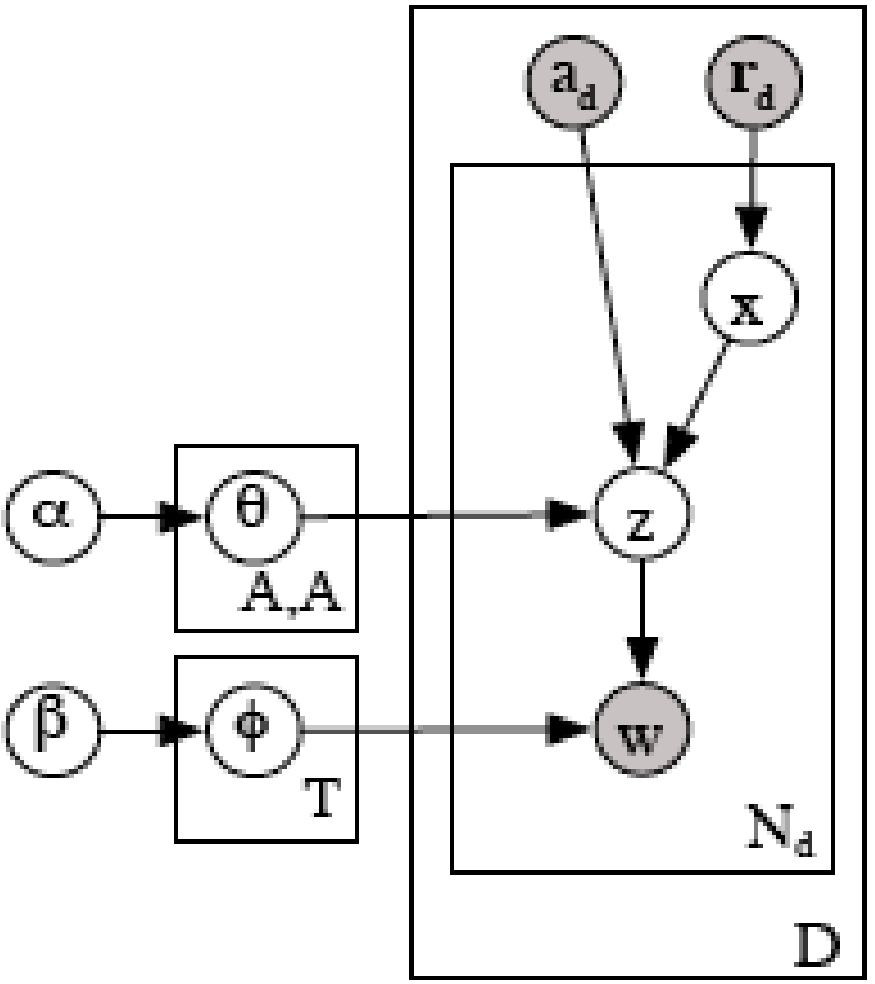
Problem

- Given a body of messages consisting of sender, message text, and a list of receivers
- Cluster the sender/receivers by role
- Can also be used to cluster documents and perform queries, but this is not generally required for this kind of data

Related Work

- Social Network Analysis – based on properties of directed graph, generally ignores content
- Latent Dirichlet Allocation, Author-Topic model – Uses content, ignores interactions

Model



- a - Author of document
- r - List of recipients
- x - Recipient assigned to word
- θ - Distribution over topics
- z - Topic assigned to word
- Φ - Distribution over words
- w - Word chosen

Algorithm

$$p(\mathbf{w}|\alpha,\beta,\mathbf{a},\mathbf{r}) = \iint p(\theta|\alpha)p(\phi|\beta)\prod_{d=1}^{D}\prod_{n=1}^{N_d}\sum_{x_{dn}}\sum_{z_{dn}} p(x_{dn}|\mathbf{r}_d)$$
$$\cdot\, p(z_{dn}|\theta_{a_d,x_{dn}})p(w_{dn}|\phi_{z_{dn}})\mathrm{d}\phi\mathrm{d}\theta.$$

- Gibbs Sampling, details left out of paper

$$P(z_i \,|\, \mathbf{z}_{-i},\mathbf{x},\mathbf{w}) \quad \propto \quad \frac{n_{z_i}^{w_v}+\beta_v}{\sum_v n_{z_i}^{w_v}+\beta_v}\frac{n_{x_i}^{z_i}+\alpha_{z_i}}{\sum_{z'} n_{x_i}^{z'}+\alpha_{z'}}$$

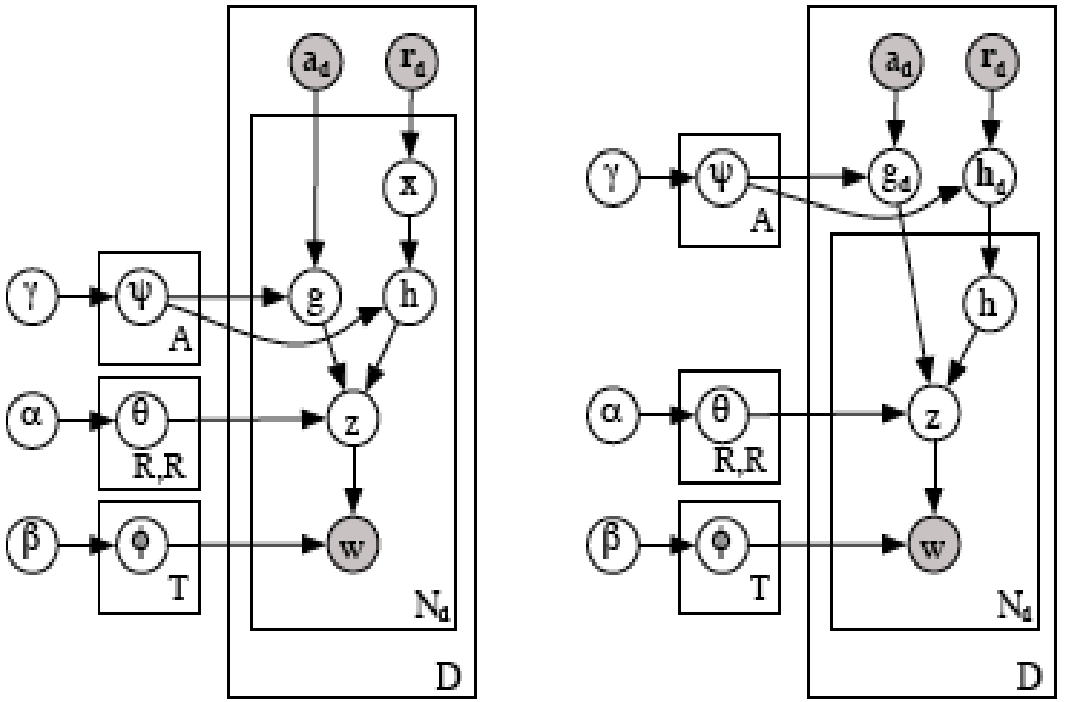
$$P(x_i \,|\, \mathbf{z},\mathbf{x}_{-i},\mathbf{w}) \quad \propto \quad \frac{n_{x_i}^{z_i}+\alpha_{z_i}}{\sum_{z'} n_{x_i}^{z'}+\alpha_{z'}}$$

Results

- Enron emails - 147 users, 23,488 messages
- Author emails - 825 users, 23,488 messages
- Role correlations noted
- Topics form coherent groups
- Performance is better than SNA
- No objective comparison

Future Work

- Role-Author-Recipient Topic Model
- Some preliminary results, but little detail



Conclusion

- The main contribution is a model which can be used to capture author/recipient data in a corpus
- Objective evaluation is difficult
- One potential approach is to split an identity in two and observe which roles the two parts are assigned

Group and Topic Discovery from Relations and Their Attributes

- Neural Information and Processing Systems 2005 Conference
- All authors from University of Massachusetts Amherst

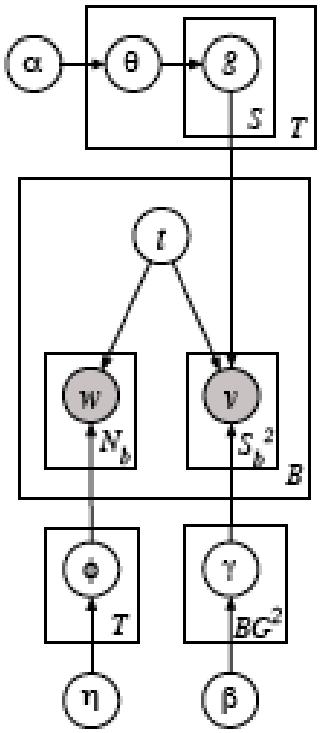
Problem

- Given a body of messages and a set of entities which act on those messages, identify groups among the entities, and topics among the messages
- Voting is the most straightforward example, but their algorithm is flexible enough to allow arbitrary sets of actions on messages

Related Work

- Blockstructures model - Given a graph, assume edge probabilities are determined by latent classes of the vertices
- Role-Author-Recipient-Topic model - Self reference
- Principal Component Analysis model - applied to Senate data for 2003, but no comparison

Model



SYMBOL	DESCRIPTION
g_{st}	entity s's group assignment in topic t
t_b	topic of an event b
$w_k^{(b)}$	the kth token in the event b
$v_{ij}^{(b)}$	entity i and j's group(s) behaved same (1) or differently (2) on the event b
S	# of entities
T	# of topics
G	# of groups
B	# of events
V	# of unique words
N_b	# of word tokens in the event b
S_b	# of entities who participated in the event b

Algorithm

$$P(g_{st}|\mathbf{v}, \mathbf{g}_{-st}, \mathbf{w}, \mathbf{t}, \alpha, \beta, \eta)$$

$$\propto \frac{\alpha_{g_{st}} + n_{t g_{st}} - 1}{\sum_{g=1}^{G} (\alpha_g + n_{t_g}) - 1} \prod_{b=1}^{B} \left(I(t_b = t) \prod_{h=1}^{G} \frac{\prod_{k=1}^{2} \prod_{x=1}^{d^{(b)}_{g_{st}hk}} \left(\beta_k + m^{(b)}_{g_{st}hk} - x\right)}{\prod_{x=1}^{\sum_{k=1}^{2} d^{(b)}_{g_{st}hk}} \left(\left(\sum_{k=1}^{2} (\beta_k + m^{(b)}_{g_{st}hk})\right) - x\right)} \right)$$

First term is general likelihood of group given topic, second term is degree to which proposed group assignment makes groups predict voting

$$P(t_b|\mathbf{v}, \mathbf{g}, \mathbf{w}, \mathbf{t}_{-b}, \alpha, \beta, \eta)$$

$$\propto \frac{\prod_{v=1}^{V} \prod_{x=1}^{e^{(b)}_v} (\eta_v + c_{t_b v} - x)}{\prod_{x=1}^{\sum_{v=1}^{V} e^{(b)}_v} \left(\sum_{v=1}^{V} (\eta_v + c_{t_b v}) - x\right)} \prod_{g=1}^{G} \prod_{h=g}^{G} \frac{\prod_{k=1}^{2} \Gamma(\beta_k + m^{(b)}_{ghk})}{\Gamma(\sum_{k=1}^{2} (\beta_k + m^{(b)}_{ghk}))}$$

First term reflects word probabilities, second term reflects the degree to which the proposed topic assignment makes groups predict voting

Results

- Baseline is find a single topic for each event, and used Blockstructures to find groups within topics
- US Senate voting records, 1989-2004, using votes and index terms
- UN General Assembly Resolutions, 1990-2003
- There is slight improvement over Blockstructures
- They show several improvements in topic generation over the mixture of unigrams model

Conclusion

- Again, evaluation is difficult
- The main contribution is the ability to including voting data in the clustering decisions
- Expert oversight in the groups would have been useful - do the groups found by the algorithm correspond to actual political groups?