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

\[ p(w | \alpha, \beta, a, r) = \int \int p(\theta | \alpha) p(\phi | \beta) \prod_{d=1}^{D} \prod_{n=1}^{N_d} \sum_{x_{dn}} \sum_{z_{dn}} p(x_{dn} | r_d) \]
\[ \cdot p(z_{dn} | \theta_{a_d}, x_{dn}) p(w_{dn} | \phi_{z_{dn}}) d\phi d\theta. \]

- Gibbs Sampling, details left out of paper

\[
P(z_i | z_{-i}, x, w) \propto \frac{n_{wv}^z + \beta^z}{\sum_v n_{wv}^z + \beta^v} \frac{n_{x}^z_i + \alpha^z_i}{\sum_z n_{x}^z_i + \alpha_z}
\]
\[
P(x_i | z, x_{-i}, w) \propto \frac{n_{x}^z_i + \alpha^z_i}{\sum_z n_{x}^z_i + \alpha_z}
\]
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

<table>
<thead>
<tr>
<th>SYMBOL</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>$g_{st}$</td>
<td>entity $s$’s group assignment in topic $t$</td>
</tr>
<tr>
<td>$t_b$</td>
<td>topic of an event $b$</td>
</tr>
<tr>
<td>$w_k^{(b)}$</td>
<td>the $k$th token in the event $b$</td>
</tr>
<tr>
<td>$v_i^{(b)}$</td>
<td>entity $i$ and $j$’s group(s) behaved same (1) or differently (2) on the event $b$</td>
</tr>
<tr>
<td>$S$</td>
<td># of entities</td>
</tr>
<tr>
<td>$T$</td>
<td># of topics</td>
</tr>
<tr>
<td>$G$</td>
<td># of groups</td>
</tr>
<tr>
<td>$B$</td>
<td># of events</td>
</tr>
<tr>
<td>$V$</td>
<td># of unique words</td>
</tr>
<tr>
<td>$N_b$</td>
<td># of word tokens in the event $b$</td>
</tr>
<tr>
<td>$S_b$</td>
<td># of entities who participated in the event $b$</td>
</tr>
</tbody>
</table>
Algorithm

\[ P(g_{st}|v, g_{-st}, w, t, \alpha, \beta, \eta) \]

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

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

\[ P(t_b|v, g, w, t_{-b}, \alpha, \beta, \eta) \]

\[ \propto \frac{\prod_{v=1}^{V} \prod_{b=1}^{e(v)} (\eta_v + c_{tbv} - x)}{\prod_{x=1}^{V} e(v)} \prod_{g=1}^{G} \prod_{h=g}^{G} \frac{\prod_{k=1}^{2} \Gamma(\beta_k + m^{(b)}_{ghk})}{\Gamma((\sum_{k=1}^{2} (\beta_k + m^{(b)}_{ghk})))} \]

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