Joint Group and Topic Discovery from Relations and Text

Andrew McCallum, Xuerui Wang, and Natasha Mohanty

University of Massachusetts, Amherst, MA 01003, USA
{mccallum,xuerui,nmohanty}@cs.umass.edu

Abstract. We present a probabilistic generative model of entity relationships and textual attributes; the model simultaneously discovers groups among the entities and topics among the corresponding text. Block models of relationship data have been studied in social network analysis for some time, however here we cluster in multiple modalities at once. Significantly, joint inference allows the discovery of groups to be guided by the emerging topics, and vice-versa. We present experimental results on two large data sets: sixteen years of bills put before the U.S. Senate, comprising their corresponding text and voting records, and 43 years of similar data from the United Nations. We show that in comparison with traditional, separate latent-variable models for words or block structures for votes, our Group-Topic model's joint inference improves both the groups and topics discovered. Additionally, we present a non-Markov continouous-time group model to capture shifting group structure over time.

1 Introduction

Research in the field of social network analysis (SNA) has led to the development of mathematical models that discover patterns in interaction between entities [1, 2, 3]. One of the objectives of SNA is to detect salient groups of entities. Group discovery has many applications, such as understanding the social structure of organizations [4] or native tribes [5], uncovering criminal organizations [6], and modeling large-scale social networks in Internet services such as Friendster.com or LinkedIn.com.

Social scientists have conducted extensive research on group detection, especially in fields such as anthropology [5] and political science [7, 8]. Recently, statisticians and computer scientists have begun to develop models that specifically discover group memberships [9, 10, 11, 12]. One such model is the stochastic block structures model [11], which discovers the latent structure, groups or classes based on pair-wise relation data. A particular relation holds between a pair of entities (people, countries, organizations, etc.) with some probability that depends only on the class (group) assignments of the entities. The relations between all the entities can be represented with a directed or undirected graph. The class assignments can be inferred from a graph of observed relations or link data using Gibbs sampling [11]. This model is extended in [12] to automatically

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2006		**00-00-2006 to 00-00-2006**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Joint Group and Topic Discovery from Relations and Text	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Massachusetts Amherst,Department of Computer Science,Amherst,MA,01003**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

We present a probabilistic generative model of entity relationships and textual attributes; the model simultaneously discovers groups among the entities and topics among the corresponding text. Block models of relationship data have been studied in social network analysis for some time, however here we cluster in multiple modalities at once. Significantly, joint inference allows the discovery of groups to be guided by the emerging topics, and vice-versa. We present experimental results on two large data sets: sixteen years of bills put before the U.S. Senate, comprising their corresponding text and voting records, and 43 years of similar data from the United Nations. We show that in comparison with traditional, separate latent-variable models for words or block structures for votes, our Group-Topic model's joint inference improves both the groups and topics discovered. Additionally, we present a non-Markov continouous-time group model to capture shifting group structure over time.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**17**	

select an arbitrary number of groups by using a Chinese Restaurant Process prior.

The aforementioned models discover latent groups only by examining whether one or more relations exist between a pair of entities. The Group-Topic (GT) model presented in this paper, on the other hand, considers not only the relations between objects but also the attributes of the relations (for example, the text associated with the relations) when assigning group membership.

The GT model can be viewed as an extension of the stochastic block structures model [11, 12] with the key addition that group membership is conditioned on a latent variable associated with the attributes of the relation. In our experiments, the attributes of relations are words, and the latent variable represents the topic responsible for generating those words. Unlike previous methods, our model captures the *(language) attributes* associated with interactions between entities, and uses distinctions based on these attributes to better assign group memberships.

Consider a legislative body and imagine its members forging alliances (forming groups), and voting accordingly. However, different alliances arise depending on the topic of the resolution up for a vote. For example, one grouping of the legislators may arise on the issue of taxation, while a quite different grouping may occur for votes on foreign trade. Similar patterns of topic-based affiliations would arise in other types of entities as well, e.g., research paper co-authorship relations between people and citation relations between papers, with words as attributes on these relations.

In the GT model, the discovery of groups is guided by the emerging topics, and the discovery of topics is guided by emerging groups. Both modalities are driven by the common goal of increasing data likelihood. Consider the voting example again; resolutions that would have been assigned the same topic in a model using words alone may be assigned to different topics if they exhibit distinct voting patterns. Distinct word-based topics may be merged if the entities vote very similarly on them. Likewise, multiple different divisions of entities into groups are made possible by conditioning them on the topics.

The importance of modeling the *language* associated with interactions between people has recently been demonstrated in the Author-Recipient-Topic (ART) model [13]. In ART the words in a message between people in a network are generated conditioned on the author, recipients and a set of topics that describes the message. The model thus captures both the network structure within which the people interact as well as the language associated with the interactions. In experiments with Enron and academic email, the ART model is able to discover role similarity of people better than SNA models that consider network connectivity alone. However, the ART model does not explicitly capture groups formed by entities in the network.

The GT model simultaneously clusters entities to groups and clusters words into topics, unlike models that generate topics solely based on word distributions such as Latent Dirichlet Allocation [14]. In this way the GT model discovers

SYMBOL	DESCRIPTION
g_{it}	entity i'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 groups behaved same (1) or differently (2) on the event b
S	number of entities
T	number of topics
G	number of groups
B	number of events
V	number of unique words
N_b	number of word tokens in the event b
S_b	number of entities who participated in the event b

Table 1. Notation used in this paper

salient topics relevant to relationships between entities in the social network—topics which the models that only examine words are unable to detect.

We demonstrate the capabilities of the GT model by applying it to two large sets of voting data: one from US Senate and the other from the General Assembly of the UN. The model clusters voting entities into coalitions and simultaneously discovers topics for word attributes describing the relations (bills or resolutions) between entities. We find that the groups obtained from the GT model are significantly more cohesive (p-value $< .01$) than those obtained from the block structures model. The GT model also discovers new and more salient topics in both the Senate and UN datasets—in comparison with topics discovered by only examining the words of the resolutions, the GT topics are either split or joined together as influenced by the voters' patterns of behavior.

2 Group-Topic Model

The Group-Topic Model is a directed graphical model that clusters entities with relations between them, as well as attributes of those relations. The relations may be either directed or undirected and have multiple attributes. In this paper, we focus on undirected relations and have words as the attributes on relations.

In the generative process for each event (an interaction between entities), the model first picks the topic t of the event and then generates all the words describing the event where each word is generated independently according to a multinomial (discrete) distribution ϕ_t, specific to the topic t. To generate the relational structure of the network, first the group assignment, g_{st} for each entity s is chosen conditionally from a particular multinomial (discrete) distribution θ_t over groups for each topic t. Given the group assignments on an event b, the matrix $V^{(b)}$ is generated where each cell $V_{ij}^{(b)}$ represents if the groups of two entities (i and j) behaved the same or not during the event b, (e.g., voted the same or not on a bill). Each element of V is sampled from a binomial (Bernoulli)

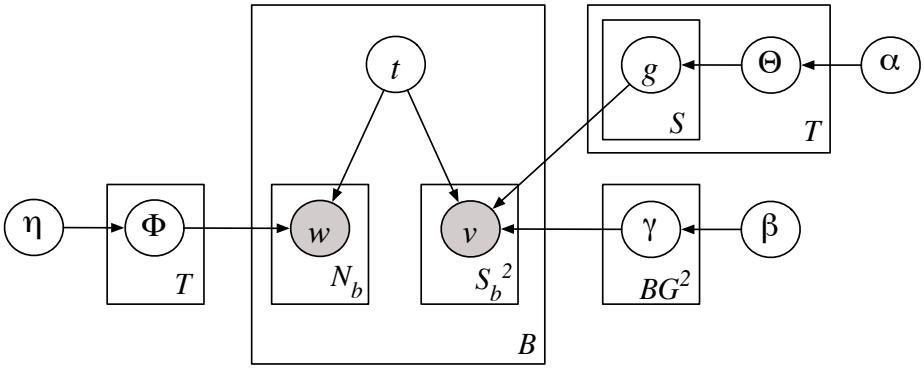


Fig. 1. The Group-Topic model

distribution $\gamma^{(b)}_{g_i g_j}$. Our notation is summarized in Table 1, and the graphical model representation of the model is shown in Figure 1.

Without considering the topic of an event, or by treating all events in a corpus as reflecting a single topic, the simplified model (only the right part of Figure 1) becomes equivalent to the stochastic block structures model [11]. To match the block structures model, each event defines a relationship, *e.g.*, whether in the event two entities' groups behave the same or not. On the other hand, in our model a relation may have multiple attributes (which in our experiments are the words describing the event, generated by a per-topic multinomial (discrete) distribution).

When we consider the complete model, the dataset is dynamically divided into T sub-blocks each of which corresponds to a topic. The complete GT model is as follows,

$$
\begin{aligned}
t_b &\sim \text{Uniform}(\frac{1}{T}) \\
w_{it}|\phi_t &\sim \text{Multinomial}(\phi_t) \\
\phi_t|\eta &\sim \text{Dirichlet}(\eta) \\
g_{it}|\theta_t &\sim \text{Multinomial}(\theta_t) \\
\theta_t|\alpha &\sim \text{Dirichlet}(\alpha) \\
V^{(b)}_{ij}|\gamma^{(b)}_{g_i g_j} &\sim \text{Binomial}(\gamma^{(b)}_{g_i g_j}) \\
\gamma^{(b)}_{gh}|\beta &\sim \text{Beta}(\beta).
\end{aligned}
$$

We want to perform joint inference on (text) attributes and relations to obtain topic-wise group memberships. Since inference can not be done exactly on such complicated probabilistic graphical models, we employ Gibbs sampling to conduct inference. Note that we adopt conjugate priors in our setting, and thus we can easily integrate out θ, ϕ and γ to decrease the uncertainty associated with them. This simplifies the sampling since we do not need to sample θ, ϕ

and γ at all, unlike in [11]. In our case we need to compute the conditional distribution $P(g_{st}|\mathbf{w}, \mathbf{V}, \mathbf{g}_{-st}, \mathbf{t}, \alpha, \beta, \eta)$ and $P(t_b|\mathbf{w}, \mathbf{V}, \mathbf{g}, \mathbf{t}_{-b}, \alpha, \beta, \eta)$, where $\mathbf{g}_{-st}$ denotes the group assignments for all entities except entity s in topic t, and $\mathbf{t}_{-b}$ represents the topic assignments for all events except event b. Beginning with the joint probability of a dataset, and using the chain rule, we can obtain the conditional probabilities conveniently. In our setting, the relationship we are investigating is always symmetric, so we do not distinguish R_{ij} and R_{ji} in our derivations (only $R_{ij}(i \leq j)$ remain). Thus

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

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

where n_{tg} represents how many entities are assigned into group g in topic t, c_{tv} represents how many tokens of word v are assigned to topic t, $m^{(b)}_{ghk}$ represents how many times group g and h vote same ($k = 1$) and differently ($k = 2$) on event b, $I(t_b = t)$ is an indicator function, and $d^{(b)}_{g_{st}hk}$ is the increase in $m^{(b)}_{g_{st}hk}$ if entity s were assigned to group g_{st} than without considering s at all (if $I(t_b = t) = 0$, we ignore the increase in event b). Furthermore,

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

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

where $e^{(b)}_v$ is the number of tokens of word v in event b. Note that $m^{(b)}_{ghk}$ is not a constant and changes with the assignment of t_b since it influences the group assignments of all entities that vote on event b. We use a weighting parameter λ to rescale the likelihoods from different modalities, as is also common in speech recognition when the acoustic and language models are combined. The GT model uses information from two different modalities. In general, the likelihood of the two modalities is not directly comparable, since the number of occurrences of each type may vary greatly (e.g., there may be far more pairs of voting entities than word occurrences).

3 Related Work

There has been a surge of interest in models that describe relational data, or relations between entities viewed as links in a network, including recent work in group discovery. One such algorithm, presented by Bhattacharya and Getoor [10], is a bottom-up agglomerative clustering algorithm that partitions links in a network into clusters by considering the change in likelihood that would occur if two clusters were merged. Once the links have been grouped, the entities connected by the links are assigned to groups.

Another model due to Kubica et al. [9] considers both link evidence and attributes on entities to discover groups. The Group Detection Algorithm (GDA) uses a Bayesian network to group entities from two datasets, demographic data describing the entities and link data. Unlike our model, neither of these models [10, 9] consider attributes associated with the links between the entities. The model presented in [9] considers attributes of an entity rather than attributes of relations between entities.

The central theme of GT is that it simultaneously clusters entities and attributes on relations (words). There has been prior work in clustering different entities simultaneously, such as information theoretic co-clustering [15], and multi-way distributional clustering using pair-wise interactions [16]. However, these models do not also cluster attributes based on interactions between entities in a network.

In our model, group membership defines pair-wise relations between nodes. The GT model is an enhancement of the stochastic block structures model [11] and the extended model of Kemp et al. [12] as it takes advantage of information from different modalities by conditioning group membership on topics. In this sense, the GT model draws inspiration from the Role-Author-Recipient-Topic (RART) model [13]. As an extension of ART model, RART clusters together entities with similar roles. In contrast, the GT model presented here clusters entities into groups based on their relations to other entities.

Exploring the notion that the behavior of an entity can be explained by its (hidden) group membership, Jakulin and Buntine [17] develop a discrete PCA model for discovering groups. In the model each entity can belong to each of the k groups with a certain probability, and each group has its own specific pattern of behaviors. Therefore, an entity's behavior depends on the probability of belonging to a group and the probability that the group has that behavior. They apply this model to voting data in the 108th US Senate where the behavior of an entity is its vote on a resolution. A similar model is developed in [18] that examines group cohesion and voting similarity in the Finnish Parliament. We apply our GT model also to voting data. However, unlike [17, 18], since our goal is to cluster entities based on the similarity of their voting patterns, we are only interested in whether a pair of entities voted the same or differently, not their actual yes/no votes. Two resolutions on the same topic may differ only in their goal (e.g., increasing vs. decreasing budget), thus the actual votes on one could be the converse of votes on the other. However, pairs of entities who vote the same on one resolution would tend to vote same on the other resolution. To capture this, we model relations as *agreement* between entities, not the yes/no vote itself. This kind of "content-ignorant" feature is similarly found in some work on web log clustering [19].

There has been a considerable amount of previous work in understanding voting patterns [20, 7, 8], including research on voting cohesion of countries in the EU parliament [7] and partisanship in roll call voting [8]. In these models roll call data are used to estimate *ideal points* of a legislator (which refers to a legislator's preferred policy in the Euclidean space of possible policies). The models assume

Datasets	Avg. AI for GT	Avg. AI for Baseline	p-value	Block Structures
Senate	0.8294	0.8198	$< .01$	0.7850
UN	0.8664	0.8548	$< .01$	0.7934

Table 2. Average AI for different models for both Senate and UN datasets. The group cohesion in (joint) GT is significantly better than in (serial) baseline, as well as the block structures model that does not use text at all.

that each vote in the roll call data is independent of the remaining votes, i.e., each individual is not connected to anyone else who is voting. However, in reality, legislation is shaped by the coalitions formed by like-minded legislators. The GT model attempts to capture this interaction.

4 Experimental Results

We present experiments applying the GT model to the voting records of members of two legislative bodies: the US Senate and the UN General Assembly. We set $\alpha = 1$, $\beta = 5$, and $\eta = 1$ in all experiments. To make sure of convergence, we run the Markov chains for 10,000 iterations, (which by inspection are stable after a few hundred iterations), and use the topic and group assignments in the last Gibbs sample.

For comparison, we present the results of a baseline method that first uses a mixture of unigrams to discover topics and associate a topic with each resolution, and then runs the block structures model [11] separately on the resolutions assigned to each topic. This baseline approach is similar to the GT model in that it discovers both groups and topics, and has different group assignments on different topics. However, whereas the GT model performs joint inference simultaneously, the baseline performs inference serially. Note that our baseline is still more powerful than the block structures models, since it models the topic associated with each event, and allows the creation of distinct groupings dependent on different topics.

In this paper, we are interested in the quality of both the groups and the topics. In the political science literature, group cohesion is quantified by the *Agreement Index (AI)* [17, 18], which measures the similarity of votes cast by members of a group during a particular roll call. The AI for a particular group on a given roll call i is based on the number of group members that vote Yes(y_i), No(n_i) or Abstain(a_i) in the roll call i. Higher AI index means better cohesion.

$$AI_i = \frac{\max\{y_i, n_i, a_i\} - \frac{y_i+n_i+a_i-\max\{y_i,n_i,a_i\}}{2}}{y_i + n_i + a_i}$$

The block structures model assumes that members of a legislative body have the same group affiliations irrespective of the topic of the resolution on vote. However, it is likely that members form their groups based on the topic of the resolution being voted on. We quantify the extent to which a member s switches

Economic	Education	Military Misc.	Energy
federal	education	government	energy
labor	school	military	power
insurance	aid	foreign	water
aid	children	tax	nuclear
tax	drug	congress	gas
business	students	aid	petrol
employee	elementary	law	research
care	prevention	policy	pollution

Table 3. Top words for topics generated with the mixture of unigrams model on the Senate dataset. The headers are our own summary of the topics.

Economic	Education + Domestic	Foreign	Social Security + Medicare
labor	education	foreign	social
insurance	school	trade	security
tax	federal	chemicals	insurance
congress	aid	tariff	medical
income	government	congress	care
minimum	tax	drugs	medicare
wage	energy	communicable	disability
business	research	diseases	assistance

Table 4. Top words for topics generated with the GT model on the Senate dataset. The topics are influenced by both the words and votes on the bills.

groups with a *Group Switch Index* (GSI).

$$GSI_s = \sum_{i,j}^{T} \frac{\text{abs}(\boldsymbol{s_i} - \boldsymbol{s_j})}{|G(s,i)| - 1 + |G(s,j)| - 1}$$

where $\boldsymbol{s_i}$ and $\boldsymbol{s_j}$ are bit vectors of the length of the size of the legislative body. The k_{th} bit of $\boldsymbol{s_i}$ is set if k is in the same group as s on topic i and similarly $\boldsymbol{s_j}$ corresponds to topic j. $G(s,i)$ is the group of s on topic i which has a size of $|G(s,i)|$ and $G(s,j)$ is the group of s on topic j. We present entities that frequently change their group alliance according to the topics of resolutions.

Group cohesion from the GT model is found to be significantly greater than the baseline group cohesion under a pairwise t-test, as shown in Table 2, which indicates that the GT's joint inference is better able to discover cohesive groups. We find that nearly every document has a higher Agreement Index across groups using the GT model as compared to the baseline. As expected, stochastic block structures without text [11] is even worse than our baseline.

4.1 The US Senate Dataset

Our Senate dataset consists of the voting records of Senators in the 101st-109th US Senate (1989-2005) obtained from the Library of Congress THOMAS

Group 1	Group 3	Group 4
73 Republicans	Cohen(R-ME)	Armstrong(R-CO)
Krueger(D-TX)	Danforth(R-MO)	Garn(R-UT)
Group 2	Durenberger(R-MN)	Humphrey(R-NH)
90 Democrats	Hatfield(R-OR)	McCain(R-AZ)
Chafee(R-RI)	Heinz(R-PA)	McClure(R-ID)
Jeffords(I-VT)	Kassebaum(R-KS)	Roth(R-DE)
	Packwood(R-OR)	Symms(R-ID)
	Specter(R-PA)	Wallop(R-WY)
	Snowe(R-ME)	Brown(R-CO)
	Collins(R-ME)	DeWine(R-OH)
		Thompson(R-TN)
		Fitzgerald(R-IL)
		Voinovich(R-OH)
		Miller(D-GA)
		Coleman(R-MN)

Table 5. Senators in the four groups corresponding to Topic Education + Domestic in Table 4.

Senator	Group Switch Index
Shelby(D-AL)	0.6182
Heflin(D-AL)	0.6049
Voinovich(R-OH)	0.6012
Johnston(D-LA)	0.5878
Armstrong(R-CO)	0.5747

Table 6. Senators that switch groups the most across topics for the 101st-109th Senates

database. During a roll call for a particular bill, a Senator may respond *Yea* or *Nay* to the question that has been put to vote, else the vote will be recorded as *Not Voting*. We do not consider *Not Voting* as a unique vote since most of the time it is a result of a Senator being absent from the session of the US Senate. The text associated with each resolution is composed of its index terms provided in the database. There are 3423 resolutions in our experiments (we excluded roll calls that were not associated with resolutions). Each bill may come up for vote many times in the U.S. Senate, each time with an attached amendment, and thus many relations may have the same attributes (index terms). Since there are far fewer words than pairs of votes, we adjust the text likelihood to the 5th power (weighting factor 5) in the experiments with this dataset so as to balance its influence during inference.

We cluster the data into 4 topics and 4 groups (cluster sizes are suggested by a political science professor) and compare the results of GT with the baseline. The most likely words for each topic from the traditional mixture of unigrams model is shown in Table 3, whereas the topics obtained using GT are shown in Table 4. The GT model collapses the topics Education and Energy together into Education and Domestic, since the voting patterns on those topics are quite similar. The new topic Social Security + Medicare did not have strong enough word coherence to appear in the baseline model, but it has a very distinct voting

Everything Nuclear	Human Rights	Security in Middle East
nuclear	rights	occupied
weapons	human	israel
use	palestine	syria
implementation	situation	security
countries	israel	calls

Table 7. Top words for topics generated from mixture of unigrams model with the UN dataset (1990-2003). Only text information is utilized to form the topics, as opposed to Table 8 where our GT model takes advantage of both text and voting information.

GROUP ↓	Nuclear Arsenal	Human Rights	Nuclear Arms Race
	nuclear	rights	nuclear
	states	human	arms
	united	palestine	prevention
	weapons	occupied	race
	nations	israel	space
1	Brazil	Brazil	UK
	Columbia	Mexico	France
	Chile	Columbia	Spain
	Peru	Chile	Monaco
	Venezuela	Peru	East-Timor
2	USA	Nicaragua	India
	Japan	Papua	Russia
	Germany	Rwanda	Micronesia
	UK...	Swaziland	
	Russia	Fiji	
3	China	USA	Japan
	India	Japan	Germany
	Mexico	Germany	Italy...
	Iran	UK...	Poland
	Pakistan	Russia	Hungary
4	Kazakhstan	China	China
	Belarus	India	Brazil
	Yugoslavia	Indonesia	Mexico
	Azerbaijan	Thailand	Indonesia
	Cyprus	Philippines	Iran
5	Thailand	Belarus	USA
	Philippines	Turkmenistan	Israel
	Malaysia	Azerbaijan	Palau
	Nigeria	Uruguay	
	Tunisia	Kyrgyzstan	

Table 8. Top words for topics generated from the GT model with the UN dataset (1990-2003) as well as the corresponding groups for each topic (column). The countries listed for each group are ordered by their 2005 GDP (PPP) and only the top 5 countries are shown in groups that have more than 5 members.

pattern, and thus is clearly found by the GT model. Thus GT discovers topics that are salient in that they correlate with people's behavior and relations, not simply word co-occurrences.

Examining the group distribution across topics in the GT model, we find that on the topic Economic the Republicans form a single group whereas the Democrats split into 3 groups indicating that Democrats have been somewhat divided on this topic. With regard to Education + Domestic and Social Security + Medicare, Democrats are more unified whereas the Republicans split into 3 groups. The group membership of Senators on Education + Domestic issues is shown in Table 5. We see that the first group of Republicans include a Democratic Senator from Texas, a state that usually votes Republican. Group 2 (majority Democrats) includes Sen. Chafee who is known to be pro-environment and is involved in initiatives to improve education, as well as Sen. Jeffords who left the Republican Party to become an Independent and has championed legislation to strengthen education and environmental protection.

Nearly all the Senators in Group 4 (in Table 5) are advocates for education and many of them have been awarded for their efforts (e.g., Sen. Fitzgerald has been honored by the NACCP for his active role in Early Care and Education, and Sen. McCain has been added to the ASEE list as a *True Hero* in American Education). Sen. Armstrong was a member of the Education committee; Sen. Voinovich and Sen. Symms are strong supporters of early education and vocational education, respectively; and Sen. Roth has constantly voted for tax deductions for education. It is also interesting to see that Sen. Miller (D-GA) appears in a Republican group; although he is in favor of educational reforms, he is a conservative Democrat and frequently criticizes his own party—even backing Republican George W. Bush over Democrat John Kerry in the 2004 Presidential election.

Many of the Senators in Group 3 have also focused on education and other domestic issues such as energy, however, they often have a more liberal stance than those in Group 4, and come from states that are historically less conservative. Senators Hatfield, Heinz, Snowe, Collins, Cohen and others have constantly promoted pro-environment energy options with a focus on renewable energy, while Sen. Danforth has presented bills for a more fair distribution of energy resources. Sen. Kassebaum is known to be uncomfortable with many Republican views on domestic issues such as education, and has voted against voluntary prayer in school. Thus, both Groups 3 and 4 differ from the Republican core (Group 2) on domestic issues, and also differ from each other.

The Senators that switch groups the most across topics in the GT model are shown in Table 6 based on their GSIs. Sen. Shelby(D-AL) votes with the Republicans on Economic, with the Democrats on Education + Domestic and with a small group of maverick Republicans on Foreign and Social Security + Medicare. Both Sen. Shelby and Sen. Heflin are Democrats from a fairly conservative state (Alabama) and are found to side with the Republicans on many issues.

4.2 The United Nations Dataset

The second dataset involves the voting record of the UN General Assembly [21]. We focus first on the resolutions discussed from 1990-2003, which contain votes of 192 countries on 931 resolutions. If a country is present during the roll call, it may choose to vote *Yes*, *No* or *Abstain*. Unlike the Senate dataset, a country's vote can have one of three possible values instead of two. Because we parameterize agreement and not the votes themselves, this 3-value setting does not require any change to our model. In experiments with this dataset, we use a weighting factor 500 for text (adjusting the likelihood of text by a power of 500 so as to make it comparable with the likelihood of pairs of votes for each resolution). We cluster this dataset into 3 topics and 5 groups (again, numbers are suggested by a political science professor).

The most probable words in each topic from the mixture of unigrams model is shown in Table 7. For example, Everything Nuclear constitutes all resolutions that have anything to do with the use of nuclear technology, including nuclear weapons. Comparing these with topics generated from the GT model shown in Table 8, we see that the GT model splits the discussion about nuclear technology into two separate topics, Nuclear Arsenal which is generally about countries obtaining nuclear weapons and management of nuclear waste, and Nuclear Arms Race which focuses on the arms race between Russia and the US and preventing a nuclear arms race in outer space. These two issues had drastically different voting patterns in the U.N., as can be seen in the contrasting group structure for those topics in Table 8. The countries in Table 8 are ranked by their GDP in 2005.[1] Thus, again the GT model is able to discover salient topics—topics that reflect the voting patterns and coalitions, not simply word co-occurrence alone.

As seen in Table 8, groups formed in Nuclear Arms Race are unlike the groups formed in the remaining topics. These groups map well to the global political situation of that time when, despite the end of the Cold War, there was mutual distrust between Russia and the US with regard to the continued manufacture of nuclear weapons. For missions to outer space and nuclear arms, India was a staunch ally of Russia, while Israel was an ally of the US.

Overlapping Time Intervals In order to understand changes and trends in topics and groups over time, we run the GT model on resolutions that were discussed during overlapping time windows of 15 years, from 1960-2000, each shifted by a period of 5 years. We consider 3823 unique resolutions in this way. The topics as well as the group distribution for the most dominant topic during each time period are shown in Table 9.

Over the years there is a shift in the topics discussed in the UN, which corresponds well to the events and issues in history. During 1960-1975 the resolutions focused on countries having the right to self-determination, especially countries

[1] http://en.wikipedia.org/wiki/List_of_countries_by_GDP_%28PPP%29. In Table 8, we omit some countries (represented by ...) in order to incorporate other interesting but relatively low ranked countries (for example, Russia) in the GDP list.

Time Period	Topic 1	Topic 2	Topic 3	Group distributions for Topic 3				
				Group 1	Group2	Group3	Group4	Group5
60-75	Nuclear	Procedure	Africa Indep.	India	USA	Argentina	USSR	Turkey
	operative	committee	calling	Indonesia	Japan	Colombia	Poland	
	general	amendment	right	Iran	UK	Chile	Hungary	
	nuclear	assembly	africa	Thailand	France	Venezuela	Bulgaria	
	power	deciding	self	Philippines	Italy	Dominican	Belarus	
65-80	Independence	Finance	Weapons	Cuba	India	Algeria	USSR	USA
	territories	budget	nuclear	Albania	Indonesia	Iraq	Poland	Japan
	independence	appropriation	UN		Pakistan	Syria	Hungary	UK
	self	contribution	international		Saudi	Libya	Bulgaria	France
	colonial	income	weapons		Egypt	Afghanistan	Belarus	Italy
70-85	N. Weapons	Israel	Rights	Mexico	China	USA	Brazil	India
	nuclear	israel	africa	Indonesia		Japan	Turkey	USSR
	international	measures	territories	Iran		UK	Argentina	Poland
	UN	hebron	south	Thailand		France	Colombia	Vietnam
	human	expelling	right	Philippines		Italy	Chile	Hungary
75-90	Rights	Israel/Pal.	Disarmament	Mexico	USA	Algeria	China	India
	south	israel	UN	Indonesia	Japan	Vietnam	Brazil	
	africa	arab	international	Iran	UK	Iraq	Argentina	
	israel	occupied	nuclear	Thailand	France	Syria	Colombia	
	rights	palestine	disarmament	Philippines	USSR	Libya	Chile	
80-95	Disarmament	Conflict	Pal. Rights	USA	China	Japan	Guatemala	Malawi
	nuclear	need	rights	Israel	India	UK	St Vincent	
	US	israel	palestine		Russia	France	Dominican	
	disarmament	palestine	israel		Spain	Italy		
	international	secretary	occupied		Hungary	Canada		
85-00	Weapons	Rights	Israel/Pal.	Poland	China	USA	Russia	Cameroon
	nuclear	rights	israeli	Czech R.	India	Japan	Argentina	Congo
	weapons	human	palestine	Hungary	Brazil	UK	Ukraine	Ivory C.
	use	fundamental	occupied	Bulgaria	Mexico	France	Belarus	Liberia
	international	freedoms	disarmament	Albania	Indonesia	Italy	Malta	

Table 9. Results for 15-year-span slices of the UN dataset (1960-2000). The top probable words are listed for all topics, but only the groups corresponding the most dominant topic are shown (Topic 3). We list the countries for each group ordered by their 2005 GDP (PPP)and only show the top 5 countries in groups that have more than 5 members. We do not repeat the results in Table 8 for the most recent window (1990-2003).

in Africa which started to gain their freedom during this time. Although this topic continued to be discussed in the subsequent time period, the focus of the resolutions shifted to the role of the UN in controlling nuclear weapons as the Cold War conflict gained momentum in the late 70s. While there were few resolutions condemning the racist regime in South Africa between 1965-1980, this was the topic of many resolutions during 1970-1985—culminating in the UN censure of South Africa for its discriminatory practices.

Other topics discussed during the 70s and early 80s were Israel's occupation of neighboring countries and nuclear issues. The reduction of arms was primarily discussed during 1975-1990, the time period during which the US and Soviet Union had talks about disarmament. During 1980-1995 the central topic of discussion was the Israeli-Palestinian conflict; this time period includes the beginning of the *Intifada* revolt in Palestine and the Gulf War. This topic continued to be important in the next time period (1985-2000), but in the most recent slice (1990-2003, Table 8) it has become a part of a broader topic on human rights by combining other human rights related resolutions that appear as a separate topic during 1985-2000. The human rights issue continues to be the primary topic of discussion during 1990-2003.

Throughout the history of the UN, the US is usually in the same group as Europe and Japan. However, as we can see in Table 9 during 1985-2000, when the Israeli-Palestinian conflict was the most dominant topic, US and Israel form a group of their own separating themselves from Europe. In other topics discussed

during 1985-2000, US and Israel are found to be in the same group as Europe and Japan.

Another interesting result of considering the groups formed over the years is that, except for the last time period (1990-2003), countries in eastern Europe such as Poland, Hungary, Bulgaria, etc., form a group along with USSR (Russia). However, in the last time window on most topics they become a part of the group that consists of the western Europe, Japan and the US. This shift corresponds to the end of the communist regimes in these countries that were supported by the Soviet Union. It is also worth mentioning that before 1990, our model assigned East Germany to the same group as other eastern European countries and USSR (Russia), while it assigned West Germany to the same group as western European countries.[2]

5 Groups over Time

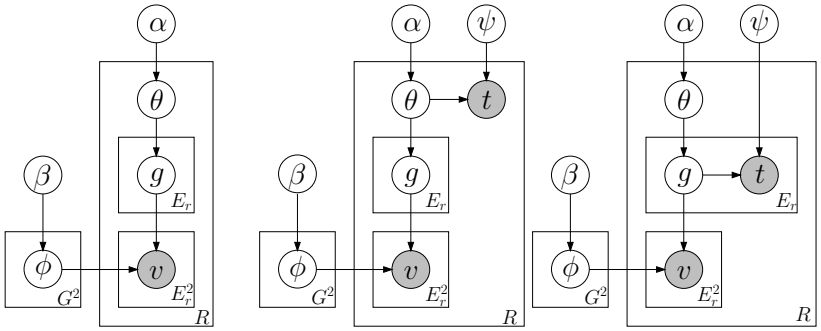
In Table 9 in the previous section, we show that group formation changes over time by simply pre-dividing the dataset into disjoint subsets based on timestamps. In this section, by contrast, we investigate the dynamic changes of groups using a model that explicitly incorporates time—jointly discovering groups and their continuous time profiles.

Traditional transition-based Markov models have played a major role in modeling various dynamic systems including social networks. For example, recent work by Sarkar and Moore [22] proposes a latent space model that accounts for friendships drifting over time. Blei and Lafferty propose dynamic topic models (DTMs) in which the alignment among topics across time steps is captured by a Kalman filter [23].

Instead we propose here a new model that does not make the Markov assumption, rather, we treat timestamps as observed random variables, as in [24]. In comparison to more complex alternatives, the relative simplicity of our model is a great advantage—not only for the relative ease of understanding and implementing it, but also because this approach can in the future be naturally applied into other more richly structured models. In our model, which we call *Groups over Time (GOT)*, group discovery is influenced not only by relational co-occurrences, but also by temporal information. Rather than modeling a sequence of state changes with a Markov assumption on the dynamics, GOT models (normalized) absolute timestamp values. This allows GOT to see long-range dependencies in time, and to predict group distributions given a timestamp. It also helps avoid a Markov model's risk of inappropriately dividing a group in two when there is a brief gap in its appearance.

The graphical model representation of our model is shown in Figure 2. For comparison, the stochastic block structures model is shown in Figure 2(a). Groups over Time is a generative model of timestamps and the relational structures of a social network. There are two ways of describing its generative process.

[2] This is not shown in Table 9 because they are missing from the 2005 GDP data.



(a) Multi-relation stochastic block structures (b) GOT model, alternate view (c) GOT model, for Gibbs sampling

Fig. 2. Three group models: stochastic block structures and two perspectives on GOT

The first, which corresponds to the process used in Gibbs sampling for parameter estimation (Figure 2(c)), is as follows:

1. Draw G^2 binomials ϕ_{gh} from a Dirichlet prior β, one for each group pair (g, h);
2. For each relation r in total R relations, draw a multinomial θ_r from a Dirichlet prior α;
   (a) For each entity e_{ri} of E_r entities in relation r, draw a group g_{ri} from multinomial θ_r and draw a timestamp t_{ri} from Beta $\psi_{g_{ri}\cdot}$;
   (b) For each entity pair (i, j), draw their agreement v_{ij} from binomial $\phi_{g_i g_j}$;

Although, in the above generative process, a timestamp is generated for each entity, all the timestamps of the entities in a relation are observed as the same, because there is typically only one timestamp associated with a relation. When fitting our model from typical data, each training relation's timestamp is copied to all the entities appearing in the relation. However, after fitting, if actually run as a generative model, this process would generate different time stamps for the entities appearing in the same relation. An alternative generative process description of GOT, (better suited to generate an unseen relation), is one in which a single timestamp is associated with each relation, generated by rejection or importance sampling, from a mixture of per-group Beta distributions over time with mixtures weight as the per-relation θ_r over groups. As before, this distribution over time would be parameterized by the set of timestamp-generating Beta distributions, one per group. The graphical model for this alternative generative process is shown in Figure 2(b).

5.1 Dynamic Group Discovery in UN

We apply the Group over Time (GOT) model to the UN data set described in Section 4.2, and compare it with the stochastic block structures model. Because of space limitation, we do not show example group distributions from the two

models. However, when we calculate the Agreement Index (AI, defined in Section 4) of the groups discovered by the two models on the UN data set; we find that the average AI for the stochastic block structures model is 0.7934, and 0.8169 for GOT. We conclude that the groups obtained from the GOT model are significantly more cohesive (p-value $< .01$) than those obtained from the block structures model. Note that the average AI from GOT (0.8169) is also lower than the one from GT (0.8664) due to the lack of textual attributes.

6 Conclusions

We present the Group-Topic model that jointly discovers latent groups in a network as well as clusters of attributes (or topics) of events that influence the interaction between entities in the network. The model extends prior work on latent group discovery by capturing not only pair-wise relations between entities but also multiple attributes of the relations (in particular, the model considers words describing the relations). In this way the GT model obtains more cohesive groups as well as fresh topics that influence the interaction between groups. The model could be applied to variables of other data types in addition to voting data. We are now using the model to analyze the citations in academic papers to capture the topics of research papers and discover research groups. It would also apply to a much larger network of entities (people, organizations, etc.) that frequently appear in newswire articles.

The model can be altered suitably to consider other attributes characterizing relations between entities in a network. In ongoing work we are extending the Group-Topic model to capture a richer notion of topic, where the attributes describing the relations between entities are represented by a mixture of topics.

The Group over Time model provides a simple and novel way to take advantage of the temporal information as continuous observations, in contrast to the traditional transition-based Markov models. We believe that the simplicity of this approach is an advantage in certain applications.

7 Acknowledgments

This work was supported in part by the National Science Foundation under grant #IIS-0326249, and by the Defense Advanced Research Projects Agency under contract #NBCHD030010 and #HR0011-06-C-0023. We would also like to greatly thank Prof. Vincent Moscardelli, Chris Pal and Aron Culotta for helpful discussion.

References

[1] Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press (1994)

[2] Carley, K.: A theory of group stability. American Sociological Review **56**(3) (1991) 331–354

[3] Krackhardt, D., Carley, K.M.: A PCANS model of structure in organization. In: Int. Sym. on Command and Control Research and Technology. (1998)
[4] Carley, K.: A comparison of artificial and human organizations. Journal of Economic Behavior and Organization **56** (1996) 175–191
[5] Denham, W.W., McDaniel, C.K., Atkins, J.R.: Aranda and Alyawarra kinship : A quantitative argument for a double helix model. American Ethnologist **6**(1) (1979) 1–24
[6] Sparrow, M.: The application of network analysis to criminal intelligence: an assessment of prospects. Social Networks **13** (1991) 251–274
[7] Hix, S., Noury, A., Roland, G.: Power to the parties: Cohesion and competition in the European Parliament, 1979-2001. British Journal of Political Science **35**(2) (2005) 209–234
[8] Cox, G., Poole, K.: On measuring the partisanship in roll-call voting: The U.S. House of Represenatatives, 1887-1999. American Journal of Political Science **46**(1) (2002) 477–489
[9] Kubica, J., Moore, A., Schneider, J., Yang, Y.: Stochastic link and group detection. In: AAAI. (2002)
[10] Bhattacharya, I., Getoor, L.: Deduplication and group detection using links. In: LinkKDD. (2004)
[11] Nowicki, K., Snijders, T.A.: Estimation and prediction for stochastic blockstructures. Journal of the American Statistical Association **96**(455) (2001)
[12] Kemp, C., Griffiths, T.L., Tenenbaum, J.: Discovering latent classes in relational data. Technical report, MIT CSAIL (2004)
[13] McCallum, A., Corrada-Emanuel, A., Wang, X.: Topic and role discovery in social networks. In: IJCAI. (2005)
[14] Blei, D., Ng, A., Jordan, M.: Latent Dirichlet allocation. JMLR **3** (2003) 993–1022
[15] Dhillon, I.S., Mallela, S., Modha, D.S.: Information-theoretic co-clustering. In: SIGKDD. (2003)
[16] Bekkerman, R., Yaniv, R.E., McCallum, A.: Multi-way distributional clustering via pairwise interactions. In: ICML. (2005)
[17] Jakulin, A., Buntine, W.: Analyzing the US Senate in 2003: Similarities, networks, clusters and blocs (2004)
[18] Pajala, A., Jakulin, A., Buntine, W.: Parliamentary group and individual voting behavior in Finnish Parliamentin year 2003 : A group cohesion and voting similarity analysis (2004)
[19] Beeferman, D., Berger, A.: Agglomerative clustering of a search engine query log. In: SIGKDD. (2000)
[20] Fenn, D., Suleman, O., Efstathiou, J., Johnson, N.: How does Europe make its mind up? Connections, cliques, and compatibility between countries in the Eurovision song contest. arXiv:physics/0505071 (2005)
[21] Voeten, E.: Documenting votes in the UN General Assembly (2003) http://home.gwu.edu/∼voeten/UNVoting.htm.
[22] Sarkar, P., Moore, A.: Dynamic social network analysis using latent space models. In: The 19th Annual Conference on Neural Information Processing Systems. (2005)
[23] Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: Proceedings of the 23rd International Conference on Machine Learning. (2006)
[24] Wang, X., McCallum, A.: Topics over time: A non-markov continuous-time model of topical trends. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. (2006)