The Author-Recipient-Topic Model for Topic and Role Discovery in Social Networks: Experiments with Enron and Academic Email

Andrew McCallum, Andrés Corrada-Emmanuel, Xuerui Wang
Department of Computer Science
University of Massachusetts Amherst
Amherst, MA 01003 USA
{mccallum,corrada,xuerui}@cs.umass.edu

Technical Report UM-CS-2004-096

December 11, 2004

Abstract

Previous work in social network analysis (SNA) has modeled the existence of links from one entity to another, but not the language content or topics on those links. We present the Author-Recipient-Topic (ART) model for social network analysis, which learns topic distributions based on the the direction-sensitive messages sent between entities. The model builds on Latent Dirichlet Allocation and the Author-Topic (AT) model, adding the key attribute that distribution over topics is conditioned distinctly on both the sender and recipient—steering the discovery of topics according to the relationships between people. We give results on both the Enron email corpus and a researcher's email archive, providing evidence not only that clearly relevant topics are discovered, but that the ART model better predicts people's roles.

1 Introduction

Social network analysis (SNA) is the study of mathematical models for interactions among people, organizations and groups. With the recent availability of large datasets of human interactions (Shetty & Adibi, 2004; Wu et al., 2003), the popularity of services like Friendster.com and LinkedIn.com, and the salience of the connections among the 9/11 hijackers, there has been growing interest in social network analysis.

Historically, research in the field has been led by social scientists and physicists (Lorrain & White, 1971; Albert & Barabási, 2002; Watts, 2003; Wasserman & Faust, 1994), and previous work has emphasized binary interaction data, sometimes with directed edges, sometimes with weights on the edges. There has not, however, yet been significant work by researchers with backgrounds in statistical natural language processing, nor analysis that captures the richness of the *language contents* of the interactions—the words, the topics, and other high-dimensional specifics of the interactions between people.

Using pure network connectivity properties, SNA often aims to discover various categories of nodes in a network. For example, in addition to determining that a node-degree distribution is heavy-tailed, we can also find those particular nodes with an inordinately high number of connections, or with connections to a particularly well-connected subset of the network. Furthermore, using these properties we can assign "roles" to certain nodes, *e.g.* (Lorrain & White, 1971; Wolfe & Jensen, 2003). However, it is clear that

network properties are not enough to discover all the roles in a social network. Consider email messages in a corporate setting, and imagine a situation where a tightly knit group of users trade email messages with each other in a roughly symmetric fashion. Thus, at the network level they appear to fulfill the same role. But perhaps, one of the users is in fact a manager for the whole group—a role that becomes obvious only when one accounts for the language content of the email messages.

Outside of the social network analysis literature, there has been a stream of new research in machine learning and natural language models for clustering words in order to discover the few underlying topics that are combined to form documents in a corpus. Latent Dirichlet Allocation (Blei et al., 2003) robustly discovers multinomial word distributions of these topics. Hierarchical Dirichlet Processes (Teh et al., 2004) can determine an appropriate number of topics for a corpus. The Author-Topic Model (Steyvers et al., 2004; Rosen-Zvi et al., 2004) learns topics conditioned on the mixture of authors that composed a document. However, none of these models are appropriate for social network analysis, in which we aim to capture the directed interactions and relationships between people.

The paper presents the *Author-Recipient-Topic* (ART) model, a directed graphical model of words in a message generated given their author and a set of recipients. The model is similar to the Author-Topic (AT) model, but with the crucial enhancement that it conditions the per-message topic distribution jointly on both the author and individual recipients, rather than on individual authors. Thus the discovery of topics in the ART model is influenced by the social structure in which messages are sent and received. Each topic consists of a multinomial distribution over words. Each author-recipient pair has a distribution over topics. We can also easily calculate marginal distributions over topics conditioned solely on an author, or solely on a recipient, in order to find the topics on which each person is most likely to send or receive.

Most importantly, we can also effectively use these person-conditioned topic distributions to measure similarity between people, and thus discover people's roles by clustering using this similarity.[1] For example, people who receive messages containing requests for photocopying, travel bookings, and meeting room arrangements can all be said to have the role "administrative assistant," and can be discovered as such because in the ART model they will all have these topics with high probability in their receiving distribution. Note that we can discover that two people have similar roles even if in the graph they are connected to very different sets of people.

We demonstrate this model on the Enron email corpus comprising 147 people and 24k messages, and also on 9 months of incoming and outgoing mail of the first author, comprising 825 people and 23k messages. We show not only that ART discovers extremely salient topics, but also give evidence that ART predicts people's roles better than AT. Furthermore we show that the similarity matrix produced by AT is drastically different than the SNA matrix, but ART's is similar, while also having some interesting differences.

We also describe an extension of the ART model that explicitly captures *roles* of people, by generating role associations for the authors and recipients of a message, and conditioning the topic distributions on those role assignments. The model, which we term Role-Author-Recipient-Topic (RART), naturally represents that one person can have more than one role. We present three possible RART variants, and describe preliminary experiments with one of these variants.

2 Author-Recipient-Topic Models

Before describing the Author-Recipient-Topic model, we first describe three related models. Latent Dirichlet Allocation (LDA) is a Bayesian network that generates a document using a mixture of topics (Blei et al., 2003). In its generative process, for each document d, a multinomial distribution θ over topics is randomly

[1]The clustering may either external to the model by simple greedy-agglomerative clustering, or internal to the model by introducing latent variables for the sender's and recipient's roles, as described in the Role-Author-Recipient-Topic (RART) model toward the end of this paper.

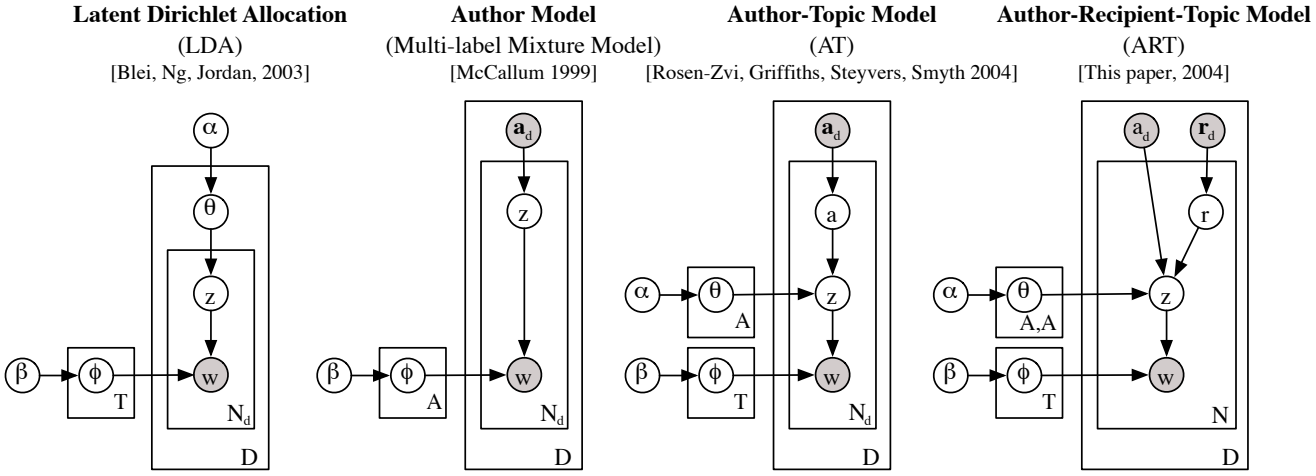

Figure 1: Three related models, and the Author-Recipient-Topic model. In all models, each observed word, w, is generated from a multinomial word distribution, ϕ_z, specific to a particular topic, z, however topics are selected differently in each of the models. In LDA, the topic is sampled from a per-document topic distribution, θ, which in turn is sampled from a Dirichlet over topics. In the Author Model, there is one topic associated with each author (or category), and authors are sampled uniformly. In the Author-Topic model, there is a separate topic-distribution, θ, for each author, and the selection of topic-distribution is determined by uniformly sampling an author from the observed list of the document's authors. In the Author-Recipient-Topic model, there is a separate topic-distribution for each author-recipient pair, and the selection of topic-distribution is determined from the observed author, and by uniformly sampling from the set of recipients for the document.

sampled from a Dirichlet with parameter α, and then to generate each word, a topic z is chosen from this topic distribution, and a word, w, is generated by randomly sampling from a topic-specific multinomial distribution ϕ_z. The robustness of the model is greatly enhanced by integrated out uncertainty about the per-document topic distribution θ.

The Author model (also termed a Multi-label Mixture Model) (McCallum, 1999), is a Bayesian network that simultaneously models document content and its authors' interests with a one-to-one correspondence between topics and authors. The model was originally applied to multi-label document classification, with categories acting as authors. In its generative process, for each document a set of authors $\mathbf{a}_d$ is observed. To generate each word, an author, z, is sampled uniformly from the set, and then a word, w, is generated by sampling from an author-specific multinomial distribution ϕ_z.

The Author-Topic (AT) model is a similar Bayesian network, in which each authors' interests are modeled with a *mixture* of topics (Steyvers et al., 2004; Rosen-Zvi et al., 2004). In its generative process for each document, a set of authors, $\mathbf{a}_d$, is observed. To generate each word, an author x is chosen at uniform from this set, then a topic z is selected from a topic distribution θ_x that is specific to the author, and then a word w is generated by sampling from a topic-specific multinomial distribution ϕ_z.

However, as described previously, neither of these models are suitable for modeling message data.

An email message has one sender and in general more than one recipients. We could treat both the sender and the recipients as "authors" of the message, and then employ the AT model, but it does not distinguish the author and the recipients of the message. This is undesirable in many real-world situations. A manager may send email to a secretary and vice versa, but the nature of the requests and language used may be quite different. Even more dramatically, consider the large quantity of junk email that we receive; modeling the

topics of these messages as undistinguished from the topics we write about as authors would be extremely confounding and undesirable since they do not reflect our expertise or roles.

Alternatively we could still employ the AT model by ignoring the recipient information of email and treating each email document as if it only has one author. However, in this case (which is similar to the LDA model) we are lose all information about the recipients, and the connections between people implied by sender-recipient relationships.

Thus, we propose an Author-Recipient-Topic (ART) model for message data. The ART model captures topics and the directed social network of senders and receivers by conditioning the multinomial distribution over topics distinctly on both the author and one recipient of a message. Unlike the AT, the ART model takes into consideration both author and recipients distinctly, in addition to modeling the email content as a mixture of topics.

The ART model is a Bayesian network that simultaneously models message content, as well as the directed social network in which the messages are sent. In its generative process for each message, an author, a_d, and a set of recipients, $\mathbf{r}_d$, are observed. To generate each word, a recipient, x, is chosen at uniform from $\mathbf{r}_d$, and then a topic z is chosen from a multinomial topic distribution $\theta_{a_d,x}$, where the distribution is specific to the author-recipient pair (a_d, x). (This distribution over topics could also be smoothed against a distribution conditioned on the author only, although we did not find that to be necessary in our experiments.) Finally, the word w is generated by sampling from a topic-specific multinomial distribution ϕ_z. The result is that the discovery of topics is guided by the social network in which the collection of message text was generated.

The Bayesian network for all four models is shown in Figure 1.

In the ART model, for a particular message d, given the hyperparameters α and β, the author a_d, and the set of recipients $\mathbf{r}_d$, the joint distribution of an author mixture θ, a topic mixture ϕ, a set of N_d recipients $\mathbf{x}_d$, a set of N_d topics $\mathbf{z}_d$ and a set of N_d words $\mathbf{w}_d$ is given by:

$$p(\theta, \phi, \mathbf{x}_d, \mathbf{z}_d, \mathbf{w}_d | \alpha, \beta, a_d, \mathbf{r}_d) = p(\theta|\alpha)p(\phi|\beta) \prod_{n=1}^{N_d} p(x_{dn}|\mathbf{r}_d)p(z_{dn}|\theta_{a_d,x_{dn}})p(w_{dn}|\phi_{z_{dn}})$$

Integrating over θ and ϕ, and summing over $\mathbf{x}_d$ and $\mathbf{z}_d$, we get the marginal distribution of a document:

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

Finally, we take the product of the marginal probabilities of single documents, and the probability of a corpus is:

$$p(\mathbf{D}|\alpha, \beta, \mathbf{a}, \mathbf{r}) = \prod_{d=1}^{D} p(\mathbf{w}_d|\alpha, \beta, a_d, \mathbf{r}_d)$$

2.1 Monte Carlo Gibbs sampling

Inference on models in the LDA family cannot be performed exactly. Three standard approximations have been used to obtain practical results: variational methods (Blei et al., 2003), Gibbs sampling (Griffiths & Steyvers, 2004; Steyvers et al., 2004; Rosen-Zvi et al., 2004), and expectation propagation (Griffiths & Steyvers, 2004; Minka & Lafferty, 2002). We chose Gibbs sampling for its ease of implementation.

To carry out the Gibbs sampling we need to derive a formula for $P(z_i, x_i|\mathbf{z}_{-i}, \mathbf{x}_{-i})$, the conditional distribution of a topic and recipient for the i_w word given all other words topic and recipient assignments, $\mathbf{z}_{-i}$ and $\mathbf{x}_{-i}$. To understand why, let us try to calculate $P(\mathbf{z}, \mathbf{x} \mid \mathbf{w})$, the posterior distribution of topic and recipient assignments given the words in the corpus.

We begin by calculating $P(\mathbf{w} \mid \mathbf{z}, \mathbf{x})$. Using $P(\mathbf{w} \mid \mathbf{z}, \mathbf{x}, \Phi)$, we can integrate out the unknown Φ distributions to obtain:

$$P(\mathbf{w} \mid \mathbf{z}, \mathbf{x}, \Phi) = \prod_{i_w=1}^{W} \phi_{z_{i_w}}(w_{i_w})$$

Rearranging the product over the W word tokens present in the corpus to collect words that are assigned to the same topic, we obtain,

$$P(\mathbf{w} \mid \mathbf{z}, \mathbf{x}, \Phi) = \prod_{z=1}^{T} \prod_{v=1}^{V} \phi_z^{n_z^{w_v}},$$

where $n_z^{w_v}$ is the number of times that a vocabulary word, w_v was assigned to a topic. And finally, we integrate out the ϕ distributions by using the Dirichlet distribution,

$$\begin{aligned} P(\mathbf{w} \mid \mathbf{z}, \mathbf{x}) &= \int \prod_{z=1}^{T} \left(\frac{\Gamma(\sum_{v=1}^{V} \beta_v)}{\prod_{v=1}^{V} \Gamma(\beta_v)} \left(\prod_{v=1}^{V} \phi_z^{n_z^{w_v} + \beta_v - 1}(w_v) d\phi_z(w_v) \right) \right) \\ &= \prod_{z=1}^{T} \left(\frac{\Gamma(\sum_{v=1}^{V} \beta_v)}{\prod_{v=1}^{V} \Gamma(\beta_v)} \left(\frac{\prod_{v=1}^{V} \Gamma(n_z^{w_v} + \beta_v)}{\Gamma(\sum_{v=1}^{V} \beta_v + \sum_{v=1}^{V} n_z^{w_v})} \right) \right) \end{aligned}$$

Similarly, we can calculate $P(\mathbf{z}, \mathbf{x})$ using a procedure analogous to that used for $P(\mathbf{w} \mid \mathbf{z}, \mathbf{x})$. We collect terms from vocabulary words assigned to the same topic and author-recipient pair and integrate out the Θ distributions corresponding to all the different author-recipient pairs, P:

$$P(\mathbf{z}, \mathbf{x}) = \left(\prod_{i_w=1}^{W} \frac{1}{n_R(d_{i_w})} \right) \prod_{p=1}^{P} \left(\frac{\Gamma(\sum_z \alpha_z)}{\prod_{z=1}^{T} \Gamma(\alpha_z)} \frac{\prod_z \Gamma(n_p^z + \alpha_z)}{\Gamma(\sum_z \alpha_z + \sum_z n_p^z)} \right),$$

where $n_R(d_{i_w})$ is the number of recipients corresponding to a word in a given email.

Putting together our equations for $P(\mathbf{w} \mid \mathbf{z}, \mathbf{x})$ and $P(\mathbf{z}, \mathbf{x})$ we can obtain an expression for $P(\mathbf{w}, \mathbf{z}, \mathbf{x})$. This allows us to write an expression for the posterior distribution of $\mathbf{z}$ and $\mathbf{x}$ given the corpus,

$$P(\mathbf{z}, \mathbf{x} \mid \mathbf{w}) = \frac{P(\mathbf{w}, \mathbf{z}, \mathbf{x})}{\sum_{\mathbf{z},\mathbf{x}} P(\mathbf{w}, \mathbf{z}, \mathbf{x})}$$

However, we cannot calculate the denominator directly.

Gibbs sampling gets around this intractability by using the conditional distribution $P(z_i, x_i, w_i \mid \mathbf{z}_{-i}, \mathbf{x}_{-i}, \mathbf{w}_{-i})$ to run a Markov chain Monte Carlo calculation. We can calculate this conditional as,

$$\begin{aligned} P(z_i, x_i, w_i \mid \mathbf{z}_{-i}, \mathbf{x}_{-i}, \mathbf{w}_{-i}) &= \frac{P(\mathbf{z}, \mathbf{x}, \mathbf{w})}{P(\mathbf{z}_{-i}, \mathbf{x}_{-i}, \mathbf{w}_{-i})} \\ &= \frac{1}{n_R} \frac{\frac{\Gamma(n_p^t + \alpha_t)}{\Gamma(\sum_z n_p^z + \sum_z \alpha_z)} \frac{\Gamma(n_t^{w_v} + \beta_v)}{\Gamma(\sum_v n_t^{w_v} + \sum_v \beta_v)}}{\frac{\Gamma(n_p^t - 1 + \alpha_t)}{\Gamma(\sum_z n_p^z - 1 + \sum_z \alpha_z)} \frac{\Gamma(n_t^{w_v} - 1 + \beta_v)}{\Gamma(\sum_v n_z^{w_v} - 1 + \sum_v \beta_v)}} \\ &= \frac{1}{n_R} \frac{n_{p,-i}^t + \alpha_t}{\sum_z n_{p,-i}^z + \sum_z \alpha_z} \frac{n_{t,-i}^{w_v} + \beta_v}{\sum_v n_{t,-i} + \sum_v \beta_v}, \end{aligned}$$

where the recipient, r, is part of the author-recipient pair, p, the $-i$ subscript is used to denote that the counts are taken by excluding the assignment of word i itself, and n_R is the number of recipients for the email to which word i belongs.

Further manipulation can turn equation 1 into update equations for the topic and recipient of each corpus token, $P(z_i \mid \mathbf{z}_{-i}, \mathbf{x}, \mathbf{w})$ and $P(x_i \mid \mathbf{z}, \mathbf{x}_{-i}, \mathbf{w})$ suitable for random or systematic scan updates:

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

3 Related Work

The use of social networks to discover "roles" for the people (or nodes) in the network goes back over three decades to the work of Lorrain and White (1971). It is based on the hypothesis that nodes on a network that relate to other nodes in "equivalent" ways must have the same role. This equivalence was given a probabilistic interpretation by Holland et al. (1983): nodes assigned to a class/role are stochastically equivalent if the probabilities of the relationships with all other nodes are the same for nodes in the same class/role.

The limitation of a single class/role label for each node in the network was relaxed in recent work by Wolfe and Jensen (2003). They consider a model that assigns multiple role labels to a given node in the network. One advantage of multiple labels is that in this factored model, fewer parameters are required to be estimated than then in a non-factored, single label obliged to represent more values. They find that, two labels with three values (giving $3^2 = 9$ possible labelings for each node) is a better estimator for synthetic data produced by a two-label process than a model using one label with nine possible values. This is, of course, the advantage of *mixture models*, such as LDA and the ART model presented here.

The study of email social networks has been hampered by the unavailability of a public corpus. The research that has been published has used email to-from logs. Logs are easier to obtain and are less intrusive on user's privacy. This means that previous research has focused on the topological structure of email networks, and the dynamics of the email traffic between users. Wu et al. (2003) looked at how information flowed in an email network of users in research labs (mostly from HP Labs). They conclude that epidemic models of information flow do not work for email networks and thus identifying hubs in the network may not guarantee that information originating at a node reaches a large fraction of the network. This finding serves as an example that network properties are not sufficient to optimize flow on an email network. Adamic and Adar (2004) studied the efficiency of "local information" search strategies on social networks. They find that in the case of an email network at HP Labs, a greedy search strategy works efficiently as predicted by Kleinberg (2000) and Watts et al. (2002).

All these approaches, however, limit themselves to the use of network topology to discover roles. The ART model complements these approaches by using the content of the "traffic" between nodes to create language models that can bring out differences invisible at the network level.

As discussed above, the ART model is a direct offspring of Latent Dirichlet Allocation (Blei et al., 2003), the Multi-label Mixture Model (McCallum, 1999), and the Author-Topic Model (Steyvers et al., 2004; Rosen-Zvi et al., 2004), with the distinction that ART is specifically designed to capture language used in a directed network of correspondents.

4 Experimental Results

We present results with the Enron email corpus and the personal email of one of the authors of this paper (McCallum). The Enron email corpus, is a large body of email messages subpoenaed as part of the investigation by the Federal Energy Regulatory Commission (FERC), and then placed in the public record. The

original dataset contains 517,431 messages, however MD5 hashes on contents, authors and dates show only 250,484 of these to be unique.

Although the Enron Email Dataset contains the email folders of 150 people, two people appear twice with different usernames, and we remove one person who only sent automated calendar reminders, resulting in 147 people for our experiments. We hand-corrected variants of the email addresses for these 147 users to capture the connectivity of as much of these users' email as possible. The total number of email messages traded among these users is 23,488. We did not model email messages that were not received by at least one of the 147.

In order to capture only the new text entered by the author of a message, it is necessary to remove "quoted original messages" in replies. We eliminate this extraneous text by a simple heuristic: all text in a message below a "forwarded message" line or timestamp is removed. This heuristic certainly incorrectly looses words that are interspersed with quoted email text. Words are formed as sequences of alphabetic characters. To remove sensitivity to capitalization, all text is downcased.

Our second dataset consists of the personal email sent and received by McCallum between January and October 2004. It consists of 23,488 unique messages written by 825 authors. In typical power-law behavior, most of these authors wrote only a few messages, while 128 wrote ten or more emails. After applying the same text normalization filter (lowercasing, removal of quoted email text, etc.) that was used for the Enron data, we obtained a text corpus containing 457,057 word tokens, and a vocabulary of 22,901 unique words.

4.1 Topics and Prominent Relations from ART models

Table 1 shows the highest probability words from eight topics in an ART model trained on the 147 users with 50 topics. (The quoted titles are our own interpretation of a summary for the topics.) The clarity and specificity of these topics are typical of the topics discovered by the model. For example, Topic 17 comes from message discussing review and comments on documents; Topic 27 comes from messages negotiating meeting times.

Beneath the word distribution for each topic are the three author-recipient pairs with highest probability of discussing that topic—each pair separated by a horizontal line, with the author above the recipient. For example, Mary Hain, the top author of messages in the "Legal Contracts" topic, was an in-house lawyer at Enron. By inspection of other messages, Eric Bass seems to have been the coordinator for a fantasy basketball league among Enron employees.

4.2 Stochastic Blockstructures and Roles

The stochastic equivalence hypothesis from SNA states that nodes in a network that behave stochastically equivalently must have similar roles. In the case of an email network consisting of message counts, the natural way to measure equivalence is to examine the probability that a node communicated with other nodes. If two nodes have similar probability distribution over their communication partners, we should consider them role-equivalent. Lacking a true distance measure between probability distributions, we can use some symmetric measure, such as the Jensen-Shannon (JS) divergence, to obtain a symmetric matrix relating the nodes in the network. Since we want to consider nodes/users that have a small JS divergence as equivalent, we can use the inverse of the divergence to construct a symmetric matrix in which larger numbers indicate higher similarity between users.

Standard recursive graph-cutting algorithms on this matrix can be used to cluster users, rearranging the rows/columns to form approximately block-diagonal structures. This is the familiar process of 'block-structuring' used in SNA. We perform such an analysis on two datasets: a small subset of the Enron users consisting mostly of people associated with the Transwestern Pipeline Division within Enron, and the entirety of McCallum's email.

Topic 5 "Legal Contracts"		Topic 17 "Document Review"		Topic 27 "Time Scheduling"		Topic 45 "Sports Pool"	
section	0.0299	attached	0.0742	day	0.0419	game	0.0170
party	0.0265	agreement	0.0493	friday	0.0418	draft	0.0156
language	0.0226	review	0.0340	morning	0.0369	week	0.0135
contract	0.0203	questions	0.0257	monday	0.0282	team	0.0135
date	0.0155	draft	0.0245	office	0.0282	eric	0.0130
enron	0.0151	letter	0.0239	wednesday	0.0267	make	0.0125
parties	0.0149	comments	0.0207	tuesday	0.0261	free	0.0107
notice	0.0126	copy	0.0165	time	0.0218	year	0.0106
days	0.0112	revised	0.0161	good	0.0214	pick	0.0097
include	0.0111	document	0.0156	thursday	0.0191	phillip	0.0095
M.Hain	0.0549	G.Nemec	0.0737	J.Dasovich	0.0340	E.Bass	0.3050
J.Steffes		B.Tycholiz		R.Shapiro		M.Lenhart	
J.Dasovich	0.0377	G.Nemec	0.0551	J.Dasovich	0.0289	E.Bass	0.0780
R.Shapiro		M.Whitt		J.Steffes		P.Love	
D.Hyvl	0.0362	B.Tycholiz	0.0325	C.Clair	0.0175	M.Motley	0.0522
K.Ward		G.Nemec		M.Taylor		M.Grigsby	

Topic 34 "Operations"		Topic 37 "Power Market"		Topic 41 "Government Relations"		Topic 42 "Wireless"	
operations	0.0321	market	0.0567	state	0.0404	blackberry	0.0726
team	0.0234	power	0.0563	california	0.0367	net	0.0557
office	0.0173	price	0.0280	power	0.0337	www	0.0409
list	0.0144	system	0.0206	energy	0.0239	website	0.0375
bob	0.0129	prices	0.0182	electricity	0.0203	report	0.0373
open	0.0126	high	0.0124	davis	0.0183	wireless	0.0364
meeting	0.0107	based	0.0120	utilities	0.0158	handheld	0.0362
gas	0.0107	buy	0.0117	commission	0.0136	stan	0.0282
business	0.0106	customers	0.0110	governor	0.0132	fyi	0.0271
houston	0.0099	costs	0.0106	prices	0.0089	named	0.0260
S.Beck	0.2158	J.Dasovich	0.1231	J.Dasovich	0.3338	R.Haylett	0.1432
L.Kitchen		J.Steffes		R.Shapiro		T.Geaccone	
S.Beck	0.0826	J.Dasovich	0.1133	J.Dasovich	0.2440	T.Geaccone	0.0737
J.Lavorato		R.Shapiro		J.Steffes		R.Haylett	
S.Beck	0.0530	M.Taylor	0.0218	J.Dasovich	0.1394	R.Haylett	0.0420
S.White		E.Sager		R.Sanders		D.Fossum	

Table 1: An illustration of several topics from a 50-topic run for the Enron Email Dataset. Each topic is shown with the top 10 words and their corresponding conditional probabilities. The quoted titles are our own summary for the topics. Below are prominent author-recipient pairs for each topic. For example, Mary Hain was an in-house lawyer at Enron; Eric Bass was the coordinator of a fantasy basketball league within Enron. In the "Operations" topic it is satisfying to see Beck, who was the Chief Operating Officer at Enron; Kitchen was President of Enron Online; and Lavorato was CEO of Enron America. In the "Government Relations" topic, we see Dasovich, who was a Government Relation Executive, Shapiro who was Vice President of Regulatory Affairs, Steffes, who was Vice President of Government Affairs, and Sanders, who was Vice President of WholeSale Services. In "Wireless" we see that Haylett, who was Chief Financial Officer and Treasurer, was an avid user of the Blackberry brand wireless, portable email system.

Topic 5 "Grant Proposals"		Topic 31 "Meeting Setup"		Topic 38 "ML Models"		Topic 41 "Friendly Discourse"	
proposal	0.0397	today	0.0512	model	0.0479	great	0.0516
data	0.0310	tomorrow	0.0454	models	0.0444	good	0.0393
budget	0.0289	time	0.0413	inference	0.0191	don	0.0223
work	0.0245	ll	0.0391	conditional	0.0181	sounds	0.0219
year	0.0238	meeting	0.0339	methods	0.0144	work	0.0196
glenn	0.0225	week	0.0255	number	0.0136	wishes	0.0182
nsf	0.0209	talk	0.0246	sequence	0.0126	talk	0.0175
project	0.0188	meet	0.0233	learning	0.0126	interesting	0.0168
sets	0.0157	morning	0.0228	graphical	0.0121	time	0.0162
support	0.0156	monday	0.0208	random	0.0121	hear	0.0132
smyth mccallum	0.1290	ronb mccallum	0.0339	casutton mccallum	0.0498	mccallum culotta	0.0558
mccallum stowell	0.0746	wellner mccallum	0.0314	icml04-webadmin icml04-chairs	0.0366	mccallum casutton	0.0530
mccallum lafferty	0.0739	casutton mccallum	0.0217	mccallum casutton	0.0343	mccallum ronb	0.0274
mccallum smyth	0.0532	mccallum casutton	0.0200	nips04workflow mccallum	0.0322	mccallum saunders	0.0255
pereira lafferty	0.0339	mccallum wellner	0.0200	weinman mccallum	0.0250	mccallum pereira	0.0181

Table 2: The four topics most prominent in McCallum's email exchange with Padhraic Smyth, from a 50-topic run of ART on 10 months of McCallum's email. The topics provide an extremely salient summary of McCallum and Smyth's relationship during this time period: they wrote a grant proposal together; they set up many meetings; they discussed machine learning models; they were friendly with each other. Each topic is shown with the 10 highest-probability words and their corresponding conditional probabilities. The quoted titles are our own summary for the topics. Below are prominent author-recipient pairs for each topic. The people other than smyth also appear in very sensible associations: stowell is McCallum's proposal budget administrator; McCallum also wrote a proposal with John Lafferty and Fernando Pereira; McCallum also sets up meetings, discusses machine learning and has friendly discourse with his graduate student advisees: ronb, wellner, casutton, and culotta; he does not, however, discuss the details of proposal-writing with them.

We begin with the Enron TransWestern Pipeline Division. Our analysis here employed a "closed-universe" assumption—only those messages traded among authors in the dataset were used.

The traditional SNA similarity measure (in this case JS divergence of distributions on recipients from each person) is shown in the left matrix in Figure 2. Darker shading indicates that two users are considered more similar. A related matrix resulting from our ART model (JS divergence of recipient-marginalized topic distributions for each email author) appears in the middle of the Figure. Finally, the results of the same analysis using topics from the AT model rather than our ART model can be seen on the right. The three matrices are similar, but have interesting differences.

Consider Enron employee Geaccone (user 9 in all the matrices in Figure 2). According to the traditional SNA role measurement, Geaccone and McCarty (user 8) have very similar roles, however, both the AT and ART models indicate no special similarity. Inspection of the email messages for both users reveals that Geaconne was an Executive Assistant, while McCarty was a Vice-President—rather different roles—and, thus output of ART and AT is more appropriate. We can interpret these results as follows: SNA analysis shows that they wrote email to similar sets of people, but the ART analysis illustrates that they used very different language when they wrote to these people.

Comparing ART against AT, both models provide similar role distance for Geaccone versus McCarty, but ART and AT show their differences elsewhere. For example, AT indicates a very strong role similarity between Geaconne and Hayslett (user 6), who was her boss (and CFO & Vice President in the Division); on the other hand, ART more correctly designates a low role similarity for this pair—in fact, ART assigns low similarity between Geaconne and all others in the matrix, which is appropriate because she is the only executive assistant in this small sample of Enron employees.

Another interesting pair of people is Blair (user 4) and Watson (user 14). ART predicts them to be role-similar, while the SNA and AT models do not. ART's prediction seems more appropriate since Blair worked on "gas pipeline logistics" and Watson worked on "pipeline facility planning", two very similar jobs.

McCarty, a Vice-President and CTO in the Division, also highlights differences between the models. The ART model puts him closest to Horton (user 5), who was President of the Division. AT predicts that he is closest to Rapp (user 12), who was merely a lawyer that reviewed business agreements, and also close to Harris (user 15), who was only a mid-level manager.

Using ART in this way emphasizes role similarity, but not group membership. This can be seen by considering Thomas (user 3, an energy futures trader), and his relation to both Rapp (user 12, the lawyer mentioned above), and Lokey (user 16, a regulatory affairs manager). These three people work in related areas, and both ART and AT fittingly indicate a role similarity between them, (ART marginally more so than AT). On the other hand, SNA emphasizes *group memberships* rather than role similarity by placing users 1 through 3 in a rather distinct block structure; they are the only three people in this matrix who were not members of the Enron Transwestern Division group, and these three exchanged more email with each other than with the people of the Transwestern Division. In pending work we are developing a model that integrates both ART and SNA metrics to jointly model both role and group memberships.

Based on the above examples, and other similar examples, we posit that the ART model is more appropriate than the SNA and AT in predicting role similarity.

We thus would claim that the ART model is clearly better than the SNA model in predicting role-equivalence between users, and somewhat better than the AT model in this capacity.

We also carried out this analysis with the personal email for McCallum to further validate the difference between the ART and SNA predictions. There are 825 users in this email corpus. Table 3 shows the closest pairs, as calculated by the ART model and SNA model. The difference in quality between the ART and SNA halves of the table is striking.

Almost all the pairs predicted by the ART model look reasonable while many of those predicted by SNA are the opposite. For example, ART matches editor and reviews, two email addresses that send messages managing journal reviews. User mike and mikem are actually two different email addresses for the same person. Most other correferent email addresses were pre-collapsed by hand during preprocessing; here ART has pointed out out a mistaken omission, indicating the potential for ART to be used as a helpful component of an automated coreference system. Users aepshtey and smucker were students in a class taught by McCallum. Users coe, laurie and kate are all UMass CS Department administrative assistants; they rarely send email to each other, but they write about similar things. User ang is Andrew Ng from Stanford; joshuago is Joshua Goodman of Microsoft Research; they are both on the organizing committee of a new conference along with McCallum.

On the other hand, the pairs declared most similar by the SNA model are mostly extremely poor. Most of the pairs include donna, and indicate pairs of people who are similar only because in this corpus they appeared mostly sending email only to McCallum, and not others. User donna is McCallum's spouse. Other pairs are more sensible. For example, aepshtey, smucker and rasmith were all students in McCallum's class. User elm is Erik Learned-Miller who is correctly indicated as similar to editor since he is the Production Editor for the Journal of Machine Learning Research.

To highlight the difference between the SNA and ART predictions, we present Table 4, which was obtained by using both ART and SNA to rank the pairs of people by similarity, and then listing the pairs

Pairs considered most alike by ART	
User Pair	*Description*
editor reviews	Both journal review management
mike mikem	Same person! (manual coref error)
aepshtey smucker	Both students in McCallum's class
coe laurie	Both UMass admin assistants
mcollins tom.mitchell	Both ML researchers on SRI project
mcollins gervasio	Both ML researchers on SRI project
davitz freeman	Both ML researchers on SRI project
mahadeva pal	Both ML researchers, discussing hiring
kate laurie	Both UMass admin assistants
ang joshuago	Both on org committee for a conference
Pairs considered most alike by SNA	
User Pair	*Description*
aepshtey rasmith	Both students in McCallum's class
donna editor	Spouse is unrelated to journal editor
donna krishna	Spouse is unrelated to conference organizer
donna ramshaw	Spouse is unrelated to researcher at BBN
donna reviews	Spouse is unrelated to journal editor
donna stromsten	Spouse is unrelated to visiting researcher
donna yugu	Spouse is unrelated grad student
aepshtey smucker	Both students in McCallum's class
rasmith smucker	Both students in McCallum's class
editor elm	Journal editor and its Production Editor

Table 3: Pairs considered most alike by ART and SNA on McCallum email. All pairs produced by the ART model are accurately quite similar. This is not so for the top SNA pairs. Many users are considered similar by SNA merely because they appear in the corpus mostly sending email only to McCallum. However, this causes people with very different roles to be incorrectly declared similar—such as McCallum's spouse and the JMLR editor.

User Pair	*Description*
editor reviews	Both journal editors
jordan mccallum	Both ML researchers
mccallum vanessa	A grad student working in IR
croft mccallum	Both UMass faculty, working in IR
mccallum stromsten	Both ML researchers
koller mccallum	Both ML researchers
dkulp mccallum	Both UMass faculty
blei mccallum	Both ML researchers
mccallum pereira	Both ML researchers
davitz mccallum	Both working on an SRI project

Table 4: Pairs with the highest rank difference between ART and SNA on McCallum email. The traditional SNA metric indicates that these pairs of people are different, while ART indicates that they are similar. There are strong relations between all pairs.

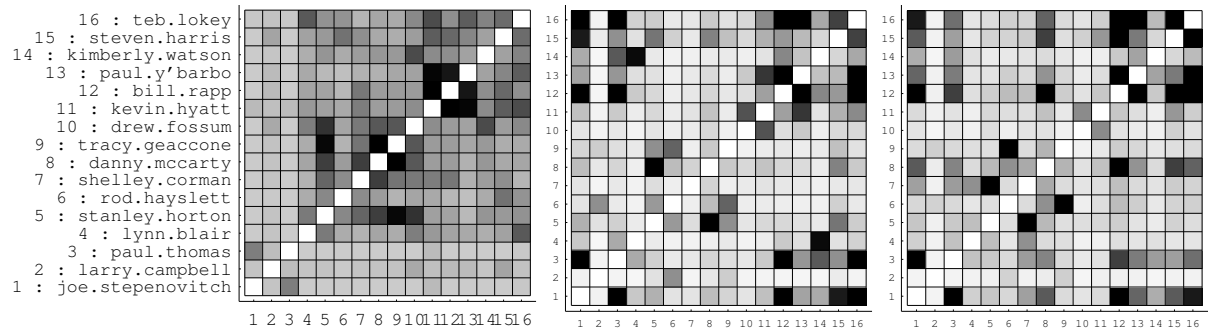

Figure 2: **Left:** SNA Inverse JS Network. **Middle:** ART Inverse JS Network. **Right:** AT Inverse JS Network. Darker shades indicate higher similarity.

with the highest rank *differences* between the two models. These are pairs that SNA indicated were different, but ART indicated were similar. In every case, there are role similarities between the pairs.

5 Role-Author-Recipient-Topic Models

To better explore the roles of authors, an additional level of latent variable can be introduced to explicitly model roles. Of particular interest is capturing the notion that a person can have multiple roles simultaneously —- a person can be both a professor and a hiker. Each role is associated with a set of topics, and these topics may overlap. For example, professors' topics may prominently feature research, meeting times, grant proposals, and friendly relations; hikers topics may prominently feature mountains, climbing equipment, and also meeting times and friendly relations.

We incorporate into the model a new set of variables that take on values indicating role, and term this augmented model the the Role-Author-Recipient-Topic (RART) model. In RART, authors, roles, and message contents are modeled simultaneously. Each author has a multinomial distribution over roles. Authors and recipients are mapped to a role assignments, and then a topic is selected based on these roles. Thus we have a clustering model, in which appearances of topics are the underlying data, and sets of correlated topics gather together clusters that indicate role. Each sender-role and recipient-role pair has a multinomial

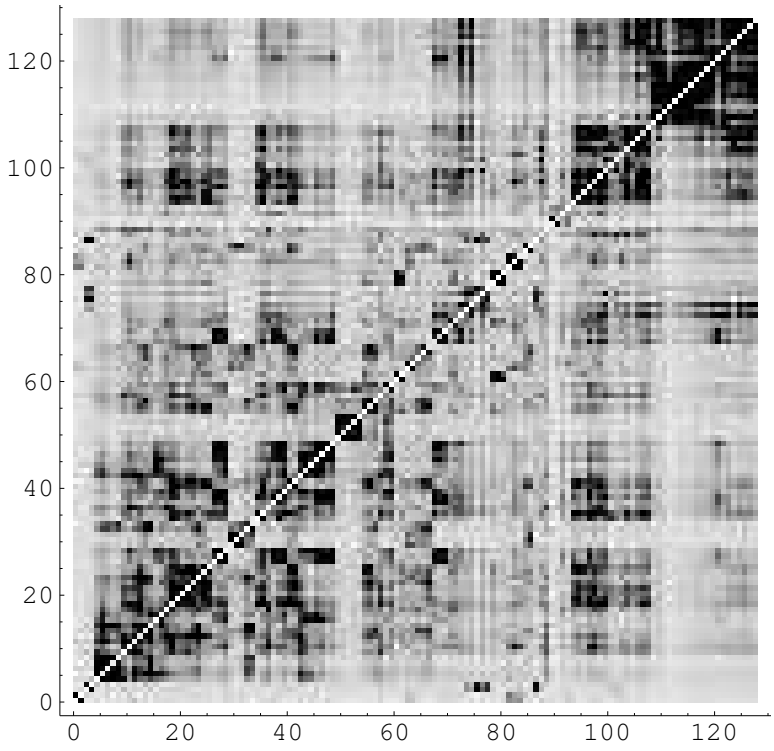

Figure 3: SNA Inverse JS Network for a 10 topic run on McCallum Email Data. Darker shades indicate higher similarity. Graph partitioning was calculated with the 128 authors that had ten or more emails in McCallum's Email Data. The block from 0 to 30 are people in and related to McCallum's research group at UMass. The block from 30 to 50 includes other researchers around the world.

distribution over topics, and each topic has a multinomial distribution over words.

As shown in Figure 4, different strategies can be employed to incorporate the "role" latent variables. First in RART1, role assignments can be made separately for each word in a document. This model represents that a person can change role during the course of the email message. In RART2, on the other hand, a person chooses one role for the duration of the message. Here each recipient of the message selects a role assignment, and then for each word, a recipient (recipient-role) is selected on which to condition the selection of topic. In RART3, the recipients together result in the selection of a common, shared role, which is used to condition the selection of every word in the message. This last model may help capture the fact that a person's role may depend on the other recipients of the message, but also restricts all recipients to a single role.

We describe the generative process of RART1 in this paper in detail, and leave the other two for subsequent work. In its generative process for each message, an author, a_d, and a set of recipients, $\mathbf{r}_d$, are observed. To generate each word, a recipient, x, is chosen at uniform from $\mathbf{r}_d$, and then a role g for the author, and a role h for the recipient x are chosen from two multinomial role distributions ψ_{a_d} and ψ_x, respectively. Next, a topic z is chosen from a multinomial topic distribution $\theta_{g,h}$, where the distribution is specific to the author-role recipient-role pair (g, h). Finally, the word w is generated by sampling from a topic-specific multinomial distribution ϕ_z.

In the RART1 model, for a particular message d, given the hyperparameters α, β and γ, the author a_d, and the set of recipients $\mathbf{r}_d$, the joint distribution of an author mixture θ, a role mixture ψ, a topic mixture ϕ, a set of N_d recipients $\mathbf{x}_d$,a set of N_d sender roles $\mathbf{g}_d$, a set of N_d recipient roles $\mathbf{h}_d$, a set of N_d topics $\mathbf{z}_d$ and a set of N_d words $\mathbf{w}_d$ is given by:

$$\begin{aligned} p(\theta, \phi, \psi, \mathbf{r}_d, \mathbf{g}_d, \mathbf{h}_d, \mathbf{z}_d, \mathbf{w}_d | \alpha, \beta, \gamma, a_d, \mathbf{r}_d) &= p(\psi|\gamma)p(\theta|\alpha)p(\phi|\beta) \\ &\quad \prod_{n=1}^{N_d} p(x_{dn}|\mathbf{r}_d)p(g_{dn}|a_d)p(h_{dn}|x_{dn})p(z_{dn}|\theta_{g_{dn},h_{dn}})p(w_{dn}|\phi_{z_{dn}}) \end{aligned}$$

Integrating over ψ, θ and ϕ, and summing over $\mathbf{x}_d$, g_d, h_d and $\mathbf{z}_d$, we get the marginal distribution of a document:

$$\begin{aligned} p(\mathbf{w}_d|\alpha, \beta, \gamma, a_d, \mathbf{r}_d) &= \iiint p(\psi|\gamma)p(\theta|\alpha)p(\phi|\beta) \\ &\quad \prod_{n=1}^{N_d} \sum_{x_{dn}} \sum_{g_{dn}} \sum_{h_{dn}} \sum_{z_{dn}} p(x_{dn}|\mathbf{r}_d)p(g_{dn}|a_d)p(h_{dn}|x_{dn})p(z_{dn}|\theta_{g_{dn},h_{dn}})p(w_{dn}|\phi_{z_{dn}}) \mathrm{d}\psi \mathrm{d}\phi \mathrm{d}\theta \end{aligned}$$

Finally, we take the product of the marginal probabilities of single documents, and the probability of a corpus is:

$$p(\mathbf{D}|\alpha, \beta, \gamma, \mathbf{a}, \mathbf{r}) = \prod_{d=1}^{D} p(\mathbf{w}_d|\alpha, \beta, \gamma, a_d, \mathbf{r}_d)$$

To perform inference on RART models, the Gibbs sampling formulae can be derived in a similar way as in Section 2.1, but in a more complex form.

6 Experimental Results with RART

No significant experiments have been conducted on RART models. Based upon our preliminary experimental results with the RART model, properly setting the smoothing parameters is crucial. To make inference more efficiently, we can do inference in distinct parts. For example, because we introduce two additional

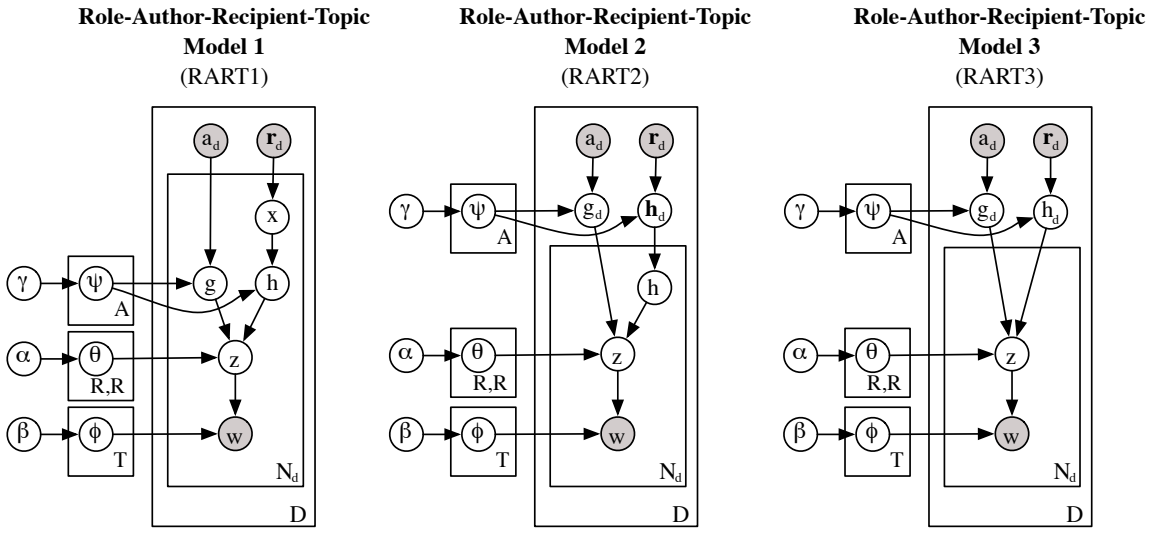

Figure 4: Three possible variants for the Role-Author-Recipient-Topic (RART) model.

latent variables (author role and recipient role), the sampling procedure at each iteration is significantly more complicated. One strategy we have found useful is that we can train an ART model first, and use this to fix the topic assignments for each word token. At the next stage, we treat topic as observed, and in this way the RART model can be trained more simply. Although such a strategy may not be recommended for arbitrary graphical models, we feel this is reasonable because we find that a single sample from Gibbs sampling on the ART model yields useful results.

7 Conclusions

We have presented the Author-Recipient-Topic model, a Bayesian network for social network analysis that discovers discussion topics conditioned on the sender-recipient relationships in a corpus of messages. To the best of our knowledge, this model combines for the first time the directionalized connectivity graph from social network analysis with the clustering of words to form topics from probabilistic language modeling.

The model can be applied to discovering topics conditioned on message sending relationships, clustering to find social roles, and summarizing and analyzing large bodies of message data. The model would form a useful component in systems for routing requests, expert-finding, message recommendation and prioritization, and understanding the interactions in an organization in order to make recommendations about improving organizational efficiency.

Additional work on the Role-Author-Recipient-Topic (RART) and other models that explicitly capture roles and groups is ongoing.

Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval, the National Science Foundation under NSF grant #IIS-0326249, and by the Defense Advanced Research Projects Agency, through the Department of the Interior, NBC, Acquisition Services Division, under contract #NBCHD030010.

References

Adamic, L., & Adar, E. (2004). How to search a social network. http://arXiv.org/abs/cond-mat/0310120.

Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, *74*, 47–97.

Blei, D. M., Ng, A. Y., & Jordan, M. J. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topis. *Proceedings of the National Academy of Sciences* (pp. 5228–5235).

Holland, P., Laskey, K. B., & Leinhardt, S. (1983). Stochastic blockmodels: Some first steps. *Social Networks*, *5*, 109–137.

Kleinberg, J. (2000). Navigation in a small world. *Nature*, *406*, 845.

Lorrain, F., & White, H. C. (1971). The structural equivalence of individuals in social networks. *Journal of Mathematical Sociology*, *1*, 49–80.

McCallum, A. (1999). Multi-label text classification with a mixture model trained by EM. *AAAI Workshop on Text Learning*.

Minka, T., & Lafferty, J. (2002). Expectation-propagation for the generative aspect model. *In Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*. New York: Elsevier.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*. Banff, Alberta, Canada.

Shetty, J., & Adibi, J. (2004). *The Enron email dataset database schema and brief statistical report* (Technical Report). Information Sciences Institute.

Steyvers, M., Smyth, P., Rosen-Zvi, M., & Griffiths, T. (2004). Probabilistic author-topic models for information discovery. *The Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Seattle, Washington.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2004). *Hierarchical dirichlet processes* (Technical Report). UC Berkeley Statistics.

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. New York: Cambridge University Press.

Watts, D. J. (2003). *Six degrees: The science of a connected age*. Norton.

Watts, D. J., Dodds, P. S., & Newman, M. E. J. (2002). Identify and search in social networks. *Science*, *296*, 1302–1305.

Wolfe, A. P., & Jensen, D. (2003). Playing multiple roles: Discovering overlapping roles in social networks. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Seattle, Washington, USA.

Wu, F., Huberman, B. A., Adamic, L. A., & Tyler, J. R. (2003). Information flow in social groups. http://arXiv.org/abs/cond-mat/0305305.