Topic and Role Discovery in Social Networks

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

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 direction-sensitive messages sent between entities. The model builds on Latent Dirichlet Allocation (LDA) 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 and Related Work

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, with directed and/or weighted edges. There has not, however, previously 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] learns topics conditioned on the mixture of authors that composed a document. However, none of these models are appropriate for SNA, 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. 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 23k messages, and also on about 10 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 gives evidence that ART predicts people's roles better than AT and SNA. Furthermore we show that the similarity matrix produced by ART is different from both the SNA matrix and the AT matrix in several appropriate ways.

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 role assignments. The model, which we term *Role-Author-Recipient-Topic* (RART), naturally represents that one person can have more than one role. We describe several possible RART variants, and describe preliminary experiments with one of these variants.

2 Author-Recipient-Topic Models

Before describing the ART 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 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 integrating 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 1-1 correspondence between topics and authors. For each document d, 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]. In its generative process for each document d, a set of authors, $\mathbf{a}_d$, is observed. To generate each word, an author x is chosen uniformly 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 from a topic-specific multinomial distribution ϕ_z. However, as described previously, none of these models is 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 this does not distinguish the author and the recipients of the message, which 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.

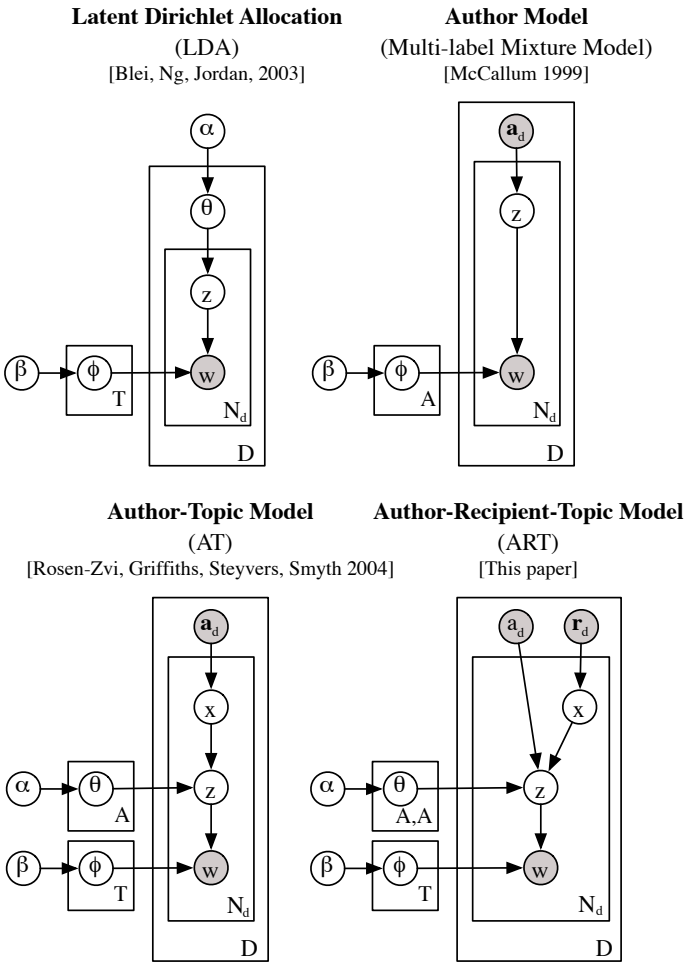

Figure 1: Three related models, and the ART model. In all models, each observed word, w, is generated from a multinomial word distribution, ϕ_z, specific to a particular topic/author, z, however topics are selected differently in each of the models.

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 losing 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 recipients 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 d, an author, a_d, and a set of recipients, $\mathbf{r}_d$, are observed. To generate each word, a recipient, x, is chosen uniformly 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). 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 models is shown in Figure 1.

In the ART model, given the hyperparameters α and β, an author $\mathbf{a}$, and a set of recipients $\mathbf{r}$, the joint distribution of the topic mixtures θ, the word mixtures ϕ, a set of recipients $\mathbf{x}$, a set of topics $\mathbf{z}$ and a set of words $\mathbf{w}$ in the corpus is given by:

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

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

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

3 Experimental Results

We present results with the Enron email corpus and the personal email of the first author of this paper (McCallum). After preprocessing, the Enron corpus we use contains 147 users and 23,488 emails, and the McCallum dataset consists of 23,488 messages written by 825 authors, sent or received by McCallum during Jan.-Oct., 2004. Gibbs sampling is employed to conduct all experiements (as detailed in [McCallum et al., 2004]).

3.1 Topics and Prominent Relations from ART

Table 1 shows the highest probability words from six topics in an ART model trained on the 147 Enron 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.

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, Hain, the top author of messages in the "Legal Contracts" topic, was an in-house lawyer at Enron. In the "Operations" topic, it is satisfying to see Beck, who was the Chief Operating Officer at Enron. In the "Government Relations" topic, we see Dasovich, who was a Government Relation Executive, Shapiro who was Vice President of Regulatory Affairs, and Steffes, who was Vice President of Government Affairs. Results on the McCallum email dataset are reported in Table 2.

Topic 5 "Legal Contracts"		**Topic 17 "Doc. Review"**		**Topic 27 "Time Scheduling"**	
section	0.0299	attached	0.0742	day	0.0419
party	0.0265	agreement	0.0493	friday	0.0418
language	0.0226	review	0.0340	morning	0.0369
contract	0.0203	questions	0.0257	monday	0.0282
date	0.0155	draft	0.0245	office	0.0282
enron	0.0151	letter	0.0239	wednesday	0.0267
parties	0.0149	comments	0.0207	tuesday	0.0261
notice	0.0126	copy	0.0165	time	0.0218
days	0.0112	revised	0.0161	good	0.0214
include	0.0111	document	0.0156	thursday	0.0191
M.Hain	0.0549	G.Nemec	0.0737	J.Dasovich	0.0340
J.Steffes		B.Tycholiz		R.Shapiro	
J.Dasovich	0.0377	G.Nemec	0.0551	J.Dasovich	0.0289
R.Shapiro		M.Whitt		J.Steffes	
D.Hyvl	0.0362	B.Tycholiz	0.0325	C.Clair	0.0175
K.Ward		G.Nemec		M.Taylor	
Topic 34 "Operations"		**Topic 37 "Power Market"**		**Topic 41 "Gov. Relations"**	
operations	0.0321	market	0.0567	state	0.0404
team	0.0234	power	0.0563	california	0.0367
office	0.0173	price	0.0280	power	0.0337
list	0.0144	system	0.0206	energy	0.0239
bob	0.0129	prices	0.0182	electricity	0.0203
open	0.0126	high	0.0124	davis	0.0183
meeting	0.0107	based	0.0120	utilities	0.0158
gas	0.0107	buy	0.0117	commission	0.0136
business	0.0106	customers	0.0110	governor	0.0132
houston	0.0099	costs	0.0106	prices	0.0089
S.Beck	0.2158	J.Dasovich	0.1231	J.Dasovich	0.3338
L.Kitchen		J.Steffes		R.Shapiro	
S.Beck	0.0826	J.Dasovich	0.1133	J.Dasovich	0.2440
J.Lavorato		R.Shapiro		J.Steffes	
S.Beck	0.0530	M.Taylor	0.0218	J.Dasovich	0.1394
S.White		E.Sager		R.Sanders	

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.

3.2 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. We can measure a similarity symmetrically by calculating the Jensen-Shannon (JS) divergence, and inverting it.

Standard recursive graph-cutting algorithms on this matrix can be used to cluster users, rearranging the rows/columns to

Topic 5 "Grant Proposals"	Topic 31 "Meeting Setup"	Topic 38 "ML Models"	Topic 41 "Friendly Discourse"
proposal	today	model	great
data	tomorrow	models	good
budget	time	inference	don
work	ll	conditional	sounds
year	meeting	methods	work
glenn	week	number	wishes
nsf	talk	sequence	talk
project	meet	learning	interesting
sets	morning	graphical	time
support	monday	random	hear
smyth	ronb	casutton	mccallum
mccallum	mccallum	mccallum	culotta
mccallum	wellner	icml04-webadmin	mccallum
stowell	mccallum	icml04-chairs	casutton
mccallum	casutton	mccallum	mccallum
lafferty	mccallum	casutton	ronb
mccallum	mccallum	nips04workflow	mccallum
smyth	casutton	mccallum	saunders
pereira	mccallum	weinman	mccallum
lafferty	wellner	mccallum	pereira

Table 2: The four topics most prominent in McCallum's email exchange with smyth (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. 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 lafferty (John Lafferty) and pereira (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.

form approximately block-diagonal structures. This is the familiar process of 'blockstructuring' 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.

Beginning with the Enron data, Figure 2 shows the results from traditional SNA (in this case, JS divergence of distributions on recipients from each sender), ART (JS divergence of recipient-marginalized topic distributions for each sender) and AT (using the topics distributions from AT instead of our ART). Darker shading indicates higher similarity between people.

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 data reveals that Geaconne was an Executive Assistant, while McCarty was a Vice-

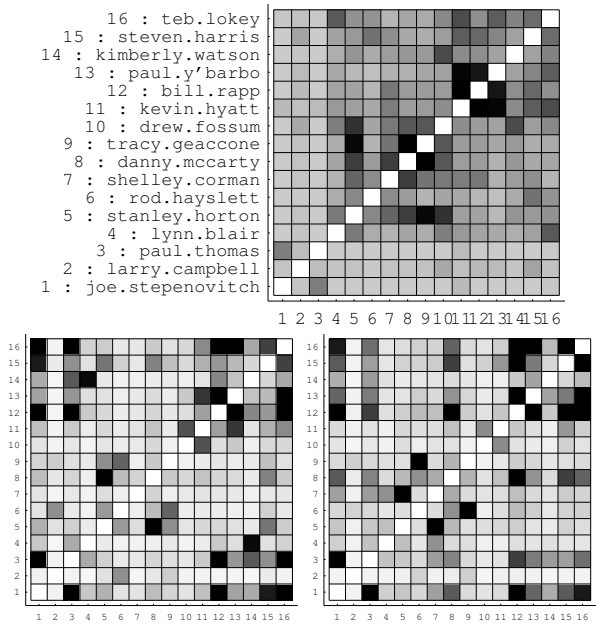

Figure 2: **Top:** SNA Inverse JS Network. **Left Bottom:** ART Inverse JS Network. **Right Bottom:** AT Inverse JS Network. Darker shades indicate higher similarity.

President—rather different roles—and, thus output of ART and AT is more appropriate. Thus, SNA analysis shows that they wrote email to similar sets of people, but the ART analysis illustrates that they used very different language.

Here ART and AT provide similar role distance, but they 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 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.

Based on the above examples, and other similar examples, we 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 not at all. For example, ART matches mike and mikem are actually two different email addresses for the same person. (Most

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 mitchell	Both ML researchers on SRI project
Pairs considered most alike by SNA	
User Pair	*Description*
aepshtey rasmith	Both students in McCallum's class
donna editor	Spouse is unrel. to journal editor
donna krishna	Spouse is unrel. to conf. organizer
donna ramshaw	Spouse is unrel. to researcher at BBN
donna reviews	Spouse is unrel. to journal 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 (donna) and the JMLR editor.

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 coe and laurie are both UMass CS Department administrative assistants; they rarely send email to the same people, but they write about similar things. 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.

4 Role-Author-Recipient-Topic Models

To better explore the roles of authors, an additional level of latent variables can be introduced to explicitly model roles. Of particular interest is capturing the notion that a person can have multiple *roles* simultaneously—for example, a person can be both a professor and a mountain climber. 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; climbers' topics may prominently feature mountains, climbing equipment, and also meeting times and friendly relations.

We incorporate into the ART model a new set of variables that take on values indicating role, and we term this augmented model 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 some role assignments, and 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 roles. Each sender-role and recipient-role pair has a multinomial distribution over topics, and each topic has a multinomial distribution over words.

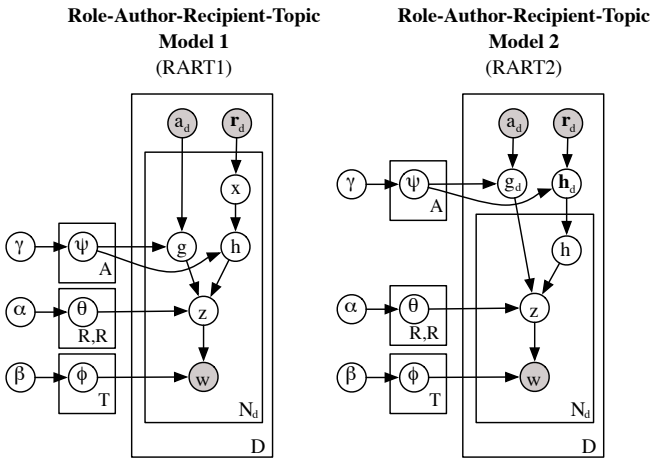

Figure 3: Two possible variants for the Role-Author-Recipient-Topic (RART) model.

As shown in Figure 3, 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 (with corresponding role) is selected on which to condition the selection of topic. Some other variants are possible, for example (not shown in Figure 3), 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.

The generative process of RART models is similar to that for ART, as described in Section 2. The Gibbs sampling formulae for RART models can be derived in the same way as for ART, but in a more complex form.

5 Experimental Results with RART

Preliminary experiments have been conducted with the RART1 model. Because we introduce two sets of additional latent variables (author role and recipient role), the sampling procedure at each iteration is significantly more complex. To make inference more efficient, we can instead perform it in two distinct parts. One strategy we have found useful is to first train an ART model, and use a sample to obtain topic assignments and recipient assignments for each word token. Then, in the next stage, we treat topics and recipients as observed (locked). Although such a strategy may not be recommended for arbitrary graphical models, we feel this is rea-

sonable here because we find that a single sample from Gibbs sampling on the ART model to yield good assignments. The following results are based on a 15-group, 50-topic run of RART1 on McCallum email dataset.

Our results show that the RART model indeed clearly discovers automatically person-role information by its explicit inclusion of a role variables. For example, the users most prominent in Role 3 include, (in probability order): olc, gauthier, irsystem, system, allan, valerie, tech, and steve. They are all employees (or mailing lists) of the IT support staff at UMass CS, except for allan, who, however, was the professor chairing the department's computing committee. Role 4 seems to represent "working on the SRI CALO project." Its most prominent members include (in probability order): pereira, claire, israel, moll, mgervasio, melinda.gervasio, majordomo, and colin.evans. Most of them are researchers working on CALO project, many of them at SRI. The sender majordomo sends messages from an SRI CALO mailing list. The users mgervasio and melinda.gervasio are actually the same person; satisfyingly RART found that they have very similar role distributions.

One objective of the RART model is to capture the multiple roles that a person has. For example, user allan (James Allan) mentioned above has a role in "IT support," but also has a role as "researcher." Consider also user pereira (Fernando Pereira); his top five role assginments are Role 2 "NLP research", Role 4 "SRI CALO", Role 6 "proposal writing," Role 10 "grant issues," and Role 8 "guests at McCallum's house"—all exactly appropriate, as viewed through McCallum's email. Note that the difference between roles can be subtle, for example, Role 6 and Role 10 overlap.

As expected, one can observe interesting differences in the sender versus recipient topic distributions associated with each role. For instance, in Role 4 "SRI CALO," the top three topics for a sender role are Topic 27 "CALO information," Topic 11 "mail accounts," and Topic 36 "program meetings," but for its recipient roles, most prominent are Topic 48 "task assignments," Topic 46 "a particular CALO-related research paper," and Topic 40 "java code". Space limitations prevent inclusion of tables showing the full distributions associated many topics, roles and people, however these will be available in an accompanying technical report.

6 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.

The Role-Author-Recipient-Topic (RART) models explicitly capture the multiple roles of people, based on messages sent and received. Additional work on 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 Central Intelligence Agency, the National Security Agency, 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

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.

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*.

McCallum, A., Corrada-Emmanuel, A., & Wang, X. (2004). *The author-recipient-topic model for topic and role discovery in social networks: Experiments with enron and academic email* (Technical Report). University of Massachussets, Amherst, UM-CS-2004-096.

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*.

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*. Cambridge University Press.

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

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*.

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