Journal of Artificial Intelligence Research 30 (2007) 249-272 Submitted 12/06; published 10/07

Topic and Role Discovery in Social Networks with Experiments on Enron and Academic Email

Andrew McCallum MCCALLUM@CS.UMASS.EDU
Xuerui Wang XUERUI@CS.UMASS.EDU
Department of Computer Science
University of Massachusetts
140 Governors Drive
Amherst, MA 01003 USA

Andrés Corrada-Emmanuel ACORRADA@PHYSICS.UMASS.EDU
Department of Physics
University of Massachusetts
666 North Pleasant Street
Amherst, MA 01003 USA

Abstract

Previous work in social network analysis (SNA) has modeled the existence of links from one entity to another, but not the attributes such as 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 and gives lower perplexity on previously unseen messages. We also present the Role-Author-Recipient-Topic (RART) model, an extension to ART that explicitly represents 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 data sets of human interactions (Shetty & Adibi, 2004; Wu, Huberman, Adamic, & Tyler, 2003), the popularity of services like MySpace.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 (group or block) of the network (Nowicki & Snijders, 2001; Kemp, Griffiths, & Tenenbaum, 2004; Kemp, Tenenbaum, Griffiths, Yamada, & Ueda, 2006; Kubica, Moore, Schneider, & Yang, 2002; Airoldi, Blei, Fienberg, & Xing, 2006; Kurihara, Kameya, & Sato, 2006). Furthermore, using these properties we can assign "roles" to certain nodes (Lorrain & White, 1971; Wolfe & Jensen, 2004). 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 in which 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. Probabilistic Latent Semantic Indexing (Hofmann, 2001) and Latent Dirichlet Allocation (Blei, Ng, & Jordan, 2003) robustly discover multinomial word distributions of these topics. Hierarchical Dirichlet Processes (Teh, Jordan, Beal, & Blei, 2004) can determine an appropriate number of topics for a corpus. The Author-Topic Model (Steyvers, Smyth, Rosen-Zvi, & Griffiths, 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 multinomial 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

1. The clustering may be 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.

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 9 months of incoming and outgoing mail of the first author, comprising 825 people and 14k 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. Also, we show that the similarity matrix produced by ART is different from both the SNA matrix and the AT matrix in several appropriate ways. Furthermore, we find that the ART model gives a significantly lower perplexity on previously unseen messages than AT, which shows that ART is a better topic model for email messages.

We also describe an extension of the ART model that explicitly captures *roles* of people, by generating role associations for the author and recipient(s) of a message, and conditioning the topic distributions on the 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 experiments with one of these variants.

The importance of modeling the *language* associated with social network interactions has also recently been demonstrated in the Group-Topic (GT) model (Wang, Mohanty, & McCallum, 2006). Unlike ART, which discovers roles, GT discovers groups. Like ART, it uses text data to find interesting and useful patterns that would not be possible with edge relations alone. GT simultaneously clusters entities into groups that share similar interaction patterns, and also clusters text (or other attributes) of their interactions into topics—doing so in such a way that clustering in each dimension informs the other. When applied to the voting records and corresponding text of resolutions from the U.S. Senate and the U.N., the Group-Topic model shows that incorporating the votes results in more salient topic clusters, and that different groupings of legislators emerge from different topics. Both role discovery and group discovery are primary areas of SNA research.

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 author's interests are modeled with a *mixture* of topics (Steyvers et al., 2004). In

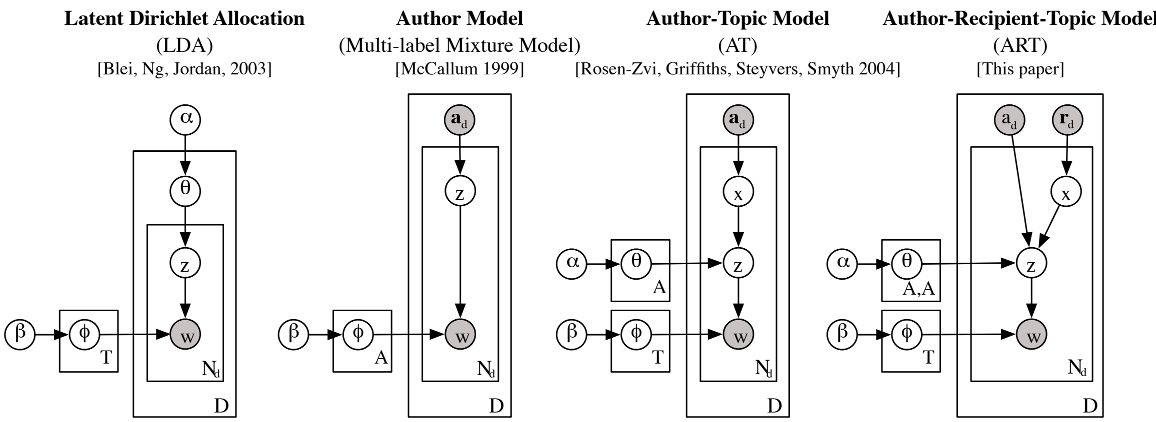


Figure 1: Three related models, and the ART model. In all models, each observed word, w, is generated from a multinomial word distribution, ϕ_z, specific to a particular topic/author, z, however topics are selected differently in each of the models. In LDA, the topic is sampled from a per-document topic 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, the topic is sampled from a per-author multinomial distribution, θ, and authors are sampled uniformly 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 a recipient from the set of recipients for the document.

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.

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 the sender-recipient relationships.

SYMBOL	DESCRIPTION
T	number of topics
D	number of email messages
A	number of email accounts (senders and recipients)
V	number of unique words (vocabulary size)
N_d	number of word tokens in message d

Table 1: Notation used in this paper

Thus, we propose an Author-Recipient-Topic (ART) model for email messages. 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 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). 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 graphical model representations for all models are shown in Figure 1. In the ART model, given the hyper-parameters α and β, an author a_d, and a set of recipients $\mathbf{r_d}$ for each message d, the joint distribution of the topic mixture θ_{ij} for each author-recipient pair (i, j), the word mixture ϕ_t for each topic t, 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:

$$P(\Theta, \Phi, \mathbf{x}, \mathbf{z}, \mathbf{w}|\alpha, \beta, \mathbf{a}, \mathbf{r}) = \prod_{i=1}^{A}\prod_{j=1}^{A} p(\theta_{ij}|\alpha) \prod_{t=1}^{T} p(\phi_t|\beta) \prod_{d=1}^{D}\prod_{i=1}^{N_d} (P(x_{di}|\mathbf{r}_d) P(z_{di}|\theta_{a_d x_{di}}) P(w_{di}|\phi_{z_{di}}))$$

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

$$\begin{aligned} & P(\mathbf{w}|\alpha, \beta, \mathbf{a}, \mathbf{r}) \\ = & \iint \prod_{i=1}^{A}\prod_{j=1}^{A} p(\theta_{ij}|\alpha) \prod_{t=1}^{T} p(\phi_t|\beta) \prod_{d=1}^{D}\prod_{i=1}^{N_d} \sum_{x_{di}=1}^{A} (P(x_{di}|\mathbf{r}_d) \sum_{z_{di}=1}^{T} (P(z_{di}|\theta_{a_d x_{di}}) P(w_{di}|\phi_{z_{di}}))) \mathrm{d}\Phi \mathrm{d}\Theta \end{aligned}$$

2.1 Inference by Gibbs Sampling

Inference on models in the LDA family cannot be performed exactly. Three standard approximate inference methods have been used to obtain practical results: variational methods

Algorithm 1 Inference and Parameter Estimation in ART

initialize the author and topic assignments randomly for all tokens
repeat
  for $d = 1$ to D **do**
    for $i = 1$ to N_d **do**
      draw x_{di} and z_{di} from $P(x_{di}, z_{di}|\mathbf{x}_{-di}, \mathbf{z}_{-di}, \mathbf{w}, \alpha, \beta, \mathbf{a}, \mathbf{r})$
      update $n_{a_d x_{di} z_{di}}$ and $m_{z_{di} w_{di}}$
    end for
  end for
until the Markov chain reaches its equilibrium
compute the posterior estimates of θ and ϕ

(Blei et al., 2003), Gibbs sampling (Griffiths & Steyvers, 2004; Steyvers et al., 2004; Rosen-Zvi, Griffiths, Steyvers, & Smyth, 2004), and expectation propagation (Griffiths & Steyvers, 2004; Minka & Lafferty, 2002). We choose Gibbs sampling for its ease of implementation. Note that we adopt conjugate priors (Dirichlet) for the multinomial distributions, and thus we can easily integrate out θ and ϕ, analytically capturing the uncertainty associated with them. In this way we facilitate the sampling—that is, we need not sample θ and ϕ at all. One could estimate the values of the hyper-parameters of the ART model, α and β, from data using a Gibbs EM algorithm (Andrieu, de Freitas, Doucet, & Jordan, 2003). In some applications, topic models are very sensitive to hyper-parameters, and it is extremely important to set the right values for the hyper-parameters. However, in the particular applications discussed in this paper, after trying out many different hyper-parameter settings, we find that the sensitivity to hyper-parameters is not very strong. Thus, again for simplicity, we use fixed symmetric Dirichlet distributions ($\alpha = 50/T$ and $\beta = 0.1$) in all our experiments.

We need to derive $P(x_{di}, z_{di}|\mathbf{x}_{-di}, \mathbf{z}_{-di}, \mathbf{w}, \alpha, \beta, \mathbf{a}, \mathbf{r})$, the conditional distribution of a topic and recipient for the word w_{di} given all other words' topic and recipient assignments, $\mathbf{x}_{-di}$ and $\mathbf{z}_{-di}$, to carry out the Gibbs sampling procedure for ART. We begin with the joint probability of the whole data set, and by the chain rule, the above conditional probability can be obtained with ease:

$$P(x_{di}, z_{di}|\mathbf{x}_{-di}, \mathbf{z}_{-di}, \mathbf{w}, \alpha, \beta, \mathbf{a}, \mathbf{r}) \propto \frac{\alpha_{z_{di}} + n_{a_d x_{di} z_{di}} - 1}{\sum_{t=1}^{T}(\alpha_t + n_{a_d x_{di} t}) - 1} \frac{\beta_{w_{di}} + m_{z_{di} w_{di}} - 1}{\sum_{v=1}^{V}(\beta_v + m_{z_{di} v}) - 1}$$

where n_{ijt} is the number of tokens assigned to topic t and the author-recipient pair (i, j), and m_{tv} represent the number of tokens of word v assigned to topic t.

The posterior estimates of θ and ϕ given the training set can be calculated by

$$\hat{\theta}_{ijz} = \frac{\alpha_z + n_{ijz}}{\sum_{t=1}^{T}(\alpha_t + n_{ijt})}, \hat{\phi}_{tw} = \frac{\beta_w + m_{tw}}{\sum_{v=1}^{V}(\beta_v + m_{tv})} \tag{1}$$

Detailed derivation of Gibbs sampling for ART is provided in Appendix A. An overview of the Gibbs sampling procedure we use is shown in Algorithm 1.

3. Related Work

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

The limitation of a single class/role label for each node in a network is relaxed in recent work by Wolfe and Jensen (2004). 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 in a non-factored model using a 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) look 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 in an email network. Adamic and Adar (2004) study 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, Dodds, and Newman (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" among nodes to create language models that can bring out differences invisible at the network level.

As discussed in the introduction, we have also recently developed a model for group discovery. In addition to relation-edge data, our Group-Topic (GT) model also takes into consideration the textual attributes of relations, and allows the discovery of groups to be guided by emerging textual topics and vice-versa (Wang et al., 2006). Experiments on voting data show the Group-Topic model's joint inference improves both the groups and topics discovered. Other modalities of information can be combined to discover hidden structure. For example, time and text are combined in the Topics over Time (TOT) model (Wang & McCallum, 2006), which finds trends in time-sensitive topics using a continuous distribution over time-stamps. Dynamic Topic Models (Blei & Lafferty, 2006b) incorporate time into topic models through transitions in a Markov process. The ART model could be easily extended to incorporate temporal information.

As discussed earlier, 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), with the distinction that ART is specifically designed to capture language used in a directed network of correspondents. Another more recent model that associates topics with people is the Author-Persona-Topic (APT) model (Mimno & McCallum, 2007). APT is designed specifically to capture the expertise of a person, modeling expertise as a mixture of topical intersections, and is demonstrated on the task of matching reviewers to submitted research papers.

New topic models have been actively studied in recent years for many different tasks, including joint modeling of words and research paper citations (Erosheva, Fienberg, & Lafferty, 2004), capturing correlations among topics (Blei & Lafferty, 2006a; Li & McCallum, 2006), taking advantage of both topical and syntactic dependencies (Griffiths, Steyvers, Blei, & Tenenbaum, 2004), and discovering topically-relevant phrases by Markov dependencies in word sequences (Wang, McCallum, & Wei, 2007). Many of these models could be easily combined with the ART model, and would likely prove useful.

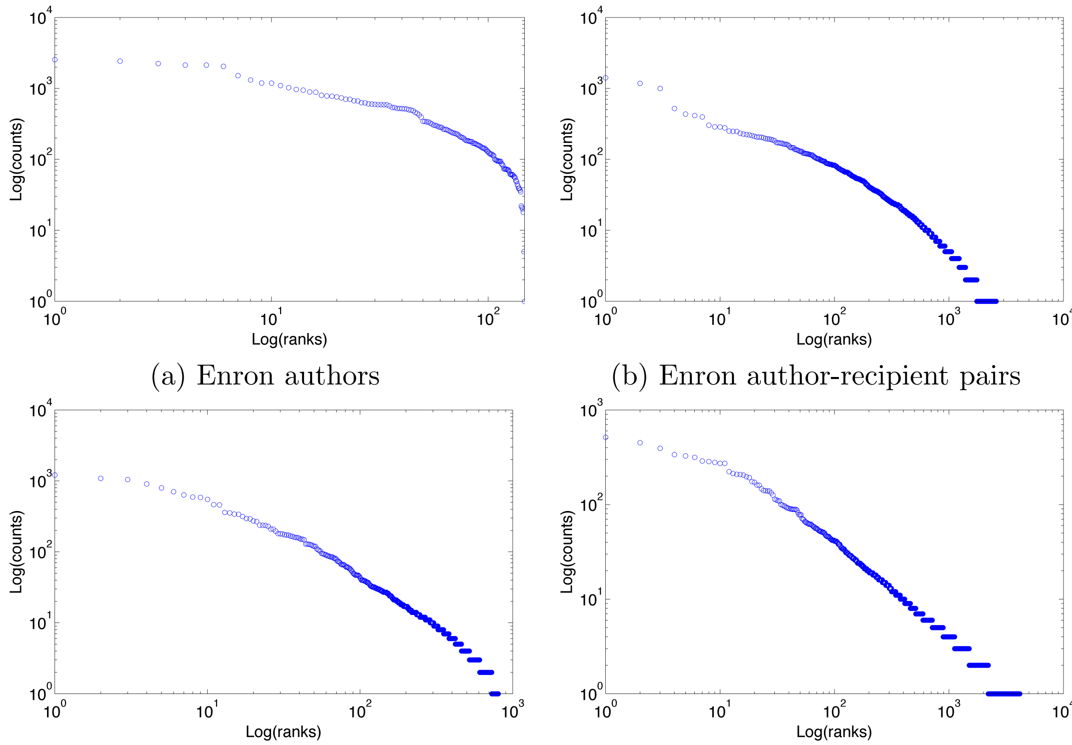
4. Experimental Results

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 data set 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 data set 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' emails 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 users.

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 time stamp is removed. This heuristic certainly incorrectly looses words that are interspersed with quoted email text. Only words formed as sequences of alphabetic characters are kept, which results in a vocabulary of 22,901 unique words. To remove sensitivity to capitalization, all text is downcased.

Our second data set consists of the personal email sent and received by McCallum between January and September 2004. It consists of 13,633 unique messages written by 825 authors. In a 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 set, we obtained a text corpus containing 457,057 word tokens, and a vocabulary of 22,901 unique words.



(a) Enron authors (b) Enron author-recipient pairs

(c) McCallum authors (d) McCallum author-recipient pairs

Figure 2: Power-law relationship between the frequency of occurrence of of an author (or an author-recipient pair) and the rank determined by the above frequency of occurrence. In the author plots, we treat both the sender and the recipients as authors.

By conditioning topic distributions on author-recipient pairs instead of authors, the data we have may look sparser considering that we have substantially more author-recipient pairs than authors. However, as shown in Figure 2, we can find that the number of emails of an author-recipient pair and its rank determined by the count still follow a power-law behavior, as for authors. For example, in the McCallum data set, 500 of possible 680,625 author-recipient pairs are responsible for 70% of the email exchange. That is, even though the data are sparser for the ART model, the power-law behavior makes it still possible to obtain a good estimation of the topic distributions for prominent author-recipient pairs.

We initialize the Gibbs chains on both data sets randomly, and find that the results are very robust to different initializations. By checking the perplexity, we find that usually the Gibbs chain converges after a few hundred iterations, and we run 10,000 iterations anyway to make sure it converges.

4.1 Topics and Prominent Relations from ART

Table 2 shows the highest probability words from eight topics in an ART model trained on the 147 Enron users with 50 topics. The quoted titles are our own interpretation of

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 2: An illustration of several topics from a 50-topic run for the Enron email data set. 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 football 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.

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 (Document Review) comes from the messages discussing review and comments on documents; Topic 27 (Time Scheduling) comes from the 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, Hain, the top author of messages in the "Legal Contracts" topic, was an in-house lawyer at Enron. By inspection of messages related to "Sports Pool", Eric Bass seems to have been the coordinator for a fantasy football league among Enron employees. 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. Results on the McCallum email data set are reported in Table 3.

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, a 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 'blockstructuring' used in SNA. We perform such an analysis on two data sets: 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.

We begin with the Enron TransWestern Pipeline Division. Our analysis here employed a "closed-universe" assumption—only those messages traded among considered authors in the data set 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 3. 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 Figure 3. 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.

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 3: The four topics most prominent in McCallum's email exchange with Padhraic Smyth, from a 50-topic run of ART on 9 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.

Consider Enron employee Geaccone (user 9 in all the matrices in Figure 3). 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 the 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.

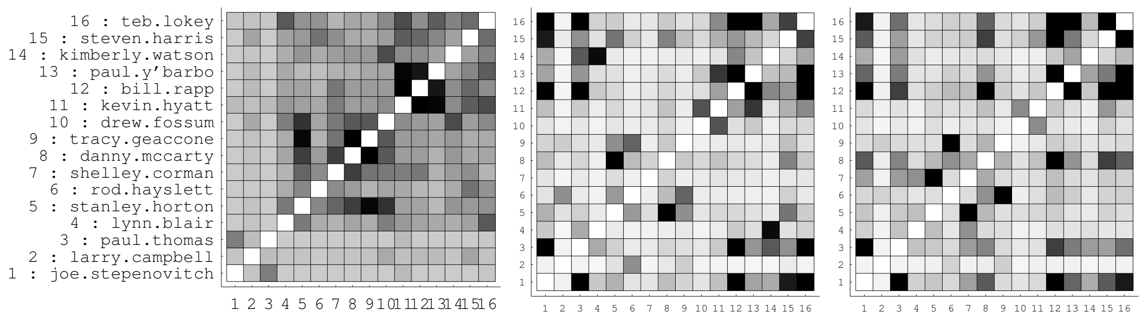


Figure 3: **Left:** SNA Inverse JS Network. **Middle:** ART Inverse JS Network. **Right:** AT Inverse JS Network. Darker shades indicate higher similarity.

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, traditional SNA results (Figure 3 left) 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 separate work we have also developed the Group-Topic (GT) model, which explicitly discovers groups in a way that leverages accompanying text (Wang et al., 2006). In the future we may also develop 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 SNA and AT in predicting role similarity. We thus would claim

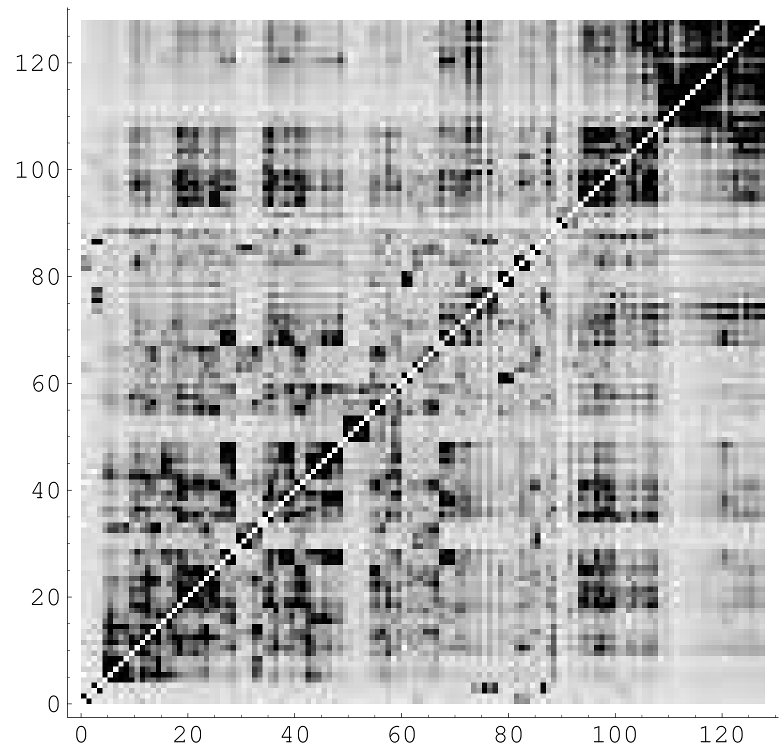


Figure 4: SNA Inverse JS Network for a 10 topic run on McCallum Email Data. Darker shades indicate higher similarity. Graph partitioning was calculated with the 128 authors that had ten or more emails in McCallum's Email Data. The block from 0 to 30 are people in and related to McCallum's research group at UMass. The block from 30 to 50 includes other researchers around the world.

that the ART model yields more appropriate results 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, while only 128 wrote ten or more emails. We perform the blockstructure analysis with these 128 users, shown in Figure 4. The blocks discovered are quite meaningful, e.g., the block from 0 to 30 are people in and related to McCallum's research group at UMass, and the block from 30 to 50 includes other researchers around the world.

Table 4 shows the closest pairs in terms of JS divergence, as calculated by the ART model and the 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 coreferent email

Pairs considered most alike by ART	
User Pair	*Description*
editor reviews	Both journal review management
mike mikem	Same person! (manual coreference 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 organizing 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 4: 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.

addresses were pre-collapsed by hand during preprocessing; here ART has pointed 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

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 5: 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.

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 was the Production Editor for the Journal of Machine Learning Research.

To highlight the difference between the SNA and ART predictions, we present Table 5, which was obtained by using both ART and SNA to rank the pairs of people by similarity, and then listing the pairs 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.

4.3 Perplexity Comparison between AT and ART

Models for natural languages are often evaluated by perplexity as a measure of the goodness of fit of models. The lower perplexity a language model has, the better it predicts the unseen words given the words we previously saw.

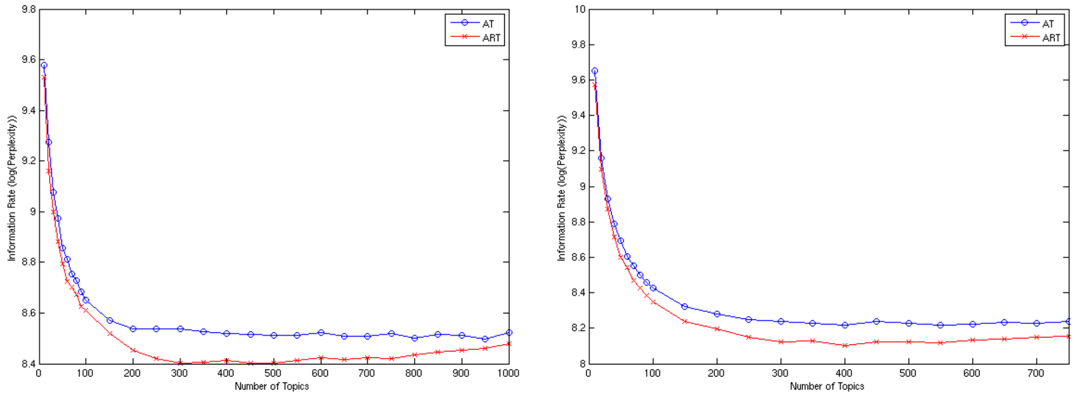
The perplexity of a previously unseen message d consisting of words $\mathbf{w}_d$ can be defined as follows, when the author a_d and the recipient(s) $\mathbf{r}_d$ are given:

$$\text{Perplexity}(\mathbf{w}_d) = \exp\left(-\frac{\log(p(\mathbf{w}_d|a_d, \mathbf{r}_d))}{N_d}\right),$$

where ($\hat{\theta}$ and $\hat{\phi}$ defined in Equation 1)

$$p(\mathbf{w}_d|a_d, \mathbf{r}_d) = \prod_{i=1}^{N_d}\left(\frac{1}{|\mathbf{r}_d|}\sum_{r\in\mathbf{r}_d}\sum_{t=1}^{T}\hat{\theta}_{a_d r t}\hat{\psi}_{t w_{di}}\right).$$

We randomly split our data sets into a training set (9/10) and a test set (the remaining 1/10). In the test sets, 92.37% (Enron) and 84.51% (McCallum) of the author-recipient pairs also appear in the training sets. Ten Markov chains are run with different initializations,



(a) Enron data set (b) McCallum data set

Figure 5: Perplexity comparison of AT and ART on two data sets. We plot the information rate (logarithm of perplexity) here. The difference between AT and ART is significant under one-tailed t-test (Enron data set: p-value < 0.01 except for 10 topics with p-value $= 0.018$; McCallum data set: p-value $< 1e - 5$).

and the samples at the 2000th iteration are used to estimate $\hat{\theta}$ and $\hat{\phi}$ by Equation 1. We report the average information rate (logarithm of perplexity) with different number of topics on two data sets in Figure 5.

As clearly shown in the figure, ART has significantly better predictive power than AT over a large number of randomly selected test documents on both data sets under one-tailed t-test. Particularly on the Enron data set, ART uses much fewer number of topics to achieve the best predictive performance. We can also find that the lowest perplexity obtained by ART is not achievable by AT with any parameter setting on both data sets. Both these results provide evidence that ART discovers meaningful topics in the context of a social network and is indeed more appropriate to message data than AT.

Here we do not compare perplexity between ART and LDA, however AT (which ART dominates in perplexity) has already been shown to have better perplexity than LDA (Rosen-Zvi, Griffiths, Smyth, & Steyvers, 2005). Due to the much simpler model structure, the author model (McCallum, 1999) has much worse perplexity. Measured on both data sets, the information rates (log perplexity) are larger than 10, whereas ART's information rates are mostly between 8 and 9.

5. 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.

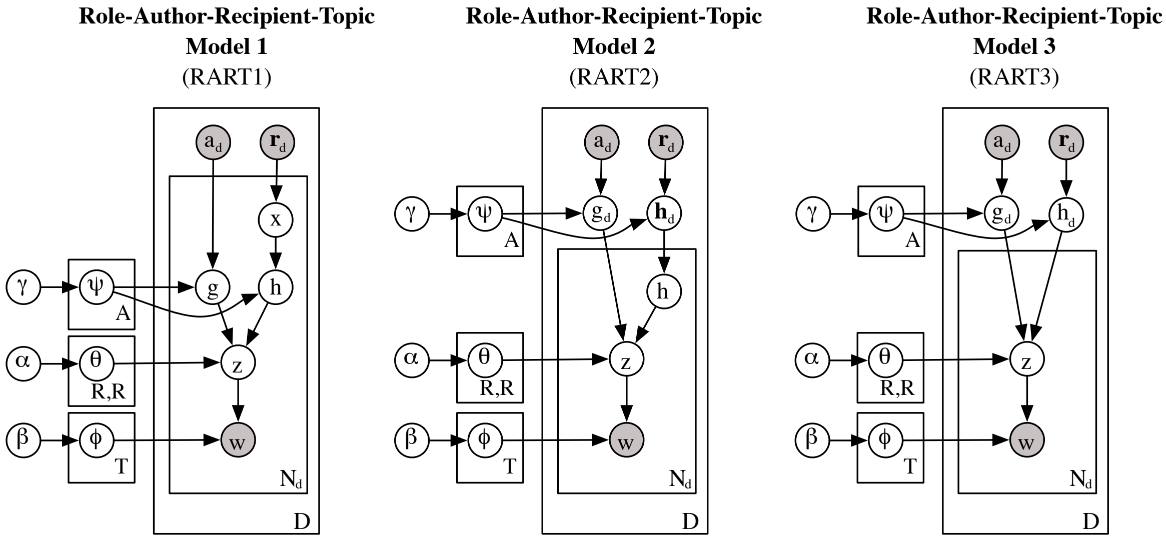


Figure 6: Three possible variants for the Role-Author-Recipient-Topic (RART) model.

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.

As shown in Figure 6, 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. 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 exploration elsewhere. 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 θ_{gh}, where the distribution is specific to the

Role 3 "IT Support at UMass CS"		Role 4 "Working on the SRI CALO Project"	
olc (lead Linux sysadmin)	0.2730	pereira (prof. at UPenn)	0.1876
gauthier (sysadmin for CIIR group)	0.1132	claire (UMass CS business manager)	0.1622
irsystem (mailing list CIIR sysadmins)	0.0916	israel (lead system integrator at SRI)	0.1140
system (mailing list for dept. sysadmins)	0.0584	moll (prof. at UMass)	0.0431
allan (prof., chair of computing committee)	0.0515	mgervasio (computer scientist at SRI)	0.0407
valerie (second Linux sysadmin)	0.0385	melinda.gervasio (same person as above)	0.0324
tech (mailing list for dept. hardware)	0.0360	majordomo (SRI CALO mailing list)	0.0210
steve (head of dept. of IT support)	0.0342	collin.evans (computer scientist at SRI)	0.0205

Table 6: An illustration of two roles from a 50-topic, 15-group run for the McCallum email data set. Each role is shown with the most prominent users (their short descriptions in parenthesis) and the corresponding conditional probabilities. The quoted titles are our own summary for the roles. For example, in Role 3, the users 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.

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, given the hyper-parameters α, β and γ, an author a_d, and a set of recipients $\mathbf{r}_d$ for each message d, the joint distribution of the topic mixture θ_{ij} for each author-role recipient-role pair (i, j), the role mixture ψ_k for each author k, the word mixture ϕ_t for each topic t, a set of recipients $\mathbf{x}$, a set of sender roles $\mathbf{g}$, a set of recipient roles $\mathbf{h}$, a set of topics $\mathbf{z}$ and a set of words $\mathbf{w}$ is given by (we define R as the number of roles):

$$
\begin{aligned}
& P(\Theta, \Phi, \Psi, \mathbf{x}, \mathbf{g}, \mathbf{h}, \mathbf{z}, \mathbf{w} | \alpha, \beta, \gamma, \mathbf{a}, \mathbf{r}) \\
= & \prod_{i=1}^{R} \prod_{j=1}^{R} p(\theta_{ij}|\alpha) \prod_{t=1}^{T} p(\phi_t|\beta) \prod_{k=1}^{A} p(\psi_k|\gamma) \prod_{d=1}^{D} \prod_{i=1}^{N_d} P(x_{di}|\mathbf{r}_d) P(g_{di}|a_d) P(h_{di}|x_{di}) P(z_{di}|\theta_{g_{di}h_{di}}) P(w_{di}|\phi_{z_{di}})
\end{aligned}
$$

Integrating over Ψ, Θ and Φ, and summing over $\mathbf{x}$, $\mathbf{g}$, $\mathbf{h}$ and $\mathbf{z}$, we get the marginal distribution of a corpus, similar to what we showed for ART.

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

6. Experimental Results with RART

Extensive 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 reasonable here because we find that a single sample from Gibbs

allan (James Allan)		**pereira (Fernando Pereira)**	
Role 10 (grant issues)	0.4538	Role 2 (natural language researcher)	0.5749
Role 13 (UMass CIIR group)	0.2813	Role 4 (working on SRI CALO Project)	0.1519
Role 2 (natural language researcher)	0.0768	Role 6 (proposal writing)	0.0649
Role 3 (IT Support at UMass CS)	0.0326	Role 10 (grant issues)	0.0444
Role 4 (working on SRI CALO Project)	0.0306	Role 8 (guests at McCallum's house)	0.0408

Table 7: An illustration of the role distribution of two users from a 50-topic, 15-group run for the McCallum email data set. Each user is shown with his most prominent roles (their short descriptions in parenthesis) and the corresponding conditional probabilities. For example, considering user *pereira* (Fernando Pereira), his top five role assignments are all appropriate, as viewed through McCallum's email.

sampling on the ART model yields good assignments. The following results are based on a 15-group, 50-topic run of RART1 on McCallum email data set.

Our results show that the RART model does indeed automatically discover meaningful person-role information by its explicit inclusion of a role variable. We show the most prominent users in two roles in Table 6. For instance, the users most prominent in Role 3 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." Most of its top prominent members are researchers working on CALO project, many of them at SRI. The sender *majordomo* sends messages from an SRI CALO mailing list. Users *claire* and *moll* were, however, unrelated with the project, and we do not know the reason they appear in this role. 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. The role distribution of two users are shown in Table 7. For example, user *allan* (James Allan) mentioned above has a role in "IT support," but also has a role as a "member of the Center for Intelligent Information Retrieval," as a "grant proposal writer," and as a "natural language researcher." Although not a member of the "SRI CALO Project," *allan*'s research is related to CALO, and perhaps this is the reason that CALO appears (weakly) among his roles. Consider also user *pereira* (Fernando Pereira); his top five role assignments are all exactly appropriate, as viewed through McCallum's email.

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 CALO-related research paper," and Topic 40 "java code".

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.

The Role-Author-Recipient-Topic (RART) models explicitly capture the multiple roles of people, based on messages sent and received. Future work will develop models that explicitly capture both roles and groups.

Acknowledgments

Some of the material in this paper was presented in part at the 19th International Joint Conference on Artificial Intelligence (IJCAI 2005) in Edinburgh, Scotland, July 30–August 5, 2005. 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.

Appendix A. Gibbs Sampling Derivation for ART

We need to derive $P(x_{di}, z_{di}|\mathbf{x}_{-di}, \mathbf{z}_{-di}, \mathbf{w}, \alpha, \beta, \mathbf{a}, \mathbf{r})$, the conditional distribution of a topic and recipient for the word w_{di} given all other words' topic and recipient assignments, $\mathbf{x}_{-di}$ and $\mathbf{z}_{-di}$, to carry out the Gibbs sampling procedure for ART. We begin with the joint probability of the whole data set. Note here that we can take advantage of conjugate priors to simplify the integrals.

$$\begin{aligned}
& P(\mathbf{x}, \mathbf{z}, \mathbf{w}|\alpha, \beta, \mathbf{a}, \mathbf{r}) \\
= & \iint \prod_{i=1}^{A}\prod_{j=1}^{A} p(\theta_{ij}|\alpha) \prod_{t=1}^{T} p(\phi_t|\beta) \prod_{d=1}^{D}\prod_{i=1}^{N_d} P(x_{di}|\mathbf{r}_d) \cdot P(z_{di}|\theta_{a_d x_{di}}) P(w_{di}|\phi_{z_{di}}) \mathrm{d}\Phi \mathrm{d}\Theta \\
= & \prod_{d=1}^{D} (\frac{1}{|\mathbf{r}_d|})^{N_d} \int \prod_{i=1}^{A}\prod_{j=1}^{A} \left(\frac{\Gamma(\sum_{t=1}^{T}\alpha_t)}{\prod_{t=1}^{T}\Gamma(\alpha_t)} \prod_{t=1}^{T} \theta_{ijt}^{\alpha_t - 1} \right) \prod_{i=1}^{A}\prod_{j=1}^{A}\prod_{t=1}^{T} \theta_{ijt}^{n_{ijt}} \mathrm{d}\Theta \\
& \times \int \prod_{t=1}^{T} \left(\frac{\Gamma(\sum_{v=1}^{V}\beta_v)}{\prod_{v=1}^{V}\Gamma(\beta_v)} \prod_{v=1}^{V} \phi_{tv}^{\beta_v - 1} \right) \prod_{t=1}^{T}\prod_{v=1}^{V} \phi_{tv}^{m_{tv}} \mathrm{d}\Phi \\
\propto & \prod_{i=1}^{A}\prod_{j=1}^{A} \int \prod_{t=1}^{T} \theta_{ijt}^{\alpha_t + n_{ijt} - 1} \mathrm{d}\theta_{ij} \prod_{t=1}^{T} \int \prod_{v=1}^{V} \phi_{tv}^{\beta_v + m_{tv} - 1} \mathrm{d}\phi_t
\end{aligned}$$

$$\propto \prod_{i=1}^{A}\prod_{j=1}^{A} \frac{\prod_{t=1}^{T}\Gamma(\alpha_t + n_{ijt})}{\Gamma(\sum_{t=1}^{T}(\alpha_t + n_{ijt}))} \prod_{t=1}^{T} \frac{\prod_{v=1}^{V}\Gamma(\beta_v + m_{tv})}{\Gamma(\sum_{v=1}^{V}(\beta_v + m_{tv}))}$$

where $|\mathbf{r}_d|$ is the number of recipients in message d, n_{ijt} is the number of tokens assigned to topic t and the author-recipient pair (i, j), and m_{tv} represent the number of tokens of word v assigned to topic t.

Using the chain rule, we can obtain the conditional probability conveniently. We define $\mathbf{w}_{-di}$ as all word tokens except the token w_{di}.

$$\begin{aligned}
& P(x_{di}, z_{di}|\mathbf{x}_{-di}, \mathbf{z}_{-di}, \mathbf{w}, \alpha, \beta, \mathbf{a}, \mathbf{r}) \\
= \; & \frac{P(x_{di}, z_{di}, w_{di}|\mathbf{x}_{-di}, \mathbf{z}_{-di}, \mathbf{w}_{-di}, \alpha, \beta, \mathbf{a}, \mathbf{r})}{P(w_{di}|\mathbf{x}_{-di}, \mathbf{z}_{-di}, \mathbf{w}_{-di}, \alpha, \beta, \mathbf{a}, \mathbf{r})} \propto \frac{P(\mathbf{x}, \mathbf{z}, \mathbf{w}|\alpha, \beta, \mathbf{a}, \mathbf{r})}{P(\mathbf{x}_{-di}, \mathbf{z}_{-di}, \mathbf{w}_{-di}|\alpha, \beta, \mathbf{a}, \mathbf{r})} \\
\propto \; & \frac{\frac{\Gamma(\alpha_{z_{di}} + n_{a_d x_{di} z_{di}})}{\Gamma(\alpha_{z_{di}} + n_{a_d x_{di} z_{di}} - 1)}}{\frac{\Gamma(\sum_{t=1}^{T}(\alpha_t + n_{a_d x_{di} t}))}{\Gamma(\sum_{t=1}^{T}(\alpha_t + n_{a_d x_{di} t}) - 1)}} \frac{\frac{\Gamma(\beta_{w_{di}} + m_{z_{di} w_{di}})}{\Gamma(\beta_{w_{di}} + m_{z_{di} w_{di}} - 1)}}{\frac{\Gamma(\sum_{v=1}^{V}(\beta_v + m_{z_{di} v}))}{\Gamma(\sum_{v=1}^{V}(\beta_v + m_{z_{di} v}) - 1)}} \propto \frac{\alpha_{z_{di}} + n_{a_d x_{di} z_{di}} - 1}{\sum_{t=1}^{T}(\alpha_t + n_{a_d x_{di} t}) - 1} \frac{\beta_{w_{di}} + m_{z_{di} w_{di}} - 1}{\sum_{v=1}^{V}(\beta_v + m_{z_{di} v}) - 1}
\end{aligned}$$

If one wants, further manipulation can turn the above formula into separated update equations for the topic and recipient of each token, suitable for random or systematic scan updates:

$$\begin{aligned}
P(x_{di}|\mathbf{x}_{-di}, \mathbf{z}, \mathbf{w}, \alpha, \beta, \mathbf{a}, \mathbf{r}) &\propto \frac{\alpha_{z_{di}} + n_{a_d x_{di} z_{di}} - 1}{\sum_{t=1}^{T}(\alpha_t + n_{a_d x_{di} t}) - 1} \\
P(z_{di}|\mathbf{x}, \mathbf{z}_{-di}, \mathbf{w}, \alpha, \beta, \mathbf{a}, \mathbf{r}) &\propto \frac{\alpha_{z_{di}} + n_{a_d x_{di} z_{di}} - 1}{\sum_{t=1}^{T}(\alpha_t + n_{a_d x_{di} t}) - 1} \frac{\beta_{w_{di}} + m_{z_{di} w_{di}} - 1}{\sum_{v=1}^{V}(\beta_v + m_{z_{di} v}) - 1}
\end{aligned}$$

References

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

Airoldi, E., Blei, D., Fienberg, S., & Xing, E. (2006). Stochastic blockmodels of mixed-membership: General formulation and nested variational inference. In *ICML Workshop on Statistical Network Analysis*.

Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, *74*(1), 47–97.

Andrieu, C., de Freitas, N., Doucet, A., & Jordan, M. (2003). An introduction to MCMC for machine learning. *Machine Learning*, *50*, 5–43.

Blei, D., & Lafferty, J. (2006a). Correlated topic models. In *Advances in Neural Information Processing Systems 18*.

Blei, D. M., & Lafferty, J. D. (2006b). Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*.

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

Erosheva, E., Fienberg, S., & Lafferty, J. (2004). Mixed membership models of scientific publications. *Proceedings of the National Academy of Sciences*, *101(Suppl. 1)*.

Griffiths, T., Steyvers, M., Blei, D., & Tenenbaum, J. (2004). Integrating topics and syntax. In *Advances in Neural Information Processing Systems (NIPS) 17*.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101 (suppl. 1)*, 5228–5235.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, *42*(1), 177–196.

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

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., & Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In *Proceedings of the 21st National Conference on Artificial Intelligence*.

Kemp, C., Griffiths, T. L., & Tenenbaum, J. (2004). Discovering latent classes in relational data. Tech. rep., MIT AI Memo 2004-019.

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

Kubica, J., Moore, A., Schneider, J., & Yang, Y. (2002). Stochastic link and group detection. In *Proceedings of the 18th National Conference on Artificial Intelligence*, pp. 798–804.

Kurihara, K., Kameya, Y., & Sato, T. (2006). A frequency-based stochastic blockmodel. In *Workshop on Information Based Induction Sciences*.

Li, W., & McCallum, A. (2006). Pachinko allocation: DAG-structured mixture models of topic correlations. In *Proceedings of the 23rd International Conference on Machine Learning*.

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. In *the 16th National Conference on Artificial Intelligence Workshop on Text Learning*.

Mimno, D., & McCallum, A. (2007). Expertise modeling for matching papers with reviewers. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 500–509.

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

Nowicki, K., & Snijders, T. A. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, *96*(455), 1077–1087.

Rosen-Zvi, M., Griffiths, T., Smyth, P., & Steyvers, M. (2005). Learning author-topic models from text corpora. *Submitted to Journal of Machine Learning Research.*

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

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

Steyvers, M., Smyth, P., Rosen-Zvi, M., & Griffiths, T. (2004). Probabilistic author-topic models for information discovery. In *Proceedings of the 10th 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. Tech. rep., UC Berkeley Statistics.

Wang, X., & McCallum, A. (2006). Topics over time: A non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Know ledge Discovery and Data Mining*, pp. 424–433.

Wang, X., McCallum, A., & Wei, X. (2007). Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Proceedings of the 7th IEEE International Conference on Data Mining.*

Wang, X., Mohanty, N., & McCallum, A. (2006). Group and topic discovery from relations and their attributes. In *Advances in Neural Information Processing Systems 18*, pp. 1449–1456.

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.* W. W. Norton & Company.

Watts, D. J., Dodds, P. S., & Newman, M. E. J. (2002). Identify and search in social networks. *Science*, *296*(5571), 1302–1305.

Wolfe, A. P., & Jensen, D. (2004). Playing multiple roles: Discovering overlapping roles in social networks. In *the 21st International Conference on Machine Learning Workshop on Statistical Relational Learning and its Connections to Other Fields.*

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