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

Andrew McCallum[1], Xuerui Wang[1], Andrés Corrrada-Emmanuel[2]

[1]Department of Computer Science , [2]Department of Physics
University of Massachusetts Amherst
JAIR Oct-2007
Discussion lead by: Miao Liu

March 3, 2011

Outline

- Availability of large human interactions dataset, popularity of online service (MySpace.com, LinkedIn.com etc), salient connection of the 9/11 hijackers arouse growing interest in SNA.
- Social network analysis (SNA):
  - models directed interactions among people, organizations and groups

- Availability of large human interactions dataset, popularity of online service (MySpace.com, LinkedIn.com etc), salient connection of the 9/11 hijackers arouse growing interest in SNA.
- Social network analysis (SNA):
  - models directed interactions among people, organizations and groups
  - emphasizes the existence of links among individuals (binary interaction), with directed and/or weighted edges

- Availability of large human interactions dataset, popularity of online service (MySpace.com, LinkedIn.com etc), salient connection of the 9/11 hijackers arouse growing interest in SNA.
- Social network analysis (SNA):
  - models directed interactions among people, organizations and groups
  - emphasizes the existence of links among individuals (binary interaction), with directed and/or weighted edges
  - limits to the use of network topology to roles discovery without sufficiently exploiting the language content of interactions (messages).
- Outside SNA literature, models of words clustering for topic discovery are developed, including
  - Probabilistic Latent Semantic Indexing (Hofmann, 2001)

- Availability of large human interactions dataset, popularity of online service (MySpace.com, LinkedIn.com etc), salient connection of the 9/11 hijackers arouse growing interest in SNA.
- Social network analysis (SNA):
  - models directed interactions among people, organizations and groups
  - emphasizes the existence of links among individuals (binary interaction), with directed and/or weighted edges
  - limits to the use of network topology to roles discovery without sufficiently exploiting the language content of interactions (messages).
- Outside SNA literature, models of words clustering for topic discovery are developed, including
  - Probabilistic Latent Semantic Indexing (Hofmann, 2001)
  - Latent Dirichlet Allocation (LDA) (Blei, et al, 2003)

- Availability of large human interactions dataset, popularity of online service (MySpace.com, LinkedIn.com etc), salient connection of the 9/11 hijackers arouse growing interest in SNA.
- Social network analysis (SNA):
  - models directed interactions among people, organizations and groups
  - emphasizes the existence of links among individuals (binary interaction), with directed and/or weighted edges
  - limits to the use of network topology to roles discovery without sufficiently exploiting the language content of interactions (messages).
- Outside SNA literature, models of words clustering for topic discovery are developed, including
  - Probabilistic Latent Semantic Indexing (Hofmann, 2001)
  - Latent Dirichlet Allocation (LDA) (Blei, et al, 2003)
  - Hierarchical Dirichlet Processes (Teh, et al, 2004)

- Availability of large human interactions dataset, popularity of online service (MySpace.com, LinkedIn.com etc), salient connection of the 9/11 hijackers arouse growing interest in SNA.
- Social network analysis (SNA):
  - models directed interactions among people, organizations and groups
  - emphasizes the existence of links among individuals (binary interaction), with directed and/or weighted edges
  - limits to the use of network topology to roles discovery without sufficiently exploiting the language content of interactions (messages).
- Outside SNA literature, models of words clustering for topic discovery are developed, including
  - Probabilistic Latent Semantic Indexing (Hofmann, 2001)
  - Latent Dirichlet Allocation (LDA) (Blei, et al, 2003)
  - Hierarchical Dirichlet Processes (Teh, et al, 2004)
  - Aurthor-Topic Model (Steyvers, et al, 2004)
- The above language models are inappropriate for SNA.

- Author-Recipient-Topic Models (ART)
  - models the language associated with social network interactions
  - conditions topics distribution jointly on author and recipient pairs
  - discover people's roles by clustering using author-recipient conditioned topic distribution.
- Role-Author-Recipient-Topic Models (RART)
  - extends ART and explicit captures roles of people, by generating role associations for the author and recipients, and conditioning the topic distribution on the role assignment.
  - represents that one person can have multiple roles
  - has three variants.

ART model

- d: document(message)
- a_d: an author
- $\mathbf{r}_d$: the set of recipients for each message
- x: a recipient
- z: a topic
- $\theta_{a_d x}$: the topic distribution specific to the author-recipient pair (a_d, x)
- ϕ_z: word distribution specific to the topic z

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

ART model

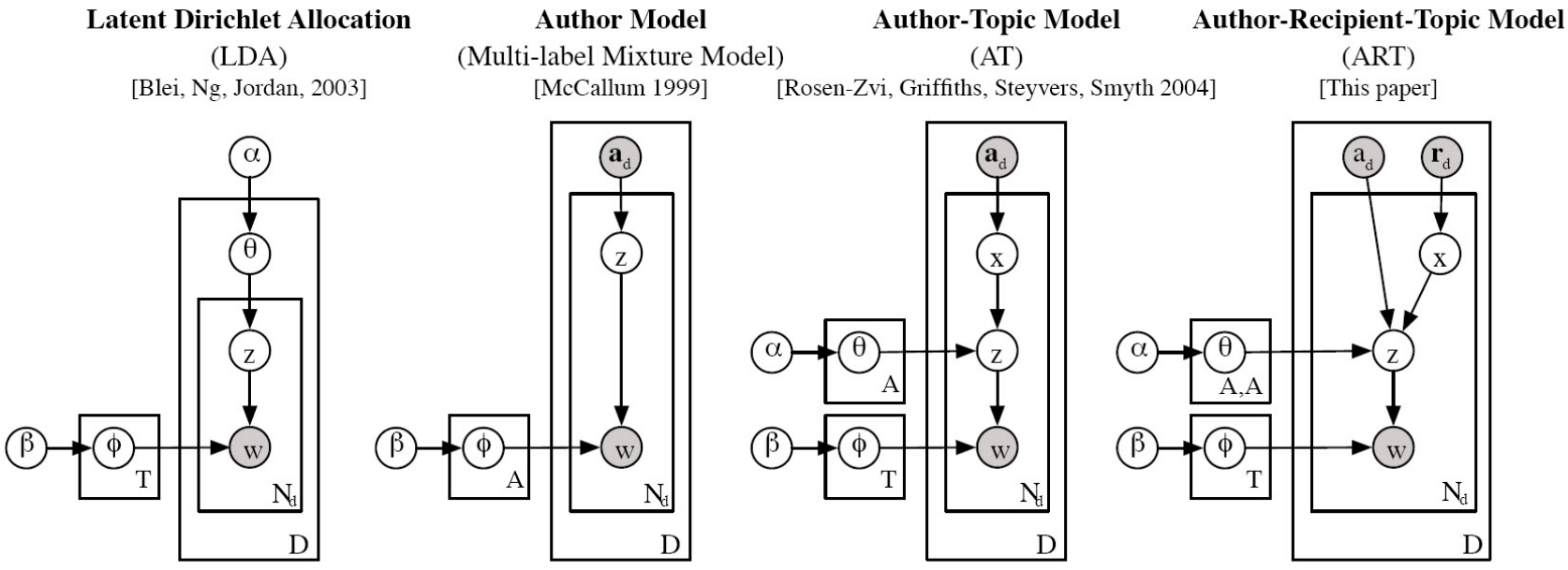


Figure 1: Three related models, and the ART model.

Inference

- The joint distribution of 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 z and a set of words w in the corpus is given by

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

- Marginal distribution of a corpus:

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

- Inference by Gibbs Sampling

Inference

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

- the conditional distribution of a topic and recipient for the word w_{di} given all other words's topic and recipient assignments, $\mathbf{x}_{di}$ and $\mathbf{z}_{-di}$

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

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

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

RATR models

- To better explore the roles of authors, an additional level of latent variables can be introduced to model roles.
- Of particular interest, a person can have multiple roles simultaneously. e.g. professor and mountain climber.
- Each role is associated with a set of topics, and these topics may overlap.
  - professors' topics: meeting times, grant proposals, and friendly relations
  - climbers' topics: mountains, climbing equipment, and also meeting times and friendly relations
- In RART, authors, roles and message-contents are modeled simultaneously.

Graphical Models for RART Model

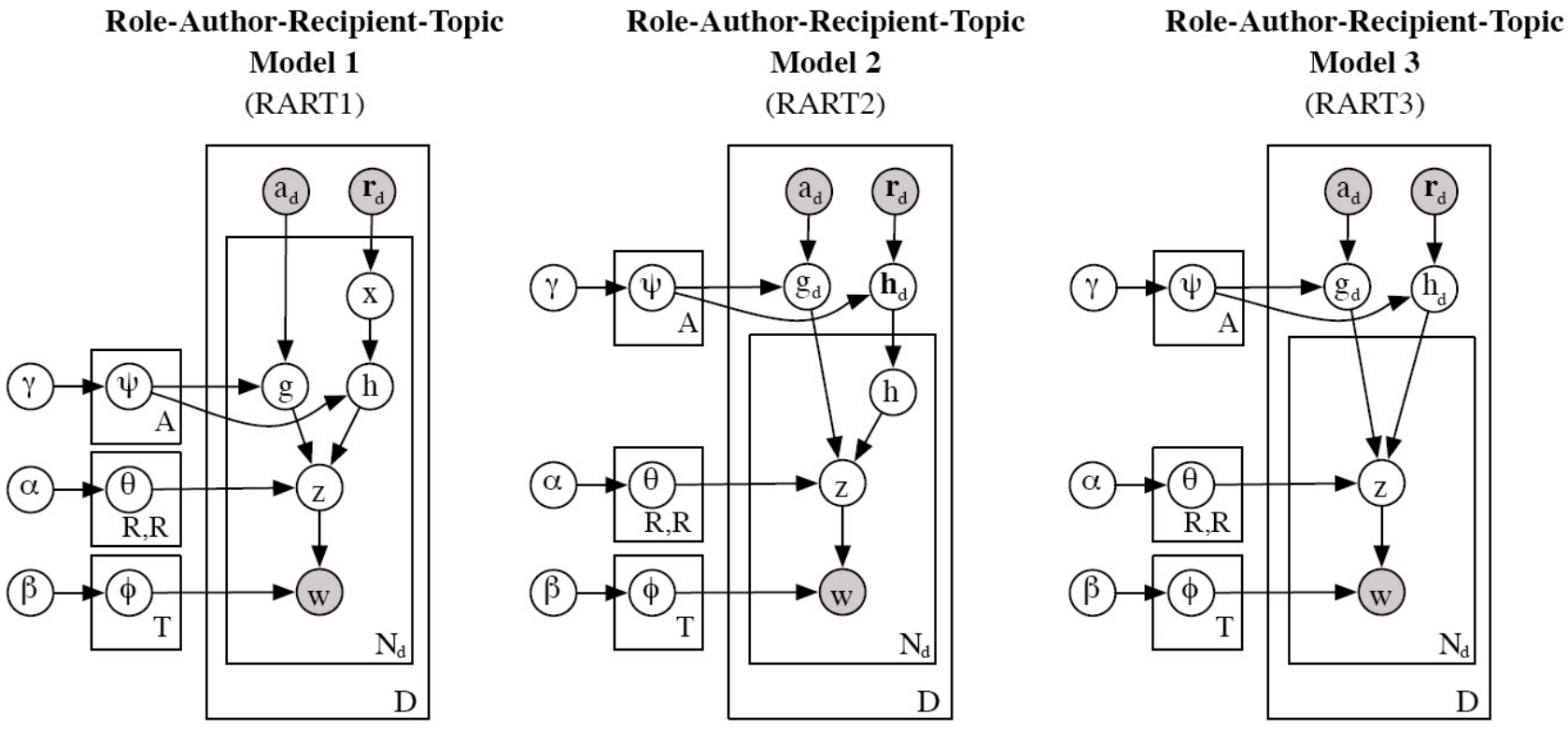


Figure 2: Three possible variants for the Role-Author-Reipient-Topic (RART) model.

Inference for RART1

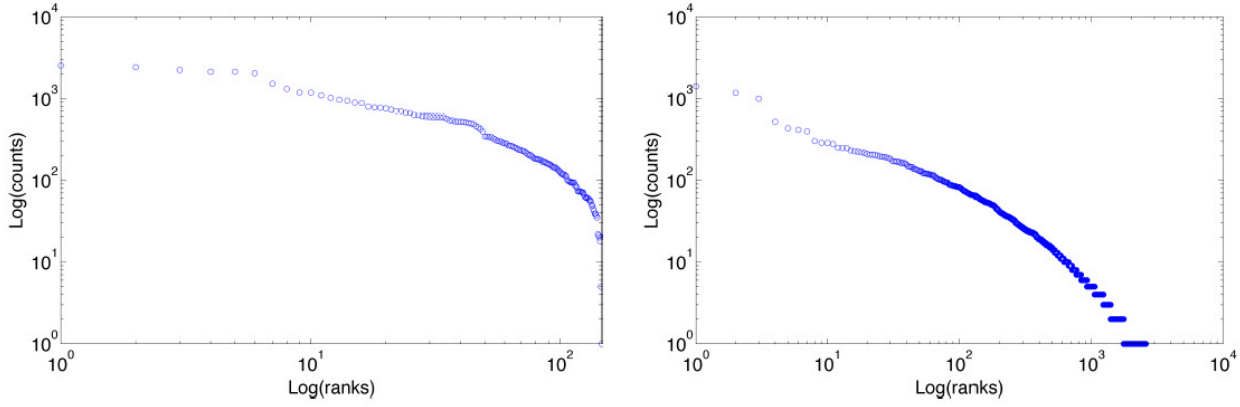
- For the RART1 model, the joint distribution of a topic mixture θ_{ij} for each author-role recipient-role pair (i, j), the role mixture ψ_k for each author k, the word mixture ϕ_t for each topic t, a set of sender roles $\mathbf{g}$, a set of recipients roles $\mathbf{h}$, a set of topics $\mathbf{z}$ and a set of words $\mathbf{w}$ is given by

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

- The marginal distribution of a corpus can be obtained by Integrating over Ψ,Θ and Φ, and summing over $\mathbf{x}, \mathbf{g}, \mathbf{h}, \mathbf{z}$.
- Gibbs sampling formulae can be derived in a similar ways as ART model.

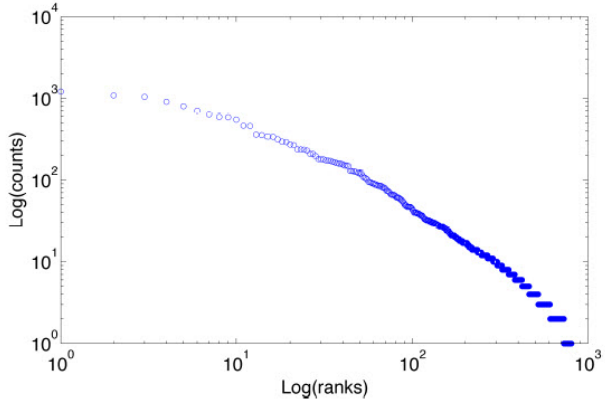
Data sets

1. Enron email corpus
   - is subpoenaed as part of the investigation by the Federal Energy Regulatory Commission (FERC), and then placed in the public record;
   - after text normalization filter (lowercasing, removal of quoted email text, etc), contains 147 users, 23,488 email messages, 22,901 unique words.
2. Personal email sent and received by McCallum between Jan. and Sept. 2004.
   - 13,633 unique messages written by 825 authors;
   - after applying the same normalization procedure, obtain a text corpus containing 457,057 word tokens, and a vocabulary of 22,901 unique words.

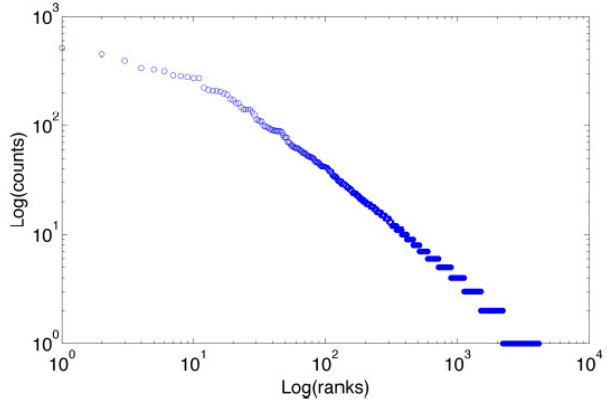


(a) Enron authors

(b) Enron author-recipient pairs



(c) McCallum authors



(d) McCallum author-recipient pairs

Figure 3: 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.

Topics and Prominent Relations from ART

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 J.Steffes	0.0549	G.Nemec B.Tycholiz	0.0737	J.Dasovich R.Shapiro	0.0340	E.Bass M.Lenhart	0.3050
J.Dasovich R.Shapiro	0.0377	G.Nemec M.Whitt	0.0551	J.Dasovich J.Steffes	0.0289	E.Bass P.Love	0.0780
D.Hyvl K.Ward	0.0362	B.Tycholiz G.Nemec	0.0325	C.Clair M.Taylor	0.0175	M.Motley M.Grigsby	0.0522

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 L.Kitchen	0.2158	J.Dasovich J.Steffes	0.1231	J.Dasovich R.Shapiro	0.3338	R.Haylett T.Geaccone	0.1432
S.Beck J.Lavorato	0.0826	J.Dasovich R.Shapiro	0.1133	J.Dasovich J.Steffes	0.2440	T.Geaccone R.Haylett	0.0737
S.Beck S.White	0.0530	M.Taylor E.Sager	0.0218	J.Dasovich R.Sanders	0.1394	R.Haylett D.Fossum	0.0420

Figure 4: An illustration of several topics from a 50-topic run for the Enron email data set.

Topics and Prominent Relations from ART

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	0.1290	ronb	0.0339	casutton	0.0498	mccallum	0.0558
mccallum		mccallum		mccallum		culotta	
mccallum	0.0746	wellner	0.0314	icml04-webadmin	0.0366	mccallum	0.0530
stowell		mccallum		icml04-chairs		casutton	
mccallum	0.0739	casutton	0.0217	mccallum	0.0343	mccallum	0.0274
lafferty		mccallum		casutton		ronb	
mccallum	0.0532	mccallum	0.0200	nips04workflow	0.0322	mccallum	0.0255
smyth		casutton		mccallum		saunders	
pereira	0.0339	mccallum	0.0200	weinman	0.0250	mccallum	0.0181
lafferty		wellner		mccallum		pereira	

Figure 5: The four topics most prominent in McCallums email exchange with Padhraic Smyth, from a 50-topic run of ART on 9 months of McCallums email.

Stochastic Blockstructures and Roles (a small subset of Enron Data)

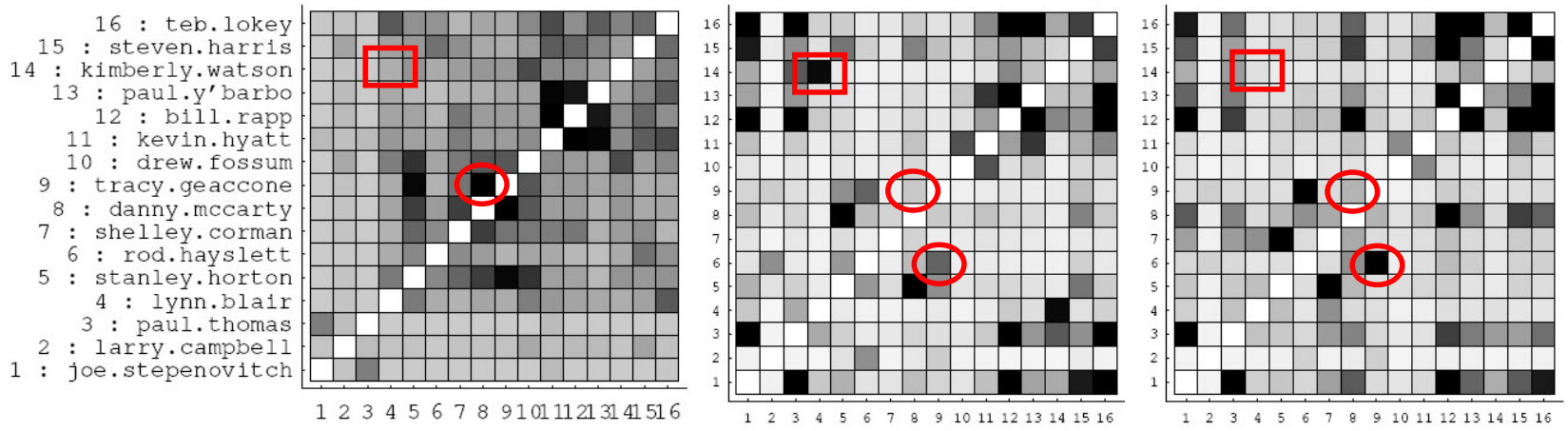
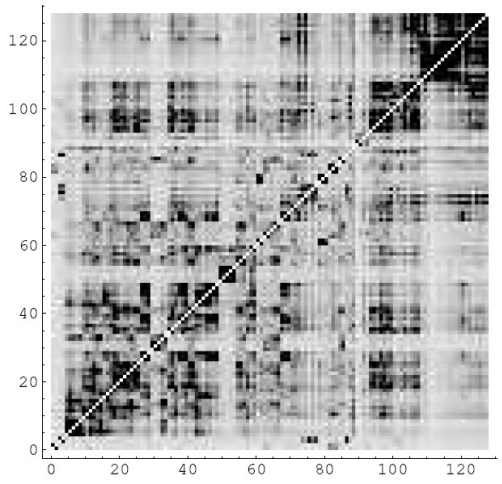


Figure 3: **Left:** SNA Inverse JS Network. **Middle:** ART Inverse JS Network. **Right:** AT Inverse JS Network. Darker shades indicate higher similarity.

- Compare Geaccone (User 9, Executive assistant), McCarty (CFO & Vice President,User 8), Hayslett (User 6)
- Compare Blair (user 4, gas pipeline logistics) and Watson (user 14 pipeline facility planning)
- Compare McCarty (user 8), Horton (user 5, President), Rapp (user 12 lawyer), Harris (user 15, mid-level manager)
- ART emphasizes role similarity, but not group membership: Considering Thomas(user 3, an energy future trader) to both Rapp (user 12), and Lokey (user 16, a regulatory affairs manager)
- The results of Traditional SNA is opposite to ART: User 1-3 (the only people in this matrix who were not member of Enron Transwestern Division) are put in different blocks.

Stochastic Blockstructures and Roles (McCallum's email)



(a) SNA Inverse JS Network for a 10 topic run on McCallum Email Data.

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

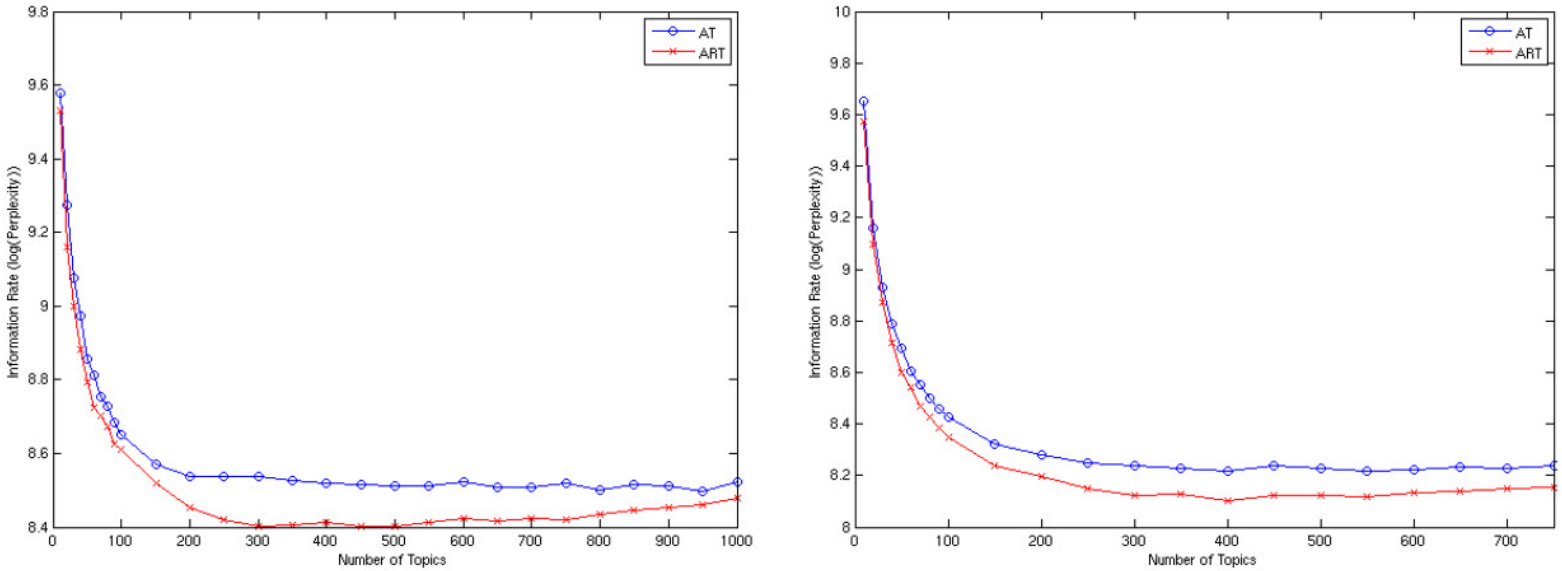
User Pair	*Description*
aepshtey rasmith	Both students in McCallum's class
donna editor	Spouse is unrelated to journal editor
donna krishna	Spouse is unrelated to conference organizer
donna ramshaw	Spouse is unrelated to researcher at BBN
donna reviews	Spouse is unrelated to journal editor
donna stromsten	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

(b) Pairs considered most alike by ART and SNA on McCallum email.

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

(c) Pairs with the highest rank difference between ART and SNA on McCallum email.

Perplexity Comparison between AT and ART



(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 $< 1e-5$).

Experimental Results with RART

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

(d) table 6 Each role is shown with the most prominent users and the the corresponding conditional probabilities.

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

(e) table 7 Each user is shown with his most prominent roles and the corresponding conditional probabilities.

Illustrations of two roles from a 50-topic, 15-group run for the McCallum email data set.