Modeling and Predicting Personal Information Dissemination Behavior

Xiaodan Song, Ching-Yung Lin
Department of Electrical Engineering, University of Washington, Box 352500, Seattle, WA 98195, USA
{song, cylin}@ee.washington.edu

Belle L. Tseng
NEC Labs America, 10080 N. Wolfe Road, SW3-350, Cupertino, CA 95014, USA
belle@sv.nec-labs.com

Ming-Ting Sun
Department of Electrical Engineering, University of Washington, Box 352500, Seattle, WA 98195, USA
sun@ee.washington.edu

ABSTRACT

In this paper, we propose a new way to automatically model and predict human behavior of receiving and disseminating information by analyzing the contact and content of personal communications. A personal profile, called CommunityNet, is established for each individual based on a novel algorithm incorporating contact, content, and time information simultaneously. It can be used for personal social capital management. Clusters of CommunityNets provide a view of informal networks for organization management. Our new algorithm is developed based on the combination of dynamic algorithms in the social network field and the semantic content classification methods in the natural language processing and machine learning literatures. We tested CommunityNets on the Enron Email corpus and report experimental results including filtering, prediction, and recommendation capabilities. We show that the personal behavior and intention are somewhat predictable based on these models. For instance, "to whom a person is going to send a specific email" can be predicted by one's personal social network and content analysis. Experimental results show the prediction accuracy of the proposed adaptive algorithm is 58% better than the social network-based predictions, and is 75% better than an aggregated model based on Latent Dirichlet Allocation with social network enhancement. Two online demo systems we developed that allow interactive exploration of CommunityNet are also discussed.

Categories and Subject Descriptors
I.2.6 [Artificial Intelligence]: Learning

General Terms: algorithms, experimentation

Keywords: user behavior modeling, personal information management, information dissemination

1. INTRODUCTION

Working in the information age, the most important is not what you know, but who you know [1]. A social network, the graph of relationships and interactions within a group of individuals, plays a fundamental role as a medium for the spread of information, ideas, and influence. At the organizational level, personal social networks are activated for recruitment, partnering, and information access. At the individual level, people exploit their networks to advance careers and gather information.

Informal network within formal organizations is a major, but hard to acquire, factor affecting companies' performance. Krackhardt [2] showed that companies with strong informal networks perform five or six times better than those with weak networks, especially on the long-term performance. Friend and advice networks drive enterprise operations in a way that, if the real organization structure does not match the informal networks, then a company tends to fail [3]. Since Max Weber first studied modern bureaucracy structures in the 1920s, decades of related social scientific researches have been mainly relying on questionnaires and interviews to understand individuals' thoughts and behaviors for sensing informal networks. However, data collection is time consuming and seldom provides timely, continuous, and dynamic information. This is usually the biggest hurdle in social studies.

Personal Social Network (PSN) could provide an organizing principle for advanced user interfaces that offer information management and communication services in a single integrated system. One of the most pronounced examples is the networking study by Nardi *et al.* [4], who coined the term *intensional networks* to describe personal social networks. They presented a visual model of user's PSN to organize personal communications in terms of a social network of contacts. From this perspective, many tools were built such as LinkedIn [5], Orkut [6], and Friendster [7]. However, all of them only provide tools for visually managing personal social networks. Users need to manually input, update, and manage these networks. This results in serious drawbacks. For instance, people may not be able to invest necessary efforts in creating rich information, or they may not keep the information up-to-date as their interests, responsibilities, and network change. They need a way to organize the relationship and remember who have the resources to help them. We coin the terminology of managing these goals as *personal social capital management*[1].

In this paper, we develop a *user-centric* modeling technology, which can dynamically describe and update a

KDD'05, August 21–24, 2005, Chicago, Illinois, USA.

[1] Social capital infers to the accumulated contacts' human capital (asset, power, resources) that a person can explore through social network [8].

person's personal social network with context-dependent and temporal evolution information from personal communications. We refer to the model as a *CommunityNet*. Senders and receivers, time stamps, subject and content of emails contribute three key components – ***content semantics***, ***temporal information,*** and ***social relationship***. We propose a novel Content-Time-Relation (CTR) algorithm to capture dynamic and context-dependent information in an unsupervised way. Based on the *CommunityNet* models, many questions can be addressed by inference, prediction and filtering. For instance, 1) Who are semantically related to each other? 2) Who will be involved in a special topic? Who are the important (central) people in this topic? 3) How does the information flow? and 4) If we want to publicize a message, whom should we inform?

Figure 1 shows the procedure of our proposed scheme. First, topic detection and clustering is conducted on training emails in order to define topic-communities. Then, for each individual, *CommunityNet* is built based on the detected topics, the sender and receiver information, and the time stamps. Afterwards, these personal *CommunityNets* can be applied for inferring organizational informal networks and predicting personal behaviors to help users manage their social capitals. We incorporate the following innovative steps:

1) *Incorporate content analysis into social network in an unsupervised way*
2) *Build a CommunityNet for each user to capture the context-dependent, temporal evolutionary personal social network based on email communication records*
3) *Analyze people's behaviors based on CommunityNet, including predicting people's information sending and receiving behaviors*
4) *Show the potential of using automatically acquired personal social network for organization and personal social capital management*

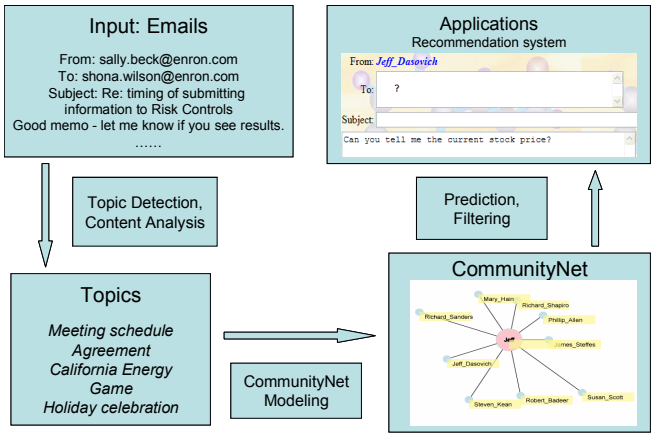


Figure 1. An Overview of CommunityNet

We tested the *CommunityNet* model on the Enron email corpus comprising the communication records of 154 Enron employees dating from Jan. 1999 to Aug. 2002. The Enron email dataset was originally made available to public by the Federal Energy Regulatory Commission during the investigation [9]. It was later collected and prepared by Melinda Gervasio at SRI for the CALO (A Cognitive Assistant that Learns and Organizes) project. William Cohen from CMU has put up the dataset on the web for research purpose [9]. This version of the dataset contains around 517,432 emails within 150 folders. We clean the data and extract 154 users from those 150 folders with 166,653 unique messages from 1999 to 2002. In the experiments, we use 16,873 intra-organizational emails which connect these 154 people.

The primary contributions of this paper are three-fold. First we develop an algorithm incorporating content-time-relation detection. Second, we generate an application model which describes personal dynamic community network. Third, we show how this model can be applied to organization and social capital management. To the best of our knowledge, this is among the first reported technologies on fusing research in the social network analysis field and the content analysis field for information management. We propose the CTR algorithm and the *CommunityNet* based on the Latent Dirichlet Allocation algorithm. In our experiments, we observed clear benefit of discovering knowledge based on multi-modality information rather than using only single type of data.

The rest of the paper is organized as follows. In Section 2, we present an overview of related work. In Section 3, we present our model. We discuss how to use CommunityNet to analyze communities and individuals in section 4 and 5, respectively. In Section 6, we show two demo systems for query, visualization and contact recommendation. Finally, conclusions and future work are addressed in Section 7.

2. RELATED WORK

2.1 Social Network Analysis

To capture relationships between entities, social network has been a subject of study for more than 50 years. An early sign of the potential of social network was perhaps the classic paper by Milgram [10] estimating that on average, every person in the world is only six edges away from each other, if an edge between *i* and *j* means "*i* knows *j*". Lately, introducing social network analysis into information mining is becoming an important research area. Schwartz and Wood [11] mined social relationships from email logs by using a set of heuristic graph algorithms. The *Referral Web* project [12] mined a social network from a wide variety of publicly-available online information, and used it to help individuals find experts who could answer their questions based on geographical proximity. Flake *et al.* [13] used graph algorithms to mine communities from the Web (defined as sets of sites that have more links to each other than to non-members). Tyler *et al.* [14] use a betweenness centrality algorithm for the automatic identification of communities of practice from email logs within an organization. The Google search engine [15] and Kleinberg's HITS algorithm of finding hubs and authorities on the Web [16] are also based on social network concepts. The success of these approaches, and the discovery of widespread network topologies with nontrivial properties, have led to a recent flurry of research on applying link analysis for information mining.

A promising class of statistical models for expressing structural properties of social networks is the class of Exponential Random Graph Models (ERGMs) (or p* model) [17]. This statistical model can represent structural properties that define complicated dependence patterns that cannot be easily modeled by deterministic models. Let *Y* denote a random graph on a set of *n* nodes and let *y* denote a particular graph on those nodes. Then, the probability of *Y* equals to *y* is

$$P_\theta(Y=y)=\frac{\exp\left(\theta^T s(y)\right)}{c(\theta)} \quad (1)$$

where $s(y)$ is a known vector of graph statistics (Density, Reciprocity, Transitivity, etc) on y, θ is a vector of coefficients to model the influence of each statistics for the whole graph, T means "transpose", $c(\theta)$ is a normalization term to satisfy $\sum_y P_\theta(Y=y)=1$. The parameters θ are estimated based on the observed graph y^{obs} by maximum likelihood estimation.

All the research discussed above has focused on using static properties in a network to represent the complex structure. However, social networks evolve over time. Evolution property has a great deal of influence; *e.g.*, it affects the rate of information diffusion, the ability to acquire and use information, and the quality and accuracy of organizational decisions.

Dynamics of social networks have attracted many researchers' attentions recently. Given a snapshot of a social network, [19] tries to infer which new interactions among its members are likely to occur in the near future. In [20], Kubica *et al.* are interested in tracking changes in large-scale data by periodically creating an agglomerative clustering and examining the evolution of clusters over time. Among the known dynamical social networks in literature, Snijder's dynamic actor-oriented social network [18] is one of the most successful algorithms. Changes in the network are modeled as the stochastic result of network effects (density, reciprocity, *etc.*). Evolution is modeled by continuous-time Markov chains, whose parameters are estimated by the Markov chain Monte Carlo procedures. In [21], Handcock *et al.* proposed a curved ERGM model and applied it to the new specifications of ERGMs This latest model uses nonlinear parameters to represent structural properties of networks.

The above mentioned dynamic analyses show some success in analyzing longitudinal stream data. However, most of them are only based on pure network properties, without knowing what people are talking about and why they have close relationships.

2.2 Content Analysis

In statistical Natural Language processing, one common way of modeling the contributions of different topics to a document is to treat each topic as a probability distribution over words, viewing a document as a probability distribution over words, and thus viewing a document as a probabilistic mixture over these topics. Given T topics, the probability of the ith word in a given document is formalized as:

$$P(w_i)=\sum_{j=1}^{T} P(w_i \mid z_i=j)P(z_i=j) \tag{2}$$

where z_i is a latent variable indicating the topic from which the ith word was drawn and $P(w_i \mid z_i=j)$ is the probability of the word w_i under the jth topic. $P(z_i=j)$ gives the probability of choosing a word from topics j in the current document, which varies across different documents.

Hofmann [22] introduced the aspect model Probabilistic Latent Semantic Analysis (PLSA), in which, topics are modeled as multinomial distributions over words, and documents are assumed to be generated by the activation of multiple topics. Blei *et al.* [23] proposed Latent Dirichlet Allocation (LDA) to address the problems of PLSA that parameterization was susceptible to overfitting and did not provide a straightforward way to infer testing documents. A distribution over topics is sampled from a Dirichlet distribution for each document. Each word is sampled from a multinomial distribution over words specific to the sampled topic. Following the notations in [24], in LDA, D documents containing T topics expressed over W unique words, we can represent $P(w \mid z)$ with a set of T multinomial distributions ϕ over the W words, such that $P(w \mid z=j)=\phi_j^{(w)}$, and *P(z)* with a set of D multinomial distribution θ over the T topics, such that for a word in document d, $P(z=j)=\theta_j^{(d)}$. Recently, the Author-Topic (AT) model [25] extends LDA to include authorship information, trying to recognize which part of the document is contributed by which co-author. In a recent unpublished work, McCallum *et al.* [26] further extend the AT model to the Author-Recipient-Topic model by regarding the sender-receiver pair as an additional author variable for topic classification. Their goal is role discovery, which is similar to one of our goals as discussed in Sec. 4.1.2 without taking the temporal nature of emails into consideration.

Using LDA, ϕ and θ are parameters that need to be estimated by using sophisticated approximation either with variational Bayes or expectation propagation. To solve this problem, Griffiths and Steyvers [24] extended LDA by considering the posterior distribution over the assignments of words to topics and showed how Gibbs sampling could be applied to build models. Specifically,

$$P(z_i=j \mid z_{-i},w) \propto \frac{n_{-i,j}^{(w_i)}+\beta}{n_{-i,j}^{(\cdot)}+W\beta}\frac{n_{-i,j}^{(d_i)}+\alpha}{n_{-i,\cdot}^{(d_i)}+T\alpha} \tag{3}$$

where $n_{-i}^{(\cdot)}$ is a count that does not include the current assignment, $n_j^{(w)}$ is the number of times word w has been assigned to topic j in the vector of assignments z, $n_j^{(d)}$ is the number of times a word from document d has been assigned to topic j, $n_j^{(\cdot)}$ is a sum of $n_j^{(w)}$, $n_{\cdot}^{(d)}$ is a sum of $n_j^{(d)}$. Further, one can estimate $\phi_j^{(w)}$, the probability of using word w in topic j, and $\theta_j^{(d)}$, the probability of topic j in document d as follows:

$$\hat{\phi}_j^{(w)}=\frac{n_j^{(w)}+\beta}{n_j^{(\cdot)}+W\beta} \tag{4}$$

$$\hat{\theta}_j^{(d)}=\frac{n_j^{(d)}+\alpha}{n_{\cdot}^{(d)}+T\alpha} \tag{5}$$

In [24], experiments show that topics can be recovered by their algorithm and show meaningful aspects of the structure and relationships between scientific papers.

Contextual, relational, and temporal information are three key factors for current data mining and knowledge management models. However, there are few papers addressing these three components simultaneously. .In our recent paper, we built user models to explicitly describe a person's expertise by a relational and evolutionary graph representation called *ExpertisetNet* [27]. In this paper, we continue exploring this thread, and build a *CommunityNet* model which incorporates these three components together for data mining and knowledge management.

3. COMMUNITYNET

In this section, we first define terminologies. Then, we propose a Content-Time-Relation (CTR) algorithm to build the

personal *CommunityNet*. We also specifically address the prediction of the user's behaviors as a classification problem and solve it based on the *CommunityNet* models.

3.1 Terminology

Definition 1. Topic-Community: *A topic community is a group of people who participate in one specific topic.*

Definition 2: Personal Topic-Community Network (PTCN): *A personal topic-community network is a group of people directly connected to one person about a specific topic.*

Definition 3. Evolutionary Personal Social Network: *An evolutionary personal social network illustrates how a personal social network changes over time.*

Definition 4. Evolutionary Personal Topic-Community Network: *An evolutionary network illustrates how a person's personal topic-community network changes over time.*

Definition 5. Personal Social Network Information Flow: *A personal social network information flow illustrates how the information flows over a person's personal social network to other people's personal social networks*

Definition 6: Personal Topic-Community Information Flow: *A personal Topic-CommunityNet information flow illustrates how the information about one topic flows over a person's personal social network to other people's personal social networks.*

3.2 Personal Social Network

We build people's personal social networks by collecting their communication records. The nodes of a network represent whom this person contacts with. The weights of the links measure the probabilities of the emails he sends to the other people: A basic form of the probability that an user *u* sending email to a recipient r is:

$$P(r \mid u) = \frac{\text{number of times } u \text{ sends emails to } r}{\text{total number of emails sent out by } u} \tag{6}$$

We build evolutionary personal social networks to explore the dynamics and the evolution. The ERGM in Eq. (1) can be used to replace Eq. (6) for probabilistic graph modeling. A big challenge of automatically building evolutionary personal social network is the evolutionary segmentation, which is to detect changes between personal social network cohesive sections. Here we apply the same algorithm as we proposed in [27]. For each personal social network in one time period *t*, we use the exponential random graph model [17] to estimate an underlying distribution to describe the social network. An ERGM is estimated from the data in each temporal sliding window. With these operations, we obtain a series of parameters which indicates the graph configurations.

3.3 Content-Time-Relation Algorithm

We begin with email content, sender and receiver information, and time stamps, and use these sources of knowledge to create a joint probabilistic model. An observation is *(u, r, d, w, t)* corresponds to an event of a user *u* sending to receivers *r* an email *d* containing words *w* during a particular time period *t*. Conceptually, users choose latent topics *z*, which in turn generate receivers *r*, documents *d*, and their content words *w* during time period *t*.

$$P(\langle u,r\rangle \mid d,t) = \sum_{z} P(\langle u,r\rangle \mid z,t) P(z \mid d,t) \tag{7}$$

where $\langle u,r\rangle$ is a sender-receiver pair during time period *t*. $\langle u,r\rangle$ can be replaced by any variable to indicate the user's behavior, as long as it is also assumed to be dependent on latent topics of emails.

In order to model the PTCN, one challenge is how to detect latent topics dynamically and at the same time track the emails related to the old topics. This is a problem similar to topic detection tracking [28]. We propose an incremental LDA (ILDA) algorithm to solve it, in which the number of topics is dynamically updated based on the Bayesian model selection principle [24]. The procedures of the algorithm are illustrated as follows:

Incremental Latent Dirichlet Allocation (ILDA) algorithm:

Input: *Email streams with timestamp t*

Output: $\phi_{j,t}^{(w)}$, $\theta_{j,t}^{(d)}$ *for different time period t*

Steps:

1) *Apply LDA on a data set with currently observed emails in a time period t to generate latent topics* z_j *and estimate* $P(w \mid z_j, t_0) = \phi_{j,t_0}^{(w)}$ *and* $P(z_j \mid d, t_0) = \theta_{j,t_0}^{(d)}$ *by equation (4) and (5). The number of topics is determined by the Bayesian model selection principle.*
2) *When new emails arrive during time period k, use Bayesian model selection principle to determine the number of topics and apply* $P(z_i = j \mid z_{-i}, w, t_k) \propto P(w \mid z_i = j, t_{k-1}) \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\square}^{(d_i)} + T\alpha}$ *to estimate* $P(z \mid d, t_k)$, $P(w \mid z, t_k)$, *and* $P(z \mid w, t_k)$.
3) *Repeat step 2) until no data arrive.*

Based on this ILDA algorithm, we propose a Content-Time-Relation (CTR) algorithm. It consists of two phases, the training phase and the testing phase. In the training phase, emails as well as the senders, receivers and time stamps are available. $P(w \mid z, t_{old})$ and $P(\langle u,r\rangle \mid z, t_{old})$ are learnt from the observed data. In the testing phase, we apply ILDA to learn $P(z \mid d, t_{new})$. Based on $P(\langle u,r\rangle \mid z, t_{old})$, which is learnt from the training phase, $\langle u,r\rangle$ can be inferred. Again, $\langle u,r\rangle$ represents a sender-receiver pair or any variable to indicate the user's behavior, as long as it is dependent on the latent topics of emails.

Content-Time-Relation (CTR) algorithm:

1) Training phase

Input: Old emails with content, sender and receiver information, and time stamps t_{old}

Output: $P(w \mid z, t_{old})$, $P(z \mid d, t_{old})$, and $P(\langle u,r\rangle \mid z, t_{old})$

Steps:

a) Apply Gibbs Sampling on the data according to equation (3).

b) Estimate $P(w \mid z_j, t_{old}) = \phi_{j,t_{old}}^{(w)}$ *and* $P(z_j \mid d, t_{old}) = \theta_{j,t_{old}}^{(d)}$ *by equation (4), and (5).*

c) Estimate

$$\begin{aligned} P(\langle u,r\rangle \mid z, t_{old}) &= \sum_{d} P(\langle u,r\rangle \mid d, t_{old}) P(d \mid z, t_{old}) \\ &\propto \sum_{d} P(\langle u,r\rangle \mid d, t_{old}) P(z \mid d, t_{old}) \end{aligned} \tag{8}$$

2) Testing phase

Input: *New emails with content and time stamps* t_{new}

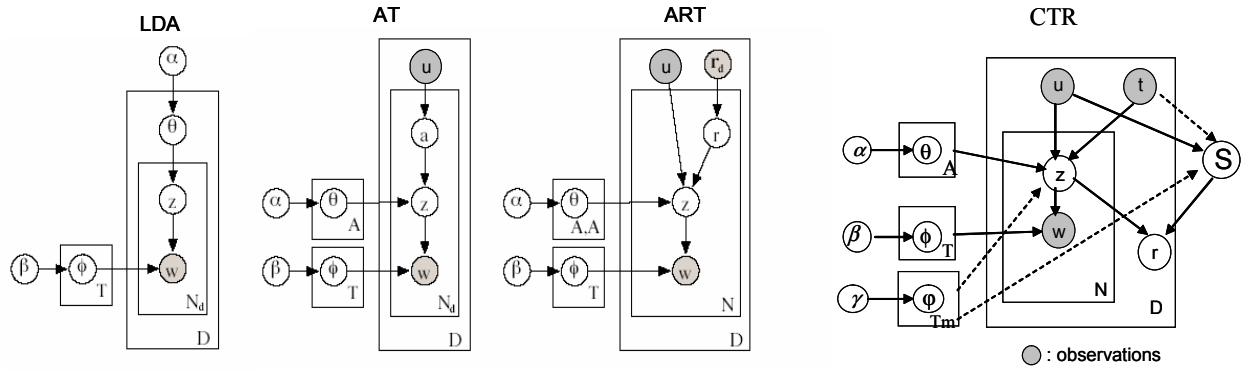


Figure 2. The graphical model for the CTR model comparing to LDA, AT and ART models, where *u*: sender, *t*: time, *r*: receivers, *w*: words, *z*: latent topics, *S*: social network, *D*: number of emails, *N*: number of words in one email, *T*: number of topics, *Tm*: size of the time sliding window, *A*: number of authors, θ, ϕ and φ are the parameters we want to estimate with the hyperparameters α, β, γ

Output: $P(\langle u,r\rangle \mid d, t_{new})$, $P(w \mid z, t_{new})$, and $P(z \mid d, t_{new})$

Steps:

a) Apply incremental LDA by Gibbs Sampling based on $P(z_i = j \mid z_{-i}, w, t_{new}) \propto P(w, z_i = j \mid t_{old}) \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,\square}^{(d_i)} + T\alpha}$ *to estimate* $P(w \mid z_j, t_{new})$, *and* $P(z \mid d, t_{new})$ *by equation (4) and (5).*

b) If the topics are within the training set, estimate $\hat{P}(\langle u,r\rangle \mid d, t_{new}) = \sum_z P(\langle u,r\rangle \mid z, t_{old}) P(z \mid d, t_{new})$, *else if the sender and receivers are within the training set, estimate* $\hat{P}(\langle u,r\rangle \mid d, t_{new})$ *by topic-independent social network* $P(\langle u,r\rangle \mid t_{old})$.

c) If there are new topics detected, update the model by incorporating the new topics.

Inference, filtering, and prediction can be conducted based on this model. For the CTR algorithm, sender variable *u* or receiver variable *r* is fixed. For instance, if we are interested in $P(r \mid u, d, t)$, which is to answer a question of whom we should send the message *d* to during the time period *t*. The answer will be

$$\arg\max_r \hat{P}(r \mid u, d, t_{new}) =$$

$$\arg\max_r \left(\sum_{z_{t_{old}}} P(r \mid u, z_{t_{old}}, t_{old}) P(z_{t_{old}} \mid u, d, t_{new}) + \sum_{z_{t_{new}} / z_{t_{old}}} P(r \mid u, t_{old}) P(z_{t_{new}} \mid u, d, t_{new}) \right) \quad (9)$$

where $z_{t_{new}} / z_{t_{old}}$ represents the new topics emerging during the time period *t*. Another question is if we receive an email, who will be possibly the sender?

$$\arg\max_u \hat{P}(u \mid r, d, t_{new}) =$$

$$\arg\max_u \left(\sum_{z_{t_{old}}} P(u \mid r, z_{t_{old}}, t_{old}) P(z_{t_{old}} \mid r, d, t_{new}) + \sum_{z_{t_{new}} / z_{t_{old}}} P(u \mid r, t_{old}) P(z_{t_{new}} \mid r, d, t_{new}) \right) \quad (10)$$

Eq. (9) and Eq. (10) integrate the PSN, content and temporal analysis. Social network models such as ERGM in Eq. (1) or the model in Sec. 3.2 can be applied to the $P(\langle u,r\rangle \mid d, t)$ terms.

Figure 2 illustrates the CTR model and compares to the LDA, AT and ART models. In CTR, the observed variables not only include the words *w* in an email but also the sender *u* and the timestamp on each email *d*.

3.4 Predictive Algorithms

For the sake of easier evaluation, we focus on prediction schemes in details. Specifically, we address the problem of predicting receivers and senders of emails as a classification problem, in which we train classifiers to predict the senders or receivers and other behavior patterns given the observed people's communication records. The trained classifier represents a function in the form of:

$$f : Comm(t-i, t) \rightarrow Y \quad (11)$$

where $Comm(t-i, t)$ is the observed communication record during the interval from time *t-i* to *t*, *Y* is a set of receivers or senders or other user behavior patterns to be discriminated, and the value of $f(Comm(t-i, t))$ is the classifier prediction regarding which user behavior patterns gave rise to the observed communication records. The classifier is trained by providing the history of the communication records with known user behaviors.

3.4.1 Using Personal Social Network Model

We aggregate all the communication records in the history of a given user, and build his/her personal social network. We choose those people with the highest communication frequency with this person as the prediction result.

3.4.2 Using LDA combined with PSN Model

We use the LDA model and combine it with PSN to do the prediction, which is referred as LDA-PSN in the paper. Latent topics are detected by applying original LDA on the training set and LDA is used for inference in testing data without incorporating new topics when time passes by. The possible senders and receivers when new emails arrive, $P(\langle u,r\rangle \mid d, t_{new})$ is estimated as $\hat{P}(\langle u,r\rangle \mid d, t_{new}) = \sum_z P(\langle u,r\rangle \mid z, t_{old}) P(z \mid d, t_{new})$. People are ranked by this probability as the prediction results.

3.4.3 Using CTR Model

People tend to send emails to different group of people under different topics during different time periods. This is the assumption we made for our predictive model based on CTR.

$P(\langle u,r\rangle \mid d,t_{new})$ is estimated by applying the CTR model discussed in section 3.3. The prediction results are people with highest scores calculated by equation (9) and (10).

3.4.4 Using an Adaptive CTR Model

Both the personal social network and the CTR model ignore a key piece of information from communication records -- the dynamical nature of emails. Both personal social network and Topic-Community dynamically change and evolve. Only based on the training data which are collected in history will not get the optimal performance for the prediction task. Adaptive prediction by updating the model with newest user behavior information is necessary. We apply several strategies for the adaptive prediction. The first strategy is aggregative updating the model by adding new user behavior information including the senders and receivers into the model. Then the model becomes:

$$\hat{P}(\langle u,r\rangle \mid d,t_i) = \sum_{k=1}^{K} P(\langle u,r\rangle \mid z_k,t_{old}) P(z_k \mid d,t_i) + \sum_{z_{t_i}/z_{old}} P(\langle u,r\rangle \mid t_{old}) P(z_{t_i} \mid d,t_i) \quad (12)$$

where K is the number of old topics. Here, we always use the data from t_{old} , including t_0 to t_{i-1} to predict the user behavior during t_i .

In the second strategy, we assume the correlation between current data and the previous data decays over time. The more recent data are more important. Thus, a sliding window of size *n* is used to choose the data for building the prediction model, in which the prediction is only dependent on the recent data, with the influence of old data ignored. Here in equation (12), t_{old} consists of t_{i-n} to t_{i-1}.

3.5 CommunityNet Model

We then build a *CommunityNet* model based on the CTR algorithm. The *CommunityNet* model, which refers to the personal Topic-Community Network, draws upon the strengths of the topic model and the social network as well as the dynamic model, using a topic-based representation to model the content of the document, the interests of the users, the correlation of the users and the receivers and all these relationship changing over time. For prediction, *CommunityNet* incorporates the adaptive CTR model as described in Section 3.4.4.

4. COMMUNITY ANALYSIS

The first part of our analysis focuses on identifying clusters of topics, and the senders and receivers who participated in those topics. First, we analyze the topics detected from the Enron Corpus. Then, we study the topic-community patterns.

4.1 Topic Analysis

In the experiment, we applied Bayesian model selection [24] to choose the number of topics. In the Enron intra-organization emails, there are 26,178 word-terms involved after we apply stop-words removal and stemming, We computed $P(w \mid T)$ for T values of 30, 50, 70, 100, 110, 150 topics and chose T = 100 with the maximum value of $\log(P(w \mid T))$ for the experiment.

4.1.1 Topic Distribution

After topic clustering based on words, for each document, we have *P(z|d)*, which indicates how likely each document belongs to each topic. By summing up this probability for all the documents, we get the topic distribution of how likely each topic occurs in this corpus. We define this summed likelihood as "Popularity" of the topic in the dataset. From this topic distribution, we can see that some topics are hot - people frequently communicate with each other about them, while some others are cold, with only few emails related to them. Table 1 illustrates the top 5 topics in Enron corpus. We can see that most of them are talking about regular issues in the company like meeting, deal, and document. Table 2 illustrates the bottom 5 topics in Enron corpus. Most of them are specific and sensitive topics, like "Stock" or "Market". People may feel less comfortable to talk about them broadly.

Table 1. Hot Topics

meeting	deal	Petroleum	Texas	document
meeting	deal	Petroleum	Houston	letter
plan	desk	research	Texas	draft
conference	book	dear	Enron	attach
balance	bill	photo	north	comment
presentation	group	Enron	America	review
discussion	explore	station	street	mark

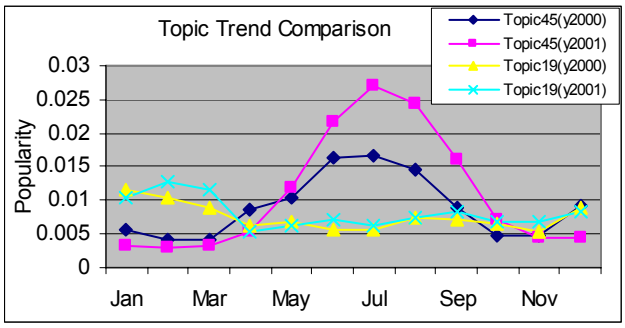
Table 2. Cold Topics

Trade	stock	network	Project	Market
trade	Stock	network	Court	call
London	earn	world	state	market
bank	company	user	India	week
name	share	save	server	trade
Mexico	price	secure	project	description
conserve	new	system	govern	respond

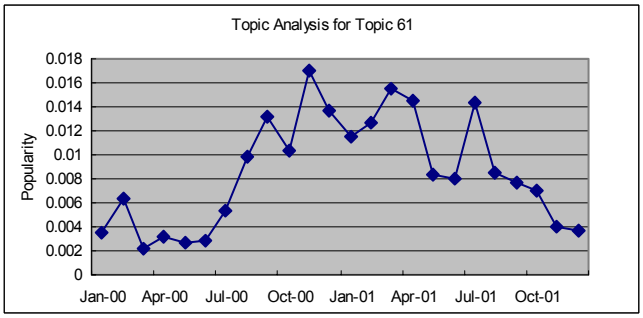
4.1.2 Topic Trend Analysis

To sense the trend of the topics over time, we calculate the topic popularity for year 2000 and 2001, and calculate the correlation coefficients of these two series. For some topics, the trends over years are similar. Figure 3(a) illustrates the trends for two topics which have largest correlation coefficients between two years. Topic 45, which is talking about a schedule issue, reaches a peak during June to September. For topic 19, it is talking about a meeting issue. The trend repeats year to year.

Figure 3(b) illustrates the trend of Topic "California Power" over 2000 to 2001. We can see that it reaches a peak from the end of year 2000 to the beginning of year 2001. From the timeline of Enron [29], we found that "California Energy Crisis" occurred at exactly this time period. Among the key people related to this topic, Jeff Dasovich was an *Enron government relations executive*. His boss, James Steffes was *Vice President of Government Affairs.* Richard Schapiro was *Vice President of Regulatory Affairs*. Richard Sanders was *Vice President and Assistant General Counsel*. Steven Kean was *Executive Vice President and Chief of Staff.* Vincent Kaminski was a *Ph.D. economist and Head of Research for Enron Corp*. Mary Han was a *lawyer at Enron's West Coast trading hub*. From the timeline, we found all these people except Vince were very active in this event. We will further analyze their roles in Section 5.



(a) Trends of two yearly repeating events.



Keywords with $P(w\|z)$	power 0.089361 California 0.088160 electrical 0.087345 price 0.055940 energy 0.048817 generator 0.035345 market 0.033314 until 0.030681
Key people with $P(u\|z)$	Jeff_Dasovich 0.249863 James_Steffes 0.139212 Richard_Shapiro 0.096179 Mary_Hain 0.078131 Richard_Sanders 0.052866 Steven_Kean 0.044745 Vince_Kaminski 0.035953

(b) The trend of "California Power" and most related keywords and people.

Figure 3. Topic trends

4.2 Predicting Community Patterns

We assume that, people communicate with certain people only under certain few topics. People in the same community under a topic would share the information. Thus, if there is something new about one topic, people in that topic-community will most likely get the information and propagate it to others in the community. Finally, many people in the community will get the information.

To evaluate our assumption and answer the question of who will be possibly involved in an observed email, we collect the ground truth about who are the senders and receivers for the emails and use the CTR algorithm to infer $P(\langle u,r\rangle | z_j, t_{new})$ by $P(\langle u,r\rangle | z_j, t_{old})$. We partitioned the data into training set and testing set. We tried two strategies for this experiment. First is to randomly partition the data into a training set with 8465 messages and a testing set with 8408 messages. Prediction accuracy is calculated by comparing the inference results and the ground truth (*i.e.*, receiver-sender pair of that email). We found that 96.8446% people stick in the old topics they are familiar with. The second strategy is to partition data by time: emails before 1/31/2000 as the training data (8011) and after that as the testing data (8862). We found 89.2757% of the people keep their old topics. Both results are quite promising. It is found that people really stick in old topics they are familiar with.

5. INDIVIDUAL ANALYSIS

In this section, we evaluate the performance of *CommunityNet*. First, we show how people's roles in an event can be inferred by *CommunityNet*. Then, we show the predicting capability of the proposed model in experiments.

5.1 Role Discovery

People with specific roles at company hierarchy behave specifically on specific topics. Here we show it is possible to infer people's roles by using *CommunityNet*.

In Section 4.1.2, we show there are some key people involved in "California Energy Crisis". In reality, Dasovich, Steffes, Schapiro, Sanders, and Kean, were in charge of government affairs. Their roles were to "solve the problem". Mary Hain was a lawyer during the worst of the crisis and attended meetings with key insiders. We calculated the correlation coefficients of the trends of these people and the overall trend of this topic. Jeff Dasovich got 0.7965, James Steffes got 0.6501, Mary Hain got 0.5994, Richard Shapiro got 0.5604, Steven Kean got 0.3585 (all among the 10 highest correlation scores among 154 people), and Richard Sanders got 0.2745 (ranked 19), while Vince Kaminski had correlation coefficient of -0.4617 (Figure 4). We can see that all the key people except Vince Kaminski have strong correlation with the overall trend of "California Energy Crisis". From their positions, we can see that all of them were sort of politicians while Vince Kaminski is a researcher. Thus, it is clear to see the difference of their roles in this topic.

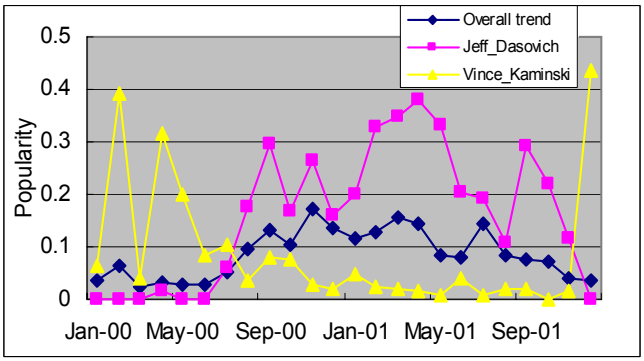


Figure 4. Personal topic trend comparison on "California Power"

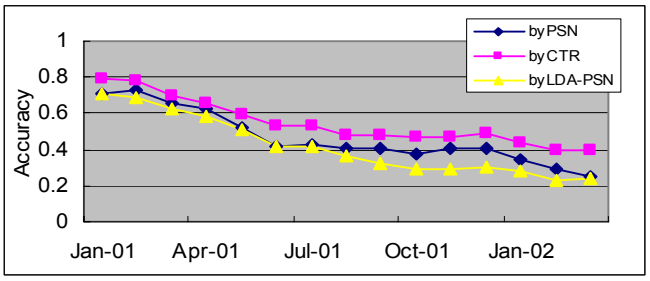
5.2 Predicting Receivers

Here we want to address the problem of whether it is possible to infer who will possibly be the receivers by a person's own historic communication records and the content of the email-to-send. One possible application is to help people organize personal social capital. For instance, if a user has some information to send or a question to ask, *CommunityNet* can recommend the right persons to send the info or get the answer.

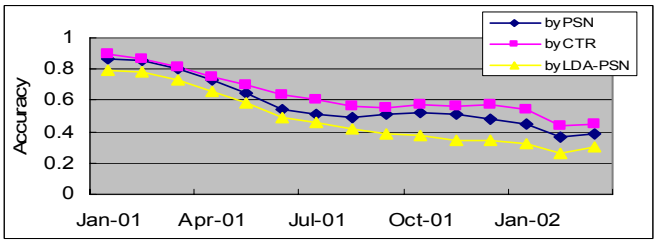
We conduct experiments by partitioning the dataset into a training set with the emails from 1999 to 2000, and a testing set with the emails from 2001 to 2002. The testing set is further partitioned into sub-sets with emails from one month as a subset. With this, we have 15 testing sets. (We exclude the emails after March 2002 because the total number of emails after that is only 78.) One issue we want to mention is that the number of people from 1999 to 2000 is 138, while from 2001 to 2002 is 154. In this study, we test each email in the training set by using its content,

sender, and time as prior information to predict the receiver, which is compared to the real receiver of that email.

In Figure 5, we illustrate the prediction performance by comparing the CTR algorithm, PSN, and the aggregated LDA-PSN model. The result shows that CTR beats PSN by 10% on accuracy. The aggregated LDA-PSN model performs even worse than PSN, because of the inaccurate clustering results. The performance gain is 21%. Moreover, intuitively, personal contacts evolve over time. Models built at a specific time should have decreasing predicting capability over time. In this figure, we obtain strong evidence of this hypothesis by observing that the performance of these models monotonically decays. This also implies our models well match the practice.



(a) Accuracy based on the top 5 most likely people



(b) Accuracy based on the top 10 most likely people

Figure 5. Prediction Accuracy comparisons. Accuracy is measured by testing whether the "real" receiver is among the prediction list of the top 5 or 10 most likely people

5.3 Inferring Senders

We test whether it is possible to infer who will possibly be the senders given a person's *CommunityNet* and the content of the email. One possible application is to exclude spam emails or detect identification forgery. Figure 6 illustrates the prediction result, which also shows the prediction accuracy decays over time.

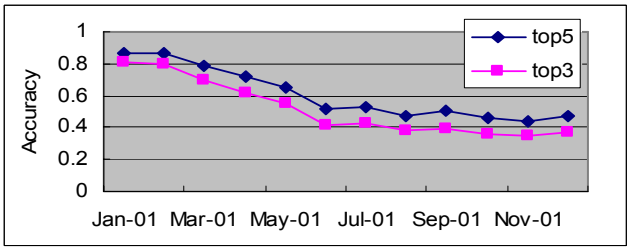
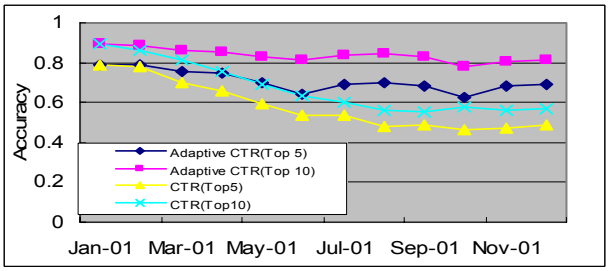


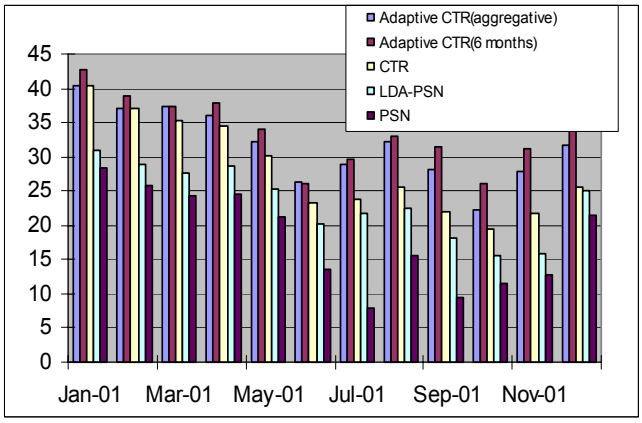
Figure 6. Predicting senders given receiver and content

5.4 Adaptive Prediction

We observed the prediction performance decays over time from the results of 5.2 and 5.3, which reflects the changes of the nature of email streams. Here we apply adaptive prediction algorithms we mentioned in 3.4.3, in which we incrementally and adaptively estimate statistical parameters of the model by gradually forgetting out-of-state statistics.



(a). Comparison between Adaptive CTR and CTR models



(b) Comparison of algorithms using Breese evaluation metrics

Figure 7. Performance evaluation for adaptive prediction algorithm and overall comparison

Figure 7 (a) illustrates the performance of the Adaptive CTR algorithm and compares it to the CTR algorithm. For the data far away from the training data, the improvement is more than 30%. And, if we compare it to the PSN and LDA-PSN algorithms, the performance gains are 58% and 75%, respectively. Evaluation by this accuracy metric tells us how related the top people ranked in the prediction results are. To understand the overall performance of the ranked prediction results, we apply the evaluation metric proposed by Breese [30], and illustrate the overall comparison in Figure 7(b). This metric is an aggregation of the accuracy measurements in various top-*n* retrievals in the ranked list. Among all predictive algorithms, adaptive CTR models perform best and PSN performs worst. In adaptive CTR models, estimating from recent data of six months beats aggregative updating the model from all the data from the history.

6. COMMUNITYNET APPLICATIONS

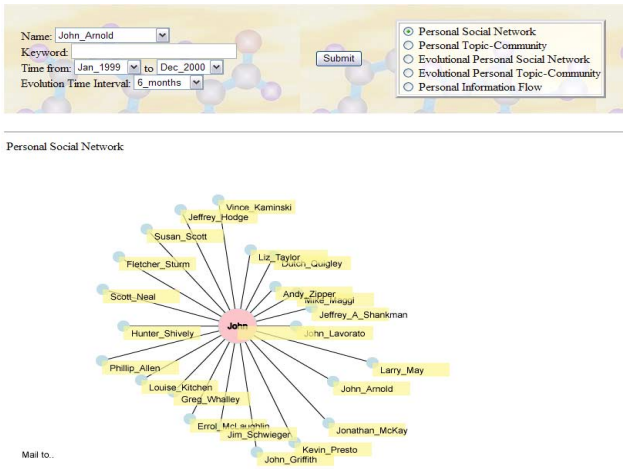
In this section, we show two application systems we built based on the *CommunityNet*. The first one is a visualization and query tool to demonstrate informal networks incorporation. The second one is a receiver recommendation tool which can be used in popular email systems. These demos can be accessed from http://nansen.ee.washington.edu/CommunityNet/.

6.1 Sensing Informal Networks

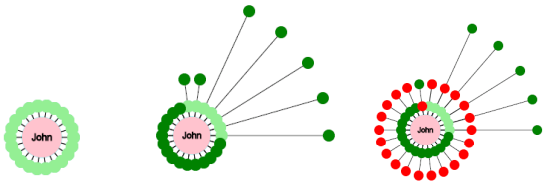
6.1.1 Personal Social Network

Figure 8 illustrates the interface of a visualization and query system of CommunityNet. The distance of nodes represents the closeness (measured by the communication frequencies) of a person to the center person. Users can click on the node to link to

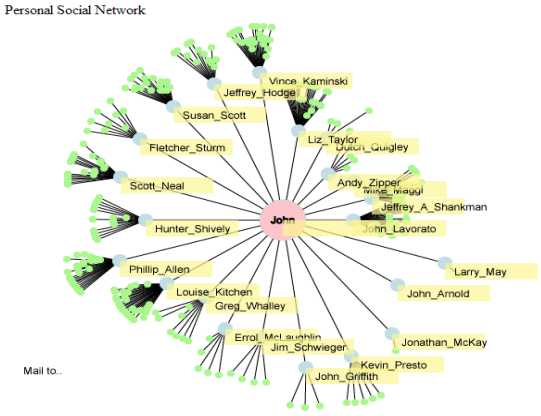
the CommunityNet of another person. This system can show personal social networks, which includes all the people a user contacts with during a certain time period. For instance, Figure 8(a) illustrates the personal social network of Vice President John Arnold from January 1999 to December 2000. During this period, there were 22 people he sent emails to, regardless what they were talking about. An evolutionary personal social network is illustrated in Figure 8(b), in which we show people's personal social network changes over time. From Jan. 1999 to Dec. 2000, no new contact was added to John's PSN. However, people's relationship changed in 2000. A Personal Social Network Information Flow is illustrated in Figure 8(c), in which we show how the information flows through the network (here we illustrate the information in two levels.)



(a) Personal Social Network of John Arnold



(b1) Jan-'99 to Dec-'99 (b2) Jan-'00 to Jun-'00 (b3) Jul-'00 to Dec-'00
(b) Evolutionary Personal Social Network



(c) Personal Social Network Information Flow with two-level personal social network of John Arnold

Figure 8. Personal social networks of John Arnold

6.1.2 Personal Topic-Community Network

Personal topic-community network can show whom this user will contact with under a certain topic. On retrieval, keywords are required for inferring the related topics. Figure 9 illustrates several personal topic-community networks for John Arnold. First, we type in "Christmas" as the keyword. CommunityNet infers it as "holiday celebration" and shows the four people John contacted with about this topic. About "Stock", we find John talked with five people on "Stock Market" and "Company Share" from Jan. 1999 to Dec. 2000. Personal Topic-Community network can be depicted by the system, too.

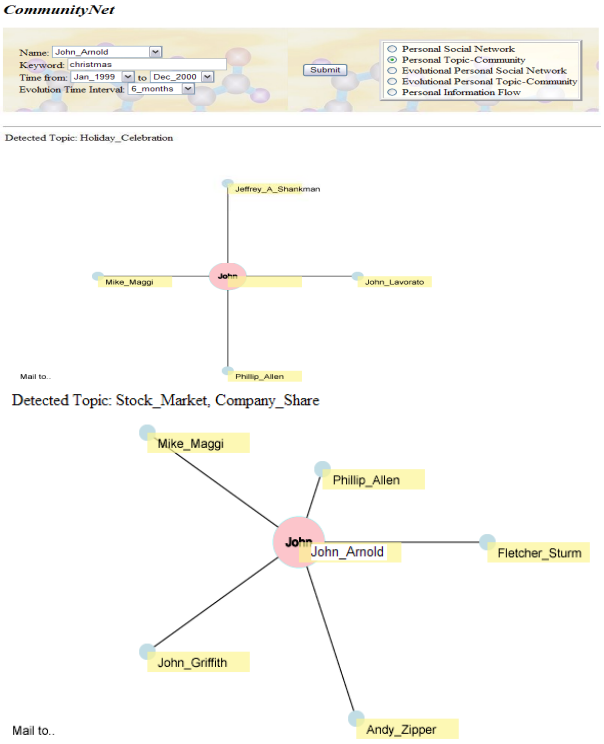


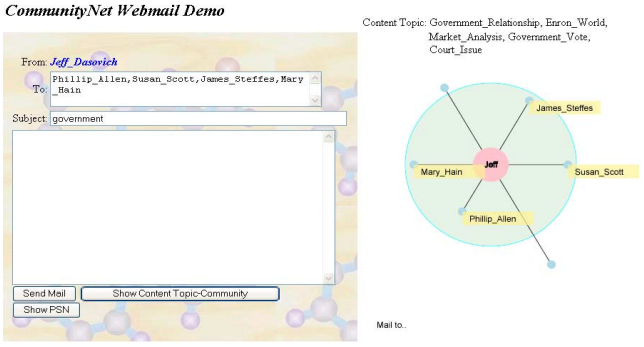
Figure 9. Personal Topic-Community Networks when we type in "Christmas" and "Stock"

6.2 Personal Social Capital Management - Receiver Recommendation Demo

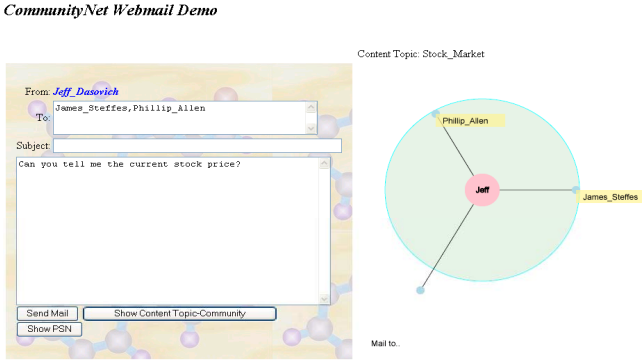
When a user has some questions, he/she may want to know whom to ask how to find an expert and who may tell him/her more details because of their close relationships. In our second demo, we show a *CommunityNet* application which addresses this problem. This tool can be incorporated with general email systems to help users organize their personal social capitals. First, after a user login a webmail system, he can type in content and/or subject then click on the "Show Content Topic-Community". This tool shall recommend appropriate people to send this email to, based on the learned personal social network or personal topic-community. The distances of nodes represent the closeness of the people to the user. Users can click on the node to select an appropriate person to send email to. If the center node is clicked, then a sphere grows to represent his ties to a group of experts. Click on "Mail To", then the people in the sphere will be included in the sender list.

In the examples in Figure 10, we log in as Jeff Dasovich. He can ask his closest friends whenever he has questions or wants to disseminate information. If he wants to inform or get informed on

"Government" related topics, the system will suggest him to send emails to Steffes, Allen, Hain, or Scott. The topics are inferred by matching the terms from the Subject as well as the content of the email. He can also type in "Can you tell me the current stock price?" as the email content. This system will detect "Stock Market" as the most relevant topic. Based on Dasovich's *CommunityNet*, it shows three possible contacts. He then chooses appropriate contact(s).



(a) Receiver recommendation for "Government"



(b) Receiver recommendation for "Can you tell me the current stock price?"

Figure 10. Receiver recommendation demo system

7. CONCLUSIONS AND FUTURE WORK

In this paper, we propose a new way to automatically model and predict human behavior of receiving and disseminating information. We establish personal *CommunityNet* profiles based on a novel *Content-Time-Relation* algorithm, which incorporates contact, content, and time information simultaneously from personal communication. *CommunityNet* can model and predict the community behavior as well as personal behavior. Many interesting results are explored, such as finding the most important employees in events, predicting senders and receivers of emails, *etc.* Our experiments show that this multi-modality algorithm performs better than both the social network-based predictions and the content-based predictions. Ongoing work includes studying the response time of each individual to emails from different people to further analyze user's behavior, and also incorporating nonparametric Bayesian methods such as hierarchical LDA with contact and time information.

8. ACKNOWLEDGMENTS

We would like to thank D. Blei, T. Griffiths, Yi Wu and anonymous reviewers for valuable discussions and comments. This work was supported by funds from NEC Labs America.

9. REFERENCES

[1] B. A. Nardi, S. Whittaker, and H. Schwarz. "It's not what you know, it's who you know: work in the information age," First Mon., 5, 2000.

[2] D. Krackhardt, "Panel on Informal Networks within Formal Organizations," XXV Intl. Social Network Conf., Feb. 2005.

[3] D. Krackhardt and M. Kilduff, "Structure, culture and Simmelian ties in entrepreneurial firms," Social Networks, Vol. 24, 2002.

[4] B. Nardi, S. Whittaker, E. Isaacs, M. Creech, J. Johnson, and J. Hainsworth, "ContactMap: Integrating Communication and Information Through Visualizing Personal Social Networks," *Com. of the Association for Computing Machinery*. April, 2002.

[5] https://www.linkedin.com/home?trk=logo.

[6] https://www.orkut.com/Login.aspx.

[7] http://www.friendster.com/.

[8] N. Lin, "Social Capital," Cambridge Univ. Press, 2001.

[9] W. Cohen. http://www-2.cs.cmu.edu/~enron/.

[10] S. Milgram. "The Small World Problem," Psychology Today, pp 60-67, May 1967.

[11] M. Schwartz and D. Wood, "Discovering Shared Interests Among People Using Graph Analysis", Comm. ACM, v. 36, Aug. 1993.

[12] H. Kautz, B. Selman, and M. Shah. "Referral Web: Combining social networks and collaborative filtering," Comm. ACM, March 1997.

[13] G. W. Flake, S. Lawrence, C. Lee Giles, and F. M. Coetzee. "Self-organization and identification of Web communities," IEEE Computer, 35(3):66–70, March 2002.

[14] J. Tyler, D. Wilkinson, and B. A. Huberman. "Email as spectroscopy: Automated Discovery of Community Structure Within Organizations," Intl. Conf. on Communities and Technologies., 2003.

[15] L. Page, S. Brin, R. Motwani and T. Winograd. "The PageRank Citation Ranking: Bringing Order to the Web," Stanford Digital Libraries Working Paper, 1998.

[16] J. Kleinberg. "Authoritative sources in a hyperlinked environment," In Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998.

[17] S. Wasserman, and P. E. Pattison, "Logit models and logistic regression for social networks: I. An introduction to Markov graphs and p*", Psychometrika, 61: 401– 425, 1996.

[18] T. A.B. Snijders. "Models for Longitudinal Network Data," Chapter 11 in Models and methods in social network analysis, New York: Cambridge University Press, 2004.

[19] D. L.-Nowell and J. Kleinberg, "The Link Prediction Problem for Social Networks," In Proceedings of the 12th Intl. Conf. on Information and Knowledge Management, 2003.

[20] J. Kubica, A. Moore, J. Schneider, and Y. Yang. "Stochastic Link and Group Detection," In Proceedings of the 2002 AAAI Conference. Edmonton, Alberta, 798-804, 2002.

[21] M. Handcock and D. Hunter, "Curved Exponential Family Models for Networks," XXV Intl. Social Network Conf., Feb. 2005.

[22] T. Hofmann, "Probabilistic Latent Semantic Analysis," Proc. of the Conf. on Uncertainty in Artificial Intelligence, 1999.

[23] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," Journal of Machine Learning Research, 3:993-1022, January 2003.

[24] T. Griffiths and M. Steyvers, "Finding Scientific Topics," Proc. of the National Academy of Sciences, 5228-5235, 2004.

[25] M. R.-Zvi, T. Griffiths, M. Steyvers and P. Smyth, "The Author-Topic Model for Authors and Documents", Proc. of the Conference on Uncertainty in Artificial Intelligence volume 21, 2004.

[26] A. McCallum, A. Corrada-Emmanuel, and X. Wang, "The Author-Recipient-Topic Model for Topic and Role Discovery in Social Networks: Experiments with Enron and Academic Email," Technical Report UM-CS-2004-096, 2004.

[27] X. Song, B. L. Tseng, C.-Y. Lin, and M.-T. Sun, "ExpertiseNet: Relational and Evolutionary Expert Modeling," 10th Intl. Conf. on User Modeling, Edinburgh, UK, July 24-30, 2005.

[28] J. Allan, R. Papka, and V. Lavrenko. "On-line New Event Detection and Tracking," Proc. of 21st ACM SIGIR, pp.37-45, August 1998.

[29] http://en.wikipedia.org/wiki/Timeline_of_the_Enron_scandal.

[30] J. Breese, D. Heckerman, and C. Kadie. "Empirical analysis of predictive algorithms for collaborative filtering," Conf. on Uncertainty in Artificial Intelligence, Madison,WI, July 1998.