Relationship Identification for Social Network Discovery

Christopher P. Diehl
Applied Physics Laboratory
Johns Hopkins University
Laurel, MD 20723

Galileo Namata and Lise Getoor *
Computer Science Department / UMIACS
University of Maryland
College Park, MD 20742

Abstract

In recent years, informal, online communication has transformed the ways in which we connect and collaborate with friends and colleagues. With millions of individuals communicating online each day, we have a unique opportunity to observe the formation and evolution of roles and relationships in networked groups and organizations. Yet a number of challenges arise when attempting to infer the underlying social network from data that is often ambiguous, incomplete and context-dependent. In this paper, we consider the problem of collaborative network discovery from domains such as intelligence analysis and litigation support where the analyst is attempting to construct a validated representation of the social network. We specifically address the challenge of relationship identification where the objective is to identify relevant communications that substantiate a given social relationship type. We propose a supervised ranking approach to the problem and assess its performance on a manager-subordinate relationship identification task using the Enron email corpus. By exploiting message content, the ranker routinely cues the analyst to relevant communications relationships and message traffic that are indicative of the social relationship.

Introduction

The Internet provides an increasing number of avenues for communication and collaboration. From instant messaging and email to wikis and blogs, millions of individuals are generating content daily that reflects their relationships with others in the world, both online and offline. Now that storage has become vast and inexpensive, much of this data will be archived for years to come. This provides new opportunities and new challenges. As networked groups and organizations increasingly leverage online means of communication and collaboration, there is an opportunity to observe the formation and evolution of roles and relationships from the communications archives. Such data provides a rich collection of evidence from which to infer the structure, attributes and dynamics of the underlying social network. Yet numerous challenges emerge as one contends with data that is often ambiguous, incomplete and context-dependent.

If we wish to analyze the underlying social network that is at least partially represented by a collection of informal, online communications, it is important to think carefully about the data transformations required prior to conducting any type of analysis. At the highest level, we are fundamentally interested in discovering entities and the types of relationships they share. This implies that we must do more than simply adopt the communications (hyper)graph as a surrogate for the social network. Entities can and often do use more than one account online and not all communications relationships are equivalent. In fact, the social network can be thought of as a collection of networks with different relationship types (e.g. friendship, trust, advice, management). Human relations are multi-faceted and context-dependent. Therefore it is important to tease the communications apart and understand what types of relationships are being expressed among the entities.

We view the network discovery process of identifying the entities and their relationships as being inherently a collaborative process between human and machine. In this paper, we consider the scenario from domains such as intelligence analysis and litigation support where an analyst is attempting to reconstruct a representation of the social network from the data with minimal context. This involves mapping the communications graph, which represents communication events among network references (email addresses, telephone numbers, etc.), to a validated social network expressing typed relationships among the known entities that the analyst believes are substantiated by the data. Within this process, there are two distinct tasks: *entity resolution* and *relationship identification*. Entity resolution refers to the mapping of network references to their corresponding entities. Relationship identification refers to the identification of relevant communications that are indicative of a given relationship type.

In this paper, we propose a supervised ranking approach to address the relationship identification problem. Our goal is to focus the analyst's attention on relevant communications relationships that express a given social relationship along with relevant message traffic that supports this association. We begin the discussion in the following section with a formal definition of the problem. We discuss our approach to learning a relationship ranker from traffic statistics and message content and present an evaluation of these methods on a manager-subordinate relationship identification task in email. We then review related work and con-

*This work was supported by NSF Grant #0423845.

clude with thoughts on future directions.

Problem Definition

Informal, online communications such as instant messaging, text messaging and email are composed of structured and unstructured data. At the most basic level, this includes the network references corresponding to the sender and one or more recipients, the date and time of the communication and the message content. We will define a *communications archive* $\mathcal{C}$ as a set of observed messages exchanged among a set of network references N:

$$\mathcal{C} = \{m_k = (n_k^s, N_k^r, d_k, b_k) : n_k^s \in N, N_k^r \subseteq N\}. \quad (1)$$

For each message m_k, n_k^s is the sender's network reference, N_k^r is the set of recipient network references, d_k is the date and time and b_k is the body of the message. Every archive has a corresponding *communications graph* $\mathcal{C}_g = \{N, L\}$ that represents the message data as a set of dyadic communication relationships

$$L = \left\{l_{ij} = (n_i^s, n_j^r, M_{ij}) : n_i^s, n_j^r \in N, M_{ij} \subseteq \mathcal{C}\right\}. \quad (2)$$

among the network references N. For each directed relationship l_{ij}, n_i^s is the sender's network reference, n_j^r is the recipient's network reference and M_{ij} is the set of messages sent by n_i^s that include n_j^r as one of the recipients.

The task of relationship identification involves identifying a mapping from the dyadic communications relationships L to one or more social relationships from a predefined set S. To emphasize the collaborative nature of our approach to the task, it is not our intention to develop an algorithm that automatically maps communications relationships to social relationships without intervention. A validated social network is one that the analyst believes is supported by evidence in the data. Therefore the machine's role in a collaborative approach to the task is to focus the analyst's attention on potentially relevant relationships along with supporting evidence in the message traffic.

We envision the analyst navigating the communications graph by following paths and incrementally investigating relationships in the ego networks corresponding to network references along the path. The *ego network* for a given entity in a network is generally defined as the subgraph that represents all of the direct relationships between the selected entity (the ego) and others (the alters). Formally in the case of the communications graph, the ego network $\mathcal{E}(n_i)$ for a given network reference $n_i \in N$ can be defined as

$$\mathcal{E}(n_i) = \mathcal{E}_o(n_i) \cup \mathcal{E}_i(n_i) \quad (3)$$

where

$$\mathcal{E}_o(n_i) = \{l_{ij} = (n_i, n_j, M_{ij}) \in L\}. \quad (4)$$

is the set of directed communications relationships from the ego to the alters and

$$\mathcal{E}_i(n_i) = \{l_{ji} = (n_j, n_i, M_{ji}) \in L\}. \quad (5)$$

is the set of directed communications relationships from the alters to the ego. For the purposes of ranking communications relationships within an ego network, we will initially restrict our attention to the set $\mathcal{E}_o(n_i)$ to avoid training and testing on the same message traffic.

Relationships in a given ego network $\mathcal{E}_o(n_i)$ will be ranked with a learned scoring function h that assigns a real-valued score to the relationship indicating its relative likelihood of expressing the social relationship of interest. If multiple social relationships are defined in the set S, there will be a corresponding scoring function for each social relationship. The task therefore is to learn a scoring function from a set of known relationships that successfully ranks relevant communications relationships higher than irrelevant relationships.

Learning to Rank Relationships

Objective

From initial exploration of the data or external sources of information, we assume a set of ego networks in the communications graph have been labeled, indicating whether or not the communications relationships exhibit the social relationship of interest. Initially we will approach the problem of learning multiple scoring functions independently. Therefore in each learning exercise, our goal is to learn a single scoring function for the given social relationship.

For a subset $N_t \subseteq N$ of network references in the collection, we assume the corresponding set of ego networks

$$\bar{\mathcal{E}} = \{\bar{\mathcal{E}}(n_i) : n_i \in N_t\} \quad (6)$$

are fully labeled

$$\bar{\mathcal{E}}(n_i) = \{(l_{ij}, s_{ij}) : l_{ij} \in L, s_{ij} \in \{0, 1\}\} \quad (7)$$

where s_{ij} indicates whether the communications relationship exhibits the given social relationship. Given a feature extraction process $f(l) \in \mathbb{R}^p$ that maps a specified communications relationship r to a p-dimensional feature vector, we can reexpress the labeled training data as

$$\bar{\mathcal{F}} = \{\bar{\mathcal{F}}(n_i) : n_i \in N_t\} \quad (8)$$

where

$$\bar{\mathcal{F}}(n_i) = \{(f_{ij}, s_{ij}) : l_{ij} \in L, f_{ij} = f(l_{ij}), s_{ij} \in \{0, 1\}\}. \quad (9)$$

The goal is to estimate a scoring function h that yields good generalization performance in terms of the *mean reciprocal rank* of relevant relationships on unseen ego networks. The rank of a relevant relationship is defined with respect to the irrelevant relationships within the corresponding ego network. For the ego network $\bar{\mathcal{E}}(n_i)$,

$$\mathcal{F}_r(n_i) = \{f_{ij} : (f_{ij}, s_{ij}) \in \bar{\mathcal{F}}(n_i), s_{ij} = 1\} \quad (10)$$

is the set of feature vectors corresponding to the relevant communications relationships and

$$\mathcal{F}_o(n_i) = \{f_{ij} : (f_{ij}, s_{ij}) \in \bar{\mathcal{F}}(n_i), s_{ij} = 0\} \quad (11)$$

is the set of feature vectors for the irrelevant communications relationships. The *rank* $r(f_r, n_i)$ of a relevant relationship $f_r \in \mathcal{F}_r(n_i)$ is therefore defined as

$$r(f_r, n_i) = 1 + |\{f_o : h(f_o) \geq h(f_r), f_o \in \mathcal{F}_o(n_i)\}| \quad (12)$$

where $h(f) \in \mathbb{R}$. The mean reciprocal rank $MRR(\bar{\mathcal{F}})$ for the scoring function on the labeled ego networks is then

$$MRR(\bar{\mathcal{F}}) = \frac{1}{R} \sum_{n \in N_t} \sum_{f_r \in \mathcal{F}_r(n)} \frac{1}{r(f_r, n)} \tag{13}$$

where $R = |\cup_{n \in N_t} \mathcal{F}_r(n)|$.

Approach

Given the complexity of learning a scoring function that directly optimizes the mean reciprocal rank, we will indirectly optimize a bound on this criteria by minimizing the number of *rank violations* committed by the scoring function. The ranking performance of the scoring function can be assessed by considering how well the function satisfies a series of pairwise ranking constraints. For every possible pairing of relevant and irrelevant relationships in an ego network, we desire a scoring function that scores the relevant relationships higher than the irrelevant relationships so that

$$h(f_r) - h(f_o) > 0$$

$$\forall f_r \in \mathcal{F}_r(n), f_o \in \mathcal{F}_o(n), n \in N_t. \tag{14}$$

A violation of one of these constraints is what we will refer to as a rank violation. Clearly the number of rank violations maps directly to the rank as implied by equation 12. Appendix A clarifies the connection between the number of rank violations and the mean reciprocal rank. The important observation is that the minimization of rank violations leads to maximization of a lower bound on mean reciprocal rank.

We pursue a large-margin approach to learning the scoring function following in the spirit of prior large-margin ranking work (Herbrich, Graepel, & Obermayer 1999; Joachims 2002; Yan & Hauptmann 2006). We define the *rank margin* as

$$m(f_r, f_o) = h(f_r) - h(f_o) \tag{15}$$

for a pair of relevant and irrelevant relationships (f_r, f_o). A positive rank margin implies the rank constraint for the pair is satisfied. The magnitude of the rank margin gives a measure of the degree of satisfaction.

We will assume the scoring function h takes a generalized linear form

$$h(f) = w \cdot \Phi(f) : \mathbb{R}^p \rightarrow \mathbb{R} \tag{16}$$

where Φ is an arbitrary nonlinear mapping. We will estimate the scoring function through minimization of the following regularized objective function

$$C(w) = \frac{1}{2}||w||^2 + \lambda \sum_{n \in N_t} \sum_{f_r \in \mathcal{F}_r(n)} \sum_{f_o \in \mathcal{F}_o(n)} g\left(m(f_r, f_o)\right) \tag{17}$$

where g is a convex margin loss function. At the optimum of this objective function,

$$w^* = \sum_{n \in N_t} \sum_{f_r \in \mathcal{F}_r(n)} \sum_{f_o \in \mathcal{F}_o(n)} \alpha(f_r, f_o) \left(\Phi(f_r) - \Phi(f_o)\right) \tag{18}$$

where $\alpha(f_r, f_o) = -\lambda g'(m^*(f_r, f_o))$ and $m^*(f_r, f_o)$ are the rank margins at the optimum. Substituting into equation 16, we find the optimum scoring function takes the form

$$h(f) = \sum_{n \in N_t} \sum_{f_r \in \mathcal{F}_r(n)} \sum_{f_o \in \mathcal{F}_o(n)} \alpha(f_r, f_o) \left(\Phi(f_r) - \Phi(f_o)\right) \cdot \Phi(f). \tag{19}$$

Given the transformed feature vectors enter the expansion solely as dot product terms, we can employ kernel functions $K(x, y) = \Phi(x) \cdot \Phi(y)$ satisfying Mercer's Theorem which provides a range of functional forms. This ultimately yields the general scoring function

$$h(f) = \sum_{n \in N_t} \sum_{f_r \in \mathcal{F}_r(n)} \sum_{f_o \in \mathcal{F}_o(n)} \alpha(f_r, f_o) \left(K(f_r, f) - K(f_o, f)\right). \tag{20}$$

The corresponding dual objective function for the general nonlinear case is obtained by substituting equations 18 and 20 into equation 17 yielding

$$\begin{aligned} C(\alpha) \;=\; & \frac{1}{2} \sum_{n \in N_t} \sum_{f_r \in \mathcal{F}_r(n)} \sum_{f_o \in \mathcal{F}_o(n)} \\ & \sum_{n' \in N_t} \sum_{f'_r \in \mathcal{F}_r(n')} \sum_{f'_o \in \mathcal{F}_o(n')} \\ & \alpha(f_r, f_o)\alpha(f'_r, f'_o)(K(f_r, f'_r) \\ & -K(f_o, f'_r) - K(f_r, f'_o) + K(f_o, f'_o)) \\ & +\lambda \sum_{n \in N_t} \sum_{f_r \in \mathcal{F}_r(n)} \sum_{f_o \in \mathcal{F}_o(n)} g\left(m(f_r, f_o)\right) \end{aligned} \tag{21}$$

Message Ranking

After ranking communications relationships with the scoring function, a natural question to ask is how does each message contribute to the overall score for a given relationship? If we define a scoring function with the form

$$h(f) = w \cdot \Phi(f) = w \cdot \sum_{m_i \in M} \Phi'(f_m) = \sum_{m_i \in M} h_m(f_{m_i}) \tag{22}$$

where the relationship score can be expressed as a linear combination of message scores $h_m(f_{m_i})$, we can immediately assess the relative contributions and sort the messages based on the message scores. We will employ a feature space and kernel function for content-based relationship ranking that admits this decomposition.

Manager-Subordinate Relationship Identification

To evaluate the utility of the proposed approach, we consider the problem of manager-subordinate relationship identification within an email archive. For this task, the goal is to identify relationships within each ego network where the alter is the ego's manager. In the following, we present two relationship summarization methods for exploiting relationship traffic statistics and message content.

From	*Recipients Include*	*From*	*Recipients Include*
n_a	n_b	n_b	n_a
n_a	n_c and not n_b	n_b	n_c and not n_a
n_c	n_a and not n_b	n_c	n_b and not n_a
n_c	n_a and n_b		

Table 1: List of possible communications events corresponding to a dyadic relationship (n_a, n_b). N_c is the common set of network references with whom both n_a and n_b communicate. n_c is a generic reference to any network reference in N_c.

Traffic-Based Relationship Ranking

In a hierarchical organization, it seems reasonable to believe that traffic patterns alone can provide significant indicators of organizational structure, assuming that issues of observability do not unduly complicate matters. Within the literature, there is evidence that group structure evident in email communications corresponds well to organizational constructs (Tyler, Wilkinson, & Huberman 2003). Similarly, we investigate whether management behavior is evident in the traffic statistics.

For a given dyadic relationship (n_a, n_b), we compute a number of traffic-based features between the network references n_a, n_b and the set of network references N_c with whom both n_a and n_b communicate. n_c is a generic reference to any network reference in N_c. The common associates are included to allow the ranker to key on potential differences in communication patterns with fellow colleagues and the manager. For each type of communication event listed in Table 1, from the specified network reference that includes/excludes the specified recipients, we compute the number of messages of this type and the quartiles for the distribution of the number of recipients observed across those messages. Including summary statistics for the number of recipients is potentially important for capturing differences in information distribution behavior and indications of group communications/directives.

Content-Based Relationship Ranking

Although traffic statistics alone may be sufficient for ranking relationships, they do not provide insight to the analyst that enables her to make a judgement about the type of social relationship expressed. Ultimately message content must be identified that substantiates the social relationship. Therefore we will assess the performance of a ranker that directly exploits the message content and allows us to rank the messages within the communications relationship.

Prior to computing feature vectors for individual relationships, we perform filtering steps on the message content to remove spurious characters and eliminate text from previous messages in the thread. Then we construct a master term list for the communications archive to define the feature space. For each communications relationship, we summarize the traffic by simply counting the term frequencies across the set of messages corresponding to the communications relationship. No stop word removal or term weighting was applied prior to learning the ranker.

Approach	*MRR*
Content-Based with Attribute Selection	0.719
Content-Based	0.660
Traffic-Based (From n_a including n_c and not n_b)	0.613
Traffic-Based	0.518
Random	0.211
Worst-Case	0.141

Table 2: Mean reciprocal rank for the various approaches. The MRR reported for the learned rankers results from the best performing regularization parameter.

Results

To assess the performance of our approach, we utilized the Enron email dataset along with organizational ground truth derived from an internal Enron document. This dataset is the collection of email from the folders of 151 Enron employees released as part of the government investigation into Enron's financial practices. Our results are based on the UC Berkeley version of the collection containing approximately 250,000 unique email messages mainly occurring in the 2000-2002 time frame.

Using an internal Enron document specifying the direct reports for Enron employees over 2000-2001, we identified 43 individuals in the collection with observable manager-subordinate relationships and nontrivial ego networks. We constructed the ego networks corresponding to each employee over this time frame and retained only those relationships where a minimum of 5 emails were sent in each direction. The resulting ego networks range in size from 2 to 107 relationships.

For both traffic and content-based relationship ranking, we use a linear kernel function and evaluate generalization performance using leave-one-ego-network-out cross-validation. We report the mean reciprocal rank (MRR) for the best performing regularization parameter. We also provide results for the worst case, where all the rank constraints are violated, and the average case for random selection. The results are provided in Table 2.

Traffic-Based Relationship Ranking

The linear ranker trained on all of the traffic statistics performs well relative to the baselines. By reducing the feature space to a single dimension, we achieve a significant additional improvement. Ranking relationships solely based on the number of emails sent from the ego to the common network references and not to the alter yields the best performance. After some reflection on group dynamics, this result is intuitively appealing. First the feature emphasizes relationships where there is a large set of common network references. For a manager and subordinate, these will likely correspond to fellow members of the group that the manager leads. At the same time, the feature deemphasizes relationships where more emails are sent from the ego to the common set and the alter. When both ego and alter are colleagues, these events are more likely than when the alter is the manager.

Content-Based Relationship Ranking

We explored two content-based ranking approaches. In the first approach, a linear ranker was trained on the relationship term frequencies for all 19,067 terms. Examining the absolute value of the resulting weight vector, we determined the 1000 most discriminative terms. Then we trained another linear ranker only on the term frequencies for the selected terms.

We found that content-based ranking consistently outperforms traffic-based ranking. We also found that attribute selection provides a significant additional performance improvement. As shown in Table 2, the content-based ranker trained in the constrained term space yields the highest MRR of 0.719. Examining the top ranked terms in the weight vector, we find terms indicative of the relationship of interest. Some notable words appearing in the top 20 include "please", "report", "project", "termination", and "executed".

We note that there are some ego networks in which content-based ranking performs worse than traffic-based ranking. The messages in these relationships suggest that the problem may be caused by more complex relationships. For example, in one ego network where content-based ranking performs significantly worse, the ego is a senior legal analyst. Although this individual had only one assigned manager, she performed tasks for other individuals, such as writing and analyzing legal documents, similar to those performed for her direct manager.

Content-Based Message Ranking

To qualitatively evaluate message ranking, we examined the highly ranked messages identified by the top performing content-based ranker. In cases where the manager-subordinate relationship achieved rank 1, we found that definitive evidence was usually contained within the top 10 messages. Definitive evidence for this type of social relationship includes emails with weekly reports, vacation requests, and project assignments. For example:

From: Cheryl Nelson [cheryl.nelson@enron.com]
To: Mark E Taylor [.taylor@enron.com]
Subject: Holiday Vacation
Hi Mark,
I would like to take Wednesday, December 27th as a vacation day because I could not get a flight on the 26th. Since I do not plan to leave town until December 24, I could catch up with my work by working on sat. December 23rd. Let me know if this is okay with you.

Although the analyst may have some preconceived notions about the nature of the relationship that are accurate, there are other aspects that may be specific to the domain or organization and therefore difficult to anticipate. For example, message ranking revealed "workload updates" requested by one manager from subordinates. Workload updates are weekly reports. This process also identified emails that provide evidence for the social relationship in ways one would not expect. For example:

From: Christian Yoder [christian.yoder@enron.com]
To: Elizabeth Sager [elizabeth.sager@enron.com],
Genia Fitzgerald [genia.fitzgerald@enron.com]
Subject: Happiness
Happiness is looking at the new legal org chart (which Jan just now dropped on my desk). I always approach these dry documents as though they were trigrams resulting from throwing the coins and consulting the I-Ching. At the top of the trigram which I find myself listed in I see a single name: Elizabeth Sager, and at the bottom I see the name Genia FitzGerald. ... cgy

As this example hopefully illustrates, message ranking may help the analyst gain additional insights and move beyond evidence that can be discovered through simple keyword queries.

Related Work

In the scenario we are considering, where an analyst is examining a collection of online communications with minimal context a priori, it will be important to have a number of tools to examine the data from varying perspectives. By focusing solely on the communications events through analysis of the communications graph, we can identify groups/communities and key individuals that are influential based on their position in the graph. Yet in general, we can conclude little about the nature of the relationships without exploiting the corresponding content.

Within the context of email exploitation, McCallum et al. (McCallum, Corrada-Emmanuel, & Wang 2004) took the first step toward a richer model of email relationships by proposing a generative model that captures the dependencies between topics of conversation and relationships. Since then, several other generative models have been proposed (Wang, Mohanty, & McCallum 2005; Song *et al.* 2005; Zhou *et al.* 2006; Zhang *et al.* 2006) that support joint relationship-topic clustering or group-topic clustering. These algorithms provide utility when initially exploring the data. Yet as the analyst discovers various relationship types of interest, these approaches do not provide a mechanism to capture and exploit the analyst's relationship labels so that additional relevant content tailored to her information needs can be identified. Our approach therefore provides a complimentary capability by leveraging the context provided by the analyst.

Other related approaches in the literature have focused primarily on processing the communications events to understand the structure of the social network. Eckmann et al. (Eckmann, Moses, & Sergi 2003) develop an information-theoretic approach to email exchange that allows for separating static and dynamic structure which appears to correspond to formal and ad-hoc organizational structure. Tyler et al. (Tyler, Wilkinson, & Huberman 2003) present a group detection algorithm that segments the communications graph by eliminating edges with low betweenness centrality. The validity of the groups detected within HP Labs was verified through interviews. Diesner and Carley (Diesner & Carley 2005) analyze global properties of the Enron communications graph and rank network references using various centrality measures from social network analysis to identify influential individuals. O'Madadhain and

Smyth (O'Madadhain & Smyth 2005) propose an approach for ranking vertices in graphs representing event data and demonstrate a weak correlation between network reference rank and position in the organizational hierarchy using a corporate archive of email events.

In recent years, there has been increasing interest in defining learning methods that address ranking tasks (Herbrich, Graepel, & Obermayer 1999; Joachims 2002; Freund *et al.* 2003; Burges *et al.* 2005). Our approach is inspired by earlier work on large-margin methods for ranking (Herbrich, Graepel, & Obermayer 1999; Joachims 2002; Yan & Hauptmann 2006) that learn a scoring function through minimization of the number of rank violations on the training data. Similar to (Yan & Hauptmann 2006), our general objective is to learn a ranker that successfully ranks relevant objects higher than irrelevant objects across a set of object sets. In the case of (Yan & Hauptmann 2006), the object sets are collections of retrieved documents corresponding to various queries. In our scenario, the object sets are the communications relationships in the labeled ego networks. We chose to minimize the number of rank violations in order to indirectly maximize the mean reciprocal rank. As we have established in Appendix A, minimization of rank violations maximizes a lower bound on the mean reciprocal rank (MRR). Recent work (Joachims 2005; Burges, Ragno, & Le 2007) has examined the problem of directly optimizing multivariate performance measures similar to MRR that more accurately represent ranking performance across object sets of varying size. Additional work is needed to define suitable methods for direct optimization of the MRR.

Conclusions and Future Work

The overall process of inferring the underlying social network from a communications archive involves two main tasks: entity resolution and relationship identification. In this paper, we presented a formal definition of the relationship identification task and proposed a supervised ranking approach to the problem. We showed that through minimization of rank violations, we can indirectly learn a relationship ranker that maximizes a lower bound on the mean reciprocal rank. Through experimentation on the Enron email dataset, we demonstrated the utility of this approach on a manager-subordinate relationship identification task. Using traffic and content-based features, the ranking method is able to routinely cue the analyst to relevant communications relationships. Message ranking using the content-based ranker provided additional guidance by illuminating compelling evidence within the message traffic substantiating the social relationship.

Cueing the analyst to relevant relationships and message content is an important first step; yet it is only half of the collaborative cycle we envision. As the user navigates the communications graph, she will make judgements about relationships and message content. These judgements can be exploited to incrementally refine the scoring function as her exploration proceeds. The goal is to enable continuous learning behind the scenes that supports her in the discovery process. To realize this capability, a number of challenges must be addressed such as automated model selection (feature selection and hyperparameter tuning) and learning from multiple types of rank constraints indicating what relationships and message content are relevant. Other questions emerge about how to most effectively leverage unlabeled relationships in the communications graph and direct labeling efforts to rapidly accelerate the learning. These are some of the issues we will focus on in future research.

Appendix A: Lower Bound on the Mean Reciprocal Rank

In order to show that minimizing the number of rank violations is a reasonable proxy for maximizing the mean reciprocal rank (MRR), we need to understand how these quantities are related. For a fixed number of rank violations, the resulting MRR varies depending on how the rank violations are distributed across the relevant relationships. If the rank violations are concentrated, so that a small number of relevant relationships are low ranked, the MRR will be higher than the case where the same number of rank violations are distributed across a larger number of relevant relationships. It is this line of thought that leads to bounds on the MRR for a given number of rank violations.

Let us assume that there are M relevant relationships and that the maximum possible rank for the ith relationship is $N_r^i + 1$. This implies that the maximum number of rank violations that can be associated with the ith relevant relationship is N_r^i.

A useful analogy for this discussion is to imagine we have M bags and N balls. Each bag has one ball prior to assigning any of the N balls. The ith bag can hold $N_r^i + 1$ balls. In this scenario, the mean reciprocal rank is the mean reciprocal number of balls in a bag. To lower bound the MRR, we need to determine the assignment of N balls to the bags that minimizes the mean reciprocal number of balls in a bag.

To minimize the MRR for N balls, consider a process whereby the balls are incrementally assigned to the bags so that at each step the MRR is minimized. This implies that we want to assign the next ball to the bag that maximizes the incremental reduction in MRR. If there are b balls in a bag already, the incremental reduction in MRR from an additional ball is

$$\frac{1}{M}\left(\frac{1}{b} - \frac{1}{b+1}\right) = \frac{1}{Mb(b+1)}. \tag{23}$$

Therefore, at each step, we should add the next ball to a bag with the least number of balls that can accept an additional ball. By uniformly adding balls to bags that can accept them, we will maintain a minimum MRR throughout the process.

Let $S_k (k \geq 2)$ be the number of bags that can hold k or more balls. We will make p passes down the line of bags adding one ball to each bag that can accept one until all N balls are placed. On the ith pass,

$$B_i = \min\left(S_{i+1}, N - \sum_{j=0}^{i-1} B_j\right) \tag{24}$$

balls are placed. $B_0 = 0$ by definition. The lower bound on MRR is therefore

$$\begin{aligned} MRR_{\min} &= \frac{M - B_1}{M} + \frac{1}{M}\left(\frac{1}{p+1}B_p \right. \\ &\quad \left. + \sum_{i=1}^{p-1} \frac{1}{i+1}(B_i - B_{i+1})\right) \\ &= 1 - \frac{1}{M}\sum_{i=1}^{p} \frac{1}{i(i+1)}B_i. \end{aligned} \tag{25}$$

The key observation here is that all of the B_i for $i < p$ remain constant and B_p decreases as N decreases. Therefore the lower bound on MRR is strictly monotonically increasing with a decreasing number of rank violations.

References

Burges, C.; Shaked, T.; Renshaw, E.; Lazier, A.; Deeds, M.; Hamilton, N.; and Hullender, G. 2005. Learning to rank using gradient descent. In *Proc. of the Intl. Conf. on Machine Learning*.

Burges, C.; Ragno, R.; and Le, Q. 2007. Learning to rank with nonsmooth cost functions. In Schölkopf, B.; Platt, J.; and Hoffman, T., eds., *Advances in Neural Information Processing Systems 19*. MIT Press.

Diesner, J., and Carley, K. M. 2005. Exploration of communications networks from the Enron email corpus. In *Proc. of the Workshop on Link Analysis, Counterterrorism and Security, SIAM Intl. Conf. on Data Mining*, 3–14.

Eckmann, J.-P.; Moses, E.; and Sergi, D. 2003. Dialog in e-mail traffic. *ArXiv Condensed Matter E-Prints*.

Freund, Y.; Iyer, R.; Schapire, R.; and Singer, Y. 2003. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research* 933–969.

Herbrich, R.; Graepel, T.; and Obermayer, K. 1999. Support vector learning for ordinal regression. In *Proc. of the Ninth Intl. Conf. on Articial Neural Networks*, 97–102.

Joachims, T. 2002. Optimizing search engines using clickthrough data. In *Proc. of ACM SIGKDD*, 133–142.

Joachims, T. 2005. A support vector method for multivariate performance measures. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, 377–384. New York, NY, USA: ACM Press.

McCallum, A.; Corrada-Emmanuel, A.; and Wang, X. 2004. 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, University of Massachusetts Amherst.

O'Madadhain, J., and Smyth, P. 2005. EventRank: A framework for ranking time-varying networks. In *Proc. of the Intl. Workshop on Link Discovery, KDD Conf.*, 9–16.

Song, X.; Lin, C.-Y.; Tseng, B. L.; and Sun, M.-T. 2005. Modeling and predicting personal information dissemination behavior. In *Proc. of ACM SIGKDD*, 479–488.

Tyler, J. R.; Wilkinson, D. M.; and Huberman, B. A. 2003. Email as spectroscopy: Automated discovery of community structure within organizations. In *Communities and Technologies*. Kluwer, B.V. 81–96.

Wang, X.; Mohanty, N.; and McCallum, A. 2005. Group and topic discovery from relations and text. In *Proc. of the Intl. Workshop on Link Discovery, KDD Conf.*, 28–35.

Yan, R., and Hauptmann, A. G. 2006. Efficient margin-based rank learning algorithms for information retrieval. In *Intl. Conf. on Image and Video Retrieval*.

Zhang, D.; Gatica-Perez, D.; Roy, D.; and Bengio, S. 2006. Modeling interactions from email communications. In *Proc. of the IEEE Intl. Conf. on Multimedia and Expo*.

Zhou, D.; Manavoglu, E.; Li, J.; Giles, C. L.; and Zha, H. 2006. Probabilistic models for discovering e-communities. In *Proc. of the Intl. World Wide Web Conf.*, 173–182.