Inferring Relevant Social Networks from Interpersonal Communication

Munmun De Choudhury*
Arizona State University, Tempe, USA
munmun@asu.edu

Winter A. Mason
Yahoo! Research, NY, USA
winteram@yahoo-inc.com

Jake M. Hofman
Yahoo! Research, NY, USA
hofman@yahoo-inc.com

Duncan J. Watts
Yahoo! Research, NY, USA
djw@yahoo-inc.com

ABSTRACT

Researchers increasingly use electronic communication data to construct and study large social networks, effectively inferring unobserved *ties* (e.g. i is connected to j) from observed communication *events* (e.g. i emails j). Often overlooked, however, is the impact of tie definition on the corresponding network, and in turn the relevance of the inferred network to the research question of interest. Here we study the problem of network inference and relevance for two email data sets of different size and origin. In each case, we generate a family of networks parameterized by a threshold condition on the frequency of emails exchanged between pairs of individuals. After demonstrating that different choices of the threshold correspond to dramatically different network structures, we then formulate the relevance of these networks in terms of a series of prediction tasks that depend on various network features. In general, we find: a) that prediction accuracy is maximized over a non-trivial range of thresholds corresponding to 5–10 reciprocated emails per year; b) that for any prediction task, choosing the optimal value of the threshold yields a sizable ($\sim 30\%$) boost in accuracy over naïve choices; and c) that the optimal threshold value appears to be (somewhat surprisingly) consistent across data sets and prediction tasks. We emphasize the practical utility in defining ties via their relevance to the prediction task(s) at hand and discuss implications of our empirical results.

Categories and Subject Descriptors

H.1.2 [**Models and Principles**]: User/Machine Systems; J.4 [**Social and Behavioral Sciences**]: Sociology

General Terms

Algorithms, Experimentation, Human Factors

Keywords

Communication networks, email, learning, network structure, network thresholds, social networks, social network analysis, ties.

*Part of this research was performed while the author was visiting Yahoo! Research, New York.

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

1. INTRODUCTION

The rapidly growing volume of electronic communication data, such as that derived from email exchange, instant messaging, mobile phones, online games, social networking or social media sites, has been a great benefit to social network analysis, enabling researchers to study networks at very large scales and over extended time periods [19, 15, 16, 18, 20, 9, 5, 30]. However, the excitement generated by this explosion of available data has overshadowed two distinct but related problems: first, the *inference problem,* that "real" social ties are not directly observable and hence must be inferred from observations of events, like physical interactions or communication records; and second the *relevance problem,* that there is no one "true" social network, but rather many such networks, each corresponding to a different definition of a tie, and each relevant to different social processes.

To illustrate the interdependence of the inference and relevance problems, consider three possible definitions for an edge deriving from observed communication data[1]: (1) an edge exists between i and j if either has communicated with the other at least once in the past year; (2) an edge exists if each has communicated with the other at least once in the past week; and (3) an edge exists if each has communicated with the other at least once per week for the past year. Stated in isolation, each of these definitions is plausible; yet each of them could potentially yield very different networks, not only in terms of average density (i.e. number of edges), but also in terms of important structural features like path lengths, local clustering [33] and motifs, degree distribution, and community structure [26]. Moreover, which of these networks is the "relevant" one will in general depend on the research question of interest. For instance, if one is interested in a process like communication between trusted peers, where the relevant network is only made up of "strong ties", one might prefer definition (3) above; whereas if one is interested only in short term diffusion of information, one might prefer definition (2); or finally, if one is interested in communities that persist over extended time intervals, one may prefer definition (1). Of course, even these theoretically motivated edge definitions are unfortunately vague, as

[1]As we discuss in section 6, the inference/relevance problem applies quite generally to observational data [13], not just communication data, and even to networks generated by other methods like surveys [24, 9].

no empirically grounded theory yet exists that links some quantifiable definition of tie strength to, say, the transmission of some specific type of information or influence; indeed tie strength itself remains an ambiguous concept with multiple, possibly inconsistent definitions [12, 21]. Nevertheless, the example serves to illustrate that the inference problem (mapping observations to ties) cannot be resolved independently of the relevance problem.

Typically, however, in network analysis research these problems *are* addressed separately. That is, the researcher first nominates some plausible but ad-hoc definition of a tie, say in terms of a "threshold" condition for an edge (e.g. "a tie exists between i and j if and only if they have communicated at least once in the observed data"), and only then analyzes the network corresponding to that definition. In other words, rather than asking "For this problem, what is the most relevant network?," the researcher is asking, in effect "How is this (ad hoc) network relevant to my problem?" Although in any one instance this approach seems reasonable, because the researcher considers only one possible definition of the network, he or she has no way of knowing whether other possible definitions would have been more relevant to the problem at hand. Moreover, when different authors studying different data sets make different assumptions about how to infer network ties, comparisons across studies cannot be made in a meaningful way.

To address the combined inference/relevance problem, we propose that rather than defining ties based on intuitions about the data and only then studying the properties of the corresponding network, network analysts should instead define ties explicitly in terms of their relevance to the particular objective of interest. For example, if one is interested is in social influence and the diffusion of innovations or culture, we propose that the "relevant" network be inferred directly from some observed pattern of influence. Alternatively, if the objective is to partition the network into like-minded communities, one could identify the network that best captures the communities with shared beliefs. Or if one is interested in predicting which pairs are likely to communicate in the future, one might identify the network that best predicts previous communication activity.

In this paper we outline a primitive version of this approach, which proceeds in two stages. First, we first introduce a simple method for inferring networks from pair-wise communication data that admits an edge between two individuals only when their communication exceeds a certain threshold τ of intensity. The method is primitive because, in relying on an ad-hoc definition of a threshold, it suffers from some of the same problems that we seek to resolve. Nevertheless, as we illustrate with two email data sets—one, based on two years of server logs at a major US university [15]); and the other drawn from the publicly available Enron email corpus [29]—it is general enough to demonstrate that network structure can change dramatically as a function of the tie definition. This dependence is visually apparent in Figure 1, which depicts the largest component in networks for different values of the threshold.

Having demonstrated the potential impact of the choice of threshold, we then consider the issue of how to select the "correct" value in terms of its relevance to some empirically observed pattern of interest, where "relevance" is formalized as a prediction task. To illustrate the method, we again study the same two email datasets, identifying in each case the value of τ that best predicts (a) observable individual attributes like gender or status (e.g. for the UNIVERSITY dataset, "status" corresponds to student, faculty, affiliate, etc.), (b) the likelihood of future communication between pairs of individuals, or (c) (for the UNIVERSITY data) their co-membership in known communities. We emphasize that there is nothing special about these particular prediction tasks, which were chosen largely as a matter of convenience given the available data. Nevertheless, our results highlight the main contributions of our proposed approach:

1. We find that there exists a non-trivial threshold on edge weights—corresponding to about 5–10 reciprocated emails per year over which our set of chosen prediction tasks seem to yield maximum accuracy.
2. Choosing the above "optimal" threshold produces a sizeable ($\sim$ 30%) boost in accuracy over naïve approaches of network inference.
3. Finally, we observe that this optimal range of threshold values appears to be relatively consistent across both datasets and the various prediction tasks.

The rest of the paper is organized as follows. Section 2 gives an overview of the related work, highlighting the differences with our approach here. In Section 3, we introduce the two data sets that we use to illustrate our argument, and describe our method for inferring networks from communication data. In Section 4, we then study the properties of these families of networks, examining a range of network- and node-level metrics. Next, in Section 5, we conduct several prediction tasks on the inferred networks. Finally in section 6, we present our conclusions and discuss directions towards future work.

2. RELATED WORK

Use of interpersonal communication for network inference has been of interest to researchers for several decades. Early work [28] utilized email traffic to infer social networks for the purpose of discovering communities of shared interest. To arrive at their network, the authors developed a highly customized approach, discarding messages not thought to be relevant to shared interests (e.g. bulk emails, emails sent from administrative accounts, etc.), discarding nodes and edges not part of the main connected component, and pruning the main connected component to focus on "core" nodes. More recent work using email data has focused on the frequency of email exchange as an indicator of relevance. where in some cases [31, 10, 1] the authors specify a fixed threshold, mapping observed frequencies to binary (i.e. $w_{ij} \in \{0, 1\}$) edge weights, while in other cases directed, weighted edges are constructed [8]. Finally, some authors have applied a time-dependent threshold condition [6, 15, 5] to communication data in order to detect tie creation and deletion in dynamic networks.

Although motivated by different research questions, the approaches taken in these studies are consistent with that outlined in the introduction: a single definition of what constitutes a tie is chosen, largely on the basis of intuitive plausibility, and only the properties of that particular network are considered. The question of whether other possible definitions might have generated different results, and if so, which results are the relevant ones, is therefore rarely raised. Even

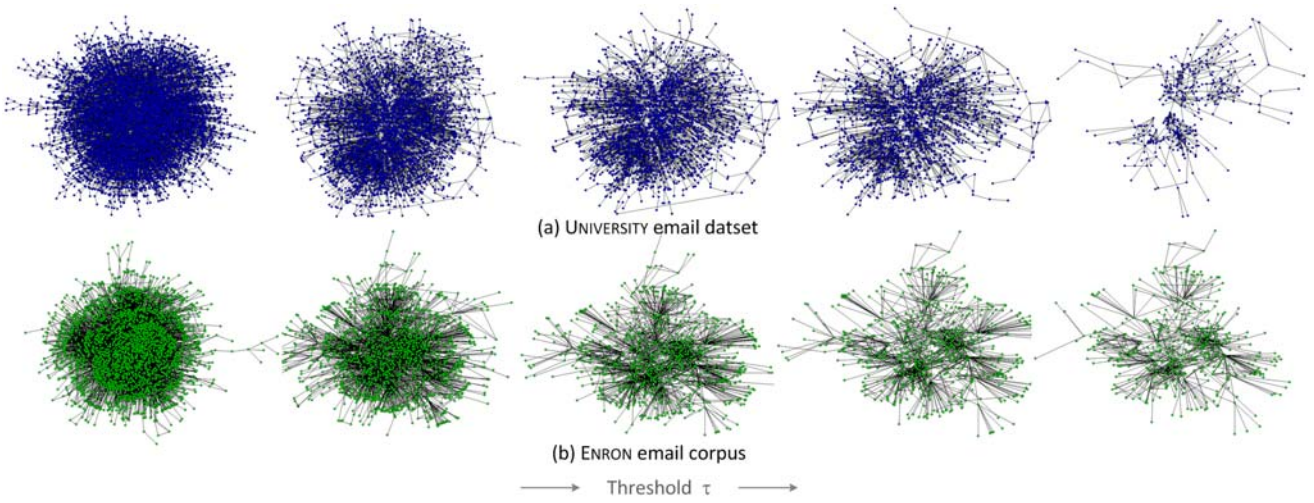

Figure 1: Topology of the largest components over various choices of threshold conditions for (a) a dataset based on email server logs at a US university, and (b) the Enron email corpus. Significant changes in topology are observed as the thresholding condition of the network is varied.

where alternative definitions are considered [15, 17], the purpose is exclusively to serve as a robustness check on the findings; thus the scope of possibilities is typically limited to within some range of the original choice of threshold. Most closely related to the current work are two recent studies using mobile phone data [27, 9]. In [27], the authors systematically deleted edges as a function of call frequency in order to investigate the connectivity of the network, and its impact on information diffusion. The main distinction between this study and the current work—aside from our broader focus on network properties other than connectivity—is that we not only show that different choices of threshold generate different networks, but also that some inferred networks are more relevant than others with respect to social processes (e.g. gender and status homophily) of interest. In [9], the authors use communication data to predict self-reported ties generated by a survey tool. Although this paper is similar to ours in its emphasis on relevance, [9] treats self-reported ties as the "ground truth," whereas we make no such assumption.

3. INFERRING SOCIAL NETWORKS

In this section we first describe the two communication datasets and then discuss our method for inferring the social networks from communication events. We emphasize that this method is by no means completely general—indeed, it embodies a number of assumptions that impose ad-hoc restrictions on the range of possible inferred networks. Nevertheless, it is sufficiently general to admit a broad family of networks, all of which are based on the same underlying set of observations but which vary dramatically in structure, as we show in Section 4.

3.1 Datasets

UNIVERSITY. Our first dataset is a complied registry of all email (incoming and outgoing, as recorded in server logs) associated with individuals at a large university in the United States, comprised of undergraduate and graduate students, faculty, and staff spanning over a period of two years (i.e. in the order: Fall, Spring, Summer, Fall, Spring, Summer). The emails contain encrypted IDs of the sender and recipient(s) of each email and the timestamp, but do not contain the content. The dataset also features several (anonymized) personal attributes, including status, gender, age, departmental affiliation, number of years in the community, dorm and home zipcode information for the students, as well as course affiliations for the students at each semester.

In order to focus on a population of users who use emails as a major communication mode, we have considered only the individuals who have email IDs associated with the university domain (non-university emails were excluded because we did not have complete information about external accounts). In addition, we excluded individuals who did not send at least one email in each of the six semesters under consideration; thus ensuring that we observed a consistent set of individuals engaged in regular inter-personal communication. After applying both of these restrictions, our data comprises 19,817 individuals with a total of 1,098,285 emails over the two year period.

ENRON. Our second dataset is a repository of the emails exchanged internally among the employees at the Enron Corporation, obtained through a subpoena as part of an investigation by the Federal Energy Regulatory Commission (FERC) and then made public. The data set comprises 4,736 individuals (including both Enron executive officers as well as individuals external to Enron but involved in communication), who sent 1,063,352 emails over the period 1998-2002. In addition, information about the status of the individuals in the corporation (e.g., "Director", "Trader", "Manager", etc.) was made available for public use. In contrast with the UNIVERSITY data, no filtering of the ENRON data was required.

3.2 Constructing Thresholded Networks

For the chosen set of individuals V in both UNIVERSITY and ENRON, we define the weight $w^s_{ij} = \sqrt{w_{ij}.w_{ji}}$ of an edge between two individuals i and j as the geometric mean[2] of

[2]Geometric mean of the number of emails between two users u_i and u_j ensures that there is *no* tie if communication

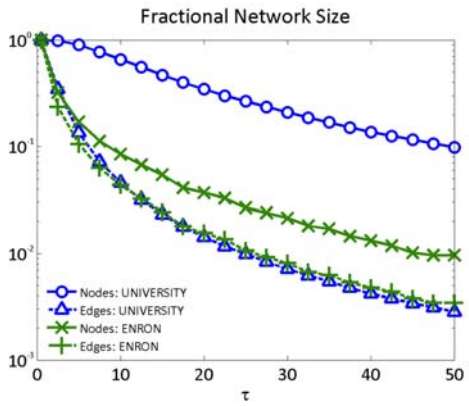

Figure 2: Fractional network size in terms of nodes and edges over different thresholds τ, and with respect to the network corresponding to minimum threshold ($\tau = 1$). Results are shown for the UNIVERSITY email dataset and the ENRON corpus.

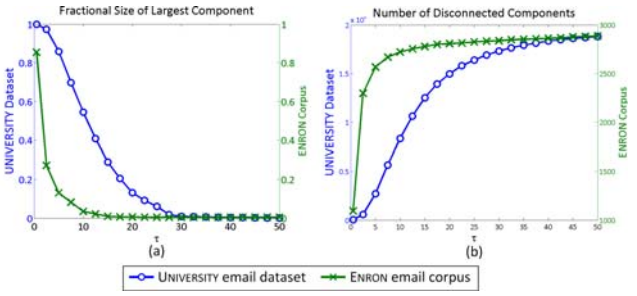

Figure 3: Changes in characteristics of the network components for the two email datasets.

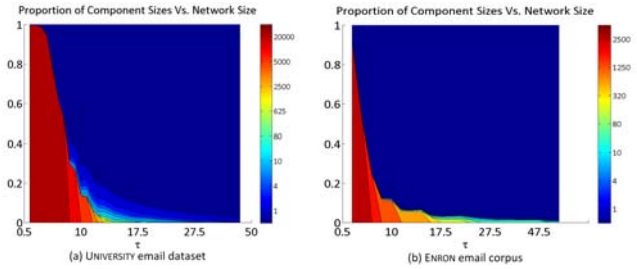

Figure 4: Sizes of different network components as a fraction of the entire network, shown over different τ, for the two datasets. The colorbar corresponds to the different component sizes.

the annualized rate of messages exchanged over the span of two and four years respectively, where w_{ij} is the number of emails sent per year from i to j. We then define a network $G(V, E_s; \tau)$ comprising the edges E_s between the pairs of nodes i and j in V whose edge weights w_{ij}^s exceed a specified threshold τ. By systematically varying τ, therefore, we can then obtain a family of networks, $\{G(\tau_1), G(\tau_2), \cdots, G(\tau_K)\}$ corresponding to more or less stringent definitions of what counts as a "relationship".

4. NETWORK DESCRIPTIVE STATISTICS

Having defined the family of networks $\{G(\tau_1), G(\tau_2), \cdots, G(\tau_K)\}$, we now investigate the variation in structural characteristics as a function of the threshold τ. To do this, we generate networks for the UNIVERSITY and ENRON email datasets, for values of τ ranging between 0.5 and 50[3]. For each of these networks, we then consider two sets of features: "network-level" features that capture properties of overall network size and connectivity; and "node-level" features like local clustering, bridging, and connectivity, that characterize individual nodes. The set of features we have considered is not intended to be exhaustive, nor are they necessarily more revealing of network structure than other possible choices. Nevertheless, they are commonly studied by network analysts, and are often invoked to justify substantive conclusions about outcomes of interest like the potential for information flow through a network or the relative influence of nodes (e.g. [4]).

4.1 Network-level Features

We first investigate the variation in the network size as a function of the threshold τ. Figure 2 shows the number of edges and (non-singleton) nodes for both email data sets, from which it is clear that the choice of τ makes a sizeable impact on the inclusion and exclusion of both nodes and edges. For example, by increasing the threshold from $\tau = 1$ to $\tau = 5$, the number of edges in the network are reduced by an order of magnitude.

There are two notable features apparent in Figure 2: first, the drop in the number of edges over increasing τ looks strikingly similar in the two datasets; and second the number of nodes included in the two networks diminishes at very different rates. The explanation for these different results is as follows. The distribution of edge weights is similar between the two datasets; thus the rate at which edges are removed with increasing τ is also similar. The ENRON dataset, however, contains many more "peripheral" nodes in the sense that nodes are connected to the core of the network by only a single edge; thus the rate at which nodes become isolates with increasing τ is initially much greater than in the UNIVERSITY dataset.

This can be seen in Figure 3, which shows the size and number of connected components as a function of the threshold. Figure 3(a) shows a dramatic drop in the fractional size in both datasets, and Figure 3(b) a correspondingly dramatic increase in the number of disconnected components; but the changes happen at a lower value of the threshold (around $\tau = 5$) for ENRON.

This can also be seen in Figure 4, which shows, as a function of τ, how the sizes of the different network components change as a fraction of the entire network. We observe that for both datasets, at $\tau = 0.5$ the majority of the nodes are in the largest component (size $\sim 80\%$). Initially as τ increases, nearly all of the deleted edges disconnect singletons from the giant component, and only after τ approaches ~ 10 are larger components disconnected.

between u_i and u_j is unidirectional. This eliminates any one-way communications, such as mass bulk emails from an administrator at a university.

[3]The natural starting point for τ is the lowest value for which both networks are defined. This corresponds to $\tau = 0.5$, or one email over the period of two years for the UNIVERSITY dataset

4.2 Node-level Features

A wide variety of structural features of nodes have been used for purposes such as understanding the structure of communities within a population [16, 25, 31], studying the flow of information between individuals and groups [19], and predicting the relative similarity of friends and strangers [28]. Following prior work, we now consider a selected set of features that can be roughly grouped into three categories: those that measure the *reach* of a node (node degree, average neighbor degree, size of two-hop neighborhood) [15, 18]; those that measure the *closure* of the ego network (embeddedness, clustering coefficient) [15, 18]; and those that measure how much the node is *bridging* communities (network constraint, number of ego components) [4, 22]. We first briefly review the definitions of these features and then present the results for both datasets.

Reach

- *Node Degree:* The degree of a node is defined as the total number of neighbors, or immediate contacts, given by the set $\Gamma_i = \{u_j : e_{ij} \in E_s\}$. For individual $u_i \in V$, $k_i = \|\Gamma_i\|$.
- *Average Neighbor Degree:* The average neighbor degree $k_i^{(n)}$ of a node i is defined as the mean degree over all of its immediate contacts.
- *Size of Two-hop Neighborhood:* Size of two hop neighborhood $k_i^{(2)}$ of a node i is the count of all of the node's neighbors plus all of the node's neighbor's neighbors. Note that this is a count of the nodes, and therefore is agnostic to how many edges there are between these nodes.

Closure

- *Embeddedness:* We define the embeddedness of a node with respect to its neighborhood as the mean of the ratio between the set of common contacts and the set of all contacts for the node and each neighbor. It is given as,

$$\mathcal{E}_i = \frac{1}{k_i} \sum_{u_j \in \Gamma_i} \frac{|\Gamma_i \cap \Gamma_j|}{|\Gamma_i \cup \Gamma_j|} \tag{1}$$

- *Normalized Clustering Coefficient:* The clustering coefficient of a node is a standard notion of local density[4] (i.e. "the average probability that two of my neighbors are neighbors of each other"), given by $c_i = \frac{2|e_{jm}|}{k_i(k_i-1)}$, where e_{jm} are the edges connecting $u_j, u_m \in \Gamma_i$ and Γ_i is the "neighborhood" of i. As we have seen, however, the graphs we are studying vary dramatically in terms of their global density as a function of the threshold; thus it is more informative to see how local density varies relative to global density, rather than in absolute terms. We therefore define the "normalized clustering coefficient" of a node as the ratio of the clustering coefficient and the graph density:

$$C_i = \frac{c_i}{k_i/(N-1)}, \tag{2}$$

where N is the number of nodes in the graph. Note that in contrast to c_i which varies between 0 and 1, C_i has no upper bound.

Bridging

- *Network Constraint:* We define network constraint of a certain node i as given in Burt [4]:

$$\chi_i = \sum_{j \in \Gamma_i} \left(p_{ij} + \sum_{q \in \Gamma_i, q \neq j} p_{iq} p_{qj} \right)^2 \tag{3}$$

Here $p_{ij} = w_{ij} / \sum_i w_{ij}$ denotes the amount of direct attention that node i gives to node j. The sum $\sum_{q \in \Gamma_i, q \neq j} p_{iq} p_{qj}$ is the total amount of indirect attention that i gives to j through some intermediary q. Thus, as i's contacts become more connected, i's attention becomes more redundant and i's network constraint increases. This measure is minimized when none of i's neighbors are neighbors with each other, in which case it evaluates to $\frac{1}{k_i}$.

- *Ego Components:* Restricting attention solely to a node i's immediate neighborhood (i.e. its neighbors and all the edges between them), this measure η_i is a count of the number of connected components that remain when the focal node and its incident edges are removed. It is maximal if none of the node's neighbors have connections between them and minimal if there is a path connecting all of the node's neighbors that does not include the node itself.

As before, we study these features for both datasets for the family of networks $\{G(\tau_1), G(\tau_2), \cdots, G(\tau_K)\}$, where τ varies between 0.5 and 50. Figure 5(a–f) shows the values of six out of the seven of these features (average neighbor degree behaves almost indistinguishably from two-hop neighborhood, so is omitted for clarity), averaged over the population of non-isolated nodes.

For all of the measures of reach the values are necessarily monotonically decreasing because increasing τ can only delete edges, which means every node's degree can only go down. As can be seen in Figure 5, the average degree of the nodes decreases more sharply in the UNIVERSITY than it does in the ENRON dataset, though the change in the number of nodes reachable in two hops is very similar across datasets.

In contrast with reach, the measures of bridging do not necessarily change monotonically. Depending on which edges are deleted—those that connect nodes to different groups or those that tie groups together—both network constraint and the number of ego components can increase or decrease. Empirically, however, Figure 5 indicates that the overall trends are mostly monotonic: in general, network constraint increases, while number of ego components decreases (where in contrast, the UNIVERSITY dataset exhibits a slight increase for low values of τ). The explanation for these trends appears to be that for low τ the graph comprises a number of densely connected components, between which nodes can act as bridges. As we increase the threshold, however, the bridges between these clusters are preferentially severed, suggesting that bridging edges are not as strong as those within clusters, consistent with Granovetter's conjecture on the strength of weak ties [12].

[4]Note, in this definition of clustering coefficient, we ignore nodes whose clustering coefficients are undefined for a certain threshold τ.

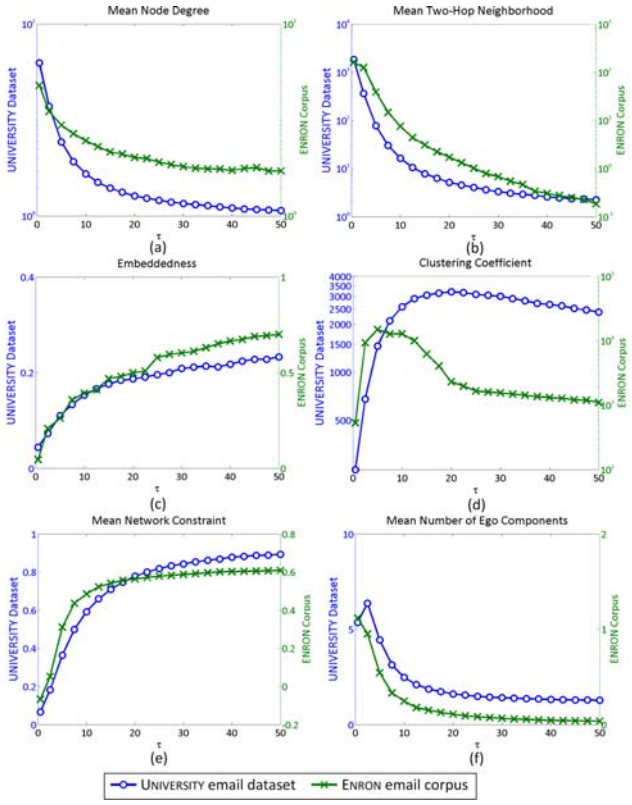

Figure 5: Changes in aggregated node-level features for the two email datasets.

The change in the measures of closure provide further support for this hypothesis. Embeddedness shows similar variation with τ to network constraint, indicating that as edges are deleted the neighborhoods of adjacent nodes have substantial overlap. The change in normalized clustering coefficient, however, is somewhat less intuitive and arises from two competing effects. On the one hand, if locally embedded "strong ties" arise out of a process of homophily and triadic closure [15], then one might suspect clustering coefficient would increase with τ, as weaker, less embedded ties are successively pruned away. On the other hand, if locally embedded edges arise out individuals sharing common "social foci" [11], then dense clusters in the network may be mostly made up of weak ties, in which case clustering would decrease with increasing τ. As Figure 5(d) indicates, we find evidence for both of these conjectures: at first, the normalized clustering coefficient C_i increases, consistent with Granovetter's [12] intuition that the weakest ties are bridges. Normalized clustering coefficient, however, peaks around $\tau = 15$ for the UNIVERSITY dataset and around $\tau = 5$ for the ENRON dataset, after which it decreases monotonically, suggesting that above a certain threshold most remaining ties are associated with dense clusters, and thus commensurate with Feld's social foci hypothesis [11], that pruning weaker ties reduces clustering.

4.3 Discussion

Table 1 summarizes the results of this section, displaying sample values of τ for the features discussed above for the UNIVERSITY dataset. To interpret this table in concrete terms, we note that at least three of these choices of threshold correspond closely to definitions of ties that have been invoked by previous authors: $\tau = 0.5$ [10], $\tau = 5$ [1], and $\tau = 15$ [31] respectively. That all three choices of τ have been made in prior work suggests that all three are defensible; yet Table 1 shows clearly that the networks we would infer from them would have vastly different properties, in terms of its density, connectivity, and clustering, among other properties. Average node degree, for example, varies between $k = 39.3$ and $k = 2.7$. How then should one choose the "correct" value of τ? Clearly one cannot do so on intuitive grounds alone; nor do Figures 2, 3, and 5 provide much insight—the features clearly change, but not in a way that suggests any obviously preferred value of τ. In the next section, therefore, we propose a method for choosing τ that depends explicitly on its relevance to some empirically observed pattern or a social process of interest.

Table 1: Summary of different node-level features as a function of τ, shown for the UNIVERSITY dataset.

	k_i	$k_i^{(2)}$	$\mathcal{E}_i$	c_i	C_i	χ_i	η_i
$\tau = 0.5$	39.3	1,845.2	0.01	0.48	244.7	0.1	5.4
$\tau = 5$	25.3	956.3	0.1	0.48	379.5	0.1	6.2
$\tau = 10$	5.9	77.8	0.1	0.36	1,461.2	0.4	4.4
$\tau = 15$	2.7	16.4	0.2	0.30	2,578.7	0.6	2.5
$\tau = 20$	2.0	7.8	0.2	0.26	3,046	0.7	1.9
$\tau = 50$	1.0	1.8	0.3	0.14	1,888.9	0.9	1.2

5. NETWORK-BASED PREDICTION

As noted in the introduction, if one were interested in, say, social influence, our proposed approach would be to infer the network that is most relevant to some empirically observed pattern of influence. To illustrate this approach, we study features present in our data: the distribution of individual attributes (gender and status), future communication between pairs, and membership in known communities. In all cases, we formalize our notion of relevance as a prediction task, where the desired value of τ is the one that maximizes the prediction of the observed property of interest for the network under consideration. For example, the homophily principle [23] implies that two individuals sharing the same status (e.g. undergraduate, graduate student, faculty, staff) are more likely to have an edge between them. Rather than choosing some definition of the threshold on some other grounds, therefore, we propose that the appropriate choice of threshold is the one for which the induced network provides the best predictor of an individual's status, given the statuses of his or her network neighbors (who, in turn, are defined by that choice of network).

5.1 Prediction Tasks

We now specify in more detail the four prediction tasks for which can we empirically evaluate the relevance of the networks: status; gender; future communication activity between pairs; and community membership. For all of the prediction tasks, with the exception of community detection, we utilize the node's attributes (e.g., affiliation, communication activity) and structural features (e.g., degree, normalized clustering coefficient) as well as the corresponding attributes/activities of its neighbors. In the following subsections, we discuss the prediction techniques for the different tasks in detail.

Node Status / Gender Prediction. Node status prediction deals with predicting whether an individual (1) in the UNIVERSITY email dataset is a student (undergraduate or graduate), faculty, staff, affiliate or other; or (2) in the ENRON dataset has a designation such as "Director", "Trader", "Manager", etc. within the company. Similarly, our second prediction task deals with predicting the gender of a particular node.

Let us represent the features for a node i in $G(\tau)$ as $\mathbf{f}_i^\tau = \{k_i^\tau, k_i^{(n),\tau}, k_i^{(2),\tau}, \mathcal{E}_i^\tau, C_i^\tau, \chi_i^\tau, \eta_i^\tau, \omega_1 \cdot |N_i(a_1)|, \omega_2 \cdot |N_i(a_2)|, \cdots, \omega_q \cdot |N_i(a_q)|\}$, where ω_j gives the mean edge weight of i with respect to the neighbors having attribute value j ($1 \leq j \leq q$) and $N_i(a_j)$ is the subset of i's neighbors whose attribute value is j. In our experiments we also consider an unweighted version, where we set all ω_j to 1. Based on the feature vectors $\mathbf{f}_i^\tau$ for all nodes i in the network $G(\tau)$, we construct the feature matrix, $\mathbf{F}^\tau \in \mathbb{R}^{d_1 \times |V|}$ and a vector of the actual homophily attributes (status / gender) of each node i in $G(\tau)$, given as, $\mathbf{A} \in \mathbb{R}^{1 \times |V|}$.

The prediction task over a network $G(\tau)$ can now be defined as a learning problem where $\mathbf{F}^\tau$ and $\mathbf{A}$ can be split into training set ($\mathbf{F}_\mathcal{R}^\tau$, $\mathbf{A}_\mathcal{R}$), $\sim 90\%$, and test set ($\mathbf{F}_\mathcal{S}^\tau$, $\mathbf{A}_\mathcal{S}$), $\sim 10\%$, and used in a multi-class Support Vector Machine (SVM) [3] framework (with a Gaussian RBF kernel) to predict the attributes of nodes. The prediction technique is described as follows. For every $G(\tau)$, we perform a k-fold cross-validation over the training set ($\mathbf{F}_\mathcal{R}^\tau$, $\mathbf{A}_\mathcal{R}$) to learn the optimal model parameters, including feature weights and the kernel width. These parameters are then used on the test set ($\mathbf{F}_\mathcal{S}^\tau$, $\mathbf{A}_\mathcal{S}$) to predict the node attributes, $\widehat{\mathbf{A}}_\mathcal{S}$.

Predicting Future Communication. The purpose of this prediction task is to determine the probability of future communication activity of a certain node (i.e. the number of emails sent). To predict activity at time t_{m+1}, we use a similar feature-based representation of a node i in the network $G(\tau)$, i.e. the structural features, and the mean weighted activities of its neighbors from time t_0 to t_m; but we augment the feature space by also using the node i's communication over the past, from t_0 to t_m. Hence the feature vector for prediction at time t_{m+1} can be written as $\mathbf{f}_{i,m+1}^\tau = \{k_{i,0:m}^\tau, k_{i,0:m}^{(n),\tau}, k_{i,0:m}^{(2),\tau}, \mathcal{E}_{i,0:m}^\tau, C_{i,0:m}^\tau, \chi_{i,0:m}^\tau, \eta_{i,0:m}^\tau, \sum_{j \in \Gamma_i} w_{ij} \cdot \alpha_{j,0:m}, \alpha_{i,0}, \alpha_{i,1}, \cdots, \alpha_{i,m}\}$, where $\alpha_{j,0:m}$ is the activity of node j (i.e. number of emails sent by node j) from time t_0 to t_m and $\alpha_{i,l}$ is the activity of node i at time t_l.

We fit a linear model of communication activity as a function of the node level features $\mathbf{F}_{0:m}^\tau$, i.e.

$$\mathcal{A}_m = \beta_{0:m}^\tau \cdot \mathbf{F}_{0:m}^\tau + \xi_{0:m}^\tau, \tag{4}$$

where $\beta_{0:m}^\tau$ are the regression coefficients and $\xi_{0:m}^\tau$ is additive noise. The best-fit coefficients $\beta_{0:m}^\tau$ are used along with the feature vector at t_{m+1}, to predict future node activity given as $\widehat{\mathcal{A}}_{m+1} \in \mathbb{R}^{1 \times |V|}$:

$$\widehat{\mathcal{A}}_{m+1} = \beta_{0:m}^\tau \cdot \mathbf{F}_{m+1}^\tau \tag{5}$$

For the prediction of future communication activity in the UNIVERSITY dataset, we divide the data over the span of two years into the six different semesters and regress over the first five semesters to predict the activity at the sixth semester. In the case of the ENRON email corpus, we divide the span of activity over four years (1998-2002) into time intervals of $t_i = 3$ months each. We incrementally train over the duration from t_0 to t_m, and predict the activity of each node at t_{m+1}, based on the technique discussed above.

Community Detection. In the final prediction task considered, we investigate the correlation between known community structure in the UNIVERSITY dataset with that inferred from network topology. For each threshold τ, we fit a stochastic block model [32] to the unweighted network $G(\tau)$ using variational Bayesian inference [14]. We then compare the resulting (soft) partition of nodes to the partition given by university affiliation of individuals to different schools, as reported in the node metadata. We quantify the correlation between the ground truth and inferred partitions using normalized mutual information as described in section 5.2.

This method for community detection assumes a model in which each node i belongs to one of Z latent groups (or "blocks"), indicated by z_i, with probability π_μ, $\mu \in 1, \ldots, Z$. The probability of an edge A_{ij} between nodes i and j depends only on the group assignments z_i and z_j: if the nodes are in the same group ($z_i = z_j$), an edge exists between them with probability θ_+; if they are in different groups ($z_i \neq z_j$), an edge exists between them with with probability θ_-. Given only the observed edges $e_{ij} \in E_s$ in the graph $G(\tau)$, distributions over the group assignments $p(z_i)$ are inferred via variational Bayesian inference.

Here we fix the number of groups to $Z = 5$, corresponding to the number of partitions given by university affiliation. Affiliations for singletons are assigned a uniform distribution over group assignments, i.e. $p(z_i = \mu) = 1/Z$. We note that, in contrast to the other prediction tasks at the node and edge level, this task involves the global structure of the network.

5.2 Results

For the tasks of predicting node status and gender, we quantify performance via classification accuracy, i.e. the fraction of nodes for which the predicted and actual values agree. Likewise, we quantify agreement between the (real-valued) predicted and actual future communication activity using percent error between the number of future emails predicted and observed.

In the case of community detection, however, the algorithm returns a probability distribution over group membership for each node. We quantify the agreement between this distribution and the actual group assignment (i.e. affiliation in the UNIVERSITY dataset) via normalized mutual information (NMI). This standard measure for evaluating performance of community detection algorithms [7] is given by the mutual information of the joint distribution over actual and predicted assignments, normalized by the entropy of the marginal distribution over actual distribution. Confined to lie between 0 (minimum agreement) and 1 (maximum agreement), NMI is similar to the number of correctly classified nodes, but penalizes misclassified nodes more heavily.

UNIVERSITY. Prediction results for the UNIVERSITY email dataset for the four tasks described above are shown in Figure 6. First we observe that the predictions using weighted features perform better than the corresponding unweighted version; that is, the frequencies of communication (the edge weights) are informative for all of the prediction tasks, even in the thresholded graphs where infrequent communication is discarded. Second, we observe that in all cases the ac-

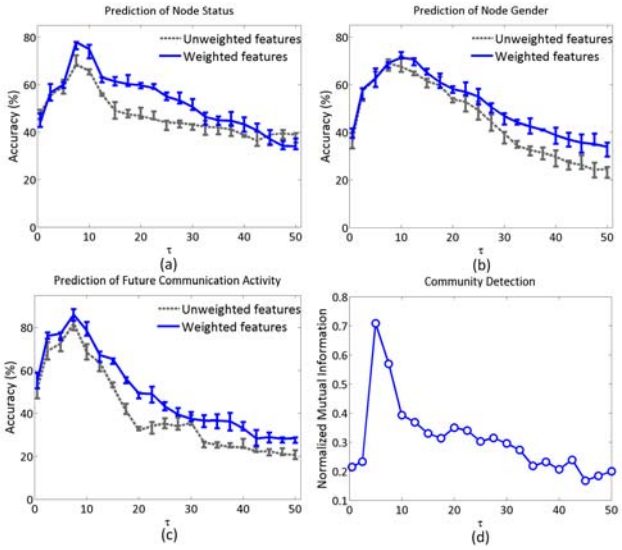

Figure 6: Mean prediction accuracies (over all nodes) for three different prediction tasks on the UNIVERSITY email dataset—(a) node status (e.g. undergraduate, graduate, faculty etc.) and (b) gender prediction, (c) predicting future communication activity and NMI in detection of community structure (i.e. affiliation to schools). Two cases per task (a-c) are shown, one with unweighted features and the other with weighted features. The error bars in the plots (a-b) correspond to the k-fold cross validation performed in the prediction process. The error bars in plot (c) correspond to the deviation in prediction error across all the users at each τ.

curacy peaks at a non-trivial value of τ, (i.e., at a value greater than the minimum τ at which no threshold condition is applied). This result suggests that there is some optimal balance to be struck between removing noisy edges and retaining sufficient information about a node's neighborhood when making predictions. We note also that the gain associated with discarded edges is nontrivial, corresponding to as much as 30% performance gain over the naïve strategy of retaining all the edges. Surprisingly, the same rule seems to apply equally to weighted and unweighted networks; that is, even when weights are retained on the edges, one still gains a large boost by discarding the lowest-weight edges. Third, although the prediction accuracy peaks at different numerical values of τ (node status and future communication activity peak at $\tau = 7.5$, whereas accuracy for gender peaks at $\tau = 10$, and community structure peaks at $\tau = 5$), the peaks all fall within a relatively small range.

ENRON. With respect to the ENRON email corpus, differences in the available data restrict us to just two of the above four tasks: (1) prediction of node status (Figure 7); and (2) prediction of future communication activity of the nodes (Figure 8(a–b)). The results for node status are similar to the UNIVERSITY dataset: first, the maximum accuracy appears at a non-trivial value of τ both for the unweighted features ($\tau = 7.5$) and for the weighted features ($\tau = 10$); and second, the weighted features improve the prediction accuracies. For future communication activity we find the accuracy in the prediction of future activity peaks in roughly the same range as for the UNIVERSITY data ($\tau = 2.5$ to $\tau = 7.5$

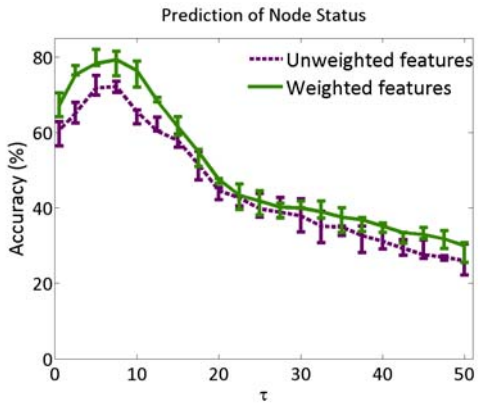

Figure 7: Mean accuracies in prediction of node status (i.e. designation at the company) for the ENRON email corpus; both unweighted and weighted features are shown.

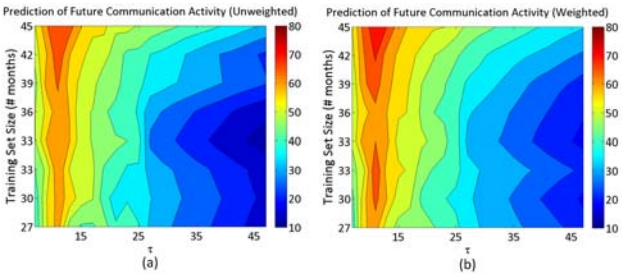

Figure 8: Mean accuracies (over all nodes) in prediction of future communication activity for ENRON dataset, (a) unweighted case, and (b) weighted case.

for the unweighted and the weighted cases respectively), as shown in Figure 8(a–b). We also observe that training over extended durations improves the accuracy across all thresholds. Once again, however, the range of τ where the accuracy peaks seems to be reasonably consistent across training set sizes.

5.3 Discussion

To summarize our findings, the threshold values that are most predictive for the tasks we have considered are not obvious, either on the basis of intuition (how would one choose $\tau = 10$ versus $\tau = 15$?) or from the descriptive statistics present in section 4.2. Nevertheless, the choice of τ has a substantial impact: networks corresponding to optimal values of τ perform as much as 30% better than naïve choices (e.g. defining an edge whenever two individuals have exchanged at least one email over the entire observation period). Finally, we observe that as expected, different values of τ optimize the prediction task for different empirical patterns; however, on this point, we also note an intriguing and unexpected secondary finding—that although different, the optimal values of τ seem to fall in a surprisingly narrow range between $\tau = 5$ and $\tau = 10$. A partial explanation for this result may be that the particular prediction tasks we have examined all involve predicting a certain attribute of an individual (status, gender, measure of communication and community membership), given (a) her node features; and (b) corresponding attribute values of her neighbors. Pos-

sibly, therefore, all four tasks are reflections of the same general principle of homophily [23]—that similar nodes are more likely to be connected by social ties than dissimilar nodes—in which case the small range of optimal τ may not be as surprising as it initially appears.

That the range of optimal τ is also similar across datasets is, however, puzzling. The two datasets were collected several years apart in very different organizations, and involved very different people who were presumably communicating about very different topics. Therefore there is no *a priori* reason to suspect that the same, or even a remotely similar definition of an edge, as reflected by the intensity of reciprocated communication, should satisfy our prediction tasks. Possibly the observed correspondence is simply a coincidence, and will not generalize to other cases. If such a finding does hold, however, it holds out the promise that networks inferred on the basis of one empirically observed pattern are also relevant to other patterns that have not been observed; that, for example, a network inferred on the basis of gender association could be used to predict the diffusion of social influence, or that the definition of a tie relevant to diffusion in one network for which diffusion data may be available could be applied to another network for which it isn't. Clearly claims of this nature are speculative; nevertheless, they suggest interesting directions for future research efforts.

6. CONCLUSIONS

Returning to our original motivation, network analysis of communication data takes as input some set of observations and infers from these data a set of relations to which social and psychological meaning is attached. We argue here that this inference procedure, which heretofore has been defined in a largely separate and often ad-hoc manner, should be as much a part of the analysis as the measurement of structural features. In this paper, we have addressed a narrow version of this general problem; that is, how to determine an optimal threshold condition for edges so as to predict particular node attributes (e.g. gender, status) or behavior. Starting with a baseline network of communication based on email exchanges in two different datasets, we constructed a family of networks by consistently removing edges with weights below a series of specified thresholds. We then studied a range of commonly used descriptive statistics, finding dramatic differences in network- and node-level features depending on the choice of threshold. Finally, we introduced a method for selecting among all these possible networks on the basis of a series of prediction tasks. The prediction accuracies peak in a non-obvious—yet relatively narrow—threshold range across both datasets. We conclude with a discussion of several limitations of the work presented above, as well as possible directions for future studies.

First, and most importantly, the general problem of "data relevance" is considerably more difficult than we have allowed for within our narrow framework. Obviously, the outcome of any network inference procedure seems likely to be influenced by the manner in which one generates the family of possible networks to begin with; thus a more general approach than the one we have adopted here might be advisable. For example, although we have allowed ties to be weighted, these weights refer only to the average frequency of communication, and so capture "strength" at best incompletely; and our use of the geometric mean of email exchanges as the basis on which to apply the threshold condition, although reasonable, is clearly not the only sensible approach. We have also not allowed ties to be directed, or "multiplex", or to have time varying properties [20]; yet all are arguably important features of real-world social relations. At a minimum, therefore, it would be desirable to establish methods for inferring weighted, directed, multiplex, and time-varying networks from observational data.

A second, related limitation in this work is that we have used a simple binary threshold function to generate candidate network structures. One can imagine a more sophisticated means of transforming edge weights via arbitrary functions and learning function parameters while simulatenously optimizing for predictive performance. Finally, it is not clear why we find such consistency of the optimal choice of threshold across different prediction tasks in our experiments, and especially across different networks. Further investigation across a wider range of communication data and prediction tasks may provide insight into whether there is any special significance to the range of threshold values observed here.

In closing, we note that although the focus in this paper has been on networks inferred from communication data, social networks may be constructed from other kinds of observable data too, such as the joint participation of actors in scientific collaborations, social events, informal organizations, corporate boards or even movies. As with communication data, researchers typically infer the presence of social networks from data of this type by choosing some ad-hoc threshold condition: for example, "i and j will be considered connected if they share at least one group." And, as with communication data, one may ask how relevant a particular shared affiliation is: just because two directors sit on the same board does not necessarily indicate how often they talk or how much they trust each other; nor is it clear what one should infer from the existence of a co-authored paper, a "friend" nomination on Facebook or any similar observation. Even for network data generated by survey tools, an analogous problem arises: survey respondents presumably apply some criteria for whom they report as a contact; yet because these criteria are generally not known to the researcher (or even necessarily to the respondents themselves), it can be difficult to interpret significance of the reported ties [2, 9].

The fundamental issue raised in this paper—that is, how to infer relations of social and psychological relevance from observable events and event participant reports—is therefore an extremely general one that applies well beyond the scope of communication data, impacting a much wider range of network problems than we have considered here. Extending the methods introduced here to apply to different classes of interactional data, possibly in combination (e.g. email, mobile phone calls, and affiliation data), and to more general classes of network-related phenomena (e.g. the dynamics of collective social behavior) therefore ought to provide ample opportunities for future work.

7. ACKNOWLEDGEMENT

We thank Siddharth Suri for assistance with computing some of the node-level features.

8. REFERENCES

[1] Lada Adamic and Eytan Adar. How to search a social network. *Social Networks*, 27(3):187–203, July 2005.

[2] Peter Bearman and Paolo Parigi. Cloning headless frogs and other important matters: Conversation topics and network structure. *Social Forces*, 83(2):535–557, December 2004.

[3] Christopher J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.*, 2(2):121–167, 1998.

[4] Ronald S. Burt. Structural holes and good ideas. *The American Journal of Sociology*, 110(2):349–399, 2004.

[5] Aaron Clauset and Nathan Eagle. Persistence and periodicity in a dynamic proximity network. In *DIMACS Workshop on Computational Methods for Dynamic Interaction Networks*, 2007.

[6] Corinna Cortes, Daryl Pregibon, and Chris Volinsky. Computational methods for dynamic graphs. *Journal of Computational and Graphical Statistics*, 12(4):950–970, December 2003.

[7] Leon Danon, Albert Díaz-Guilera, Jordi Duch, and Alex Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, page P09008, 2005.

[8] Jana Diesner, Terrill L. Frantz, and Kathleen M. Carley. Communication networks from the enron email corpus "it's always about the people. enron is no different". *Comput. Math. Organ. Theory*, 11(3):201–228, 2005.

[9] N. Eagle, A. Pentland, and D. Lazer. Inferring social network structure using mobile phone data. *PNAS*, 106(36):15274–15278, 2009.

[10] Jean-Pierre Eckmann, Elisha Moses, and Danilo Sergi. Entropy of dialogues creates coherent structures in e-mail traffic. *Proceedings of the National Academy of Sciences of the United States of America*, 101(40):14333–14337, October 2004.

[11] Scott L. Feld. The focused organization of social ties. *The American Journal of Sociology*, 86(5):1015–1035, 1981.

[12] M. S. Granovetter. The strength of weak ties. *The American Journal of Sociology*, 78(6):1360–1380, 1973.

[13] M. Hammer. Social access and the clustering of personal connections. *Social Networks*, 2(4):305–325, 1980.

[14] Jake M. Hofman and Chris H. Wiggins. A bayesian approach to network modularity. *Phys Rev Lett.*, 100(5), June 2008.

[15] Gueorgi Kossinets and Duncan J. Watts. Empirical analysis of an evolving social network. *Science*, 311(5757):88–90, January 2006.

[16] Ravi Kumar, Jasmine Novak, and Andrew Tomkins. Structure and evolution of online social networks. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 611–617, New York, NY, USA, 2006. ACM.

[17] Ravi Kumar, Andrew Tomkins, and Erik Vee. Connectivity structure of bipartite graphs via the knc-plot. In *WSDM '08: Proceedings of the international conference on Web search and web data mining*, pages 129–138, New York, NY, USA, 2008. ACM.

[18] Jure Leskovec and Eric Horvitz. Planetary-scale views on a large instant-messaging network. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 915–924, New York, NY, USA, 2008. ACM.

[19] David Liben-Nowell, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11623–11628, August 2005.

[20] R. Dean Malmgren, Jake M. Hofman, Luis A.N. Amaral, and Duncan J. Watts. Characterizing individual communication patterns. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 607–616, New York, NY, USA, 2009. ACM.

[21] Peter V. Marsden. Network data and measurement. *Annual Review of Sociology*, 16:435–463, 1990.

[22] Winter Mason and Sid Suri. Predicting individual success in social networks. In preparation.

[23] Miller Mcpherson, Lynn S. Lovin, and James M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

[24] Theodore Mead Newcomb. *The acquaintance process*. Holt, Rinehart and Winston, New York, NY, 1961.

[25] M. E. Newman. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Phys Rev E Stat Nonlin Soft Matter Phys*, 64(1 Pt 2), July 2001.

[26] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113+, Feb 2004.

[27] J.-P. Onella, J. Saramaki, J. Hyvonen, M. Argollo de Menezes, K. Kaski, A.-L. Barabási, and J. Kertész. Analysis of a large-scale weighted network of one-to-one human communication. *New Journal of Physics*, 9(6):179–204, February 2007.

[28] Michael F. Schwartz and David C. M. Wood. Discovering shared interests using graph analysis. *Commun. ACM*, 36(8):78–89, 1993.

[29] J. Shetty and J. Adibi. Enron email dataset. Technical report, USC Information Sciences Institute, 2004.

[30] Eric Sun, Itamar Rosenn, Cameron Marlow, and Thomas Lento. Gesundheit! modeling contagion through facebook news feed. In *ICWSM '09: Proceedings of the Third International Conference on Weblogs and Social Media*, San Jose, CA, May 2009. AAAI Press.

[31] Joshua R. Tyler, Dennis M. Wilkinson, and Bernardo A. Huberman. Email as spectroscopy: Automated discovery of community structure within organizations, Mar 2003.

[32] Y Wang and G Wong. Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 1987.

[33] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, June 1998.