Who Thinks Who Knows Who? Socio-cognitive Analysis of Email Networks
— Source link ↗

Nishith Pathak, Sandeep Mane, Jaideep Srivastava

Institutions: University of Minnesota

Related papers:

- Topic and role discovery in social networks with experiments on enron and academic email

- Measuring the Importance of Users in a Social Network Based on Email Communication Patterns

- Dynamic organizations. How to measure evolution and change in organizations by analyzing email communication networks

- Automated social hierarchy detection through email network analysis

- Detecting Communities in Organizational Social Network Based on E-mail Communication

Technical Report

Department of Computer Science
and Engineering
University of Minnesota
4-192 EECS Building
200 Union Street SE
Minneapolis, MN 55455-0159 USA

TR 06-023

Who Thinks Who Knows Who? Socio-cognitive Analysis of Email Networks

Nishith Pathak, Sandeep Mane, and Jaideep Srivastava

July 21, 2006

Who Thinks Who Knows Who? Socio-cognitive Analysis of Email Networks

Nishith Pathak
Dept. of Computer Science,
University of Minnesota,
Minneapolis, USA
npathak@cs.umn.edu

Sandeep Mane
Dept. of Computer Science,
University of Minnesota,
Minneapolis, USA
smane@cs.umn.edu

Jaideep Srivastava
Dept. of Computer Science,
University of Minnesota,
Minneapolis, USA
srivasta@cs.umn.edu

Abstract

Interpersonal interaction plays an important role in organizational dynamics, and understanding these interaction networks is a key issue for any organization, since these can be tapped to facilitate various organizational processes. However, the approaches of collecting data about them using surveys/interviews are fraught with problems of scalability, logistics and reporting biases, especially since such surveys may be perceived to be intrusive. Widespread use of computer networks for organizational communication provides a unique opportunity to overcome these difficulties and automatically map the organizational networks with a high degree of detail and accuracy. This paper describes an effective and scalable approach for modeling organizational networks by tapping into an organization's email communication. The approach models communication between actors as non-stationary Bernoulli trials and Bayesian inference is used for estimating model parameters over time. This approach is useful for socio-cognitive analysis (who knows who knows who) of organizational communication networks. Using this approach, novel measures for analysis of (i) closeness between actors' perceptions about such organizational networks (agreement), (ii) divergence of an actor's perceptions about organizational network from reality (misperception) are explained. Using the Enron email data, we show that these techniques provide sociologists with a new tool to understand organizational networks.

Keywords

Socio-cognitive network, email communication network, belief divergence, Enron email data

1. Introduction

Organization dynamics plays an important role in the functioning of an enterprise. Understanding the dynamics of organizational processes empowers managers and enables them to effectively manage an enterprise's resources. Informal social and socio-cognitive networks in an organization play an important role in such processes and significant effort has been made to study them. However, most research has relied on data collected manually (e.g. using surveys and observing communication between individuals in meetings) and hence is subject to a variety of noise (e.g. biased opinions). The emergence of computer networks has enabled new methods of communication, e.g. e-mail and instant messaging, between individuals in an organization, providing a unique opportunity to study social networks in a detailed and unbiased manner by collecting such data. In addition, the ease of use and small costs of electronic communication have enabled geographically dispersed communication between individuals, leading to the creation of geographically-unrestricted social networks. The current social network analysis models like latent space model [3] and p* model [16] suffer from computational efficiency and scalability issues. Thus, there exists a need for new scalable, efficient computational techniques to study such organizational networks.

In email communication, an actor observes only those emails which are addressed to that actor, i.e., the actor is in either To, Cc or Bcc fields of those emails. From a socio-cognitive perspective, different actors have different perceptions about the email communication network. Thus, email communication motivates as well as enables the study of socio-cognitive networks in an organization, i.e., understanding *who knows who knows who* in a social network. No prior research exists for

such an analysis of email communication networks. Thus, this paper proposes a novel model for representing the communication between actors in a social network, using non-stationary Bernoulli probabilities. Such probabilities are derived based on the observed email communication. A Markov time window based approach is described to handle the non-stationary nature of Bernoulli probabilities. As against a more sophisticated model, the proposed simple model provides a scalable approach, in addition to being less affected by data sparseness as well as providing reasonable performance on real data. Data sparseness exists in email communication networks since an actor (on the average) communicates with only a few other actors (limited social bandwidth observed in social networks [5]). Thus, such a model can be used for both socio-centric as well as ego-centric analysis of a social network.

Using the proposed non-stationary Bernoulli model, each actor's perceptions about the total email communication is modeled using the respective subset of emails observed by that actor. To quantify the difference in perceptions of actors, a novel measure, *a-closeness*, which uses KL-divergence, is proposed. This measure is similar to the *perceptual congruence* measure in social science literature [1]. In addition to the actors' perceptions, the email server observes all the email communication, which is also represented using the proposed model and thus forms the baseline for the real communication network. The divergence of an actor's perceptions from the real communication network is quantified using a novel measure, called *r-closeness*. No counterpart for such a measure exists in social science research due to lack of availability of such real data. Experimental results using the proposed model and measures on real-world Enron email dataset show interesting results and illustrate that these techniques provide a powerful computational tool for social network analysis.

The rest of the paper is organized as follows: Section 2 provides background on the problem addressed in this paper. Section 3 describes the non-stationary Bernoulli model for constructing a social network from email data. Section 4 explains two different socio-cognitive analyses of an email communication network using the proposed model and then describes new measures for such analyses. Section 5 presents experimental results of socio-cognitive analyses on the Enron email dataset. Section 6 summarizes the paper, explains the applications of this research and discusses future research directions.

2. Background

The first sub-section explains basic terminology on social and socio-cognitive network analysis; the next sub-section analyzes the impact of email communication on social network analysis; and the final sub-section describes the problem addressed in this research.

2.1 Social and Socio-Cognitive Network Analysis

Social network analysis is an active field of study in sociology as well as anthropology. A *social network* is a social structure of individuals (people) called *actors*, related (directly or indirectly) to each other through a common relation of interest [15]. A social network plays an important role in the dissemination of ideas, information or influences among the individuals. However, in any social network, it is not possible for everyone to be connected to everyone else, nor is it desirable [1]. Thus, the main motivation of social network analysis is to study "who knows who" in a social network. There are two types of social networks analysis: (i) *Socio-centric (whole) network analysis*, where the interactions between the entire well-defined set of people are studied; and (ii) *Egocentric (personal) network analysis*, where one studies the interactions between an actor (called "ego") and only those actors related (directly or indirectly) to the ego.

Substantial research has illustrated the importance of such analyses in organizations. In an organization, informal networks are formed by relationships between employees across functions and/or divisions in order to accomplish tasks quickly [8]. Such informal networks can cut through formal reporting procedures to jump start stalled initiatives and meet extraordinary deadlines. Informal networks can just as easily sabotage companies' best laid plans by blocking communication and fomenting opposition to change unless managers know how to identify and direct them. Social network analysis enables the understanding of which actors are perceived as "friends" or "adversaries" by an actor, and which actors are aware of the presence of which other actors.

Taking this a step further is socio-cognitive network analysis, which analyzes "who knows who knows who" in the social network. This analysis is useful as it affects the perceptions of an actor about other actors, and hence the behavior of actors towards other actors. This is of prime importance to a manager in an

organization. The beliefs for each actor are translated into a weighted digraph corresponding to the social network that exists from that actor's perspective. Using these digraphs, one can determine who thinks who is influential in the organization. This information is highly valuable for a manager to understand the existing informal network in the organization. Traditionally, researchers have relied on actor interviews and surveys for socio-cognitive data. Hence, even though techniques from simple graph-based to sophisticated multilevel models ([1], [13] and [14]) exist for analyzing these responses, there has been no research on extracting interesting socio-cognitive patterns from large observable email communication logs.

2.2. Organizational Email Communication

One of the main reasons for computer networks (and Internet) to come into existence is to foster collaborative work between geographically dispersed researchers. These computer networks have now turned into an infrastructure that supports social networks; connecting people, organizations as well as knowledge [13]. The widespread use of internet and the growing online community of users have enabled the formation of social networks based on different relations of interest. For example, Usenet – a widely used online newsgroup – had more than 80,000 topic-oriented discussion groups (or social networks) in 2000. These discussion groups allow individuals to form geographically dispersed, loosely-bound, social networks. On the other hand, computer networks also facilitate an actor to participate in different social networks (communities), thus enabling the actor to know many more other actors and increase his/her social capital.

In an organization, an email server logs all emails exchanged between employees, thus capturing an unbiased view of all email communication between them. In an organization, it is possible to map the online actor (e.g. email address) to a real-world actor (e.g. employee), and analysis of these interactions has the potential of providing unbiased measures about social relationships between real-world actors. However, to analyze such gigabytes of data about emails exchanged between employees (considering a medium scale organization) requires new scalable, computational techniques. With the availability of the Enron email corpus, there has been a growing interest in applying computational techniques to analyze email-based social networks. Initial research on analysis of
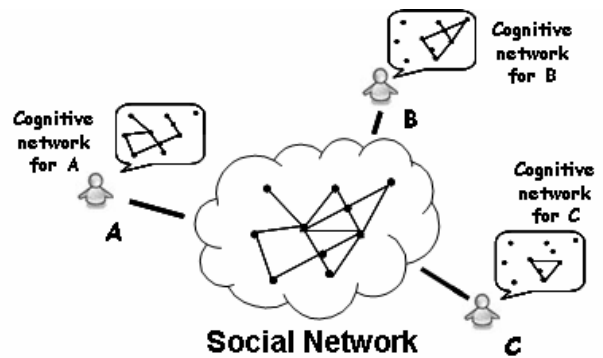
Figure 1. Actor's perceptions of a social network (Socio-cognitive network).

such email data has concentrated mainly on applying traditional social network techniques and/or graph-based measures [3] [6], [12].

2.3 Problem description

This research takes a step further by providing novel computational techniques for socio-cognitive analysis of email data. Consider an e-mail sent by actor A to B, with Cc to C and Bcc to D. The analysis of the header reveals the following: B and C know that A and B communicated, and that all three (A, B and C) know about this communication. However, neither B nor C know that D was also sent this e-mail. Actors A and D know D received the email, and both also know that B and C do not know that D received that e-mail. This illustrates that an e-mail can create different beliefs about communication among different actors, depending on whether and how they are included in the email recipient list. Based on the observed emails, an actor forms his/her beliefs of probabilities of communication between different actors. An email communication network is defined using the actors as the nodes and edges between actors representing email communication between them. Each actor in the network maintains his/her communication network based on the emails observed by him/her. Such a communication network defines the actor's beliefs regarding the social network and the set of such networks for all actors is defined as a *socio-cognitive network* in this paper. (see Figure 1).

This paper thus addresses the problem of representing, using an intuitive, simple yet scalable model, the email communication networks in a socio-cognitive network and then illustrates the use of that model for novel, interesting socio-cognitive network analysis.

3. An Approach for Socio-Cognitive Network Modeling

This section presents a novel approach for automated construction of a communication network in a socio-cognitive network by analyzing of an organization's email communication.

3.1 Basic concepts

As explained in previous section, an actor participating in the email communication network maintains beliefs regarding the email communications in that network, i.e. beliefs about who communicates with whom, based on the emails that the actor observes. Basic concepts, which enable modeling of such communication probabilities, are explained here. Consider an email communication network consisting of N actors denoted by the set, $\{A_i \mid 1 \leq i \leq N\}$. Let $P_i = Pr(Sender = A_i)$ denote the probability that an email in the communication network is sent by the actor A_i. Thus,

$$P_i = \frac{Number\ of\ emails\ sent\ by\ A_i}{Total\ number\ of\ emails\ sent\ in\ the\ network}$$

Since each email has a unique (single) sender, the events corresponding to an email being sent by different actors are mutually exclusive. Hence, the following condition must always hold -

$$P_i \geq 0, \forall\ A_i\ and\ \sum_{\forall\ A_i} P_i = 1 \quad \dots \quad (1)$$

Let $P_{j|i} = Pr(A_j \in Recipients \mid Sender = A_i)$ denote the probability of A_j being a recipient of an email, given that A_i has sent that email, i.e.,

$$P_{j|i} = \frac{Number\ of\ emails\ sent\ by\ A_i\ and\ received\ by\ A_j}{Total\ number\ of\ emails\ sent\ by\ A_i}$$

Thus, P(i,j), the probability that an actor A_i sends an email to an actor A_j, is defined as,

$$P_i \times P_{j|i} = P(i,j) = \frac{Number\ of\ emails\ sent\ by\ A_i\ and\ received\ by\ A_j}{Total\ number\ of\ emails\ sent\ in\ the\ network}$$

This represents the "strength" of the actor A_i's communication with actor A_j. The events corresponding to different actors being recipients of an email are not mutually exclusive since an email may have multiple recipients. Thus the marginal probabilities of different actors being recipients, are dependent and so do not add up to one. Another approach is to consider it as point to point communication, i.e. an email with multiple recipients assumed as multiple emails with one recipient for each such email. But in that case, the marginal probabilities are forced to be independent, which may be a strong assumption that may not hold in most cases. Hence, in order to preserve the dependencies between marginals, the probabilities $P(i,j)$ of an actor A_i being a sender and another actor A_j being a recipient are not mutually exclusive events for different pairs of senders and recipients.

3.2 Modeling Communication Network in a Socio-cognitive Network

The event of an actor A_i being the sender and A_j being the recipient of an email is mutually exclusive to its complement, i.e. the event where for an email either A_i is not the sender or A_j is not a recipient or both. The probabilities of these two events are represented as $P(i,j)$ and $1-P(i,j)$ respectively. We define a Bernoulli distribution over the two events corresponding to email communication between actors A$_i$ and A$_j$, i.e.,

$$L(i,j) = [P(i,j),\ 1-P(i,j)].$$

where $P(i,j)$ is the parameter of the Bernoulli distribution $L(i,j)$. For the communication network perceived by an actor, there will $N(N-1)$ such distributions, one for every ordered pair of actors (A_i, A_j), $A_i \neq A_j$. Every email exchanged in the network is a Bernoulli trial, i.e. either a given email is sent by an actor and the other actor is one of the email's recipient(s) or it's complement (see Figure 2).

Figure 2. Communication between actors expressed as Bernoulli distributions.

Based on such observations, every actor maintains a distribution over all possible probabilities $P(x,y)$ for a given ordered pair (A_x, A_y), i.e. a distribution over all possible values for the parameter $P(x,y)$ of each Bernoulli distribution $L(x,y)$.
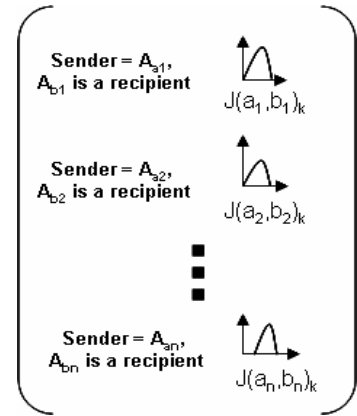
Figure 3. Belief State of actor A$_k$, with beliefs as Beta distributions.

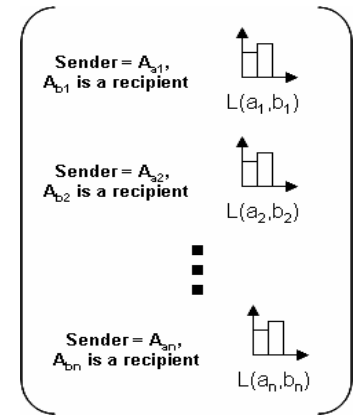For maintaining this distribution over all possible parameters of a Bernoulli distribution, a Beta

---
Algorithm 1 Belief update for an actor

Input:
- A_x is an actor
- E_x is the set of all emails observed by A_x
- B_x is the belief state of A_x

Output:
- Updated belief state B_x for actor A_x

Pseudo-code:
1. **For each** $e \in E_x$ **do**
2.   **For** the sender A_s and each recipient A_r of e **do**
3. $B_x[s][r]++.$
4.   **End for**
5. $A[x].number\_of\_observed\_emails++$
6. **End for**
7. **Return** B_x

---

distribution is used. As the Beta distribution is the conjugate prior for the Bernoulli distribution, a Bayesian update on the parameters of a Beta distribution is used for maintaining actors' "beliefs".

DEFINITION 1 (**Belief State**): A *belief state* of an actor is defined as a set of *N(N-1)* Beta distributions, where each Beta distribution *J(i,j)* is defined over the corresponding Bernoulli distribution *L(i,j) representing email communication between actors A_i and A_j.*

Thus, the belief state B_k for a given actor A_k is given as,

$B_k = \{ J(i,j)_k \mid \forall$ ordered (A_i, A_j) such that $A_i \neq A_j \}$

where $J(i,j)_k$, is a Beta distribution over the parameter of $L(i,j)$ and is defined as A_k's *belief* about probability of email communication from A_i (sender) to A_j (recipient) (see Figure 3). Each such Beta distribution $J(i,j)_k$ in belief state B_k *of an actor* A_k has two parameters, $\alpha(i,j)_k$ and $\beta(i,j)_k$. Based on the communication A_k observes, A_k updates the parameters for all $J(i,j)_k$ in B_k. We associate the parameter $\alpha(i,j)_k$ with the number of successes, i.e. the number of emails, observed by A_k, that have been sent by A_i to A_j, and parameter $\beta(i,j)_k$. with failures, i.e. number of emails observed by A_k for which either A_i is not the sender or A_j is not the recipient or both. Thus, for each email observed by A_k to be sent from A_i to A_j, the corresponding $\alpha(i,j)_k$ parameter is incremented whereas for each failure, the parameter $\beta(i,j)_k$. is incremented.

Algorithm 1 provides the methodology for updating an actor's belief state based on the set of emails observed in a particular time window. An actor A_k ($1 \leq k \leq N$) starts with an initial belief state B_k, with parameters for all distributions having default prior values. As actors observe email communication, the actor updates his/her

belief state. To maintain the belief state of an actor A_k, $N(N-1)$ counters, corresponding to $\alpha(i,j)_k$ of each Beta distribution, are maintained. In addition, a counter is maintained for the total number of emails observed by each actor. The $\beta(i,j)_k$ parameter is computed by subtracting the corresponding $\alpha(i,j)_k$ counter from the total number of emails. Counters for each of the $\alpha(i,j)_k$ parameters are initialized with their corresponding priors and the "total number of emails" counter starts with an initial value of $(\alpha(i,j)_k + \beta(i,j)_k)$, as will be explained later.

3.3 Non-stationarity and Time Windows

As more emails are exchanged over time, the communication probabilities may change. Thus, as A_k observes more email communication in the network over time, he/she updates his/her belief state using Bayesian inference. Since the underlying Bernoulli probabilities are non-stationary in nature, we choose to capture this dynamic nature of the communication probabilities using a time window based approach. At the beginning of each time window, the parameters for all Beta distributions in a given actor's belief state are scaled down by a parameter λ ($0 \leq \lambda \leq 1$). For each email, the corresponding $\alpha(i,j)_k$ and $\beta(i,j)_k$ are updated, and thus an actor's belief state of all communication relations are maintained. At the beginning of the next time window, the posterior parameters from previous time window are scaled down and are used as priors for the next time window. The model parameter λ regulates how much of history is "remembered" by an actor, i.e. the degree of the Markovian chain. Higher the value of λ, more is the importance given to history. If $\lambda = 1$, each observation is given the same importance and all the history is remembered. If $\lambda = 0$, then the previous probability estimates are completely washed out at the beginning of each time window and new priors ($\alpha(i,j)_k > 0$ and $\beta(i,j)_k > 0$) are chosen. Thus, there is an exponential decay of history, where the rate of decay is controlled by the parameter λ.

Another important parameter of interest in this model is the length of the time window. This problem is similar to the classical problem of segmenting time series in temporal data analysis, since the vector of communication probabilities is analogous to a time series dataset and each segment is analogous to a time window. A Bayesian belief update for each actor occurs at the end each time window. The choice of length of time window affects the number of emails observed in the time window and hence the interpretation of results. This paper assumes that the

length of the time window is a user-specified parameter, but it provides sufficient number of emails within each time window. Other approaches such as varying time window length and/or updating different actors' belief states at end of different time windows can also be adopted, but they are left as open problems for future research.

To model the temporally varying nature of beliefs, we denote the belief state of an actor at time t as $B_{k,t}$.

DEFINITION 2 (**Belief State at time t**): Formally, the belief state for the given actor A_k at the given time t, is defined as,

$$B_{k,t} = \{ J(i,j)_{k,t} \mid \forall\ (A_i, A_j)\ \text{such that}\ A_i \neq A_j \}$$

where, $J(i, j)_{k,t}$ is the Beta distribution for an ordered pair of actors (A_i, A_j), maintained by the actor A_k at time t.

The belief state of a given actor at time t reflects what the actor believes to be the probabilities of the possible strengths of different actor communications in the network at time t. A socio-cognitive network at a given time is the set of belief states of all actors at that time.

3.4 Priors Selection

This sub-section addresses the issue of selecting priors for the parameters of each of the distributions $J(x,y)_k$ in a given belief state B_k. The priors are chosen such that

$$\alpha(i,j) = \delta_i \varepsilon_{j|i}\ \text{and}\ \beta(i,j) = 1 - \alpha(i,j),$$

where δ_i is the prior probability for A_i being the sender of an email and $\varepsilon_{j|i}$ is the prior probability for A_j being a recipient given that A_i has sent the email. Each probability in an actor's belief state is expressed as a fraction of the communication in the network. Hence, the sum of the expected probabilities for all communications must always be greater than or equal to *1* and less than or equal to *(N-1)* (see appendix A). Since, the events of different actors being senders is mutually exclusive, the following condition must hold, $\sum_i \delta_i = 1$. Thus, a simple solution chosen is to use uniform priors, where each $\delta_i = 1/N$, N being the number of actors. For $\varepsilon_{j|i}$, a closed world assumption is made, i.e. since an actor has not observed any communication in the prior state, the initial prior probability for the event of A_j being a recipient given that the email has been sent by A_i, is some small ε_+. For example, assigning $\varepsilon_{j|i} = 0.01$ gives the following simple solution for priors is $\alpha(i,j)_k = 0.01/N$ and $\beta(i,j)_k = 1 - (0.01/N)$. An advantage of small initial values for both $\alpha(i,j)_k$ and $\beta(i,j)_k$ is the low influence of the priors in the updated belief states, because as the number of observations (emails) is usually relatively large compared to the priors, it results in "washing out" of priors.

3.5 Time Complexity Analysis

This sub-section analyzes the computational complexity for belief update an actor needs to perform on observing an email. Consider an email sent or received by an actor in the communication network. Let the number of recipients in the email be n. The actor needs to update parameters for all sender-recipient pairs. For this purpose the actor increments the "total mails observed counter" and the α parameter counter for each sender-recipient pair observed by the actor. This requires a maximum of $(n+1)$ updates. Thus, the complexity for belief update for every email an actor observes is $O(n+1)$. In case of the socio-cognitive network, since n<<N (N is total number of actors), the time complexity is practically also approximately linear.

4. Socio-cognitive Network Analysis

This section presents two useful socio-cognitive analyses which can be performed using the model described in the previous section.

4.1 Divergence between Beliefs

Given the belief states $B_{x,t}$ and $B_{y,t}$ for two actors A_x and A_y at time t, there is a need to measure the similarity between these belief states in order to quantify how similar the perceptions of the two actors. Since $B_{x,t}$ and $B_{y,t}$ are vectors of probability distributions, in this paper, for computing the divergence between $B_{x,t}$ and $B_{y,t}$, the divergence between respective pairs of beliefs in the two sets are computed and then combined. In this paper, the divergence between respective beliefs of two actors is defined as the KL-divergence [9] across the expected Bernoulli distributions for the two respective beliefs. The expected Bernoulli distribution for a belief is the expectation of the Beta distribution corresponding to that belief. If $J(i,j)_{x,t}$ is the Beta distribution, then the corresponding expected Bernoulli distribution is denoted as $E[J(i,j)_{x,t}]$, which is obtained by normalizing the parameters of Beta distribution $J(i,j)_{x,t}$ as follows,

$$E[J(i,j)_{x,t}] = [\frac{\alpha(i,j)_{x,t}}{\alpha(i,j)_{x,t} + \beta(i,j)_{x,t}}, \frac{\beta(i,j)_{x,t}}{\alpha(i,j)_{x,t} + \beta(i,j)_{x,t}}]$$

KL-divergence is an information-theoretic measure for quantifying directed divergence between two probability distributions. KL-divergence of a

probability distribution p from a probability distribution q, denoted as $KL(q||p)$, is given as,

$$KL(q \parallel p) = \sum_{\forall x} q(x) \log \frac{q(x)}{p(x)}$$

Since it is an asymmetric measure, the symmetric KL-divergence $KL_{sym}(q||p)$ is defined as,

$$KL_{sym}(q \parallel p) = KL(q \parallel p) + KL(p \parallel q).$$

Thus,

DEFINITION 3. The similarity between beliefs of email communication from A_i to A_j for actors A_x and A_y, expressed by the Beta distributions $J(i,j)_{x,t}$ and $J(i,j)_{y,t}$, at time t, is defined as,

$$Sim(J(i,j)_{x,t}, J(i,j)_{y,t}) = \frac{1}{1 + KL_{symm}(J(i,j)_{x,t} \parallel J(i,j)_{y,t})} \; and$$

$$KL(E[J(i,j)_{x,t} \parallel E[J(i,j)_{y,t}) = p \log \frac{p}{q} + (1-p)\log \frac{1-p}{1-q} \;...(4)$$

where, $p = \dfrac{\alpha(i,j)_{x,t}}{\alpha(i,j)_{x,t} + \beta(i,j)_{x,t}}$ *and*

$$q = \frac{\alpha(i,j)_{y,t}}{\alpha(i,j)_{y,t} + \beta(i,j)_{y,t}}$$

This similarity between two beliefs ranges from 0 to 1, with 0 and 1 indicating minimum and maximum similarity respectively. Definition 3 is used to measure the similarity between belief states of two actors in the following sections.

4.2 a-closeness Measure

An important analysis using belief states for each actor is to measure the similarity between actors' perceptions of email communication networks. This paper proposes a novel measure, *a-closeness*, to quantify such similarity in perceptions of two actors. This measure is based on the previous definition 3 of divergence between belief states of two actors at time t.

*DEFINITION 4 (**a-closeness**). The a-closeness measure is defined as the agreement between belief states $B_{x,t}$ and $B_{y,t}$ of actors A_x and A_y respectively at time t, and is given by,*

$$a-closeness(B_{x,t}, B_{y,t}) = \frac{\sum_{\forall (i,j) \in B_{x,t} \cap B_{y,t}} Sim(J(i,j)_{x,t} \parallel J(i,j)_{y,t})}{\sqrt{n(B_{x,t}) \cdot n(B_{y,t})}} \;...(5)$$

where $n(B_{x,,t})$ represents the number of beliefs (communication links) for which actor A_x has observed at least one email and $(B_{y,,t})$ represents the number of beliefs (communication links) for which actor A_y has observed at least one email.

The a-closeness for two belief states is symmetric and ranges between 0 and 1, with lower values representing lesser closeness and higher values representing more closeness. It attains a maximum similarity of 1 only when the two belief states are identical.

The numerator in definition 4 sums up the similarity between only those beliefs for which both A_x and A_y have observed at least one email.[1] The intuitive reasoning for this is now explained. An email communication network is usually quite sparse, i.e. out of all possible ordered pairs of actors, only a few of them will actually communicate. Hence, the belief states of the actors being compared will be even sparser and for both the actors, the beliefs associated with majority of communications will indicate very low probability of occurring (since no instances of these interactions have been observed). In such a case, it is desirable to disregard such beliefs while measuring similarity between actors' belief states. The situation analogous to computing document similarity, where one computes similarity based only on those words that are present in both the documents. Also, if the whole set of beliefs is considered for every actor, one implicitly assumes that the every actors is equally aware of the presence of all actors as well as all relations in the social network, which may be quite unrealistic. The denominator normalizes the numerator using the geometric mean of the number of beliefs for which each actor has observed at least one email.

The first application of a-closeness measure is to use it to construct a graph, called *"agreement graph"*, where nodes represent actors while an edge exists between two nodes if the a-closeness measure between those actors is greater than a user-specified threshold μ. This graph captures information about which pairs of actors have similar perception about email communication network. Classical social network analysis techniques can be applied to such a graph. For example, cliques represent groups of actors having similar beliefs of email communication networks, bow-ties represent articulation points, star structures identify the central actors, whereas bridges identify actors with similar beliefs to two or more other groups.

A second application of a-closeness is to compute the mean a-closeness across all ordered pairs of actors. This represents the *consensus* among the actors. Lower mean a-closeness indicates lower agreement within the social network while higher mean a-closeness represents higher agreement between the actors. In

---

[1] Other interpretations of closeness between belief states are possible and remains an interesting open research problem

addition, the standard deviation of across all actors quantifies the variance in agreement of actors in network.

4.3 r-closeness Measure

For second analysis, this paper introduces the concept of a *"super-actor"*, i.e. an actor who observes all the communication in the network. An email server is an example of a super-actor. A closed world assumption is made wherein all email communication is said to be sent through the email server, hence observed by it and no other email communication occurs between the actors.[2] Thus, the super-actor's belief state for email communication is a benchmark for reality, under the closed world assumption and the study of similarity between an actor's belief state and the super actor's belief state (reality) is a novel and interesting analysis. To quantify this divergence, this paper proposes r-closeness measure as defined below.

DEFINITION 5 (**r-closeness**). *The r-closeness measure is defined as the closeness of an actor A_x's belief state $B_{x,t}$ to super-actor's belief state (reality) $B_{S,t}$ at a time t and is given by,*

$$r-closeness(A_x) = a-closeness(B_{S,t}, B_{x,t}) \dots (6)$$

Higher is the r-closeness for an actor, more realistic are the actor's perceptions about email communication in the network. The mean *r-closeness* across all actors provides an aggregate measure of the "overall knowledge" or "level of perception" in the network. Higher is the mean r-closeness, then more actors in the network actually know about other actors' communications, i.e. the communication is transparent. A lower mean value for *r-closeness* indicates that actors generally have "misperceptions" regarding other actors' communications. The later is usually expected to be observed for a large social network consisting of various diverse groups, where it is difficult for a single actor to capture all communication in the network. . The standard deviation for *r-closeness* across different actors indicates the variance in the levels of perception.

Other application includes testing new hypotheses for socio-cognitive networks. For example, Krackhardt [6] explains that an actor's perception of "who communicates with whom" is a function of the actor's social position. In an organizational environment, it is believed that "top actors in the formal organizational hierarchy have better knowledge about communication than lesser actors and hence better perceptions about the social network", i.e executive management have a better perception of the social network as compared to employees. In addition, intuitively it is expected that, "more is the communication an actor observes, the better are actor's perceptions about the social interactions occurring in organization". Such hypotheses can be tested using the r-closeness measure computed for all actors.

5 Experimental Work

This section describes the experimental results for socio-cognitive network analysis of Enron email dataset using the proposed model and measures.

5.1 Enron Email Corpus

The Enron email corpus (http://www.cs.cmu.edu/~enron/) is a set of emails between 151 users, mostly senior management of Enron, exchanged between mid-1998 and mid-2002 (approximately 4 years), which includes the Enron crisis that broke out in October 2001. In the current experimental setup, a cleaned version is chosen, in which duplicate, erroneous and junk emails have been removed (Shetty and Abidi [11]). The data consists of 252,759 email messages for the set of 151 users. For this experimental analysis, first the entire set of 151 users is chosen, and then only those emails (approx. 20,311) which are exchanged between these 151 users were selected. The length for the time window was chosen to be one month. Results for different values of $\lambda \in \{0, 0.5, 1\}$ are compiled, where $\lambda=0$ represents no history, $\lambda=1$ includes all history and $\lambda=0.5$ represents an exponential decay of history.

5.2 Experimental Results

5.2.1 a-closeness

An agreement graph for socio-cognitive network is constructed using the a-closeness of actors (employees) at the end of October 2000 and October 2001. An edge is drawn between two actors only if the a-closeness between them was more than a certain threshold μ ($\mu \in \{0.25, 0.5, 0.7\}$). The a-closeness values between actors are observed to be low in general and interesting trends are observed only for $\mu=0.25$. Figures 4 (a), (b) and (c) show the agreement graph for October, 2000, for different values of λ and $\mu=0.25$. It is observed that each graph consists of many small, disjoint components of users. A possible reason for this is because big

---

[2] This assumption will be relaxed in future research.

organizations like Enron usually have many organizational groups with high intra-group communication and low inter-group communication. Interesting structures like cliques, bowties and stars are observed in the agreement graph. Except for a few changes in edges, no significant changes are observed for different values of λ and almost the same clusters of actors are observed. But the reason for such lack of changes with λ may be because of the nature of the underlying dataset, the nature of analysis (looking mainly at macro level statistics and trends) as well as the choice of the time window length. For smaller time windows or for other datasets, interesting, unexpected changes might be observed for different values of λ.

Figures 5 (a), (b) and (c) show the agreement graph for October 2001, for $\mu=0.25$ and $\lambda \in \{0, 0.5, 1\}$. Each one of them mainly consists of one large, connected component (except for $\lambda=0$ where the large components breaks up into two large components, however, this does not affect the general conclusions drawn regarding the October 2001 a-closeness trends). This indicates that there is a considerable extent of the overlap in social perceptions during the crisis period. The connectivity of the October 2001 agreement graphs also indicates that communication (and hence information) is shared among various actors and pairs of actors are "few hops" away from each other in terms of cognitive overlap. Such a network is highly conducive towards dissemination of ideas in a social network. Indeed, in case of Enron dataset, the Enron crisis was a "hot topic" that was often discussed in the underlying social network.

Also, note that the number of nodes in the October 2001 graphs is much more than that of the October 2000 graphs. An actor is included in the agreement graph only if it's a-closeness with at least one actor, crosses the threshold. In October 2000, many actors are isolated from the rest of the network due to less email communication between most actors while in October 2001 almost all actors are part of one big component due to high overlap of email communication.

We also observe some interesting structures such as -cliques of actors having similar r-closeness and persistent cliques (cliques that exist in both October 2000 and October 2001). For example, a clique of traders such that all traders had similar low r-closeness measure shows there was agreement in the perceptions of the group, but the entire group is far removed from reality. The second example is a clique of employees which is a persistent clique (disconnected clique in the top right corner of Figure 9). For actors present in such a persistent clique, there is probably a strong

correlation in their roles, like all such actors worked on the same project. Though insufficient knowledge regarding the domain of data limits the understanding of causes for such structures, the proposed methodology holds promise in finding interesting patterns/structures of the socio-cognitive aspects of Enron email data, which traditional approaches fail to capture.

5.2.2 r-closeness

The r-closeness across actors is examined for two different months, October, 2000, a month with normal email activity in the organization, and October, 2001, a month during the Enron crisis. In each case, users are ranked in the decreasing order of r-closeness. For October, 2000, the actors can be roughly divided into three categories. The first category consists of actors who are communicatively active and observe a lot of diverse communications. These actors occupy the top positions in the rankings. These are followed by the second category actors who also observe a lot of communication; however, their observations are skewed which in turn leads to skewed perceptions. The third category consists of actors who are communicatively inactive and hardly observe any of the communication. These actors have low r-closeness values and are at the bottom of the rankings table. Table 1 summarizes the percentages of various actors (according to their formal positions) in the different ranges of r-closeness rankings. Using the rankings for October 2000, two socio-cognitive network hypotheses of interest to sociologists are studied.

H1. Higher is an actor in the organizational hierarchy, better is his/her perception of the social network.
From the r-closeness rankings, it is observed that majority of the top positions are not occupied by higher level executive employees. The top 50 ranks consist of a large chunk of the employee population (around 46.4% of the employees) along with 21.4% of the higher management and 34.4% of the executive management actors (see Table 1). A related observation is that most of the higher level executives are communicatively inactive and therefore have fewer perceptions.

H2. The more communication an actor observes, the better will be his/her perception regarding the social network.
It is observed that even though some actors observe a lot of communication, they are still ranked low in terms of r-closeness. A main reason for this is that actors tend

to participate in only certain communications and participate less in other communications. This results in perceptions about the social network that are skewed towards those "favored" communications. Executive management actors who observed a lot of communication showed a tendency for this "skewed perception" behavior.

Table 1. Users in different rank ranges of r-closeness (October 2000, λ=0.5)

Ranks	Not Avail-able	Emplo-yees	Higher Manage -ment	Executive Manage- ment	Others
1-10	10.3% (4)	4.9% (2)	7.1% (2)	3.4% (1)	7.1% (1)
11–50	17.9% (7)	41.5% (17)	14.3% (4)	31.0% (9)	21.4% (3)
51-151	71.8% (28)	53.6% (22)	78.6% (22)	65.6% (19)	71.5% (10)

Table 2. Users in different rank ranges of r-closeness (October 2001, λ=0.5)

Ranks	Not Avail-able	Empl oyees	Higher Manag ement	Executive Manage- ment	Othe rs
1-10	5.1% (2)	2.5% (1)	3.6% (1)	20.7% (6)	0% (0)
11–50	23.1 % (9)	26.8% (11)	28.6% (8)	37.9% (11)	7.1% (1)
51-151	71.8% (28)	70.7% (29)	67.8% (19)	41.4% (12)	92.9% (13)

Other observations from this socio-cognitive network analysis of Enron email data are discussed below. Table 2 summarizes statistics for r-closeness rankings for the month of October 2001. The rankings for the crisis month October 2001 are significantly different from those of October 2000. For both the months, the distribution of various actors among the r-closeness rankings was only slightly different for different λ values.[3] For all values of λ, it was observed that the percentage of management staff among the top 50 ranks increased significantly at the cost of employees being pushed down. Thus, a shift from the normal behavior is observed, indicating that communication perceived by most management level actors is more diverse and evenly distributed as compared to the skewed or no perceptions in Oct 2000. A possible reason for this may be that during the crisis month,

---

[3] Due to space constraints, only results for λ = 0.5 are illustrated here.

emails were exchanged across different levels of formal hierarchy in the organization thus exposing management level actors to more diverse communication [3]. Another possible and intuitively appealing reason [3] is that during October 2000, on an average, management people "sent" about 80% and "received" only 20% of the total communication they were exposed to. In the October 2001, there was a reversal and management people sent only 20% and received about 80% of their total communication. Since they observed a lot more communication during the later period, there was a significant increase in the r-closeness ranks of management level actors during October 2001. Finally, management level actors were also lot more communicatively active in October 2001 than in October 2000 (i.e. they were exposed to a lot more communication during the crisis period and so the 80% of October 2001 is greater than the 20% of October 2000).

Figure 6 is a plot of mean *r-closeness* of all actors over time for different values of λ. An interesting observation is that, for λ=0, the mean *r-closeness* peaks during the crisis month of October 2001, indicating a general increase in the perception of social interactions during the crisis period. After the crisis period, mean r-closeness drops down. For λ > 0, the plots are almost identical and it is observed that r-closeness increases until the crisis period and after that it stabilizes. This can be attributed to increased communication among actors. Since almost each actor in the network was involved in some communication, as a result, the general awareness of an actor increased. The difference in observation for λ= 0 and λ > 0 is due to the "memory" effect introduced by taking λ > 0.

6 CONCLUSIONS AND FUTURE DIRECTIONS

The growing popularity of computer network-based social networks and the ability to collect gigabytes of *unbiased* social information provides a unique opportunity for computer scientists to develop new computational techniques for mining social network patterns. This paper makes important contributions to this research by (i) providing a scalable computational for modeling socio-cognitive networks for email communication network, (ii) proposing a measure to quantify similarities in individual actors' perceptions of social network in such a socio-cognitive network and using it to construct agreement graphs between actors, (iii) identifying a novel analysis, enabled by social network on computer networks, for quantifying how well an actor's perceptions reflect reality and proposing

a new measure for the same, and (iv) illustrating the use of these techniques using a real-world Enron email data. These techniques provide a handy computational tool for sociologists to analyze large datasets and will enable in advancing the understanding of such social networks. This paper will motivate research in developing new computational tools (e.g. more sophisticated, scalable approaches) for email based social networks. Future research directions include (i) incorporating semantic information about the contents of email and (ii) different weights of importance for actors in To, Cc and Bcc fields of the email.

7. Acknowledgements

8. References

[1]   N. Contractor (1998) Formal and Emergent Predictors of Coworkers' Perceptual Congruence on an Organization's Social Structure. Human Communications Research, 24, 536-563.

[2]   R. Cross, N. Nohria, A. and A. Parker. Six Myths About Informal Networks — and How To Overcome Them. *Sloan Management Review*, 43(3), pp. 67-76, 2002.

[3]   J. Diesner, and K. Carley. (2005). Exploration of Communication Networks from the Enron Email Corpus. *Proc. of Workshop on Link Analysis, Counterterrorism and Security, SIAM International Conference on Data Mining* 2005, pp. 3-14. Newport Beach, CA, April 21-23, 2005.

[4]   P. Hoff, A. Raftery and M. Handcock. (2002) Latent Space Approaches to Social Network Analysis. *Journal of American Statistical Association,* Vol. 97(460), 1090-1098.

[5]   E.M. Jin, M. Girvan, M.E.J. Newman (2001) The structure of growing social networks. *Phys. Rev.*, E 64, 046132.

[6]   B. Klimt and Y. Yang. (2004). Introducing the Enron corpus. *First Conference on Email and Anti-Spam (CEAS).*

[7]   D. Krackhardt. (1990). Assessing the political landscape: structure, cognition, and power in organizations. *Administrative Science Quarterly* 35, 342-369.

[8]   D. Krackhardt and J. Hanson. Informal Networks: The Company behind the Chart. *Harvard Business Review*, 104-111, July-August, 1993.

[9]   S. Kullback and R. Leibler. (1951) On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79-86.

[10] N. Pathak, S. Mane and J. Srivastava. (2006) Who Thinks Who Knows Who ? Socio-Cognitive Analysis of Email Network. *CSE Technical Report, University of Minnesota, Minneapolis, USA.*

[11] J. Shetty and J. Adibi (2004). The Enron email dataset database schema and brief statistical report. Technical Report, ISI, University of Southern California.

[12] J. Shetty and J. Adibi (2005) Discovering Important Nodes through Graph Entropy - The Case of Enron Email Database. *In Proc. of LinkKDD, in conjunction with the 11th ACM SIGKDD.*

[13] M. Van Duijn, J. Van Busschbach, T. Snijder (1999) Multilevel Analysis of Personal Networks as dependent Variables. *Social Networks*, 21, 187-209

[14] J. Vermunt, M. Kalmijn (2006) Random Effects models for personal networks, An application to marital status homogeneity. Methodology, 2, 34-41

[15] S. Wasserman and K. Faust. (1994) Social Network Analysis – Methods and Applications. Cambridge University Press.

[16] S. Wasserman and P. Pattison (1996) Logit Models and Logistic Regression for Social Networks: I An Introduction to Markov Graphs and p*. *Psychometrika*, 61, 401-425.

[17] B. Wellman. (2001). Computer Networks as Social Networks. *Science*, 293(14).

APPENDIX
A. VALID BELIEF STATES

Consider the expected Bernoulli distribution $E[J(x,y)]$ using the Beta distribution $J(x,y)$, in an actor's belief state. The parameter of $E[J(x,y)]$ is the expected communication probability $E[P(x,y)]$, according to the actor, given by,

$$E[P(x,y)] = \frac{\alpha(x,y)}{\alpha(x,y) + \beta(x,y)} \quad \text{where } \alpha(x,y) \text{ and } \beta(x,y)$$

are the parameters of Beta distribution $J(x,y)$. Since, communication probabilities are defined as fractions of the total communication,

We must have, $\sum_{\forall(i,j),i\neq j} E[P(i,j)] \geq 1$

Recall, $P(i,j) = P_i P_{j|i}$

Each $P_{j|i}$ has a maximum value of *1*, which gives

$$\sum_{\forall j, i\neq j} P_{j|i} \leq (N-1)$$

Since, $\sum_{\forall i} P_i = 1$, we must have,

$1 \leq \sum_{\forall(i,j),i\neq j} E[P(i,j)] \leq (N-1)$, where N is the number of actors

If the expected communication probabilities in the belief state of an actor do not satisfy the above inequality then we say that the actor's belief state is

"invalid" (i.e. the particular set of expected communication probabilities inferred by the actor cannot actually exist).

PROPOSITION 1. If the prior probabilities of a belief state are such that the belief state is valid, then the posterior probabilities will also result in a valid belief state.

Let the communication probability $P(i,j)$ have prior probability x_{ij}. Suppose $\alpha(i,j) = x_{ij}$ and $\beta(i,j) = 1-x_{ij}$. Then the expected communication probability will be,

$$E[P(i,j)] = \frac{\alpha(i,j)}{\alpha(i,j) + \beta(i,j)} = x_{ij}$$

Also assume that the priors correspond to a valid belief state i.e.

$$1 \leq \sum_{\forall(i,j),i \neq j} x_{ij} \leq (N-1) \quad \cdots (8)$$

Suppose, in a time interval, an actor observes M emails out of which k_{ij} are emails from actor A_i to A_j. We have,

$$M \leq \sum_{\forall(i,j),i \neq j} k_{ij}$$

$$\Rightarrow M + 1 \leq \sum_{\forall(i,j),i \neq j} (x_{ij} + k_{ij}) \quad \cdots (9) \text{ (from 8)}$$

$\sum_{\forall(i,j),i \neq j} k_{ij}$ is maximum when every email, from some actor, is addressed to every other actor.

$$\Rightarrow \sum_{\forall(i,j),i \neq j} k_{ij} \leq M(N-1)$$

$$\Rightarrow \sum_{\forall(i,j),i \neq j} (x_{ij} + k_{ij}) \leq M(N-1) + (N-1) \quad \cdots (10) \text{ (from 8)}$$

From (9) and (10) we have,

$$1 \leq \sum_{\forall(i,j),i \neq j} \left(\frac{x_{ij} + k_{ij}}{M + 1}\right) \leq (N-1)$$

But, $\dfrac{x_{ij} + k_{ij}}{M + 1} = x'_{ij}$, where x'_{ij} is the expected posterior probability, $E[P(i,j)]_{posterior}$.

$$\therefore 1 \leq \sum_{\forall(i,j),i \neq j} x'_{ij} \leq N-1 \cdot$$

The above proof also holds for the case when priors for Beta distribution parameters do not sum up to *1* i.e. $\alpha(x,y) = rx_{ij}$ and $\beta(x,y) = r - \alpha(x,y)$, where r is some real valued scaling factor indicating the confidence in the prior probability x_{ij} ∎

The priors x_{ij} can be expressed a product $\delta_i \varepsilon_{j|i}$, where δ_i is prior for P_i and $\varepsilon_{j|i}$ is prior for $P_{j|i}$. In some cases instead of directly working with x_{ij}, it might be easier to fix δ_i and $\varepsilon_{j|i}$ such that (8) is satisfied.
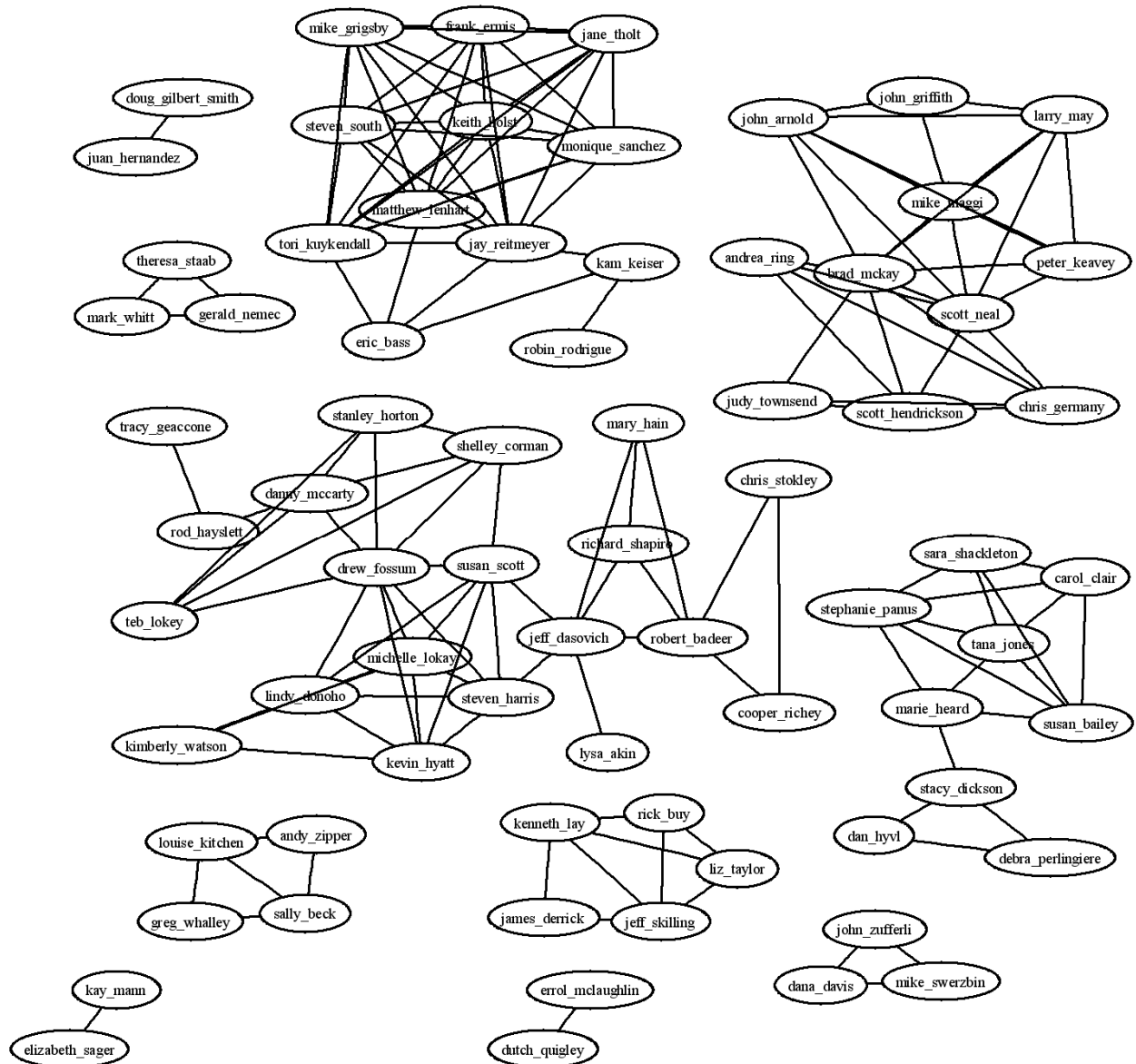
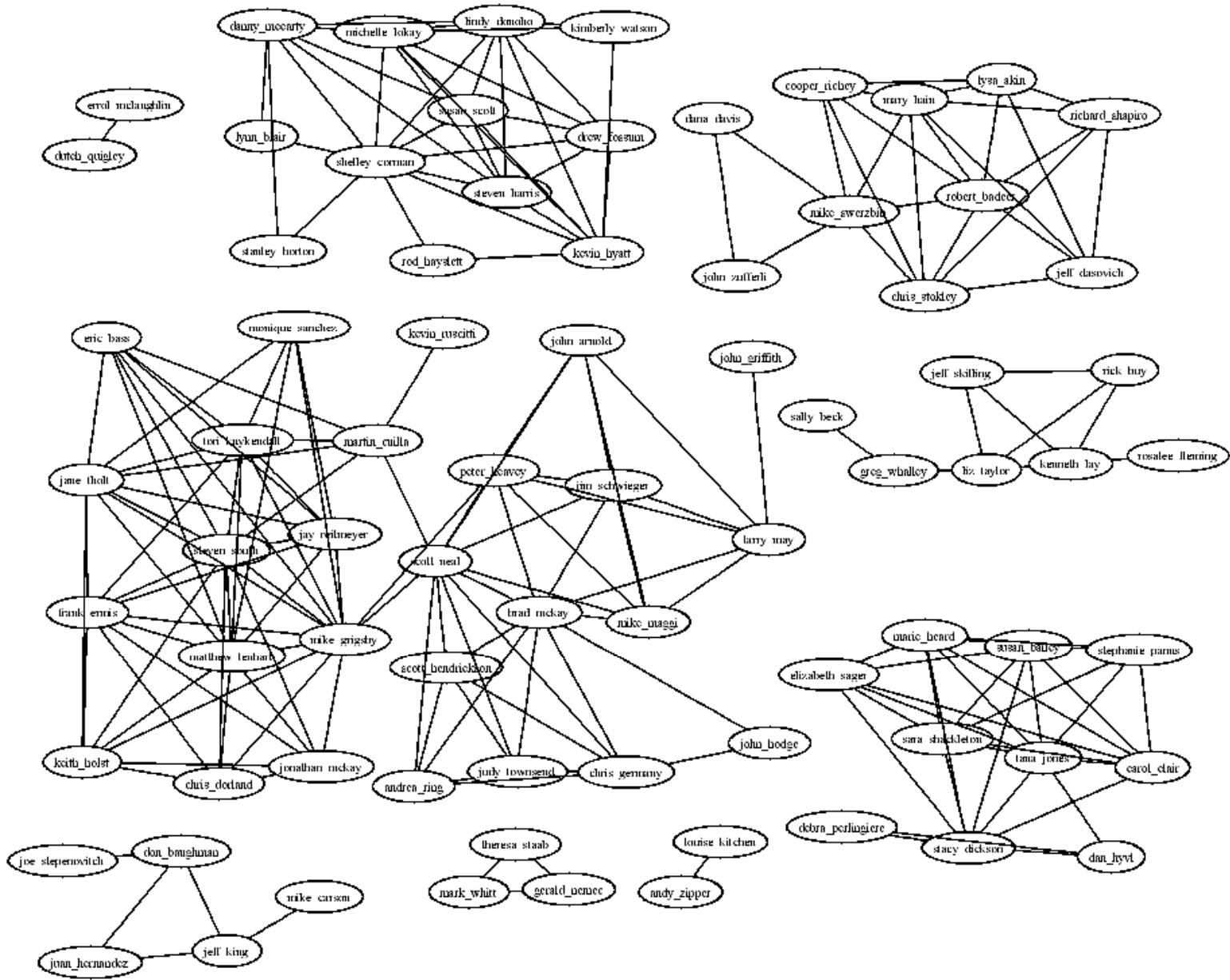**Figure 4 (a). Agreement graph for October 2000 (μ = 0.25 and λ = 0)**

Figure 4 (b). Agreement graph for October 2000 ($\mu = 0.25$ and $\lambda = 0.5$)

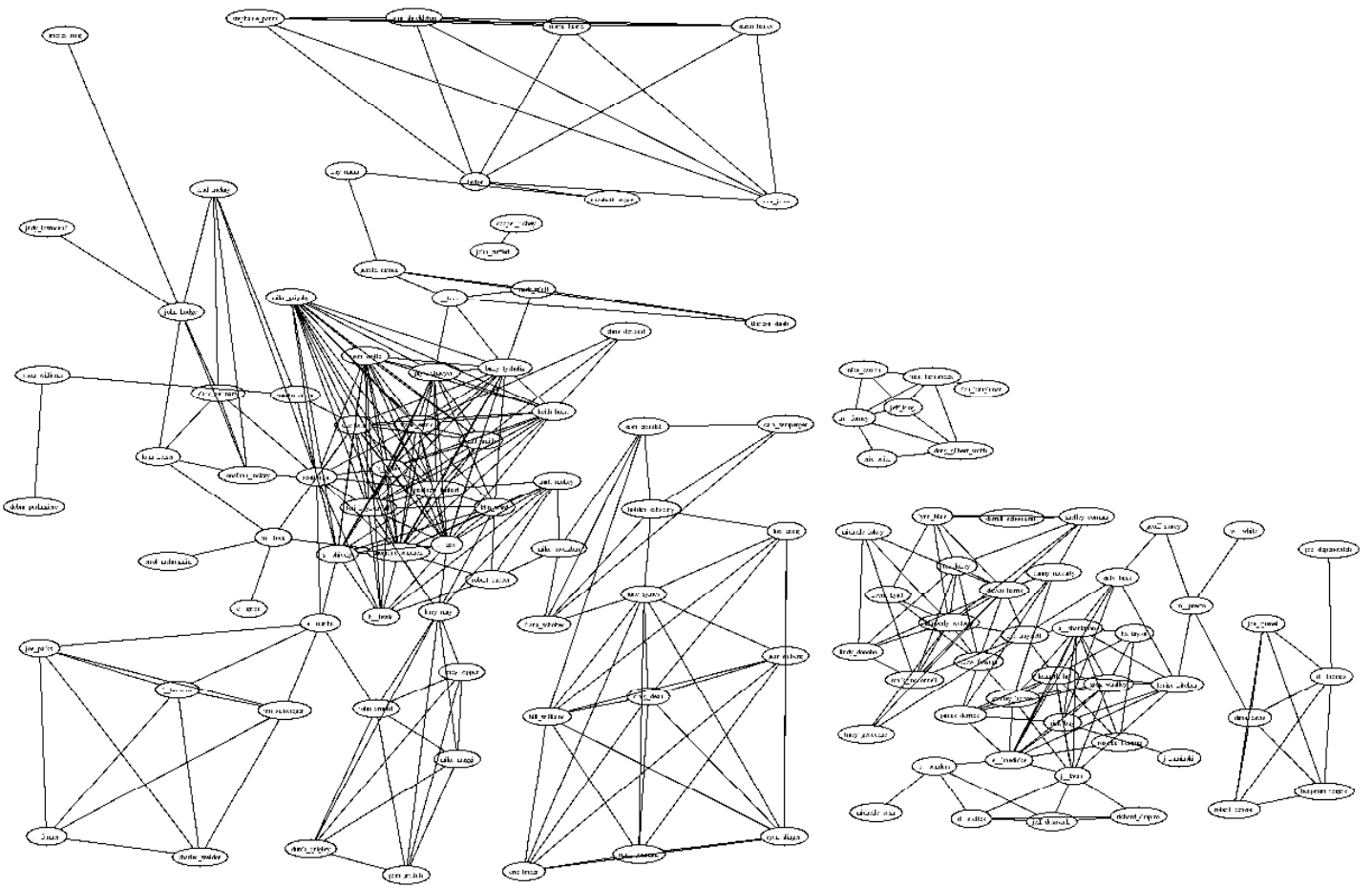**Figure 4 (c). Agreement graph for October 2000 ($\mu = 0.25$ and $\lambda = 1$)**

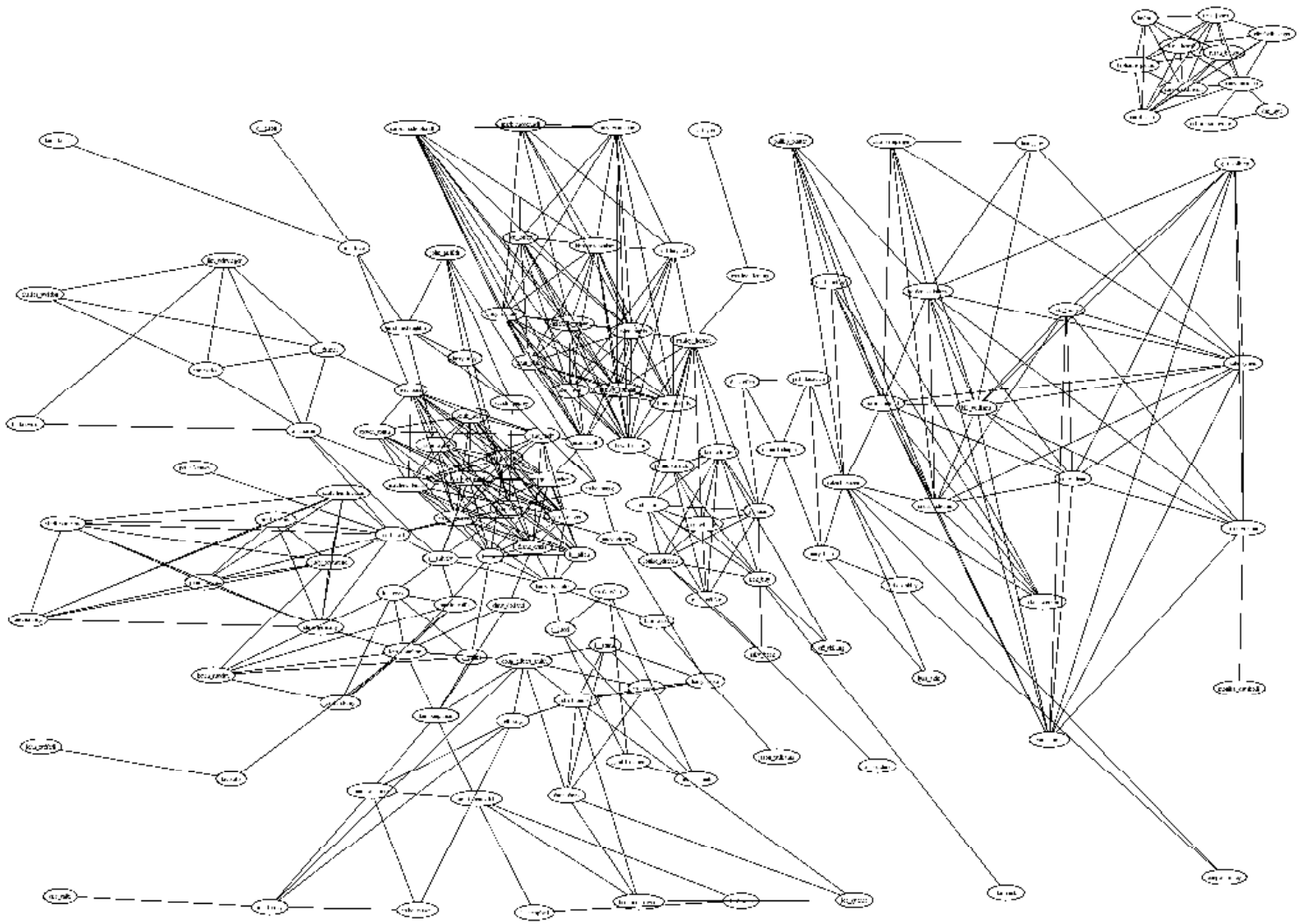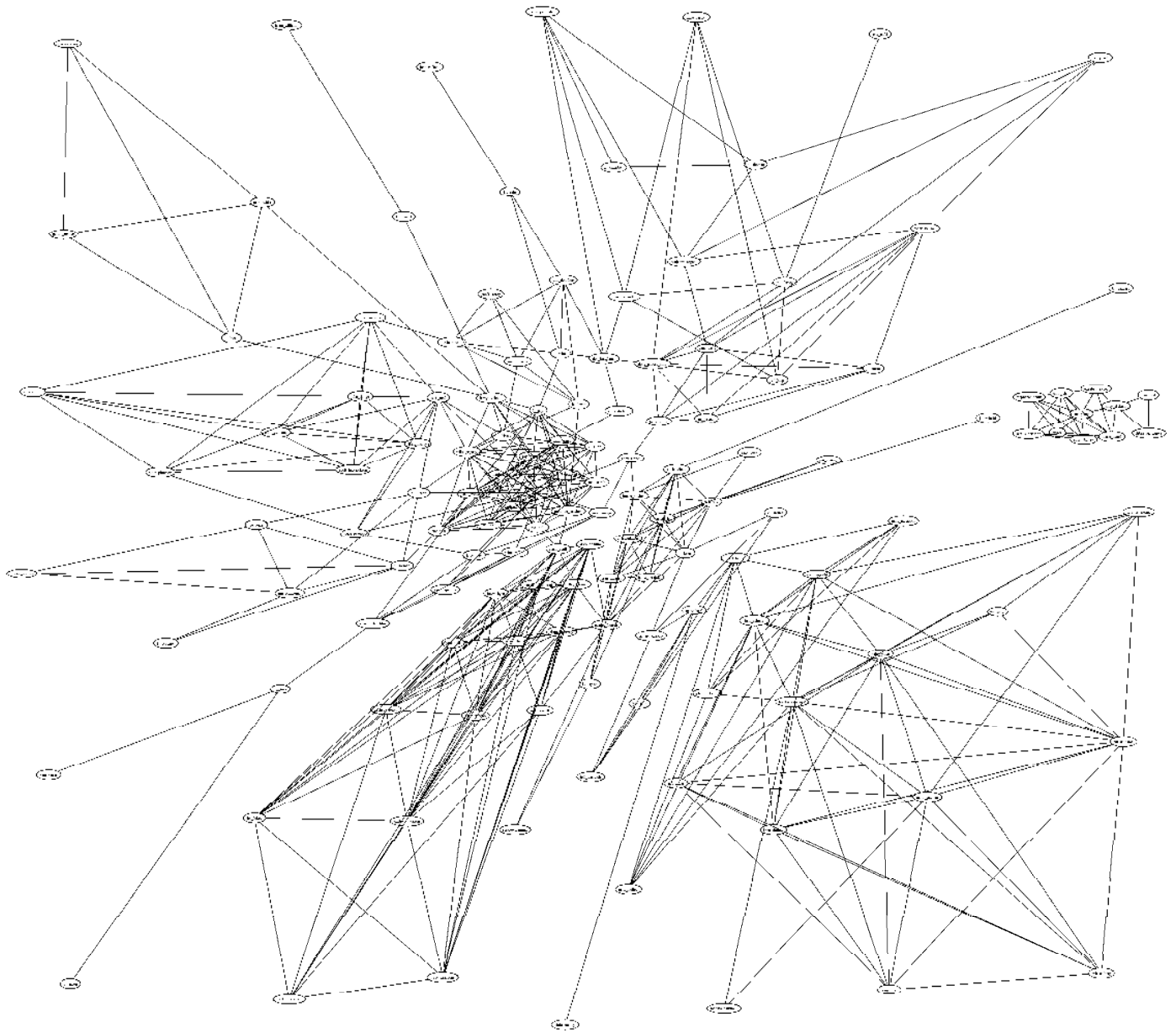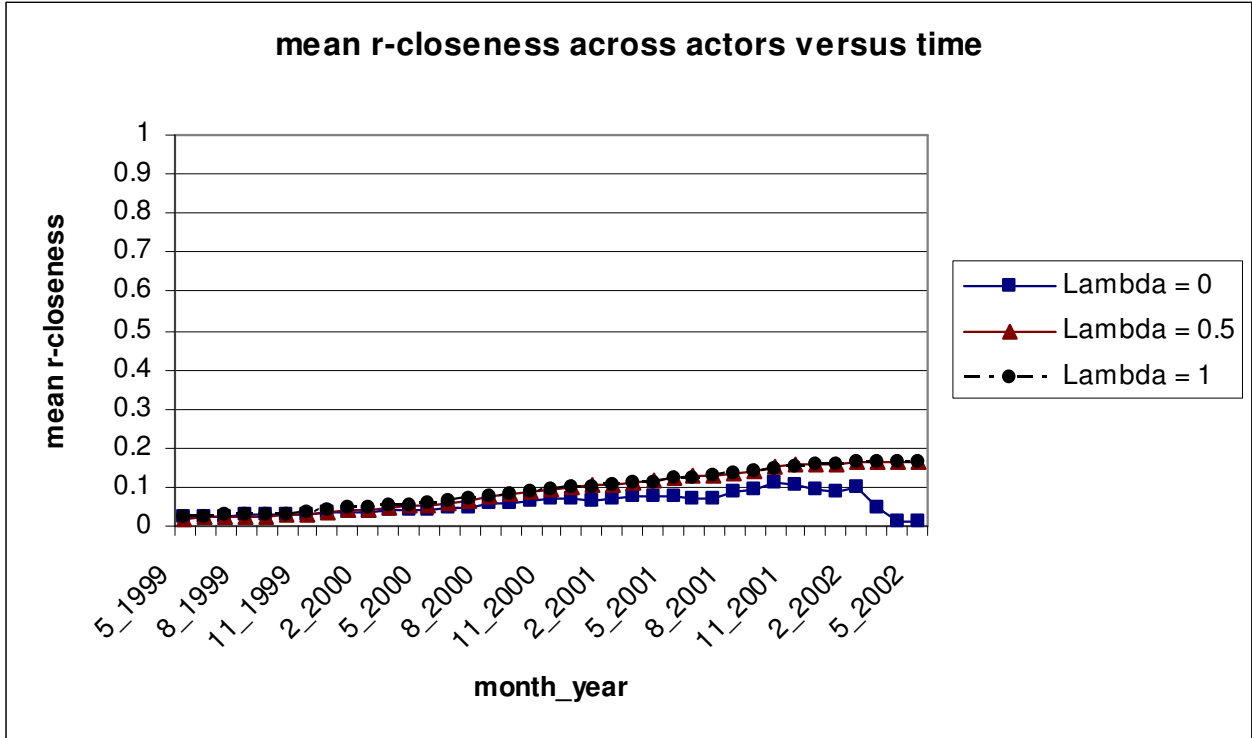**Figure 5 (a). Agreement graph for October 2001 ($\mu = 0.25$ and $\lambda = 0$)**

Figure 5 (b). Agreement graph for October 2001 (μ = 0.25 and λ = 0.5)

Figure 5 (c). Agreement graph for October 2001 ($\mu = 0.25$ and $\lambda = 1$)

Figure 6. Mean r-closeness across actors