Using Email-Based Network Analysis to Determine Awareness Foci

Adriana S. Vivacqua[1] and Jano Moreira de Souza[1,2]

[1] COPPE/UFRJ, Graduate School of Computer Science,
[2] DCC/IM, Institute of Mathematics
Federal University of Rio de Janeiro,
Rio de Janeiro, Brazil
{avivacqua, jano}@cos.ufrj.br

Abstract. A number of studies have indicated that awareness of others' activities plays an important part in collaboration. Consequently, awareness has been a frequent theme in cooperative work research. Researchers have acknowledged that proximity has a strong effect on collaboration, and that maintaining awareness of peers becomes harder in distributed environments. Many awareness systems require configuration by the user and work only in predefined shared environments. In this paper, we present an investigation into the determination of awareness targets, through email-based user interaction analysis. The final goal is to be able to draw inferences as to who and what a user would be interested in maintaining awareness of, enabling a system to automatically determine awareness foci and adjust itself according to its user.

1 Introduction

The dissemination of network technology and adoption of distributed work teams by companies has led to a move towards remote work: individuals that used to be collocated might now be spread throughout the world. Studies have shown increased adoption of virtual work teams, in which members are geographically dispersed and communicate and coordinate mainly via electronic communication tools [16]. People participate in several projects at the same time, dividing their time and attention accordingly [21]. Individuals must therefore organize themselves and their work to accomplish different tasks, very often with different collaborators. Participation in different groups usually means that, depending on the situation, a person might have different roles and obligations, perform different activities and work towards different goals, all of which must be managed so they do not conflict with each other.

This leads to the notion of supporting individual work and tying it to the group as appropriate [25]. We work with looser collaborative environments, in which individuals need tools that enable them to quickly switch into closer interaction when necessary, and to easily relate their work to that of others.

In collocated environments, individuals are capable of observing others' actions, thereby gathering awareness information [14]. With increased distribution and implementation of virtual teams, opportunities for collaboration, interaction and information

Y.A. Dimitriadis et al. (Eds.): CRIWG 2006, LNCS 4154, pp. 78 – 93, 2006.

exchange may be compromised: in these environments, casual interactions seldom happen and observation of others becomes harder.

A looser structure and distance sometimes lead to fragmentation: members may not communicate very often or be kept up-to-date of the latest evolution in others' work, resulting in rework, delays or confusion. The focus of this research is on improving awareness of the work environment in order to facilitate the group's work. This paper describes a method to automatically distribute task awareness information among group members and an initial reflection upon some of the assumptions underlying this method. Our system has been conceived as a means of integrating individual work with the shared group context, with the final goal of improving cohesion and reducing fragmentation. We expect such a system will promote informal interaction and facilitate opportunistic collaboration when deployed.

The remainder of this paper is organized as follows: in the next section we present a brief literature review of the area, followed by the envisioned system in section 3. Section 4 contains a preliminary analysis, followed by a discussion in section 5.

2 Background Literature

2.1 Self Governing Groups

In self-governing groups, actors have control over job allocation, day-to-day production planning and control [2]. These groups emerge out of a need to handle unpredictable events or contingencies (and usually dissolve when they are no longer necessary), and enable an organization to quickly adapt to new demands generated by the environment, sometimes deviating from pre-established norms and rules. In many cases, groups are composed of peers, where there is no formal hierarchical structure. This means that many of the decisions are the result of arrangements between peers, as is the work that finally gets done [1].

Due to the underlying interdependence between tasks, workers have to articulate (i.e., divide, allocate, coordinate, schedule, interrelate, etc.) their activities [28]. When individuals collaborate, they often shift back and forth between individual and shared work, and between loosely and tightly coupled collaboration [14]. This is especially true when there is low interdependence between them [15]. A reasonable approach in these cases is to provide individual work support and add collaboration support to the individual work tools, enabling collaboration when necessary. Awareness of current and past efforts becomes necessary, since one individual might work on a shared artifact for a while and another may pick it up later [5]. This looser structure and distance may lead to a decrease in involvement and interaction. As a consequence, individuals miss opportunities for collaboration, and sometimes end up working individually because they are unaware of each other's activities, performing overlapping tasks or duplicating work.

2.2 Awareness

Situation awareness involves perception and interpretation of relevant elements of the environment. The basic set of elements that compose workspace awareness information are those that address the "who, what, where, when and how" questions: who are we working with, what are they doing, where are they working, when and how certain events happen.

Awareness is knowledge about the environment that must be maintained as it changes. It is maintained through perceptual information gathered from the environment; and it is generally secondary to other goals. Staying aware of others is taken for granted in everyday interactions, but becomes hard in distributed systems, where communication and interaction resources are poor [14]. This information facilitates collaboration by simplifying communication and coordination, allowing better management of coupling and determination of the need to collaborate: prior research has established that awareness of others is important in integrating a group [23], creating and maintaining shared context [13], and establishing contact [11].

2.2.1 Focus and Nimbus Theory

The Focus and Nimbus model of awareness for shared applications is based on spatial models of interaction [26]. It considers a set of objects in space, which interact based on their levels of awareness. Awareness, in turn, is manipulated via focus and nimbus, subspaces within which an object directs its presence or attention. Awareness is the overlap between nimbus and focus, where:

- Nimbus is the information given out by each object in the space, which can be perceived by others, and
- Focus describes the objects at which a user directs his or her attention.

In a collocated environment, individuals give out a large amount of information while working, which can be picked up by anyone paying attention to it. In computational settings, users give out information via the applications they interact with and the operating system, which is normally not relayed to others. We believe some of this information might be of use to help the group coordinate and conduct its work: other users should be able to pick up part of the information generated, depending on their focus. In our approach, we determine a user's focus through an analysis of his or her ongoing interactions. We are also working on a privacy scheme to automatically determine a user's nimbus.

2.3 Social Worlds and the Locales Framework

The Locales Framework [8] provides a set of abstractions to support the design and analysis of collaborative work. It is based primarily on the notion of continually evolving action and of *Social Worlds.* A Social World is a group of people who share a commitment to collective action, and it forms the prime structuring mechanism for interaction (as defined by Strauss, cited in [9]). Individuals are usually involved in multiple social worlds at a time, which means that different social worlds are interconnected and that actions in one social world may reflect in another. Each individual typically engages in multiple activities that span more than one social world.

In this framework, a *Locale* is an abstract concept that arises from the use of space and resources by a group. It maps the relationship between a Social World (and its interaction needs) and the *sites* and *means* its members use to meet those needs. Sites are the spaces (e.g. shared file systems) and means are objects contained in these spaces (e.g. the files and documents stored in this file system) [7].

Following these lines of thought, we have been working on collaboration support systems that take into account the emergent and situated nature of work, and the fact that individuals constantly reorganize to perform their tasks. We are working on systems to help the individual connect his work to others who relate to them.

2.4 Social Network and Interaction Analysis

Social network analysis is used widely in the social and behavioral sciences, as well as economics. It concerns the study of social entities and their relationships: communication among individuals, trade between businesses or treaties between nations. It considers structures such as the sociogram, a graph that represents individuals and the relations between them [31]. These relations can be of diverse nature (communication, party attendance, etc.), and are usually expressed as graphs and matrixes; upon which network analysis can be performed [29]. Social network analysts look at the world in terms of patterns or regularities in relationships between actors.

Sociocentric analysis looks at relationship structures from a global perspective (e.g., a graph of the communication between all members of a department or group). Egocentric network analysis, on the other hand, focuses on the individual (*ego*), and analyzes his or her interactions with a set of others (*alters*). This type of network has been used to study the social environment surrounding individuals or families, and social support structures [31].

Electronic interactions usually leave traces, such as email, fora or messenger logs. These interactions display certain rhythms that correspond to the individuals work patterns [24], and can be used to study the evolution collaborative endeavors. For instance, intense message exchange usually accompanies cooperative work. Additionally, individual patterns of email exchange can also indicate hierarchy and positioning in a group [6]. We construct an egocentric network based on the email records of electronic communication, and search this network to discover ongoing collaboration and need for awareness information.

3 An Approach for Awareness Information Distribution

To bridge the gap between individual and joint work, we have designed a distributed, peer to peer system to provide awareness information. In this system, agents check each user's current activities and ongoing interactions and exchange information with other peers to keep its user informed of their activities. Each agent's goal is to *maintain awareness between peers by displaying information about the activities of its user's acquaintances*. To reach this goal, the agent:

1. collects information generated by the user while working on his or her computer;
2. exchanges information with other users' agents; and
3. provides information to the user about his or her alters' activities.

This means that the agent must filter the information down to that which might be of interest to its user. In this paper, we present a method to determine awareness foci.

An egocentric network is built based on the set of user acquaintances. This network can be viewed as a tree with ego (the user) at the root and his or her alters (the acquaintances) at the first level, to which the information generated by each user (the list of tasks he or she is currently performing) is added as a second level. Selecting the appropriate information thus becomes a problem of determining which of the leaves in this tree are of interest to the user and pruning the answer space accordingly. The determination of interest foci is divided into two stages, discussed in more detail in the following subsections:

1. discovering which peers the user might be interested in (selecting nodes at the first level); and
2. deciding which of their activities the user would want to know about (selecting leaves).

In this section, we describe the reasoning used to select from the universe of available information (everything generated by other users and provided to the assistant agent) that which is relevant to the user, which we call the *focus of interest.*

3.1 Information Processing and Organization

To reason about its user's needs, the agent gathers information about ongoing interactions from email logs. This information is organized to represent ongoing relationships and interest foci. The following concepts are used: a *tie* is a relationship between two users. It is composed of *interactions* between these parties. These interactions in turn are composed of *message exchanges*, which are groups of *email messages* (raw data). A series of email messages is grouped into an interaction, and several interactions define a tie. These concepts are illustrated in Figure 1.

To construct the user's network, we take the values of the *From, To, CC* and *BCC* fields and build the user's list of acquaintances. In this network, *Ego* is the user (normally determined by looking at the *From* field of outgoing emails), and his or her *Alters* are the many peers with whom ego exchanges email. The system groups email messages and their replies (determined via *Subject, Message-ID* and *Reference-To* tags) into interactions (conversations in GMail), qualifying each reply by the time it took the user to respond (extracted from the *Date* field). An interaction contains several messages, qualified by length (number of emails) and duration (time from first to last message). A tie is characterized by the frequency of interaction between alters, i.e. how often they exchange mail.

For each user, average frequency of interaction and average response time are also calculated generically (how quickly does ego respond to email or how often he or she sends/receives email) and per alter, which we believe will turn out to be more significant (how often does ego send email to alter A and how quickly does ego reply to messages from alter B).

To reduce the search space, this network is pruned before adding the set of activities each alter is performing. It is easy to see how this search space can become quite large, which is why the system attempts to infer the need for awareness information. As an illustration, picture a user with 100 contacts in his or her address book, each of which performing 3 or 4 tasks simultaneously – if we were to provide the user with this raw data, he or she would have to keep track of 100 different people performing 300-400 different tasks to determine which ones are interesting, which would most

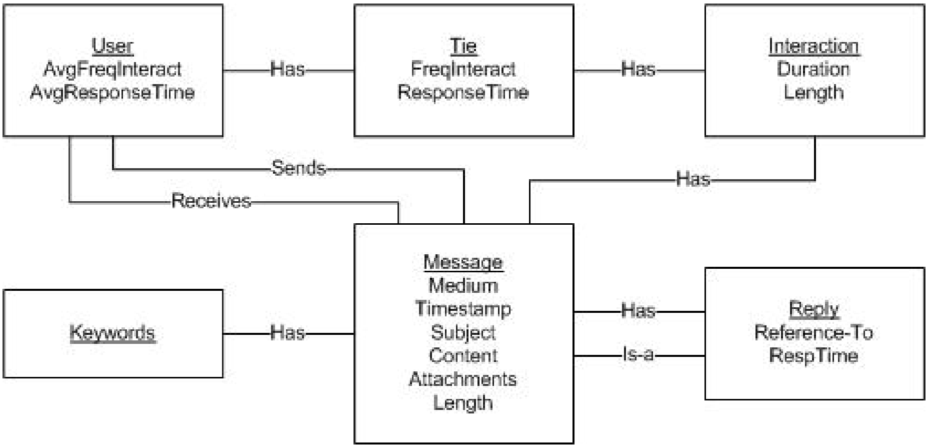


Fig. 1. Conceptual model of messages and interactions

likely result on serious information overload. We estimate only about 10-15 peers will be of interest at any given moment, depending on the groups and activities a user is engaged in.

In addition to the structural processing, the system performs content analysis on the messages, clustering them according to their topics as well as sender-recipient groups. Message bodies are processed for keywords and keyword vectors are built for each message. Interactions and ties are also classified according to the keywords found in the messages. This enables us to determine the themes of the interactions and defines the shared context for ego and each of his or her alters, which is later used for matching the group context to the individual tasks. We are also considering the use of concepts and activity ontologies to enrich the classification and matching.

The agent also keeps track of its user's activities, periodically extracting ongoing activity lists from the operating system, with application and file names. Textual files (pdfs, word documents) are processed for keywords in the same manner as messages, and compared to ongoing interactions. This ties work in progress to ongoing interactions, and should hopefully yield a relation between individual tasks and the social world a user is inserted in.

3.2 Determining Who: First Level Prune

The first level prune tries to answer the following questions: given the universe of user acquaintances, which ones would the user be interested in keeping track of? We focus on ongoing collaboration, as shared work often benefits from awareness.

Taking the values of the *From, To, CC* and *BCC* fields, a full list of acquaintances is built. Senders and recipients determine the working groups that form a user's focus of interest: individuals co-occurring in messages (e.g. multiple recipients) form the social worlds a user is part of. There are certain rhythms to work, and activity within a social world ranges according to the need. Thus, a social world may be very active for a certain period of time and slow down after a certain point (e.g., project completion or reaching a milestone). Therefore, the system must check for the formation of new social worlds or change in activity patterns. We look for discrepancies between current behavior and "normal" behavior. Variables that currently characterize email exchanges are the number of messages exchanged (*Message Quantity*) and response time (*Response Time*). For each alter, we compare the current behavior to the normal behavior (the previously calculated average).

Given that patterns of email exchange usually emerge over a length of time, we are currently experimenting with different combinations of variables to determine the appearance of new working relationships. Intense message exchange usually accompanies ongoing collaboration, so Message Quantity is one of our qualifiers. We check if there are series of replies in a period of time shorter than the average, or whether there is an intensification of the exchanges (i.e., more messages are being exchanged than usual). Response Time should also be considered, as lower response time might mean a higher priority subject. We consider that social worlds in which the user is very active will be of more interest, with activity providing an indication of the focus of attention. It is important to note that social worlds are not defined only by a group of individuals, but also by the shared context that bring them together. This means that content analysis is needed to disambiguate interactions, defining the social worlds as a set of individuals with a shared theme, goal or project.

3.3 Second Level Prune: Determining What

After determining which social worlds are of interest to the user, other peers are queried for information about alters' ongoing activities. The determination of which activities are of interest to the user will be done using keyword matching, comparing the contents of the interactions with the contents of the documents relating to the ongoing tasks, so that only activities related to ongoing interactions are shown to the user. Hopefully, the first level prune will significantly reduce the list of acquaintances and, consequently, the number of peers that need to be contacted and the amount of information that will be exchanged and processed in this stage.

Each agent periodically queries the operating system to elicit its user's task list. It then analyzes the text relating to the tasks at hand to build keyword vectors to represent these. Our first approach is to build these using the TFiDF algorithm [27], which generates weighed keyword vectors given textual documents, and match these using the vector space model, where documents are matched using the cosine measure of proximity. Given that most of the activities under consideration are information processing tasks that involve a large amount of textual information (word processing, website surfing and searching, chat, etc.), this is a feasible approach, which should elicit activities that are related to previous conversations. Being established methods for information retrieval and matching, TFiDF and cosine measures have been extensively applied and tested, with good results. However, other text matching methods that may yield better results exist, and we will be experimenting with these.

In [3], a method for eliciting speech acts from email is presented and tested. It is based on a previously constructed taxonomy of speech acts applied to email (email-acts) describing verbs and nouns, with promising results. We hope to explore this approach as well, since it would provide better descriptions of activity information.

For now, we are keeping granularity coarse, picking only high level tasks. Thus, a user sees that an alter is editing a file they have exchanged, but not what paragraph or text has been changed. We are working on more fine grained analysis and display to enable the user to “drill down” into the peers’ tasks to obtain more information.

4 Preliminary Analysis

As we construct the system, we chose to build intermediary versions that would allow us to work with some of the assumptions and get user feedback to adjust our algorithms and approach. The current implementation performs structural email parsing, extracting senders and recipients, building graphs (sociograms) and keeping count of messages exchanged between individuals. No content analysis is performed. We built an interface to display the corresponding sociograms, with which we can explore temporal boundaries, data sources and cutoff points. With this we can interview users regarding the social worlds and how they relate to ongoing work and awareness needs. In this fashion, we were able to perform a few preliminary analyses and get user feedback before proceeding with system implementation.

4.1 Working Assumptions

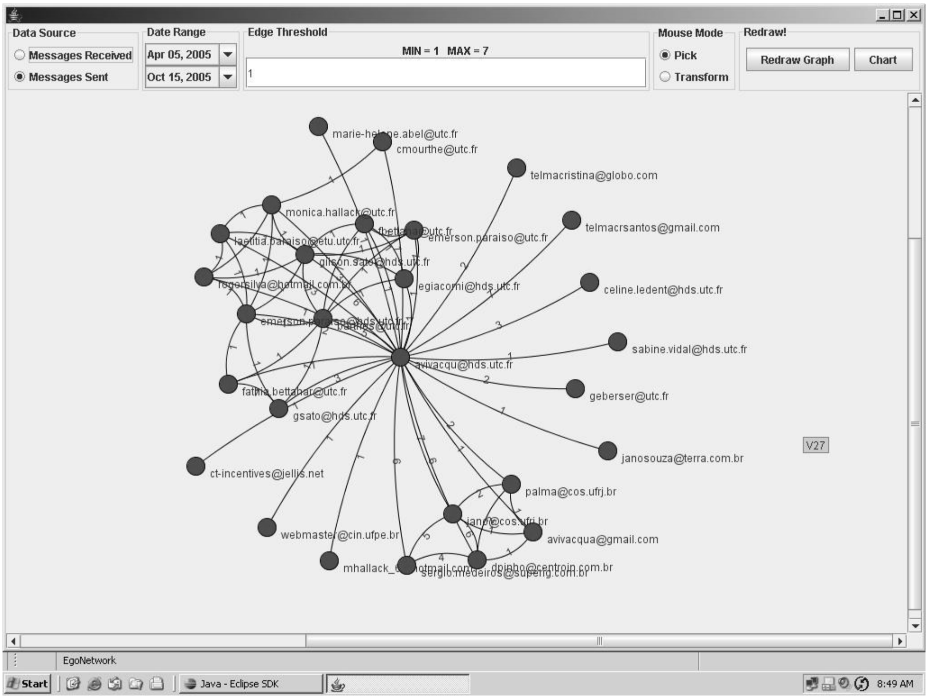
For these initial verifications, we were interested in working with four assumptions that underlie the system under construction. The first one is that social worlds are reflected in email. Thus, our first question was whether it was possible to identify social worlds through email based structural network analysis. What this means in practical terms is that cliques found in email-based social networks correspond to the different social worlds a user takes part in. If this is true, a system can infer working groups by identifying cliques in a graph (cliques are subgraphs where every node is connected to all others.) Our first assumption thus reads: *a clique represents a social world. For every clique in a sociogram, there will be a corresponding social world.*

Our second assumption is that activity patterns and social worlds change with time. These temporal patterns reflect the rhythms of group activity, from inception to project completion. By slicing the data into different timeslots, different social worlds should become apparent. This characterizes the changing patterns of collaboration a user typically engages in. The social worlds become stronger or fade away depending on the project dynamics. If this is true, by keeping track of these patterns, a system should be capable of adjusting the awareness needs of its user. We question whether *given different timeslots, different social worlds will become active;* and if *given a social world, it is possible to identify a pattern of intensification and decay in message exchange that corresponds to the activity in that social world.*

In [30], it is suggested that contents of the Outbox are more important than the contents of the Inbox in this type of analysis, since they reflect interactions the user has actually decided to engage in. Should this be true, the number of email messages to be processed and the resulting network would become considerably smaller, significantly speeding up processing time. However, Inbox contents cannot be completely discarded, since they contain valuable interaction information. Our third assumption is that a series of email messages is relevant only if a user has sent messages as well as received. It should be possible to construct a sufficiently elaborate social network to represent collaboration based on the interactions found in the user’s outbox, discarding messages from the user's inbox which have not been replied to. We assume that *if a message belongs to an interaction in which the user has not taken part, then it is of little importance to the user (it can be discarded).*

We wanted to reflect on two additional points: the first is whether short timeslots are significant for the identification of collaboration. In [24], a 15 year email log was analyzed, with data aggregated into 1-year slices. To be useful for the distribution of awareness information, this time frame needs to be significantly reduced (to days or weeks), since we are interested in collaboration at the moment it is happening. Thus our visualization tool allows us to slice time arbitrarily and check what sorts of patterns become visible, and if shorter periods (e.g. a month or two out of a 4-year email log) will display the same patterns as the ones found in the one year time slices. The second point we want to reflect on is the identification of thresholds. That is, how different does behavior have to be to be considered relevant to awareness needs? How many messages should be exchanged and how low a response time must be observed to characterize collaboration? This would help us determine how to detect ongoing collaboration on the fly. We would also like to verify how useful *Message Quantity* is as a qualifier and how well it ties into the determination of awareness needs.

To verify our assumptions, we built an interface that allows us to visualize data, slicing it into different timeslots and sources, shown in Figure 2. It implements a spring-embedded graph layout, using the Fruchterman-Rheingold force model [10],



Data Source: ○ Messages Received ● Messages Sent

Date Range: Apr 05, 2005 – Oct 15, 2005

Edge Threshold: MIN = 1 MAX = 7 — 1

Mouse Mode: ● Pick ○ Transform

Fig. 2. Visualization screenshot, where several cliques are visible

which treats a graph a set of nodes that repel each other connected by springs which attract them. The resulting graph reflects node proximity while minimizing line crossings. The visualization was built using the Java language and the JUNG library for graph construction and display. While all the aforementioned points have not been fully addressed yet, we are keeping them in mind: our current version allows us to slice time as desired, but new visualizations are needed to help analyze the data.

We processed 4 users' email histories and asked them questions regarding the resulting sociograms. We chose our users based on the fact that they were all heavy email users (with several thousand email messages in their mailboxes), but had different profiles, and we expected the data to follow different patterns. Our users came from different backgrounds, and the data they brought with them reflected as much: 2 were full time students, with many short projects and collaborations and a few longer collaborations with other students; 1 was a professor with several short and long term collaborations, some requiring close control, some not and 1 was a navy officer, with long and medium term projects requiring control and coordination. We explored the sociograms with our users, slicing the data in different ways. We took the opportunity to ask whether they believed that some sort of additional awareness information might have been beneficial to the work in progress. We asked users if:

1. the cliques they identified in their sociograms were related to projects or other collaborations going on at that time;
2. different cliques became active when the temporal range was changed;
3. patterns of message exchange reflected projects;
4. the social worlds in which the user had not participated (other than as an "observer") were of interest as far as peer task awareness.

4.2 Verification and Analysis

When asked, our users were capable of relating social worlds to the cliques that showed up in their sociograms. However, not all of these social worlds were related to ongoing work. There were situations where a clique represented a group that shared some sort of context (e.g., students in the same department), but were not in direct collaboration. There were large amounts of group emails exchanged for information only, but no actual collaboration going on. Thus, we confirm that it is possible to identify social worlds from cliques, but it doesn't follow that all these represent joint projects. Further investigation is necessary to determine how to differentiate between work and non-work messaging.

With changes in time slots, different groups became active, showing up on the visualization. It must be noted that, since the data was historical and cumulative, the social worlds don't actually disappear, they become more or less active (and a user became more or less active within them) according to the situation. While the full view was somewhat cluttered, slicing it to shorter periods considerably reduced the number of messages, making it easier to identify different subgroups. This confirms that social worlds come and go, which is reflected on email. Inspecting the temporal graph seen in Figure 3 (where time is sliced into daily email exchanges), we could easily see changes in interaction pattern. A dormant relationship suddenly springs to life, with emails being exchanged daily (sometimes several messages a day, depending on the urgency), and then dies out as abruptly when deadlines are reached. This is confirms our second assumption, and is particularly interesting since we were able to explore considerably shorter periods than those presented in [24] and still detect collaboration. However, changes were often quite abrupt, going from no interaction to 4 messages a day overnight. While we expected this to happen, we also hoped to see softer patterns, where interactions would gradually increase with time.

When asked, users said that social worlds in which they did not actively participate were not of much interest to them (as far as task awareness). Users wanted to be aware of their closer, more immediate collaborators, where there was a lot of coordination to be done. They had no desire to be aware of everybody's work, although in some cases they would like to remain superficially aware of what was going on. This indicates that, for task awareness purposes, we can leave out all incoming threads in which the user has not participated. In computational terms, this significantly reduces graph size, and, consequently, memory needed and computation time.

Within the emails, there were several instances of project-related social worlds, usually qualified by intense interaction in a shorter period of time (weeks or a few months). This suggests a way of more effectively picking activity-related groups. However, when inspecting the data, it became apparent that structure alone was not sufficient to tease these apart, especially when there were overlapping social worlds. These needed to be qualified according to the activities or themes of the interactions, so that they could be effectively set apart. There were quite a few overlapping social worlds (including temporal overlaps, where a group works together on more than one project at the same time). Within our data sets, there were also several social worlds embedded in other social worlds. Large groups who perhaps work in a same building and smaller subgroups who work closely together. While a user will probably not be interested in keeping close track of the activities of members of the larger group, he or she may want to have periodic summaries or reports on how work has been progressing. This leads us to think of awareness as a continuum, with awareness needs tied to a user's participation in a group. The user might desire to have more or less information (depth and frequency) about others, depending on his or her level of involvement with the group. We are considering the use of artificial intelligence techniques such as fuzzy sets to better represent a user's focus, and how much information he or she would want to have.

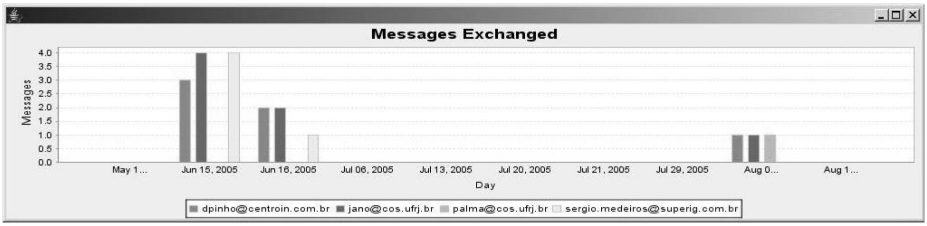


Fig. 3. Time based interaction graph

Message Quantity was a reasonable qualifier when looking at a user's outbox, but not at all when inspecting an inbox. Some alters sent 200+ messages over a 6-month period and were neither collaborators nor of any interest to the user. A user's outgoing messages, however, seemed to reflect the social worlds a user engaged in more accurately. Participation in conversations involves an investment of time and effort that indicates a certain level of interest and commitment to the group. Accordingly, we are changing our message processing algorithms to process outgoing messages first, building social worlds and then trying to fit incoming messages into these.

Users often engaged in animated discussions which were not work related. While this does denote a certain level of interest in the subject (our users were actually interested in what was going on), it doesn't mean the user would want to keep track of others' work. For instance, one of our users had quite a few discussions with a group of friends regarding TV Series, politics and movies. While the content of these interactions would not match any ongoing tasks, it might still be a costly false positive, which increases the search space. We are refining our algorithms to disregard these threads from the start. One possibility would be to perform an initial match with the user's own work to see if the user was actually working on the subject.

5 Discussion and Future Work

Lack of space precludes a lengthy discussion of related systems. Messengers in general have been widely adopted and have become a frequent means of communication. Most provide ways for a user to express whether he or she is available, busy or "out for lunch", passing that information on to their peers. A few more complex task awareness approaches exist: for instance, MultiVNC [15] displays miniatures of peers' desktops in order to improve awareness in a working group and increase collaboration. It doesn't filter or verify what is actually of interest to the user, and interface is quite busy. Community Bar [20] allows users to specify what peers they want to be aware of, organizing them into social worlds. Their focus is on media items (webcam shot, calendar, post-it, chat software, etc.), not content, so users tell the system what media they want keep track of and the user is left to sift through the information contained therein and decide which are valuable. Doc2U [22] is a shared editing environment where information about who is editing which parts of a shared document is distributed among peers. It requires logging in to a shared environment and is only applicable to one activity. A number of agent-based systems have been created to provide information that the user might otherwise not have had. Many also deal with the problem of the overwhelmingly large amount of information available at any given moment by sorting out what is useful to the user at the time [17].

A number of systems to classify emails into activities have recently been developed [4], and activity modeling has been growing as a research area. Unified Activity Modeling, for instance, proposes a generic model of activity and a framework to integrate individual, informal, work with more strict organizational workflows [21]. This research seeks to help users organize and contextualize emails within activities, and might be useful in our context as well: through an accurate classification of emails into tasks, it becomes easier to determine the activities within a social world, which should then lead us to appropriate information dissemination.

A method to construct networks of people and keywords and for discovery of people with similar interests is presented in [19]. It mines email data and constructs networks of people, which can later be used to determine who has knowledge on what topics, displaying the networks for user inspection. Another approach for social network use is presented in [12], where the author uses networks to locate individuals with a certain expertise and availability through an analysis of their activities and tasks. In the aforementioned approaches, the emphasis is on finding experts, and navigating the social network to create an awareness of who knows what. Our interest, on the other hand, is quite distinct: we aim to identify working groups and to provide activity awareness information only of these individual, as it happens. It is not meant as a system for group formation or expertise location, but as tool to assist coordination and collective action. Our intent is to use Social Networks as an active way for making inferences, monitoring and influencing collaboration, as suggested in [18]. Our emphasis is not in the display and visualization of the networks, but in what patterns can be found and what calculations and inferences can be made. In [6], a series of patterns and work rhythms are presented, we plan on building on this work to determine what these mean in terms of information needs and distribution.

Our system is currently under implementation, and at this point, this approach seems promising: it provides a way to explore awareness needs of individuals in relation to their ongoing collaborations. Email-based analysis can elicit interaction patterns that denote role attribution or the organization of a team. We expect these will have different information needs (e.g., core vs. periphery members differ in terms of nature, quantity and depth of the information needed), and further research is needed.

The system is being built using the Java language, with several specific open source libraries: so far, JUNG has been used for graph construction and display and JFreeChart for the time charts. For the following phases, we plan on using Java and JNI to monitor users' ongoing tasks (this information can be obtained directly from the Windows operating systems APIs) and JACOB, a Java library to interface with COM automation (present in all office applications and many others), to communicate with Windows-based applications. Emails are stored in an Access database.

When managing a few thousand emails, the system becomes a bit slow especially when drawing, something we are trying to work around. It currently reads Eudora mailboxes, but we are already checking on other possibilities, such as reading directly from the server. Another difficulty was dealing with raw data: in general, our users' email files were fairly disorganized and sometimes contained duplicate messages. Additionally, several individuals had more than one email address, which means they must be organized into personas so that the data makes more sense.

We will continue to explore the interplay between interaction and awareness needs. Even though our preliminary analysis was small, with only a few subjects, it indicates some directions for further research: to develop a more complete mapping between interaction levels and awareness needs, other variables need to be taken into account, such as response time and content. New experiments need to be designed, with more users and different emphasis, so that other information can be gleaned from the data. One of our next activities will be a controlled experiment to check on the effects of different types of information at different moments.

5.1 Privacy Issues

Whenever information is automatically collected or distributed, privacy becomes an issue. The automatic management of a user's nimbus is an open issue at this point, although we are experimenting with network based calculations for that as well. For the time being, we leave the choice of what to make available to the user. We are adopting a three-tiered privacy scheme, where a user can define whether a task is public (all can see), protected (some can see) or private (none can see). The user will be able to determine alters, keywords, or resources that fall within each of the tiers, and who has access to what in the protected level. When a task is found that should be propagated to other peers, it is checked against the specified restrictions to see if it falls within a specific privacy tier and whether it can be sent to the requesting agent.

We are currently fitting users' activities into one of the following categories: manipulation of shared objects, manipulation of non-shared objects and chat between members. We are working with the assumption that all shared objects and interactions within a social world should be made public to members of that social world. For instance, editing or forwarding a file that has been sent around as an attachment, or chat related to the project between members of the social world. Manipulation of non-shared objects is a more complex case. For our initial prototype, we prefer to err on the side of caution and block all non-shared material. These simple heuristics should help us decide on whether to send information around until a better privacy scheme is in place. Upon reflection, this transparency might compromise the capability of political articulation within a group, so we expect some reaction from users.

When we look beyond organizational structures, protocols and hierarchies, modern organizations are composed of networks of interacting actors [1]. More often than not, knowledge is exchanged and work is undertaken through these informal relations between workers, in networks that cut across departmental, functional and organizational boundaries. Thus, modern organizations require coordination and integration of activities across these boundaries, and information systems should provide support for distributed coordination and decision-making.

In this paper we have presented an approach to the determination of awareness foci based on egocentric email-based social network analysis. We believe this is a promising line of research that holds many possibilities for further work. Many studies have applied social network analysis to uncover relations between people and patterns of interaction, but few have used these patterns as a basis for a system to actively assist the user, choosing only to display this information.

Acknowledgements

This work was partially supported by CAPES and CNPq.

References

1. Bernoux, P. La Sociologie des Entreprises. Éditions du Seuil, Paris (1999)
2. Carstensen, P., Schmidt, K. Self Governing Production Groups: Towards Requirements for IT Support. In 5th IFIP Int. Conf. on Information Technology in Manufacturing and Services (BASYS'02), Cancun, Mexico, Kluwer Academic Publishers (2002) 49-60
3. Cohen, W.W., Carvalho, V.R., Mitchell, T.M. Learning to Classify Email into Speech Acts. In Proc. Conf. on Empirical Methods in Natural Language Processing (2004)

4. Dredze, M. Lau, T., Kushmerick, N. Automatically Classifying Email into Activities. In Proceedings of Intelligence User Interfaces (IUI 06), Sydney, Australia (2006)
5. Edwards, K., Mynatt, E. Timewarp: Techniques for Autonomous Collaboration. In Proc. CHI 1997, Atlanta, GA (1997)
6. Fisher, D., Dourish, P. Social and Temporal Structures in Everyday Collaboration. In Proc. CHI 2004, ACM Press (2004) 551-558
7. Fitzpatrick, G., Kaplan, S., Mansfield, T. Applying the Locales Framework to Understanding and Designing. In Proc. OzCHI 98. Australia, IEEE (1998) 122-129
8. Fitzpatrick, G. The Locales Framework: Understanding and Designing for Cooperative Work. PhD Thesis, University of Queensland (1998)
9. Fitzpatrick, G., Tolone, W., Kaplan, S. Worlk, Locales and Distributed Social Worlds. In Proc ECSCW 95. Stockholm, Sweden, Kluwer Academic Publishers (1995) 1-16.
10. Fruchterman, T., Reingold, E. Graph drawing by force-directed placement. Software: Practice and Experience, 21(11), John Wiley and Sons (1991) 1129–1164
11. Greenberg, S., Johnson, B. Studying Awareness in Contact Facilitation. In CHI 97 Workshop on Awareness and Collaborative Systems. Atlanta Georgia (1997)
12. Groth, K. Using Social Networks for Knowledge Management. In Proc ECSCW 03 Workshop on Moving From Analysis to Design: Social Networks in the CSCW Context. Helsinki, Finland (2003)
13. Gutwin, C. Greenberg, S. The Importance of Awareness for Team Cognition in Distributed Collaboration. In Team Cognition: Understanding the Factors that that Drive Process and Performance. Washington, APA Press (2004) 177-201
14. Gutwin, C., Greenberg, S. A Descriptive Framework of Workspace Awareness for Real-Time Groupware. In Computer Supported Cooperative Work 11. Kluwer Academic Publishers, Netherlands (2002) 411-446
15. Gutwin, C., Greenberg, S., Blum, R. Dyck, J. Supporting Informal Collaboration in Shared-Workspace Groupware. The Interaction Lab Technical Report HCI-TR-2005-01, University of Saskatchewan, Canada (2005)
16. Hertel, G., Geister, S., Konradt, U. Managing Virtual Teams: A review of current empirical research. In Human Resource Management Review 15. Elsevier (2005), 69-95
17. Maes, P. Agents that Reduce Work and Information Overload. In Communications of the ACM Vol. 37, No. 7. ACM Press (1994) 31-40
18. Martinez, A., Dimitriadis, Y. Tardajos, J., Velloso, O., Villacorta, M. Integration of SNA in a Mixed Evaluation Approach for the Study of Participatory Aspects of Collaboration. In Proc ECSCW 03 Workshop on Moving From Analysis to Design: Social Networks in the CSCW Context. Helsinki, Finland (2003)
19. McArthur, R. Bruza, P. Discovery of Social Networks and Knowledge in Social Networks by Analysis of Email Utterances. In Proc ECSCW 03 Workshop on Moving From Analysis to Design: Social Networks in the CSCW Context. Helsinki, Finland (2003)
20. McEwan, G., Greenberg, S. Community Bar: Designing for Awareness and Interaction. In ACM CHI Workshop on Awareness Systems: Known Results, Theory, Concepts and Future Challenges (2005)
21. Moran, T.P. Unified Activity Management: Explicitly Representing Activity in Work Support Systems. In Proc. ECSCW 05 Workshop on Activity: From a Theoretical to a Computational Construct. Paris (2005)
22. Morán, A. L., Favela, J., Martínez-Enríquez, A. M., Decouchant, D. 2002. Before Getting There: Potential and Actual Collaboration. In Proc. CRIWG 02 Springer-Verlag (2002)
23. Narine, T. Leganchuk, A., Mantei, M., Buxton, W. Collaboration Awareness and its use to consolidate a Disperse Group. Proc. of Interact 97 Sydney Australia (1997)

24. Perer, A., Shneiderman, B., Oard, D.W. Using Rhythms of Relationships to Understand Email Archives. In Email Archives Visualization Workshop
25. Pinelle, D., Gutwin, C. A Groupware Design Framework for Loosely Coupled Groups. In Proc. ECSCW 05, Paris, France (2005)
26. Rodden, T. Populating the Application: A Model of Awareness for Cooperative Applications. In Proc. CSCW 96, ACM Press (1996) 87-96
27. Salton, G. Automatic Text Processing: the Transformation, Analysis and Retrieval of Information by Computer. Addison-Wesley Publishing (1988)
28. Schmidt, K., Bannon, L. Taking CSCW Seriously: Supporting Articulation Work. In Computer Supported Cooperative Work 1, vol. 1. Kluwer Academic Publishers, Netherlands (1992) 7-40
29. Scott, J. Social Network Analysis: A Handbook. Sage Publication, London (1991)
30. Tyler, J., Tang, J. When can I expect an Email Response? A Study of Rhythms in Email Usage. In Proc. ECSCW 03 (2003)
31. Wasserman, S., Faust, K. Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge US (1994)