Creating a model of the Dynamics of Socio-Technical Groups using Electronic Trace Data

Sean P. Goggins, Giuseppe Valetto, Christopher Mascaro, Kelly Blincoe Drexel University

ABSTRACT

Individuals participating in technologically mediated forms of organization often have difficulty recognizing when groups emerge, and how the groups they take part in evolve. This paper contributes an analytical framework that improves awareness of these virtual group dynamics through analysis of electronic trace data from tasks and interactions carried out by individuals in systems not explicitly designed for context adaptivity, user modeling or user personalization. We discuss two distinct cases to which we have applied our analytical framework. These two cases provide a useful contrast of two prevalent ways for analyzing social relations starting from electronic trace data of either artifact-mediated or direct person-to-person interactions. Our case study integrates electronic trace data analysis with analysis of other, triangulating data specific to that application. We show how our techniques fit in a general model of Group Informatics, which can serve to construct group context, and be leveraged by future tool development aimed at augmenting context adaptivity with group context and a social dimension. We describe our methods, data management strategies and technical architecture to support the analysis of individual user task context, increased awareness of group membership, and an integrated view of social, information and coordination contexts.

Keywords

Activity awareness, group awareness, virtual groups, communities of practice, networks of practice, task context

1. INTRODUCTION

User modeling and personalization have lately focused attention on building systems that adapt automatically to their context of use. A great deal of that research closely examines individual user characteristics and preferences, and describes algorithms and approaches for applying that user information to increase personal engagement, or improve effectiveness and user experience within computing and information systems. Despite the fact that most user experience and technologically-mediated context involves groups of people interacting with each other, often around artifacts, little research to date models or emphasizes units of social organization as a major component of user personalization or context adaptivity.

Social organization is not embodied by a single socio-technical system; research programs that focus on one system lack the empirical data to provide insight and guidance for generalized modeling and personalization algorithms. Because we have conducted empirical research modeling social contexts in political discourse, software engineering, online learning, recreational sports, and online dating, we are able to present a description of research informed not just by the concrete cases in a single paper like this one, but by years of multi-domain empirical study. The cases we present break through single system studies of user behavior to present a broader perspective. We show that the time distance between interactions, for example, powerfully influences adaptivity of users in online learning. The shorter the time between interactions, the more knowledge is constructed. In contrast, we show that time distance is virtually meaningless in the analysis of software engineering teams. We focus this paper's empirical contribution on these two cases, as they provide the sharpest contrast.

These social interaction outcomes are important because people working in a technological context are also situated in a social context that is not central to current user modeling or personalization. User interactions in these socio-technical contexts are an implicit record from which units of social organization can be understood, if the interactions are analyzed from multiple perspectives. For example, Maloney-Krichmar & Preece (2005) used a broad two and a half year long ethnographic study to explicate how online communities develop, form subgroups, share information and behave socially. They

"This paper or a similar version is not currently under review by a journal or conference, nor will it be submitted to such within the next three months. This paper is void of plagiarism or self-plagiarism as defined in Section 1 of ACM's Policy and Procedures on Plagiarism."

learned that strong subgroups within a larger community make substantial contributions to sustaining and developing the overall community's purpose and vitality. Since such group participation certainly informs the stance, activity and practices of users, their group participation is an integral part of their user profile. Still, little work to date integrates methods such as those used by Maloney-Krichmar & Preece with generic user modeling and systems like those described by Kobsa (2001, 2007) and others.

Burt and Conati (2003), like Maloney-Krichmar and Preece (2005), describe the limitations and hampering effects of systems designed for research on user modeling and personalization. Our goal is not to understand context adaptivity in a single, laboratory system, but to understand first, and analytically, how group context and task context can be incorporated into existing systems and lead to improvements in user modeling and personalization. Systems in wide use provide what are sometimes considered unmanageable volumes of interaction data (Cohen, Dolan, Dunlap, Hellerstein, & Welton, 2009; Lynch, 2008). But in the case of research focused on group awareness and the integration of social constructs into user modeling and personalization, these large stores of electronic trace data are in fact an opportunity, and an invaluable source for study, modeling and evaluation. Our approach enables researchers to leverage such systems and their data in a theoretically and methodologically coherent manner, which places the differences of each socio-technical contexts in the forefront. We analyze these systems using a model of Group Informatics, which we have defined iteratively, through empirical studies reporting on 16 different sets of electronic trace data (Blincoe, Valetto, & Goggins, 2012; Goggins, 2007; Goggins, Galyen, & Laffey, 2010a; Goggins, Laffey, & Tsai, 2007; Goggins, Mascaro, & Mascaro, 2012a; Mascaro & Goggins, 2011; Goggins & Erdelez, 2010).

1.1 Analytical Framework and Cases

This paper presents an analytical framework that: 1) consumes electronic trace data from individual participation of users in large computer-mediated contexts that are operational in the field; 2) extracts elements of individual context from that trace data; 3) triangulates that individual context with other data and analysis techniques specific to each domain or application; 4) analyzes the accreted information to resolve group formation, participation and evolution; and 5) associates the resulting context information with emergent groups to represent the "group context".

The analytical framework outlined above is the principal contribution of this paper. Its insights originate from empirical work on data collected from real-world socio-technical systems. To contribute a social, analytical framework to context adaptivity, we have worked our way up — and abstracted from — distinct experiments conducted with a set of techniques for capturing and analyzing group formation from traces of activities and interactions in those systems. Although the requirements of our context analysis vary with the different domains considered, leading to the usage of an array of qualitative and network analytic techniques, we show how those techniques fit in a general model of Group Informatics, which we define in Goggins, Mascaro, Valetto & Gallagher (Goggins, Mascaro, & Valetto, 2012), and we briefly review here in Section 2.

This paper applies that Group Informatics Model, through the analytical framework articulated in this paper, to two cases in domains with distinct characteristics: asynchronous online learning with context awareness tools, and open source software development. The analysis we present here demonstrates how context adaptivity can be accomplished by expanding personalization and user modeling to include membership and participation in groups. Our work is at the intersection of social navigation (Dourish, 2003; Konstan & Reidl, 2003), shell systems (Kobsa, 2007) and other research that contributes to an understanding of user modeling and personalization that takes social context into account, and will lead to context adaptive systems that provide group awareness in a range of contexts. Methodologically, we advance the vital construct of user data resolution, which Kay (1995) first defined and Kobsa (2001, 2007) later advanced. While accretion describes the accumulation of user data, resolution of user data involves its interpretation. Our use and analysis of accreted data achieves enhanced resolution — grounded in the context data associated with users' actions and interactions. This approach enables the identification of emergent groups who may not even be aware of their groupness (Goggins et al., 2007).

Awareness systems have been studied extensively, with the goal of facilitating collaboration (Carroll, Rosson, Convertino, & Ganoe, 2006; Carroll, Rosson, Farooq, & Xiao, 2009; Carroll, Neale, Isenhour, Rosson, & McCrickard, 2003). To build social context into context adaptivity, we leverage concepts of social matching first outlined by Terveen and McDonald (Terveen & McDonald, 2005) to analyze interactions between users. We also recognize how the analysis of interactions can benefit from the characterization of the collaboration of users around a task context. In the cases we present, we show that the context of user interactions (e.g. social behaviors within a discussion thread) and the context of task interactions (e.g. artifact-oriented actions of users producing a knowledge-intensive product) are equally important and equally conducive to group awareness, even if they call for different analysis techniques. Our software engineering case is focused on task context and relations between users and artifacts, while the online learning case is focused on direct user-to-user interactions.

Our investigation of these two domains provides an example that future researchers and tool developers can build upon, while our analytical framework captures and allows reconciliation of those differences, and can help guide the researcher and technology designer to make choices about the types of interaction that are most appropriate in the specific application domains. Our model of Group Informatics and the analytical framework presented here can be implemented in future systems, to offer group awareness and group context as a new dimension of user modeling and personalization that builders of shell systems, modeling servers, and social navigation sites can leverage.

In the rest of this paper we first provide an overview of the model of group informatics, which guides our analysis, followed by a synthesis of relevant literature. We then describe our research methods and present results from our two case studies. We conclude with a discussion of our field studies and method, and their implications for the design of context-adaptive systems that make group awareness and task context a more central part of user modeling and personalization.

2. A MODEL OF THE DYNAMICS OF GROUP INFORMATICS

Models have been widely used to capture and organize salient traits of personalization information accrued from technologically mediated contexts, especially from the individual user perspective (Anaya & Boticario, 2011; Bunt & Conati, 2003; Cosley, Ludford, & Terveen, 2003; Kobsa, 2001; Kobsa, 2007; Zimmermann, Specht, & Lorenz, 2005). Our fieldwork has led us to define a model of Group Informatics (Goggins et al., 2012).

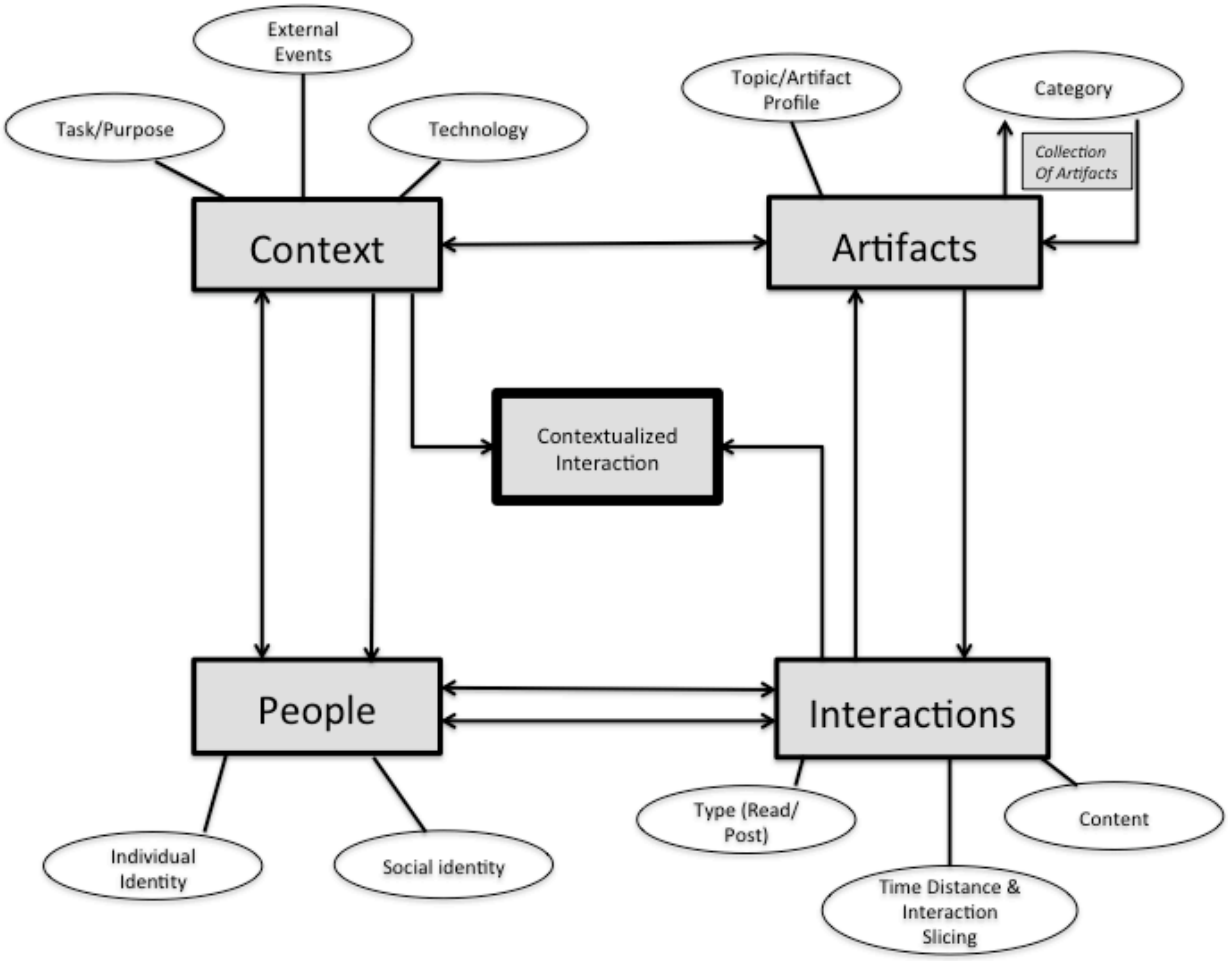

Figure 1 - Model of Group Informatics Research

Our Group Informatics model is depicted in Figure 1. It shows relationships between our principal units of analysis, the technologically-mediated interaction, and the two main dimensions of socio-technical systems: people and artifacts. The context component, which is a general model in this figure, and a context specific implementation in figure 2 (section 4), recognizes that there are factors – both internal and external to socio-technical context, that lead to the construction and persistent evolution of the user's experience of context. Both Nardi (Gonzalez, Nardi, & Mark, 2009; Kaptelinin & Nardi, 2006; Nardi & O'Day, 2000; Nardi, 1996; Nardi, 2007; Nardi, Whittaker, & Schwarz, 2002) and Dourish (Dourish & Button, 1996; Dourish, 2004; Dourish, 2006; Harrison & Dourish, 1996; Dourish, 2001) have written extensively on socio-technical theories related to the slippery notion of context. Our Group Informatics model is informed by this theoretical work. For example, when determining how we construct and understand context, we must describe the different decisions we make as researchers when applying the Group Informatics model. In Sections 4 and 5 below, we discuss how the model is applied to the two cases in this paper. Our goal is not a canonical model of context adaptivity, but the application of a Group Informatics model informed by five years of empirical work, and based upon decades of empirical and theoretical development in socio-technical research and user modeling and personalization.

The central part of the Group Informatics model is the "Contextualized Interaction". This is the locus in the model where qualitative research that takes place around the four main components of the model, context, artifact, person and interaction, are made operational. User profile information, interaction types, interaction frequency, time and the role of artifacts are examples of contextual attributes that can be accreted from interaction traces within the socio-technical context, attached to those interactions, and can then contribute to construct a representation of group and task context. Such a representation includes relations between people: for example,

in the online learning case, the interactions are various forms of discourse around ideas in discussion forums, wikis and archived chat. In the open source software engineering case, instead we look at software developer interactions with artifacts of the software product, since they inform and constrain the developer-to-developer relations. From these two cases, the reader can begin to see, conceptually, that "contextualized interactions" in figure one are the component of the model where context adaptivity begins. Our study of diverse systems in the light of this common model helps to illustrate the role that complementary, qualitative analysis of the socio-technical context can play in the setup and configuration of context adaptive tools (see Section 4, and figure 2).

2.1 Model Validation

Validation of social user modeling and personalization technology must occur against real world data, in real system contexts, and leverage methods grounded in both social science and computer science scholarship (Crowston, Wei, Li, & Howison, 2006; Crowston, Wiggins, & Howison, 2010; Howison, Wiggins, & Crowston, 2012). In each of the empirical studies we conducted, leading to the explication of our model of the dynamics of group informatics, we performed context relevant validation using empirical research methods that reflexively move back and forth between data, analysis and insight. Though these inductive methods are common in the social sciences, they are just now being recognized as a vital mechanism for validating the real world applicability of algorithms developed through computer implementation, formal analysis or laboratory experiments.

In the development of this article, we revisited data from the prior published work we refer to throughout and performed a consistent, systematic process of triangulating our social science and electronic trace data to validate our model against the results of our prior work. This process is described in section Four. In each case, without the application of our reflexive, integrated process of informing log analysis with qualitative coding and analysis of communication in each context, the statistical analyses result in networks and clusters that users do not recognize as part of their experience; the technical associations diverge from the social affiliations without our method (or, we suspect, other methods that take an integrative approach).

For example, in (Goggins, Laffey, Amelung, & Gallagher, 2010b) we describe how we comparatively evaluated a data mining algorithm based on a combination of Pearson correlation coefficients and a Jaccard distance measure to identify participant clusters and topical clusters with the network analytic techniques our model uses. We then compared the results of each algorithm with two interviews conducted with each of 42 informants who were members of seven different online groups. In the first interview, which occurred mid collaboration, we asked them to identify which individuals they collaborated with most closely on work tasks in the groups we identified through each method. In the second interview, we asked them which clusters – those from the data mining algorithm or the model based network analysis of electronic trace data – most closely represented the social structures in their online course. Over 90% of informants (39 out of 42) chose the network analytic clusters as more representative of their social experience than the data mining focused clusters. The electronic trace data used was the same, but the models and methodological approaches to identifying groups were different.

In another exemplar study validating our model of the dynamics of group informatics (Goggins, Laffey, & Gallagher, 2011b) we triangulated the groups and leadership identified through our model with student reports of who they associated with, and content analysis explicating the types of interaction. For example, we showed that interactions with a high degree of knowledge construction correlated with more cohesive subgroups identified through our application of the model and corresponding network analysis.

3. BACKGROUND AND RELATED WORK

3.1 User Modeling and Personalization

One of the great challenges for user personalization and modeling researchers seeking to incorporate a social dimension into their work is the depth and breadth of research examining socio-technical phenomena in different, but related scholarly communities. Our work emerges from a center that is more group focused, and socially focused than user modeling and personalization focused. It is through the user modeling and personalization literature, however, that we have identified ways to turn our work from understanding to action. In this section, we integrate the most salient literature that led us to model units of social organization as a major component of user personalization or context adaptivity.

Collaborative filtering systems are one example of a literature that focused heavily on the discovery of common user traits to drive what one might alternately refer to as "social context adaptivity". Cosley et al. (2003) described the challenge of discovering like-minded people in a large sea of individuals in online contexts. To build understanding of these phenomena, he structured a study using a "family feud" style online game for an experimental study, which yielded several findings. First, people with the same interests did not demonstrate differences in their ability to find information. Second, people with similar education levels demonstrated more similarity in existing knowledge, and did not demonstrate productive, complementary interests when answering diverse types of trivia questions. These are both challenges for social recommender systems because the recommender engine does not always know these traits but they have a material effect on outcomes. Social recommendation systems, user models and personalization that do not incorporate both user profile and user interaction data are, therefore, limited.

Work derived from the user modeling and personalization literature identifies gaps that our work, focused on modeling units of social organization as a component of user modeling and personalization adapt. With the rest of this section, we review prior user modeling and personalization research that informs and motivates our more socially focused analytic work in the community. Kobsa (2007, 2001) defined generic user modeling systems, which are a primary focus of the user modeling and personalization literature. His works highlight shell systems, user modeling servers and the attributes considered important for user modeling systems over the years. These technologies and their application suggest possible limitations that call for incorporation of interaction data, and its analysis, in user modeling and personalization. Interactions among users have not been central in the history of user modeling and adaptation. Some work in the field has defined new and improved ways for deriving personalization information from these data stores at the individual unit of analysis (Bunt & Conati, 2003) or by incorporating basic, individually focused measures of user interaction into the construction of context for user modeling and personalization (Zimmermann et al., 2005). Prior efforts recognize that interaction data has a role to play in user modeling and personalization, but continues to approach this data primarily from the individual user perspective. Reconsidering user modeling and personalization as a valuable research field where interactions between users play a more central role in the construction of personalization tools is therefore a promising area of inquiry.

Bunt & Conati (2003) focus on the application of Bayesian network analysis in open online learning systems, describing individual learning activities and progress. The goal of their study is to understand how to encourage and motivate learner exploration in an online learning context. The type of skill being modeled and the need for a better definition of exploratory behavior in these systems, prior to modeling, are each identified as important considerations. They also found that learners consistently overestimate the amount and quality of exploration they perform in online learning contexts, and thus propose solutions that include encouraging learners to explain their exploratory decisions, and for researchers to then factor these decisions back into the model used to encourage further exploration. In the end, they propose natural language processing as a natural extension of their work, arguing that computational linguistics offers a solution for improving models of exploration in learning. Related research focuses on the social nature of learning (Amelung, 2007; Bandura, 1977), and suggests that encouraging student social behavior in online systems could produce alternative designs and system implementations that are competitive with those possible through natural language processing. Systems focused on providing context information to learners to increase social learning (Goggins et al., 2010b; Goggins, Laffey, & Amelung, 2011a; Laffey, Amelung, & Goggins, 2009) or content-free collaborative modeling of online learning (Anaya & Boticario, 2011) are examples of such alternatives already being pursued.

In a recent user modeling and personalization study, Anaya and Boticario (2011) demonstrate a data mining approach that does not look at content specifically, as recommended by Bunt & Conati (2003), but instead focuses on detecting when collaboration has occurred, without reference to content. They demonstrate that tools can help students and teachers better manage collaboration by increasing awareness of the level of collaboration occurring, and whether or not specific contributions of students received responses. Conversations started, messages sent and replies to messages were used to measure collaboration in their study. For each of these measures, they defined four variables: number, average, variance and a ratio of the number of threads started compared to the

total overall contributions of students. These measures provide a high-level understanding of the degree of collaboration taking place in an online context with a high-degree of task structure, in their case, an online course.

3.2 Extending User Modeling and Personalization By Focusing on Interactions

The approach we take in our empirical work is most similar to that of Anaya & Boticario, but we propose three main extensions. First, we examine networks of people, identifying who is working with whom, while Anaya & Boticario focus on the amount of interaction overall. Instead of focusing on this individual behavior and comparing people with each other, we focus on statistics derived from discrete *interactions* between people, either directly, or through the mediation of work artifacts. This kind of analysis will help with the identification of group awareness, and the important aspect of group emergence in an online context. Second, we focus on contexts that have been effectively employed to support real-world endeavors, whereas Anaya & Boticario acknowledge that a significant limitation of their work is that the forum tool they used for their study hindered, instead of promoting, collaboration[1]. Third, our approach makes qualitative inputs to the model — derived from analysis by researchers — a primary practice.

Zimmerman, Specht & Lorenz (2005) focus on the integration of personalization research with contextualization information, adapted to the needs, goals and knowledge characteristics of available data in online contexts. Their view, which is consistent with our approach, is that contextualization complements personalization, so that environmental states can be taken into account. We extend this notion, by also focusing on social groups that emerge in online context as a critical component for fully understanding the context and environment of the user, augmenting other context data, and ultimately contributing to context adaptive systems in the future. Unlike Zimmerman et al (2005), we examine the interactions of users with each other and the artifacts in the system to different degrees, depending on the domain of the virtual context.

Terveen and McDonald (2005) look at interactions between users as a central component in a construct they describe as social matching. We regard this social matching research to be complementary to context-adaptive systems development. Firstly, issues of trust, reputation and interpersonal attraction identified by Terveen & McDonald (2005) can be useful considerations in the development of context-adaptive systems, since they contribute to personalization. Secondly, by framing social matching as building upon prior work in user modeling, group recommenders, online community studies, awareness systems, social visualization systems and social navigation systems, the breadth and complexity of context adaptivity overall is brought into focus.

One of the central features in social matching is the introduction of users to one other, but that plays a less central role in our model of Group Informatics applied to our cases. First, our cases describe group emergence and group structures already taking place in the virtual contexts we observe. Second, our approach regards the tasks that users perform as a key dimension of context, which limits motivational gaps as an area of study in the cases, while that is a primary concern in discussions of social matching.

3.3 Framing the Study of Emergent Groups as *Group Informatics*

Group Informatics is principally concerned with the emergence and development of small groups within larger socio-technical contexts, which may be conceptualized as communities of practice, networks of practice or, more broadly as technologically mediated social structures. In the Group Informatics model, individual relationships are implicit in the occurrence of an interaction between two people, made visible via electronic trace data. Technologically mediated groups are studied as networks (Brown & Duguid, 2000), communities of practice (Wenger, 1998), groups (Goggins et al., 2007; Rohde & Shaffer, 2003; Rohde, Reinecke, Pape, & Janneck, 2004) and individual relationships (Granovetter, 1985). Like Mitchell (Mitchell, 1969), we identify the relationship between these different organizational structures as existing on a continuum that is discernable through comparative studies of social network characteristics, such as density and size.

The Group Informatics model facilitates synthesis of concepts related to user modeling and personalization with studies of artifact-focused as well as conversationally-focused forms of socio-technical interaction. The

[1] It is common that tools that are conceived specifically for gathering high quality research data are limited for practical use

interaction is central, and can be conceptualized as between people, or people and artifacts. If artifacts are conceptualized as boundary objects (Lee, 2007; Star & Griesemer, 1989) around which interactions occur, then the Group Informatics model offers a viable and compelling approach for either person-to-person or person-to-artifact interactions derived from electronic trace data. The case studies presented here and the analytical framework we describe are an early, but, we argue, important waypoint in the development of context-adaptive systems, premised on the notion that context is derived dynamically, through user interaction in domain specific systems as initially argued by Dourish (2004).

4. METHOD AND RESEARCH QUESTIONS

User context is constructed through interaction in a socio-technical context, but every single context has a different specific purpose, different combination of people, actions and interactions, and different technology support. The goal of our analytical framework, built on our previously articulated model of Group Informatics, is to turn electronic traces of interaction into evidence of participation in emergent groups, augmented with data that represent the context relevant to the group as a social unit operating in the online context. The steps for that transformation process are explicated in Figure 2:

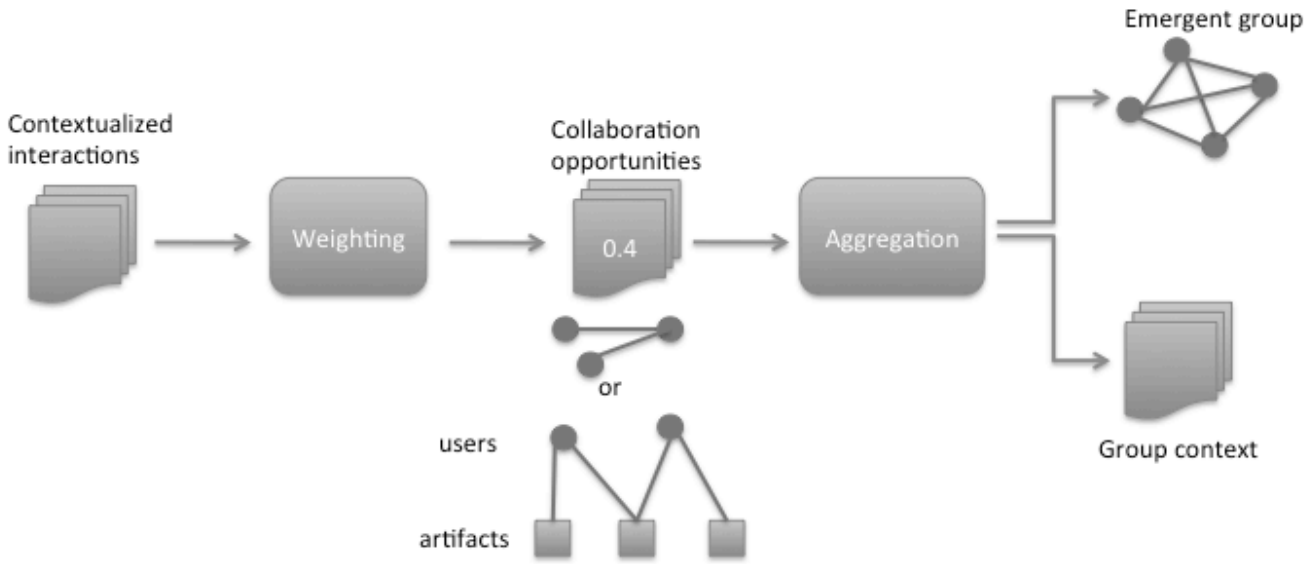

Figure 2: Analytic framework: process and by-products.

Recall that this analytical framework is an elaboration of the central component of the Group Informatics model that we describe in detail in another paper (Goggins et al., 2012), and present in summary form in section two. The contextualized interaction traces that are fed into our analytical framework emerge from the mixed methods described in the model of Group Informatics. These *contextualized interactions* are not only descriptive of the interaction that took place, but also carry attributes that situate those interactions within the socio-technical system supported by the online context. These include attributes that associate interactions to recognizable tasks being carried out within the context. This is a major focus for us; attributes may also relate each interaction to the history of the user and her experience of the context, or to other entities (people or artifacts) that populate and share the environment with the user. For example, in the case of a software development environment, a contextualized interaction does not simply record that developer Alice has made a change to some part of the software product, but also augments that basic trace with information that may speak to the artifacts touched or consulted by Alice and in what sequence, the number of times each artifact was in the focus of Alice's attention and for how long, the task to which all of that work is associated, and so on.

The definition of contextualized interactions of interest is the first step of the process outlined in figure two, and is qualitative, and — necessarily — highly domain- and context-specific; however, we can elaborate on it in a

general way. For contextualized interactions we are mainly interested in attributes that are quantifiable, since the following step of the process is their *weighting*. The Group Informatics model helps us to decide which attributes to quantify, and how to quantify them. As we noted, real-world systems capture sometimes very large volumes of user interaction data. Clearly, not all those interactions are equally significant: for example, many may capture routine activities, digressions, even mistakes, or other behaviors that do not carry particular semantics salient to what our qualitative analysis determines are significant indications of group behavior, leadership, or user behavior attributes which could contribute to user models or personalization. The weighting step is therefore based on quantitative contextual attributes of interaction traces, determined using qualitative information gleaned from triangulating data.

Our concept of triangulation is similar in purpose to other notions in phenomenological research, such as corroboration, validation and evaluation. The terms are different, but the purposes are congruent. Our analytical framework operationalizes these concepts, and helps bring to the forefront the most significant interactions while filtering noise. The weighting step assigns a relevance score (typically non-negative) to all contextualized interactions.

We look at each weighted, contextualized interaction as a *collaboration opportunity*, which we represent as a bilateral relation between two entities in the system, One of them is, in all cases, a user who participated in the interaction; the other may be another user, or an artifact. This depends on the types of interaction supported by our virtual context, as well as its purpose, determined from qualitative research methods like content analysis, digital ethnography or user interviews. In an online context that focuses on discourse, discussion and coordination, such as an online forum, collaboration opportunities connect pairs of users directly; in an context whose main purpose is instead the construction of some product or deliverable, such as an open source software development community, opportunities for collaboration between users are mediated by the artifacts upon which each user interacts with as part of her work.

In the former case (represented by our case on completely online groups in a learning context), the set of collaboration opportunities denotes a weighted social network between the users. In the latter (represented by the open source software engineering case), instead, the best representation is a bi-partite network: one mode of the network is the set of users, and the other mode collects the set of artifacts being the subject of the users' interaction. Collaboration opportunities represent the weighted arcs connecting those two modes.

Notice that – besides its weight – a collaboration opportunity relation must also have a *timestamp*, denoting the moment when that collaboration opportunity materializes. This is critical to identifying emergent group formation, which in turn is instrumental for achieving the goal of group awareness, since timeliness is a major quality attribute for awareness support, and group participation may be very dynamic on the time axis.

The next process step in our analytical framework is *aggregation*. This step aims at the identification of groups, and is fulfilled by applying suitable Social Network Analysis (SNA) algorithms to the network of collaboration opportunities. Unlike prior research that focuses attention on raw interaction data, through this aggregation step our analytical framework shifts the focus from dyadic relations, which denote the strength of the interest or requirement to collaborate (either directly, or in an artifact-mediated fashion) between two individuals, to a multi-party relation that must capture subsets of the whole set of users who have a particularly strong common interest or requirement to collaborate as a distinct social unit. The aggregation step leverages some form of clustering of the original collaboration opportunities network, and its output is a new network among the system users, which makes the multi-party relations visible and quantifies them.

A next output of our analytical framework is a representation of group context, which may include *intentional* information about the reason (interest or requirement) for the group to emerge, which sheds light upon the nature of the multi-party relation that was discovered through aggregation; it also may include *descriptive* information, i.e., details on the kind of activities the group as a whole carries out and their content. Group context information is often itself an aggregation of some of the context of participating individuals, derived from the contextualized interactions which originally contributed to the formation of the group – the key output of the Group Informatics model that feeds our analytical framework. A critical point of understanding is that the extensive literature

applying network analysis to raw electronic trace data does not typically reveal groups that exist in the data (Howison et al., 2012).

Our analytical framework, its process steps, and the data manipulated and produced through that analysis is instantiated in Section five in a way that is specific to each of the case studies we present there. In that section, the case study discussions are organized to highlight how the concepts informing our analytical framework are applied to each domain. Each case study reports on the following set of items:

- **Domain of study**: introduction to the domain and the application;
- **Contextualized interactions**: nature of the interaction traces collected and their contextual attributes;
- **Weighing procedure**: what kind of weighing is applied to the contextual attributes and how weighs are computed;
- **Collaboration opportunities network**: the nature of the network resulting from weighing, and a discussion about the semantics of the dyadic relations captured by it for the domain at hand;
- **Aggregation procedure**: the aggregation algorithms and parameters used to make groups visible, by manipulating the collaboration opportunities network;
- **Emergent groups**: the group structures emerging from aggregation, and a discussion of how they match what we know about the case study and the community of users populating the socio-technical system;
- **Group context**: what contextual information we can associate to each identified group.

To conclude each case study, we discuss how the techniques and findings described shed light on two research questions, which represent the common motivation guiding us to carry out all of those studies. Our research questions are the following:

1. **Q1 – Groups identification:** to what extent did the analytical framework succeed to identify emergent groups in this case study?
2. **Q2 – Automation:** to what extent did the analytical framework support the automation of the process, starting with the available electronic trace data produced within the socio-technical system?

5. CASE STUDIES AND FINDINGS

The case studies discussed in this Section are different with respect to the type of work the socio-technical system in each case study is set out to accomplish, the different types of participation it supports, and the kinds of trace data it captures. We have selected these case studies to demonstrate that our analytic framework can be applied in the face of significant heterogeneity.

From a socio-technical structuring perspective, we point out the contrast in types of work conducted through the systems in these case studies. In our first case study, we deal with completely online groups that are focused on learning discourse. In the second one, we study Open Source software developers, who are focused on quality production of knowledge-intensive, technical artifacts. Work type is an important distinction, which sometimes gets overlooked when performing network analysis of electronic trace data, but we find that if the reader keeps this in mind the case by case complexity associated with user adaptable context can be more easily appreciated.

As we note in the cases, the structural difference in our application of the Group Informatics model through our analytical framework is the difference between person-person and person-artifact centered analysis. This contrast is our rationale for selecting these two cases for this paper, from among the 16 specific studies we conducted over the years, which let us to our Group Informatics model and analytical framework as presented.

5.1 Completely Online Groups

5.1.1 *Domain of Study*

If you have ever joined an online group, you know the experience to be different than groups you join in your physical community. The first author studied the effects of context awareness tools on completely online graduate level courses at three major US Universities for five years. Completely online courses, like free and open source projects and Wikipedia editing occur without people ever meeting face to face. In the case of completely online courses, this is more absolute than it is with free and open source projects. Unlike large scale Internet facilitated projects, completely online graduate level courses have a managed structure with time constraints and specific tasks that are contained within a clear organizational context. The courses studied

incorporate a significant component of small group work, so in this respect, such courses are similar to organizationally situated work groups.

The technical tool used for instruction in our research is the open source course management system Sakai[2], with the JForum discussion board tool integrated. Sakai itself also provides file sharing, archived chat, wiki's and profile pages that students in the courses we studied used. As explained in the next section, activities in Sakai were logged using a tool developed at a large Midwestern US University, called CANS[3], which created participant awareness of the online context by sharing the overall activity of different members in raw form. As explained in the next section, our Group Informatics model was applied to the interactions captured by CANS, extending our understanding of activity in the Sakai environment through our analytical framework.

Completely online courses rely on technical infrastructure and pedagogical infrastructure (Goggins et al., 2010b; Goggins & Erdelez, 2010). Not every instructor uses the same technology in the same way, though a particular instructor is likely to use the available technology in similar ways for each course. This is the case in our study of the Sakai environment. By studying different instructors using the same tool, and generating interaction data using a common, context aware logging format, we are able to make general observations about the ways that groups come in to being under the direction of specific instructors, and within the constraints of the tool in general.

5.1.2 *Contextualized Interactions*

The "Context Aware Notification System" (CANS) generates electronic trace data that includes context information about each discussion board, chat, wiki and file repository where activity took place, and the sequencing of discussion board posts using timestamps in the logs. CANS captures both read and post information in discussion boards, wikis, archived chat and file sharing areas of the course management system. This bi-modal, context enhanced data capture enables us to identify new semantics for standard social network measures. For example, we know that high betweenness users identified through our analytical framework are of two types – those who are influencers, which is the classic meaning in SNA, and those who are highly active (Goggins et al., 2010a). We also observe that completely online groups like these create complete graph, dense networks; which are different in many respects from other social networks examined in the other studies reported in the wider literature. In this research, electronic trace data from CANS is woven together with interview data, content analysis and ethnographic coding to increase understanding of group development (Goggins et al., 2011b), group identity (Goggins, Laffey, Galyen, & Mascaro, 2011b) and the infrastructure required to support context enhanced online learning.

The model of Group Informatics guides the integration of these triangulating forms of data and their corresponding analysis. For example, in each course, the pedagogy of the instructor (small group focused versus lecture focused), the subject matter (interaction design, versus computer programming) contributes to the types of interactions we are able to identify in a course. When we look at software design courses, groups are often assigned a priori and also emerge over the life of the course. In the case of programming courses, there may be group work, but our analysis is that such groups' form around discussion threads focused on specific technical problems. The groups that matter in the programming courses are more dynamic than the design groups. This leads us to contextualize the traces using the group informatics model differently in these two types of courses.

The work (file sharing, wikis, artifact construction) and talk traces (discussion boards discourse) captured by CANS are from both Sakai and the JForum discussion board included in the online context. These tools were integrated seamlessly for the users, so that JForum appeared simply as a component in Sakai. Of the total interactions in the interaction warehouse, 91% occurred within the Jforum discussion board. The information in each trace includes:

- **User ID for the Event Creator**
- **Object ID –** This could be the specific JForum discussion board, a course wiki, a file or an archived chat.

[2] http://www.sakaiproject.org

[3] http://www.cansaware.org

- **Event URL –** The event URL is a unique page identifier. The Jforum discussion board limits five posts to a page, and each page in a discussion forum has its own unique event URL. It is therefore possible to perform analysis on page views for sections of the discussion forum, not merely the entire forum. We have a finer grained understanding of who is reading and responding to whom in a particular timeframe than other's conducting similar research. This data also enables the construction of implicit ties, discussed later and also unique to our research.
- **Kind**: There are seven main event types. JForum.Read, JForum.Post, Wiki.Edit, Session.Begin, Session.End, File.Create, File.Change and User.Chat.
- **Event Date:** a timestamp for the event in the log
- **Object Create Date** a timestamp for the create date of the object the event is in reference to. When creating a discussion post, this will be the same as the event date. When reading a discussion post, this will be the Event Date for the most recently created post in the thread.

Applying the group informatics model we performed further analysis of the logs in order to add context to them. CANS interactions are initially recorded between users and objects. In a discussion where knowledge is being constructed and information is being shared, however, interaction is more aptly represented as occurring between two people. Our interview data, content analysis of discussions and ethnographic following of the course unfolding during the capture of log data demonstrated that the courses we study do, in fact, function more like a conversation than like two people interacting around an artifact. In order to accommodate this reality in the Group Informatics model, we contextualized the log files as interactions between individuals. Figure three shows how the raw CANS record will note a specific individual performed a specific action at a page URL. In figure three, this is a "read" action. Few (<1%) interactions in the Sakai systems we study are between individuals and artifacts; they are almost entirely directed to one or more other individuals. The engagement is active and social. This is in contrast with our other case, in software engineering, where the interaction is truly between a software developer and an artifact. Figure three depicts how the actions of individuals are initially contextualized in the logs. The person depicted reads a set of discussion board posts provided previously by other participants. A connection is made in the log between this interaction and the individual performing it.

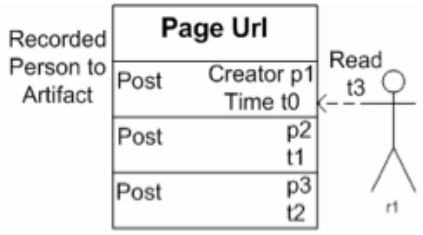

Figure 3 - Contextualizing Interactions in the Socially Focused Interaction Warehouse

However, an interaction between a person and a discussion board does not capture the full context of a discussion, particularly when the entire course of 15 to 35 people are engaged, or if a small group of 3 to 5 people are engaged in ongoing discussion. Figure four shows how we create the implicit connections associated with an online discussion thread in a completely online course. When a student views a page URL with multiple posts, there is an implicit interaction between each of the students whose contributions are on the page. Our integration of other data through the Group Informatics model enables us to operationalize these implicit connections through processing of CANS logs in time sequence, as depicted in figure four. The same explicit read act creates a connection between the user and all the other users in the page. These are different types of connections that explicit responses, and the Group Informatics model guides us to employ a specific weighting strategy, depending on the triangulating data available for the context as a whole, from prior studies of similar classes, and from ongoing study of new classes. These implicit connections inform our weighting, described next.

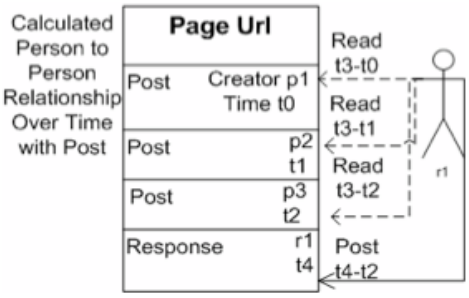

Figure 4 - Contextualizing Interactions by Accounting for Implicit, as well as explicit interactions

5.1.3 *Weighting Procedure*

Knowing the basics of who posted each item is sufficient for a rudimentary social network, but it is still not complete and does not take advantage of the model of Group informatics or our analytical framework. Weighting of ties and making connections for the appropriate number of prior items on a page are also necessary. Each of these components – how far back in a post sequence to imply a tie, and how to weight a tie – are deeply contextual. Figure 4 illustrates how we connect and weight connections between a user reading content and the user who creates it in Sakai. Similar strategies are used in each system, but contextualized in ways specific to that system.

Our methods for deriving weighted social network data from electronic traces overcome a number of the criticisms leveled against the analysis of such data using social network analysis methods (Goggins et al., 2011a). Issues of validity, reliability, and theoretical coherence in the common application of SNA to electronic trace data are described in depth in the paradigm altering work of Howison, Wiggins & Crowston (Howison et al., 2012). Questions of validity are centered on an understanding of the theoretical and methodological origins of social network analysis; which emerges from the practice of sociologists sampling co presence of people in social settings or collecting diaries of self reports. These two types of data are shown to have validity by representing both an observed network of interactions in the social settings, and self reports of who is important through the diaries. Many times people report connections to highly influential others whether this same connection is observed in the world or not. Electronic trace data, in contrast, represents a full set of technologically mediated interactions.

Calculation of weights is determined through information in the Group Informatics model. The most significant variables for weighting connections between individuals are the number of total interactions between them, compared to an average and the length of time (referred to as time distance) between those interactions. Two students with frequent interactions with long time distances may have equal connection strength with two students who interact occasionally, in short rapid bursts. Our Group Informatics Model supports analysis of what is really occurring during these two interaction scenarios, and enables us to use connection weight as a proxy for simply and understandably representing the connection between two nodes, even though the connection types may be slightly different.

5.1.4 *Collaboration Opportunities Network*

Connecting each of the students in each course context with other students by transforming raw, person-artifact interactions into a social network results in a collaboration opportunities network. The weighted incident arcs from a student Sa to another student, Sb measures the strength of connection between those two students. In most studies of online learning, these are dense social network where every member is connected to every other member to some extent. Decisions we make during aggregation procedures are central to our understanding of the collaboration networks of interest in a particular course. We describe this in the next section.

5.1.5 *Aggregation Procedure*

In the study of completely online group emergence the decision to aggregate, and how to aggregate is drawn from the context, artifact and interaction components of the model. Since each interaction contains a timestamp, and we calculate implicit interactions in discussion based on whether or not a post exists at the time a user reads or posts, and how recent the related post (up to five) in a page URL is, we aggregate differently, depending on the

research question. To study emergent group behavior and consider task and group context, we aggregate the data in two ways. First, we slice the data according to known module time periods – places in a course where work is taking place in small groups, peer to peer collaboration or individually are sliced and analyzed to represent the different socio-technical processes embodied by these different a priori configurations. For example, a priori group structures lead us to look for those groups, while in individual activities; the notion of group emergence is more likely the qualitative finding from the discovery of groups in our analytical framework. Each course is treated as a separate unit, so interactions between people who are in two courses at the same time are only measured within the course context.

Once we have created a set of discrete interactions organized and grouped as described above, we will aggregate them in to node by node pairs, with both a summed and averaged weight. These weights, which are derived from analysis and processing through the group informatics model and weighting procedure, are then visually represented in network models. The social context and relations between users are then, in this way, represented in a static diagram, though the weighting procedure represents a dynamic factor in the interaction. A visualization of a weighted social network represents dynamic behavior in a static diagram, which is enabled by application of the Group Informatics model through our analytical framework.

5.1.6 *Emergent Groups*

The Group Informatics model driven visualizations are information rich sources for understanding group emergence. The weighted edge sets generated by the weighting and aggregating of interactions in an online course, following our model of Group Informatics can then be analyzed using standard social network analysis (SNA) measures. Applying these statistics to the processed electronic trace data is theoretically coherent with the measures themselves, overcoming prior criticisms of applying network analytic techniques to raw trace data. For example, degree centrality is intended to identify significant interaction between one node and the other nodes in the network. The emergent groups will share high in degree and out degree centrality traits with each other, and tend to interact in clusters. By applying weights based on qualitative data, we have validated that the relations a social network analyst expects to be represented by a set of measures is, in fact, what is represented. While raw interaction traces may reflect different sorts of behaviors, our Group Informatics model normalizes electronic trace data to reflect what network analysts expect the measure to mean (see section 2) (Goggins et al., 2012).

The emergent groups identified in this case are triangulated back with the qualitative data and with user reports of group membership. This serves as a self check, as well as external validation that the model, as applied, works for the types of cases studied here. It is also an encouraging sign that the Group Informatics model and our analytical framework for constructing context adaptive systems will be applicable in the development of future tools.

5.1.7 *Group Context*

We know about every discussion board the user participated in, and we understand this within the context of a class. In some cases there are restricted group discussion boards. Here, individual group context becomes more apparent because of our processing through the model. As Dourish (2004) points out, context is dynamically constructed. This is especially the case when the context is a new form of technologically mediated social organization, like the two described in our paper. With group context identified using our analytical framework, we are in a position to inform future tools that enable context adaptivity by incorporating group context into user modeling and personalization research because the complex, social factors are understood through analysis of interactions through the user of our model and analytical framework.

5.1.8 *Research Questions*

We were able to effectively identify groups and group context through our analysis of completely online learning groups. Further, we were able to identify clear steps for the automation of this group awareness and task awareness; leading us to conclude that a next step in the application of the Group Informatics model is to build our analytical toolkit into live systems, and provide context adaptivity that takes into account the full, socio-technical context of use as represented not just by raw trace analysis, but by the contextualization and triangulation of that analysis with other data and data analysis. This is evident from our answering of the research questions in this case.

5.1.8.1 Research question Q1: groups identification

Through analysis of context data generated by CANS and analyzed as noted above we have discovered patterns of group identity formation and leadership in completely online learning groups (Goggins et al., 2010b; Goggins et al., 2011b; Goggins, Laffey, & Galyen, 2009). For example, in the study of one online course, we can see that a core group of team members are clearly identifiable from our processed CANS logs as the leaders of their respective groups. This is illustrated in figure five. Each group member is represented by a unique name that includes their group number (g1-g6) and a unique identifier. One group is represented with actual names. Each group has a distinct shape and shading combination, such that it is visually clear from the figure that one member of each group is in the core for the overall network. This figure, developed using our model of Group Informatics and the analytical framework described in section four demonstrates that both groups, and the leaders of those groups are identifiable using our model. These results have been replicated across 11 courses with 3 different instructors at two different institutions teaching five different courses.

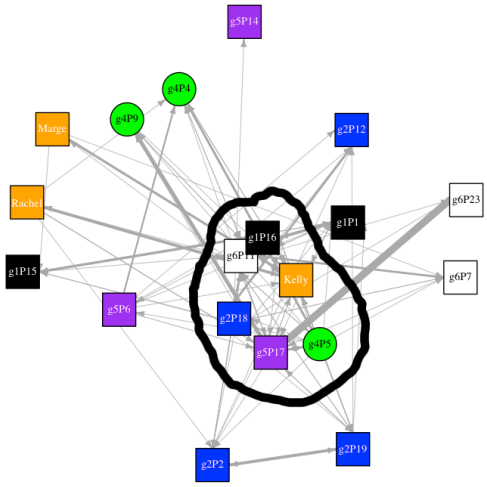

Figure 5 - Example of How Leadership is revealed through network analysis of CANS data

In another study we showed how connectivity between members differs by read and post activity. The blue lines in figure 6 represent post connections and the gray lines represent read connections. This figure illustrates two important dimensions of bimodal logging as an important component in the analytical framework presented here, and in indication that these analytical tools will support context adaptivity that incorporates group and task awareness information into tools developed by us and by other user modeling and personalization researchers. First, the heavy volume of more passive, reading behavior compared to active posting behavior in the network is clear. There is a lot more gray (reading) than blue (posting). Second, the node sizes representing the degree centrality of the nodes show how in a tightly connected network like this, the variance between the largest and smallest node is relatively small. Third, the relative proximity of different nodes to each other suggests a few different roles in this course. Nodes N04 and N09, near the upper left, small and green, have positions between clusters in the graph. These users, whose participation is relatively low as indicated by degree centrality, are located in central positions of the network because they act as gatekeepers between clusters. This graph is a visual mechanism for making high betweenness individuals – lurkers and brokers – visible in a course social network, similar to the way we accomplished this using statistical measures (Goggins, Galyen & Laffey, 2010).

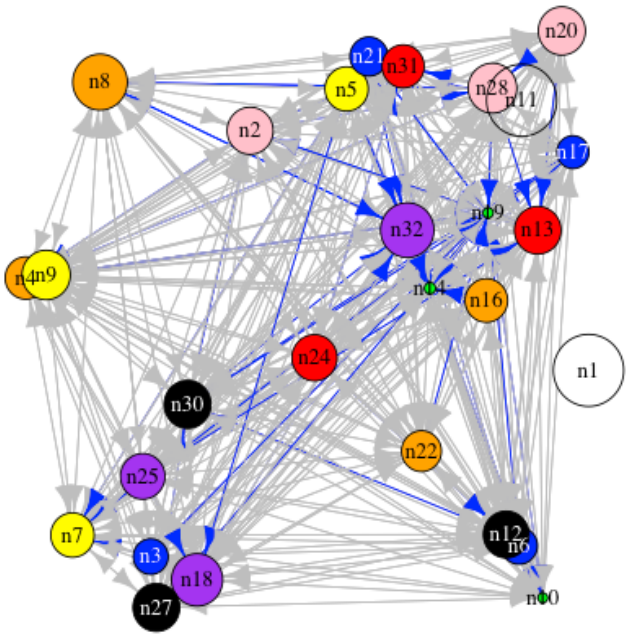

Figure 6 - Contrast of Read (gray) versus post (blue) data

5.1.8.2 Research question Q2: automation

These two visualizations provide us with a clear sense of how group task context can be used to make groups more aware of each other and how they are interacting in completely online educational contexts. The data available from CANS, processed in our analytical framework and leveraging our model of Group Informatics, could provide these same visualizations to asynchronous learning groups on a regular basis. CANS already provides basic information about frequency of login and discussion boards where people gather with each other online. Providing these sorts of visualizations could increase group awareness. If visualizations like these were centered on specific discussion boards, file sharing areas or wikis, they would also serve to help support the provision of task context and, ultimately, context adaptive systems. Since our analytical framework has been developed and operationalized across a range of courses, this is possible with relatively little effort.

5.2 Open Source Software Engineering

5.2.1 Domain of study

Open source software projects are among the most impressive examples of the ability of virtual software development organizations to create complex, robust and very large products. Work by the Apache Foundation[4], Linux[5] and other open source communities[6] provide recognized examples of projects whose output rivals top-of-the-line products produced by "traditional" software development organizations. Virtual software development organizations are also becoming increasingly prevalent in for-profit enterprises. Geographically dispersed and globally distributed teams are becoming more and more common for the execution of large-scale industrial software development projects. A distributed industrial software organization typically structures its project teams across business unit locations and boundaries, and often in partnership with other for-profit enterprises; with increasing frequency, we also see virtual organizations born from joint ventures between corporations and open source projects, leading to hybrid organizational models (Wagstrom, Herbsleb, Kraut, & Mockus, 2010).

In those software development virtual organizations, co-located work and in-person communication and coordination are the exception, if they exist at all; all work and communication occurs through the mediation of

[4] http://www.apache.org
[5] http://www.linux.org
[6] http://www.flossworld.org/

the tool set that makes up the production infrastructure of the project. Most of these tools, ranging from source code repositories, to bug and change requests databases, to mailing lists and discussion boards, etc. provide traces of the developers' activities. It is worth noticing that – among virtual software development organizations – open source communities are not only the most distributed and decentralized, but also the ones that are most open in making available and leveraging those traces. While their primary motivation is to facilitate the work of community members, and to lower the entry barrier to potential new contributors, this practice of course also makes them an good subject for trace-based studies and analyses.

We have been studying a project within the Eclipse.org open source ecology, namely Mylyn (Kersten & Murphy, 2006), which maintains one of the most popular add-ons (also known as *plugins*) to the widely used Eclipse Integrated Development Environment (IDE). An IDE is a productivity tool for software development; modern IDEs are extensible, by means of plugins that provide sets of new or advanced functionality in a shrink-wrapped component. The Mylyn plugin for Eclipse records the actions that a developer carries out within the user interface of the IDE (e.g., menu selections, mouse clicks, opening and editing of files, issuing of commands, etc.) while she works. Since a software developer often works on multiple tasks at once, the purpose of Mylyn is to ease the cognitive load of the developer whenever she switches from one task to another. As the developer takes up a past task, Mylyn uses her recorded activity to automatically modify the appearance of the IDE, and present in a prominent way to the developer the most pertinent information for that task, such as the software artifacts she has been working on (the *working set* of the task), and the IDE tools and commands she has used. By recording the actions related to each task, Mylyn thus helps the developer constructing and remembering the context of her work: in fact, Mylyn calls these traces *task contexts*, and the action recorded in them *context events*.

5.2.2 *Contextualized interactions*

Mylyn — as an open source project — represents the subject of this case study; however, Mylyn — as a technology —also provides us with the contextualized interactions that allow us to study the project itself, since the contributors to the Mylyn project routinely use the Mylyn plugin in their work, and– by convention – have been collecting for several years in the project repository hosted at Eclipse.org the task context information for their completed tasks.

The work traces in a Mylyn task context are extremely fine-grained, both temporally and with respect to the detail of observable work. They also reveal the whole working set of artifacts consulted or manipulated at any point in time during a task, whereas most software development traces only record the narrower *change set,* that is, the artifacts that are modified and then committed in fulfilling a task. Mylyn context provide the progression of work by the developer from the beginning to the completion of her task. The information within a context event includes, among other things:

- **Developer ID**
- **Task ID**
- **Kind**: type of action by the developer on the Eclipse GUI (selection, edit, command, etc.)
- **Structure Handle**: a unique ID that identifies what software artifact is subject to the action. Granularity can vary (files vs. classes vs. class inner elements, i.e., methods and attributes).
- **Start Date:** a timestamp
- **End Date** a timestamp

We postulated that individual task context traces can be leveraged to support group awareness by analyzing the contexts of developers engaged in concurrent work, and how they are related. For example, if contexts show an *intersection* – defined as an overlap between the working sets of 2 or more developers — those developers may constitute a group within the project team. In a recent study (Blincoe, Valetto & Goggins, 2012), we have analyzed these traces, and found that intersections between the contexts of *pairs* of developers provide accurate and early means to detect the *coordination requirem*ents (Cataldo, Wagstrom, Herbsleb, & Carley, 2006) between those pairs. Here, we extend that work to detect groups, by looking at the intersection of multiple contexts. An advantage of this technique is that overlapping working sets are often an antecedent to group formation, since – in a virtual software development organization — groups may form *ad hoc* once developers realize their need to

negotiate and resolve the complexities related to their concurrent manipulation of the same artifacts for different technical reasons.

For this study, we have examined the data about eight releases of the Mylyn software (v2.0 through v3.3), which span almost three years, from December of 2006 until October of 2009. The mass and kind of information made available by the Mylyn repositories for that period covers nicely the major entities of the Group Informatics model presented in Section 2. From the Bugzilla repository of the project we have extracted 1,970 software development tasks. Each of those tasks is associated to one or more context records, for a total of 588,796 context events. We have filtered those contextualized interactions, and considered only the selection and edit actions upon Java source code artifacts (450,747 events), since other types of artifacts are by-products – as opposed to subjects – of development work. Besides allowing us to establish the relations between a task and the artifacts consulted or manipulated during that task, context records, of course, also contain information about the developers who carried out work on the same artifacts, and hence worked on each task. Each of the eight Mylyn releases, whose duration typically between three and four months, saw the contribution of 13 to 18 developers. We consider all the tasks by those developers in the same release as potentially concurrent.

We have also collected other data from the project repositories, which we have used for triangulation and validation of the insight gained through the analysis of the aforementioned contextualized interactions. In particular, we have mined the recorded acts of communication archived in the project mailing lists and in discussion threads about the individual project tasks in the same release periods. The population contributing these comments is much larger than the set of actual developers. Over the 3 years considered, more than 400 distinct user IDs posted comments in these forums.

5.2.3 *Weighting procedure*

Our process of analysis of the data set described above works as follows:

1. For each release, we consider all contexts, and all events within each contexts, to determine the intersection between the artifact working sets of all development tasks;
2. We produce a set of *intersection records* in the form:
   `Ix` = (`Task_a`, `Task_b`, `Task_c`, ...) → (`Artifact_1`, `Artifact_2`, ..., `Artifact_n`);
3. We construct a bi-partite network. The developers involved in all the tasks listed in each intersection represent a mode of the network, and the intersections themselves are the other mode. The arcs of the bipartite network are valued: the weight of the arc between developer `Da` and intersection `Ix` represents the number of artifacts manipulated by `Da` in each task she has worked on and included in `Ix`. Notice that, since developer `Da` may be responsible for multiple tasks within the same intersection `Ix`, the weight of the edge between `Da` and `Ix` can be a multiple of the artifact cardinality of the intersection.

5.2.4 *Collaboration opportunities network*

The bi-partite network described above, connecting each of the Mylyn developers with a subset of the Mylyn artifacts, represents our collaboration opportunities network. The intersections represent the subject and the content of the potential collaboration between any developers that have operated on those artifacts in one or more of their assigned tasks within a software development release. The weighted incident arcs from a developer `Da` to an intersection `Ix` measures the strength of the interest — or need — of `Da` to collaborate with other developers on the set of artifacts included in `Ix`. Moving from a local to a global view of the network, the set of incident arcs to various intersections departing from the `Da` node show all the collaboration opportunities for `Da`. Several of those collaboration opportunities may involve a limited number of fellow developers, who therefore may have the most interest to come together as a group, and coordinate their concurrent work to efficiently complete their tasks.

5.2.5 *Aggregation procedure*

From the weighted bi-partite network of collaboration opportunities described above, we want to identify subgroups of the developers' community that gravitate towards common work (i.e. overlapping working sets) across several different tasks, and who have significant amount of overlap in those working sets. Many different network-analytic techniques exist for such a purpose. Our aggregation procedure focuses on the construct of bi-cliques (Borgatti & Everett, 1997), and works as follows:

1. we dichotomize the bi-partite collaboration opportunities network at a level that is above the median of the weight of all the edges in the network;
2. we compute bi-cliques, to capture what subset of the developers community tend to co-participate in the same intersections;
3. we compute the structural correlation matrix between developers, based on the relationship between the set of developers and the set of bi-cliques they are part of, which was computed in the previous step.

5.2.6 *Emergent groups*

As the result of aggregation, we obtain a new person-by-person network (in the remainder a "work network"), in which arcs denote the Pearson correlation between any two developers, and signify how similar are those two developers in terms of the bi-cliques they are part of. To identify sub-groups that are cohesive within this network of developers, we want to filter out weak correlations. Therefore, we use a cutoff point of 0.4 for the arcs in the network. An example of the networks thus obtained can be seen in Figure 7.

5.2.7 *Group context*

The aggregation procedure distils the relations between developers to discover cohesive groups; however, it is possible to "carry over" throughout that procedure some information that describes the collective context of those groups, and that is derived from the contextualized traces of interaction provided by Mylyn. The most important source of information is represented by the intersection records in the form

`Ix = (Task_a, Task_b, Task_c, ...) → (Artifact_1, Artifact_2, ..., Artifact_n)`

described above. We can easily retrieve the artifact sets that are involved in the intersections that determine each bi-clique; from there, we can also compute the sets of artifacts that underlie the bilateral developer-to-developer correlation arcs represented in the final network output by the aggregation process. The union of those artifact sets provides a look at the common field of work of each emergent group. An even more detailed view is available possible, since we can also retrieve all context events associated to those artifact sets. At that fine-grained level, we know about all actions (consultation, editing, etc.) carried out by members of the group, as well as their timestamps.

5.2.8 Research questions

On the basis of the process explained so far and its results, we now answer the research questions listed in Section 4 for the open source software engineering case study.

5.2.8.1 Research question Q1: groups identification

To answer research question Q1, we validated the outcome of the application of our analytical framework to this data set with two different kinds of triangulating information. First of all, we used qualitative information that we have collected while studying the Mylyn project, and perusing its repositories and archives that can be freely consulted on the Web, as is the custom of open source projects. Moreover, we used communication traces, which are an integral part of those archives, to construct a separate and independent social network (in the remainder, a "talk" network), and compare it with the "work" network resulting from our analysis for any given release. In the interest of brevity, we discuss in detail our results for two of the eight releases we examined: Mylyn v.2.0 (the earliest), and v.3.3 (the most recent release included in our data set).

Figure 7 shows the 13 developers active in v.2.0 (with anonymized IDs). We limit our analysis to constructs larger than a triad: two major such structures are visible. They are the 2-cliques (Alba, 1973) composed by (304, 143, 399, 373, 463) and {Seidman (304, 35, 22, 312, 319), which we consider as our "candidate" work groups, and refer to as WG1 and WG2, respectively. WG1 and WG2 have developer #304 in common; Seidman & Foster, 1978). As discussed – among others — by (Erickson, 1988), a network construct that is more robust and has redundant arcs is more likely to truly represent an actual distinct group within the organization depicted by the network as a whole; therefore, our primary candidate for an emergent work group among the developers of Mylyn v.2.0 is WG1, while we consider WG2 a more marginal candidate.

Triangulation with our knowledge of the project history helps to interpret the results in Figure 7. We know that the central figure of developer #304 is the most active in release 2.0, working on many tasks, and participating in the most intersections with colleagues' work. Moreover, we can shed light on the nature of the work relationship

made evident by the arc between developers #304 and 335. By means of data such as developers' profiles and communication information, we have learned that the involvement of developer #35 in Mylyn v.2.0 amounts to a volunteer sub-project (the *Google summer of code*[7]), organized as an individual but intense task, in which #35 worked on hundreds of code artifacts, and was mentored by #304. The summer work of developer #35 concerned a component of the Mylyn product outside of the critical project path: as such, it was kept well-separated from the main stream of work and only intersected the activities of a few other Mylyn contributors. What we know about Mylyn v.2.0 thus corroborates the structure and groups visible in Figure 7.

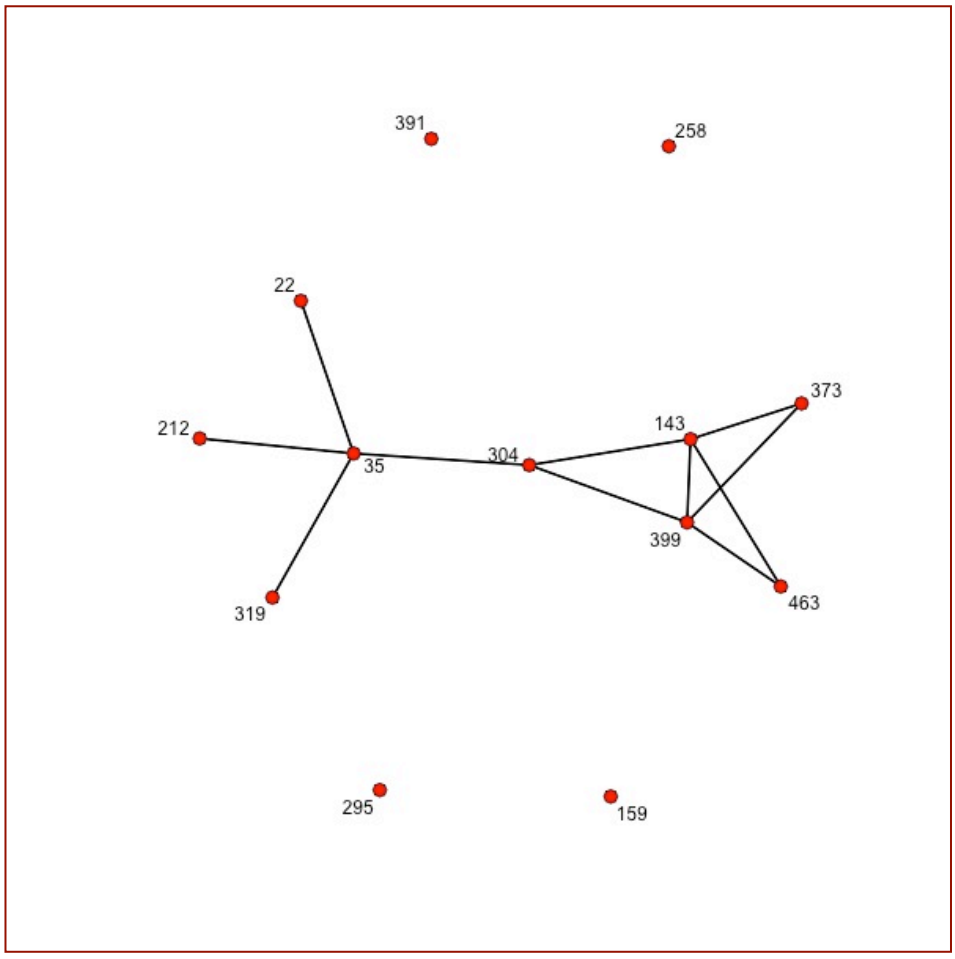

Figure 7: Mylyn v.2.0 – cohesive subgroups based on task context data.

We sought further corroboration in the communication traces from the archived developers' discussions. From those traces we have constructed the "talk" social network for Mylyn v.2.0, through the same methods discussed in the case study of Section 5.1. Project discussions in open source project may involve members that are not actively working on the software product, and also the general user audience, who may comment on quality issues of, and ideas for, the software product. For Mylyn v.2.0, the talk network includes many more actors and relations (64 nodes and 131 arcs) than the work network distilled by our analysis of contextualized interactions. Therefore, we focused on our candidate work groups, and looked at whether they are also recognizable in the talk network.

As a first step, we looked at properties of network connectivity, and seek structurally cohesive subgroups (Moody & White, 2003). For this analysis we used the `cohesive.blocks`() algorithm implemented in the R package *igraph* (Csardi & Nepusz, 2006). That procedure segments the overall graph in a hierarchy of increasingly more cohesive groups, each of which is a subset of the group that precedes it. The results for Mylyn v.2.0 are shown in

[7] http://code.google.com/soc/

Figure 8, which draws the talk network and color-codes the cohesive blocks it identifies. The most cohesive "talk" group is the set of seven red nodes TG1 = (304, 143, 399, 373, 159, 457, 391).

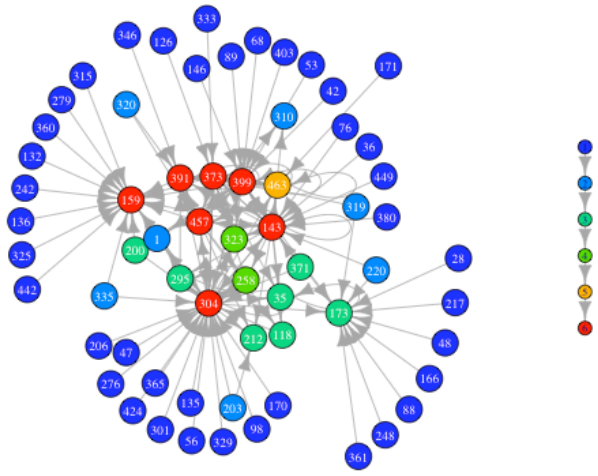

Figure 8: talk network for Mylyn v.2.0

First of all, we observe that our marginal candidate, the WG2 2-clique, does not find confirmation in the topology of the talk network: among the five members of WG2 only #304, who is in common with WG1, is included in any of the most cohesive groups. This is in line with the observation by (Erickson, 1988) that a recognizable, but sparsely connected, sub-graph may not reliably indicate an emergent group.

We can also see that four of the five WG1 members also appear in TG1; also, of the three TG1 members that are not in WG1, there is one, #457, who did not do any development work in Mylyn v.2.0, and therefore could not be included in our "work" network. Moreover, developer #463, who is in WG1 but not in TG1, is the only orange node in the cohesive blocks graph. That means that WG1 is fully represented in TG2 = (304, 143, 399, 373, 159, 457, 391, 463), which is the immediate superset of TG1 and the next most cohesive group.

Therefore, the role of WG1 as an emergent distinct group in the Mylyn development organization finds corroboration, from a structural standpoint, by the analysis of triangulation data from the communication traces. Based on the same data, we can carry out further analysis, and evaluate whether candidate group WG1 can be seen as "tighter" than other possible choices, such as highly cohesive structured like TG1 or TG2. Many social network scholars, possibly starting from (Bock & Husain, 1950), have maintained that a distinct subgroup in a network should be characterized by more, or more intense, ties between group members, in comparison to the set of ties of the group members with the rest of the network In the case of a weighted, directed graph like our talk network, that contrast can be expressed with the formula below

$$GS = \frac{\dfrac{\sum_{i \in G} \sum_{j \in G} K_{i,j} * a_{i,j}}{\bar{G} * \left(\bar{G} - 1\right)}}{\dfrac{\sum_{i \in G} \sum_{j \notin G} a_{i,j} + \sum_{i \notin G} \sum_{j \in G} a_{i,j}}{2 * \bar{G} * \left(\bar{N} - \bar{G}\right)}} \qquad K_{i,j} = 0 \text{ if } i = j \;\; / \;\; K_{i,j} = 1 \text{ if } i \neq j$$

where: GS is the ratio that denotes the group strength; a_{ij} is the weight of a tie between two nodes; $\bar{G}$ is the vertex cardinality of a candidate group G; and $\bar{N}$ is the vertex cardinality of the whole network.

This formula computes the ratio between the average strength of the group-internal arcs and the arcs towards network nodes external to the group (also known as centripetal vs. centrifugal strength, in the view of (Alba, 1973). The higher the ratio, the tighter the subgroup, and the more likely to constitute an actual group of actors with a need or incentive to coordinate within the overall network.

We computed the ratio above for WG1 vs. TG1 and TG2, and the table below displays those results. Although all groups result quite tight and the GS score are rather close, WG1 has a slight edge. We interpret this data to mean that WG1, which has been identified through our analytical framework, is at least as good a candidate for an emergent group as TG1 and TG2, which are derived instead from the topology of the communication network. Taken together, the various analysis elements we have derived from triangulating data provide a confirmation of our contextualized analysis of work traces.

v.2.0 – Avg. tie strength =5.87	WG1 (304, 143, 399, 373, 463)	TG1 (304, 143, 399, 373, 159, 457, 391)	TG2 (304, 143, 399, 373, 159, 457, 391, 463)
Centripetal ties strength (Total)	408	507	531
Centrifugal ties strength (Total)	295	245	222
Centripetal strength (Avg.)	20.40	12.07	9.48
Centrifugal strength (Avg.)	0.50	0.31	0.25
GS ratio	**40.80**	**39.32**	**37.92**

The same analysis can be applied to other project releases. We report it hereby for Mylyn v3.3, because the process highlights a different – and possibly even more interesting — lesson. In Figure 9, we see the network produced by our analytical framework. Several noticeable constructs exist, including two 4-person cliques WG1 = (391, 416, 373, 159) and WG2 = (391, 416, 1, 193). We can also observe that those two cliques, as well as the triad (399, 391, 416), all hinge upon the pair of developers (391, 416), which seem to hold them together. Therefore, we also consider the 2-clique WG3 comprising of the seven nodes (391, 416, 373, 159, 1, 193, 399).

In Figure 10, we see the cohesive blocks computed for the talk network of Mylyn v.3.3. The most cohesive subgroup TG1 is composed of the red nodes (391, 416, 373, 159, 399, 304, 118). Five of seven members of WG3 are therefore also found in TG1; moreover, if we consider the next cohesive block, TG2, it adds a single member, the orange node #1, which also appears in WG3. Furthermore, WG1 is fully represented in TG1, and three out of four members of WG2 are also members in TG2, with the exception of #193. The topology of the talk network thus largely confirms the results from the work network.

Again, we looked also at how tight the candidate groups are, using the same technique as for Mylyn v.2.0; the table below summarizes the results. With respect to WG1 and WG2, the centripetal strength of these small 4-persons candidate groups is simply too weak; that means they do not constitute a complete group *per se*, but they are more likely granular elements of a larger group. In fact, when they are "merged", as in WG3, we get much

better results. However, even if the GS score of WG3 is higher, it remains slightly lower score than TG1; more, the score of TG2 indicates it is a significantly tighter group candidate.

When we examined the adjacency matrix of the talk network to understand the reason of these results, we could see they – quantitatively – depend on two factors: the inclusion in WG1 of developer #193, and the absence of developer #304.

v3.3 – Avg. tie strength =4.39	WG1 (391, 416, 373, 159)	WG2 (391, 416, 1, 193)	WG3 (391, 416, 373, 159, 399, 1, 193)	TG1 391, 416, 373, 159, 399, 304, 118}	TG2 (391, 416, 373, 159, 399, 1, 304, 118)
Centripetal ties strength (Total)	26	11	264	299	337
Centrifugal ties strength (Total)	244	222	169	177	141
Centripetal strength (Avg.)	2.17	0.92	6.29	7.12	6.02
Centrifugal strength (Avg.)	0.61	0.55	0.26	0.27	0.19
GS ratio	**3.44**	**1.67**	**24.20**	**26.37**	**31.68**

Developer #193 is quite undistinguished in the talk network, as he is connected only to #399; therefore he does not bring a significant contribution to the GS score of WG1; actually it penalizes it. There is a reason for such a peripheral role: we know that the only two tasks performed by developer #193 deal with the integration of Mylyn with a separate commercial software product. Moreover, #193 was at the time working in the company producing said software; therefore, developer #193 was largely on the outside of the Mylyn development community, and possibly unknown to most people there. However, in the development of Mylyn release v.3.3, the role of #193 is quite noticeable. He communicated with developer #399 because they both worked directly together on one of those two tasks. In the other task, developer #193 worked alone, but the task was a large endeavor, in which he logged several thousands context events, and manipulated a large number software artifacts; as such, he had a chance to heavily impact the work of several other team members.

From the perspective of collaboration opportunities and coordination needs, therefore, the inclusion of developer #193 in a group with the other members of WG1 seems justified. In fact, it shows an example in which our analytical framework is able to highlight an emergent work group that spans organizational boundaries, thus providing a level of awareness that would be hardly available otherwise to members of separate organizations.

Developer #304 represents in a certain sense the flip side of that. In the early Mylyn releases, he was the most active developer, but later his position radically changed: he relinquished his strong hold on the Mylyn code base, and took a increasingly isolated role. By release v.3.3, developer # 304 is no longer central to the development effort, although he still works on a few tasks, mostly in tandem with one teammate (in this case #351), and serving as a mentor, as he did with #305 in v.2.0. This reflects normal dynamics in maturing open source projects, in which early leaders can take a step back from the day-to-day operations, and may choose to assume mainly a managerial role, or concentrate on specific ideas and pet projects. Consistently with his role as a senior technical lead and project manager, however, developer #304 is still quite active in communicating with the rest of the community, in particular discussing with those developers who contribute most work to the release. These frequent exchanges boost his importance in the talk network and the GS score for TG1 and TG2.

Developer #304 could thus be conceivably included — with that managerial role — in any candidate emergent group for release v.3.3, but that is something that remains opaque to an analysis focusing on work traces. His case seems to highlight some of the limits of our technique; it also serves as a reminder the importance of triangulation to capture the multi-dimensional nature of group work and collaboration.

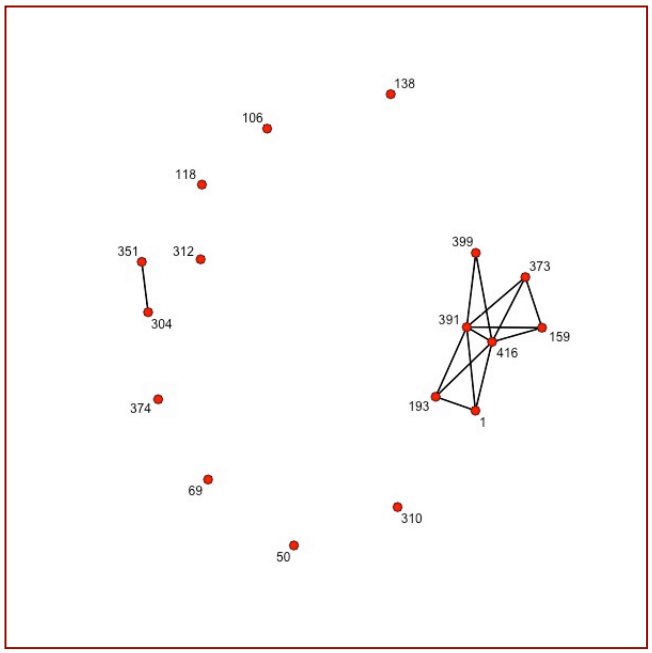

Figure 9: Mylyn v.3.3 – cohesive subgroups based on task context data.

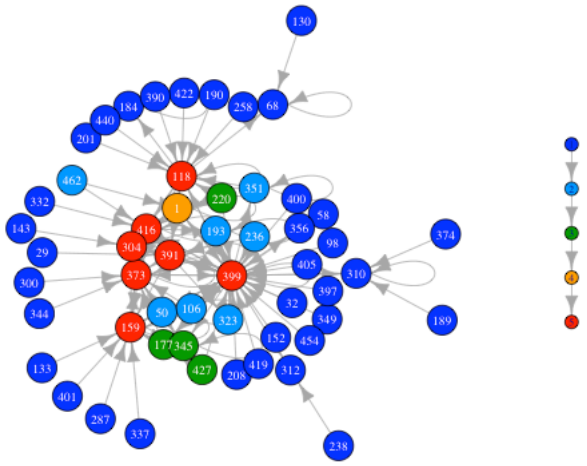

Figure 10: talk data network for Mylyn v.3.3

5.2.8.2 Research question Q2: automation

The steps constructing the work networks described in this Section are algorithmic in nature. They only have a couple of parameters, whose values can be assigned either automatically or *a priori*. Therefore, our process can be fully automated.

It is important to remark that the networks thus produced are simple, with low node and edge cardinality; they will be often amenable to direct visual examination for the extraction of groups. In fact, the whole purpose of the aggregation step of our analytical framework is to distil a large amount of information (the bi-partite collaboration opportunities network) into a much simpler network that only shows significantly strong relations that make the fabric of a work group. As a result, the corresponding network constructs can in many cases — like our case study — become immediately evident. If the final networks produced by our method were larger and less sparse, patterns and structures revealing cohesive subgroups such cliques, etc. would be easily extracted by running standard SNA algorithms.

There is also another important facet that speaks to efficiency and automation. Even if the study we have presented was retrospective, the nature of the context data we used is not. In Blincoe, Valetto & Goggins (2012), we discussed the timeliness that can be achieved in recognizing coordination requirements between developer pairs when leveraging Mylyn context data. The group detection procedure described here can occur in an equally timely fashion. Timeliness is paramount to make effective use of any automated awareness capabilities. This point is also closely related to the design implications of our work, regarding context adaptive systems and tools, which we discuss in the next Section.

6. DISCUSSION

6.1 Context Adaptivity Through Group Awareness

While user modeling and personalization research to date has focused mostly on individual profile characteristics and personal context, users of contemporary information and computing systems experience context as a dynamic, largely social phenomenon, as those systems are increasingly designed as collaborative and social virtual contexts. Making those systems context-adaptive therefore calls for heightened attention to the social (or group) context in which the users are immersed as they interact with the system and with one another.

We are not the first to argue that social information is a part of context (see (Anaya & Boticario, 2011; Bunt & Conati, 2003)), but we contribute some means to overcome some of the principal obstacles. One of the outstanding issues is the emergence of groups in virtual contexts. Our model of Group Informatics is equipped with an analytical framework that can surface group participation and group dynamics. That is, we maintain, the basis upon which any context-adaptive systems that incorporate a social dimension must be built, since group awareness is likely to fundamentally impact the behavior of the individual user towards the system and other users. Dynamic group participation information – obtained in a timely manner from the system itself – should be made part of each user's profile so that the system can provide the user with the means to exploit that awareness and manage her social sphere during her technical activity.

Electronic interaction traces represent the source of all our analysis, and are invaluable for arriving at group awareness. An important aspect surrounding our method is that – as others before us have shown (Goggins et al., 2010a; Howison et al., 2012), and the cases we described here have reinforced — raw interaction traces must be interpreted in the light of triangulating information about the specific application domain, in order for them to represent social interactions as experienced by users, and to be used to calculate meaningful social network analysis statistics like degree centrality, betweenness and others. This calls for significant attention by the researcher or practitioner involved in decisions related to the design of effective group awareness and social context adaptivity in socio-technical systems.

A final, important observation is that the notion of group participation needs not be – by itself – the only personalization trait that can be extracted and leveraged for context adaptivity. Because of the ties with underlying electronic and contextualized interaction traces promoted by our model and framework, it is possible to associate the construct of an emergent group with a wealth of information that represents the context of that group as a

social unit within a socio-technical system. Group context information can provide insight into the collective motivation for collaboration among group members as well as a description of the content of that collaboration.

6.2 Limitations

A comparative case study of our analytic method, while informed by prior empirical work, represents a limited sample and must be evaluated across more contexts and more examples. Context adaptivity requires models, methods and ultimately software implementation to prove useful. Our contribution is an analytical model, illustrated by two cases. Future work will address the limitations of a comparative case study, though we argue the description and examples described here are necessary if context adaptivity is to address the group and task context dimensions of user modeling and personalization research into context adaptivity.

The other limitations of this work center on our integration of quantitative and qualitative data using a model derived from our work across contexts. How data is collected, the noise that is in that data and the limitations inherent in the use of electronic trace data for understanding social phenomena are significant, as pointed out by Howison, Wiggins & Crowston (2012). Our work makes progress toward overcoming these limitations by making the methods, theoretical assumptions, data analysis strategy and model followed in two cases explicit. We do not claim to have devised a generalizable approach; but by documenting two cases this explicitly, our work enables other researchers to test, refine and adapt an approach to understanding the social aspects of user personalization and modeling across socio-technical contexts.

6.3 Design implications for tools

Knowledge of group participation for each user of a socio-technical system, plus information that can be construed as the context that the group as a whole uses for its work are the outputs of our analytic framework. We envision a two-step strategy to incorporate those analytic results into new collaborative and context adaptive systems and to augment existing systems with new, analytical tools.

The first step is predicated on the timeliness of our analysis, that is, the ability to carry out the analysis and discover groups as they emerge. Building on that timeliness, we can develop an incremental "live" mechanism that continuously collects the contextualized traces from the socio-technical systems and all of its participants, and periodically executes our analysis in a centralized server. Such a mechanism is an important next step in our work, and the work of the community focused on context adaptivity that incorporates a social dimension. Such tools will make users aware of how groups form and evolve in quasi-real time, based on the changing communication, coordination and work concerns of their members. By providing group awareness, a tool of this kind will in fact also modify and extend the context of the individual users with knowledge about their group participation at any point in time. Such a tool will be useful for the group members themselves, for any roles involved in the management of the community of participants in the socio-technical system, and for researchers investigating emergent group dynamics. Our analytical framework is a necessary foundation for this work.

The second step can be seen as an immediate follow-up to the group awareness support tool above. It will involve including in each user's profile not only timely updates regarding her group participation, but also the associated group context information. This would augment the personalization information about each user with a social dimension that describes the collaborative side of the user's activity within the socio-technical system she operates in. Such augmented information could be used to achieve a level of asynchronous *social translucence* (Erickson & Kellogg, 2000). Social translucence is an augmented form of awareness: a user would be aware not only of the identity of other users who are members of her same emergent groups, but also of the work being undertaken by each of those members and how it relates to her own work. Technically, those social translucence features could take the form of a *trading zone* (Galison, 1999) that is customized to the each emergent group, and whose content changes as the group evolves in its structure, as well as with respect to the changing collaboration interests and requirements of the group.

7. CONCLUSION

We have postulated that electronic traces of computer-mediated activity in emergent groups are semantically rich enough to enable the derivation of contextual information about the work of organization members. Hence trace data can be distilled into task context data using our model of Group Informatics, the analytical framework presented here and the rigorous application of multiple methods to each domain for which context adaptive tools

are being developed. Common trace data is not the equivalent of a common model to support context adaptivity. We show the benefits of recognizing both the social and the technical processes embodied in a socio-technical system for building context adaptive tools. Our analytical framework can be used to detect groups within these new forms of socio-technical organizations, which, in turn, contributes important social profile data back to user models and personalization, enabling context adaptive systems with a substantive social dimension. Technologically mediated groups are in general emergent and dynamic.

8. REFERENCES

Alba, R. D. (1973). A graph‐theoretic definition of a sociometric clique†. *Journal of Mathematical Sociology*, *3*(1), 113-126.

Amelung, C. (2007). Using Social Context and E-Learner Identity as a Framework for an E-Learning Notification System. *International Journal on E-Learning*, *6*(4), 501-517.

Anaya, A. R., & Boticario, J. G. (2011). Content-free collaborative learning modeling using data mining. *User Model User-Adap Inter*, *21*(1-2), 181-216.

Bandura, A. (1977). *Social Learning Theory.* Englewood Cliffs, NJ: Prentice-Hall.

Blincoe, K., Valetto, G., & Goggins, S. (2012). *Proximity: a measure to quantify the need for developers' coordination*. Proceedings of the ACM conference on Computer Supported Cooperative Work (CSCW), pp 1351-1360 , 2012, Seattle, WA.

Bock, Husain. (1950). An Adaptation of Holzinger's B-Coefficients for the Analysis of Sociometric Data. *Sociometry*, *13*, 146-153.

Borgatti, S. P., & Everett, M. G. (1997). Network Analysis of Two Mode Data. *Social Networks*, *19*(3), 243-269.

Brown, J. S., & Duguid, P. (2000). *The Social Life of Information.* Cambridge, MA: Harvard Business School Press.

Bunt, A., & Conati, C. (2003). Probabilistic student modelling to improve exploratory behaviour. *User Modeling and User-Adapted Interaction*, *13*(3), 269-309.

Carroll, J. M., Rosson, M. B., Farooq, U., & Xiao, L. (2009). Beyond being aware. *Information and Organization*,. doi:10.1016/j.infoandorg.2009.04.004

Carroll, J. M., Rosson, M. B., Convertino, G., & Ganoe, C. H. (2006). Awareness and Teamwork in Computer Supported Collaborations. *Interacting With Computers*, *18*, 21-46.

Carroll, J. M., Neale, D. C., Isenhour, P. L., Rosson, M. B., & McCrickard, D. S. (2003). Notification and awareness: synchronizing task-oriented collaborative activity. *International Journal of Human-Computer Studies*, *58*, 605-632.

Cataldo, M., Wagstrom, P. A., Herbsleb, J. D., & Carley, K. M. (2006). *Identification of Coordination Requirements: Implications for the Design of Collaboration and Awareness Tools*. Proceedings from CSCW 2006, Banff, Alberta, Canada.

Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J. M., & Welton, C. (2009). MAD skills: new analysis practices for big data. *Proceedings of the VLDB Endowment*, *2*(2), 1481-1492.

Cosley, D., Ludford, P., & Terveen, L. (2003). *Studying the effect of similarity in online task-focused interactions*. Proceedings from Proceedings of the 2003 international ACM SIGGROUP conference on Supporting group work.

Crowston, K., Wiggins, A., & Howison, J. (2010). *Analyzing Leadership Dynamics in Distributed Group Communications*. Proceedings from HICSS-43, Hawaii.

Crowston, K., Wei, K., Li, Q., & Howison, J. (2006). *Core and periphery in Free/Libre and Open Source software team communications*. Proceedings from System Sciences, 2006. HICSS'06. Proceedings of the 39th Annual Hawaii International Conference on.

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal Complex Systems*, *1695*(1695).

Dourish, P., & Button, G. (1996). Technomethodology: Paradoxes and Possibilities. *Proceedings of CHI96 Human Factors in Computing Systems,* 13-18.

Dourish, P. (2001). *Where the Action Is: Foundations of Embodied Interaction.* Cambridge, MA: MIT Press.

Dourish, P. (2003). Where the Footprints Lead: Tracking Down Other Roles for Social Navigation. In *Designing Information Spaces: The Social Navigation Approach* (pp. 273-292). New York, NY: Springer.

Dourish, P. (2004). What We Talk About When We Talk About Context. *Personal and Ubiquitous Computing*, *8*, 19-30.

Dourish, P. (2006). *Re-Space-ing Place: "Place" and "Space" Ten Years On*. Proceedings from CSCW '06, Banff, Alberta, Canada.

Erickson, B. H. (1988). The relational basis of attitudes. *Social structures: A network approach*, *99*, 121.

Erickson, T., & Kellogg, W. A. (2000). Social Translucence: An Approach to Designing Systems that Support Social Processes. *ACM Transactions on Computer-Human Interaction*, *7*, 59-83.

Galison, P. (1999). Trading Zone: Coordinating Action and Belief. In *The Science Studies Reader.* New York: Routledge.

Goggins, S., Laffey, J., & Amelung, C. (2011a). *Context Aware CSCL: Moving Toward Contextualized Analysis*. Proceedings from CSCL 2011, Hong Kong.

Goggins, S., Mascaro, C., & Mascaro, S. (2012a). *Relief after the 2010 Haiti Earthquake: Participation and Leadership in an Online Resource Coordination Network*. Proceedings from Computer Supported Cooperative Work, 2012, Seattle, WA.

Goggins, S., Laffey, J., Galyen, K., & Mascaro, C. (2011b). Group Awareness in Completely Online Learning Groups: Identity, Structure, Efficacy and Performance. *International Journal of Computer Supported Cooperative Work*, *Under Review*.

Goggins, S. (2007). *Exploring Small Group Collaboration and Creativity in 3D Virtual Worlds*. Proceedings from ACM Group '07, Sanibel Island, FL.

Goggins, S., & Erdelez, S. (2010). Collaborative Information Behavior in Completely Online Groups. In J. Foster (Ed.), *Collaborative Information Behavior: User Engagement and Communication Sharing.* Hershey, PA: ISI Global.

Goggins, S., Galyen, K., & Laffey, J. (2010a). *Network Analysis of Trace Data for the Support of Group Work: Activity Patterns in a Completely Online Course*. Proceedings from ACM Group 2010, Sanibel Island, FL.

Goggins, S., Laffey, J., & Galyen, K. (2009). *Social Ability in Online Groups: Representing the Quality of Interactions in Social Computing Environments*. Proceedings from IEEE Conference on Computer Science and Engineering, Vancouver, BC.

Goggins, S., Laffey, J., & Tsai, I.-C. (2007). *Cooperation and Groupness: Community Formation in Small online Collaborative Groups*. Proceedings from Proceedings of the ACM Group Conference 2007, Sanibel Island, FL.

Goggins, S. P., Laffey, J., & Gallagher, M. (2011b). Completely online group formation and development: small groups as socio-technical systems. *Information Technology & People*, *24*(2), 104-133.

Goggins, S. P., Laffey, J., Amelung, C., & Gallagher, M. (2010b). *Social Intelligence In Completely Online Groups*. Proceedings from IEEE International Conference on Social Computing, Minneapolis, MN.

Goggins, S. P., Mascaro, C., & Valetto, G. (2012). Group Informatics: A Methodological Approach and Ontology for Understanding Socio-Technical Groups. *JASIS&T*, *Under Review*.

Gonzalez, V. M., Nardi, B., & Mark, G. (2009). Ensembles: understanding the instantiation of activities. *Information Technology & People*, *22*(2), 109-131.

Granovetter, M. (1985). Economic action and social structure: the problem of embeddedness. *American journal of sociology*, *91*(3), 481.

Harrison, S., & Dourish, P. (1996). *Re-place-ing space: the roles of place and space in collaborative systems*. Proceedings from Proceedings of the 1996 ACM conference on Computer supported cooperative work.

Howison, J., Wiggins, A., & Crowston, K. (2012). Validity Issues in the Use of Social Network Analysis with Digital Trace Data. *Journal of the Association of Information Systems*, *12(2)*(2).

Kaptelinin, V., & Nardi, B. A. (2006). *Acting with Technology: Activity Theory and Interaction Design (Acting with Technology)*. The MIT Press.

Kay, J. (1995). The um toolkit for reusable, long term user models. *User Modeling and User-Adapted Interaction*, *4*(3), 149-196.

Kersten, G., & Murphy, G. C. (2006). *Using Task Context to Improve Programmer Productivity*. Proceedings from FSE.

Kobsa, A. (2001). Generic user modeling systems. *User modeling and user-adapted interaction*, *11*(1), 49-63.

Kobsa, A. (2007). *Generic user modeling systems*. Proceedings from The adaptive web.

Konstan, J. A., & Reidl, J. (2003). Collaborative Filtering: Supporting Social Navigation in Large, Crowded Infospaces. In *Designing Information Spaces: The Social Navigation Approach* (pp. 43-82). New York, NY: Springer.

Laffey, J., Amelung, C., & Goggins, S. (2009). A Context Awareness System for Online Learning: Design Based Research. *International Journal on E-Learning*, *8*(3), 313-330.

Lee, C. P. (2007). Boundary Negotiating Artifacts: Unbinding Routine of Boundary Objects and Embracing Chaos in Collaborative Work. *Computer Supported Cooperative Work*, *16*, 307-339.

Lynch, C. (2008). Big data: How do your data grow? *Nature*, *455*(7209), 28-29.

Maloney-Krichmar, D., & Preece, J. (2005). A multilevel analysis of sociability, usability, and community dynamics in an online health community. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *12*(2), 201-232.

Mascaro, C., & Goggins, S. P. (2011). *Brewing Up Citizen Engagement: The Coffee Party on Facebook*. Proceedings from Communities & Technologies, 2011, Brisbane, Australia.

Mitchell, J. C. (1969). *Social networks in urban situations: analyses of personal relationships in Central African towns.* Humanities Press Intl.

Moody, J., & White, D. R. (2003). Structural cohesion and embeddedness: A hierarchical concept of social groups. *American Sociological Review,* 103-127.

Nardi, B. A., & O'Day, V. (2000). *Information ecologies: Using technology with heart.* The MIT press.

Nardi, B. (1996). *Context and Consciousness.* Boston, MA: Massachusetts Institute of Technology.

Nardi, B. (2007). Placeless Organizations: Collaborating for Transformation. *Mind, Culture and Activity*, *14*(1-2), 5-22.

Nardi, B., Whittaker, S., & Schwarz, H. (2002). NetWORKers and their Activity in Intensional Networks. *Computer Supported Cooperative Work*, *11*, 205-242.

Rohde, M., Reinecke, L., Pape, B., & Janneck, M. (2004). Community-Building with Web-Based Systems - Investigating a Hybrid Community of Students. *Computer Suppoerted Cooperative Work*, *13*, 471-499.

Rohde, M., & Shaffer, D. W. (2003). Us, Ourselves and We: Thoughts about Social (Self-) Categorization. *SIGGROUP Bulletin*, *24*(3), 19-24.

Seidman, S. B., & Foster, B. L. (1978). A graph‐theoretic generalization of the clique concept*. *Journal of Mathematical sociology*, *6*(1), 139-154.

Star, S. L., & Griesemer, J. R. (1989). Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science*, *19*, 387-420.

Terveen, L., & McDonald, D. W. (2005). Social Matching: A Framework and Research Agenda. *ACM Transactions on Computer-Human Interaction*, *12*(3), 401-434.

Wagstrom, P., Herbsleb, J. D., Kraut, R., & Mockus, A. (2010). *The Impact of Commercial Organizations on Volunteer Participation in an Online Community*. Proceedings from Academy of Management Annual Meeting, Montreal, Canada.

Wenger, E. (1998). *Communities of Practice: Learning, Meaning and Identity.* New York: Cambridge University Press.

Zimmermann, A., Specht, M., & Lorenz, A. (2005). Personalization and Context Management. *User Model User-Adap Inter*, *15*(3-4), 275-302.

8. AUTHOR BIOGRAPHIES

Sean P. Goggins, Ph.D joined the College of Information Science and Technology at Drexel University in Philadelphia, PA as an assistant professor in September 2009. Previously he had been involved in industrial software engineering and product development in the medical device, data mining, software and publishing industries since 1993. He holds a B.S. in History from the University of Wisconsin, an M.S. in Software Engineering from the Computer Science department at the University of Minnesota and a Ph.D in Information Science and Technology from the University of Missouri. His research interests are focused on building context adaptive spaces to support distributed group work. Specifically, he seeks to build understanding of the emergence, development and dissolution of technologically mediated groups in both walled gardens like those found in online learning and software engineering, and open public spaces, like those found on social media platforms. With this understanding, he is developing a methodological approach to research that focuses on interactions, performance, identity and discourse through technology.

Giuseppe (Peppo) Valetto, Ph.D joined the Department of Computer Science at Drexel University as an assistant professor in Software Engineering in September 2007. Previously, he has been involved in industrial and academic research since 1994, working at Xerox Research in Grenoble (France), CEFRIEL – Politecnico di Milano (Italy), Telecom Italia Lab in Torino (Italy), and at IBM Research in Hawthorne, NY. He holds a Laurea in Electronic Engineering from the Politecnico di Torino (Italy), and an MS and Ph.D. in Computer Science from Columbia University, New York, NY. His research interests include tools, methods and processes for enhancing cooperation and coordination in collaborative software development, in particular in large-scale, distributed software projects. He is also interested in the engineering of autonomic and self-adaptive software systems, with a focus on algorithms and mechanisms for the self-organization of large-scale software systems, like peer-to-peer networks.

Christopher Mascaro is a doctoral candidate in Information Studies at Drexel University. He is studying under Sean P. Goggins. He received his B.A. in Political Science from The University of Michigan in 2000 and his M.A. degree in Government and Political Communication from Johns Hopkins University in 2010. His research focuses on technologically mediated group formation and how individuals in these groups interact, form identity, participate in discourse and evolve structurally over time, especially in the political domain.

Kelly Blincoe is a doctoral candidate in the Department of Computer Science at Drexel University. She is studying under Giuseppe Valetto. She holds a BE in Computer Engineering from Villanova University, an MS in Information Science from Pennsylvania State University, and an MS in Computer Science from Drexel University. Her research interests include methods for improving the coordination within software development teams.