

# Community Trajectory: Discovery of Evolutionary Collaboration Patterns Based on Event Co-participation

Yu-Ru Lin   Hari Sundaram   Aisling Kelliher

Arts Media and Engineering Program, Arizona State University, Tempe, AZ 85281

e-mail: {yu-ru.lin, hari.sundaram, aisling.kelliher}@asu.edu

## ABSTRACT

This paper presents a framework for analyzing and summarizing collaborative activities in a highly dynamic, context-rich social network. In everyday work, collaboration is essential, and people need to be aware of the activities of their peers to find opportunities for collaboration. Understanding activities of one's peers become particularly difficult when people collaborate across time, locations or even disciplines. We propose to support collaboration by extracting and representing collaborative patterns, thus improve awareness. There are three key contributions in this paper: (a) a probability model for extracting the clustering structure of social interaction within a collaborative event context, (2) collaboration analysis for determining key people, events, and the strength of ties based on conditional probabilities, and (3) a dynamic, interactive visual representation that reveals "community trajectory" which supports exploration of temporal collaborative patterns. Our experiments using real-world collaborative event data reveals subgroup evolution and typical characteristics of collaborative projects, indicating the utility of our approach for supporting reflection on collaborative activities.

## Categories and Subject Descriptors

H.3.3 [Information Systems]: *Information Search and Retrieval*;

H.3.5 [Information Systems]: *Online Information Services*;

H.5.4 [Information Systems]: *Hypertext/Hypermedia*; J.4

[Computer Applications]: *Social and Behavioral Sciences*

## General Terms

Experimentation, Measurement, Algorithms, Human Factors.

## Keywords

Community extraction, temporal dynamics, collaboration, event, clustering, visualization, visual summarization, social network analysis

## 1. INTRODUCTION

This paper presents a framework for analyzing and summarizing collaborative activities in a highly dynamic, multi-disciplinary, context-rich social network. In everyday work, people need to be aware of the activities of their peers in order to find opportunities or solutions for collaboration. However, the awareness is hard to maintain in multi-disciplinary environments, as well as when the collaborators have different tie-schedules or may work at different locations.

*Collaborative awareness* refers to having knowledge of activities and events in the collaborative workspace. It involves knowing people relevant to one's work, when one's co-workers are available or busy, the activities in which they are engaged, current collaborative groups and their collaborative history. All of these factors inform our decision making during collaboration practice.

For example, a faculty member who finds it difficult to schedule time to meet with each of his advisees may look for intelligent scheduling software to optimize his meeting time with each student. Alternatively, if he notices stronger interaction among several of his students, he could optimize his schedule by meeting with them as a group. A creative solution could emerge through reflection upon collaborative patterns.

There is limited prior work on tools or techniques that reveal collaborative patterns. Work on tools supporting collaborative work focus on reducing conflicts over shared resources such as meeting times or spaces [2,3]. These techniques do not support reflection on collaboration because they do not enable users to see or track the temporal patterns of their collaborative activities. Social interaction is fundamental to collaboration and there has been work on analyzing social interaction by extracting cohesive subgroups in networked data. Recent literature on analyzing evolution of subgroups in dynamic social networks [4,7,8]. These techniques focus on modeling and characterizing the social interaction independent of the event context. This is a key limitation since reflection on collaborative activities requires understanding the context in which they occur. There has also been work on providing spatial or temporal navigation of user created events [1,6], but they are not applicable for review of large corpuses of group events over extended periods of time.

We propose a novel framework for analyzing and summarizing collaborative activities that allows users to reflect upon their collaborative patterns. There are three key contributions:

1. A unified framework to extract the clustering structure of social interaction within a collaborative event context. We propose to extract collaborative structures via clustering of event data. To extract the clusters, we develop a probabilistic model that factors the joint distribution of people and events.
2. A method for identifying important people and events, as well as interactions between people. We determine the importance of people and events through their probabilities conditioned on a cluster. To capture the interaction between two people, we define *collaborative tie* as conditional probability of one given another.
3. A dynamic, interactive visual representation that reveals "community trajectory" which supports exploration of temporal collaborative patterns.

Our experiments using real-world collaborative event data reveals subgroup evolution and typical characteristics of collaborative projects.

The rest of the paper is organized as follows: In the next section we describe our method for analyzing structures of people and events within a collaborative context. In section 3 we present the collaborative activity representation – community trajectory. We show experimental results in section 4 and conclude in section 5.

## 2. COLLABORATION STRUCTURES

We now examine the collaboration structures emerging in an everyday work environment. We consider the set of events involved in collaboration process, e.g. a kickoff meeting for a project, a progress review meeting, etc. An *event* refers to a real-world occurrence, which may be described using attributes such as who, where, when, what or by media such as images [10]. We refer to such attributes as the event *context* – they support the understanding of everyday events. An event occurring in collaboration process usually includes multiple co-working participants. The co-presence of people in an event is a key aspect of collaboration. Therefore, we shall examine the collaboration structures using the “who” attribute, i.e. who participates in certain events.

We examine collaboration structures by comparing a sequence of events occurring over time. Let us consider a sequence of  $M$  events  $E = \{e_1, \dots, e_M\}$ , where each event  $e$  is associated with a timestamp  $t(e)$ . Two events  $e_i$  and  $e_j$  can be compared through a similarity function  $s(e_i, e_j)$ . In general, event similarity can be measured by comparing *any* aspect of the event context, but since we concentrate on the “who” attribute, we define event similarity over a set of  $N$  users  $U = \{u_1, \dots, u_N\}$ . We use a standard measure, Jaccard’s index, to measure the similarity of participants in two events  $e_i$  and  $e_j$ , thus  $s(e_i, e_j) = |U(e_i) \cap U(e_j)| / |U(e_i) \cup U(e_j)|$ , where  $U(e)$  is the set of users participating in event  $e$ . The Jaccard’s index measures the common participants of the two events relative to the total participants of both.

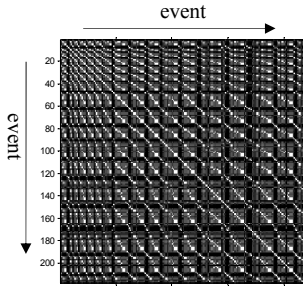


Figure 1: similarity matrix of events using Jaccard’s index of participants.

The similarity matrix provides a key insight: In everyday work, collaborative events having common participants re-occur over time. Consequently, people can be clustered due to co-occurring in common events. The observation enables us to develop a compact representation of collaborative activities.

### 2.1 Factorization of event co-participation

We propose to extract the structures of collaboration by decomposing the co-presence of people and events into clusters – specifically we decompose the joint probability of people and events. We assume that given a cluster, the events associated with the cluster, and any two people co-participate in the events, are conditionally independent. We can consider observing an instance

with two users  $u_i$  and  $v_j$  co-participating in a collaborative event  $e_l$  through a generative process, as illustrated in Figure 2:

1. Choose a user  $u_i$  with probability  $p(u_i)$ .
2. Choose a cluster  $c_k$  which  $u_i$  is mostly likely to belong to according to the conditional probability  $p(c_k|u_i)$ .
3. Choose another user  $v_j$  who is mostly likely to be seen in the cluster according to the conditional probability  $p(v_j|c_k)$ .
4. Finally choose an event  $e_l$  which is mostly likely associated with the cluster according to  $p(e_l|c_k)$ .

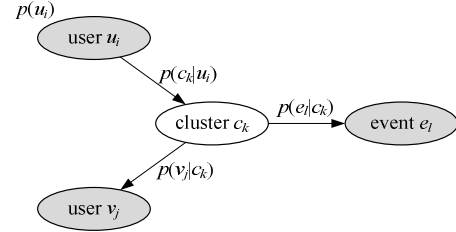


Figure 2: A generative model for event based clustering. User  $u_i$ ,  $v_j$  and event  $e_l$  are observed (shaded circles), and the cluster  $c_k$  is not observed.

In this model, the co-occurrence of  $u_i$ ,  $v_j$  and  $e_l$  is represented as:

$$p(u_i, v_j, e_l) = \sum_{k=1}^K p(u_i) p(c_k | u_i) p(v_j | c_k) p(e_l | c_k) \quad \langle 1 \rangle$$

$$= \sum_{k=1}^K p(c_k) p(u_i | c_k) p(v_j | c_k) p(e_l | c_k)$$

This equation implies a user is implicitly associated with a cluster because she tends to interact with the cluster member or members in the cluster events. The more she interacts with the cluster members, the more likely she belongs to the cluster. Thus how likely a user  $u_i$  belongs to a cluster  $c_k$  is affected by how much she interacts with people in the cluster while participating in events associated with the cluster. The likelihood is given by conditional probability  $p(c_k|u_i)$ , which indicates, given a user  $u_i$ , how likely she belongs to a cluster  $c_k$ . We then assign user  $u_i$  to group  $c_k$  if  $k = \text{argmax}_k p(c_k|u_i)$ . Similarly, we determine an event  $e_l$  to be associated with a cluster  $c_k$  if  $k = \text{argmax}_k p(c_k|e_l)$ .

We can estimate the model parameters using a standard expectation-maximization (EM) algorithm to maximize the log likelihood of data. The log likelihood of the event instances is:

$$L = \sum_i^N \sum_j^N \sum_l^M n_{ijl} \log \sum_{k=1}^K p(c_k) p(u_i | c_k) p(v_j | c_k) p(e_l | c_k) \quad \langle 2 \rangle$$

where  $n_{ijl}$  denotes the number of co-occurrences of  $u_i$ ,  $v_j$  and  $e_l$ . An event such as weekly meeting might have multiple instances.

The proposed probabilistic model is similar to the probabilistic latent semantic model (pLSI) [5] used in document clustering. pLSI represents the joint distribution over documents and words. However, instead of factorizing the joint distribution of the co-occurrences of a user and an event, i.e.  $p(u_i, e_l)$ , we factorize the joint distribution of the co-occurrences of two users co-participating in a event, i.e.  $p(u_i, v_j, e_l)$ . Our model assumes, not only the co-occurrence of a user and an event, but that the social interaction between two users, in terms of event participation, is conditionally independent given a cluster. This allows us to derive estimation of social interaction between two users suitable for the context of collaboration, as will be discussed next.

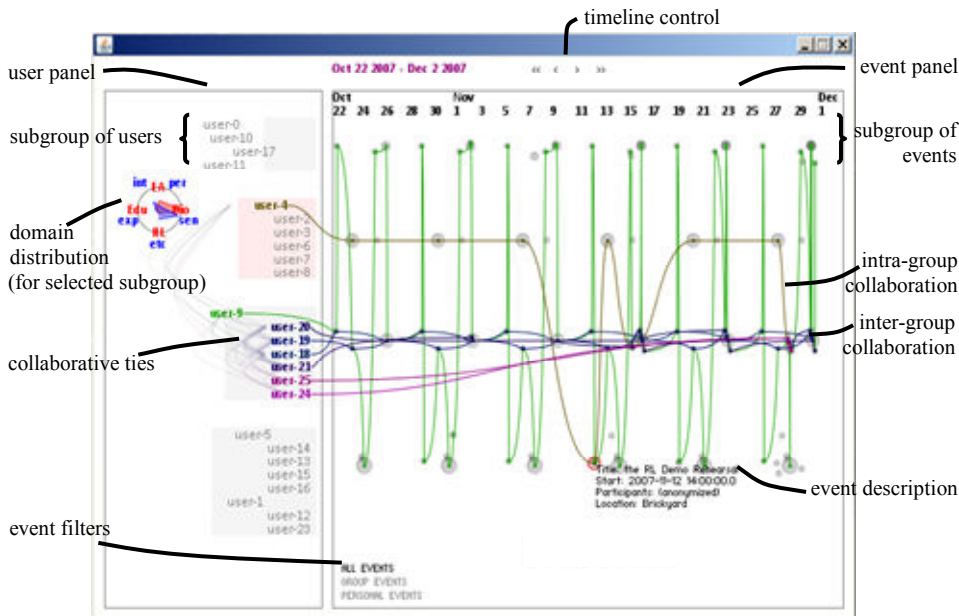


Figure 3: Our proposed representation summarizes collaborative activities in a complex collaborative environment (AME).

## 2.2 Collaboration analysis

Based on the clustering structure extracted by the proposed model, we seek to answer questions critical to collaborative awareness: (1) Who is the key person? (2) What is an important collaborative event? (3) How strong is the collaboration between two users? First, we conjecture a key person from a cluster should be involved in many events associated with the cluster because the fact that a user participates in a group of related events implies his or her participation is *needed* for collaboration. Thus we can determine the importance of a user  $u_i$  in a cluster  $c_k$  by the conditional probability  $p(u_i|c_k)$ . One measure of event importance can depend on the amount of participation. Thus we define the importance of an event  $e_i$  in a cluster  $c_k$  by the conditional probability  $p(e_i|c_k)$ . This analysis enables us to identify the representative people or events out of many collaborative activities.

We observe collaboration between two people is usually asymmetric. For example, a member in a group might interact mostly with the group leader, while the group leader might interact uniformly with the whole group. To capture this, we define the *collaborative tie* to each user  $v_j$  for an individual user  $u_i$  as follows:

$$p(v_j | u_i) = \frac{\sum_{k=1}^K p(c_k) p(u_i | c_k) p(v_j | c_k)}{\sum_{k=1}^K p(c_k) p(u_i | c_k)} \quad <3>$$

The collaborative tie measures the amount of interaction between user  $u_i$  and  $v_j$  derived from collaborative activities, normalized by those activities involving  $u_i$ . Collaborative tie can be interpreted as opportunity of collaboration between two people, where the collaboration might involve other people. The normalization allows a user to compare the strength of tie over all of his or her collaborators.

(1) **Event panel:** The set of events occurring in the given duration is shown as bubbles. The horizontal position of the event indicates when the event occurs, and the vertical position indicates which cluster the event is associated with. The size of event bubbles is proportional to the event importance within the cluster, i.e.  $p(e_i|c_k)$ . When a user clicks on an event bubble, the description of the event is shown. (2) **User panel:** The set of users participating in the events occurring in the given duration is shown by their name (anonymized in the figure). Users assigned to the same clusters are grouped together. The horizontal position of a user indicates how likely she belongs to the cluster, i.e.  $p(c_k|u_i)$ . Within each cluster, the vertical position of the user indicates the user's importance within the cluster, ranked according to  $p(u_i|c_k)$ . Higher ranked users are displayed first.

Next we discuss a visual representation that supports exploring the collaborative patterns based on the extracted clustering structure and the collaboration analysis.

## 3. COMMUNITY TRAJECTORY

This section presents an interactive visual representation that allows users to review their collaborative activities over time. The goal of our representation is to investigate the following questions, which are critical to maintaining collaborative awareness: (1) When are my colleagues busy or available? (2) What activities are they engaged in? (3) Who collaborates with whom? How strong is the collaboration across different groups? (4) How does our collaboration change? A group calendar cannot effectively answer these questions because the amount of events and the complex interdependency of social interaction obscure patterns.

We develop a compact representation called community trajectory that seeks to answer these questions by showing the co-occurrences of events and people associated with extracted clustering structure. A snapshot of the proposed representation is shown in Figure 3. We take our research program, AME, as an example to demonstrate the functionality of the representation. The research collaboration in AME is cross-disciplinary. It can be described using a matrix, where the rows comprise four major research areas and the columns comprise four major application areas. People from different research areas work together in specific application areas and each of them can be assigned to a corresponding row and column.

The dynamic representation is constructed in the following way: Given a time duration  $T$ , we extract the set of events  $E_T$  occurring during  $T$  and the set of users  $U_T$  participating in  $E_T$ . We extract a set of  $K$  clusters from  $E_T$  and  $U_T$  by using the proposed clustering method. As in Figure 3, we show  $U_T$  in the *user panel* and  $E_T$  in the *event panel*, organized by the extracted clusters. The representation of collaborative activities dynamically changes depending on the given time duration  $T$ .

When clicking on a user name, her collaborating history is shown as a curve intersecting all events she participated in during the given time. Multiple users' activities can be compared by selecting more users. The user activity curve reveals (1) how active a user is, (2) how much the user participates in or across group activities, (3) how two or more users collaborate within a group or across different groups. When multiple users are selected, their collaborative ties, i.e.  $p(v_j|u_i)$  are displayed in the user panel. A thicker curve between two names indicates stronger collaborative interaction during the time.

User can shift or zoom the time window to see events at different time scales (weekly, monthly or in arbitrary duration). As users change the time duration, the displayed events, user names and their position, as well as the activity curve change with respect to the given time. Thus the proposed representation enables users to reflect upon their collaboration at different granularities.

In addition, the characteristics of the extracted clusters can be displayed using tools such as radar charts. The radar charts represent the *domain distribution* based on the number of members from each domain of application area (e.g. "RL", "Bio", etc.) or research area (e.g. "sen", "per", etc.). The information about application/research areas for individuals can be extracted from public user profiles. The radar chart thus reveals collaboration across disciplines.

In the next section we discuss our experimental results using real-world collaborative event data.

## 4. EXPERIMENTS

This section presents our experimental results with real-world collaborative event data collected within the AME research groups.

**Dataset description.** We collect a sample of event data from two sources: (1) *eventory* [9], and (2) personal calendar feeds. Eventory is an online media archive system where users can create, explore and manage events. It is developed to support dynamic and instant interaction among AME members. Secondly, we extract events through subscription of calendar feeds provided by members who use calendar services / tools such as Google Calendar, Microsoft Outlook or Apple iCal. For privacy concerns, we only use the sets of events which users make public, and filter out events with less than two participants. In total, there are 276 events collected from the two sources. We analyze 186 events occurring during Sep. 23 to Dec. 15, 2007, totally 12 weeks. There are 29 different users appearing in these events, including 9 faculty and 20 students. The user base covers about half of all AME members and concentrated on the "RL" and "Bio" application area. The information about application or research areas can be obtained from each user's public research profile.

We apply our clustering methods on this dataset. The dataset contains 4551 user-user-event tuples. The log likelihood initially increases fast and slows down when  $K \geq 4$ , where the increase at  $K = 4$  is less than 1%. We thus select  $K = 4$  and estimate the model parameters to extract four clusters.

We examine the details of the clustering results as in Figure 4(a). The plot is similar to the interactive representation: we render each user as a single curve, but colored according to her application area. (Due to the space limit, we do not show the plot with curves colored according to research areas.) This can be viewed as a "zoom-out" of the interactive representation as in Figure 3, which facilitates our comparison. In Figure 4(a), the

curve intersects at  $(t, k)$  if the user is assigned to the  $k^{\text{th}}$  cluster at week  $t$ . If the user appears in events during week  $t$  but not before week  $t$ , the curve will come from the top because she is not assigned to any of the four groups before week  $t$ . And if the user does not appear in events after week  $t$ , the curve will go to the top after  $t$ . When a strand of curves appears to have the same color, it indicates the cluster members come from the same application area.

Figure 4(a) comprises several useful observations: (1) The thick blue strand indicates a sustained group working for the "RL" application area. This "RL" strand splits at week 3 because one of the faculty directs four of his advisees to form a subgroup for a specific task in an RL project. (2) At week 5, two branches merge into the "RL" strand. This is caused by another faculty establishing a regular meeting schedule with her students; all of them belong to the "RL" project. (3) An orange strand indicating another group working for the project "MC" – a new project initiated recently. It is not described in any user profile. (4) At week 10 a small group at (10, 1) appears. This group emerges due to several meetings occurring at this week regarding issues in the "Bio" project. These meetings involve a faculty and a student who have never appeared in previous events. Both cases suggest that the representation reveals emergent collaborative groups that cannot be directly obtained from static user profiles. (5) The importance of some roles then can also be revealed. Although not obvious in Figure 4(a), the split and merge of the "RL" strand corresponds to the activities of two faculty, both high-ranked people in their group (displayed at top of their clusters in Figure 3). This might suggest that the collaborative structures are more sensitive to the activity of higher ranked users.

We validate the clustering results by comparing with the information about application and research areas. We consider the application / research areas of individuals as class information. We use a standard measure, the normalized mutual information (NMI), as an indicator of the correspondence between the clustering membership and the class information. Given a set of clusters  $C = \{c_k\}$  and a set of classes  $A = \{a_j\}$ , NMI is defined as:

$$NMI(C, A) = \frac{I(C; A)}{(H(C) + H(A))/2} \quad <4>$$

$$I(C; A) = \sum_k \sum_j p(c_k \cap a_j) \log \frac{p(c_k \cap a_j)}{p(c_k)p(a_j)}$$

where  $H(Z) = -\sum p(z) \log p(z)$  for  $z \in Z$ . It is defined as the mutual information between clusters and the classes, normalized by the maximum of marginal entropies. NMI measures the amount of information by which our knowledge about the classes increases when we have the clusters. Higher NMI indicates a higher correspondence between the clusters and the classes (research / application areas).

Figure 4(b) shows the NMI values for the clusters derived from events occurring during each week. At the end of the curves we plot the NMI values for overall clustering according to events occurring during the entire 12 weeks. As can be seen, the NMI values between the clustering results and the information about application areas are higher than those about research areas. This implies the collaborative activities are driven more by people working together in application areas. This is consistent with the coherent colors of curves in Figure 4(a). The merges and splits of the strands in Figure 4(a) also correspond to the rise and fall of NMI application curve in Figure 4(b). We notice the two temporal

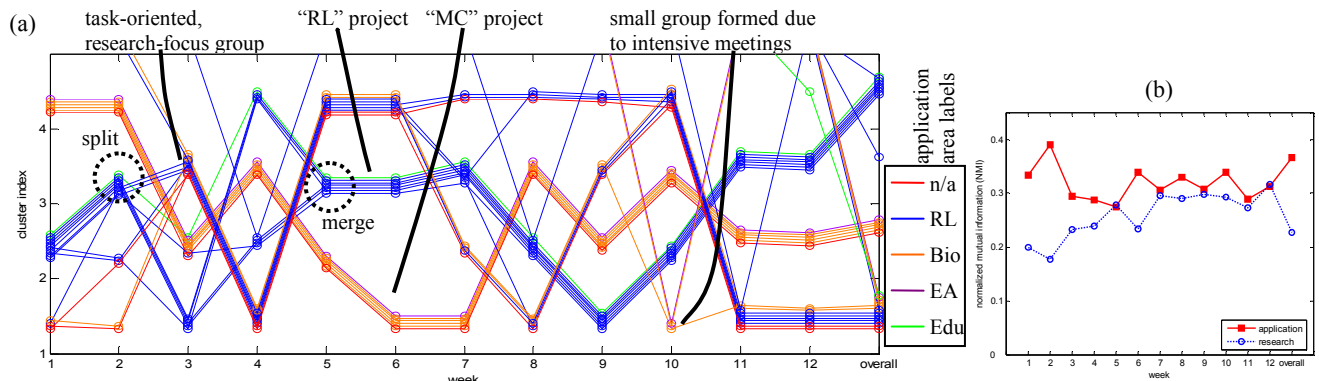


Figure 4: (a) Clustering results of interaction based method over 12 weeks, for  $K = 4$ . Each single curve indicates a user, colored according to the user's application areas. The curve intersects at  $(t, k)$  if the user is assigned to cluster  $k$  at week  $t$ . (b) Normalized mutual information (NMI) values between extracted clusters and application / research areas, over 12 weeks.

NMI curves exhibit different trends for application and research information. These counter trends are reasonable in AME because the AME application areas are developed to deliberately include people with different or even complementary research interests.

We note that the information about application / research areas is static and cannot reflect the complete picture of collaboration in practice. Our proposed dynamic representation of collaborative patterns provides a way for users to reflect on their collaboration process.

## 5. CONCLUSION

In this paper, we propose a new framework to support reflection upon collaborative activities. The framework comprises (1) a method for extracting the clustering structure of social interaction within a collaborative event context, (2) collaboration analysis for identifying key people and events, as well as the strength of ties derived from collaborative activities, and (3) the interactive community trajectory representation that allows exploring temporal collaborative patterns. Our experimental study suggests useful observation by recognizing the collaborative patterns from the visual representation.

**Discussion and future work.** We obtain event instances through the calendar tools. However, the usage of calendar is contextual – for example, a user might only specify the weekly meetings or class schedule in their calendar, while another user might update his calendar very often to manage both regular and accidental events. Usually, users specify planned events, while unplanned events are rarely recorded in the calendar. These make difficult to systematically capture the actual events. We also note the limitation about the co-occurrences based metrics for the importance of an event. Sometimes an event involved few people can have important impact on other events. As part of our future work, we will improve our data acquisition approach. To identify significant collaborative events, we plan to incorporate analysis on outcomes of collaboration, e.g. the generated document or multimedia data. We also plan to conduct a comprehensive user study to evaluate the proposed framework.

## 6. REFERENCES

- [1] P. APPAN and H. SUNDARAM (2004). *Networked multimedia event exploration*. Proceedings of the 12th annual ACM international conference on Multimedia: 40-47.
- [2] J. BIEHL, M. CZERWINSKI, G. SMITH and G. ROBERTSON (2007). *FASTDash: a visual dashboard for fostering awareness in software teams*. Proceedings of the SIGCHI

- conference on Human Factors in computing systems: 1313-1322.
- [3] M. BRZOWSKI, K. CARATTINI, S. KLEMMER, P. MIHELICH, J. HU and A. NG (2006). *groupTime: preference based group scheduling*. Proceedings of the SIGCHI conference on Human Factors in computing systems: 1047-1056.
- [4] T. FALKOWSKI, J. BARTELHEIMER and M. SPILIOPOULOU (2006). *Mining and Visualizing the Evolution of Subgroups in Social Networks*, International Conference on Web Intelligence, 2006., 52-58,
- [5] T. HOFMANN (1999). *Probabilistic latent semantic indexing*. Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval: 50-57.
- [6] T. KAPLER and W. WRIGHT (2005). *GeoTime Information Visualization*. Information Visualization 4(2): 136-146.
- [7] Y.-R. LIN, H. SUNDARAM, Y. CHI, J. TATEMURA and B. TSENG (2007). *Blog Community Discovery and Evolution Based on Mutual Awareness Expansion*, 2007 IEEE/WIC/ACM International Conference on Web Intelligence, November 2007.
- [8] G. PALLA, A. BARABASI and T. VICSEK (2007). *Quantifying social group evolution*. eprint arXiv: 0704.0744.
- [9] X.-J. WANG, S. MAMADGI, A. THEKDI, A. KELLIHER and H. SUNDARAM (2007). *Eventory - An Event Based Media Repository*, IEEE International Conference on Semantic Computing, Irvine, CA, Sep. 2007.
- [10] A. ZUNJARWAD, H. SUNDARAM and L. XIE (2007). *Contextual wisdom: social relations and correlations for multimedia event annotation*. Proceedings of the 15th international conference on Multimedia: 615-624.