Multilevel Analysis of Uptake, Sessions, and Key Actors in a Socio-Technical Network

Dan Suthers

Dept. of ICS, University of Hawaii

1680 East West Road, POST 317

Honolulu, HI 96822

1.808.956.3890

suthers@hawaii.edu

Nathan Dwyer

Ibis Networks

828 Fort Street Mall, Suite 600

Honolulu, HI 96813

1.808.485.9044

ndwyer@gmail.com

ABSTRACT

Learning in social settings is a complex phenomenon that involves multiple processes at individual and collective levels of agency. Thus, a richer understanding of learning in socio-technical networks will be furthered by analytic methods that can move between and coordinate analyses of individual, small group and network level phenomena. This paper outlines our implementation of an analytic framework intended to address these and other needs (e.g., integrating fragmented traces of activity into one analytic artifact), and gives an example using data from the Tapped In educator professional network. The methods build on empirical relationships between events to build a graph of uptake relations—how one event builds on another, which are then used to identify sessions in the space-time dimensions of a rich online environment, identify key actors within sessions using sociometrics, and find relationships between sessions that might be vectors for the transmission of ideas or practices.

Categories and Subject Descriptors

K.3 [**Computers and Education**]: General

Keywords

Networked learning, socio-technical networks, multi-level analysis, interaction analysis, social network analysis.

1. INTRODUCTION

The computational methods for analysis of learning in networked communities presented in this paper are based on a view of learning as a complex phenomenon. Theories of how learning takes place in social settings vary in the agent of learning (e.g., individual, small group, or community), and in the process or 'mechanism' of learning [16] (e.g., information transfer or knowledge communication [4], intersubjective meaning-making such as argumentation and co-construction [1, 15], shifts in participation and identity [8, 12], and accretion of cultural capital [13]). Learning takes place at all of these levels of agency and with all of these processes. Thus, understanding learning in its full richness rather than for a narrow academic purpose requires examining data that reveal the relationships between individual and collective levels of agency and potentially coordinating multiple methods of analysis [19].

This presents one analytic challenge that motivates the work reported here: how to meaningfully connect multiple levels of analysis. To preview, we match for their complementary strengths and weaknesses both interpretative and computational interaction analysis, which enable us to see what groups of individuals are doing and how they are doing it but at a level of detail that obscures larger scale patterns, with (social) network analysis, which provides summaries of ties and affiliations in a form amenable to drawing conclusions about network or community patterns but loses the details of how people actually interact.

The networked learning communities we study are technologically embedded. We examine activity in an environment that offers asynchronous threaded discussions, quasi-synchronous chat, file sharing, and other media for interaction. Participants include members of organizations and others in scheduled events and persons who come to the socio-technical network of their own accord; and in both cases, participants are free to wander between specific settings and events. Consequently, the data related to a given participant results in different kinds of traces in the log files associated with these media, at different times and different virtual spaces. The trace of what for a given participant was a unitary experience is fragmented across these logs, and needs to be reassembled to reveal this activity. This is the second analytic challenge addressed by our approach: how to reassemble fragmented traces into a single analytic artifact. To preview, log events are abstracted and merged into a single abstract transcript of events, and this is used to derive a series of representations that support levels of analysis of interaction (contingency and uptake graphs, and session graphs) and of ties ("associograms" [20] and sociograms).

Other publications have detailed some of the theory [7, 16] and analytic representations [17, 20] behind this work. In this paper we report on progress providing computational methods for transforming log files into interaction and session graphs and sociograms, and means of drawing conclusions based on these representations. First, we provide an overview of the series of computational transformations taken. Then an extended example illustrates the methods using the Tapped In data.

2. OVERVIEW OF THE FRAMEWORK

Due to space constraints, we can only give an overview of our analytic framework here: for detailed explanation we refer readers to [17, 20]. The representations used at various levels of analysis are shown in Figure 1. At the bottom we begin various traces of activity such as log files of events. (HTTP logs are shown just for illustration: our actual data includes database logs and textual transcripts of chats.) These are parsed using methods that are necessarily system-specific to abstract meaningful events into an event stream, shown in the second level. This event stream is represented using an abstract transcript format that records actor and media participants in events.
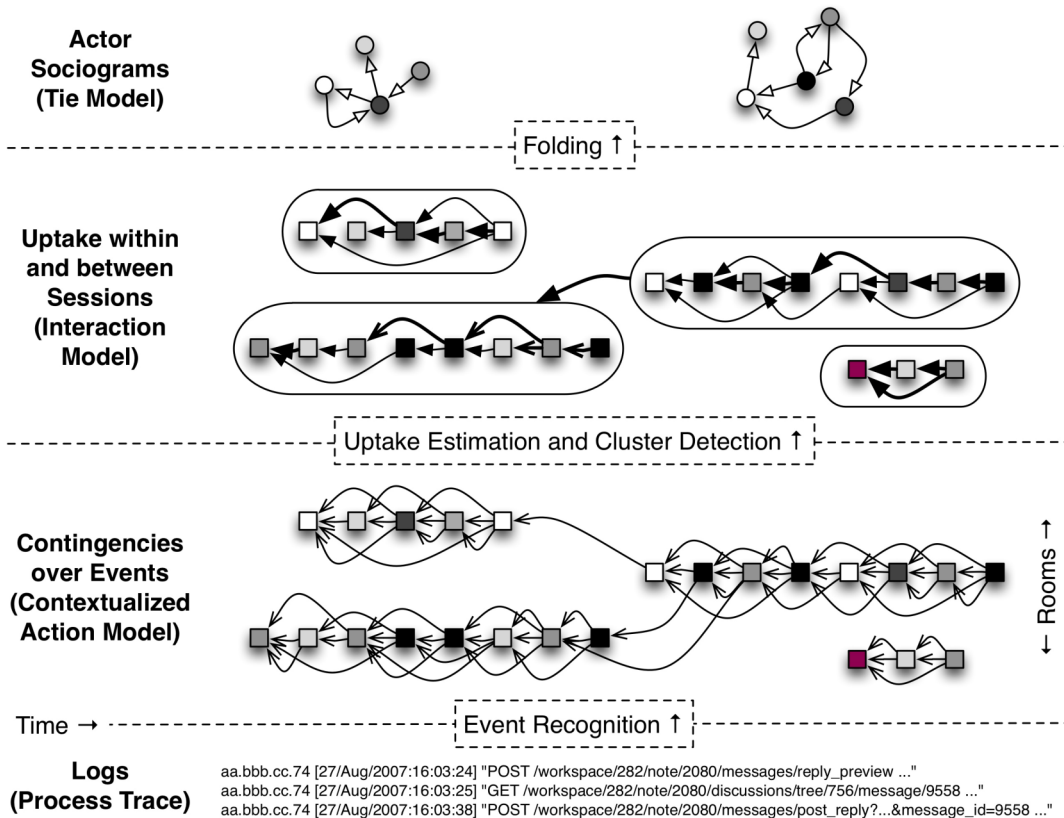
Figure 1. Levels of Analysis and their Representations.

Then at this level we compute ***contingencies*** between events, to produce a model of how acts are mutually contextualized. Human action is contingent upon its setting in diverse ways: our computational methods capture some of these contingencies that are amenable to automated detection. For example, a *temporal contingency* reflects the likelihood that events occurring close together in time are related. For example, in analyzing quasi-synchronous chat contingencies are installed to prior contributions that occur within an adjustable time window but not too recently (under a Keystroke Level Model of how long an expert typist could have typed the contribution, to ensure that the taken up contribution was already visible [11]). *Address* and *reply contingencies* are installed from an utterance mentioning a user by name to the last contribution (address) and next contribution (replay) by that participant within a time window. Contingencies are installed to *prior acts* of a participant over a larger time window to reflect the continuity of an agent's purpose. Overlap in content as represented by sets of lexical stems is used to produce a *lexical contingency* weighted by the number of terms overlapping. The resulting ***contingency graph*** is represented in our own Traces framework, the Entity-Event-Contingency graph or EEC, which permits multiple edges between vertices (events).

Most graph algorithms assume at most only one edge between two vertices. Also, and more importantly, we are interested in ***uptake***, the relationship between events in which a human action takes up some aspects of prior events as being significant in some manner. For example, replying to prior contributions in a chat or discussion is an example of uptake, but the concept of uptake is not specific to a medium (it can cross media) or limited to transactivity (one can uptake without being "other-directed"). Contingencies are of interest only as evidence for uptake. So, we abstract the contingency graph to an uptake graph, using a weighted (and presently linear) combination of the various types of contingencies between contributions (vertices) to derive a single uptake relation represented as a graph edge weighted to reflect strength of evidence in the contingencies. In our framework, different weights can be used for different purposes (e.g., finding sessions; analyzing the interactional structure of sessions): the weighting schemes are declared using XML as shown in Table 1.

As shown in the next level of Figure 1, ***uptake graphs*** are similar to contingency graphs in that they also relate events, but they collect together bundles of contingencies into uptake relations, optionally filtering out low-weighted bundles. At this point, we can do several interesting things with these uptake graphs. A graph clustering or "community detection" algorithm (e.g., modularity partitioning, [3, 5]) is then applied to the uptake graph to find clusters of related contributions that we call "sessions". A session can cross settings such as chat rooms. Inter-session and intra-session analysis procedes from here.

For inter-session analysis, we collapse each session into a single vertex representing the session, but retain the inter-session uptake links. These inter-session links indicate potential influences across time and space from one session to another. An example will be given shortly.

2

Table 1. A weighting scheme for combining contingencies into estimations of uptake.
(Lexical overlap is handled separately, weighting proportional to the number of overlapping lexical stems.)

```xml
<?xml version="1.0" encoding="UTF-8"?>
<weighter bundlename="apps.analyzer" classname="apps.output.weighting.StandardWeighter">
    <weighting defaultweight="1" threshold="2">
            <entry typesuffix="AddressContingency"              weight="3" />
            <entry typesuffix="ReplyContingency"                weight="3" />
            <entry typesuffix="MediaContingency"                weight="3" />
            <entry typesuffix="LastContributionContingency"     weight="2" />
            <entry typesuffix="TimeWindowContingency"           weight="1" />
    </weighting>
</weighter>
```

For intra-session analysis, the uptake graph for a session is isolated. Two paths are possible from here. The sequential structure of the interaction can be micro-analyzed to understand the development of group accomplishments: this part is not automated. Or we can fold the uptake network into an actor-actor sociogram (directed weighted graph), where the tie strength between actors is the sum of the strength of uptake between their contributions. This sociogram can be analyzed using conventional (social) network analysis methods such as with eigenvector centrality to identify key actors, etc. [10, 21].
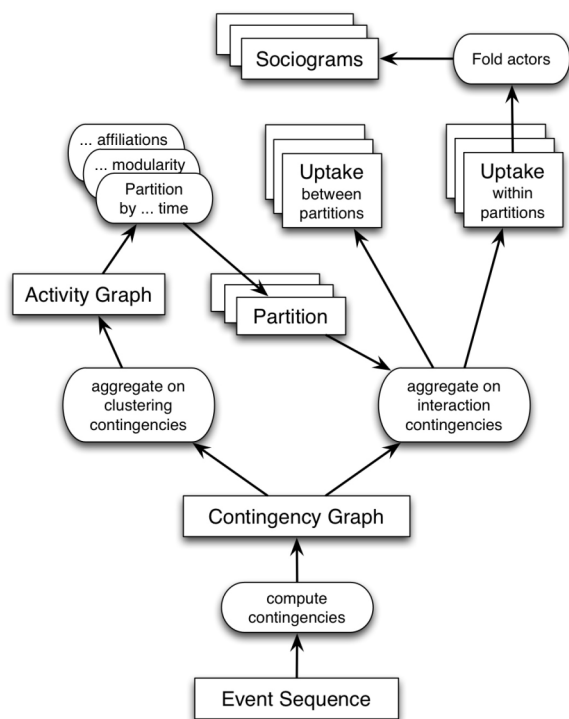

Figure 2 Process Model

Pathways for analysis in our framework are summarized in Figure 2. Contingencies are applied to events (in the EEC abstract transcript) to produce a contingency graph. Contingencies are then aggregated into uptake between the same events. A single aggregation can be used, or optionally (as shown in Figure 2) different weightings can be used for identifying sessions by graph partitioning versus for doing detailed interaction analysis (we are currently studying whether it is productive to use different weightings for these purposes). In either case, uptake that crosses partitions can be used to identify influences across space and time, and uptake within partitions can be analyzed to study the interactional structure of a session. Uptake graphs may be folded into networks where nodes are actors rather than events, to which sociometrics are applied. Another line of analysis not discussed in this paper is to fold events into actor-artifact networks, or bipartite weighted directed graphs that are called "associograms" for short, because they capture how actors are associated with each other via mutual read and write of media objects. In other work, we have undertaken community analysis of associograms to detect not only human participants in communities, but also the artifacts that reflect their mediate nature (e.g., synchronous or asynchronous) [18].

Our primary implementation is in Java, using the Hibernate object/relational model and persistence engine (hibernate.org/orm) to enable processing of large graphs. As seen above, declarative information controlling the processing (e.g., selection of type and range of source data from the EEC, sequencing of analytic steps, and weighting of contingencies) is provided in external XML scripts. We call out to the JVM-based implementation of Python (http://www.jython.org) to use the NLTK library (http://nltk.org) for lexical processing, and spawn external processes to utilize the iGraph software package (http://igraph.sourceforge.net) for graph operations such as graph partitioning.

3. EXAMPLE

Here we illustrate the approach with an analysis we conducted of data from the Tapped In network.

3.1 Tapped In

This study drew on data from SRI International's Tapped In® (tappedin.org), an international online network of educators involved in diverse forms of informal and formal professional development and peer support [6, 14]. According to its developers, Tapped In was motivated by the desire to understand how to initiate and manage large heterogeneous communities of educators, how they evolve, and the benefits that participants and sponsors derive from their involvement. This network includes activities that are sponsored by formal organizations (e.g., universities, school districts, and nonprofits) mixed with volunteer driven and other unsponsored activities, in both synchronous and asynchronous media, with participants from across all career stages and diverse occupations related to education. Thus it provides a valuable opportunity to develop and test hypotheses, tools, and techniques for understanding heterogeneous networks. Cumulatively, Tapped In has hosted the content and activities of

more than 150,000 education professionals (over 20,000 per year in our study period) in thousands of user-created spaces that contain threaded discussions, shared files and URLs, text chats, an event calendar, and other tools to support collaborative work. Over its history, more than 50 organizations, including education agencies and institutions of higher education, have consulted with Tapped In staff and became "tenants" in the system to meet the needs of students and faculty with online courses, workshops, seminars, mentoring programs, and other collaborative activities. While these organizations typically set up private spaces for people affiliated with them, there were also approximately 40-60 public activities per month designed by Tapped In members and open to anyone in the community (including tenant members). Volunteers drive the majority of Tapped In community-wide activity [6]. Extensive data collection capabilities underlying the system captured the activity of all members and groups including chat data, discussion board interactions, and file sharing.

We selected a period of peak usage that occurred from September 2005 through May 2007 for our research, and used smaller samples develop and test our methods. Here we report an example analysis of 3 days of data, centered around a session of particular interest in the second day. This session, a "Teaching Teachers" session on mentoring, had previously been chosen for micro-analysis due to its high quality of interaction. Here we wanted to see how the session was embedded in its surrounding context, and to test whether the methods described above would detect anything significant about this session.

The first step is to import data from the log files into our EEC format. Tapped In log files were in both database format and raw text files for chat transcripts: custom translators were written to import a sequence of events organized by time.

3.2 Contingencies and Uptake

Contingency analysis is run to create the contingency graph in the EEC representation, and includes installation of contingencies discussed previously. The resulting graph can have multiple contingencies between a pair of events, and is too complex to visualize here. Many network algorithms can only operate on graphs with a single edge or arc between any two vertices (nodes). Hence our next step was to collapse the multiple ties between two events into a single weighted arc representing the extent to which the second is related to the first: the "uptake graph" shown in the middle level of Figure 1. All subsequent operations described below are on the uptake graph except where noted.

3.3 Sessions

The next step is to empirically identify sessions (temporally contiguous interactions between a set of actors) in the uptake graph. Although Tapped In had scheduled "calendar events" where participants met in a particular room at a particular time, there were also many other sessions that took place spontaneously or were not announced in the calendar, and sometimes even formal events would move between rooms. So we sought to empirically identify the sessions that actually took place.

In Figure 3 we show the uptake graph for all text (chat and discussion) activity for the three-day period. All visualizations are in Gephi (gephi.org, [2]). This particular visualization uses the OpenOrd layout algorithm [9], a hierarchical version of traditional force-directed algorithms that group nodes according to their interconnectedness. Nodes are individual chat or discussion contributions. A modularity partitioning tool in iGraph based on [5] was run on this graph to identify sessions, represented by the

colors. (The modularity partitioning algorithm and the Open Ord layout algorithm use similar hierarchical strategies for clustering related nodes, as seen by the fact that nodes that are clustered spatially by OpenOrd are also assigned to the same modularity partition, as indicated by color. This suggests a general strategy: choose visualization algorithms that computationally parallel the analysis to be visualized.)
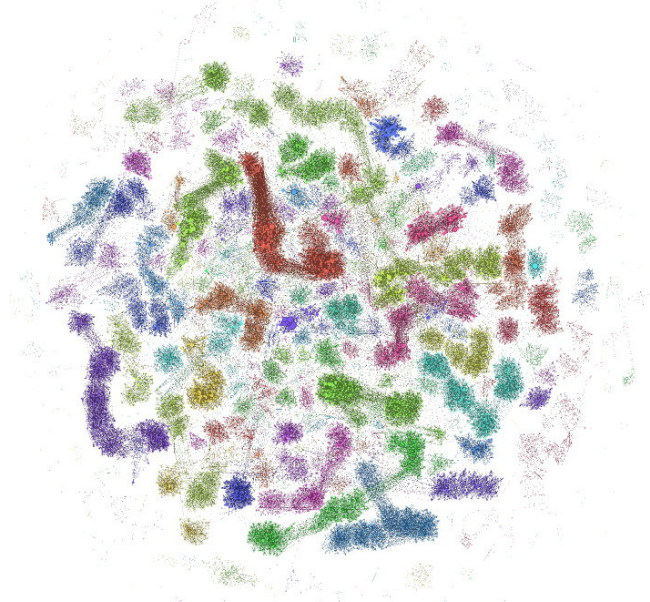

Figure 3. Uptake Graph for Three Days of Tapped In

Each connected cluster represents a session, i.e., a set of highly related chat or discussion contributions. The visualization made clear that there were a surprising number of interactive sessions taking place in Tapped In over this three-day period. Interestingly, some sessions crossed rooms: sometimes a "tour" would start in one room and move to others, or persons would meet in the Reception room and then move elsewhere. (This can be seen in Gephi by alternating the coloring between rooms and modularity classes.)

Although we find it to be interesting and sometimes useful to inspect these kinds of visualizations (e.g., to find sessions that cross rooms), such visualizations are in general difficult to interpret, and we continue to use computational tools for analysis.

3.4 Inter-Session Relations and Key Sessions

After finding modularity partitions of the uptake graph, the analysis can take two directions, as shown in Figure 2: analysis of uptake structure (interaction) within a session, and analysis of relationships between session. The latter is of particular interest for understanding how actors and ideas move between settings across time and space in a socio-technical system.

Figure 4 shows each session collapsed into a single node. Weighted degree between sessions is recomputed and shown as node size. Several of the sessions have larger node size, indicating their apparent influences on subsequent sessions. Of particular interest is the larger pink session node with a large green arrow pointing to it from a smaller light green session node (inset). What is the relationship between these sessions? First we examine the pink session, the session being "taken up" by the later one.
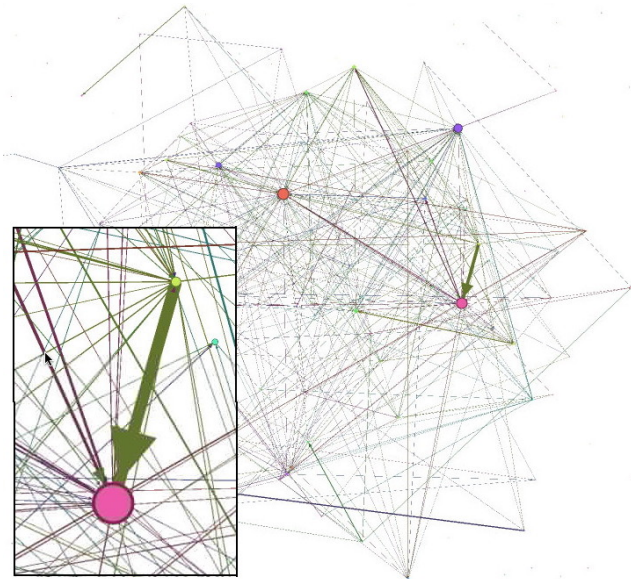
Figure 4. Inter-session uptake graph with closeup view.

session leader M and two experienced participants L and D play important roles an important role, and the sociogram helped us notice the importance of certain other actors to this session, such as A, A2 and E.
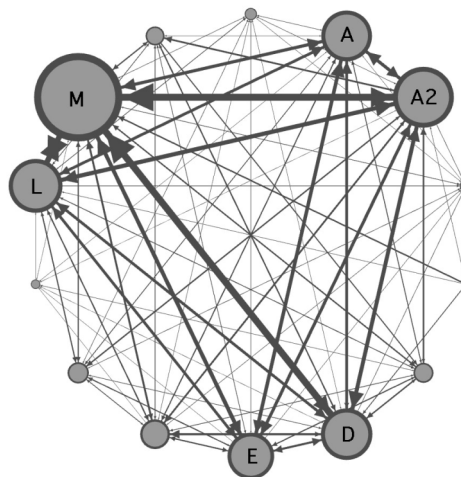

Figure 5. Session sociogram with weighted in-degree.

3.5 Session Example

It turned out that the modularity partition visualized as the large pink node captured the Teaching Teachers session on mentoring in the schools that we had been studying almost exactly. The participant contributions placed in this partition include all of the Teaching Teachers session contributions, and only a few other contributions in other locations at the beginning and end. This is remarkable because the partition was derived purely using algorithmic methods to install contingencies amongst the large number of events taking place over the three days, collapse these contingencies into uptake, and partition the graph. No information about sessions was provided to the algorithms. Thus the method shows promise as an automated way of identifying meaningful social events.

Participants in the session include the session lead who we will call M, two experienced participants L and D (one of whom may have been a volunteer facilitator assisting M), a "newbie" A trying out Tapped In for her first time, and several other participants. After introductions, an in depth discussion of peer mentoring of teachers in the school setting ensues for nearly an hour. See [CITE Cat HICSS] for a closer look at this session. Near the end, M mentions that she needs to leave for another discussion, and the others thank her and say goodbye. Interestingly, the modularity-partitioning algorithm places in this session the first few utterances of M showing up in the new session in another room, which we shall call an "In Training" group. It would be possible to filter out minority utterances from another room, but we want to preserve our ability to follow sessions that cross multiple rooms (one session we detected spanned three rooms!).

3.6 Key Actors within a Session

If we fold the uptake graph for this session into an actor-actor sociogram, we get the graph shown in Figure 5. Node size is weighted in-degree, showing the relative importance of actors in terms of the extent to which we estimate that others take up their contributions. (Other metrics such as eigenvector centrality can be used to estimate transitive importance; see also A12) Clearly the

3.7 Inter-Session Relations

As visualized in our graphs, uptake arcs are drawn from the chronologically later event to the prior event being taken up, and the same is true of the collapsed session graph in Figure 4. What is the nature of the session that is the source of the large green arrow in the inset of Figure 4, i.e., that it depends on the session we just examined? Looking at the event sequence for this session, we found that it is the In Training session that the facilitator M has just joined. Furthermore, two of the participants in her previous Teaching Teachers session, A and L, followed her there. This is the reason for the thickness (weight) of the green arrow: three actors have moved from one session to another. This relationship suggests that it might be fruitful to see whether any ideas discussed in the Teaching Teachers session were carried over to In Training. Also, a sociometric analysis of the folded graph could be conducted on this later session to see shifts in roles.

4. CONCLUSIONS AND FUTURE WORK

The prior example showed how our analytic framework and supporting algorithms can (1) find relationships between contributions in the relatively unstructured medium of chats; (2) using these relationships, parse the stream of activity into sessions defined as densely connected clusters of activity; (3) enable sociometric analysis of individual sessions; and (4) make influences between sessions across time and space visible.

More generally, this paper outlines how our work addresses two analytic challenges arising from the nature of learning in socio-technical networks. First, since learning takes place through a synergistic mixture of individual and collective agency, we need to understand aggregate phenomenon (e.g., "ties", "roles", and "communities") as both produced by and providing the setting of specific interactional events. Our framework addresses this with linked abstractions that coordinate multiple levels of analysis. Second, participant interaction is distributed across media, places and time in these environments, potentially resulting in separate traces of interaction that fragment their unitary experience. Our

framework addresses this by building on an abstract transcript of interaction.

Automating the generation of interaction and social network graphs opens up several new research approaches for relating fine-grained interaction to more aggregated levels of analysis. One approach is to expand the intra-session analysis by generating multiple social networks over the course of the session, possibly after each contribution, and track the change in actor's relational properties (e.g., reciprocity, clustering coefficient, and various forms of centrality) over the course of the session. It might be possible to typify particular types of changes in these attributes in order to recognize significant changes to the group structure or role emergence in individuals.

A more ambitious but related approach could provide insight into the dynamic development of large-scale communities. We could automate the generation of social network graphs at significant points over the entire history of the online environment, identify frequently interacting individuals, and then track the growth and fragmentation of overlapping communities over time. The sessions identified by the software represent actual periods of interaction, in contrast to external structures like chat rooms or discussion boards, and could be instrumental in determining the 'significant points' at which social network graphs would be generated. Given such a description of the community of communities it might be possible to identify critical points in the formation of healthy or unhealthy communities. This information could then be used for monitoring new environments in real-time.

Both of these approaches are enabled by our system's abstraction away from media-specific forms and the automation of mapping between levels of analysis that this abstraction enables.

5. ACKNOWLEDGMENTS

6. REFERENCES

[1] J. Andriessen, M. Baker and D. D. Suthers, eds., Arguing to Learn: Confronting Cognitions in Computer-Supported Collaborative Learning Environments., Kluwer, Dordrecht, 2003.

[2] M. Bastian, S. Heymann and M. Jacomy, Gephi: An open source software for exploring and manipulating networks, International AAAI Conference on Weblogs and Social Media, 2009.

[3] V. D. Blondel, J.-L. Guillaume, R. Lambiotte and E. Lefebvre, Fast unfolding of communities in large networks, Journal of Statistical Mechanics: Theory and Experiment, http://dx.doi.org/10.1088/1742-5468/2008/10/P10008 (2008).

[4] R. Bromme, F. W. Hesse and H. Spada, eds., Barriers and Biases in Computer-Mediated Knowledge Communication And How They May Be Overcome, Springer, New York, 2005.

[5] A. Clauset, M. Newman and C. Moore, Finding community structure in very large networks, Phys. Rev. E 70, 066111 (2004), pp. arXiv:cond-mat/0408187.

[6] U. Farooq, P. Schank, A. Harris, J. Fusco and M. Schlager, Sustaining a community computing infrastructure for online teacher professional development: A Case Study of Designing Tapped In, Computer Supported Cooperative Work, 16 (2007), pp. 397-429.

[7] S. Joseph, V. Lid and D. D. Suthers, Transcendent Communities, in C. Chinn, G. Erkens and S. Puntambekar, eds., The Computer Supported Collaborative Learning (CSCL) Conference 2007, International Society of the Learning Sciences, New Brunswick, 2007, pp. 317-319.

[8] J. Lave and E. Wenger, Situated Learning: Legitimate Peripheral Participation, Cambridge University Press, Cambridge, 1991.

[9] S. Martin, W. M. Brown, R. Klavans and K. Boyack, OpenOrd: An Open-Source Toolbox for Large Graph Layout, SPIE Conference on Visualization and Data Analysis (VDA), 2011.

[10] M. Newman, Networks: An Introduction, Oxford University Press, 2010.

[11] J. Raskin, The Humane Interface: New Directions for Designing Interactive Systems, Addison Wesley, Reading, Mass., 2000.

[12] B. Rogoff, Observing sociocultural activity on three planes: Participatory appropriation, guided participation, and apprenticeship., in J. V. Wertsch, P. D. Rio and A. Alvarez, eds., Sociocultural Studies of Mind, Cambridge University Press, New York, 1995, pp. 139-164.

[13] M. Scardamalia and C. Bereiter, Higher Levels of Agency for Children in Knowledge Building: A Challenge for the Design of New Knowledge Media, The Journal of the Learning Sciences, 1 (1991), pp. 37-68.

[14] M. Schlager, J. Fusco and P. Schank, Evolution of an Online Education Community of Practice, in K. Renninger and W. Shumar, eds., Cambridge University Press, Building Virtual Communities, 2002, pp. 129-158.

[15] G. Stahl, Group Cognition: Computer Support for Collaborative Knowledge Building, MIT Press, Cambridge, MA, 2006.

[16] D. D. Suthers, Technology affordances for intersubjective meaning-making: A research agenda for CSCL, International Journal of Computer Supported Collaborative Learning, 1 (2006), pp. 315-337.

[17] D. D. Suthers, N. Dwyer, R. Medina and R. Vatrapu, A framework for conceptualizing, representing, and analyzing distributed interaction, International Journal of Computer Supported Collaborative Learning, 5 (2010), pp. 5-42.

[18] D. D. Suthers, J. Fusco, P. Schank, K.-H. Chu and M. Schlager, Discovery of community structures in a heterogeneous professional online network, Proceedings of the Hawaii International Conference on the System Sciences (HICSS-46), January 7-10, 2013, Grand Wailea, Maui, Hawai'i (CD-ROM), Institute of Electrical and Electronics Engineers, Inc. (IEEE), New Brunswick, 2013.

[19] D. D. Suthers, K. Lund, C. P. Rosé, C. Teplovs and N. Law, Productive Multivocality in the Analysis of Group Interactions, Springer, New York, 2013.

[20] D. D. Suthers and D. Rosen, A unified framework for multi-level analysis of distributed learning Proceedings of the First International Conference on Learning Analytics & Knowledge, Banff, Alberta, February 27-March 1, 2011, 2011.

[21] S. Wasserman and K. Faust, Social Network Analysis: Methods and Applications, Cambridge University Press, New York, 1994.