Dealing with Community Data: A Report on the CSCW 2000 Workshop

Organizers: Amy Bruckman, Thomas Erickson, Danyel Fisher, Christopher Lueg

Participants: Allesandra Agnostini, Josh Berman, danah boyd, Andrew Fiore, Joshua Introne, Quentin Jones, David Millen, Emile Morse, Warren Sack, Micky Steves, Barry Wellman

ABSTRACT

A summary of the CSCW 2000 Workshop on "Dealing with Community Data," covering data collection, data analysis, community visualization, and the standardized data initiative.

Keywords

Virtual Community; Online Community; Conversation; CMC; CSCW; Quantitative Analysis; Qualitative Analysis; Social Visualization; Ethics; XML.

INTRODUCTION

Online (or virtual) communities have been a subject of increasing interest in the HCI and CSCW research communities, as well as becoming increasingly familiar elements of network-mediated social activity. However, as online communities have proliferated on the Net, the set of research techniques for collecting, analyzing and understanding the data that such communities generate has not kept pace.

The goal of this workshop was to discuss approaches to dealing with the massive amounts of data generated by online communities. In keeping with this goal (and to keep the workshop a manageable size) participants were restricted to researchers who had actually collected, analyzed, and or visualized massive amounts of community data. Topics of interest included the collection of large amounts of community data, both qualitative and quantitative analysis techniques, and ways of visualizing community data.

The remainder of this report is divided into three sections. The first section of the report summarizes the topic areas covered, and the issues, questions and observations that emerged from the discussion. The second section of the report describes one of the action items that came out of the workshop — the formation of a group to explore the possibility of defining standard XML tags for community data — and provides contact information for those interested. The third section lists the participants, and has capsule descriptions of their work.

THE WORKSHOP DISCUSSION

The organizers divided the accepted position papers into three clusters:

- Data Collection
- Data Analysis
- Community Visualization

Naturally, the papers overlapped quite a bit, many touching on all three themes. Because quite a few papers focused on data analysis, that cluster was further divided into papers that dealt with a combination of quantitative and qualitative analyses, and those that focused on the quantitative analysis of very large data sets. The resulting sets of papers defined the four workshop sessions. Each session began with an overview, and featured five-minute summaries by a selected group of workshop participants. These summaries served to ground the subsequent discussion.

Theme 1: Data Collection

One of the universals of studying online communities is that there is data — *lots* of data. And the data, or at least a good portion of it, is 'just there:' nothing special needs to be done to collect it as it accumulates in user profiles, system logs, and public archives. Ironically, this abundance of data leads to, or exacerbates, a variety of problems.

Broadly, these problems centered on the *legal and ethical* questions of data collection, on deciding *what* information to collect, and on understanding *who* the users were that the data represented.

What counts as ethical data collection varies radically according to the situation; there is no one-size-fits-all solution

Ethics of Data Collection

The most evident problem — one that every participant had struggled with — has to do with ethics of collecting and using this data. All participants felt it was important that their research data be collected only in ways with which their participants were comfortable. However, it very quickly became apparent that no single approach would suffice. Workshop participants raised a variety of problems with which they were confronted, and whose answers often

depended on very particular characteristics of the community (or its technological infrastructure).

One problem has to do with the visibility of the data collection process. While, on the one hand, it is to the researchers advantage to be unobtrusive so as not to introduce biases into community behavior, on the other hand many researchers were wary of becoming invisible. Unlike a laboratory experiment where the 'subject' consents ahead of time and is continually reminded of the situation's purpose by the very nature of the setting, the participant in an online community is caught up in an activity with its own internal processes and ends. In many cases, the online community was never designed as a research site, but was a pre-existing group that the researchers had joined. Furthermore, unlike an ethnographer engaging in participant observation of a face to face (physical) situation, the researchers studying an online community can both see 'everything' thing that happens, and are invisible—that is, their presence as observers is typically not visible not to the community.

Another concern raised by a number of participants was the issue of informed consent: what does it mean for members of an online community to give their consent to being studied, when they can neither predict the future trajectory of their own behavior over the relatively long durations during which online communities are studied, nor are they likely to understand — at the time they give their consent — the sorts of inferences which will be drawn from their data. A related issue was whether, if one or more community members decide that they no longer wish to be studied, whether the decision applies only to their behavior from that point on, or whether they can retroactively withdraw data regarding their previous actions.

Participants also raised a variety of related, but different, legal issues. Who owns community data? The individual members of the community? The community coordinator? The community as a whole? The owner of the system infrastructure? The researcher? If a group has produced a cohesive product (e.g. a collectively generated story), who owns it? whose permission is necessary for the use of that product? Can a single member of the community veto its use even if all other members are in favor of it?

Choosing Data to Collect

A second class of problem has to do with what data to collect. Although the researcher is presented with a wealth of data that is 'just there', there is, of course, a lot of data that isn't 'just there.' The amount of data available should not blind researchers to other important data. For example, a number of participants argued that data about the physical, social and cultural contexts of virtual communities were crucial. Thus, it makes a difference whether participants are logging on from home or work, and as solitary individuals or in small groups. And, to add to the problem, the members of online communities are often embedded in a diverse array of work and social contexts.

Related to the problem of what to collect is the fact that often the data collected (or that a researcher wishes to collect) changes over time. There can be two reasons for this. First, the software and network infrastructure of the community may change over time, thus changing what data can be collected (e.g. new versions of web logging software may log more or less of different information). Second, over time, emergent phenomena may occur which change the questions that the researcher is interested in addressing. The site that formerly appeared to offer a wealth of data may suddenly be seen to lack crucial information. These issues are exacerbated by the fact that online communities have, by definition, relatively long durations, and thus are often the focus of longitudinal research.

Identity

The last type of problem raised had to do with the question of identity. This problem manifests itself in two ways, as both participants and researchers have (different) interests in revealing and concealing identities. First, there is the challenge of identifying who is actually participating in the community. In many situations community members may be anonymous; to the extent that researchers wish to collect data about the social or cultural contexts in which the users are embedded, this is a problem. Another sort of problem is that in some communities users may maintain multiple identities. This may be because they wish to establish separate personalities or reputations, or because they wish violate implicit or explicit community rules, or even to try to create a counterfeit consensus or 'buzz', by making it appear that many people share a similar opinion. Depending on the research questions, researchers may wish to track behavior by online persona, or by user, or by both.

A second problem for researchers is in concealing identities when data is presented or published. While workshop participants subscribed to the usual practice of concealing names and organizations in publications for ethical reasons, there was less consensus about how far to go in concealing other potentially identifying characteristics (e.g. nicknames, dialects, workgroups, professional affiliations, community customs). A related problem is that many researchers work from sources where the archive is public. If an archive is available to the net, should one cite conversations from it anonymously, knowing that interested readers can search the net, find the conversation in question, and identify the speaker? Or should one alter the conversation enough that a search could not identify it. Each approach has its limitations. A rather different wrinkle on this problem is that sometimes, some community participants may want their real names attached to their words, a desire that goes against standard practice and may also weaken attempts to disguise the identities of other participants.

While the sorts of identity problems that arise vary considerably based on both the community being studied and the methods and goals of the researchers, it is clear that it's a crucial issue, and an important one to think about before data collection begins.

Theme 2: Data Analysis

Just as all participants had dealt with the issue of collecting data, so, as well, had all struggled with analyzing data. The second and third sessions of the workshop turned the focus of discussion to data analysis.

A Multitude of Perspectives

One of the most surprising results was the sheer diversity of methodological and theoretical perspectives in the workshop. A quick roll call of theoretical perspectives included sociology, social network theory, computational linguistics, literary criticism, rhetoric, and social theory. Methods ranged from narrative "thick description" to purely quantitative methods, though virtually everone agreed that both quantitative and qualitative methods were needed (though some argued that the distinction was fuzzy).

> **"We have a good Petri dish, but not a good microscope"**

Analysis Tools

In line with the diversity of perspectives was the wide array of tools used. About the only tool set all participants had in common in doing data analysis were hand-tuned PERL scripts. Participants used scripts to sort and clean up their log files, to count users, and to visualize data. Some had automatic scripts that updated databases; others would need to re-collect their data whenever they wanted to view their community's latest figures. Other than PERL, few participants used the same tool. Other tools used included NUDIST (text analysis), Netscan (USENET message analysis), and XXXXXX.

Comparing Different Communities

Workshop participants were very interested in the issue of being able to compare different communities — however, this turned out to be remarkably difficult. Most had developed methods of characterizing their own communities, either quantitatively or in terms of visualizations, but what was striking was how difficult it was to map one approach to characterizing a community onto a different one. It was remarkably difficult, for example, to agree upon standard metrics for the liveliness of a community. Even such an apparently simple approach as counting the number of different people who log on per day does not work if community members use multiple identities, or are anonymous, or if the notion of logging on doesn't fit (e.g. community is web-based or mailing-list based), etc. Similar problems crop up with other seemingly straightforward metrics like number of utterances, quantity of posting, quantity of activity — the nature of the community, and of the underlying technical infrastructure, impose very different interpretations on what even simple metrics mean when applied to different communities.

Analyzing USENET Threads

A number of participants were engaged in analyzing USENET data, and here, at least, it was possible to use similar metrics. One of the issues that arose with respect to the analysis of USENET data was the definition of conversation threads. More qualitatively inclined members of the workshop challenged the use of message header information as a way of identifying threads. After all, they argued, people can and do change message headers manually, or start new messages instead of using the reply command, even though they may be responding to the content of previous messages. Conversely, people may use the reply function to generate a post which then has no actual relation to the thread to which it appears to belong. The reply to such arguments was that if one is going to do an analysis of several million messages, qualitative approaches such as reading and coding responses is simply not an option.

Scale and Granularity

The discussion about USENET threads was a particular example of a more general problem: how to link low level log data to higher level goals and tasks. That is, much community data (particularly that that is collected in log files) consists of very fine grained behaviors, and the task of the researcher is to make inferences about more meaningful trends (e.g. the life cycle of community) and activities (e.g. the intentions of a user). Thus, for example, researchers wish to know whether a log entry that shows that a user spent a long period of time on a particular web page indicates that they were reading its content with interest, or were confused and got 'stuck' on the page, or simply got distracted by some other activity (and, for example, switched to another application for a while). There is no general solution to this problem, though partial solutions can be devised for particular cases. It was noted that researchers are best off if they a) know what questions they will want to answer ahead of time, and b) have some control over the data collection process and mechanism.

Survey Work

The discussions of the need for 'higher level' information, and the difficulty of inferring such information from log files, lead to a discussion of the role and usefulness of data generated by survey work. On the positive side, surveys enable researchers to examine very large data sets, use standardized questions, and to carry out comparisons among subgroups. On the negative side, the limitations of self-report data are well known, and, in particular, retrospective reports are unreliable, thus suiting survey work more to synchronic than longitudinal research. However, the work of Robert Putnam (as was described in the closing plenary) was held up as an example of the possibility of developing high quality, longitudinal findings from survey data, though the magnitude of his data sets puts his work in a rather different category from most survey work carried out in HCI and CSCW.

Theme 3: The Visualization of Community Data

While the previous sessions focused on the use and analysis of community data, this session focused on the application of community data: using it to drive visualizations of community activity. While this can, of course, be extremely useful to researchers and other analysts, a strong focus in this session was on using visualizations to provide feedback about the nature and activity of the community to the members of the community.

> **"When creating visualizations for the community, accuracy is not the point: ambiguity can be useful, and even essential."**

Workshop participants discussed two sorts of situations. One situation involved people who were searching for appropriate communities to join, who, for one reason or another, might wish to get a feeling for what a community was like without actually plunging into it. An example might be an adolescent looking for an online place where he or she might discuss issues of sexual orientation, or someone wishing to discuss a socially stigmatized disease such as depression. In these cases, it would be valuable for a person to be able to quickly (and without personal exposure) get an idea of the nature of a community: Is it supportive? Are responses empathic? Is it lively or moribund? Is there a core of community members who return, or are participants more transient in their allegiance? Trying to understand how to convey such community characteristics in a directly understandable ways raises a number of issues not typically found in other visualization scenarios.

The second situation that was discussed was using visualizations about community activity as feedback to the membership of the community. The basic assumption is that making the activity of members of a community visible can support a variety of social effects (e.g. norm establishment, peer pressure, imitation, accountability) that promote coherent behavior. As in the previous situation, such visualization goals raise some interesting issues about design such social visualizations. For example, one conjecture was that social visualizations should not be customizable, the argument being that it is important for everyone to see the same thing, and for everyone to know that everyone else was seeing the same thing. Another claim was that it was more important for a social visualization to be suggestive, than for it to be accurate, and that ambiguity in such a visualization can actually be useful. As sophisticated readers might guess, this discussion raised issues about privacy versus visibility, thus complementing the ethics thread that occurred earlier in the day.

THE STANDARDIZED DATA INITIATIVE

Responding to the frustrations expressed during the discussions of data analysis, some participants expressed interest in developing a shared code-base for community data analysis that would allow common techniques for discussing and understanding log files, community statistics, and conversation data.

Although such a system can not cover all possible types of data, there certainly are a number of commonalities between the various forms. For example, a number of workshop participants collected data that took the form of topical chat. Each of them had had to struggle with issues of threading, of distinguishing active conversations from inactive postings, and of describing their conversation. Thus, the hope is that by providing a basic library of analysis scripts designed to parse a common data format, we could reduce the number of scripts that need to be re-written and re-invented. Ultimately, we would like to provide a shared language for describing community data.

Our intent, then, is to quickly generate certain quantitative, descriptive statistics about online groups. Several members of the workshop—most notably, David Millen and Quentin Jones—presented a considerable body of statistics about the community data they had collected; both received some requests to analyze other members' data. Marc Smith's Usenet research site (http://netscan.research.microsoft.com/) also has important statistics for online group participation.

As a starting point, there are three goals for this initiative:

- Define a language for speaking about community data logs and statistics.
- Build a shared set of quantitative techniques that can be applied to this language.
- Apply technologies that have already been built to the toolset, and collect datasets to use with the toolset.

People interested in participating should contact Danyel Fisher at danyelf@cs.berkeley.edu.

PARTICIPANTS AND RESEARCH

While it's not possible to adequately summarize participants' work in this short paper, the following list provides contact information, the title of each participant's position paper, and a one sentence description of their project or interests. Abstracts and position papers may be found on line at http:/xxx.xxx.xxx.

- **Allesandra Agnostini**, agostini@cootech.disco.unimib.it. University of Milano. Position Paper: "Contextualized Traces for Multiple Purposes in Campiello". Campiello is a system for promoting interaction between inhabitants of tourist-destination cities (e.g. Venice) and tourists.
- **Josh Berman**. berman@cc.gatech.edu. Georgia Tech. "The Turing Game: An Examination of Cultural Identity in Online Environments. The Turing Game is an online environment that explores issues of online identity and diversity.

- **danah boyd**. danah@media.mit.edu. Sociable Meida, MIT Media Lab. Position Paper: "Loom2: A visual system for describing the community surrounding Usenet newsgroups." Our research goals include designing intuitive visual representations of social information and furthering our understanding of what are the most socio-culturally significant patterns in the domain of online conversations.
- **Amy Bruckman**. asb@cc.gatech.edu. Georgia Tech. Workshop organizer. She and her students in the Electronic Learning Communities (ELC) research group do research on online communities and education.

  Thomas Erickson. snowfall@acm.org. IBM T.J. Watson Research Center. Workshop organizer. "Putting the There There: Visualiing Community Data." Exploring ways of using visualizations of a community's activity as feedback to the community.
- **Andrew Fiore**. atf2@cornell.edu. Cornell University (Microsoft Research). "Visualizing Components for Persistent Conversations." We have developed a set of tools for illustrating the structure of discussion threads like those found in Usenet newsgroups and the patterns of participation within the discussions..
- **Danyel Fisher**. danyelf@cs.berkeley.edu. Computer Science, UC Berkeley. Workshop organizer. Pursuing sociologically related projects involving the visualization of newsgroup and community interaction.

  Joshua Introne. jintrone@cs.brandeis.edu. Brandeis University. "Segmenting Usage Data in Collaborative Systems". Discusses the use of "coordinating representations" in the context of a collaborative problem-solving system called VesselWorld.
- **Quentin Jones**. qgjones@acm.org. University of Haifa. "Information Overload and Virtual Public Discourse Boundaries." An analysis of 2.65 million USENET messages which examines how information overload impacts on discourse structure.
- **Christopher Lueg** lueg@it.uts.edu.au. University of Technology, Sydney. Workshop organizer. Exploring social navigation, as well as participation and identity shaping, in the context of information spaces..
- **Dave Millen**. millen@lotus.com. Lotus Research.. "New Media Challenges for Community Research." Issues that arose during a study of an online community of journalists and their use of their communication archives.
- **Emile Morse**. NIST. emile.morse@nist.gov. A Visualization Approach to Dealing with Log Data. Describes the CollabLogger, a visualization tool intended to be used for data exploration and hypothesis generation.
- **Jack Muramatsu**. jmuramat@ics.uci.edu. UC Irvine. "Multiple Virtual Identities: A General Problem for Research Applications of Community Data". Considers the problem of detecting the use of multiple virtual identities; building a tool to track the use of such identities.
- **Warren Sack.** sack@sims.berkeley.edu. SIMS, UC Berkeley. "Mapping Conversations." The Conversation Map system computes social, semantic and spatial descriptions useful for summarizing, navigating, and visualizing very large-scale email-based conversations.
- **Micky Steves**. msteves@nist.gov. NIST. Position Paper: "Mining Usability Information from Log Files: A Multi-Pronged Approach." Discusses issues that arose during the analysis of a large set of data gathered from a field study of welding engineers.
- **Barry Wellman**. wellman@chass.utoronto.ca. University of Toronto. "Physical Place and Cyber Place: The Rise of Networked Individualism." Investigates that application of social network theory and analysis techniques to the study of online community.

ACKNOWLEDGMENTS

Thanks to CSCW 2000 in general, and the workshop chairs — Christine Halverson and Yvonne Rogers — in particular, for providing support and a fine venue for the workshop.

The columns on the last page should be of equal length.