Ethno-Mining: Integrating Numbers and Words from the Ground Up

Ryan Aipperspach
Tye Lawrence Rattenbury
Allison Woodruff
Ken Anderson
John F. Canny
Paul Aoki

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2006-125
http://www.eecs.berkeley.edu/Pubs/TechRpts/2006/EECS-2006-125.html

October 6, 2006

Ethno-Mining: Integrating Numbers and Words from the Ground Up

Ryan Aipperspach[1,2], Tye Rattenbury[1], Allison Woodruff[2], Ken Anderson[3], John Canny[1], Paul Aoki[2]

{ryanaip@cs.berkeley.edu, rattenbt@cs.berkeley.edu, woodruff@acm.org, ken.anderson@intel.com, jfc@cs.berkeley.edu, aoki@acm.org}

[1]Berkeley Institute of Design
Computer Science Division
University of California, Berkeley
Berkeley, CA 94720 USA

[2]Intel Research Berkeley
2150 Shattuck Ave, #1300
Berkeley, CA 94704 USA

[3]Intel Research PaPR
2111 NE 25th Ave.
Hillsboro, OR 97124 USA

ABSTRACT

In this paper we present ethno-mining, a mixed methods approach drawing on techniques from ethnography and data mining. Ethno-mining is characterized by tight, iterative loops that integrate both the results and the processes of ethnographic and data mining techniques to interpret data. Ethno-mining provides two key benefits. First, it makes use of both qualitative and quantitative data (e.g. observations and sensor data) to study phenomena that are practically inaccessible through either data type alone. Second, it provides a means of interpreting that data which produces novel insights by exposing the biases inherent in either type of data alone. We present ethno-mining in the context of a study of mobility and laptop use in the home, discussing how findings from the study relate to the use of the method.

Author Keywords

Data mining, ethnography, qualitative methods, quantitative methods

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

We are conducting an ongoing study exploring the relationship between wireless laptops and space use in the home with the goal of developing richer understandings of behavior and motivating the design of future technologies [1,2,28]. Studying this domain presents two difficulties related to collecting and interpreting data. First, habitual domestic behaviors can be difficult for participants to self-report. They also occur at time scales (both short and long) that make manual pattern detection difficult and at times of the day that make direct or mediated (e.g. audio/video) observation inappropriate. Consequently, we chose to collect sensor data that could record very specific measurements (e.g. participants' locations) at high temporal and spatial resolution, 24 hours a day. Although rich in quantity, these sensor measurements provide a limited view of the culture of domestic life. Hence, we also collected qualitative interview and observation data with the explicit goal that the qualitative and quantitative data would inform one another.

The second, and arguably more important, issue in studying the relationship between wireless laptops and space use in the home is that building an understanding of domestic technology use (both what people do and why) involves making (potentially complicated) interpretations of the data. These interpretations, which both depend on and inform the social text and context of the home, are common in ethnographic studies. What differentiates our study from more traditional ethnography or data mining is that the interpretations we make are based equally on qualitative interview and observation data and on quantitative sensor data.

Mixed-methods approaches generally use both qualitative and quantitative data [8]; however, they maintain a separation between the two types of analyses – ultimately integrating their results but not their processes. In practice, either qualitative or quantitative analysis is typically used in service of the other. For example, quantitative questionnaires designed to assess specific issues are included in larger ethnographic studies, or brief qualitative interviews are conducted after quantitative data analysis as a common sense check on the results. Thus, while qualitative and quantitative *results* are often integrated, they are rarely used to inform the development of a single integrated *process*. While some recent studies focus on collecting sensor data to understand behavior, e.g. [21], they lack the tight integration between ethnographic and data mining techniques presented here, as is discussed later.

In addressing these issues in our study, we developed a new method we call ethno-mining. Ethno-mining, as the name suggests, combines techniques from ethnography and data mining. Specifically, the integration of ethnographic and data mining techniques in ethno-mining includes a blending of their perspectives (on what interpretations are valid and interesting and how they should be characterized) and their processes (what selections and transformations are applied to the data to find and validate the interpretations). The

benefits of ethno-mining parallel the difficulties raised above. Namely: (1) *ethno-mining relies on the collection of both qualitative and quantitative data to study phenomena that are practically inaccessible through either data type alone.* And, (2) *ethno-mining produces novel findings from and insights into the data it analyzes by integrating ethnographic and data mining techniques.* The contribution of this paper is an explication of ethno-mining, grounded in specific details from the study that generated it.

The rest of this paper is organized as follows. First, we introduce the study that generated our method. We then discuss the different components of the method through examples from our study. Next, we cover more pragmatic issues, followed by related work and conclusions.

BACKGROUND

Ethno-mining grew out of an ongoing study of wireless laptops in the home. The goal of this study is to explore the relationship between wireless laptops and space use in the home. The current study, which includes both ethnographic methods and data mining, is a follow-on to an initial study that employed only ethnographic methods. The initial study informed the types of data we chose to collect and provided initial insights about space use in the home, and our current study helped developed and expanded the findings of the initial study. Findings from the initial study and the early stages of our current study are reported in [28].

We have collected data from four households. In each household, every participant and laptop computer was outfitted with a location tracking tag from Ubisense (http://www.ubisense.net) that provides sub-meter precision location readings. In addition, software was installed on each computer to log keyboard and mouse activity, application use, and power status. These streams of sensor data were augmented with qualitative data collected through observation and interviews. The interviews were semi-structured, focusing on the locations where people spent time in their homes and how laptop use influenced and was influenced by these locations. For more information about the study, see [2].

In this paper, we reference examples from four households:

1. *Brad and Jacqueline.* Household one was a one-bedroom apartment occupied by Brad and Jacqueline, two graduate students from Australia.
2. *Jack and Margaret.* Household two was a one-bedroom apartment occupied by Jack and Margaret, a recently married couple from England.
3. *Carlo, Mareesa, and Jennifer.* Household three was a two story home occupied by Carlo and Mareesa, a thirty-something married couple and their one-year old daughter, Jennifer. Carlo was an IT professional working primarily from home, and Mareesa was a teacher taking leave to raise Jennifer.
4. *Sierra, Gaby, and Cathy.* Household four was a two-bedroom, single story home occupied by Sierra and Gaby (a female couple) and Cathy (their roommate).

The full details of the study are provided in [2,28]. In this paper we focus primarily on a single illustrative example from the study: the detection of places in the home [1].

Places in the home are created and maintained by the routine movements, and lack of movement, of the occupants. These movements have social weight – they are determined by and help determine the activities that people engage in in their homes. By tracking the location of study participants in their homes, 24-hours per day for several weeks, we were able to detect emergent places in surprising configurations. These included overlapping places varying by use, places in "empty" rooms, and different configurations of places for different residents in the same home. The places were found by processing people's location data using a custom data mining algorithm that was heavily informed by insights from ethnographic interviews and observations. Furthermore, the validation and interpretation of the found places relied equally on the quantitative characteristics output by the algorithm and on the qualitative data we collected.

Although ethno-mining emerged from a specific study, it is generally applicable to studies meeting two related conditions. First, ethno-mining combines qualitative and quantitative data sources to investigate social phenomena and thus assumes the collection of both types of data. Second, to reach an understanding of social phenomena captured in the data, ethno-mining relies on iterative loops that generate possible interpretations of the data and then seek to empirically validate those interpretations. For this to happen, the quantitative and qualitative data need to be co-informing – i.e. they need to be relevant to and capable of influencing the same set of interpretations. Practically, this means that the different types of data should be collected on the same people, situations, or settings. Although there is no strict rule about what data to collect, a general guideline is that one source of data should enable the identification of outlier data points in the other, and vice versa.

With these general study conditions in mind, we now describe how ethno-mining works.

METHOD

Ethno-mining follows the same generic steps of any study method: select a topic or question for study, collect data, analyze the data, and repeat as necessary. However, ethno-mining is unique in its integration of ethnographic and data mining techniques. This integration is carried out in *iterative loops* between the formation of interpretations of the data and the development of processes for validating those interpretations. These iterative loops are the basic building blocks of ethno-mining.

There are two key characteristics of the iterative loops in ethno-mining. First, they can be separated into three catego-

ries based on the amount of *a priori* knowledge used to find and validate interpretations of the data. Second, the results of the iterative loops are frequently, although not exclusively, represented in visualizations. Visualizations have two basic affordances: they can represent both quantitative and qualitative analyses, and they exploit the visual system to support more comprehensive data analysis, particularly pattern finding and outlier detection [5].

We use the first characteristic to organize our description of ethno-mining. Specifically, we describe three categories (simple, intermediate, and complex) of iterative loops covering the range of analyses associated with ethno-mining. For each category, we present a set of example visualizations from of our study.

Simple Analysis

This category of iterative loops concerns interpretations of the data that answer the simple question of "what happened". In essence, this category involves simply presenting data, both quantitative and qualitative, so that researchers can familiarize themselves with it and allow initial interpretations to emerge.

The interpretations of the data made in this category are distinguished from those in the other two categories by the following two points. First, they rely on well-understood abstractions like segments of measured time. Second, these abstractions have well-understood characterizations/definitions like "the hour of time between 7am and 8am." Practically, having well-understood abstractions with well-understood characterizations enables simple querying of the data that can be easily justified – e.g. "we looked at all the data between 7am and 8am to see whether people were awake or not". The downside of interpretations generated at this level is that they can be rather generic.

Examples from Our Study

The data, particularly in aggregated views, makes it easier to see emergent patterns. Visualizations at this level included density maps showing aggregate space utilization (Figure 4), animations showing the positions of participants and laptops over time, radial plots showing laptop activity over the course of the study (Figure 5), and photographs of participants, among others.

This low-level analysis culminated in the creation of a Spatial Query Tool, which supports visually browsing most of the sensor data from each household. The Spatial Query Tool proved useful in browsing for interesting events, providing a first pass at answering potential questions ("Is this interesting enough to pursue further?"), or for quickly investigating the details of an event flagged by more automated analyses (such as looking up where everyone was during a specific instance of laptop use). Like other data stream aggregation tools in the social sciences, such as Replaytool [7], we found that the creation of a tool for visualizing all of the sensor data in one place is a useful way of getting a general overview of the data during analysis and for allowing social scientists on the project team to browse the data record.

Low-level observations about individuals' space use in the home helped to confirm and enhance emergent themes in our previous qualitative study. In particular, descriptive queries supported the development of a taxonomy of place distinguishing between a small number of *favorite places* where people spend most of their time and a larger number of *kinetic places* where people carry out more specialized activities. For a more in-depth discussion of this taxonomy of place, see [28].

Intermediate Analysis

While interpretations of the data from the simple level can be used as the foundation for many types of analysis, it is also helpful to generate more nuanced and complicated interpretations. Interestingly, the interpretations and processes developed in this category often directly motivated the analyses performed in the complex category.

Intermediate interpretations of data are distinguished from the other categories by two points. First, they rely on well-known, although less clearly defined, abstractions like „morning preparation routine", „family dinner", „living room", or „party". These abstractions are nearly always defined using situationally or contextually dependent characterizations. For example, the structure of routines or the boundaries between rooms (especially in open layouts) may vary between individuals and across households. The nuanced quality of the abstractions used in this category of analyses often leads to more interesting interpretations.

Examples from Our Study

In the development of our place-finding example, we were interested in the distribution of time that participants spent in each place. Because we did not yet have a means of automatically extracting places, we used Spotfire (http://www.spotfire.com) to hand-label places (in terms of their spatial boundaries) based on cultural norms, our experiences during interviews and home visits, and diagrams made by participants of the places in their homes. This provided a quick method of labeling places and calculating the distribution of time spent in those places.

We also created "moment-in-time" diagrams (Figure 1) that highlight particular events or sequences of events that emerged from visualizations of the sensor data. They can either be a single diagram abstractly showing participant action (as in the figure) or a series of diagrams showing "snapshots" over a period of time (such as a series of diagrams showing every time a laptop moved during a party).

We also transformed continuous sequences of computer activity events into "sessions" (instances of keyboard or mouse activity separated by gaps where neither the keyboard nor mouse was used). This allowed us to ask questions about individual sessions and to create "automated"

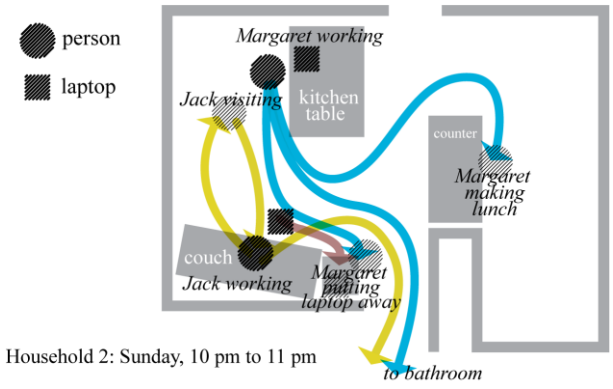

Figure 1. Moment-in-time diagram encoding interpretation of an evening in Household 2. (Note that diagrams shown to participants during interviews did not include explanatory labels.)

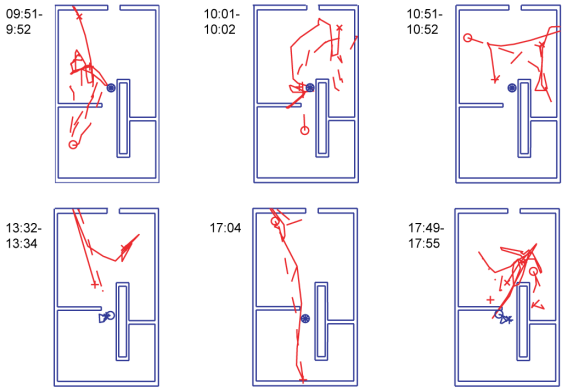

Figure 2. "Automated" moment-in-time showing Jack's use of the laptop as a music player. Each small multiple shows Jack's path before (dashed) and after (solid) using the laptop. In each case, the laptop is running music player software.

moment-in-time diagrams showing the movement of people and laptops before and after each usage session (Figure 2).

Analysis at this level made it possible to quantitatively characterize differences between favorite places and kinetic places. For example, participants spent 30% or more of the time when they were home and awake in their most popular favorite place. (Jacqueline spent 75% of her time on the couch!) Alternately, approximately 10% of time was spent in less popular favorite places, and only 1%-5% of time was spent in any given kinetic place. These measures, which would be difficult to get through observation or experience sampling (ESM) [17], highlight the magnitude of the disparity between time spent in favorite places and time spent in kinetic places.

Moment-in-time diagrams and other visualizations highlighted how different places were used by participants and revealed exceptions to use. For example, viewing plots of the paths people took immediately before and after laptop usage sessions highlighted the fact that laptops rarely moved from favorite places, as there were a small number of patterns of use for each laptop. It also highlights exceptions to typical use. For example, one sequence in Household 1 showed Jacqueline using her laptop at the table while preparing dinner and then moving it back to the couch to listen to music during dinner. This sequence stood out because Jacqueline rarely used the laptop in the kitchen or moved it between places in rapid succession. Because this event happened only once, it would be unlikely to emerge through lower-resolution sampling methods such as ESM.

Visualizations that make it easy to find visual patterns (e.g. laptop traces) or to quickly make complicated quantitative comparisons (e.g. Spotfire) help to reveal more patterns in data than looking at visualizations of raw data. They also allow analysts to quantify and characterize findings, such as revealing the actual percentage of time that users spent in favorite places. However, while making it easier to see patterns, they still require manual identification of patterns in the data and require significant amounts of intuition to narrow down the large space of potentially interesting dimensions to consider.

Complex Analysis

Whereas the interpretations of the data generated by the previous two categories rely on abstractions that are fairly well-understood, the interpretations generated at this level rely on abstractions that are not well-understood. This can be the result of novelty, i.e. no one else has considered the abstraction before, or an association with a particularly complicated phenomenon such as place. This quality has the further consequence that characterizations (i.e. definitions) of these abstractions are either very vague and hence hard to operationalize (e.g. "places are determined by people's routine behavior"), or are overly simplifying and miss interesting aspects of the abstraction they try to define (e.g. "places in the home are isomorphic to rooms").

Accordingly, ethno-mining analyses in this category can be characterized as finding interpretations of the data that rely on abstractions the researchers define. The definition of the abstractions involves specifying the characteristics, both quantitative and qualitative, that distinguish the abstraction in question from related abstractions.

Practically, this category of analysis in ethno-mining demonstrates the tightest integration of ethnographic and data mining techniques. Defining abstractions (e.g. what a place is) requires the specification of both qualitative and quantitative characteristics. Determining whether these abstractions are supported by the data in the study often involves complicated queries over the data, potentially necessitating a customized data mining algorithm. Additionally, interpreting the results of these queries requires integration of qualitative and quantitative understanding.

Examples from Our Study

Artifacts generated at this level are often visualizations of output generated by algorithms and techniques that empirically detect new conceptual structures.

Our place-finding process culminated in the development of a new algorithm that clusters the spatial data recorded for each participant into a set of high-level "places". Because of the difficulty inherent in defining place (see e.g. [14]), the process of building our place-finding algorithm involved considering previous definitions of place (as well as interacting with an environmental psychologist) in order to inform and reveal biases in the data mining algorithm we developed. This process resulted in a set of important places for each participant, but it also generated an algorithm which encodes our definition of what makes an important place (see [1]). For example, it uncovered multiple overlapping places per room and places in "empty" rooms. Additionally, because the places we found were described numerically, they enabled us to measure the similarity between places.

When examining some rooms, it was difficult to determine how many places they contained. For example, density maps from Household 1 suggested that the entire living area might contain two places: one around the couch and one around the kitchen table (Figure 4). When we asked Jacqueline, she said that she viewed the entire area as one single place.

In fact, the place-finding algorithm found three different places in the main living area: one around the couch, one around the kitchen table, and one containing both as a single place. In some circumstances, such as entertaining guests during a birthday party, Jacqueline moved around the entire living area as if it were one place rather than remaining at either the couch or the table for an extended period of time.

Similarly, we found a distinction in Brad's data between a "regular couch" place and a "deep couch" place. The regular couch place included short visits to the couch, combined with paths to other places, indicating short trips to the bathroom or the kitchen. Alternately, the deep couch place included longer visits to the couch, uninterrupted by short breaks. For both Jacqueline and Brad, the place-finding algorithm revealed subtle distinctions in the use of space within a room.

The algorithm also revealed places in "empty" rooms. When we visited Household 2, Jack and Margaret had just moved into a new apartment and had not yet purchased a bed. They slept on a futon, leaving the bedroom "empty". However, the place-finding algorithm revealed places that were not apparent on initial examination of the room. In particular, it revealed an "exercise" place for Margaret – she would regularly use the open space in the bedroom to stretch and work out.

New Laptop	Using			
Sum of duration	Place - Jack			
Place - Margaret	Home	Gone	Sleep	Grand Total
Home	63.96%	12.71%	2.82%	79.49%
Gone	18.18%	0.00%	0.03%	18.21%
Sleep	1.87%	0.00%	0.43%	2.30%
Grand Total	84.01%	12.71%	3.28%	100.00%

Figure 3. Microsoft Excel pivot table showing laptop use in Household 2. Pivot tables are two-dimensional view of a multi-dimensional data cube allowing the user to plot any two variables against each other and to filter the results by several other variables, facilitating the rapid asking and answering of a range of questions.

Additionally, having quantitative descriptions of significant places in the data allowed us to numerically compare places between people and households. In doing so we found unintuitive but interesting mappings. For example, quantitative mappings between places revealed that the kitchen table for Brad in Household 1 was most similar to the office for Sierra in Household 4. While initially non-intuitive, this makes sense upon examination of how both people use the space. Both Brad and Sierra tend to spend long term work sessions at the table and in the office, respectively, making the places similar in use.

The development of a place-finding algorithm grounded in qualitative theory provided us with more nuanced and interpretable sets of places than would the use of an arbitrary or "off-the-shelf" mining method. Additionally, because we encoded our definition of place in an algorithm, we have a numerical means of comparing places across different participants and homes, providing useful input further more qualitative analyses.

Additional Applications of Ethno-Mining from Our Study

The findings above provide a rich view of places in the home, allowing for both qualitative and quantitative analysis of movement patterns and types of places. However, the examples are part of a broader research agenda concerned with understanding the relationship between people, space, and technology in the home. We now briefly describe several other themes currently emerging from repeated applications of ethno-mining to our data set. These examples highlight how high-level findings (like places) can be successfully reintegrated into other threads of analysis.

In particular, we made use of pivot tables (Figure 3) at the intermediate level and temporal clustering at the complex level to analyze data. The temporal clustering was performed using SQL Server's built-in data analysis tools. The clustering was performed on the following variables: the positions of residents and laptops, the application running on each laptop, and the time-of-day. The key characterization of the clusters is that they reveal stable temporal-spatial patterns in the data.

Importantly, both pivot tables and temporal clustering made use of the high-level places found in our earlier analysis rather than low-level position information. These discrete place labels, along with the laptop session divisions which were manually defined, enabled easier aggregation and interpretation of data than low-level positions and streams of laptop usage events.

Social Aspects of Laptop Use

Another strong theme in our data is that laptops are social objects – we found that participants frequently shared, competed for, and were "followed by" their laptops [2]. At the simple analysis level, the "broom pattern" visible in plots of laptop activity in Household 2 (Figure 5) combined with information from interviews highlights coordinated sharing of laptops, revealing how Jack would turn on the laptop and tune into the BBC news every morning for his wife, Margaret. Patterns such as this encouraged us to explore joint use of laptops at more complex levels of analysis.

At the intermediate level, we made use of pivot tables. For example, we found examples of coordinated laptop use within the home in Household 3. Even though Carlos worked primarily from home, an instant messaging application which was used on both his desktop and his wife Mareesa's laptop showed them coordinating at the beginning (9 AM) and end of (4-5 PM) of the work day, but not during the main part of the day.

At the complex level, viewing the results of temporal clustering brought out the differences in space and computer use that happened when participants were *together versus apart* and highlighted how participants coordinated joint use of space and computer resources. For example, clustering revealed differences in the computer applications that were run when different people were home alone. In Household 1, two-thirds of Jacqueline's instant messenger use happened when Brad was gone, suggesting the use of the laptop as a companion. In Household 2, the laptop was used to play music primarily when Jack was home and rarely when Margaret was home alone. Similarly, applications such as Dreamweaver and Photoshop were used only when Margaret was home (she was developing a web page) and "work productivity" applications were used only when Jack was home alone during the day.

Laptops as Multiple Devices

During initial interviews participants would often talk about the different places in which they used laptops (e.g., using a recipe application in the kitchen). However, sensor data revealed that most laptop use happened in a relatively small number of locations and involved a small set of applications per laptop [28], suggesting that participants tended to over-report how their laptop use varied by position. The use of pivot tables reinforced this finding. Every laptop in the study was plugged in 80%-90% of the time it was in use, with the exception of the home laptop in Household 3 which frequently followed baby Jennifer around the home. Thus, the only laptop which operated primarily in an "untethered" mode was one which was required to by the activity patterns of a one year old child.

While spatial analysis suggested that laptops were fairly limited in their range of use, temporal analysis revealed that laptops were used as a *range of different devices* across time. These patterns were difficult to find in visualizations of low-level data (e.g. animations) because of the sheer volume of data involved. However, pivot tables and temporal clustering revealed patterns of use that varied both by session duration and time of day. For example, in Household 2, sessions less than 4 minutes indicated that the laptop was being used as a music player, and longer sessions indicated its use for Internet and other activity. Similarly, in Household 3, short sessions indicated use of the laptop for instant messaging coordination.

In Household 4, Carlota used the laptop primarily for Internet access during the week and primarily for instant messaging and music playing on the weekend. Similarly, in Household 3, Carlos' desktop was used for work tasks during the day and for computer games only at night and on the weekends (resulting in the only continuous use sessions longer than one hour).

While laptops had limited mobility, they were still used as a range of different devices including music players, communication tools, gaming machines. Their use varied more by time of day than by location.

Joint Patterns of Space Use

Clustering and pivot tables also revealed strong patterns in space and laptop use that emerged when participants were home together or alone. For example, in Household 1 the average amount of time Jacqueline spent in each place visit was longer when Brad was gone than when he was home. However, the specifics of these patterns emerged more strongly through temporal clustering.

These configurations highlight a relatively small set of *stable patterns of activity* within each home. For example, in Household 2 when both Jack and Margaret were together in the main living area of their apartment, Jack spent his time primarily on the couch and Margaret spent her time primarily at the table. (The laptop on the table was faster than the laptop near the couch, and Margaret did not have access to a computer while at work.) However, an alternate configuration emerged in a cluster of activity on weekend evenings (8 PM – midnight). During that time, Margaret was frequently on the couch while Jack was sitting in a lounge chair next to the table.

The interesting thing about these different configurations is that there are very strong primary patterns (Jack at the couch and Margaret at the table) and a small number of secondary patterns (Margaret at the couch and Jack on the chair, but *not* Margaret at the table and Jack on the chair,

Figure 4. Density maps showing the distribution of time spent within the home by Jacqueline (left) and her laptop (right). (Note: labels and rectangles are manually added.)

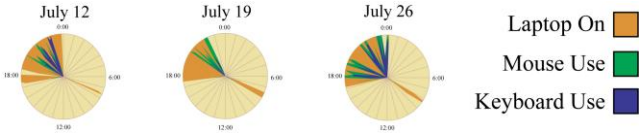

Figure 5. Radial plots of use of Household 2's "new laptop" (Tuesdays during the study). The plots reveal a visual pattern of a "broom." This pattern of short morning use and extended evening use of the laptop is typical of weekdays in this home (except for Friday evenings, which have little activity).

etc.), suggesting a joint construction of a few "workable" patterns in each household. Additionally, the distinct configuration patterns emerged more strongly in Household 2 where schedules were constrained by Margaret's job than in Household 2 where both Brad and Jacqueline had relatively unstructured summer schedules.

The differences in place visit duration and computer use when participants were alone or together highlight the fact that moment-to-moment pace of life varies depending on the presence of others [13]. They also begin to suggest a spectrum of place use patterns and coordination in the home, ranging from the individual places found using our place-finding algorithm through stable temporal patterns of joint space use and activity revealed through temporal clustering.

DISCUSSION

Having described ethno-mining, we now discuss more pragmatic concerns including the use of sensors, visualizations and interviews, and costs associated with the method. We then discuss at a general level how ethno-mining works by exposing and overcoming biases in complementary research methods before proceeding to discuss related work and conclusions.

Deploying Sensors

The use of sensors provides two advantages. First, it allows for measurement and (indirect) observation of phenomena that are practically inaccessible to methods like ESM [17] and diary studies [6]. There are a number of reasons for this. Webb [27] lists several, including: (1) observers can be selective in what they record, resulting in recording and interpretation that may be "erratic over time, as the observer learns and responds to the research phenomena observed", (2) physical observers in some cases are problematic or not allowed, and (3) self report doesn't always produce enough detail or accuracy. Using less subjective and less invasive data collection, i.e. mechanized sensor measurements, overcomes many of these issues. Moreover, sensor data can be gathered on phenomena that are outside normal human perception – occurring either too fast or slow or at spatial resolutions outside our sensory scope.

The second advantage to using sensor data is the participants' favorable orientation towards it. We found that participants were highly accommodating of the sensing infrastructure of the study. Participants readily established routines for wearing the tags. They typically put their badges on early in their morning routine, taking them off only to leave the house or to go to sleep. In fact, most participants reported forgetting that they were wearing their badges at least some of the time, occasionally walking out the door without taking the badge off.

Somewhat to our surprise, no one expressed significant concerns about wearing the badges, either with respect to safety or with respect to privacy. Participants did, however, seem to have a clear sense that they were being monitored, often making comments that began with, "You'll see… ". For example, Margaret told us, "You'll see, I always turn on a computer and then I walk away." At the same time, participants seemed to interpret the ambiguity of the data as affording them some privacy. For example, Jack told us: "I was thinking… you know where we are, but you don't know at all why we're there… if you take a specific situation, you know, we might both be asleep, or I don't know… ". The ambiguity presented a convenient filter on the data. Participants could help us to fill in missing detail from ambiguous situations, but they could also choose not to provide additional detail in sensitive situations.

This was in stark contrast to participant responses to time-lapse video data, which we recorded in two of the households (without audio) for one to two days. These households told us how glad they were when the camera was removed. As Jacqueline said, "I don't like that a lot [having a camera in the public part of the house]. I wouldn't do it for any more than that [one day]."

Visualizations and Participant Interviews

In addition to using visualizations to facilitate analysis, we used visualizations as a tool in eliciting participant feedback. Our goals in doing so are similar to other techniques

for using visualizations in interviews, such as [15,16,26]. We wanted to prompt discussion about events where self-report might be inaccurate, such as the amount of time spent in any given place, and about mundane events that weren't easily remembered.

Initially, we attempted to provide participants with as much detail as possible through animations showing participant and laptop positions. However, we found that participants had a tendency to passively watch the video and let interviewers control playback, making it difficult to stop or jump forwards and backwards for discussion. We found instead that paper copies of visualizations enabled participants to manipulate and mark up visualizations at their own pace, creating more stimulating discussion.

Visualizations prompted participants to comment on behavior patterns or reflect on surprising elements in the visualization. For example, in Household 3, Carlo first told us that he rarely went into the guest bedroom where his mother-in-law occasionally stayed. Upon examining a density map showing him spending time there, he remembered that he went into the room every day to drop off her mail. Often, a sequence of visualizations, each showing a moment–in-time arranged to cover an entire event proved useful. For example, a sequence showing the different positions of the laptop during Brad's birthday party highlighted the laptop's role as multiple devices, from its transition from being "put away" in the bedroom to its emergence as a music player to its being put away again when it was no longer needed.

Visualizations also revealed interesting interactions between participants even when data wasn't entirely correct. In Household 3, due to a miscalibration of the sensors, a density map showed Sierra spending a large amount of time by the kitchen sink (the spot actually reflected where she hung her tag when she left home). Upon seeing the visualization, Sierra exclaimed to her partner, "You see! I do wash dishes!", providing new insights into the relationship between the participants.

Costs Associated with the Method

Deploying and maintaining a set of sensors to record people's behavior is a time-consuming process. For example, we used the Ubisense location tracking system to gather quantitative data in each home. This process involved mounting 4 to 7 sensors in each home, running cables from each sensor to a central location, and going through a calibration and tuning process which often took several iterations. The setup requirements of the sensing system were not a pure cost. They did provide a unique opportunity to observe and interact with study participants during the installation, which took 2-3 days for each household.

Equally time consuming is the process of generating visualizations and developing analysis techniques. Each of the artifacts and findings described above has a different cost in terms of time and expertise required. For example, visualizations of raw data are fairly straightforward to construct, requiring knowledge about how the sensor data was collected but not about data mining techniques. However, the visualizations lack the power of automatically uncovering structures provided by data mining techniques. More automated techniques such as data mining require more expertise and time to implement.

However, researchers are developing easier to use sensing platforms. Some researchers are working on developing "tape-on-and-forget" sensors [20] and others are building sensing platforms on top of existing infrastructure, such as cell phones [9] or power lines [22]. As these systems mature, the costs associated with sensor installation and maintenance will be diminished.

Additionally, automated data mining tools are becoming more prevalent. For example, while we were required to implement a new algorithm for finding places in our data, we were able to use existing tools built into SQL Server to mine temporal patterns. Tools like [24] and [11] attempt to automate the exploration of data and may simplify the data mining process.

The emergence of new sensing and data analysis tools makes ethno-mining a more practical method than it might have been in the past. While care must be taken in choosing appropriate techniques for the phenomena being studied, requiring the involvement of researchers with an understanding of the both ethnographic and data mining issues involved, we believe that new infrastructure will reduce the time required to deploy sensors and process data, shorting the time period necessary in each iterative loop through the method.

Exposing Biases in Research Methods

In any study, making sense of and interpreting data requires the use of abstractions and concepts. In the social sciences, methods such as Grounded Theory [25] explicitly focus on generating abstractions and theories that are empirically rooted in qualitative observations (i.e. that emerge from the data) and do not rely on *a priori* conceptual constructions. However, Grounded Theory, and more generally any pragmatically oriented theory, is not un-biased. The bias is embedded in the process of selecting what "interesting" phenomena emerge from the data, or, more subtly, in selecting what phenomena to analyze.

Data Mining applies its bias in the same way – constraining the types of phenomena that can emerge but allowing the actual details of these phenomena to be empirically generated. At a meta-level, our method seeks to expose and explicitly address the selection biases in both qualitative and quantitative research methods by checking them against one another. Ethno-mining extends its scrutiny of these biases beyond simply comparing the biases embedded in standard qualitative and quantitative techniques. It does so by tightly integrating the techniques in loops, generating mutually

informed analysis techniques with complimentary sets of biases. The objective to expose and acknowledge bias is a known motivation for using mixed-methods [12, 8].

RELATED WORK

We consider any mixed method approach that explicitly combines qualitative and quantitative techniques as related work (see [8]). Here, we highlight how ethno-mining is different from existing mixed methods approaches applied in ethnography and data mining, drawing particularly on the closest examples from the CHI community.

Historically, mixed-method approaches in ethnography take the form of triangulation [18]. The goals of triangulation are central to our method, especially their focus on reducing interpretation errors by comparing them to systematic variance in any single source of data. Additionally, we draw on key components of ethnography such as intensive face-to-face involvement with participants in their own contexts and improvisational interviewing and observation techniques that allow emerging discovered realities to modify our approach. One perspective on ethno-mining is that it applies these emic techniques to the analysis of sensor data.

However, while ethnographic studies make use of multiple data sources, some of which are recorded mechanically, they tend to conduct all of their analysis in the space of human interpretable data. For example, video and audio recordings are used as proxies for human observers and then coded through human techniques, and survey results are summarized and presented to researchers for analysis. Alternatively, the process of triangulation in ethno-mining requires augments the traditional ethnographic triangulation process with quantitative and algorithmic data analysis tools that deal with machine gathered and interpreted data. To our knowledge, there are no existing ethnographic studies which integrate data mining as tightly as suggested in ethno-mining.

Likewise, more quantitative mixed-methods, which rely predominately on data mining, still use qualitative understanding of the world, if only to make sense of the experimental results [12]. For example, studies using data mining to extract behavior patterns from sensor data (e.g. [4,9,19]) must pass a certain "common sense" check based on qualitative understanding of the behavior patterns under investigation. However, while some of these methods even conduct this check by returning to participants with the results, they do not explicitly build their data mining algorithms on top of qualitative findings related to the data or use the results of their algorithms in additional qualitative analysis.

Recently, there have been a number of studies using sensors to capture social behavior. We highlight one typical example. Patel et al. [21] recently measured the distance between users and their cellular telephones using Bluetooth beacons. In the study, they showed visualizations of data back to participants and used data mining tools to model the data. However, the study lacked the integration between ethnographic observation and data mining central to ethno-mining. In particular, the key abstraction in the study, distance from the cell phones, was determined *a priori* by the limitations of the sensing devices used.

Alternately, ethno-mining could be used to unpack participant's definitions of "near" (e.g. claiming to be "always near the phone"). In Patel's study, any phenomena tied to a different characterization of distance from the cell phone – perhaps requiring finer resolution bins or bins of different sizes than the ones used, or qualitative measures such as emotional or psychological distance – are hidden. Because the goal of the study was to assess whether the cell phone was a reasonable proxy of a person's physical location, this loss was not material. However, for studies conducting more exploratory analysis, making strong *a priori* assumption about the characterization of central abstractions can be overly restrictive. This point is highlighted best in the discussion of our method, particularly in the discussion of complex analysis.

Finally, there has recently been an institutional push to integrate computational tools into the social sciences by organizations such as the UK's Centre for e-Social Science (http://www.ncess.ac.uk). The goal is to integrate computing technologies into social sciences to take advantage of high-performance computational techniques (e.g. running simulations [3]), to store and transmit data, and to facilitate remote collaboration [10]. At a high level, our method fits in line with these goals.

However, the tools developed in e-Social Science often focus on presenting and sharing data to facilitate traditional forms of manual analysis. For example, Crabtree et al. developed tools for integrating diverse streams of data into one viewing tool [7]. We suggest that automated data mining of quantitative sensor data combined with traditional observations can create stronger results. We explicitly focus on the relationship between ethnographers and data mining experts in co-constructing interpretations *and* analysis techniques.

CONCLUSIONS

In this paper we have presented a novel method integrating ethnography and data mining at multiple levels of analysis. We have discussed the details of our method and the type of results that it generates in the context of a case study exploring the movement of people and laptops in the home.

We believe that this method has the potential to reveal new understanding in a range of different areas. It can provide concrete examples to inspire design [23] or facilitate finding patterns in social behavior (e.g. in CSCW). However, in order to make its application more practical, continued development of sensing infrastructure and analysis tools is needed. One open question is the extent to which data analysis and pattern extraction from sensor data can be auto-

mated. For example, tools like iCube [24] attempt to automate the exploration of data cubes (like the ones we used in Pivot tables). Other projects are exploring means of extracting patterns directly from streams of sensor data [11]. These techniques show promise in limiting the assumptions that must be made in conducting data mining. For example, while it is easy to look for cyclical temporal patterns at the daily or weekday vs. weekend level, it is less trivial to automatically determine the time scales of other temporal patterns.

Additionally, while our method focuses on analysis of data from individual participants directly involved in the study, it may be possible to combine ethnographic observations of a small number of participants with data mining and visualization of larger found or historical datasets. The ways in which this interaction would take place and in particular the means of bridging between individual participants and larger scale data sets is left to be explored in future work.

ACKNOWLEDGEMENTS

We would like to thank everyone who gave us feedback on the development of this method or who provided input on data analysis, particularly Minos Garofalakis. We also thank Ben Hooker for his help with data visualizations. Additionally, we would like to thank our participants, without whom this study would have been impossible.

REFERENCES

1. Aipperspach, R., Rattenbury, T., Woodruff, A., and Canny, J. A Quantitative Method for Revealing and Comparing Places in the Home. Proc. Ubicomp 2006.
2. Aipperspach, R., Woodruff, A., Anderson, K., and Hooker, Ben. Maps of Our Lives: Sensing People and Objects Together in the Home. Technical Report No. EECS-2005-22, EECS Department, University of California, Berkeley, 2005.
3. Axelrod, R. Advancing the Art of Simulation in the Social Sciences. Complexity 3, 2 (1997), 16-22.
4. Begole, B., Tang, J.C., and Hill, R. Rhythm Modeling, Visualizations and Applications. Proc. UIST '03.
5. Card, S. K., Mackinlay, J. D., and Shneiderman, B. Information Visualization. Information Visualization: Using Vision to Think. Morgan-Kaufmann. San Francisco, California, 1999. 1-34.
6. Carter, S. and Mankoff, J. When Participants Do the Capturing: The Role of Media in Diary Studies. Proc. CHI 2005.
7. Crabtree, A., et al. Supporting Ethnographic Studies of Ubiquitous Computing in the Wild. Proc. DIS 2006.
8. Creswell, J. Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. SAGE Publications, Thousand Oaks, CA, 2003.
9. Eagle, N. and Pentland, A. Reality mining: Sensing Complex Social Systems. Personal and Ubiquitous Computing 10, 4 (2006).
10. Fielding, N. Qualitative Research and e-Social Science. NCeSS Commissioned Strategy Report, 2003. http://www.ncess.ac.uk/docs/qualitative_research_and_e_soc_sci.pdf.
11. Garofalakis, M. et al. Probabilistic Data Management for Pervasive Computing: The Data Furnace Project. IEEE Data Engineering 29, 1 (2006), 57-63.
12. Giddens, A. The Constitution of Society: Outline of the Theory of Structuration. University of California Press, Berkeley, CA, USA, 1984.
13. Goffman, E. Behavior in Public Places: Notes on the Social Organization of Gatherings. The Free Press, New York, 1963.
14. Harrison, S and Dourish, P. Re-place-ing space: the roles of place and space in collaborative systems. Proc. CSCW 1996.
15. Heer, J. and Boyd, D. Vizster: Visualizing Online Social Networks. Proc. InfoVis 2005.
16. Heisley, D and Levy, S. Autodriving: A Photoelicitation Technique. Journal of Consumer Research 18, 3 (1991).
17. Hormuth, S.E. The Sampling of Experiences in Situ. Journal of Personality 54, 1 (1986) 262-293.
18. Jick, T. Mixing Qualitative and Quantitative methods: Triangulation in Action. Administrative Science Quarterly 24 (1979), 602-611.
19. Lühr, S., et al. Recognition of Emergent Human Behaviour in a Smart Home: A Data Mining Approach. Pervasive and Mobile Computing, 2006 (In Press).
20. Munguia Tapia, E., Intille, S.S., and Larson, K. Activity Recognition in the Home Setting Using Simple and Ubiquitous Sensors. Proc. Pervasive, 2004.
21. Patel, S. et al. Farther Than You May Think: An Emperical Investigation of the Proximity of Users to their Mobile Phones. Proc. Ubicomp 2006.
22. Patel, S., Truong, K., and Abowd, G. PowerLine Positioning: A Practical Sub-Room-Level Indoor Location System for Domestic Use. Proc. Ubicomp, 2006.
23. Salvador, T., Bell, G., and Anderson, K. Design Ethnography. Design Management Journal 10, 4 (1999).
24. Sarawagi, S. User-Adaptive Exploration of Multidimensional Data. Proc. VLDB 2000.
25. Strauss, A. and Corbin, J. Basics of Qualitative Research. SAGE Publications, Thousand Oaks, CA, 1998.
26. Van House, N.A. and Davis, M. The Social Life of Cameraphone Images. Proc. Pervasive Image Capture And Sharing (PICS) Workshop, Ubicomp 2005.

27. Webb, E., et al. Unobtrusive Measures. SAGE Publications, Thousand Oaks, CA, USA, 2000.

28. Woodruff, A., Anderson, K., Mainwaring, S.D., and Aipperspach, R. Portable, But Not Mobile: A Study of Wireless Laptops in the Home. In submission: http://www.popular-demand.com/pervasive07.pdf