Trace ethnography: Following coordination through documentary practices

R. Stuart Geiger
School of Information
University of California at Berkeley, USA
sgeiger@berkeley.edu

David Ribes
Communication, Culture, and Technology
Georgetown University, USA
dr273@georgetown.edu

Abstract

We detail the methodology of 'trace ethnography', which combines the richness of participant-observation with the wealth of data in logs so as to reconstruct patterns and practices of users in distributed sociotechnical systems. Trace ethnography is a flexible, powerful technique that is able to capture many distributed phenomena that are otherwise difficult to study. Our approach integrates and extends a number of longstanding techniques across the social and computational sciences, and can be combined with other methods to provide rich descriptions of collaboration and organization.

1. Introduction

Institutional ethnography of distributed large-scale organizations has long been a core research area but current approaches are either staggeringly costly or fail to holistically examine such systems and their users. This paper elucidates a methodology we call *trace ethnography* that exploits the proliferation of documents and documentary traces in such highly technologically-mediated systems. In many of these socio-technical environments, traces not only document events but are also used by participants themselves to coordinate and render accountable many activities. Analysis of these detailed and heterogeneous data – which include transaction logs, version histories, institutional records, conversation transcripts, and source code – can provide rich qualitative insight into the interactions of users, allowing us to retroactively reconstruct specific actions at a fine level of granularity. Once decoded, sets of such documentary traces can then be assembled into rich narratives of interaction, allowing researchers to carefully follow coordination practices, information flows, situated routines, and other social and organizational phenomena across a variety of scales.

Trace ethnography is a powerful and flexible methodology, able to turn thin documentary traces into "thick descriptions" [10] of actors and events that are often invisible in today's distributed, networked environments. However, trace ethnography is only novel as an integration and extension of many longstanding methodological techniques and practices from across the social and computational sciences. We first review the seemingly disparate approaches which inspired our work – from multi-sited ethnography and documentary history to cultural probes and transaction log analysis. These all share a common concern with the limitations of traditional single-sited participant-observation; in this paper we assemble their various solutions.

After summarizing each of the strategies and insights we have integrated into trace ethnography, we describe the basic principles of our methodology. Our first fundamental principle is that *documentary traces abound in today's technological systems*, logging specific actions taken by uniquely-identifiable individuals with very fine levels of granularity. While these data are routinely used to quantitatively determine abstract qualities of a system (e.g., tracking the browsing habits of a website's visitors), our methodology involves decoding, or *inverting*, these traces to provide rich qualitative accounts of individual users as they act within a broader social or organizational setting.

The second fundamental principle of trace ethnography is that, explicitly or implicitly, *documentary traces are the primary mechanism in which users themselves know their distributed communities and act within them.* In stark contrast with quantitative forms of trace data analysis, our method is not mechanically analytic, as traces can only be fully inverted through an ethnographic understanding of the activities, people, systems, and technologies which contribute to their production. This approach is an extension of the 'infrastructural inversions' described by Bowker and Star [4,5] that focus on the backgrounded and often ignored elements of practice and technological support that undergird everyday activity. By bringing these documentary practices to the foreground and systematically analyzing the traces that are produced, we can give rich accounts of how distributed organizations operate.

We next demonstrate our method through a case study in which we invert multiple documentary traces to reconstruct the interactions of a fast-paced, highly-heterogeneous group in a decentralized organization: Wikipedia. In a previous paper we used trace ethnography to explore and theorize how users, editors, administrators, and bots collaborated in an event that led to the banning of a vandal [11]. In this paper we focus on our documentary methodology, showing in detail how we performed different kinds of inversions, then guiding the reader through the analytical steps and ethnographic strategies required to decode and assemble documentary traces. We conclude by providing guidance for how to extend this methodology to new kinds of traces within other organizational forms such as distributed science or open-source software communities.

2. Ethnographies of large scale globally distributed activities

Ethnography has been used for centuries to gather rich detail about the culture and practices of people. Although this methodology was born out of anthropological studies of non-Western or aboriginal societies, ethnography is used by contemporary researchers in a number of fields to give thick descriptions of how people live and work. In recent decades, ethnography has been widely used in technological fields such as human-computer interaction (HCI) to learn, for example, how users interact with technologies in their homes or workplaces. In the typical approach to ethnographic fieldwork, the researcher enters a community and becomes an active member for an extended period of time, learning organizational structures and routines, roles and responsibilities, and other local customs and practices. Taking extensive field notes during such ‘participant observation,’ the ethnographer is able to turn the people (‘ethno’) into writing (‘graphy’). This approach often reveals unexpected information that is otherwise obscured in surveys and targeted interviews, and it is well-suited for a single group that resides in a shared location.

Yet as nearly all ethnographers studying distributed phenomena note, the traditional method of participant observation has considerable cost in such settings – not only in terms of financial resources and temporal investment but also empirical and analytical focus. Additionally, when the research question involves the distributed nature of the system (and not just the qualities of individual members), local observation can directly occlude phenomena that only occur *between* local sites. In response, ethnographers have developed many different kinds of techniques and strategies for studying such populations that we have drawn on in the development of trace ethnography, which we introduce in this section. We must first note that this is not intended to be a complete review of distributed or documentary ethnographic research, nor a comprehensive typology of the methods in current use. We limit our discussion to practical concerns with ethnographically studying distributed phenomena.

2.1. Documentary approaches to distributed ethnography

Members of globally-distributed organizations regularly use many kinds of documentation to cope with their dispersion; having realized this, many ethnographers turn their attention to these documentary practices. Whether it be the trails of correspondence between scientists [24,26], the manuals and handbooks that seek to harmonize practice within corporate organizations [22], trading records in global financial markets[15], or the standards and protocols that guide technology development and use [5], documentary practices are constitutive of ‘being global.’ That is, in distributed, networked organizations, documents are the primary mechanism through which supply chain managers, open source software developers, financial traders, or climate scientists not only know their global communities, but also act within them.

In fact, as many scholars of organization have recently 'rediscovered', this is often the case even in traditional co-located organizations. For example, it is easy to imagine a manager who, while faithfully interacting with co-workers on a face-to-face basis, does not regularly read reports and as such falls ‘out of the loop.’ For this reason, Anne Beaulieu has recently suggested that ethnographers shift their focus from being co-located with their participants to being co-present, noting that “co-presence might be established through a variety of modes, physical co-location being one among them” [1]. Trace ethnography draws heavily from this insight as well as a number of ethnographic approaches that have incorporated documentary practices.

2.1.1. Single-sited, document-driven ethnography

Many classic ethnographic studies of workplaces began as traditional single-sited participant observation in which the researcher situated him or herself alongside workers and followed them around. However, as these ethnographers note, most activities in the workplace revolve around documents, some of which can become quite organizationally important. The document-driven approach to ethnography thus involves following these documents as they travel across the site, asking how, where, and by whom they are produced, edited, revised or filed. A key insight from this research is that documents are created and circulated for many reasons, thus it is difficult to study their role of 'representation' in the abstract. By focusing on the lifecycles of these documents, the researcher is able to follow workflows of activity across an organization rather than tackling the abstract 'purpose' of a document.

2.1.2. Participant-generated ethnography

Ethnographers of large-scale networks have used new technologies to capture more data from more participants in more locations than traditional fieldwork typically affords. While sketches, drawings, photos, and now even videos are commonly taken by ethnographers, some researchers are dealing with the problem of distance by having participants capture their own qualitative data. For example, in the fields of human-computer interaction and media studies, user-authored diaries or journals are used to understanding how people interact with technologies in their everyday lives [23,21]. At present, increasingly sophisticated 'cultural probes' are also being deployed to give ethnographers a richer understanding of local practices at a fraction of the cost of fieldwork [9,14].

Such methods attempt to balance the need for rich, thick, and highly-empirical data with the practical limitations involved with performing ethnographic fieldwork, and are particularly appealing when participants are numerous and highly distributed. However, the weakness of this approach is that on its own, it lacks the same holistic understanding gained though ethnographic observation. Ensuring participant compliance is difficult and there may be entire swaths of activity that are left of these logs (whether because they are embarrassing for recorders, or simply considered too mundane to be worth mentioning).

2.1.3. Historical and archival ethnography

A more specific type of participant generated ethnography focuses on the extant documentary trails contained in the vast historical records and archives of institutions. Such archives are not only produced by participants, but offer the added benefit to the ethnographer of having been endogenously ordered and classified, providing yet another 'meta layer' of data for the observer. One notable example is Diane Vaughn's historical ethnography of the events leading to the Challenger shuttle explosion [26]. As she was not able to ethnographically observe the event itself or the activities of NASA that led up the decision to launch, she systematically analyzed the plethora of archives and records that had been generated by the investigation of that explosion. This wealth of records allowed her to reconstruct the chain of decision-making from a novel perspective and produce a new explanation for how risk was assessed and negotiated across NASA's vast bureaucracy in ways that eventually led to the disastrous decision to launch.

2.1.4. Multi-Sited Ethnography ('Follow the Actors')

While 19th and early 20th century anthropologists often only visited a single field site per investigation, the past few decades has seen a sharp rise in ethnographers studying multiple locations. Methodologies like multi-sited ethnography, as expressed by George Marcus, [18] have emerged with significant vigor, especially around topics like globalization, migration, nationalism, and other issues that are not typically present in a single site. Instead of attempting to witness the effects of some global phenomena in a local site (the so-called 'Heaven in a grain of sand' approach), the researcher travels to multiple sites, following various pathways in order to assemble a narrative. In this approach, visiting multiple sites is intended not to give the ethnographer more cases, (i.e. for a broader or more representative sample), but to expand a single case beyond its immediate location.

As Marcus argues, one can perform multi-sited ethnography by following mobile populations, but an ethnographer ought to also follow materials, stories, ideologies, metaphors, and conflicts as well. For example, Julian Orr's study of photocopy repair technicians is a clear case of multi-sited ethnography in that he traveled with such teams as they visited offices around Northern California. This mobile fieldwork enabled him to locate the specific ways in which these distributed groups coped with problems of distance. However, Orr did not just follow the technicians, but also traced the roles and uses of manuals, forms, and other workplace documentation, learning the various ways in which these formal documents had to be locally interpreted (and often ignored) in order to conduct the work of photocopy repair.

Proponents of Actor-Network Theory often perform a similar kind of ethnography in taking up Bruno Latour's famous injunction to 'follow the actors', in which the category of 'actors' is quite broadly defined. For example, in a study of scientific data practices [17], Latour traces out the chains of samples that ultimately turn acres of forest and savannah into a crisp scientific chart. Such activities involve cascades of documents (tables, graphs, charts, maps and field notes), the vast majority of which are rendered invisible in the final product: a scientific publication. In these situations, the ethnographer is often focused on identifying various forms of documents and documentary practices, which can be indispensible (or utterly unhelpful) in helping participants situate themselves in these distributed settings.

2.1.5 Strategically-situated ethnography

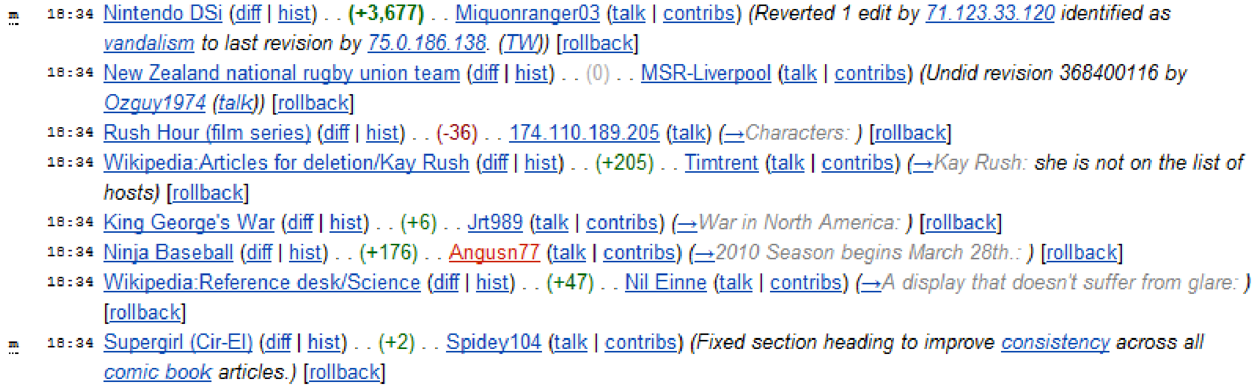
Alternatively, many scholars of large-scale distributed systems aim to locate themselves in a "strategically situated" manner [19:95]. This approach relies on the insight that participants themselves are also trying to manage scale and distribution. Instead of trying 'to be everywhere' in a large network, in this approach the ethnographer identifies key sites or events where participants are working to make intelligible their own activities. For example, in studying large-scale infrastructures, researchers have carefully and deliberately situated themselves at the particular times and places where a system is being designed, constructed, contested, broken, or repaired [25]. Similarly, social network analysts have used relational data to identify the most relevant, important, or representative local sites in a distributed organization, which are then studied in rich detail using traditional fieldwork [13,6]. Sociologists of scientific fields like demography, economics, cartography, and ecology have placed themselves in 'centres of calculation' [16] – the clearinghouses where evidence from across the world is collected, sorted and turned into knowledge.

In studies of science, focusing on a moment of controversy can be particularly revealing, as expert participants pick apart each others' evidence and arguments [7]. For example, Fujimura and Chou [8] studied the furious debates amongst biological researchers about how best to understand HIV etiology and transmission: those scientists with an 'applied' interested in managing the epidemic drew together heterogeneous epidemiologic evidence, while those interested in understanding the virus itself scoffed at those 'soft' findings and focused instead on 'hard' molecular-chemical data. In these explicit debates, many implicit assumptions and understandings often come to the fore, as the leading minds of a field will dedicate great efforts to clearly summarize and articulate their viewpoints so as to convince others. For the ethnographer, these 'controversy studies' become crucial sources for understanding otherwise arcane fields; below we explore just such a controversy in order to help understand the esoteric traces and log entries. By situating oneself at these carefully constructed vantage points it becomes possible to 'see' an entire organization, field or network through the documents and attendant practices that seek to summarize them.

3. The methodology of trace ethnography

Trace ethnography extends these forms of documentary ethnography, most heavily relying logs and records that are automatically generated in digital environments. Such traces especially abound in software platforms for the production and distribution of content where changes often need to be reverted, reviewed, or revised. These systems keep time-stamped back-up copies as well as a record of who authored the revision and often other important metadata. For example, in co-authoring a document, it is common to use the "track changes" feature present in many word processing programs. This feature embeds information about who changed what to a document and when, so that an author can reconstruct the history of a document. However, just as this information is available to authors, so too can it be valuable for ethnographic researchers.

A number of blogs, wikis, source code repositories, content management systems (CMS), and other collaboration platforms keep copies of all revisions with substantial metadata (see Figure 1). It is the increasingly rich proliferation of such traces that inspire our inquiry into the possibilities of such data. However, in investigating such traces we soon learned that the most interesting documentary traces were not those that provided seemingly full and transparent



m 18:34 Nintendo DSi (diff | hist) . . (+3,677) . . Miquonranger03 (talk | contribs) (Reverted 1 edit by 71.123.33.120 identified as vandalism to last revision by 75.0.186.138. (TW)) [rollback]

18:34 New Zealand national rugby union team (diff | hist) . . (0) . . MSR-Liverpool (talk | contribs) (Undid revision 368400116 by Ozguy1974 (talk)) [rollback]

18:34 Rush Hour (film series) (diff | hist) . . (-36) . . 174.110.189.205 (talk) (→Characters:) [rollback]

18:34 Wikipedia:Articles for deletion/Kay Rush (diff | hist) . . (+205) . . Timtrent (talk | contribs) (→Kay Rush: she is not on the list of hosts) [rollback]

18:34 King George's War (diff | hist) . . (+6) . . Jrt989 (talk | contribs) (→War in North America:) [rollback]

18:34 Ninja Baseball (diff | hist) . . (+176) . . Angusn77 (talk | contribs) (→2010 Season begins March 28th.:) [rollback]

18:34 Wikipedia:Reference desk/Science (diff | hist) . . (+47) . . Nil Einne (talk | contribs) (→A display that doesn't suffer from glare:) [rollback]

m 18:34 Supergirl (Cir-El) (diff | hist) . . (+2) . . Spidey104 (talk | contribs) (Fixed section heading to improve consistency across all comic book articles.) [rollback]

Figure 1: Documentary traces in Wikipedia – revision metadata for recent article edits

records of content creation. Embedded within these archival records of who changed what and when, we found a wide variety of codes that we initially passed over, seeing them as either incomprehensible markup or relatively non-descriptive. For example, in an edit to a Wikipedia article, a certain automatically-populated revision metadata field often contains something like "Reverted edits by UserX to last version by UserY (HG)". In combination with other similar traces, the most obvious inversion discloses who added or removed what from a Wikipedia article.

Yet hidden in that trace is also a wealth of more subtle information indicating significant differences in how edits were made at a practical level. One of the most revealing bits in this log entry is the "(HG") at the end. This information indicates that UserY is using a semi-automated vandalism detection tool called Huggle. This is especially apparent if we look at other similar revision metadata fields from Wikipedia: some have the "(HG)", some end with a "(TW)" indicating a program called Twinkle, and many have no appending marker at all. As we learned, these codes contextualize the generic action of reverting another user's edits by indicating the use of certain semi-automated vandalism detection tools (or not). Like all codes, they are sociotechnical, and therefore their meaning can only be understood in relation to their broader cultural and computational systems. Following from Bowker and Star's [4,5] call to study the mundane and everyday elements of infrastructure, we call the investigative activity of understanding these traces *inversion*.

In stark contrast to traditional documentary sources used in ethnography, the traces that we analyze can be notoriously 'thin' and, on the surface, often do not appear to hold much evidence about the actions they describe. This is because the utility of such traces does not stem from some inherent documentary quality, but rather because they are produced and circulated within a highly-standardized sociotechnical infrastructure of documentary practices. For example, a number of studies have described how many hospital workers exploit the materiality of patient charts in order to document organizational phenomena not explicitly stated in the record [12,2,3]. A doctor can quickly learn when a patient was assigned a new nurse by looking at hourly temperature readings, focusing not on the numbers but the handwriting. Other practices may be highly specific to a given organization or group: oncologists may use a different kind of notation, or the psychiatric ward may be the only section that uses staples instead of paperclips. By knowing the specificities of the sociotechnical landscape in which these documents are produced, a skilled observer can examine them to quickly trace the history of the document. In digital environments, this process is much easier, given that traces are often automatically created by logging programs, adding a higher degree of regularity to their production.

A critic may claim that trace ethnography has limitations in that it only can observe what the system or platform records, which are always incomplete. However, in highly-mediated environments such as Wikipedia, this is a benefit (or even a necessity) for the ethnographer who aims to capture the lived experience of being a member of a globally-distributed community. As we learned from lengthy participant observation as a 'vandal fighter', these users come to know both their world and each other through 'thin' documentary methods. In deciding whether a user is a vandal and should be banned, an administrator has to impute such motivation and intent using only the thin traces that the platform has recorded. Similarly, the ad-hoc team who assembled in response to the vandal do not have the luxury of being able to hear or see each other as they collaborate. In these kinds of communities, 'rich' observational data is unnecessary – or even distracting – when generating a qualitative account of their interactions.

4. Case: tracing vandals in Wikipedia

In a previously-published study [11], we used trace ethnography to detail the blocking of a vandal in Wikipedia. In that article, we focused on that user's vandalous edits and the process which led to their eventual blocking from Wikipedia after making approximately twenty inappropriate edits in a one hour period. The edits made were identified as vandalism and reverted by many different editors with many different software tools and mechanisms to coordinate their work within Wikipedia. In that article, we made many points regarding the distribution of cognition across a network, the social roles of software tools, and other topics that are not of primary concern in this paper. In this paper, we detail how we were able to generate such a rich account of activity in Wikipedia using our method of trace ethnography. We focus on our 'hands on' activity of inverting traces.

The activities we trace in our main example last approximately one hour. During that hour literally thousands of changes were made to the online encyclopedia that had nothing to do with our case and threatened to overwhelm the investigation. This is a significant problem of scale, encountered often in qualitative research that focuses on vast, distributed and highly populated phenomena such as Wikipedia. As Leigh Star notes, "the labor-intensive and analysis-intensive craft of qualitative research […] has never lent itself to ethnography of thousands." [25:383] This remains the case with trace ethnography, but the method facilitates the drawing out a single coordinated thread from the amongst the weave of ongoing activity, allowing the researcher to pull together a connected, focused stream of distributed activity.

Below, we identify a series of steps that we took to decode and invert various documentary traces.

4.1. Basics of tracing in Wikipedia

With Wikipedia, which runs on the open-source MediaWiki software platform, revision data constitute the bulk of the traces that we examine. While earlier versions of MediaWiki only allowed for these data to be generated for individual articles (constituting a step-by-step history of an article's development), the software now allows users to generate a wide variety of traces, including a history of a user's contributions, and the "recent changes" to every page in the site. From these traces, we can reconstruct the basic patterns of editing along a wide variety of vectors. In our original analysis, we used these data to trace a single vandal through all steps of the blocking routine, giving a separate account for each time a stage was escalated. Such a process was facilitated by the ability of the MediaWiki platform to list revision data and metadata chronologically by user and article. This made it easy to identify a banned user and then step back through each edit they made, looking at subsequent edits to these articles to see if they were reverted – and if so, how, when, by whom, using what tools, and at what stage in the routine.

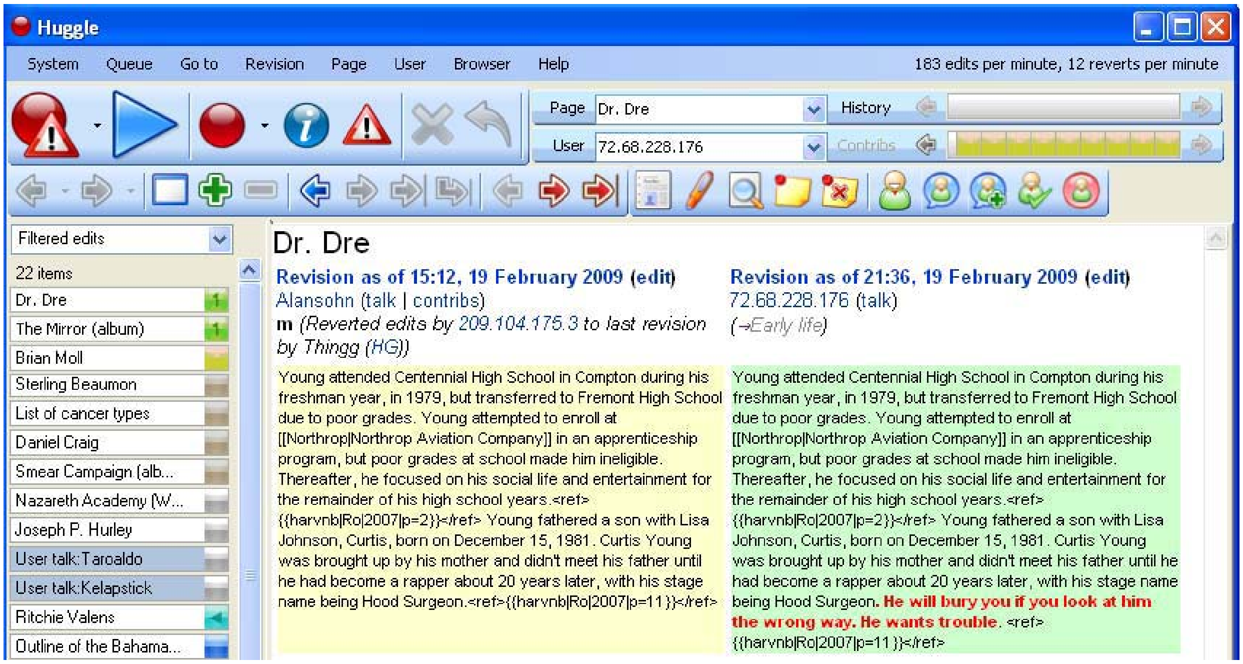
4.2. Identifying software tools

One of the most significant findings of our initial article was the growing use of semi-automated tools to enforce social order. These programs – some of which are so powerful that they are restricted to trusted editors – allow such users to revert malicious edits at immense speeds. As we described, such programs fundamentally change the social and cognitive aspects of administration in Wikipedia, making it such that a few dozen non-specialists can patrol every edit in near-real time. However, this would be completely invisible if only the top layer of the article revision data was examined, as edits made with these tools do not look any different from those performed manually. As we noted in section 3.1, it is only because of small trace markers in a metadata field that such programs reveal themselves.

The specific metadata field that drives this (and many other) inversion is the edit summary: the italicized, parenthetical text at the end of each revision in the above figures. When editing Wikipedia manually – that is, clicking the "edit this page" button in a browser – users can fill out a short field in which they can succinctly describe their actions. While these are often used sparsely and idiosyncratically by human editors, most fully-automated bots and assisted editing tools leave complex edit summaries with standard markers. When we first encountered these traces we essentially saw nothing remarkable; they were not even a part of our main analysis. Yet as we continued to see these traces in the course of observing other kinds of activity, it became apparent that some traces were highly-regularized, standing out amongst the other idiosyncratic human-authored summaries. By looking at all kinds of edit summaries, it became clear that there were several different types of regularly-occurring summaries, some with slightly different markers at the end. We gathered up the most commonly occurring types, and found that one of the most obvious distinctions was the "(HG)" and "(TW)" appended to the end of many summaries.

Becoming familiar with traces requires both immersion in the average, everyday affairs of a group and active investigation of otherwise backgrounded actors, software, and data. Once we identified a possible meaningful trace, we set out to discover why some have one marker, others have another, and most have none. This is the most classic element of our participant observation, as we had to enter the Wikipedian community to learn how and under what conditions these special traces were authored. We located users whose edits contained these markers and followed them in order to strategically situate ourselves as observers. As we found, such markers were predominantly left by users who identified themselves as 'vandal fighters', and searching through discussion spaces dedicated to this task, we came upon two terms that seemed to be related to our mysterious two-letter codes: Huggle and Twinkle. Despite the difficulty in initially locating these programs, we quickly found extensive documentation that explained what these programs were, as well as why their users were leaving highly-regularized edit summaries.

Once we discovered these tools, we began the participation stage of ethnography and installed them ourselves. In testing out the different features and then reviewing the traces we left, we were able to map out which actions lead to the production of certain kinds of traces. In the case of Huggle (Figure 2), Twinkle, and many other assisted editing tools, this was facilitated by their semi-automated nature, as such tools pre-script narrow paths of action. With such assisted editing tools the single click of a button often sets into motion a series of edits that ordinarily would have to be performed manually, such as reverting a series of edits made by a user and then sending them a pre-written warning. By participating as a vandal fighter and using



Dr. Dre

Revision as of 15:12, 19 February 2009 (edit)
Alansohn (talk | contribs)
m *(Reverted edits by 209.104.175.3 to last revision by Thingg (HG))*

Young attended Centennial High School in Compton during his freshman year, in 1979, but transferred to Fremont High School due to poor grades. Young attempted to enroll at [[Northrop|Northrop Aviation Company]] in an apprenticeship program, but poor grades at school made him ineligible. Thereafter, he focused on his social life and entertainment for the remainder of his high school years.<ref>{{harvnb|Ro|2007|p=2}}</ref> Young fathered a son with Lisa Johnson, Curtis, born on December 15, 1981. Curtis Young was brought up by his mother and didn't meet his father until he had become a rapper about 20 years later, with his stage name being Hood Surgeon.<ref>{{harvnb|Ro|2007|p=11}}</ref>

Revision as of 21:36, 19 February 2009 (edit)
72.68.228.176 (talk)
(→Early life)

Young attended Centennial High School in Compton during his freshman year, in 1979, but transferred to Fremont High School due to poor grades. Young attempted to enroll at [[Northrop|Northrop Aviation Company]] in an apprenticeship program, but poor grades at school made him ineligible. Thereafter, he focused on his social life and entertainment for the remainder of his high school years.<ref>{{harvnb|Ro|2007|p=2}}</ref> Young fathered a son with Lisa Johnson, Curtis, born on December 15, 1981. Curtis Young was brought up by his mother and didn't meet his father until he had become a rapper about 20 years later, with his stage name being Hood Surgeon. **He will bury you if you look at him the wrong way. He wants trouble.** <ref>{{harvnb|Ro|2007|p=11}}</ref>

Figure 2: Huggle screenshot (reconstructed), viewing an edit that was reverted by the vandal fighter

these tools for a substantial amount of time, we became quite familiar with the technical conditions under which each trace was produced. For example, users can revert an edit with Huggle in two nearly-identical ways: the simplest, most standard trace – Reverted edits by UserX (HG) – is left if the vandal fighter chooses to revert all recent edits to an article by a malicious editor. A slightly different edit summary – Reverted edits by UserX to last version by UserY (HG) – is left only if the vandal fighter specifically identifies a good version of an article to revert back to. This subtle distinction between two traces reveals two different work practices and rationales on the part of the user.

4.3. Decoding embedded routines

Another important finding of our research was that an important organizational routine had been embedded into many counter-vandalism programs. According to the project's guidelines, administrators will only ban users if they repeatedly edit in bad faith, disregarding enough warnings that their contributions are not constructive. While 'enough' can a subjective measure, we saw that most vandals are not referred to administrators until four escalating levels of warnings were issued and ignored. This was ultimately because of how programs like Huggle were configured: when a vandal fighter identifies a malicious edit, they can choose to also send them a pre-written warning with a single click. While earlier assisted tools required the user to manually select the appropriate warning level, current vandal fighting tools are sophisticated enough to know where the user is in the organizational routine and what the next appropriate response should be. When the vandal fighter decides to 'warn', the program retrieves all previous messages left for the malicious user and determines if any are recent warnings. If there are none, it leaves a polite first-level warning; if there are fewer than four, it leaves the next highest warning level; if there are already four, it refers the case to an administrator and requests they be blocked from editing.

One of the key insights of trace ethnography is that the same documentary resources which facilitate such automation of organizational routines also can be used to drive our analysis. So in this case, the semi-automated tool knows how many times a user has been warned by searching for traces left in pre-written warnings. Not all tools leave the same warnings (and many vandal fighters customize theirs extensively), but all tools that issue warnings leave standardized markers in HTML comment fields that all other programs can easily parse. And if a program like Huggle can automatically locate where a group of actors are within a routine or workflow using only trace data and algorithmic definitions of the process in question, then there is little reason why we could not do so as well.

The analytical process of inverting traces is similar to the description of identifying different software programs from revision metadata. Traces have to be identified and then matched up with specific stages of the routine. Once this familiarity is gained, one can quickly and easily determine where the user was situated in relation to Wikipedia's disciplinary mechanisms. In this case, the markers had a regular form: they were placed in an HTML comment form, they began with "Template:uw-", followed by a custom identifier, then a number signifying the warning level. While this would have been difficult (but not impossible) to identify inductively, we were aided by the fact that we could examine the source code of many of these algorithms to see precisely how they interpreted the traces.

This kind of analysis gets at the essence of trace ethnography's mixed-method approach: it requires initial ethnographic fieldwork to identify the possible kinds of routines present in an organization, which are then located and aggregated by performing detailed documentary analysis of trace data. Once all the relevant traces are inverted and a chain of activity assembled, the roles of various software tools can be easily identified.

4.4. Finding pre-assembled traces

While it is relatively commonplace to use revision data and metadata to trace the history of edits, it is less obvious how to trace actions taken by administrators, who use their technical access to, among other tasks, ban malicious users. Because this was the ending point of our account, it was important for us to identify the conditions under which our vandal was banned. While Wikipedia publically releases significant amounts of logging data documenting virtually every administrative action, these are note widely publicized and difficult for outsiders to notice. Because of this, one of the most helpful moments was when we came upon a specific moment in which Wikipedians had located, assembled, and inverted such logging data for their own purposes: in this case, for a disciplinary hearing regarding a controversy between administrators.

One reason why documentary traces are so useful is that they are produced and circulated in a specific sociotechnical environment, embedded with local meaning. While it is tempting to think of such data as ancillary, kept simply because computer systems log data, they are often used by members themselves to render accountable a number of social and organizational practices. In attempting to both identify sets of documentary traces as well as invert them, it is important to be alert to moments in which people assemble traces into narratives. These often occur in situations where evidence is presented – in our case, one of the first and most insightful examples of using revision metadata emerged from a disciplinary hearing from a 'wheel war'. In this controversy, a minor editorial dispute erupted as some administrators disagreed on where the discussion should take place. Administrators on opposing sides continually reversed each other, using their technical privileges to delete and re-create various discussion spaces. In determining who was at fault, the precise chronology of events (sometimes down to the seconds) became a key issue, leading one outside observer to reconstruct a detailed account of events from a variety of trace data. Because the outcome of this record was organizationally important, it was subjected to scrutiny by individuals from all sides, allowing us to be generally confident of its veracity. This allowed us to learn not only of new data sources, but also better techniques for following administrative actions.

4.5. To interview a database, talk like a bot

One of the most valuable aspects of finding this set of pre-assembled traces was that they led us to new sources of logging data and new types of traces. In the administrative controversy, revision metadata (and especially edit summaries) constituted the bulk of such evidence, but there were also links to the administrative logs as well as queries to the platform's Application Programming Interface (API). Nearly ubiquitous in contemporary software development, APIs are used by programs to communicate with each other. A tool that needs to know the time an article was last edited has two options: it can either pretend to be a web browser, download the entire article, and strip away all but the necessary metadata; or it can direct a standardized query to the API for only this information. In Wikipedia, the API is typically used not by humans but automated 'bots' that perform highly-regularized tasks. However, as we learned, APIs can be quite powerful for generating traces, allowing us to 'interview' the Wikipedia database about what it observed, in a matter of speaking.

In the administrative controversy, the API was used to gain detailed information about edits, allowing them to pinpoint actions down to the second (the web-based interface only presents hours and minutes). This was useful to us as well, and we were able to learn detailed information about the actions taken by the administrator who banned the vandal we were following. We changed the API query to retrieve all administrative logs left around the time of the vandalism incident, and searched for the major participants we had been following. Eventually, we found a rich trace record that detailed the block, and set off an entirely new round of data collection and analysis:

```
<item logid="20554611" pageid="0" ns="2"
title="User:72.68.228.176" type="block"
action="block" user="J.delanoy"
timestamp="2009-02-19T21:50:12Z"
comment="[[WP:Vandalism |Vandalism]]">
<block flags= "anononly,nocreate,noautoblock"
duration="48 hours" expiry="2009-02-
21T21:50:12Z" /></item>
```

We have only scratched the surface of the meaning that can be gleaned by fully inverting this trace. It would take significantly greater space than we have left in this paper to describe all of the details that this log entry reveals, and even more to describe how we came to learn such information. In inverting the researcher must always be guided be a set of (open and emergent) research questions, for traces have the potential to unfurl surprisingly complex narratives of coordinated activity.

5. Conclusion

This paper has shown the salience of trace ethnography for the study of distributed sociotechnical systems. The method is best for revealing the often invisible infrastructure that underlie routinized activities, allowing researchers to generate highly-empirical accounts of network-level phenomena without having to be present at every node. Trace ethnography does have its limitations, as it does not immediately allow researchers to grasp the larger sociocultural significance or history of the activities at hand. For example, the methodology described does not tell us how the organizational routine of four escalating warning was originally developed and implemented, or the attitudes vandal fighters hold towards it. In this sense trace ethnography is a cousin of ethnomethodology and actor-network theory, more suited for revealing practices, routines, distributed cognition and coordination devices than meanings, affect or 'inside the mind' cognition. However, trace ethnography is can (and should) be combined with other qualitative and quantitative methods, including traditional ethnographic, historical, archival, interview, survey, and statistical methods. It is easy to imagine a multi-methodological research project that integrates other approaches to gain 'thicker' and more holistic views of phenomena.

5.1. The invasive trace?

One of the most immediate issues trace ethnography raises are ethical, as there are significant concerns privacy and informed consent. Trace ethnography is premised on extracting more information about users and their actions than denoted by the 'purpose' of a log. Even the most privacy-sensitive users cannot imagine how a trace may be repurposed or combined with other traces. Under such conditions we cannot assume that such subjects have given their informed consent to participate in such research, even if they have freely and knowingly released such information, or if the information is public.

Such issues are not hypothetical, as there have been many widely-publicized instances in which seemingly-harmless trace data has been released to the public for research with disastrous consequences [20]. One of the first and well documented incidences is when America Online released the entire search history for 650,000 of its users over a three-month period. The data were anonymized by stripping usernames. However, since people routinely include bits of personal information in their search queries, these could be assembled together to identify particular individuals. Similarly, controversy erupted over the recent Netflix challenge, in which the movie rental company released records of its customer's movie reviews, identified only by zip code, age, and gender. Researchers quickly found that in many cases, this was all that was necessary to uniquely identify individuals using census data, especially in low-population zip codes. We have used trace ethnography to transform thin log entries to create thick images of activity and coordination, but as with many contemporary data mining techniques it could also be used in ways that challenge and reshape the boundaries of privacy.

5.2. Future research

While we have already successfully used trace ethnography to study Wikipedia, it can be easily deployed in the study of many other organizations or networks in which coordination occurs primarily through technologically-mediated channels. For us, the most promising candidates are open source software (OSS) development communities and emerging scientific cyberinfrastructure. OSS development communities share one of Wikipedia's most useful qualities for trace ethnography: a community ethos of openness and transparency that encourages public access to trace data. Such communities are often highly distributed and make significant use of software platforms to collaborate and organize. Software code repositories often keep detailed records of who changed what and when, and are often used to keep developers accountable and let maintainers know how the project has changed at a glance.

Similarly, we have found trace ethnography to be useful within our studies of scientific cyberinfrastructure. These are key sites for new configurations of interdisciplinary research, emerging models of collaboration across geographic distances and the implementation of innovative information technologies. For instance, we are currently investigating the use of grid and cloud computing models within physics. Within such approaches ticketing systems are often used extensively communities to support new forms of data sharing and analysis. For us, the logs of these ticketing systems reveal complex and detailed interrelationships of domain and computer scientists and the technicians that support their work.

In these and other distributed communities, analyzing these kinds of data, combined with ethnographic observation of how such sources are generated and used, can help us understand how such organizations are able to not merely cope with the problems of distance, but thrive in an increasingly-networked world.

6. References

[1] Beaulieu, A., From co-location to co-presence: Shifts in the use of ethnography for the study of knowledge. *Social Studies of Science 40*, 3 (2010), 453-470.

[2] Berg, M. and Bowker, G., The multiple bodies of the medical record: toward a sociology of an artifact. *Sociological Quarterly 38*, 3 (1997), 513–537.

[3] Bowker, G.C., Star, S., and Spasser, M., Classifying nursing work. *Online Journal of Issues in Nursing 6*, 2 (2001).

[4] Bowker, G.C., *Science on the run: information management and industrial geophysics at Schlumberger, 1920-1940*. MIT Press, Cambridge, Mass., 1994.

[5] Bowker, G.C. and Star, S.L., *Sorting Things Out: Classification and Its Consequences*. MIT Press, Cambridge, Mass., 2000.

[6] Cambrosio, A., Keating, P., and Mogoutov, A., Mapping Collaborative Work and Innovation in Biomedicine: A Computer-Assisted Analysis of Antibody Reagent Workshops. *Social Studies of Science 34*, 3 (2004), 325-364.

[7] Collins, H.M., Knowledge and controversy: Studies of modern natural science. *Social Studies of Science 11*, 1 (1981), 1–158.

[8] Fujimura, J.H. and Chou, D.Y., Dissent in science: Styles of scientific practice and the controversy over the cause of AIDS. *Social Science & Medicine 38*, 8 (1994), 1017-1036.

[9] Gaver, B., Dunne, T., and Pacenti, E., Design: Cultural probes. *interactions 6*, 1 (1999), 21-29.

[10] Geertz, C., *The Interpretation of Cultures*. Basic Books, New York, 1973.

[11] Geiger, R.S. and Ribes, D., The Work of Sustaining Order in Wikipedia: The Banning of A Vandal. *Proceedings of the ACM 2010 conference on Computer supported cooperative work (CSCW)*, Association for Computing Machinery (2010).

[12] Heath, C. and Luff, P., Documents and professional practice: 'bad' organisational reasons for 'good' clinical records. *Proc. CSCW 1996*, ACM (1996), 354-363.

[13] Howard, P.N., Network ethnography and the hypermedia organization: new media, new organizations, new methods. *New Media & Society 4*, 4 (2002), 550.

[14] Hutchinson, H., Mackay, W., Westerlund, B., et al., Technology probes: inspiring design for and with families. *Proc. CHI 2003*, ACM (2003), 17-24.

[15] Knorr Cetina, K. and Bruegger, U., Global Microstructures: The Virtual Societies of Financial Markets. *American Journal of Sociology 107*, 4 (2002), 905-950.

[16] Latour, B., *Science in Action: How to Follow Scientists and Engineers Through Society*. Harvard University Press, Cambridge, Mass., 1987.

[17] Latour, B., Circulating Reference: Sampling the Soil in the Amazon Forest. In *Pandora's Hope: Essays on the Reality of Science Studies*. Harvard University Press, Cambridge, Mass., 1999.

[18] Marcus, G.E., Ethnography in/of the World System: The Emergence of Multi-Sited Ethnography. *Annual Review of Anthropology 24*, 1 (1995), 95-117.

[19] Marcus, G.E., *Ethnography through thick and thin*. Princeton University Press, 1998.

[20] Narayanan, A. and Shmatikov, V., Robust De-anonymization of Large Sparse Datasets. *2008 IEEE Symposium on Security and Privacy*, (2008), 111-125.

[21] O'Day, V.L. and Jeffries, R., Orienteering in an information landscape: how information seekers get from here to there. *Proc. CHI 93*, ACM (1993), 438-445.

[22] Orr, J., *Talking about machines : an ethnography of a modern job*. ILR Press, Ithaca N.Y., 1996.

[23] Rieman, J., The diary study: a workplace-oriented research tool to guide laboratory efforts. *Proc. CHI 93*, ACM (1993), 321-326.

[24] Shapin, S. and Schaffer, S., *Leviathan and the Air-pump*. Princeton University Press, Princeton, 1985.

[25] Star, S.L., The Ethnography of Infrastructure. *American Behavioral Scientist 43*, 3 (1999), 377-391.

[26] Vaughan, D., *The Challenger Launch Decision*. University Of Chicago Press, Chicago, 1997.