Semantic social networks: a new approach to scaling digital ethnography

Alberto Cottica[1], Amelia Hassoun[2], Jason Vallet[3], and Guy Melançon[3]

[1] Edgeryders, EE; University of Alicante, SP
alberto@edgeryders.eu,
[2] University of Oxford, UK
[3] Université de Bordeaux, CNRS UMR 5800 LaBRI, FR

Abstract. We propose a data-based approach to doing ethnographic research in a digital environment. It has three main components. First, it treats online conversational environments as human communities that ethnographers can engage with as they would in onsite fieldwork. Second, it represents those conversations and the fieldnotes made by researchers thereon in network form. We call these networks *semantic social networks*, as they incorporate information on social interaction and their meaning. They encode a map of the associations between key concepts as perceived by informants as a group. Third, it uses methods borrowed from network science to process these data.
We present an application of this method to a large online conversation about community provision of health and social care, and discuss its potential for harnessing collective intelligence.

Keywords: digital ethnography, network science, collective intelligence

1 Introduction

The Internet Age has brought about a wave of exploration and innovation into digital ethnographic research methods. Substantial work has been devoted to methods that mine social networking platforms for user-generated content to analyse, often automatically. We propose an alternative approach based on convening an online conversation on the topic of interest. Such conversations function as virtual communities [17]. As such, they lend themselves to participant observation.

The digital nature of the conversational medium transforms the ethnographic evidence into structured data. This offers two opportunities. First, it allows ethnographers to do quantitative analysis on their own qualitative analysis. Secondly, as quantitative analysis functions as an aggregation layer, it allows ethnographers to handle larger volumes of evidence in coherent, replicable ways.

In what follows, we show how ethnographic evidence maps onto a type of network that encodes both social interaction and semantic association. We call these networks *semantic social networks (SSNs)*, and claim that a methodology based on them is highly accountable to ethnography as a a discipline. Its steps, save for the final quantitative analysis layer, carry naturally over from onsite field research to the digital domain. So, then, does ethnography's distinctive focus on groups of humans and their worldviews.

Additionally, SSNs are much more scalable than traditional ethnography. Coupled with open data and open standards, they can work well with thousands of informants a scale large enough for most applications, and much larger than that achieved by traditional ethnography. In sum, SSNs have the potential to evolve in a research method that (a) discovers *collective* worldviews of groups of humans; (b) can address open questions (unlike surveys) and (c) scales reasonably well (unlike onsite ethnography). They show promise as tools to harness *collective intelligence*, the processing of information by connected groups of humans [14].

In what follows, we describe the approach and present the results obtained by applying it to a digital ethnography dataset. We first introduce a data model for SSNs (section 2). Next, we present data in SSN form from a study on community-provided health and social care services (section 3. We then illustrate how we used SSNs to aggregate and navigate a large corpus of ethnographic evidence (section 4). Finally, we reflect on some possible extensions to SSNs and their potential for allowing digital ethnography to scale, while still maintaining its methodological advantages(section 5).

2 Semantic social networks: a generalised data model for digital ethnography:

Ethnography is a qualitative research technique aimed at discovering how a certain group of humans perceives a set of issues. Its unique value lies in that its findings encode the culture and worldview of the group being studied. This makes it especially suited for applications like foresight [1] and democratic stakeholder dialogue [7,21], where social and cultural meanings that arise organically from human interactions are the main objects of research rather than pre-conceived, researcher-imposed analytical categories.

Field-based ethnography treats access to informants embedded in communities as its most precious resource [10]. As the discipline expanded in topical scope and methodology, it retained its focus on extracting meaningful information by seeking analytical depth through engagements with relatively small numbers of informants [9]. This depth is typically achieved in part with long, repeated interviews with informants. Researchers then transcribe the text and associate transcripts to keywords, called *codes*, which form an ontology of concepts relevant to describe the problem at hand. These codes emerge from the ethnographer's embeddedness in the community she studies, gleaned through extended participant-observation which contextualises interview data in informants' larger environment [8,11].

Confronted with the rise of the global Internet, qualitative social scientists followed two main paths. One consists in mining social networking websites and applying quantitative analysis techniques to the retrieved material ([15,12]). We do not discuss it here.

The other consists in convening online conversations specifically to debate the issue at hand, and treating those conversations as ethnographic data. In this approach, informants co-construct and sustain visible themes of conversation through interaction with the researcher. Further, when an ethnographer is synchronically doing research with informants, she can contextualise the temporal unfolding of information rather than getting lost in noise as in other methods that analyse aggregated digital data after the fact [5]. This approach relates to works such as participant-observation with UNIX user-groups [13], online research with Anonymous hackers [6], and fieldwork in virtual worlds like World of Warcraft [16] and Second Life [3]. In these studies, anthropologists conducted long-term ethnography, interacting with informants in-setting, asking questions, and generating context-specific data that evolved through interactions with informants over time. Some projects included offline components [13], while others were completely undertaken online [3], but all pay close attention to the ways informants make sense of their own worlds and define their terminology.

To process ethnographic evidence at scale, we recast it as *data*. Data are characterised by a structure, common to all datapoints in a given dataset, that makes it amenable to being processed by machines. Machine processing, in turn, paves the way to research at scale. The specific challenge for ethnographers is to fit their evidence into a data structure without compromising its rich, contextual character. In this section, we describe the data structure we implemented in the course of a project called OpenCare.[4] It explores how communities of any kind provide health and social care, when neither states nor business can or will serve them. It consists mostly of an online interaction environment, where individuals share their experiences of care with others, discuss, and compare notes.

2.1 Primary data: contributions

SSN-based ethnographies start with the posts/comments on the social networking platform or online forums that hosts the conversation. We call *contribution* a testimony in written form (interview transcript, post on an online forum, etc.). A contribution is a datapoint of the study's primary dataset, the one generated by the informants themselves. A minimum viable structure for encoding a contribution as primary data includes:

Contribution ID The contribution's unique identifier.
Text The contribution's complete text.
Author ID A unique identifier for the informant that contributed the text.
Target ID A unique identifier for the informant that the text is addressed to.
Date and time

[4] http://opencare.cc

2.2 Secondary data: annotations

As we noted in section 1, ethnographers work by associating snippets of texts in the primary data to codes (keywords). This generates an ontology representative of the corpus of evidence at hand. In doing so, researchers produce secondary data. We call *annotation* the atomic result of this activity. A minimum viable structure for annotations as secondary data includes:

Annotation ID The annotation's unique identifier.
Contribution ID The unique identifier of the post or comment that this annotation refers to.
Snippet The part of the text in the contribution that the researcher wishes to associate with the code.
Code The ethnographic code associated to the snippet.
Author ID Unique identifier for the researcher that produced the annotation. It is useful in the case of multi-author studies.
Date and time .

This representation is sufficient to induce a network where the nodes are informants, and edges represent interactions. Codes associated to the interaction via annotations encode the semantics of that interaction. This is what we call a SSN (Figure 1). It proved to be easy to implement with most forum or blogging software applications; simple to process in meaningful ways (see below); and scalable. We propose it is general enough to fit the evidence from most ethnographies, while still rigid enough to encode it into well-formed datasets.

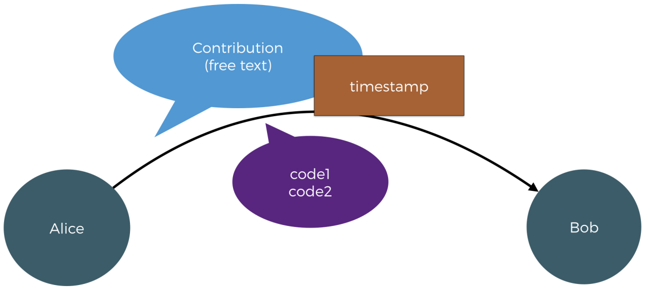

Fig. 1: An interaction between two informants that carries an ethnographic code id the atom of SSNs.

3 An application: the OpenCare data

The OpenCare project explores how communities of any kind provide health and social care, when neither states nor businesses can or will serve them. Data are gathered from an online forum where individuals share and discuss their experiences of community-provided care. At the time of writing, the forum consists of 439 long-form posts with 2,082 comments, authored by 254 unique informants. These were uploaded onto the online forum in the period between January 2016 and May 2017. This corpus was enriched with 4,555 annotations, employing 1,035 unique codes.

3.1 The OpenCare social network

Online conversation in OpenCare induces a social network where nodes are community members, and edges encode interaction. For two users A and B, we produce a connection $A \to B$ if A has commented B's content at least once. This network is directed ($A \to B \neq B \to A$) and weighted (the edge $A \to B$ has a weight of k if A has commented B's content k times). It has 249 nodes and 1,007 edges The main feature of this network is that there are no signs of polarisation, nor of balkanisation. Almost all participants are connected to the giant component, so that information is allowed to flow freely across the network. The giant component itself is not obviously resolved into distinct sub-communities (the Louvain modularity value is 0.33). These structural features allow us to conditionally accept that most opinions expressed in the forum have been run past someone (other than the proponent) in the conversation.

3.2 The OpenCare semantic network

The network representation that proved the most useful to ethnographic research is what we call the *co-occurrence network*. Its nodes are codes. Its edges are induced between two codes that occur in annotations that refer to the same post or content. This network is undirected ($A \to B \equiv B \to A$) and weighted (the edge has a weight of k if A co-occurs with B on k different posts or comments).

We can think of the co-occurrence network as an association map between the concepts expressed by the codes. A higher edge weight k indicates a stronger group-level association between the two codes connected by the edge.

The annotations on the OpenCare corpus induce a co-occurrence network with 1,035 nodes and 12,785 edges. The main component is formed of 990 nodes and 12,777 edges, and is an perfect example of small-world network as defined in [20], with a high average clustering coefficient $\bar{C} = 0.711$.

4 Results and discussion

4.1 Filtering the co-occurrence network for a high-level view

We can think of the codes co-occurrence network as a concept association map. Rather than representing the point of view of an individual, it encodes contributions from all informants, since informants are known to be in conversation with each other about the topic at hand. The resulting concept map, therefore, does not simply *aggregate* the association patterns of individuals, like a survey; it is the product of the *interaction* across participants. Edge weight k, then, represents the strength with which the conversation (as opposed to its participants) associates the codes connected by that edge.

Filtering the graph by higher value of k allows the researcher to single out the strongest associations between codes made by the informants as a group. Figure 2 shows the OpenCare code co-occurrence network for $k >= 5$ (69 codes, 96 edges). The color coding was generated by applying to it the Louvain community detection algorithm [2]. It is a highly modular network (modularity = 0.59), with clearly distinguishable groups of codes. This suggests that the debate self-organised into sub-issues.

Even with no access to the primary data, one can glimpse the structure of the debate among informants. Consider for example the cluster around `legality` (in blue): we find `existing system failure` and `regulation`, reflecting the preoccupation of some informants that community health care initiatives (much needed when systems fail), turn out to be illegal. We also find `safety`, reflecting the acknowledgement that regulation is there for a reason.

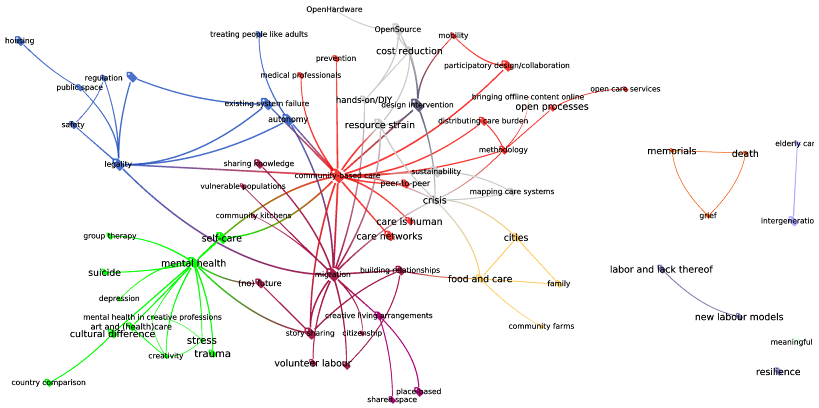

Fig. 2: The OpenCare code co-occurrence network (filtered for $k >= 5$).

4.2 Exploring the co-occurrence network for novelty

Connections that spark interest can be explored at a more granular level, either by focusing on lower-k co-occurrences between a chosen code and others (not shown), or by reading the original stories and posts in which the co-occurrences took place. Around `mental health`, `stress` and `trauma`, one finds the specific (often low-cost and social) tools that people use to manage stress, for example, learning about the usefulness of `group therapy` and online support groups, alternative therapies like `holistic healthcare`, or simply `getting outdoors`. The method allows for rich analysis on multiple levels, and never loses the granularity that makes ethnographic research so powerful.

Even at high levels, meaning comes through clearly and, crucially, high level connections illuminate connections that would be invisible at a smaller scale of analysis (for example the high-k connections between `mental health` and `trauma`, and `creativity`, `art and health care`and `story sharing`. In OpenCare, for example, these high-level connections made visible by the co-occurrence graph have enabled us to theorise that people are each other's healthcare technologies, and we have been able to detail and verify that theory through engagement with the granular details present in the stories. Without the co-occurrence network, vital interconnections would have been missed; without the detailed ethnographic data, the meaning behind those connections would be lost.

5 Extensions and future research

5.1 Open data and large-scale collaboration in ethnography

Social Semantic Networks hold the potential to make ethnography a large-scale collaborative discipline. For this to happen, we propose ethnographers embrace the practice of using and publishing open data. Open data are data that are (a) machine-readable, (b) published under licenses that allow their re-use, and (c) documented with appropriate metadata[5]. This paves the way for:

1. *Replication.* An ethnographer could pull in a colleague's primary and secondary data and check that the latter's process is clear. This increases the accountability of the research process.
2. *Large scale studies.* Accurate documentation of the code ontology allows ethnographers to work consistently on projects that would be too large for a single ethnographer to tackle. This would allow ethnographic studies at the scale of the thousands of informants.
3. *Multilingual studies.* The code ontology can be structured as a hierarchy, so that codes with the same meaning in different languages are entered in the secondary data as children of the same parent code. For example, `labour` could have `travail` and `arbeit` as children. The code co-occurrence network would be drawn between parent nodes, thus allowing both an all-languages view on data and across-languages comparisons.
4. *Reuse and extension.* An ethnographer could pull in a colleague's primary and secondary data, add her own coded corpus and use the combination of annotated corpora to produce a completely new study for example on responding to the refugee crisis, one of the care issues taken up by the OpenCare community.

[5] We have done this with OpenCare: https://doi.org/10.5281/zenodo.164970; https://github.com/opencarecc/opencare-data-documentation

5. *"Longitudinal ethnography"*. An online conversation could be revamped over time (for example every year) to keep track of how its collective point of view evolves.

The combination of these elements requires a cultural shift from practitioners. Ethnographers tend to work alone, and there is as yet no culture of open data in the discipline, as access to coded interviews and fieldnotes belongs to the ethnographer alone. To the best of our knowledge, the OpenCare dataset is the first-ever open dataset including primary and secondary data.

Yet, the payoff of such a shift is substantial. Ethnography could bring its unique methodological advantages to new problems, that demand a scale and consistency it now cannot supply. For example, we could imagine a version of Eurobarometer based on an open online conversation. Instead of answering multiple choice questions (vulnerable to framing biases [19], informants would discuss their perception of Europe, allowing researchers to discover novel patterns of association and detect the fading of old ones.

5.2 Other methodological improvements

In the future, we plan to test with at least three improvements:

1. Weighing contributions (and consequently annotations) by a "reliability score" derived by applying social theory on the social network topology [4].
2. Applying alternative ways to measure edge (association) strength k in the co-occurrence network; for example, $k(A \to B)$ could encode the number of informants that have authored contributions coded with both codes A and B, or the number of separate threads which contain at least one contribution with it, and so on. Different measures of edge strength have different interpretations, so they would allow different perspectives on the data corpus.
3. Observing and modelling the online conversation as a dynamic system. Stochastic Actor-Oriented Models might be a good place to start, despite known limitations ([18]).

6 Conclusions

Semantic social networks show some promise as a method for social research aimed at capturing collective intelligence ([14], with some of the advantages of both purely qualitative traditional ethnography and quantitative surveys. Like traditional ethnography (but unlike surveys), they deal well with open questions and novelty. Like surveys (but unlike traditional ethnography), they can handle hundreds of informants. When combined with open standards and open data, they could perhaps attempt to handle thousands of informants. We look forward to exploring this potential further.

Acknowledgment The OpenCare project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement no 688670).

References

1. Daniel Barben, Erik Fisher, Cynthia Selin, and David H Guston. 38 anticipatory governance of nanotechnology: Foresight, engagement, and integration. *The handbook of science and technology studies*, page 979, 2008.
2. Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.
3. T. Boellstorff. *Coming of Age in Second Life: An Anthropologist Explores the Virtually Human.* Princeton University Press, 2008.
4. R.S. Burt. *Brokerage and Closure:An Introduction to Social Capital: An Introduction to Social Capital.* Clarendon Lectures in Management Studies. OUP Oxford, 2005.
5. Gabriella Coleman. Ethnographic approaches to digital media. *Annual Review of Anthropology*, 39(1):487–505, 2010.
6. Gabriella Coleman. *Hacker, Hoaxer, Whistleblower, Spy: The Many Faces of Anonymous.* Verso, 2015.
7. John M Conley and Cynthia A Williams. Engage, embed, and embellish: Theory versus practice in the corporate social responsibility movement. *J. Corp. L.*, 31:1, 2005.
8. Robert Emerson, Rachel Fretz, and Linda Shaw. *Writing Ethnographic Fieldnotes.* University of Chicago Press, 2011.
9. Richard G. Fox, editor. *Recapturing Anthropology: Working in the Present.* School of American Research Press, 1991.
10. Clifford Geertz. *The Interpretation of Cultures.* Basic Books, 1973.
11. Erving Goffman. On fieldwork. *Journal of Contemporary Ethnography*, 123(18):123–132, 1989.
12. Betina Hollstein. *Qualitative Approaches*, pages 404–416. SAGE Publications Ltd, 2011.
13. Christopher Kelty. *Two Bits: The Cultural Significance of Free Software.* Duke University Press, 2008.
14. Pierre Lévy. Collective intelligence, 1997.
15. Anders K. Munk, Mette S. Abildgaard, Andreas Birkbak, and Morten K. Petersen. (Re-)Appropriating Instagram for Social Research: Three Methods for Studying Obesogenic Environments. In *Proc. of the 7th 2016 Int. Conf. on Social Media & Society*, SMSociety '16, pages 19:1–19:10. ACM, 2016.
16. Bonnie Nardi. *My Life as a Night Elf Priest:An Anthropological Account of World of Warcraft.* University of Michigan Press, 2010.
17. H. Rheingold. *The Virtual Community: Homesteading on the Electronic Frontier.* MIT Press, 2000.
18. Tom AB Snijders. Stochastic actor-oriented models for network change. *Journal of mathematical sociology*, 21(1-2):149–172, 1996.
19. Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. In *Environmental Impact assessment, technology assessment, and risk analysis*, pages 107–129. Springer, 1985.
20. Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.
21. Wendy Wolford. From confusion to common sense: using political ethnography to understand social mobilization in the brazilian northeast. In *New Perspectives in political ethnography*, pages 14–36. Springer, 2007.