Comparing techniques to reduce networks of ethnographic codes' co-occurrence

Alberto Cottica[⋆], Amelia Hassoun[∘], Jan Kubik[∘], Guy Melançon[•], Richard Mole[×], Bruno Pinaud[•] and Benjamin Renoust[⊙]

[⋆]Edgeryders, Estonia; [∘]Univ. of Oxford, UK; [•]Univ. of Bordeaux, France; [∘]The State Univ. of New Jersey, USA; [×]Univ. College London, UK; [⊙]University of Osaka, Japan

Keywords: ethnography, anthropology, networks, reduction, visual analytics

Extended Abstract

Semantic Social Network Analysis (SSNA) is a research method conceived for the social sciences. It consists in a combination of techniques drawn from digital ethnography and network science. The digital ethnography lineage ensures that SSNA remains open-ended and exploratory [1]. The network science lineage provides quantitative insight on the extent to which a statement vouched by one informant is shared by the others [2]. SSNA aims to combine the depth of ethnography with the breadth of surveys. To achieve this, it encodes annotated ethnographic corpora as structured data. Raw data consists of *Contributions*, recorded testimonies authored by informants and recorded in a database; *Annotations*, database objects created as ethnographers associate snippets of texts they find in contributions to keywords, called *Codes*; and codes themselves. We represent these data as a social network where the nodes are informants and edges represent conversational interactions. Codes – associated to edges via annotations – encode the semantics of that interaction. We call this a *semantic social network (SSN)*, and its analysis semantic social network analysis (SSNA). Here we focus on a transformation of the SSN, called the *codes co-occurrence network* (CCN) where nodes represent ethnographic codes and undirected edges represent co-occurrence. An edge between two codes means that both codes were used to code the same contribution.

We can think of CCNs as patterns of free associations, specifying the connections between the concepts encoded in ethnographic codes [6]. Ethnographers find them highly intuitive [2]. However, in a typical study (100–500 informants, 1,000–5,000 contributions) 1,000–2,000 codes might arise. Thus, CCNs tend to be both fairly large and dense, with tens of thousands of edges making them a dense networks known to be difficult to visualize [4]. Reducing a CCN could make it amenable to visual analysis by ethnographers. However, any network reduction entails a loss of information, and has to be regarded as a necessary evil. Reduction methods should always be theoretically founded, and applied with caution. We compare some reduction techniques and their theoretical groundings, and discuss their interpretations. First, we induce a CCN from a corpus obtained from an online forum discussing populist politics in Eastern Europe (336 informants, 2,284 contributions, 5,863 annotations and 1,445 codes, connected by 85,174 co-occurrence edges). Next, we attack it with alternative reduction techniques. Finally, we systematically assess each one in terms of criteria of quality prevalent in the literature on methods for qualitative research. Following [5], we evaluate the extent to which each reduction

technique: usefully supports inference, understood as an interpretation of the emerging intersubjective picture of the world; reinforces reproducibility and transparency that help to increase the researcher's ability to assess equivalence between any two implementations; does not foreclose the possibility of updating via abductive reasoning (algorithms alone do not decide how parameters should be set to get optimal readability); combines harmoniously with other parts of SSNA, such as coding and network construction.

Reduction techniques include: (i) reduction by dropping co-occurrences that occur only once or few times (Figure 1); (ii) reduction by dropping co-occurrences associated to a low number of informants; reduction by dropping edges not belonging to high-k k-cores [3].We propose an interdisciplinary approach in evaluating techniques to process ethnographic data. These techniques are, in themselves, purely mathematical, but are evaluated by an interdisciplinary team in terms of how well they support qualitative research in the social sciences.

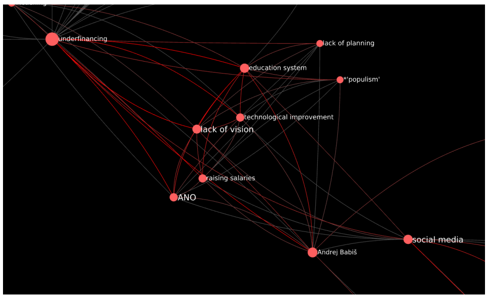

Figure 1: Detail of a CCN. Edge color maps to number of co-occurrences.

Funding

Work by Cottica, Hassoun, Kubik and Mole was partially supported by the Euuropean Commission's Horizon 2020 program under grant n. 822682.

References

[1] M. H. Agar et al. *The professional stranger: An informal introduction to ethnography*, volume 2. Academic press San Diego, CA, 1996.

[2] A. Cottica, A. Hassoun, M. Manca, J. Vallet, and G. Melançon. Semantic social networks: A mixed methods approach to digital ethnography. *Field Methods*, page 1525822X20908236, 2020.

[3] S. N. Dorogovtsev, A. V. Goltsev, and J. F. F. Mendes. K-core organization of complex networks. *Physical review letters*, 96(4):040601, 2006.

[4] M. Ghoniem, J.-D. Fekete, and P. Castagliola. On the readability of graphs using node-link and matrix-based representations: a controlled experiment and statistical analysis. *Information Visualization*, 4(2):114–135, 2005.

[5] G. King, R. O. Keohane, and S. Verba. *Designing social inquiry: Scientific inference in qualitative research*. Princeton university press, 1994.

[6] M. Stella, S. De Nigris, A. Aloric, and C. S. Siew. Forma mentis networks quantify crucial differences in stem perception between students and experts. *PloS one*, 14(10):e0222870, 2019.