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

(9,000 words)

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

## 1 Introduction

Since their inception, the social sciences have been split between qualitative and quantitative approaches and one of the most challenging undertakings has been to develop truly multi-method approaches that would combine the strengths of both and minimize their weaknesses.

We are working on a method that relies on both qualitative and quantitative techniques to increase the benefits of their complementarity. The former are employed at the stage of data collection – via in-depth interviews – and at the stage of analysis, when the ethnographically established contextual knowledge is employed in an iterative conversation with the patterns of thought revealed in our materials. Ethnographic coders – who are deeply immersed in the studied cultures – generate rich, hierarchically organized sets of codes. We analyze them not just to calculate frequencies of themes and motifs, but also to reveal their pattern of connectivity as multi-level networks that are also represented in compelling visualizations, that are the object of more analysis. In these visualizations, an ethnographic corpus is represented as a network. The nodes in the network correspond to ethnographic codes; the links connecting them represent the co-occurrence of codes in the same part of the corpus [10]. We call this network a *codes co-occurrence network* (henceforth CCN).

A problem that commonly arises is that the resulting networks are too large and dense for human analysts to process visually. Network science has come up with several (quantitative) techniques to reduce networks, based on identifying and discarding the most important edges in a network. It is relatively easy to apply them to this type of graphs. What is harder is to justify the choice of one or the other of these techniques, and of the values assigned to the parameters that they usually require. These choices are all the more important in the current context of growing doubts about the epistemological status of data processing [3]. In this paper, we propose criteria that ”good” reduction techniques for a CCN must meet. We next consider and compare – in the light of those criteria - four candidate techniques, using data from three ethnographic research projects that attack different research questions with similar methods. In doing so, we highlight the affinity of each of the four techniques with a prominent method of analysis associated in turn with an identifiable school of thought in sociology or anthropology. Our objective is to contribute to the rigor and transparency of the methodological choices of researchers when dealing with large ethnographic corpora.

We proceed as follows. In section 2 we discuss work related to our own. In section 3 we introduce the codes co-occurrence network, the network to be reduced. In section 4 we present our data. Next, we lay out criteria for choosing a technique to reduce a CCN for qualitative analysis, propose four such techniques, and proceed to apply them to the data (section 5). In section 6 we discuss our results, compare techniques with each other and propose a mapping of reduction techniques onto methods of analysis widely used in sociology or anthropology. Section 7 concludes.

# 2 Related work

The turn towards big data, fueled by radical improvements in computing power, has led to renewed faith in the ability of quantitative work to provide more generalizable and yet valid knowledge (that is, knowledge that preserves some of the richness of case-derived insights) than that obtainable by qualitative studies or quantitative projects relying on smaller numbers of cases [3].

This has led to exciting progress. At the same time, however, it has highlighted a pressing need for methodological robustness. As scientific work based on large datasets becomes methodologically innovative, more steps are needed to move from raw data[1] to final result. As a consequence, the methods themselves may be hard to check against the insights derived from intimate familiarity with specific cases. In combination with ”publish or perish” and with the premium placed by journals on counterintuitive, glamorous results, this has led to various epistemological crises. The replication crisis in psychology is the most famous of them [23], but not the only one. For example, it is claimed that half of the total expenditure on preclinical research in the US goes towards non-replicable studies [14]. Other tendencies that worry quantitative scientists, and data scien-

[1] Though the concept itself of ”raw data” is deemed problematic [3].

tists in particular, are: the persistence of citations of retracted papers [1]; the use of biased, bad-quality data in machine learning papers [28]; and the acritical acceptance of raw data as representative of base reality, when in practice data are constructed [3, 20]. All this leads to researchers obtaining divergent results depending on ostensibly innocent choices about data cleanup prior to analysis [12]. Even controlled experiments with different researchers working with the exact same datasets on the same research questions have led to spectacularly divergent results, for reasons that are not yet entirely clear [5]. Qualitative sociological and anthropological research is not expected to be replicable; rather, it derives its status as reliable knowledge from the rigor and accountability of the methods it applies. Therefore, careful, transparent choices about one's method is necessary at every step of the way, even more so when a research applies mixed methods [3].

This paper is meant as a contribution to making such choices in the particular case: that of reducing semantic networks that express qualitative data. The literature on semantic networks originates in computer science [32, 33, 36, 29]: its main idea is to use mathematical objects – graphs – to support human reasoning. Branching out from this tradition, we focus on the idea of network reduction. The latter is useful because it makes the networks in question more amenable to visual analysis, and helps researchers to appreciate, and interpret, the pattern of connectivity across the codes in their data. In doing so, we factor in previous work on the cognitive limits of humans to correctly infer the topological characteristics of a network from visual inspection [15, 24, 25, 31]. Such work confirms that large and dense networks are hard to process visually, and support the case for network reduction.

It is important to maintain full awareness of the implications of applying each technique. In this sense, this work is inscribed in the tradition of scholars who aim to apply systematic visualization techniques, while still retaining sensitivity to informants' contextual, interactional, and socioculturally specific understandings of concepts [13, 17, 34, 6]. In doing so, we are aware of the potential accountability issues – and even crises – that could come with the adoption of mixed methods. To prevent them, we fashion our mathematical techniques so that they do not violate the specific requirements of knowledge creation in anthropology and its chief method, ethnography.

# 3 The codes co-occurrence network and its interpretation

Consider an annotated ethnographic corpus. In what follows, we call any text data encoding the point of view of one informant (interview transcript, field notes, post on an online forum and so on) a *contribution*. Contributions are then coded by one or more ethnographers. Coding consists of associating snippets of the contribution's text to keywords, called *codes*. The set of all codes in a study constitutes an ontology of the key concepts emerging from the community being observed and pertinent to that study's research questions [2].

We can think of such an annotated corpus as a two-mode network. Nodes are of two types, contributions and codes. By associating a code to a contribution, the ethnographer creates an edge between the respective nodes[3].

From the two-mode network described above, we induce, by projection, the one-mode *codes co-occurrence network* (henceforth CCN). This is a network where each node represents an ethnographic code. An edge is induced between any two codes for every contribution that is annotated with both those codes. This network is undirected ($A \rightarrow B \equiv B \rightarrow A$). There can be more than one edge between each pair of nodes.

This representation is both intuitive and useful. It is intuitive because it has a clear-cut interpretation. We interpret co-occurrence as association. If two codes co-occur, it means that one informant has made references to the concepts or entities described by the codes in the same contribution, seen as a unit. Hence, this person thinks there is an association between the two. It is useful because it shows an association pattern for the whole conversation.

The downside is that CCNs tend to be resistant to visual analysis. This because they are large and dense. They are large because a large study is likely to use one or two thousand codes. They are dense as a result of the interaction of two processes. The first one is ethnographic coding. A rich contribution might be annotated 10 or 20 times, with as many codes associated to it. The second one is the projection from the 2-mode codes-to-contribution network to the 1-mode co-occurrence network. Recall that, in the latter, two codes are connected with an edge whenever they occur on annotations that annotate the same contribution. So, by construction, each contribution gives rise to a complete network (also called a clique) of all the codes associated to it, each of which is connected to all the others. Large, dense networks are known to be difficult to interpret by the human eye [15, 24].

[2] For a complete description of the data generation process, see [10], section 3.

[3] In graph theory, *nodes* and *edges* are the fundamental unit of which graphs are formed. Nodes are the entities being connected; edges, each linking two nodes, are the connecting entities.

However, the existence of parallel edges in the one-mode CCN offers theoretically grounded approaches to reduction.

# 4 Data and pre-processing

We use as data the annotated corpora from three ethnographic studies. One (OPENCARE) concerns community-produced health and social care services [11]; the second (NGI FORWARD), a policy-oriented discussion on the future of the Internet [8]; the third (POPREBL), the lived experience of Eastern European populist politics [9]. Though very different in scope, the communities being studied, and the languages of the contributions, the corpora are roughly similar in size, each with about 4,000 contributions by 300-400 informants. Their coding intensity is also roughly similar, with about 6,000 annotations and 1,000-1,500 codes each (table 1).

| | **OPENCARE** | **NGI FORWARD** | **POPREBEL** |
|---|---|---|---|
| informants | *276* | *331* | *366* |
| contributions | *3,737* | *4,068* | *3,686* |
| annotations | *5,731* | *5,871* | *6,660* |
| codes | *1,391* | *1,109* | *1,605* |

Table 1: The datasets used: some descriptive statistics

We proceed as follows: first, from each dataset we induce the relative (unreduced) CCN. The resulting CCNs are too large and dense for visual analysis (Table 2). Second, we apply to each of these stacked CCNs different techniques for network reductions. The techniques and the rationale for choosing them are the subject of Section 5. All techniques considered apply a reduction algorithm, the effects of which can be calibrated using a tuning parameter (two parameters in the case of the Simmelian backbone).

| | **OPENCARE** | **NGI FORWARD** | **POPREBEL** |
|---|---|---|---|
| nodes | *1,391* | *1,109* | *1,605* |
| edges | *25,720* | *149,971* | *106,369* |

Table 2: Size and order of the unreduced and the stacked CCNs

For each corpus and each technique, we then observe how varying the value of the tuning parameter influences the resulting reduced network. We attempt to find interpretations for choosing specific values of the tuning parameter.

Next, for each corpus and each technique we compute the maximal interpretable reduced network. By this, we mean the largest possible network that is still amenable to visual analysis, based on the relevant literature on network visualization [15, 24, 25].

Finally, for each corpus we assess the extent to which different reduction techniques select the same codes. We do this by computing the pairwise Jaccard coefficients between the maximal interpretable reduced networks that obtain from applying the different techniques.

# 5 Techniques for network reduction

## 5.1 What makes a good technique for network reduction?

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 as needed, and with caution. We propose four reductions techniques, each one related to a distinct theoretical traditions in the social sciences, particularly anthropology.

Following [18], we propose that a good reduction technique should:

1. Usefully support inference, understood as a simplifying interpretation of the emerging intersubjective picture of the world. The main contribution of network reduction to ethnographic inference is that it makes the CCN small and sparse enough to be processed visually [24, 15]. A substantial part of the human brain's capacity is allocated to processing images, so it makes sense to invest in good visualizations. A well-established literature – and techniques such as layout algorithms – help us define what a ”good” network visualization is.

2. Reinforce reproducibility and transparency. Reproducibility means that applying the same technique to the same dataset will always produce the same interpretive result (even if the technique has a stochastic component). Transparency means that how the technique operates is clear to the researcher, who can therefore assess which technique best suits her purpose, and explain that assessment to her peers.

3. Not foreclose the possibility of updating via abductive reasoning. Algorithms alone do not decide how parameters should be set to get optimal readability. Rather, the values of the parameters are co-determined by the ethnographers who possess rich empirical and theoretical knowledge of relevant contexts.

4. Combine harmoniously with other steps of the data processing cycle, such as coding and network construction.

With that in mind, we turn to the discussion of candidate techniques. We claim that all of them satisfy more or less equally conditions 3 (parameters are set by the researchers), 4 (they use in a natural way research data and the way they organize in network), and the reproducibility condition in 2 (the only stochastic component come into play in layout algorithms, and they produce visually equivalent layouts). Most of the discussion below therefore focuses on how well candidate techniques support inference (condition 1) and how transparent they are (condition 2). In other words, how useful the visualizations they produce are, and how intuitive the method of building them is to ethnographers.

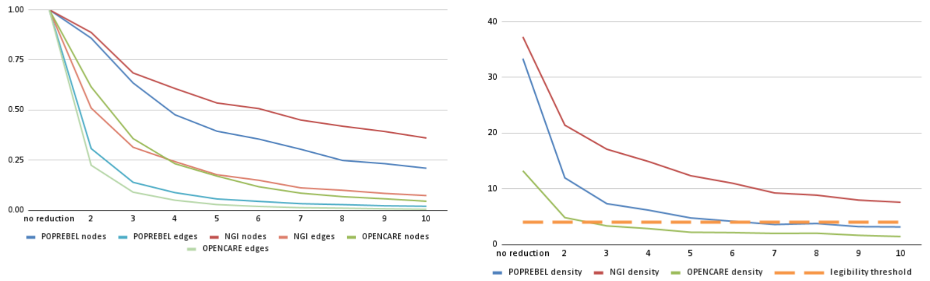
## 5.2 Association depth

A first way to reduce the CCN is the following. For each pair of nodes in the network connected by at least one edge, remove all $d$ edges connecting them, and replace them with one single edge of weight $d$, regardless of how many informants contributed to the analysed discourse. This yields a weighted, undirected network with no parallel edges.

$d$ has an intuitive interpretation in the context of ethnographic research. Consider an edge $e = code1 \leftrightarrow code2$. $d(e)$ is the count of the number of contributions in which $code1$ and $code2$ co-occur. Since we interpret co-occurrence as association, it makes sense to interpret $d(e)$ as the depth of the association encoded in $e$. This gives us a basis for ranking edges according to the value of $d$. The higher the value of $d$ of an edge, the more important that edge.

There is also a straightforward interpretation of the special case $d(e) = 1$. It means the association between `code1` and `code2` occurs only once in the corpus. It might be profoundly insightful, but it did not echo in the rest of the corpus. In a sense, it could represent the discursive isolate, an analog of a statistical anomaly, an outlier. Dropping all edges $e : d(e) = 1$ reduces the network at what seems to be an acceptable cost.

Generalizing, we can drop all edges for which $d(e) \leq d^*$. As the value of $d^*$ increases, so does the degree to which the reduced network encodes high-depth associations between codes. Choosing an appropriate level below which to drop edges means managing a trade-off. The higher the threshold, the greater the information loss. At the same time, though, the higher the threshold, the greater the legibility of the reduced network, and the clearer the picture of the basic structure of discourse in a given community, within which our respondents create meaning and make sense of the world around them.



(a) Proportional reduction in numbers of nodes and edges of the CCN for different values of $d^*$

(b) Visual density of the reduced CCN for different values of $d^*$

Figure 1: Reducing codes co-occurrences networks, according to association depth, in three ethnographic studies.

Figure 1a shows how the number of nodes and edges in the reduced co-occurrences networks of three SSNA studies decrease as we increase the value of $d*$. The unreduced weighted networks for our three datasets have 1,000 to 1,500 nodes and 18,000 to 55,000 edges each. As $d*$ increases, these numbers decrease rapidly.

Just setting $d^* = 2$ – which means only discarding one-off edges – leads to a 50-75% decrease in the number of edges. $d^* = 10$ leads to a decrease of about two orders of magnitude in the number of edges.

Figure 1b shows the decrease in network density (number of edges divided by the number of nodes) as the technique is applied with increasing values of association depth $d^*$. Unreduced networks are very dense, with 15-40 edges per node. Discarding edges $e, d(e) = 1$ reduces density by about half, but in two out of our three datasets they remain well above the value of 4 edges per node, sometimes quoted as the one that makes for comfortable visual processing [24, 25].

## 5.3 Association breadth

A second way of reducing the CCN is the following. For all pairs of nodes $code1, code2$ in the network, remove all edges $e : code1 \leftrightarrow code2$ connecting them, and replace them with one single edge of weight $b$, where $b$ is the number of informants who have authored the contributions underpinning those edges. Like in section 5.2, this yields a weighted network of codes with no parallel edges, but now edge weight has a different interpretation.

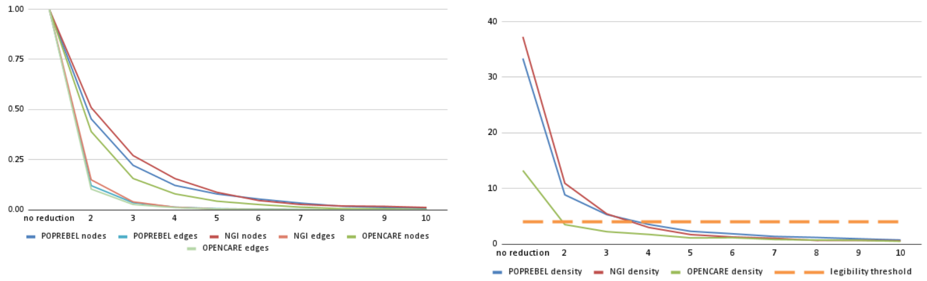
Recall that each edge $e$ in the unweighted CCN is induced by one, and only one, contribution in the corpus, which was coded with both $code1$ and $code2$. This contribution has only one author. Instead of counting contributions to the corpus, like interviews or forum posts, we are counting the related informants. This also has a straightforward interpretation for ethnographic analysis. The greater the value of $b(e : code1 \leftrightarrow code2)$, the more widespread the association between $code1$ and $code$ is in the community that we are studying. We interpret it as association breadth.

In ethnographic work, it is not rare that the same two codes occur over multiple contributions of the same informant. This happens when some informants are focused on a set of ideas, which recur when they make multiple contributions. This means that there is a mathematical relation between association breadth $b$ and association depth $d$:

$$\forall e : b(e) \leq d(e) \tag{1}$$

Like for association depth $d$, the case where association breadth $b(e) = 1$ has a straightforward interpretation. It means the association between `code1` and `code2` is endorsed by only one single informant. Again, it might be profoundly insightful, but it did not occur to anyone else in the community. It could reflect an idiosyncrasy of that particular person. Dropping all edges $e : b(e) = 1$ reduces the network at what seems to be an acceptable cost.

As we did for depth, we can drop all edges for which $b(e) \leq b^*$. As the value of b* increases, so does the degree to which the reduced network encodes broadly shared associations between codes. And again, the higher the threshold, the greater the information loss, and the greater the legibility of the reduced network, but thus also the clearer the picture of the cultural or ideological homogeneity in a studied community of discourse.



(a) Proportional reduction in numbers of nodes and edges of the CCN for different values of $b^*$

(b) Visual density of the reduced CCN for different values of $b^*$

Figure 2: Reducing codes co-occurrences networks, according to association breadth, in three ethnographic studies.

Figure 2a shows how the number of nodes and edges in the reduced co-occurrences networks of three SSNA studies decreases as we increase the value of $e^*$. Setting $e^* = 2$ – which means only discarding ”idiosyncratic” edges – leads to a 85-90% decrease in the number of edges. $d^* = 4$ leads to a decrease of about two orders of magnitude in the number of edges.

Figure 2b shows the decrease in network density (number of edges divided by the number of nodes) as the technique is applied with increasing values of association depth $b^*$. Discarding edges $e, b(e) = 1$ reduces density by 70-75%, but again in two out of our three datasets they remain well above the value of 4 edges per node.

## 5.4 Highest core values

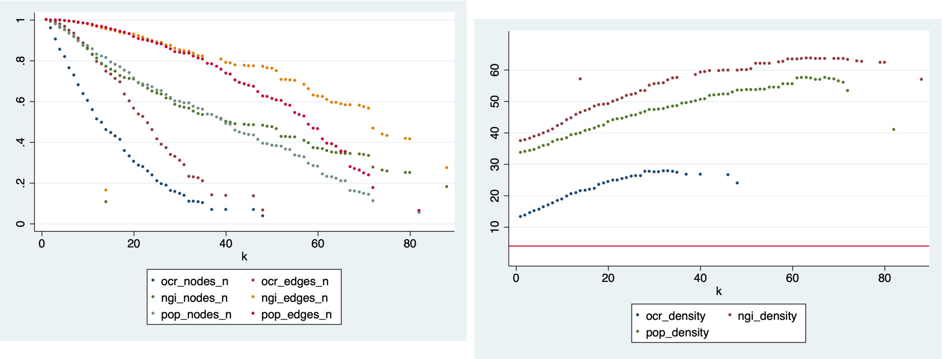
An alternative way of identifying the most important edges in a CCN is to exploit the topology of the network. For example, a co-occurrence edge could be considered important if it connects two nodes that are both connected to a large number of other nodes. A community of such nodes can be identified by computing the CCN's $k$-cores. $k$-cores are subgraphs that include nodes of degree at least $k$, where $k$ is an integer. They are used to identify cohesive structures in graphs [16]. Random graphs have the property that a giant $k$-core appears in them when their edge density becomes high enough [27].

After computing all the $k$-cores of a network, its nodes can be assigned a core value. A node's core value is the highest value of $k$ for which that node is part of a $k$-core.

To find the most important edges in the CCN, we again replace all edges between any pair of connected codes $code1$ and $code2$ with one single edge $e(code1, code2)$. Next, we remove all the codes whose core values $k$ are smaller than 1, as well as their incident edges. If the graph thus reduced is still too large and dense, we increase the value of $k$ to the next integer and repeat, until the reduced graph is interpretable by visual analysis. Notice that this method ignores edge weight; co-occurrence edges are included in the reduced network only on the basis of the number of codes that the two co-occurring codes are connected to.

In contrast to the techniques of reduction presented in sections 5.2 and 5.2, this approach to network reduction is not very effective at low levels of the tuning parameter $k$. Only for high values of $k$ does the CCN reach a substantial reduction in nodes (under 100), and even then it maintains a very high number of edges (1,000 to 10,000). As for edge density, it increases with $k$, staying well over the legibility threshold of 4. This is shown in figure 3.

Persistent high density and limited reduction are artifacts of the way in which $k$-core decomposition works. High-degree nodes are discarded last, so the highest-$k$ core is composed only of nodes with many connections to one another.



(a) Proportional reduction in numbers of nodes and edges of the CCN for different values of *k*

(b) Visual density of the reduced CCN for different values of *k*

Figure 3: Reducing codes co-occurrences networks, according to the core values of nodes, in three ethnographic studies.

In any affiliation network, like networks of co-authorship of academic papers or CCNs, the distribution of nodes' core values is disproportionately influenced by the presence of very large ”outlier” cliques. Some authors recommend dropping these cliques from the data manually [16]. In our case, a long and interesting contribution might be coded with as many as 50 codes. Each of those codes becomes connected to the other 49 in the CCN, driving their core values to at least 49, even if they do not appear anywhere else in the corpus. Moreover, in general, when the distribution of core values is fat-tailed, higher-*k* cores tend to be dominated by codes in the most heavily coded contributions, and so by the most vocal informants, who are able to deliver long and dense contributions. This is not necessarily what analysts want.

## 5.5 Simmelian backbone extraction

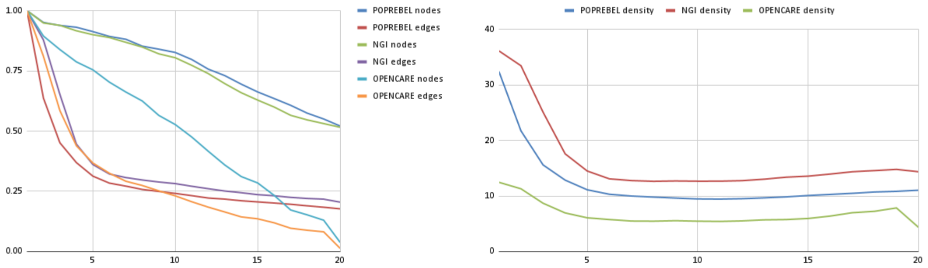
Another way to exploit the network's topology to identify its most important edges is to extract its Simmelian backbone. A network's Simmelian backbone is the subset of its edges which display the highest values of a property called redundancy [26]. An edge is redundant if it is part of multiple triangles. The idea is that, if two nodes have many common neighbors, the connection between the two is structural. This method applies best to weighted graphs. Both association depth and association breadth are natural measures of edge weight in CCNs. In what follows, we use the former.

In order to reduce a network by extracting its Simmelian backbone, we proceed as follows. First, we choose a value for the granularity parameter *k*. For each pair of nodes $n_1, n_2$ in the network, the incident edge $e(n_1, n_2)$ is considered as strong depending on the overlap between the $k$ strongest-tied neighbors of $n_1$ and those of $n_2$, called redundancy. We set $k$ to be approximately equal to the average degree in each dataset. At this point, the network can be reduced based on the redundancy value of each edge. We start dropping the lowest-redundancy edges, then gradually increase the redundancy threshold to obtain smaller and smaller networks.

As we filter for increasing values of minimum redundancy, the number of nodes decreases, but not very rapidly and with a more or less linear pattern for all datasets. The number of edges drops rapidly for low values of the minimum redundancy, but then decreases much more slowly when the network's minimum redundancy rises above 5. Consequently, edge density sees a rapid drop in the early phases of the reduction, after which it becomes more or less constant. Throughout the reduction process, the density of all three datasets stays over 4 (figure 4).

However, networks reduced with this method appear more legible to human analysts than those reduced with the highest core values method. This is because, by construction, Simmelian backbone extraction tends to leave dense communities of nodes intact, while discarding edges that connect different communities. As a consequence, reduced networks are highly modular, and feature connected components[4] breaking off the network's main body: they can be visually interpreted as small networks of communities of nodes, instead of as large networks of individual nodes (fig 5d. This appears to be semantically justified; the codes within each of the communities are closely related. However, the same process tends to break the reduced network down into many densely connected components, which destroys structural information. So, with this technique, there is a tradeoff between the reduction in the number of nodes and edges, on the one hand, and the preservation of a recognizable overall structure, on the other.

[4] A connected component of a network is a subnetwork in which any two nodes are connected to each other by paths, and which is connected to no additional vertices in the rest of the graph. They look like "islands" of nodes. Some connected components are visible at the top of Figure 5d.



(a) Proportional reduction in numbers of nodes and edges of the CCN for different values of minimum edge redundancy

(b) Visual density of the reduced CCN for different values of minimum edge redundancy

Figure 4: Reducing codes co-occurrences networks, by the extraction of their Simmelian backbones, in three ethnographic studies.

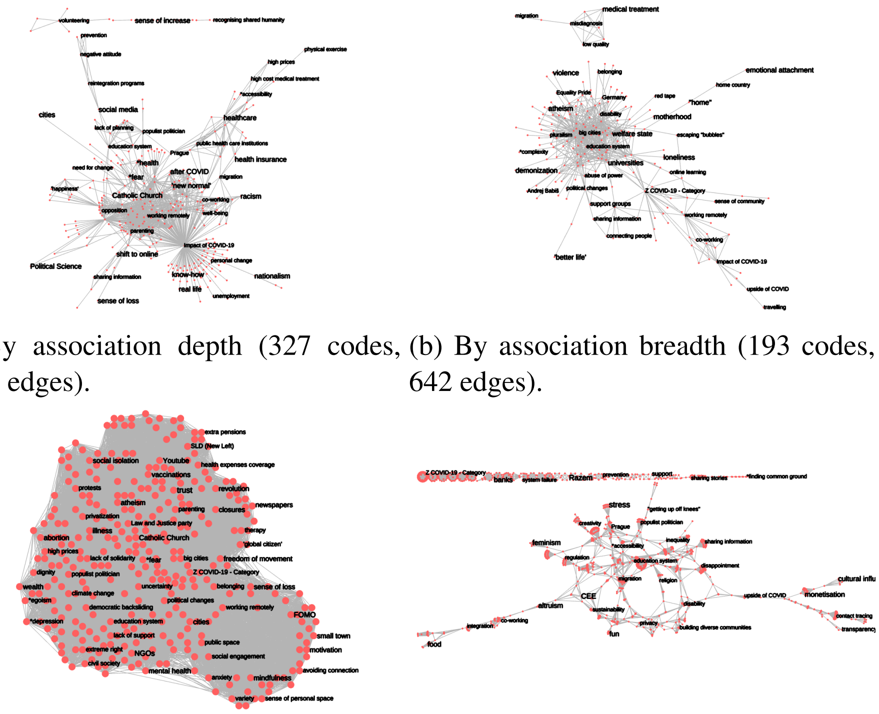
# 6 Discussion

## 6.1 Comparing reduction techniques

Reducing any network implies ranking its edges in order of importance, so that the least important edges can be dropped to simplify visual analysis. In section 5 we introduced four techniques for reducing a CNN. These techniques employ two different strategies to discover the CNN's most important edges. Two of them – by association depth and by association breadth – rank edges according to the value that a chosen property, interpreted as edge weight, assumes for each individual edge. The other two techniques – by highest core values and by Simmelian backbone extraction – use the topology of the network to rank the importance of the edges (although the latter also employs a measure of edge weight to do so). In this section we compare the relative merits of the two strategies and four techniques, based on the criteria set in section 5.1. We discuss four aspects: interpretability of the reduction techniques themselves; harmonious integration with the pre- and post-reduction phases of the data processing cycle; quantitative effectiveness; and preservation of structural information in the reduced networks. The discussion is summarized in table 3 and exemplified by figure 5.

| Criteria | ass. depth | ass. breadth | core values | Simmelian backbone |
|---|---|---|---|---|
| 1 | yes | yes | somewhat | yes |
| 2 | yes | yes | somewhat | somewhat |
| 3 | yes | yes | no | yes |
| 4 | yes | yes | no | somewhat |

Table 3: A comparison of four CCN reduction techniques against the criteria set in section 5.1. 1: Usefully supports inference; 2: Reinforces reproducibility and transparency; 3: Does not foreclose abductive reasoning; 4. Combines harmoniously with the other steps of the data processing cycle.



(a) By association depth (327 codes, 1,059 edges).

(b) By association breadth (193 codes, 642 edges).

(c) By highest core values (344 codes, 19,716 edges).

(d) By Simmelian backbone extraction (1,368 codes, 13,428 edges).

Figure 5: Reduced networks of the POPREBEL CCN applying four techniques.

The two reduction techniques based on edge weight are likely to be more intuitive to qualitative researchers without extensive training in network analysis. The measures of edge weight we adopted are grounded in the familiar practice of ethnographic coding; this results in a straightforward definition of what a strong edge is, and how reduced networks obtain. By implication, these techniques effectively combine network reduction with the early (ethnographic coding and network induction) and late (analysis of the reduced network) phases of the data processing cycle. Ethnographic coding drives the entire cycle.

Conversely, interpreting network reduction based on the highest core values of nodes or on Simmelian backbone extraction requires a certain amount of topological thinking. While doing so is certainly possible to qualitative researchers, it does require some extra effort. In this sense, these two techniques are a little less transparent than the former two. These techniques also combine with ethnographic coding upstream of reduction, but less so in the reduction phase itself, as edge weight is irrelevant to the highest core value and only in part relevant to ”simmelianness”.

In terms of quantitative effectiveness, the techniques based on edge weight outperform those based on topology. Both allow for fine-tuning between high information loss and readability; and, for all datasets, they allow to produce small, low-density networks (figures 1 and 2). Reducing by highest core values does not allow such granularity, because the highest-$k$ cores are still a substantial part of the unreduced networks; neither does it allow readability, because they are also dense, far above the threshold for human interpretability [25]. The extraction of a Simmelian backbone has the same issues, but they are mitigated by two factors. First, the granularity parameter $k$ can be increased to allow a more drastic reduction of the network. And second, the reduced networks are highly modular, and that allows for better readability for a given network size and the identification of clusters, if any, within it.

We now turn to how well structural information is preserved through network reduction. The first two methods yield reduced networks that preserve structural information, in the sense that they do not predetermine it: for example, the modularity of the reduced networks varies across our different datasets. The third and fourth method use topological information for the reduction process itself, and they both predetermine the structure of the reduced network. Reducing a CCN to the subnetwork formed by the codes with the highest core value invariably leads to a very dense network. The best a human analyst can do with it is ignore the edges altogether, and treat it as a list of important codes. Reducing it to a Simmelian backbone invariably leads to highly modular (thus legible) reduced networks. Unfortunately, as the value of the reduction parameter increases, communities of codes break off from the main body of the network and form entirely separate connected components; this destroys information about the overall pattern of connectivity in the corpus.

## 6.2 Mapping network reduction techniques onto methods of analysis in sociology and anthropology

Deciding which network reduction technique is best suited to a particular research project will largely depend on the researcher's ontological and epistemological beliefs, i.e. on assumptions about the nature of social reality and how this social reality can be known, as well as on the nature of the project itself, particularly the questions it asks.

Each of the four reduction techniques reveals a different set of attributes semantic networks have. It also turns out that each technique bears a family resemblance to a prominent method of analysis associated in turn with an identifiable school of thought in sociology or anthropology.

Determining association depth is in its essence a method of uncovering the structure of culture (discourse or thought), a task placed in the center of anthropology most prominently by Claude Levi-Strauss. His classical works *Anthropologie structurale* [22] and *La Pensée sauvage* [21] initiated a whole host of structuralist and post-structuralist approaches.

For post-structuralist sociologists and anthropologists, social relations can only be understood by analysing how they are constituted and organised through discourse. In other words, social hierarchies, norms and practices are legitimised (or delegitimised) by granting the meaning attached to specific concepts a dominant position, enabling certain ideas to become hegemonic, i.e. widely accepted as the 'Truth'. For example, the idea that ethnic nations are natural entities growing out of shared kinship ties (all academic evidence to the contrary) is used to legitimise political control by the core nation and the marginalisation of minority ethnicities. Moreover, discourse scholars work from the assumption that the meaning respondents attach to floating signifiers is relational within a discourse. Within a patriarchal discourse, the meaning attached to 'woman' is directly determined by the meaning attached to 'man', for instance. To understand the meaning of concepts, it is thus essential to understand their interrelationships; discerning which meanings are hegemonic further requires us to understand which interrelationships between concepts are dominant. Focusing on association depth is thus a useful way of bringing into sharper focus the interrelationships between concepts that are most commonly used by our respondents, thereby providing a picture of the basic structure of discourse in a given community, within which our respondents create meaning and make sense of the world around them. It is important to note that the decision to stop data collection upon reaching what is deemed sufficient saturation, a common strategy in qualitative research, would skew the results of this method.

For anthropologists, the concept of association breadth is most closely associated with network analysis, an approach whose classical formulations came from

Jeremy Boissevain and his followers [4]. Since it helps to identify an important attribute of networks not just among concepts but also actors who employ them, it seems to be particularly useful in reconstructing the structures of communities of discourse [37] or discursive fields [30]. In short, this reduction method is designed to simultaneously capture information about connections between concepts and between people who employ them; it reveals networks emerging among the concepts used by the largest number of participants.

The technique based on core values of codes is designed to determine the centrality of certain concepts in a discourse. While it does not allow for the reduction of edges (as the number of edges is the information at the center of this approach), it shows which concepts have most edges associated with them. It facilitates, therefore, a more systematic determination which discursive elements constitute what is known in cultural anthropology as root paradigms, key metaphors, dominant schemas or central symbols of a given culture [35, 2].

Finally, the Simmelian backbone extraction can contribute to the discovery of hegemonic and counter-hegemonic clusters (subcultures) of meaning in an analyzed body of discourse [7, 19]. No culture is fully integrated and each is subjected to centripetal and centrifugal forces simultaneously. As a result, even in the most "homogenous" cultures one can identify at least embryonic subcultures or – in another formulation – for every hegemony there is a budding or fully articulated counter-hegemony. The point is that a hegemony is usually built not on single symbols or concepts but on their interconnected clusters. This reduction technique helps to identify such clusters and facilitates the operationalisation of their internal coherence.

## 6.3 Do different techniques select the same codes and edges?

*A priori*, we expect different reduction techniques to select into the reduced networks codes and co-occurrence edges that are different, but not completely different from technique to technique. Different techniques prioritize different edges, and, therefore, codes. At the same time, the key co-occurrences are likely to meet the criteria of every technique. In order to quantify the extent to which different techniques converge onto the same set of codes and edges, we proceed as follows.

First, we apply each of the four techniques to each of our three datasets. For each technique-dataset pair, we compute a maximal interpretable network (MIN). By this we mean the largest network that is still interpretable by a human analyst. We then take the four MINs of each dataset, and compare them pairwise by computing the Jaccard indices on their nodes and edges.

The main difficulty with the above is to define the MINs. While graph layout algorithms have focused on minimizing edge crossing, symmetry, and other such layout properties, there is little research on how the visual representation of a graph influences the perception of quantitative properties of that graph [31]. Some attempts have been made to correlate graph attributes (like density and order) with the ability of humans to correctly perceive basic graph properties like diameter or shortest paths [15, 31]. We would instead like to use the CCN mostly to derive insights on the overall shape of the association patterns in a large corpus. In the absence of a systematic literature on the readability of graphs, we fall back on the result that graphs become difficult to interpret once their number of edges rises above four times the number of their nodes, confirmed by several authors [15, 24, 25]. The MIN, then, becomes the largest reduced graph for which

$$\frac{E}{N} < 4 \tag{2}$$

Where $E$ is the number of edges in the reduced network, and $N$ the number of its nodes.

This criterion does indeed provide a MIN when applied to reduction based on $d(e)$ and $b(e)$. However, no amount of reduction based on highest core values and Simmelian backbone extraction yields a reduced network that satisfies condition 2. For those techniques, we have to adopt other definitions of MIN.

For highest core values, we simply define the MIN as the size of the $k$–core with the largest value of $k$ in the unreduced network. This MIN is much too dense to be interpreted as a graph, but it does provide the ethnographer with a list of codes, that constitute the highest cohesion group of codes in the corpus.

For Simmelian backbone extraction, we exploit the property of Simmelian backbones to filter out edges connecting different communities of nodes, preserving those that connect different nodes in the same community. This greatly increases the visual legibility of communities of nodes [26]. However, as the value of the reduction parameter increases, it produces reduced networks that break down into several connected components, which destroys information on how these communities connect to each other. The latter is clearly valuable to ethnographers, because it is a part of the structure of the discourse in a corpus. So, we define the MIN as the smallest Simmelian backbone of the original network that still has a giant component.

# 7 Conclusions

# References

[1] Zombie research haunts academic literature long after its supposed demise. *The Economist*, 2021.

[2] M. J. Aronoff and J. Kubik. *Anthropology and political science: A convergent approach*, volume 3. Berghahn Books, 2013.

[3] A. Beaulieu and S. Leonelli. *Data and Society: A Critical Introduction*. SAGE, 2021.

[4] J. Boissevain and J. C. Mitchell. *Network analysis: Studies in human interaction*. Walter de Gruyter GmbH & Co KG, 2018.

[5] N. Breznau, E. M. Rinke, A. Wuttke, M. Adem, J. Adriaans, A. Alvarez-Benjumea, H. K. Andersen, D. Auer, F. Azevedo, O. Bahnsen, et al. Observing many researchers using the same data and hypothesis reveals a hidden universe of data analysis. 2021.

[6] J. Burrell. The field site as a network: A strategy for locating ethnographic research. *Field Methods*, 2(21):181–199, 2009.

[7] J. A. Buttigieg. *Prison Notebooks*. Columbia University Press, 2010.

[8] A. Cottica and A. Hassoun. The ngi forward semantic social network data, 2020.

[9] A. Cottica and A. Hassoun. The poprebel semantic social network data, May 2021.

[10] A. Cottica, A. Hassoun, M. Manca, J. Vallet, and G. Melançon. Semantic social networks: A mixed methods approach to digital ethnography. *Field Methods*, 32(3):274–290, 2020.

[11] A. Cottica and G. Melançon. The opencare semantic social network data, 2016.

[12] A. Decuyper, A. Browet, V. Traag, V. D. Blondel, and J.-C. Delvenne. Clean up or mess up: the effect of sampling biases on measurements of degree distributions in mobile phone datasets. *arXiv preprint arXiv:1609.09413*, 2016.

[13] W. W. Dressler, C. D. Borges, M. C. Balierio, and J. E. dos Santos. Measuring cultural consonance: Examples with special reference to measurement theory in anthropology. *Field Methods*, 17(4):331–355, 2005.

[14] L. P. Freedman, I. M. Cockburn, and T. S. Simcoe. The economics of reproducibility in preclinical research. *PLoS biology*, 13(6):e1002165, 2015.

[15] 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.

[16] C. Giatsidis, D. M. Thilikos, and M. Vazirgiannis. Evaluating cooperation in communities with the k-core structure. In *2011 International conference on advances in social networks analysis and mining*, pages 87–93. IEEE, 2011.

[17] U. Hannerz. The global ecumene as a network of networks. In A. Kuper, editor, *Conceptualizing society*, pages 34–56. Routledge, 1992.

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

[19] D. D. Laitin, J. T. Watkins IV, et al. *Hegemony and culture: Politics and change among the Yoruba.* University of Chicago Press, 1986.

[20] S. Leonelli. What distinguishes data from models? *European Journal for Philosophy of Science*, 9(2):22, 2019.

[21] C. Lévi-Strauss et al. *La pensée sauvage*, volume 289. Plon Paris, 1962.

[22] C. Lévi-Strauss and C. Lévi-Strauss. *Anthropologie structurale*, volume 171. Plon Paris, 1958.

[23] S. E. Maxwell, M. Y. Lau, and G. S. Howard. Is psychology suffering from a replication crisis? what does "failure to replicate" really mean? *American Psychologist*, 70(6):487, 2015.

[24] G. Melancon. Just how dense are dense graphs in the real world? a methodological note. In *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*, pages 1–7, 2006.

[25] T. Munzner. *Visualization analysis and design.* CRC press, 2014.

[26] B. Nick, C. Lee, P. Cunningham, and U. Brandes. Simmelian backbones: Amplifying hidden homophily in facebook networks. In *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*, pages 525–532, Aug 2013.

[27] B. Pittel, J. Spencer, and N. Wormald. Sudden emergence of a giantk-core in a random graph. *Journal of Combinatorial Theory, Series B*, 67(1):111–151, 1996.

[28] M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A. I. Aviles-Rivero, C. Etmann, C. McCague, L. Beer, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans. *Nature Machine Intelligence*, 3(3):199–217, 2021.

[29] S. C. Shapiro. Representing and locating deduction rules in a semantic network. *ACM SIGART Bulletin*, (63):14–18, 1977.

[30] D. A. Snow, S. A. Soule, and H. Kriesi. *The Blackwell companion to social movements*. John Wiley & Sons, 2008.

[31] U. Soni, Y. Lu, B. Hansen, H. C. Purchase, S. Kobourov, and R. Maciejewski. The perception of graph properties in graph layouts. In *Computer Graphics Forum*, volume 37, pages 169–181. Wiley Online Library, 2018.

[32] J. F. Sowa. *Conceptual structures: information processing in mind and machine*. Addison-Wesley Pub., Reading, MA, 1983.

[33] J. F. Sowa et al. *Knowledge representation: logical, philosophical, and computational foundations*, volume 13. Brooks/Cole Pacific Grove, CA, 2000.

[34] M. Strathern. Cutting the network. *The Journal of the Royal Anthropological Institute*, 2(3):517–535, 1996.

[35] V. Turner. Liminal to liminoid, in play, flow, and ritual: An essay in comparative symbology. *Rice Institute Pamphlet-Rice University Studies*, 60(3), 1974.

[36] W. A. Woods. What's in a link: Foundations for semantic networks. In *Representation and understanding: Studies in Cognitive Science*, pages 35–82. Elsevier, 1975.

[37] R. Wuthnow. *Communities of discourse: Ideology and social structure in the Reformation, the Enlightenment, and European socialism*. Harvard University Press, 2009.