FUNDAMENTAL PRINCIPLES OF DATA SCIENCE MASTER'S THESIS

A study of polarisaton in bimodal, online social networks

Author:
Adrián FERNÁNDEZ CID

Supervisors:
Dr. Emanuele COZZO
Dr. Oriol PUJOL VILA

A thesis submitted in partial fulfillment of the requirements
for the degree of MSc in Fundamental Principles of Data Science

in the

Facultat de Matemàtiques i Informàtica

January 7, 2022

UNIVERSITAT DE BARCELONA

Abstract

Facultat de Matemàtiques i Informàtica

MSc

A study of polarisaton in bimodal, online social networks

by Adrián FERNÁNDEZ CID

Social polarisation is a central issue in social science, and it has acquired mainstream interest in recent years. Current methods in computational social science go beyond traditional distributional analysis by leveraging the tools of network science and the availability of digital data, the typical procedure being 1) constructing a graph representation (consisting of nodes and links) of the social system; 2) determining a partition of such a graph in terms of groups that feature inter-group dissimilarity and intra-group identification; and 3) evaluating the polarisation of the partitioned graph. Such structural polarisation measures present the advantage that they can be applied more or less automatically, although some of them have been shown to present undesirable features. In this master's thesis, I explore a hybrid approach that implements an axiomatic, theoretically-sound distributional polarisation measure (Esteban and Ray, 1994) with hierarchical clustering. This novel pipeline is applied here to bimodal networks, which are less studied in the literature. In the validation use case, on the standard Southern Women dataset (Davis, Gardner, and Gardner, 1941), results reasonably agree with the expected separation in two communities (Homans, 1950; Breiger, 1974) for the Ward and centroid distance update schemes of the clustering. On the other hand, the application use case, on data from the Conference on the Future of Europe, shows no significant dipoles neither in the topic-specific nor the global analyses, which (given the previous pipeline validation) points to a lack of polarisation in the platform of the Conference. Further analysis on such data in terms of higher-order multiples is underway and may yet reveal some structure. Current results show the proposed pipeline remains a promising candidate for the study of polarisation in bimodal social networks and should be further explored.

Acknowledgements

The acknowledgments and the people to thank go here, don't forget to include your project advisor...

Chapter 1

Introduction

1.1 Motivation

The phenomenon of polarisation is believed to lie at the root of social conflict/misinformation, whose influence on societal matters like elections or public policies regarding issues like climate change or the present pandemic has attracted increasing academic and social attention. PolThy paper

Plus it's a hot topic in computational social science... (refs Emanuele)

1.2 A network theory approach

One approach to the study of polarisation is that offered by complex network analysis, the typical procedure being 1) constructing a graph representation (consisting of nodes and links) of the (social) system; 2) determining a partition of such a graph in terms of groups, or clusters, that ideally (usually) feature intra-cluster similarity as well as inter-cluster dissimilarity under some criterion; and 3) evaluating the polarisation of the partitioned graph by means of a so-called polarisation measure.

The graph representation of the system is determined by the definition of nodes and links: for instance, in a set of Twitter users one may take individual users as nodes and retweets as links, or take subsets of similar users as nodes, or take "follows" or "likes" as links. Furthermore, both nodes and links may be filtered by a measure of significance, or comprise more than one class (e.g. one can consider popular users as a special kind of node, yielding what is known as a bimodal graph: a graph with two different types of nodes), and links may be binary (either there is or there is not a link between two given nodes) or weighted. As one might expect, the finer the representation, the more complex (and difficult to study) the graph becomes.

The clustering step, so-called community-detection problem in the field of complex networks, is a current subject of active research. The optimal partition crucially depends on what one defines as a community, giving existing algorithms based on modularity maximisation, hierarchical clustering, stochastic approaches and others. (com-detect refs here, ill-posed problem etc)

Finally, the polarisation measure offers some freedom as well: it may depend on the topology of the... (ref mikko)

1.3 This master's thesis

In this project, we decided to study bimodal social networks, namely the Southern Women dataset and several networks extracted from the Conference on the Future

of Europe data[1].

The first, intended for a validation phase for the clustering and (to a lesser extent) the polarisation measure, is a well known dataset used as a standard for community detection in bimodal networks: it consists of two classes of nodes representing women and social events, and the links encode which women attended which event. The second includes 5 previously unstudied datasets made for the purpose of this thesis, where the nodes represent either policy proposals (by any European citizen or group) or users of the online platform, a link meaning that a given user endorsed a given proposal.

The partitioning of the graph is provided by the last stage of a so-called sequential, agglomerative, hierarchic, nonoverlapping (SAHN) clustering algorithm, which gradually builds up clusters from the initial data by merging the two closest (according to a dissimilarity measure) clusters at a time (the last stage being that of only two clusters remaining). Such algorithms allow for the exploration of different dissimilarity measures, of which we consider three: the well-knwon Ward and centroid distances, and the novel measure we termed *poldist*, based on the measure used for evaluating polarisation (see below and Section...).

Finally, the polarisation measure stems from the axiomatic proposal of [1], which presents the advantage of ...

The rest of the thesis is organised as follows:...

---

[1]Originally, the object of study was envisaged as a bimodal network representation of a Twitter community consisting of crowd-sourced elite users (also called influencers in regular language) and their audiences (regular users). However, the identification of such crowd-sourced elites had already been done for a previous project, and we ultimately deemed it more appropriate to choose for the object of a master's thesis data that I would have to study from 0.

Chapter 2

Background

2.1 An axiomatic polarisation measure

Traditional approaches for measuring polarisation rely on the study of distributional properties (such as bimodality or dispersion) of survey data. More recently, however, the development of network science along with the availability of digital social data have motivated methods based on the structure of the graph representation of the social system.

Such structural polarisation scores include ... The pipeline typically consists of building the graph representation, obtaining a suitable partition thereof and evaluating polarisation. These methods bear the advantage of providing an automatic means of polarisation evaluation for any system easily representable as a network, and they may in principle leverage the tools of network science to account for complex, yet relevant features in the study of polarisation.

However, structural polarisation measures also have their shortcomings. First, as already mentioned, the crucial step of finding an optimal graph partition is ill-defined, and many community-detection methods are built to invariable find a partition, whether one such partition is really present in the system or not (Fortunato and Hric, 2016). This may hinder the subsequent polarisation measurement.

Second, (Salloum, Chen, and Kivelä, 2021) showed that 8 state-of-the-art structural measures depend on presumably unrelated network properties like network size, average degree or degree distribution, which also makes it difficult to compare different networks in terms of polarisation. The authors also found that such measures failed to assign null polarisation to synthetic, random networks, and pointed out that the fact that most current structural measures require a bimodal partition may constitute a limitation.

Although the authors in (Salloum, Chen, and Kivelä, 2021) do improve the performance of the measures they study by introducing a normalisation to extract the contribution of random features, they finish by calling to the development of novel approaches, possibly fundamentally different from the ones they study. Such is precisely the objective of this master's thesis.

In (Esteban and Ray, 1994), the authors present the following polarisation measure:

$$P = K \sum_{i=1}^{N_c} \sum_{j=1}^{N_c} n_i^{1+\alpha} n_j d(i,j) \tag{2.1}$$

where K is a normalisation constant, $\alpha \in (0, 1.6]$ is the *polarisastion sensitivity* parameter, n_i is the size of cluster i, N_c the total number of clusters and $d(i,j)$ is the distance between i and j. The role of α is to enforce identification within a given group, while distance to other groups d accounts for inter-group alienation.
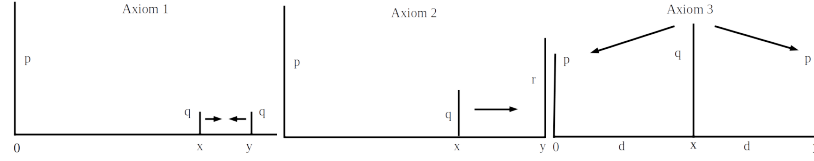
FIGURE 2.1: Illustrations of each of the three axioms imposed on the
polarisation measure used [].

Such a measure is defined on a one-dimensional distribution of a given attribute (income, opinion, etc.), and it is constructed by imposing the following three intuitive axioms, illustrated in Fig. 2.1:

- *Axiom 1.* Take $p, q \gg 0$, $p > q$, $0 < x < y$. There exists $\epsilon > 0$ and $\mu > 0$ (possibly depending on p and x) such that if $\delta(x, y) < \epsilon$ and $q < \mu p$, then the joining of the two clusters at their mid-point, $(x + y)/2$, increases polarisation.

- *Axiom 2.* Take $(p, q, r) \gg 0$, $p > r$ and $x > d(x, y)$. There is $\epsilon > 0$ such that if the cluster q is moved towards r by an amount not exceeding ϵ, polarisation increases.

- *Axiom 3.* Take $p, q \gg 0$ and $x = d(x, y) \equiv d$. Any new distribution formed by shifting population mas from the central cluster q equally to the two lateral ones, each d units of distance away, increases polarisation.

The authors also require that the sorting produced by the polarisation measure over two distributions be independent of population size, i.e.:

- *Condition H*: If $P(\pi, y) \geq P(\pi', y')$ for two distributions (π, y) and (π', y'), then $P(\lambda \pi, y) \geq P(\lambda \pi', y') \ \forall \lambda > 0$.

Theorem 1 in (Esteban and Ray, 1994) establishes that the expression in (2.1) satisfies all three of the above axioms and Condition H. The authors of (2.1) also show that the maximum polarisation is reached for a perfectly balanced bimodal, maximally separated configuration (i.e. half the points of a given population on either end of the distribution). Thus, if distance is normalised to 1, we have $P_{max} = 2(N_c/2)^{\alpha+2}$. This makes $K = 1/P_{max} = (2/N_c)^{\alpha+2}/2$ a reasonable choice for the normalisation constant in (2.1), and such is the value we use. We therefore have $P \in [0, 1]$.

There are a number of advantages to (2.1):

1. It is well founded: its axioms guarantee certain intuitively desirable properties of a polarisation measure.

2. Although it is a distributional and not a structural measure, it can also be implemented automatically in a network science pipeline provided a notion of distance.

3. Normalisation to the maximal polarisation enables straightforward and meaningful comparisons among different systems, independently of their size.

4. Although here we will focus on dipole polarisation, the formula allows for an arbitrary number of clusters, contrary to most structural polarisation measures (Salloum, Chen, and Kivelä, 2021).

2.2 Graph partitioning

As I mentioned before, community detection is an ill-posed problem, and there is no all-purpose solution. In (ref) they mention ...

In the present case, however, choice is facilitated by the constraints of the pipeline: namely, our polarisation measure (2.1) requires distances between clusters. A class of methods that implement cluster distances quite naturally is that of sequential, agglomerative, hierarchic, nonoverlapping (SAHN) clustering algorithms, and we use such methods here.

2.2.1 SAHN clustering algorithms

A SAHN algorithm (Müllner, 2011) gradually builds up clusters from the initial unit cluster data by merging the two closest clusters at a time (the last stage being that of only two clusters remaining). The procedure has therefore $N - 1$ steps, each defining a different partition, for N initial observations.

The merging order is obtained by minimising a pairwise dissimilarity measure between the clusters that is updated according to a given scheme (e.g. any of those in Fig. 2) whenever a new cluster is formed. More concretely:

Definition. A *dissimilarity measure* on a set S is a map $d : S \times S \rightarrow [0, \infty)$ which is reflexive and symmetric, i.e. we have $d(x, x) = 0$ and $d(x, y) = d(y, x)$ for all $x, y \in S$.

In general, these algorithms take as input either the pairwise dissmilarities between the initial observations (*stored matrix approach*) or the set of observations S (*stored data approach*): in the present work, I follow the stored matrix approach.

Regarding the output, a standard is a data structure that has been called *stepwise dendogram* in (Müllner, 2011):

Definition. Given a finite set S_0 with cardinality $N = |S_0|$, a *stepwise dendrogram* is a list of $N - 1$ triples (a_i, b_i, δ_i) $(i = 0, \dots, N - 2)$ such that $\delta_i \in [0, \infty)$ and $a_i, b_i \in S_i$, where S_{i+1} is recursively defined as $(S_i \setminus \{a_i, b_i\}) \cup n_i$ and $n_i \notin S \setminus \{a_i, b_i\}$ is a label for a new node.

The procedural definition of the above class of SAHN algorithms is presented in Fig. 1.

2.2.2 Distance update schemes

SAHN algorithms allow for the exploration of different dissimilarity measures as clustering criteria, of which we consider three: the well-knwon Ward and centroid distances, and the novel measure we termed *poldist*, based on the measure used for evaluating polarisation (2.1). Figure 2 shows the iterative formulas of each of these three distance update schemes, as well as their closed-form definitions.

Note that both d_W and d_p are just centroid distance weighted by a function of cluster sizes, which is ≥ 1 for d_W and ≥ 2 for d_p. For Ward, it makes clustering favour merging two unit clusters (for which $d_W = d_c$) and postpone merging two large ones. For poldist, the behaviour is qualitatively the same, although more pronounced, since we have at least (depending on α) a quadratic order on cluster sizes.

Both Ward and poldist will therefore tend to make the resulting tree more "balanced" and stretched towards the root, with poldist's effect being greater.

Another feature of the *poldist* formula is that it coincides with global polarisation when there are only two clusters (e.g. at the last step of the clustering).

Figure 1 Algorithmic definition of a hierarchical clustering scheme. Taken from (Müllner, 2011).

1: **procedure** PRIMITIVE_CLUSTERING(S, d)   ▷ S: node labels, d: pairwise dissimilarities
2:      $N \leftarrow |S|$ ▷ Number of input nodes
3:      $L \leftarrow [\,]$ ▷ Output list
4:      $size[x] \leftarrow 1$ for all $x \in S$
5:      **for** $i \leftarrow 0, \ldots, N - 2$ **do**
6:          $(a, b) \leftarrow \text{argmin}_{(S \times S) \setminus \Delta}\, d$
7:          Append $(a, b, d[a, b])$ to L.
8:          $S \leftarrow S \setminus \{a, b\}$
9:          Create a new node label $n \notin S$.
10:         Update d with the information

$$d[n, x] = d[x, n] = \text{FORMULA}(d[a, x], d[b, x], d[a, b], size[a], size[b], size[x])$$

            for all $x \in S$.
11:         $size[n] \leftarrow size[a] + size[b]$
12:         $S \leftarrow S \cup \{n\}$
13:     **end for**
14:     **return** L ▷ the stepwise dendrogram, an $((N - 1) \times 3)$-matrix
15: **end procedure**

(As usual, Δ denotes the diagonal in the Cartesian product $S \times S$.)

Figure 2 Agglomerative clustering schemes.

Name	Distance update formula FORMULA for $d(I \cup J, K)$	Cluster dissimilarity between clusters A and B
Ward (d_W)	$\sqrt{\dfrac{(n_I + n_K)d(I, K)^2 + (n_J + n_K)d(J, K)^2 - n_K d(I, J)^2}{n_I + n_J + n_K}}$	$\sqrt{\dfrac{2|A||B|}{|A| + |B|}} \cdot \|\vec{c}_A - \vec{c}_B\|_2$
centroid (d_c)	$\sqrt{\dfrac{n_I d(I, K)^2 + n_J d(J, K)^2}{n_I + n_J} - \dfrac{n_I n_J d(I, J)^2}{(n_I + n_J)^2}}$	$\|\vec{c}_A - \vec{c}_B\|_2$
poldist (d_p)	$K[(n_I + n_J)^{\alpha+1}n_K + n_K^{\alpha+1}(n_I + n_J)]d_c(I \cup J, K)$	$K(|A|^{\alpha+1}|B| + |B|^{\alpha+1}|A|) \cdot \|\vec{c}_A - \vec{c}_B\|_2$

Legend: Let I, J be two clusters joined into a new cluster, and let K be any other cluster. Denote by n_I, n_J and n_K the sizes of (i.e. number of elements in) clusters I, J, K, respectively.

The update formulas for the "Ward" and "centroid" methods assume Euclidean distance as dissimilarity measure (which is the one we use for all three methods). The expression $\vec{c}_X$ denotes the centroid of a cluster X. The factor K in front of the poldist formulas is the same constant used in for polarisation in (2.1).

All three measures defined above are dissimilarity measures in the sense that they (trivially) satisfy for all i, j the conditions of positiveness $d(i, j) \geq 0$, reflexiveness $d(i, i) = 0$ and symmetry $d(i, j) = d(j, i)$. I will also refer to them as distances, although they are not proper distances insofar as the distance of different elements may be 0 (e.g. for two clusters that have the same centroid), and even if d_c does satisfy the Triangle Inequality ($d(i, j) \leq d(i, k) + d(k, j)\, \forall i, j, k$), d_p and d_W do not (e.g. take $n_i = n_j \gg 1$, $n_k = 1$ as a counterexample).

2.2.3 Implementation

In (Müllner, 2011), the author studies different SAHN algorithms and compares their performance for several distance update schemes. In particular, he recommends the *nearest-neighbour chain* (*nn_chain*) algorithm for Ward clustering (having a worst case time complexity of $O(n^2)$) and the *generic clustering* (*generic_clustering*) algorithm for the centroid method (of $O(n^3)$ worst case time complexity, but usually closer to $O(n^2)$ in practice). I follow such recommendations in this work.

The reason for using a less efficient algorithm with centroid distance is that the particular optimisation strategy of the *nn_chain* algorithm relies on a post-processing step that sorts the stepwise dendogram by increasing distances. This approach does not produce a valid solution for the centroid update scheme, since it does not necessarily produce a monotonically increasing sequence of distances (in other words, the distance to a newly merged cluster may be smaller that to any of its components).

The *generic_clustering* algorithm, on the other hand, can be applied to any clustering distance, which is the reason why I also apply it with the poldist method.

Both algorithms are implemented by the Python library Scipy (Virtanen et al., 2020), with the particularity of producing a slightly modified version of a stepwise dendogram they call a *linkage matrix*: instead of having as rows only (a_i, b_i, δ_i) triplets with the merged clusters and their distance at each step, the output matrix features tetrads $(a_i, b_i, \delta_i, n_i)$, where n_i is the size of the newly formed cluster.

The implementations used in this work are an adaptation of Scipy's[1], with the main modification being that polarisation (2.1) is computed at every step of the clustering. Such a computation adds an $O(n^3)$ worst-case time complexity term to both the *nn_chain* and *generic_clustering* algorithms, so their upper bounds are here more similar than in their pure versions (Müllner, 2011).

Fig. 3 shows the adapted implementation of the *generic_clustering* algorithm, while that of the *nn_chain* algorithm is featured in Figs. 4, 5.

Finally, implementation of the polarisation formula (2.1) and of poldist (see Fig. 2) uses d_c (centroid distance) as distance. There are two reasons for this: i) it comes as a natural candidate; ii) other possibilities may depend on cluster size, such as Ward distance, which in general will complicate the normalisation of P because the maximum distance may change at every step due to increase from greater sizes.

2.2.4 Input

As mentioned previously, the SAHN implementations used here take as input a list of the pairwise distances between all points.

Given that application of the polarisation measure (2.1) in principle assumes one kind of entity, the natural approach to obtain such distances is to consider only one type of node of the bimodal network for clustering. A way to do that while keeping some information of its bimodal character is by means of the ϕ coefficient (Pearson, 1900) (known as the Matthews correlation coefficient in machine learning), which can later be turned into a Euclidean distance. The ϕ correlation is defined as

$$\phi(v_1, v_2) = \frac{f_{11}f_{00} - f_{10}f_{01}}{\sqrt{f_{1*}f_{0*}f_{*1}f_{*0}}} = \frac{f_{11}F - f_{1*}f_{*1}}{\sqrt{f_{1*}f_{0*}f_{*1}f_{*0}}} \tag{2.2}$$

for any pair of observation binary vectors v_1, v_2, where F is the total number of observations, f_{11} is the number of coincidental positive (1) observations, f_{10} that of

---

[1]The adaptations are written in the Python language, both for ease of implementation and for readibility.

Figure 3 The generic clustering algorithm.

 1: **procedure** GENERIC_LINKAGE(N, d)   ▷ N: input size, d: pairwise dissimilarities
 2:     $S \leftarrow (0, \ldots, N - 1)$
 3:     $L \leftarrow [\,]$ ▷ Output list
 4:     $size[x] \leftarrow 1$ for all $x \in S$
 5:     **for** x in $S \setminus \{N - 1\}$ **do**   ▷ Generate the list of nearest neighbors.
 6:         $n\_nghbr[x] \leftarrow \text{argmin}_{y>x}\, d[x, y]$
 7:         $mindist[x] \leftarrow d[x, n\_nghbr[x]]$
 8:     **end for**
 9:     $Q \leftarrow$ (priority queue of indices in $S \setminus \{N - 1\}$, keys are in *mindist*)
10:     **for** $i \leftarrow 1, \ldots, N - 1$ **do**   ▷ Main loop.
11:         $a \leftarrow$ (minimal element of Q)
12:         $b \leftarrow n\_nghbr[a]$
13:         $\delta \leftarrow mindist[a]$
14:         **while** $\delta \neq d[a, b]$ **do**   ▷ Recalculation of nearest neighbors, if necessary.
15:             $n\_nghbr[a] \leftarrow \text{argmin}_{x>a}\, d[a, x]$
16:             Update *mindist* and Q with $(a, d[a, n\_nghbr[a]])$
17:             $a \leftarrow$ (minimal element of Q)
18:             $b \leftarrow n\_nghbr[a]$
19:             $\delta \leftarrow mindist[a]$
20:         **end while**
21:         Remove the minimal element a from Q.
22:         Append (a, b, δ) to L.   ▷ Merge the pairs of nearest nodes.
23:         $size[b] \leftarrow size[a] + size[b]$ ▷ Re-use b as the index for the new node.
24:         $S \leftarrow S \setminus \{a\}$
25:         **for** x in $S \setminus \{b\}$ **do**   ▷ Update the distance matrix.
26:             $d[x, b] \leftarrow d[b, x] \leftarrow$ FORMULA$(d[a, x], d[b, x], d[a, b], size[a], size[b], size[x])$
27:         **end for**
28:         **for** x in S such that $x < a$ **do**   ▷ Update *candidates* for nearest neighbors.
29:             **if** $n\_nghbr[x] = a$ **then**   ▷ Deferred search; no nearest
30:                 $n\_nghbr[x] \leftarrow b$ ▷ neighbors are searched here.
31:             **end if**
32:         **end for**
33:         **for** x in S such that $x < b$ **do**
34:             **if** $d[x, b] < mindist[x]$ **then**
35:                 $n\_nghbr[x] \leftarrow b$
36:                 Update *mindist* and Q with $(x, d[x, b])$ ▷ Preserve a lower bound.
37:             **end if**
38:         **end for**
39:         $n\_nghbr[b] \leftarrow \text{argmin}_{x>b}\, d[b, x]$
40:         Update *mindist* and Q with $(b, d[b, n\_nghbr[b]])$
41:     **end for**
42:     **return** L ▷ The stepwise dendrogram, an $((N - 1) \times 3)$-matrix.
43: **end procedure**

Figure 4 The nearest-neighbor clustering algorithm.

1: **procedure** NN-CHAIN-LINKAGE(S, d)   ▷ S: node labels, d: pairwise dissimilarities
2:     $L \leftarrow$ NN-CHAIN-CORE(N, d)
3:     Stably sort L with respect to the third column.
4:     $L \leftarrow$ LABEL(L)   ▷ Find node labels from cluster representatives.
5:     **return** L
6: **end procedure**

1: **procedure** NN-CHAIN-CORE(S, d)   ▷ S: node labels, d: pairwise dissimilarities
2:     $S \leftarrow (0, \ldots, N-1)$
3:     $chain = [\,]$
4:     $size[x] \leftarrow 1$ for all $x \in S$
5:     **while** $|S| > 1$ **do**
6:         **if** length($chain$) ≤ 3 **then**
7:             $a \leftarrow$ (any element of S)   ▷ E.g. $S[0]$
8:             $chain \leftarrow [a]$
9:             $b \leftarrow$ (any element of $S \setminus \{a\}$)   ▷ E.g. $S[1]$
10:         **else**
11:             $a \leftarrow chain[-4]$
12:             $b \leftarrow chain[-3]$
13:             Remove $chain[-1], chain[-2]$ and $chain[-3]$ ▷ Cut the tail (x, y, x).
14:         **end if**
15:         **repeat**
16:             $c \leftarrow \mathrm{argmin}_{x \neq a} d[x, a]$ with preference for b
17:             $a, b \leftarrow c, a$
18:             Append a to $chain$
19:         **until** length($chain$) ≥ 3 and $a = chain[-3]$ ▷ a, b are reciprocal
20:         Append $(a, b, d[a, b])$ to L ▷ nearest neighbors.
21:         Remove a, b from S
22:         $n \leftarrow$ (new node label)
23:         $size[n] \leftarrow size[a] + size[b]$
24:         Update d with the information

$$d[n, x] = d[x, n] = \text{FORMULA}(d[a, x], d[b, x], d[a, b], size[a], size[b], size[x])$$

        for all $x \in S$.
25:         $S \leftarrow S \cup \{n\}$
26:     **end while**
27:     **return** L ▷ an unsorted dendrogram
28: **end procedure**

(We use the Python index notation: $chain[-2]$ is the second-to-last element in the list *chain*.)

Figure 5 A union-find data structure suited for the output conversion.

 1: **procedure** LABEL(L)
 2:      $L' \leftarrow [\,]$
 3:      $N \leftarrow$ (number of rows in L) $+ 1$ ▷ Number of initial nodes.
 4:      $U \leftarrow$ new UNION-FIND(N)
 5:      **for** (a, b, δ) in L **do**
 6:          Append $(U.\text{EFFICIENT-FIND}(a), U.\text{EFFICIENT-FIND}(b), \delta)$ to L'
 7:          $U.\text{UNION}(a, b)$
 8:      **end for**
 9:      **return** L'
10: **end procedure**

11: **class** UNION-FIND
12:      **method** CONSTRUCTOR(N)   ▷ N is the number of data points.
13:          *parent* $\leftarrow$ new int$[2N - 1]$
14:          *parent*$[0, \dots, 2N - 2] \leftarrow$ None
15:          *nextlabel* $\leftarrow N$ ▷ SciPy convention: new labels start at N
16:      **end method**

17:      **method** UNION(m, n)
18:          *parent*$[m] = $ *nextlabel*
19:          *parent*$[n] = $ *nextlabel*
20:          *nextlabel* $\leftarrow$ *nextlabel* $+ 1$ ▷ SciPy convention: number new labels
     consecutively
21:      **end method**

22:      **method** FIND(n)   ▷ This works but the search process is not efficient.
23:          **while** *parent*$[n]$ is not None **do**
24:              $n \leftarrow$ *parent*$[n]$
25:          **end while**
26:          **return** n
27:      **end method**

28:      **method** EFFICIENT-FIND(n)   ▷ This speeds up repeated calls.
29:          $p \leftarrow n$
30:          **while** *parent*$[n]$ is not None **do**
31:              $n \leftarrow$ *parent*$[n]$
32:          **end while**
33:          **while** *parent*$[p] \neq n$ **do**
34:              $p,$ *parent*$[p] \leftarrow$ *parent*$[p], n$
35:          **end while**
36:          **return** n
37:      **end method**
38: **end class**

---

positive observations for v_1 that are also negative (0) for v_2, f_{1*} the total of positive observations in v_1 regardless of v_2, and so on.

Note that $\phi \in [-1, 1]$, the minimum being attained when $f_{11} = f_{00} = 0$ and the maximum when $f_{10} = f_{01} = 0$ (this is apparent in the first definition).

The observation vectors, in this case, correspond to one kind of nodes, the primary node of the analysis, each vector encoding the connectivity value (yes: 1, no: 0) to each node of the secondary kind. F is therefore the number of secondary nodes.

As can be seen above, the ϕ coefficient is a way to have some weighting of bimodal connectivity. The required distances can be computed from ϕ as:

$$d = \sqrt{2(1 - \phi)} \tag{2.3}$$

Note that $d \in [0, 2]$: in the present application, I normalise input distances to 1.
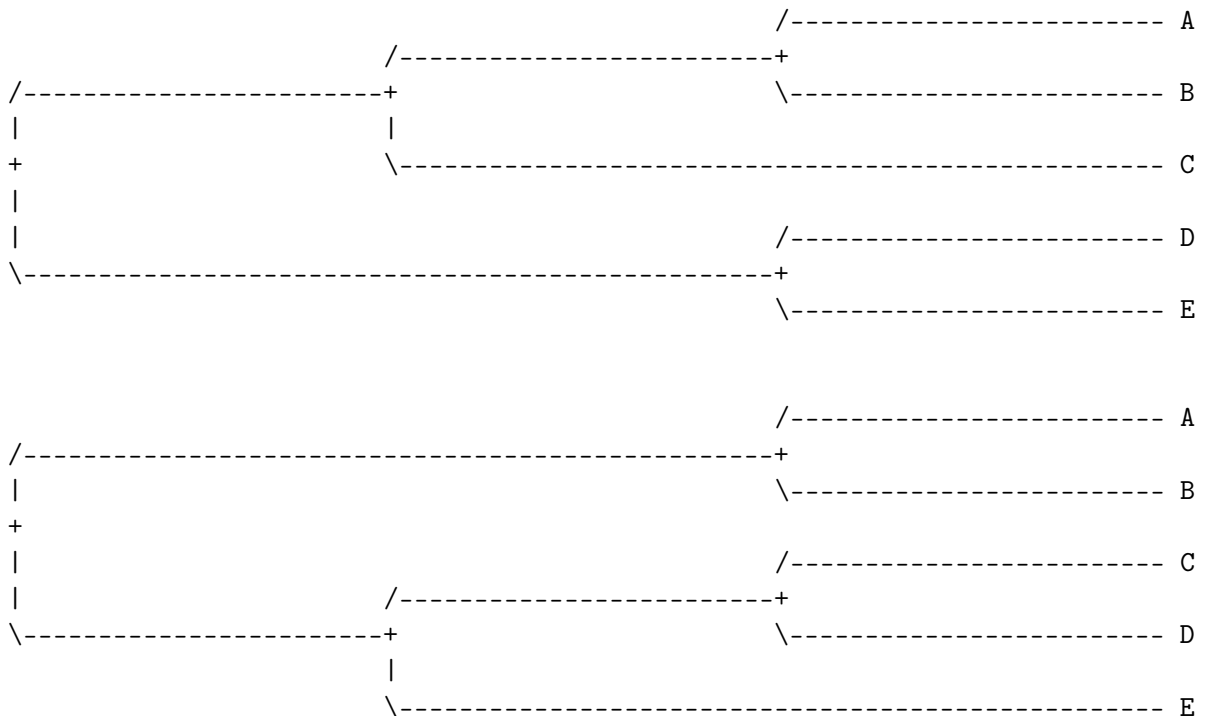
2.3 Analysis tools

2.3.1 Robinson-Foulds distance

In order to compare the trees produced with different distance update schemes (including poldist with different values of α) I use the (unweighted) Robinson-Foulds distance (Robinson and Foulds, 1981), a measure that captures topological differences between two trees T_1 and T_2 as the sum of the number of junctions that appear in only one of them, i.e.

$$D_{RF} = A + B, \tag{2.4}$$

where A is the sum of junctions found in T_1 but not in T_2 and B is the sum of junctions present in t_2 but not in T_1.

As is apparent from (2.4), D_{RF} is a sort of edit distance for trees, and it has the advantage of being quite intuitive: one can quickly see that it takes 4 steps (removing or adding one junction at a time) to go from the first to the second tree (or vice versa) below.

```
                                              /----------------------- A
                   /------------------------+
/------------------------+                   \----------------------- B
|                        |
+                        \------------------------------------------- C
|
|                                            /----------------------- D
\--------------------------------------------------+
                                            \----------------------- E


                                              /----------------------- A
/----------------------------------------------------+
|                                            \----------------------- B
+
|                                            /----------------------- C
|                   /------------------------+
\------------------------+                   \----------------------- D
                   |
                   \------------------------------------------- E
```

2.3.2   Modified Jaccard distance

For the SW dataset there are two reference 2-cluster partitions (Homans, 1950; Breiger, 1974) to compare with the final partition given by the hierarchical clustering. Since the comparison involves sets, I make use of an adaptation of *Jaccard distance*.

Jaccard distance (Jaccard, 1912) is a measure of dissimilarity bewteen sets. For sets A and B it is defined as:

$$d_J(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

Note however that we do not want to compare two sets, but two *pairs of sets* (the final state of any clustering with either of the partitions obtained by Homans and Breiger). Since our clustering always contains all women and the two reference partitions do not, one can argue for using only the set of possible coincidental elements (instead that of all possible elements) in the denominator. Therefore, if we have the pairs (R_1, R_1) and (C_1, C_2), the first being the reference, one way to adapt d_J is:

$$d_J^*((R_1, R_2), (C_1, C_2)) = 1 - S_J^*((R_1, R_2), (C_1, C_2))$$

(2.5)

$$S_J^*((R_1, R_2), (C_1, C_2)) = \sigma^{-1} \left[\frac{\max(|R_1 \cap C_1| + |R_2 \cap C_2|, |R_1 \cap C_2| + |R_2 \cap C_1|)}{|R_1 \cup R_2|} - \mu \right]$$

(2.6)

and μ and σ are just to re-center and rescale S_J^* (that we may call the modified Jaccard similarity) so that it varies between 0 and 1. I will call d_J^* the *modified Jaccard distance*.

2.3.3   Word shifts

Since that of the Conference on the Future of Europe is current, unstudied data, we lack a reference for comparison. However, given that such data consist of text (the titles and bodies of the proposals), one can qualitatively evaluate the final state through such text. Note that this would not be possible, or at least not directly, if one did the analysis on endorsers.

There are a number of ways to go about such an evaluation. The most obvious (and cumbersome) is to read all the proposals (more than 10^3) and apply some predefined labelling criterion that allows to evaluate the final partition. But since one would rather have at least a quantitative component and reduce subjectivity as much as possible, and given that this is not a master's thesis on sociometrics but on data science, we may leverage the power of natural language processing methods. One of such methods is *word shifts*, as those implemented in the *shifterator* package (Gallagher et al., 2021).

Word shifts are a tool for pairwise comparisons between texts that captures which words contribute to their difference and how. The contributions of the most relevant words are then visualized through horizontal bar charts called *word shift graphs*.

Shifterator's main input is the bag-of-words (BOW) representation of each of the two text to be compared, in the form of a dictionary whose keys are word types and whose values are the frequence of the corresponding word in the given text.

The package provides several text comparison measures, which include relative frequency, Shannon entropy, Tsallis entropy, the Kullback-Leibler divergence, and the Jensen-Shannon divergence.

The approach for evaluation in chapter 5 will thus be: i) retrieve the total text of each cluster of a final partition; ii) build their respective BOW representations; iii) obtain their word shift graph; iv) analyse the graph.

A specific aspect of the BOW representation (steo ii)) is worth noting: it may be improved by ignoring words that are deemed irrelevant for the subsequent analysis, known as *stop words* (prepositions, determiners, etc.). However, whether a given word is irrelevant depends highly on the application, so these must be handled with care. Specifically, in the application in chapter 5 I use a predefined minimal list of stop words provided by the *nltk* package that I checked previously, to which I add some extra stop words found in the word shift plots. There are also some stop words candidates I found for which was not quite sure, so I do not use them for the result reported here.

Figure **??** shows an example of the word shift graphs produced by shifterator. The top of the plot shows the bars corresponding to the total JSD of each text relative to one another, while the below bar with the Σ shows the direction of their difference. By the title and Σ, we see that Cluster 1 has a JSD of about 3x that of cluster 2.

The subplot to the left is the cumulative contribution plot, which traces how the total JSD shift changes as we add more words according to their rank. The horizontal line shows the cutoff of the contributions of the top 50 words plotted versus those that are not, which means that in this case about 25% of the overall difference is explained by the top 50 words.

We can also see the relative size (in number of word occurrences, including repeated ones) of the texts belonging to each cluster at the right, showing that Cluster 1 is (excluding stop words) less than half the size of Cluster 2.
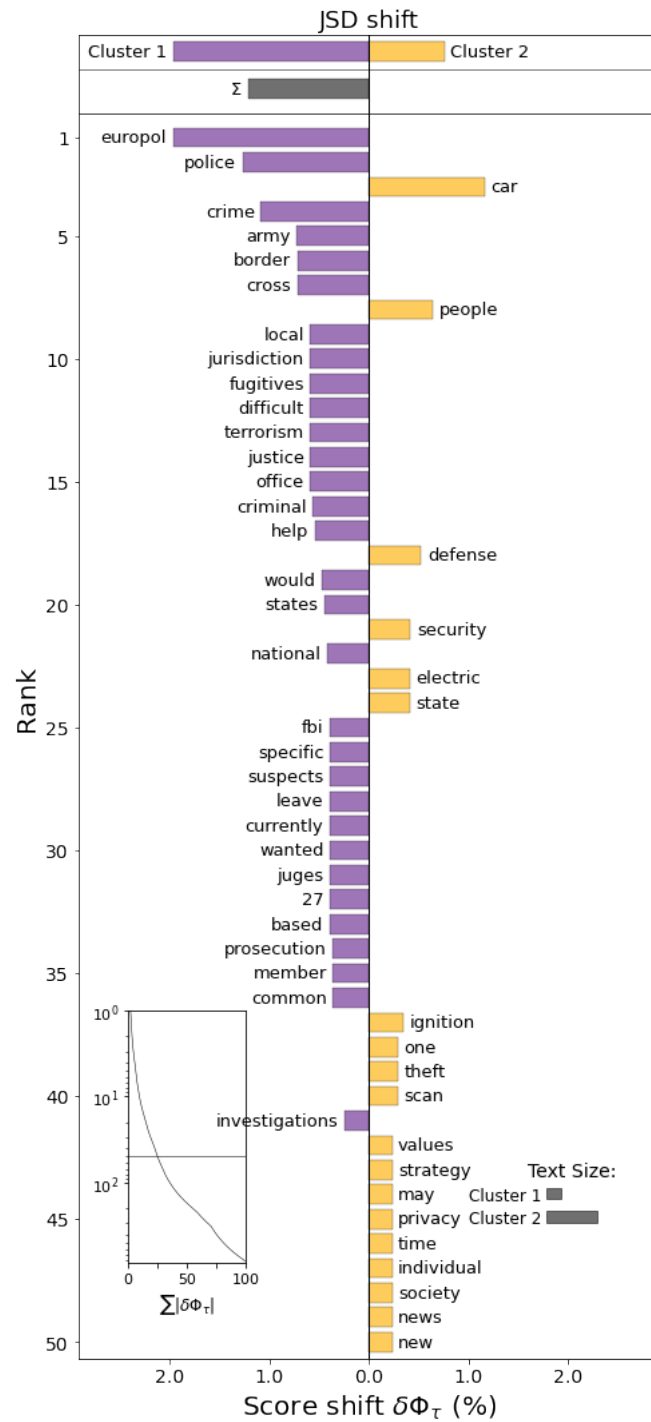
FIGURE 2.2: Example of a word shift graph.

Chapter 3

Methodology

3.1 Pipeline

Now that we have introduced the methods used, let us quickly review the pipeline. For a given system describable as a bimodal network:

1. Build the bimodal network representation.

2. Compute the pairwise ϕ coefficient among the nodes of interest. Then compute $d(\phi)$.

3. Apply hierarchical clustering with polarisation evaluation at every step. This and the rest of the steps are followed for the three distance update schemes considered (Ward, centroid and poldist).

4. Descriptive analysis: compare final polarisation to clustering heatmap of initial d; check evolution of P for $\alpha \in (0, 1.6]$; and compute D_{RF} between trees of different clusterings.

5. Evaluation. For the SW dataset, compare clustering results with references (Homans, 1950; Breiger, 1974). For the CFE data, qualitatively evaluate the clustering through the word shift of the texts of the two final clusters.

Chapter 4

Use Case 1: The Southern Women dataset

4.1 Data

The Southern Women (henceforth SW) dataset is a table containing binary values encoding the attendance (or absence thereof) of 18 women to a series of 14 social events (a card party, a club meeting, etc.) that took place throughout a year, originally compiled (Davis, Gardner, and Gardner, 1941) with the purpose of determining the influence of social status on the forming of communities. Although not apparent from the beginning, two groups of women became distinguishable in the reapraisals of Homans, 1950 and, in an alternative, more straigthforward way, of Breiger, 1974.

The SW has since become a standard dataset in computational social science, namely for testing clustering/community-detection methods in bimodal networks, and is easily available on the usual network repositories[1]. Figure 4.1 shows the original bimodal social network.

4.2 Experimental setting

The purpose of this use case is to validate to some extent the pipeline proposed here (see Chapter 3), also used in Chapter 5. Validation is made possible by previous studies (Homans, 1950; Breiger, 1974) that show there is a polarised structure and provide a reference for comparison.

The primary nodes for the analysis are the women. Although we could apply the same treatment event-wise, from a conceptual (social) perspective it is more interesting to do it for women (even if the two communities are there in either case).

Evaluation ???

4.3 Results

4.3.1 Clustering

Figure 4.2 shows the trees obtained with the Ward, centroid and poldist ($\alpha = 1$) distance update schemes, coupled with a heatmap of the input matrix distance to enhance visual interpretation. As is apparent, all three methods present a pronounced polarisation, coinciding (at least qualitatively) with previous results.

Note the inversion in one branch of the left cluster in the centroid tree: indeed, as mentioned previously, the centroid update scheme allows for such a behaviour, whereas Ward's does not (Müllner, 2011). As has also already been mentioned, that

---

[1]See, for instance, the CASOS repository: http://casos.cs.cmu.edu/computational_tools/datasets/external/davis/index2.htm
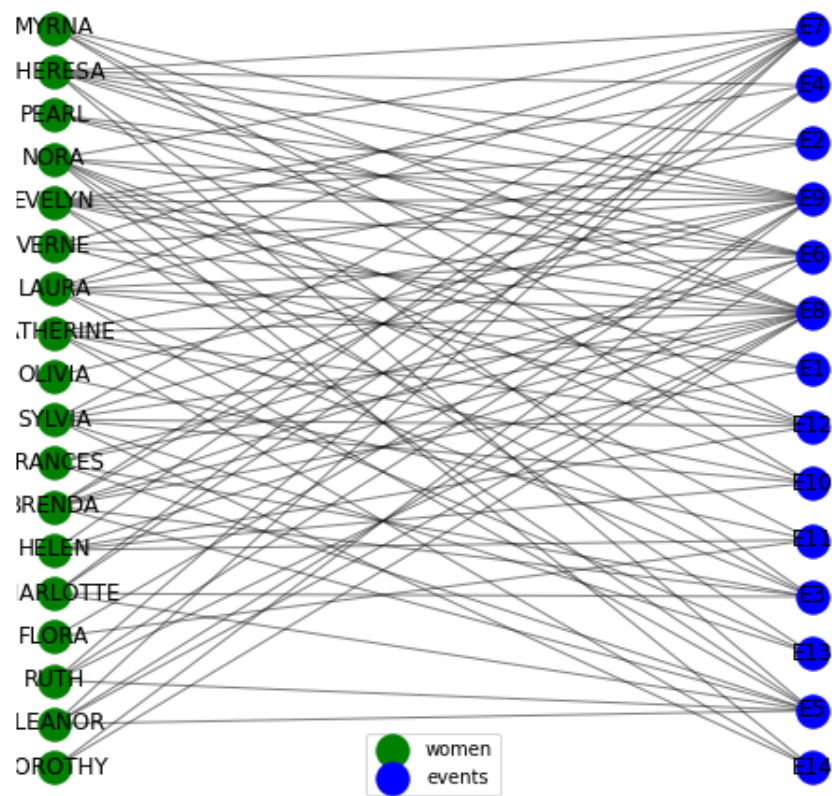
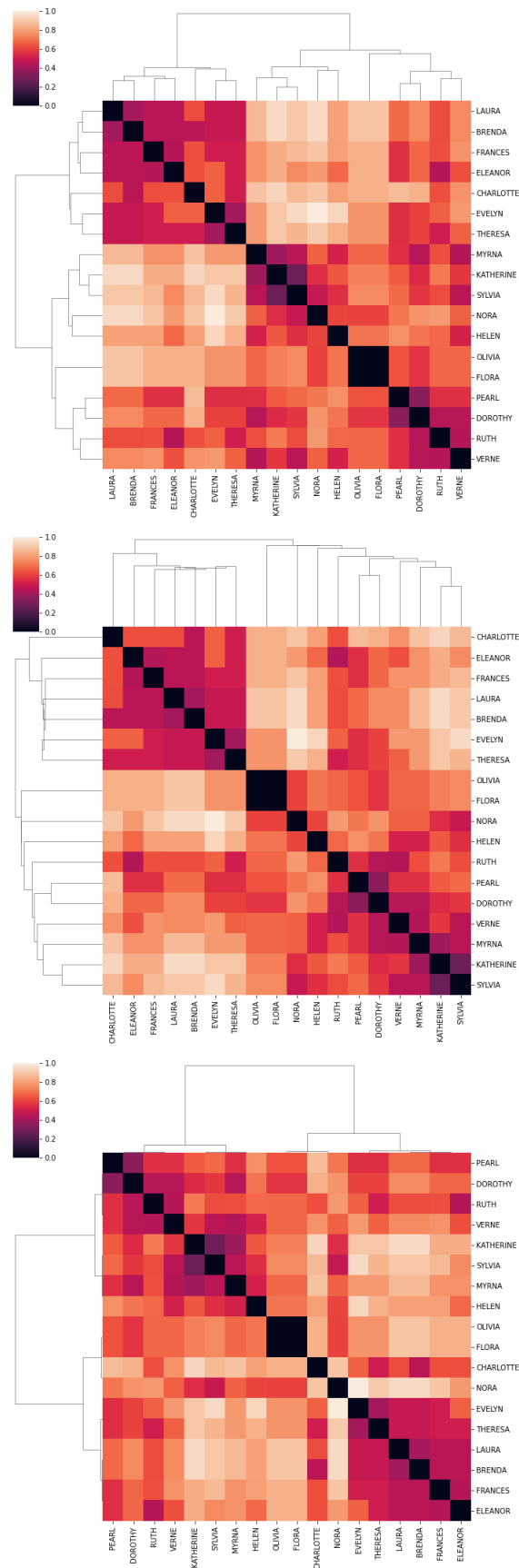FIGURE 4.1: Southern Women network.

FIGURE 4.2: Southern Women clustering for the three distance update schemes considered: Ward, centroid and *poldist*.
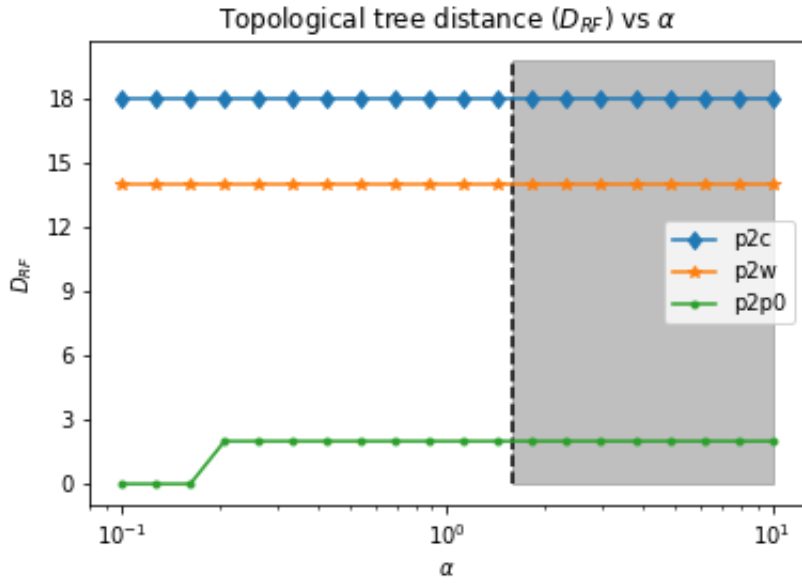
FIGURE 4.3: Robinson-Foulds (Robinson and Foulds, 1981) distance of Ward, centroid and poldist-0 ($\alpha = 0$) trees to the poldist tree for different α, on the Southern Women dataset.

is precisely why the *nn_chain* algorithm does not produce a valid solution for the centroid method.

Figure 4.3 displays the Robinson-Foulds distance of Ward, centroid and poldist-0 ($\alpha = 0$) trees to the poldist tree for different α. Looks like, aside from the change from $\alpha \approx 0.16$ to $\alpha \gtrsim 0.26$ (that yields an RF distance of 2 to the tree obtained with $\alpha = 0$), there seem to be no topological changes in the *poldist* tree for the different α allowed ($\alpha \in (0, 1.6]$, to the left of the vertical line), and not even up to $\alpha = 10$. The distances between branches do change (I checked), but the structure is the same.

Note that renormalising d_c in poldist would change nothing, because it's a proportionality constant: the only thing that could presumably make the allowed range of α more relevant for the structure of the poldist clustering is a greater network size (which would allow for the forming of greater clusters, thus increasing the contribution of α in the exponent).

4.3.2 Polarisation

Figure 4.4 shows the evolution of the polarisation through the clustering for the three methods considered and the allowed range of the polarisation sensitivity ($\alpha \in (0, 1.6]$).

In all three methods, we observe a convergence in the final state of the clustering for every α.

In addition we see that, overall, for low α P starts high and decreases with clustering step, while the behaviour is the opposite for higher values: α starts low and increases.

This is consistent with the description in the original paper (Esteban and Ray, 1994): the polarisation formula 2.1 reduces to the standard inequality formula for $\alpha \approx 0$, and inequality is generally greater in a system with only point clusters than when they start to merge; on the other hand, the contribution of a non-vanishing
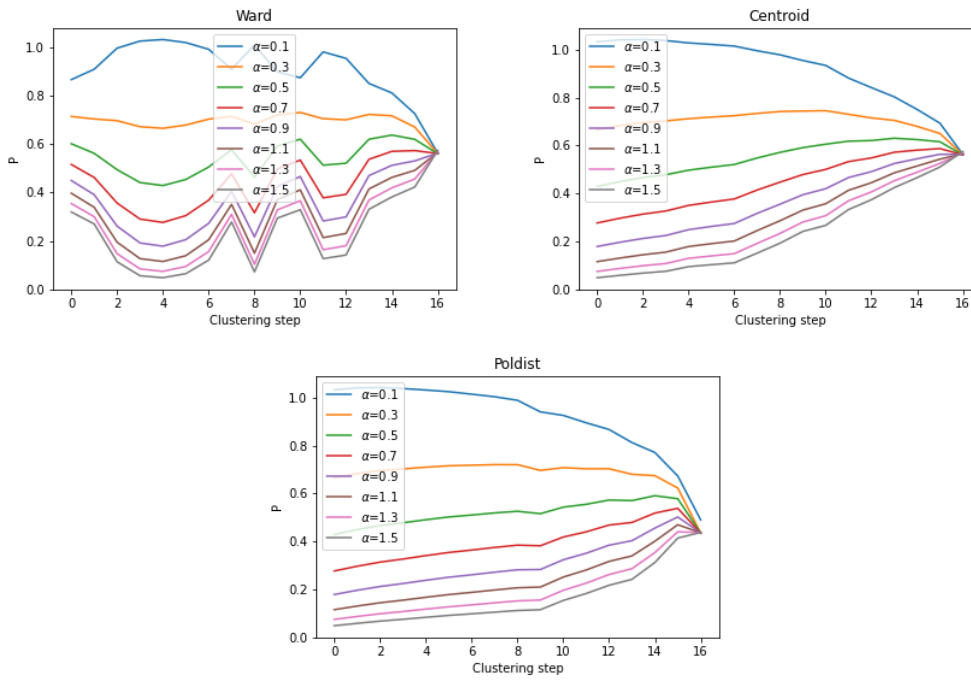
FIGURE 4.4: Polarisation throughout the clustering for different α in the Southern Women dataset, for the three distance update schemes considered: Ward, centroid and *poldist*.

α is to allow for intra-group identification to increase inter-group alienation, hence favoring greater cluster sizes. Figure 4.4 provides the intuition of the effect of α, and why it is called the *polarisation sensitivity*.

4.3.3 Evaluation

Clustering

Davis, Gardner, and Gardner, 1941 were not able to determine any meaningful women communities from the women-events network, since nearly all women were connected to each other by attendance to at least one common event (as pointed out by Breiger, 1974, there are 139 such links from a total of $(1/2)(18)(17) = 153$ possible ones, which makes a connectivity as high as 91%).

However, Homans, 1950 and Breiger, 1974 do report a specific community structure.

Homans finds that by rearranging the woman-event matrix so that women that coincided most often are grouped and events attended by most women are near the center (a tedious approach indeed, considering it had to be done manually at the time) two groups become distinct: one formed by Charlotte, Eleanor, Brenda, Theresa, Evelyn, Laura and Frances; and the other by Nora, Katherine, Helen, Sylvia and Myrna. The rest of the women are considered not to clearly belong to either group.

In the case of Breiger, by eliminating those events connected to every other event by at least one woman (in order to leave only events that help discriminate among women) he finds two groups formed by: Charlotte, Ruth, Eleanor, Brenda, Theresa, Evelyn, Laura and Frances; and Verne, Nora, Flora, Katherine, Olivia, Helen, Sylvia

and Myrna. Note that these contain Homans' groups, but add Ruth, Verne, Olivia and Flora, whom Homans judged not to belong to any group.

This means that we have two reference partitions to compare our clustering with, by means of the previously defined *modified Jaccard distance d_J^** (2.6).

Polarisation

Since Breiger, 1974 performs some operations on the original data, namely eliminating events connected to all other events, and the resulting groups of women are disconnected, it is of interest to evaluate the polarisation of such a system for reference. We expect such a polarisation to be rather high (close to 1).

To that end, I reproduced the steps followed by Breiger[2].

Polarisation is indeed rather close to 1, and greater than when using the raw data (0.68 vs 0.56).

Note that, since we have an even configuration $(8 + 8)$, what keeps the polarisation from its maximum is the distance between the two clusters: if the distance were 1 (the maximum) instead of 0.68, polarisation would also be 1.

The reason why the distance between Breiger's disconnected clusters is not maximal in our pipeline comes from the formula of the ϕ correlation (2.2): indeed, such a lack of connection in Breiger's case only means that the women in question do not share positive attendance values ($f_{11} = 0$), but the actual minimum in ϕ (i.e. the maximum in d) requires also that $f_{00} = 0$, i.e. that there is no overlap in positive *nor negative* values. Such is rarely the case even with Breiger's configuration, as we can see from $P2$ in (Breiger, 1974).

---

[2]Since we are looking for a kind of maximal reference value for the polarisation, I consider only the women in Breiger's two groups: Pearl and Dorothy, not belonging to either of them, I dismiss as noise.

Chapter 5

Use Case 2: The Conference on the Future of Europe data

5.1 Data

The Conference on the Future of Europe is a project of the European Union lead by the European Parliament, the Council and the European Commission to engage European citizens and more widely the European civil society in democratic deliberation on EU policies through the making and evaluation of concrete proposals and the organisation of and participation in events. By the end of the Conference (expected in the spring of 2022) the organisers commit to capture the proposals and discussions of the whole process into concrete policy recommendations.

Parcipants may be European citizens; European, national, regional or local authorities; or civil society organisations. They can engage through the multilingual platform of the Conference (featuring the 24 official languages of the EU), which is an instance of the *Decidim* application for mass deliberation [3]. It allows users to propose and vote "ideas" (that one can follow or endorse), organise and attend events ("meetings"), or leave comments on any of the two, among other things.

The digital platform was launched in mid-April 2021 and is foreseen to remain accessible until spring 2022, when the Conference is expected to reach conclusions [2]. At the time the data was collected for this project, the 20th October 2021 at 03h16min54s (GMT+1), there were a total of approximately 9000 proposals, 10000 endorsers (users that gave their support to at least one proposal) and 45000 proposal endorsements.

The data, which include information on proposals, meetings and their respective comments, present the opportunity to apply the pipeline proposed here on a current process of mass deliberation: the bimodal network is here composed of proposals and endorsers.

5.2 Experimental setting

5.2.1 Preprocessing

Before beginning the analysis on the CFE data, the following preprocessing actions were taken:

- Consider only proposals originally posted in English. This is to make sure there are no spurious distance or polarisation effects due to language[1].

---

[1]As expected from an EU platform, the proposals may be in any language of a member state. They are also optionally translated so that every user can vote and comment on every proposal, and the translations are included in the data.

- Remove amends. Proposals may be amends of previous ones, but I removed amends to avoid spurious endorser overlap.

- Take proposals as primary nodes. Although it seems more intuitive to cluster on endorsers rather than on proposals, it is a technical constraint that the number of the former is usually much larger than that of the latter. Additionally, performing the analysis on proposals allows to make qualitative sense of the results by actually reading and evaluating them, which cannot be done with anonymous endorsers, and any appreciative polarisation in the system should be present for both proposals and endorsers anyway (albeit with a relation that is not self-evident).

5.2.2   Analysis by topic

Proposals are categorised by topic ("Security", "Education", "Disinformation", "Coronavirus", etc.). Since it is expected that some users restrain their participation to the topics of their interest (for instance, an ecologist organisation might focus on climate change related proposals) and some topics may be more controversial than others, a topic-specific analysis imposes itself.

This part of the analysis on the CFE data follows the proposed pipeline, including the word shift graphs of the final partitions for evaluation.

5.2.3   Global analysis

The global perspective allows one to see whether users participate more or less homogeneously across topics (arguably the ideal case from a participatory point of view) or, as we might expect, they stick to proposals on one or a few given topics.

Given that α showed practically no influence on the final polarisation nor the topology of the poldist tree in the validation use case (chapter 4) nor in the topic-specific analysis of the CFE data and that the size of the data in this section makes the clustering rather slow, I only consider $\alpha = 1$.

5.3   Results

5.3.1   Analysis by topic

This section presents and discusses the results by topic. Given the quantity of plots involved and the qualitative homogeneity of results, I explicitly show here only a sample thereof: the full results are available in the project's Github repository[2].

Figure 5.1 shows the clustermaps of proposals on "Culture", ...

In general, the clustering rarely reveals any clear dipole structure (at least around the main diagonal, which is arguably its goal). There are, however, some exceptions, namely that of "Culture", but they are not prominent in any case.

The overall negative results may mean one of two things: either the systems are not significantly polarised, or our pipeline fails to capture such a polarisation. Given that the pipeline worked rather well in identifying the polarisation in the Southern Women dataset (at least the Ward and centroid methods), I lean towards the first interpretation.

In the cases where some interesting structure is revealed by the clustering, it seems the Ward method performs better than the other two: centroid tends to add

---

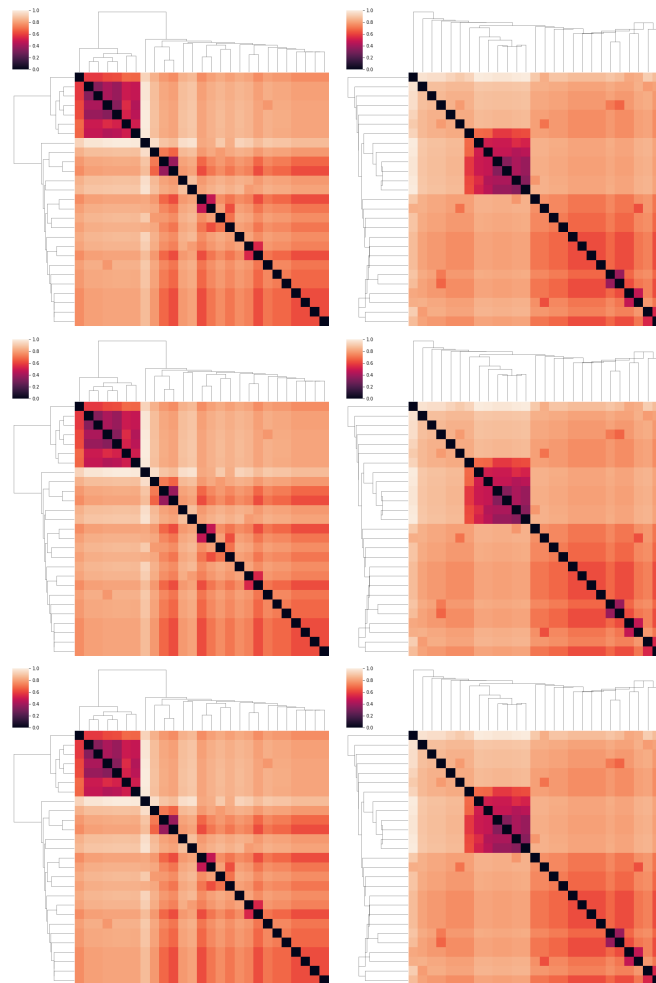[2]See https://github.com/adrifcid/polarisation.

FIGURE 5.1: CFE topic clustermaps.

up points to form one super-cluster, yielding a very low final polarisation (which is, on the other hand, a sign that our polarisation measure is working as it should); and poldist often features a heatmap showing the most off-diagonal structure, which a proper clustering should avoid. This appears to confirm the result of Chapter 4 that our distance update candidate features too strong a tendency to avoid merging large clusters to be functional.

Another feature we observe above is that a lower number of proposals tends to yield a higher final polarisation: indeed, the greater the population, the greater the number of possible configurations of the system, which means that a lesser population is in general more likely to end up with a configuration closer to the one that maximises polarisation that a greater one. This must be taken into account when comparing populations of different sizes.

Figures 5.2-5.2 show the word shift graphs of the final 2-cluster configurations of the topics displayed. The word shift graphs are overall not very relevant due to the general lack of polarisation of the clustering, but we can at least note that they show mostly meaningful words thanks to our extracting stop words beforehand. In addition, plotted words are in general ostensibly related to the corresponding topic (contrary to what would happen with spam or otherwise ill-intentioned content), which is good news regarding the participatory intent of the platform.

Regardless of the overall lack of polarisation, in the particular case of the Ward clustering of proposals on "Culture" one could argue we do see a distinction in the word shift graph, showing a more left-wing touch in the left cluster, which invokes words like "art", "education", "network" or "children" (although "economics" is also prominent); while the right one looks rather formal, perhaps more right-wing, featuring words like "Schuman" (probably referring to the Schuman Declaration), "debate" or "official".

Regarding the evolution of polarisation and tree distance for diferent α, we observe the same overall qualitative behaviour we saw with the Southern Women dataset.

For polarisation, we observe a general convergence at the end of the clustering, with the Ward lines being more wobbly thoughout the clustering; and $\alpha \approx 0$ gives initially high and decreasing polarisation, while higher values of α yield initially low and increasing polarisation (except in the cases where the clustering happens about one super-cluster, as happens frequently with the centroid distance update; there, the increase is followed by a sharp decrease).
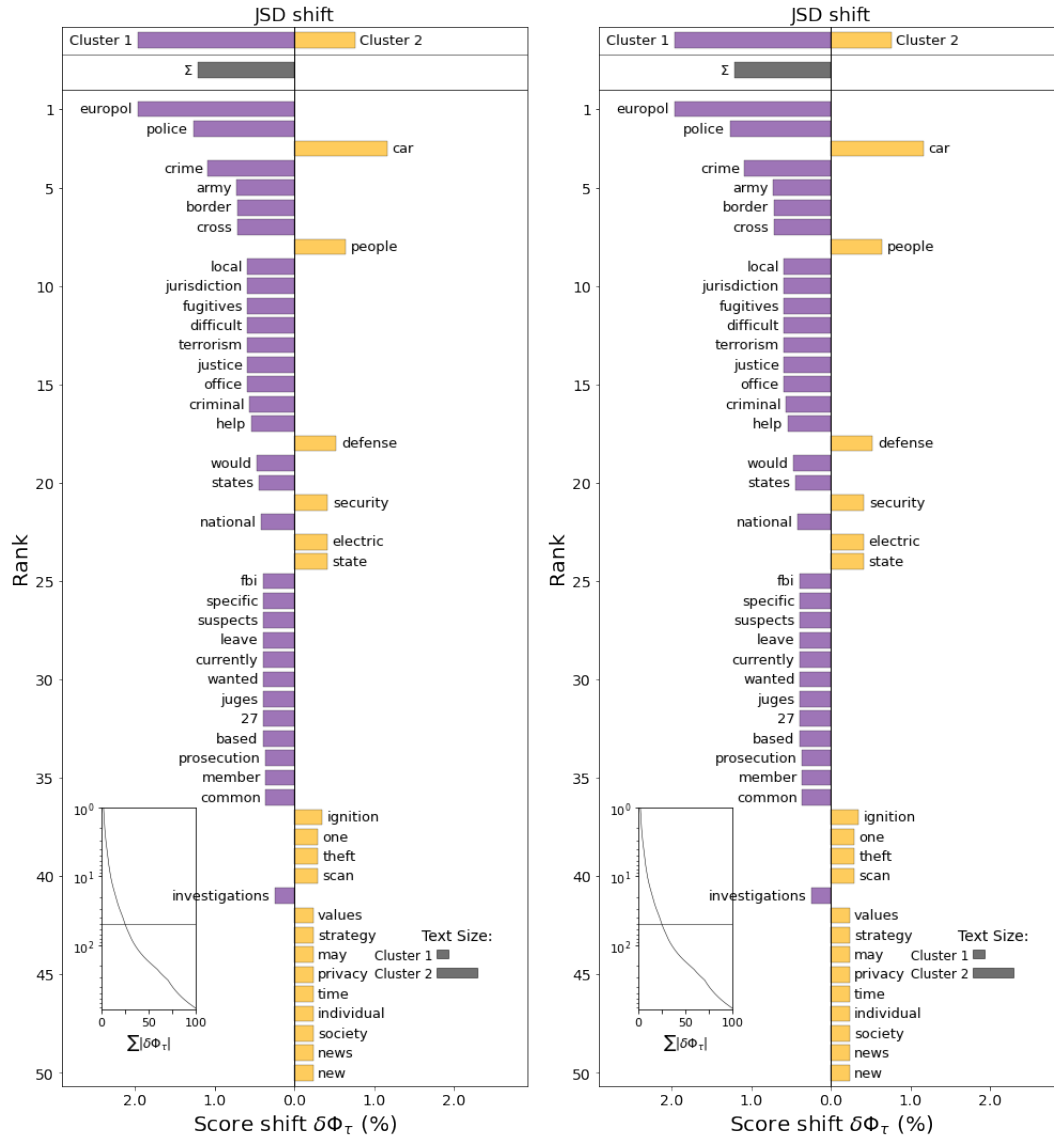
5.3.2   Global analysis

FIGURE 5.2: Word shift graphs of the final partition in the topic CFE data, for the two validated distance update schemes : Ward and centroid.
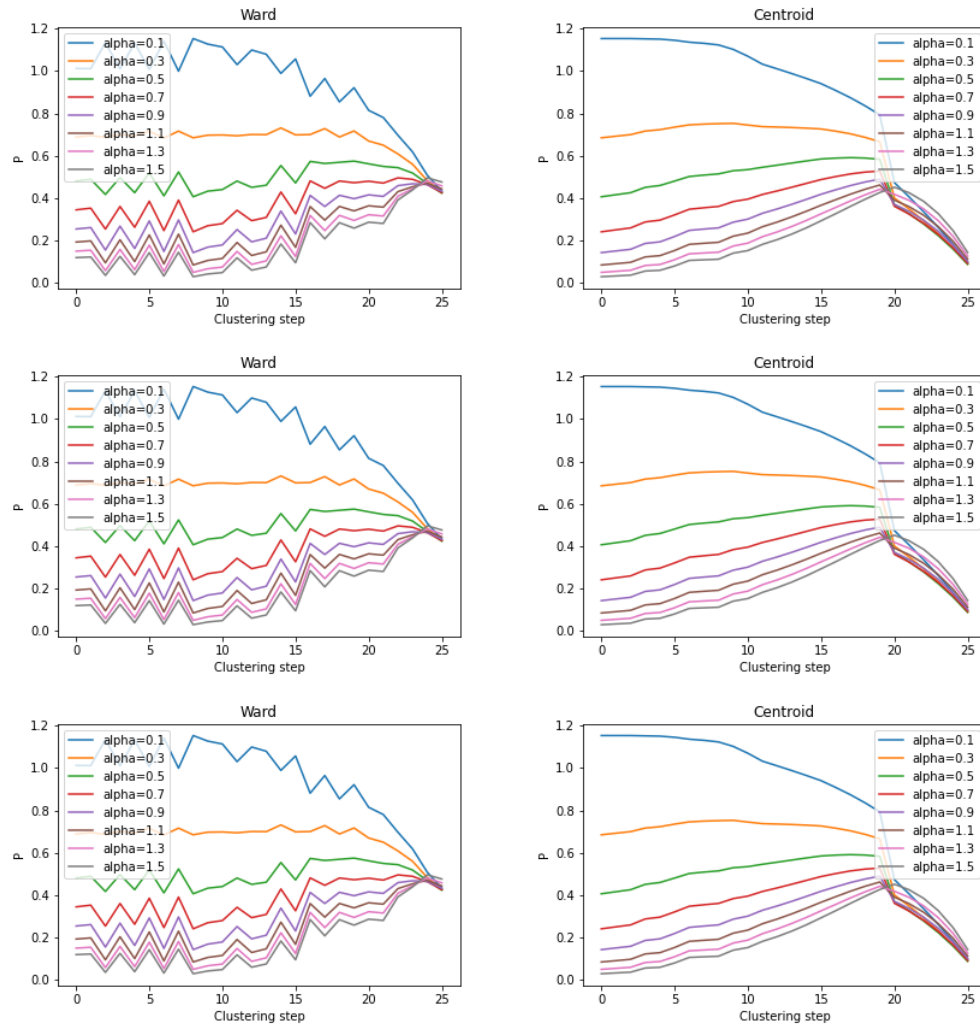
FIGURE 5.3: Polarisation throughout the clustering for different α in the topic CFE data, for the two validated distance update schemes : Ward and centroid.
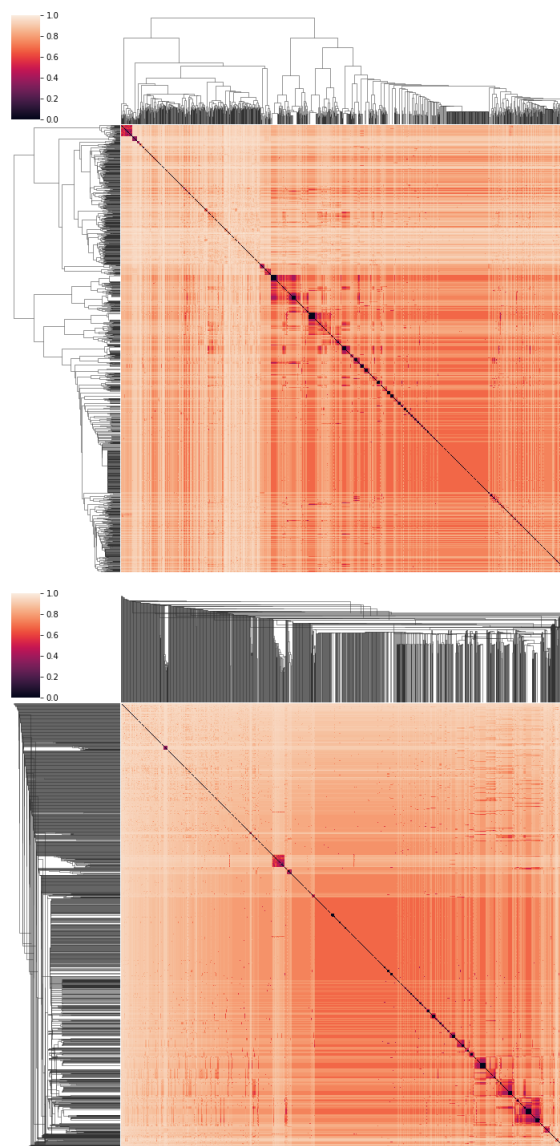
FIGURE 5.4: CFE global clustermaps.

Chapter 6

Conclusion

Social polarisation is a central issue in social science, and it has acquired mainstream interest in recent years.

Current methods in computational social science go beyond traditional distributional analysis by leveraging the tools of network science and the availability of digital data. Such structural polarisation measures present the advantage that they can be applied more or less automatically, although some of them have been shown to present undesirable features.

In this master's thesis, I have explored a hybrid approach that implements an axiomatic, theoretically-sound distributional polarisation measure (Esteban and Ray, 1994). The advantages of such a polarisation measure (2.19 are ...

The core pipeline studied here, intended for bimodal networks, is:

1. Build the bimodal network representation.

2. Compute the pairwise ϕ coefficient among the nodes of interest. Then compute $d(\phi)$.

3. Apply hierarchical clustering with polarisation evaluation at every step.

4. Compare final polarisation to clustering heatmap of initial d.

This novel pipeline is applied here to bimodal networks, which are less studied in the literature.

In the validation use case, on the standard Southern Women dataset (Davis, Gardner, and Gardner, 1941), results reasonably agree with the expected separation in two communities (Homans, 1950; Breiger, 1974) for the Ward and centroid distance update schemes of the clustering.

On the other hand, the application use case, on data from the Conference on the Future of Europe, shows no significant dipoles neither in the topic-specific nor the global analyses, which (given the previous pipeline validation) points to a lack of polarisation in the platform of the Conference. Such lack of polarisation is arguably a positive symptom in a mass deliberation platform.

Further work on the CFE data and/or the proposed pipeline may focus on any subset of the following paths:

- Higher-order multiple analysis. Contrary to most structural polarisation measures, (2.1) allows for the evaluation of multipole systems, and such an analysis may yet reveal some structure in the CFE data

- Refinement of word shifts (for CFE or any new text data). The text representations that provide the input for the word shifts do not necessarily have to be

straight, frequency-based BOW representations: more sophisticated representations may be weighted, such as the *tf-idf* representation, or include a preprocessing step that accounts for features like synonymity in the considered text, e.g. through *topic modelling*.

- Consider other distance update schemes for the hierarchical clustering (even a newly parameterised poldist to try and correct its present bias), or perhaps a different clustering method altogether. As I have mentioned, clustering methods are far from consensual and may influence the polarisation measurement independently of the polarisation measure used. Even if the current method does not seem flawed, it would be a useful check to compare it to others, both at the distance update and more fundamental levels.

- For the CFE data, perhaps refreshing it would provide new, relevant information: in any case, it would yield a more accurate representation of the system and allow to further explore the behaviour of the proposed pipeline with respect to network size.

- Pipeline optimisation. This point is of particular necessity if one is to consider greater network sizes, as the global analysis on the CFE data already takes $\sim 10h$ in my personal laptop. A most straightforward way to optimise the clustering in a dramatic way, now that the code is stable, is to re-write it using the Cython library.

- Mathematical checks, developments, adaptations? Of polarisation measure.

Finally, current results show the proposed pipeline remains a promising candidate for the study of polarisation in bimodal social networks and should be further explored.

Bibliography

Breiger, Ronald L (1974). "The duality of persons and groups". In: *Social forces* 53.2, pp. 181–190.

Davis, A, BB Gardner, and MR Gardner (1941). "Deep South; a social anthropological study of caste and class." In:

Esteban, Joan-Maria and Debraj Ray (1994). "On the measurement of polarization". In: *Econometrica: Journal of the Econometric Society*, pp. 819–851.

Fortunato, Santo and Darko Hric (2016). "Community detection in networks: A user guide". In: *Physics reports* 659, pp. 1–44.

Gallagher, Ryan J et al. (2021). "Generalized word shift graphs: a method for visualizing and explaining pairwise comparisons between texts". In: *EPJ Data Science* 10.1, p. 4.

Homans, George C (1950). "The human group." In:

Jaccard, Paul (1912). "The distribution of the flora in the alpine zone. 1". In: *New phytologist* 11.2, pp. 37–50.

Müllner, Daniel (2011). "Modern hierarchical, agglomerative clustering algorithms". In: *arXiv preprint arXiv:1109.2378*.

Pearson, Karl (1900). "I. Mathematical contributions to the theory of evolution.—VII. On the correlation of characters not quantitatively measurable". In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 195.262-273, pp. 1–47.

Robinson, David F and Leslie R Foulds (1981). "Comparison of phylogenetic trees". In: *Mathematical biosciences* 53.1-2, pp. 131–147.

Salloum, Ali, Ted Hsuan Yun Chen, and Mikko Kivelä (2021). "Separating Controversy from Noise: Comparison and Normalization of Structural Polarization Measures". In: *arXiv preprint arXiv:2101.07009*.

Virtanen, Pauli et al. (2020). "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17, pp. 261–272. DOI: 10.1038/s41592-019-0686-2.