University of Macerata

Department of Economics and Law

PhD Program in
Quantitative Methods for Economic Policy
Cycle XXXIII

Essays on the Network Analysis of Culture

Supervisors
Prof. Luca De Benedictis
Prof. Giancarlo Ragozini

PhD student
Roberto Rondinelli

Coordinator
Prof. Luca De Benedictis

Year 2019-2020

Contents

Preface **7**

Introduction **10**

Networks and symbols **21**

1 The network component of countries' cultural distances **24**
1.1 Introduction 24
1.2 Data: WVS/EVS Joint 2017 28
1.2.1 Inglehart-Welzel contribution: Cultural Map 30
1.2.2 Selected questions/variables 36
1.3 Empirical definition of culture 38
1.4 Cultural traits interdependence 41
1.4.1 Inferred countries networks of cultural values 44
1.5 Definition of distance measures 52
1.6 A new index of cultural distance 54
1.6.1 The compromise cultural distance 56
1.7 Conclusions 61
Annex 63

2 Comparing cultural distances with other distances among countries **76**
2.1 Introduction 76
2.2 Proposed cultural distance versus other distance measures 78
2.2.1 Facebook Social Connectedness Index 79
2.2.2 Combined ethnic and linguistic distance 80
2.2.3 Genetic distance 82
2.2.4 Climatic distance 84
2.3 Between-distances analysis 86
2.4 A model for the GDP per capita 91

2.5 Conclusions 94

3 Mapping networks: evidences from a simulation study 99
3.1 Introduction 99
3.2 Network simulation 102
3.2.1 Random networks 103
3.2.2 Scale-free networks 103
3.2.3 Small-world networks 104
3.2.4 Stochastic block model networks 105
3.3 Network descriptors 106
3.3.1 Micro-level descriptors 108
3.3.2 Meso-level descriptors 108
3.3.3 Macro-level descriptors 109
3.4 Find the best subset of descriptors 110
3.4.1 Subgroup Discovery 110
3.4.2 PCA: mapping networks 115
3.5 Case study: binary cultural networks 119
3.6 Conclusions 123
Annex 125

Conclusions 138

List of Figures

1.1 Inglehart-Welzel Cultural Map 35
1.2 The Network Structure of National Cultures (Sweden and Tunisia, IW variables 46
1.3 The Network Structure of National Cultures (Italy and Slovakia, IW variables 47
1.4 The Network Structure of National Cultures (Sweden and Tunisia, first battery variables) 48
1.5 The Network Structure of National Cultures (Italy and Slovakia, first battery variables) 49
1.6 Projection of the cultural distances on the first two dimensions of the non centered PCA over their *cosine matrix* 59

2.1 Cultural distance vs. facebook Social Connectedness Index 80
2.2 Cultural distance vs. Fearon Ethnic-linguistic index 81
2.3 Cultural distance vs. Genetic distance 83
2.4 Cultural distance vs. Pemberton genetic distance 84
2.5 Cultural distance vs. Climatic distance 86
2.6 Cultural distance vs. Helpman distance in GDP per capita 92

3.1 Mapping of simulated networks on the first two dimensions of the PCA using the overall set of descriptors 115
3.2 Correlation Circle of the first two dimensions of the PCA on simulated networks using the selected subset of descriptors 117
3.3 Mapping of simulated networks on the first two dimensions of the PCA using the selected subset of descriptors 118
3.4 Correlation Circle of the first two dimensions of the PCA on binary cultural networks using the selected subset of descriptors 120
3.5 Mapping of binary cultural networks on the first two dimensions of the PCA using the selected subset of descriptors 121
3.6 Non centered PCA of the *cosine matrix* of the network distances . 122

List of Tables

1.1 Example of *Battery dataset* . . . 40
1.2 Example of symbolic generalization for *Battery dataset* . . . 40
1.3 Country 1 . . . 42
1.4 Country 2 . . . 42
1.5 Country 1 . . . 42
1.6 Country 2 . . . 42
1.7 Networks statistics . . . 50
1.8 Vectorized *cosine matrices* by dataset . . . 57
1.9 Results of the non centered PCA over the *cosine matrices* associated to each dataset . . . 58
1.10 *Cosine matrix* of the cultural distances . . . 61
1.11 Differences between BDgraph with different number of iterations, applied to 60 cultural traits for Italy . . . 67
1.12 Differences between BDgraph with different number of iterations, applied to 60 cultural traits for Italy . . . 68
1.13 List of the 40 selected variables . . . 68
1.13 List of the 40 selected variables . . . 69
1.13 List of the 40 selected variables . . . 70

2.1 *Cosine matrix* of the joint comparison between cultural distances and other distances . . . 88
2.2 Associated weights to distance measures for different DISTATIS . 90
2.3 Regression Modeling over distance in GDP per capita . . . 93

3.1 Micro-level descriptors . . . 108
3.2 Meso-level descriptors . . . 109
3.3 Macro-level descriptors . . . 110
3.4 Subgroup Discovery: main results . . . 113
3.5 Subgroup Discovery for Random networks (ER) . . . 125
3.6 Subgroup Discovery for Scale-free networks (SF) . . . 125
3.7 Subgroup Discovery for Small-world networks (SW) . . . 126

3.8 Subgroup Discovery for Stochastic block model networks (BM) . . 126
3.9 Contributes and squared cosine of the PCA on the overall set of descriptors 128
3.9 Contributes and squared cosine of the PCA on the overall set of descriptors 129
3.9 Contributes and squared cosine of the PCA on the overall set of descriptors 130
3.10 Simulation mapping timing 130

Preface

In their definition, Networks are mathematical objects - composed by a set of vertices (nodes), a set of links (edges or arcs) and some vertex-level and link-level metadata. Networks can be abstract constructs but they can also formally represent real structures. You don't need to be a practitioner of mathematical statistics to be acquainted of expressions such as "six degrees of separation", "global village", "small world" or "networks are everywhere". The first idea of "six degrees of separation" was introduced by the Hungarian writer Frigyes Karinthy in a 1929 novel, but the expression is due to the title of a John Guare's play made in 1991, where, following studies of psychologist Stanley Milgram, the author talks about the possibility to reach each people around the world using four intermediaries, that means five steps, that define six degree of connection or of separation, depending on the mood of the observer. Under this prospective, the human society can be seen as a complex social network (Barabasi, 2014) and examples of connectedness in real life come to mind, easily: international trade flows, business relations between companies, interpersonal relationships (any relationship between individuals), internet communications and social networks (Facebook, Twitter, etc.), transport infrastructure, exchanges of letters, biological systems, chemical compounds, epidemiological events. The list could continue for pages and pages.

Having said that, the application of Network Analysis methods is indisputably an interdisciplinary field (Wasserman and Faust, 1994), which is positive for the dissemination of the approach and the diffusion and the validation of techniques, but at the same time it has contributed to the segmentation and dispersion of methodologies and to the difficulty of dialogue between different fields (Newman, 2018). In the recent scientific literature, the number of contributions that explicitly refer to Network Analysis are exponentially growing, to such an extent that one might think they are conveyed by the vogue of the times we are living in. But the interest of scholars of different disciplines is more than justified by recent theoretical and empirical advances. For example, underlying networks act in digitisation and all the IT processes, which have became so automatic in our everyday life: fingerprint and facial recognition by the smartphone, the use of the web, cyber security, videogames, etc. In short, in all those situations where it is more convenient considering attributes and properties of pairs of entities (dyadic attributes) instead single entities (monadic attributes), Network Analysis meth-

ods lead to a more accurate understanding of reality (Borgatti and Everett, 1997).

If there was a need for further evidence that human activities are centered on interactions and spatial movements - that essentially create ties between entities - just think about the current COVID-19 pandemic. The evolution of society itself has been influenced by this human peculiarity, so much so that in emergency conditions people suffer from the impossibility of having contacts and not being able to persevere in their favorite activity, which is to feed their social networks. As evidence of this, a study by the Italian start-up "enuan" shows that during the pandemic empathic interactions with bots, such as Alexa and the smartphone's voice assistant, have greatly increased in Italy. More generally in the world have increased the demand for pet robots, such as "ElliQ" produced by the Israeli Intuit Robotics, "Hatsune Miku" who depopulates in Japan, and Paro who acts as a pet animal.

The current situation is also the perfect example to highlight the importance of studying networks. Looking at the network of human relationships, virus prefer a lot of connections for its subsistence, while humans prefer to have few connections between them to keep the virus from spreading, but at the same time they would like to continue their usual activities: different points of view, different desired network structures.

Analyzing the objective organization of the network, one can firstly evaluate how it affects real events and secondly one can intervene on it for changing the way phenomena spreads. For example, companies like Facebook can organize data from their Whatsapp, Instagram and Facebook users in multi-layer networks[1] to propose topics and advertising relevant to them. This operation exploits the network and the current behaviors/preferences of its nodes (users) to predict future behaviors/preferences, in some way affecting the user and therefore also acting on the network, polarizing, extending or clustering it.

One component that is closely linked to network ties formation, disruption and preservation along time, is *distance*, not only in its physical/geographical meaning, but even more concerning not easily measurable quantities, such as opinions, attitudes, behaviors or more generally cultures. The quantitative evaluation of

[1] A multi-layer network (in specific cases called multiplex) is a particular kind of network where nodes are organized in different layers based on different types of relationships they have with each other. See Kivelä et al. (2014) for the complete definition and terminology.

the distance between two entities is a dyadic property and as such, the presence, intensity, direction and sign of their tie is a way to undertake it. Since entities can be individuals, objects, companies, countries, planets, as well as networks referring to specific contexts, and the way to measure similarity between them is various, a peculiarity thing of distances is their changeable nature. While physical distances are almost objectively computable, in case of culture (and even other more or less broad concepts) using a method rather than another could radically change the proximity relationship between entities, especially if they have a high degree of complexity.

Description and measurement of culture concept is one of the most debated topic along the main literature of different fields, Kroeber and Kluckhohn (1952) offer to this extent a wide collection of definitions in which authors distinguish seven typologies: descriptives, historical, normative, psychological, structural, genetic and incomplete. Sociologists and anthropologists as Edward Burnett Tylor (Tylor, 1871), Emile Durkheim (Durkheim, 1912), Max Weber (Weber, 1904), Claude Levi-Strauss (Lévi-Strauss, 1987), Talcott Parsons and Alfred Louis Kroeber (Kroeber and Parsons, 1958), Anthony Giddens (Giddens 2000, Baecker 1997), Pierre Bourdieu (Bourdieu, 1977), Franco Ferrarotti (Ferrarotti, 1986), etc. with their contributions provided a huge effort to enunciate in a discursive and conceptual way the intrinsic meaning of culture, which is perhaps the most complex and articulated concept present in literature. By incorporating all the definitions into one big element, thus considering *culture* as a large box containing the broad sense of human being, everything we daily do turns into its concept, therefore, as recent empirical papers suggest, it is unavoidable that culture matters for economics and politics attitudes (Alesina and Giuliano, 2015). To study these effects scientifically and make adequate decisions, the statistical quantification of culture is a pivotal moment for empirical analysis.

In this thesis work, distances between networks, network inference and network features are blended to obtain an accurate representation of the complex cultural heritage of countries, to map cultural networks in a reference space, to examine similarities of cultural distance together with other socio-economic distances indicators and to converge these distances into a regressive model over the distance between countries in GDP per capita.

Introduction

Since the end of the 19th century, several social researchers have focused on the concept of culture, proposing - in relation to their historical and environmental context - a comprehensive definition for them.
A clear example of how the context can affect the critical thinking of scholars is given by socioligists Emile Durkheim and Max Weber, who, although contemporary, lived different scientific backgrounds in different nations. They have developed an opposite awareness to each other towards the concept of culture, in fact the former depicted it as a superior organism objectively present (Durkheim, 1912), while the latter defined it as closely related to the perceptions of the human being (Weber, 1904).
Before them, Edward Burnett Tylor offered a broad definition, stating that culture is *"that complex whole which includes knowledge, belief, art, morals, law, custom, and any other capabilities and habits acquired by man as a member of society."* (Tylor, 1871).
Claude Levi-Strauss, anthropologist as Edward Burnett Tylor, catching up on work by Marcel Mauss (anthropologist and sociologists) imagining any culture as a set of interlinked symbolic systems (art, science, religion, economic relations, etc.), each of them represent a part of the physical and social reality and their relation (Lévi-Strauss, 1987). Even for the sociologist Anthony Giddens culture is an extended and tangible construct, in fact he identifies it with *"the way of life of the members of the sociery, or of groups within society"* (Giddens, 2000).
Recent points of view are offered by other researchers like Franco Ferrarotti and Pierre Bourdieu. Looking at culture not in an elitist way, Franco Ferrarotti conceives it as a set of shared and lived together experiences and values, fundamental for the society cohesion. Furthermore, the contemporary individual is seen as the custodian of an experience constituted by all that the past has sedimented in him, then he can analyse his background as reflection of his present and to plan the future (Ferrarotti, 1986).
An important and delicate notion as old as Aristotle, is proposed and reorganized by the sociologist Pierre Bourdieu. It is the habitus. Placing himself in a position as objective as possible, the scholar realizes that habitus gathers within itself a series of characteristics and attitudes (religion, ethnicity, level of education, etc.) capable of creating regular behaviors within a social group, then, it carries tradition and cultural reproduction. Anyway, habitus is closely linked to the structures constitutive of a particular type of environment, namely the

reference context of a social group (Bourdieu, 1977). After all, taking up the first lines of this introduction, Emile Durkheim and Max Weber are themselves a product of the habitus. In statistical terms, habitus is identifiable by all features that define a cluster of individuals, an object of varying complexity according to the specificity (size) of the social group being investigated. For istance, to unite the culture of countries or regions geographically defined, it means to define their habitus.

In addition to purely theoretical definitions of the overall concept of culture, over the years an empirical trend has developed, which, making use of survey data[2], is aimed at unpacking the construct of culture in sub-constructs, called dimensions. In this regard, Shalom H. Schwartz, Geert Hofstede, the GLOBE team and Inglehart-Welzel's contributions are well known[3]. Shalom H. Schwartz, moving from a theoretical point of view, proposed two main dimensions: Egalitarianism vs Hierarchy, Autonomy vs Embeddedness (Schwartz 2008, Schwartz and Bilsky 1987). Geert Hofstede's work led to six cultural dimensions: Power Distance Index, Individualism vs Collectivism, Masculinity vs Femininity, Uncertainty Avoidance Index, Long Term Orientation vs Short Term Normative Orientation, Indulgence vs Restraint (Hofstede, 2011). Even eighteen dimensions are contained within the GLOBE model, of which nine are due to cultural values and nine are due to practices[4]. Whilst, Ronald Inglehart and Christian Welzel, from their World Values Survey, built the Cultural Map (where countries are mapped) affirming the preponderance of two dimensions, which are in parallel with those of Shalom H. Schwartz: Secular vs Traditional, Survival vs Self-Expression[5] (Inglehart and Welzel, 2005).

The characteristic of the databases derived from these surveys (EVS and WVS, for example) to be freely available, in addition to the direct work done by the

[2] Some of them focus on one country, e.g. the General Social Survey (GSS) for the US; others on a group of countries referring to a single geographical or political area, e.g. Eurobarometers and European Values Survey (EVS) for EU area; others on worldwide countries, e.g. The Life in Transition Survey (LITS), the World Values Survey (WVS) (Inglehart 1997), the IBM study of Geert Hofstede (Hofstede 1984), the GLOBE project (House et al. 2004), the study conducted by Shalom H. Schwartz (Schwartz 1992).

[3] An extensive review of this methods used for measuring the concept of culture (often synonymous with country) is provided by Taras et al. (2009).

[4] Following the definition of Reckwitz (2002), *"a practice ... is a routinized type of behaviour which consists of several elements, interconnected to one other: forms of bodily activities, forms of mental activities, 'things' and their use, a background knowledge in the form of understanding, know-how, states of emotion and motivational knowledge. A practice - a way of cooking, of consuming, of working, of investigating, of taking care of oneself or of others, etc. - forms so to speak a 'block' whose existence necessarily depends on the existence and specific interconnectedness of these elements, and which cannot be reduced to any one of these single elements."*

[5] Other scholars consider cultural distance as isomorphic to "psychic distance" (Sivakumar and Nakata 2001, Gomes and Ramaswamy 1999, Lee 1998), which describes tangible and non-tangible characteristics that establish the closeness between two subjects (individuals, companies, countries, etc.) (Beckerman, 1956). Others demonstrate as cultural distance - mainly in the sense of Geert Hofstede's dimensions, but indirectly also in the way of Shalom H. Schwartz and Inglehart-Welzel - is only one component of psychic distance (Dow and Karunaratna, 2006).

authors of the surveys, allowed the explosion of work on secondary data that - given the popularity, vastness and complexity of the cultural topic - have tried to clarify new and old questions about the study of culture, such as its definition and measurement, the comparison between different works, the relationship with economic activities, etc. Together with the qualitative antecedent work of anthropologists, sociologists, archeologists and philosophers, the inevitable recent depopulation of quantitative works on the concept of culture, on the thousand facets of the cultural background of each person or group of people and on the relations between it and human activities, has led to the birth of new branches, among all, those we are most interested are the Cultural Economics[6] (Throsby 2001, Towse and Hernández 2020) and Cultural Networks[7] (McLean 2016, Breiger and Puetz 2015), and an exceptional expansion of literature, which organization has demanded more and more reviews. For example, Kroeber and Kluckhohn (1952) offered a review about definitions of culture, while Taras et al. (2009) analyzes different instruments for quantitatively measure culture, draws the common points between the various operational definitions of culture and proposes future developments. Alesina and Giuliano (2015) provides a review of the relationship between culture and institutions. Koltko-Rivera (2004), organizing the theory about the broad concept of Worldview[8], implicitly provides sparks to show the complexity of the word "culture", in fact it, in addition to its numerous definitions, can also be forged from those of other constructs used in its same way.
The different shades of culture, mostly related to economics, are for example captured by Guldenmund (2000) that organized the literature about safety culture[9], Miroshnik (2002) that analyzes the relationship between multinational companies behavior and national cultures, Nakata and Sivakumar (1996) which reviews the relationship between national culture and new product development, Hayton et al. (2002) with the relationship between national cultural characteristics and aggregate measures of entrepreneurship, Salant and Lauderdale (2003) which examines how acculturation affects the health of Asian immigrants in United States, Canada, Australia, New Zealand, and United Kingdom.
In the same way, there are many complex social phenomena that affect or are closely related to the individual or community cultural background, like the glob-

[6] Cultural Economics is a so broad field of study *"relevant to arts organizations, creative industries, cultural policy and, increasingly, to economic policy for growth and development"*, such as Towse and Hernández (2020) organized it in Handbook, which chapters are written by experts of a specific subfield.

[7] Cultural Networks embrace the various ways in which we can study cultural phenomena via Social Network Analysis. For our knowledge, Networks of cultural traits (values) have never been inferred in literature.

[8] As denoted by Koltko-Rivera (2004), worldview have been labelled in literature in different ways. The most common are: "world hypotheses" (Pepper, 1942), "world outlook" (Maslow, 1981), "visions of reality" (Messer, 1992).

[9] Safety cultures reflect the attitudes, beliefs, perceptions, and values that employees share in relation to safety (safety culture) (Cox and Cox, 1991).

alization (Pieterse et al. 2019, Olivier et al. 2008, Tomlinson 2012, Anthias 2001), which goes hand in hand with immigration (Rapoport et al. 2020) and economic preferences (Falk et al., 2018) fundamental for the production and the market functioning. Obviously cultural status is related with ethnicity (Desmet et al., 2017) and diversity (Ashraf and Galor, 2013). A place in the world where the complexity of culture and its evolution is particularly evident is the United States, that in this sense represents a strong stimulus for scholars (Giavazzi et al. 2019, Bertrand and Kamenica 2018).

This brief but intense excursus on some of the most popular definitions and features, inevitably shows how the articulation and permeability of culture within economic activities and relationships is obvious and implicit. They are so entangled with each other that after being absent from mainstream economics for a long time, culture - in the broad sense of local norms, customs, attitudes, values, and their subsets and *interactions*[10] - has entered economic analysis again, especially as far as measurement is concerned. In the last twenty years or so, dozens of empirical papers did support the claim that culture matters for economics, over and above the role of institutions (Alesina and Giuliano, 2015).

Moving the attention from the conceptual construction of the word culture to the similarity between entities with respect to it, we get to the core of the question that concerns this dissertation, namely the *Cultural Distance*. Already inside the Inglehart-Welzel Cultural Map or on the measurement of the Shalom H. Schwartz and Geert Hofstede's dimensions, it is possible to estimate the cultural distance between geographically defined entities. But is it completely reliable? Can cultural dimensions lack a component of interdependence between cultural traits?

In economic relations, in international agreements and in institutional dialogue, the word distance is one of the most enunciated. There are exogenous distances to be bridged to ignite a bond, sometimes there are necessary cracks and other times unavoidable breaks, but this may depend, as well as geographical and physical distances, and implicit interests, largely on the cultural status of groups of individuals. Moreover, behaviours and attitudes of a group of people can influence the type of economic system adopted and its functioning. This is why the

[10] This definition follows mostly that of the English anthropologist, and founder of cultural anthropology, Sir Edward Burnett Tylor, who already in 1871 offered a similar encompassing definition cited above. However, it summarises most of the definitions in the literature, which in the form of rhetorical figures and different words depict culture as a complex set of components. In this context another generic definition of culture, from which we can appreciate dynamism of culture along time, is expressed by Guiso et al. (2006) and reported in Alesina and Giuliano (2015): *"those customary beliefs and values that ethnic, religious, and social groups transmit fairly unchanged from generation to generation".*

measurement precision of the cultural content of an entity and of the cultural distance between them (in our case are world countries) is essential to study the mechanisms that regulate certain phenomena and to make the right decisions, especially in this historical period where cultural changes due to the immediacy of information and social media interactions, are sudden.

As in the case of national cultures, sometimes the elements on which to assess distances can be complex and therefore require more detailed procedures. Discrimination between networks of varying sizes (Van Wijk et al., 2010; Smith et al., 2016) and related to a given research field is one such case. Common distance measurements between networks often fail to grasp the complex of information within two or more networks, then we arrive to another empirical objective of this dissertation that is to propose a set of network descriptors that can capture the important information contained in the networks and discriminate them based to the generative models they come from.

In the spirit of Inglehart-Welzel (Inglehart and Welzel, 2005), this thesis - which does not intend to discuss or argue about the issues of cultural dimensions or psychic distance, nor to arrive at demonstrative/mathematical conclusions - aims to show empirical evidence from data-driven approach, making use of new data from the World Values Survey (Wave 7) and European Values Survey (Wave 5) joint survey. Data are available free of charge on the official survey website and cover a total of 79 countries, providing individual specifications for each of them. Complexity of data, concept of culture and measurement of cultural distance requires a complex approach involving Network Analysis methods firstly, as well as special distance measurements, multidimensional analysis, bayesian methods, simulations and clustering.

Chapter 1 is based on and extends the work of De Benedictis et al. (2021) on data from WVS Wave 6. The main objective is the same, namely to identify the interdependence between the cultural traits from the WVS and show their importance in defining the cultural distance between countries. To discern this task, here, we use data from the newly published of the WVS/EVS Joint 2017 survey, furthermore, in reference to De Benedictis et al. (2021), who concentrated and used the 10 indicators selected by Inglehart and Welzel for the construction of their Cultural Map, here four subsets of data with a different number of variables are considered: the 6 variables from the first battery of questions, the 10

Inglehart-Welzel Cultural Map variables, the 14 Inglehart-Welzel Cultural Map variables (for Y002 and Y003 indicators[11] we use the original variables) and 60 variables (of these, 14 are the variables previously defined, 6 are those from the first battery of questions and the others are selected to get a number that can cope with the trade off between processing time and the minimum number of missings per country).
Chapter 2 is devoted, on the one hand, to the comparison between the cultural distance measurement found in Chapter 1 and different distance measurements. These include similarities in terms of climate, ethnics and linguistics, genetics and recent phenomena like Facebook. On the other hand, to a model of distance between countries in GDP per capita, where the variety of distance measures considered in this Chapter acts as regressors.
Finally, Chapter 3, by mapping simulated binary networks in a reference space through a subset of descriptors, analyzes their clusterization by their generative process and the problem related to the discrimination power of descriptors over a set of non-isomorphic networks. Procedure involves the calculation of a large set of descriptors on 2400 networks generated by different models (Random, Scale-free, Small-world and Stochastic block model) and, via Subgroup Discovery, the choice of the best subset of descriptors, which preserves the distinction between networks from different generative models. The binary cultural networks with 60 variables estimated in Chapter 1 are used as case study and their mapping is compared with some renowned measurements of distance between networks.

[11] They respectively describe Post-materialism and Autonomy indices.

References

Alesina, A. and Giuliano, P. (2015). Culture and Institutions. *Journal of Economic Literature*, 53(4):898–944.

Anthias, F. (2001). New hybridities, old concepts: the limits of 'culture'. *Ethnic and racial studies*, 24(4):619–641.

Ashraf, Q. and Galor, O. (2013). Genetic diversity and the origins of cultural fragmentation. *American Economic Review*, 103(3):528–33.

Baecker, D. (1997). The meaning of culture. *Thesis Eleven*, 51:37–51.

Barabasi, A.-L. (2014). *Linked-how Everything is Connected to Everything Else and what it Means F.* Perseus Books Group.

Beckerman, W. (1956). Distance and the pattern of intra-european trade. *The review of Economics and Statistics*, pages 31–40.

Bertrand, M. and Kamenica, E. (2018). Coming apart? cultural distances in the united states over time. Technical report, National Bureau of Economic Research.

Borgatti, S. P. and Everett, M. G. (1997). Network analysis of 2-mode data. *Social networks*, 19(3):243–270.

Bourdieu, P. (1977). *Outline of a Theory of Practice*, volume 16. Cambridge University Press.

Breiger, R. L. and Puetz, K. (2015). Culture and networks. *International encyclopedia of social and behavioral sciences*, 5:557–62.

Cox, S. and Cox, T. (1991). The structure of employee attitudes to safety: A european example. *Work & stress*, 5(2):93–106.

De Benedictis, L., Rondinelli, R., and Vinciotti, V. (2021). The network structure of cultural distances. *arXiv:2007.02359*.

Desmet, K., Ortuño-Ortín, I., and Wacziarg, R. (2017). Culture, ethnicity, and diversity. *American Economic Review*, 107(9):2479–2513.

Dow, D. and Karunaratna, A. (2006). Developing a multidimensional instrument to measure psychic distance stimuli. *Journal of International Business Studies*, 37(5):578–602.

Durkheim, É. (1912). *Les formes élémentaires de la vie religieuse: le système totémique en Australie*, volume 4. Alcan.

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., and Sunde, U. (2018). Global evidence on economic preferences. *The Quarterly Journal of Economics*, 133(4):1645–1692.

Ferrarotti, F. (1986). *Manuale di sociologia*. Laterza.

Giavazzi, F., Petkov, I., and Schiantarelli, F. (2019). Culture: Persistence and evolution. *Journal of Economic Growth*, 24(2):117–154.

Giddens, A. (2000). *Fondamenti di sociologia*. il Mulino.

Gomes, L. and Ramaswamy, K. (1999). An empirical examination of the form of the relationship between multinationality and performance. *Journal of International Business Studies*, 30(1):173–187.

Guiso, L., Sapienza, P., and Zingales, L. (2006). Does culture affect economic outcomes? *Journal of Economic Perspectives*, 20(2):23–48.

Guldenmund, F. W. (2000). The nature of safety culture: a review of theory and research. *Safety science*, 34(1-3):215–257.

Hayton, J. C., George, G., and Zahra, S. A. (2002). National culture and entrepreneurship: A review of behavioral research. *Entrepreneurship theory and practice*, 26(4):33–52.

Hofstede, G. (1984). *Culture's consequences: International differences in work-related values*, volume 5. sage.

Hofstede, G. (2011). Dimensionalizing cultures: The Hofstede model in context. *Online readings in Psychology and Culture*, 2(1):8.

House, R. J., Hanges, P. J., Javidan, M., Dorfman, P. W., and Gupta, V. (2004). *Culture, leadership, and organizations: The GLOBE study of 62 societies*. Sage publications.

Inglehart, R. (1997). *Modernization and postmodernization: Cultural, economic, and political change in 43 societies*. Princeton university press.

Inglehart, R. and Welzel, C. (2005). *Modernization, cultural change, and democracy: The human development sequence*. Cambridge University Press.

Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., and Porter, M. A. (2014). Multilayer networks. *Journal of complex networks*, 2(3):203–271.

Koltko-Rivera, M. E. (2004). The psychology of worldviews. *Review of general psychology*, 8(1):3–58.

Kroeber, A. L. and Kluckhohn, C. (1952). Culture: A critical review of concepts and definitions. *Papers. Peabody Museum of Archaeology & Ethnology, Harvard University*.

Kroeber, A. L. and Parsons, T. (1958). The concepts of culture and of social system. *American Sociological Review*, 23(5):582–583.

Lee, D.-J. (1998). The effect of cultural distance on the relational exchange between exporters and importers: the case of australian exporters. *Journal of Global Marketing*, 11(4):7–22.

Lévi-Strauss, C. (1987). *Introduction to the work of Marcel Mauss*. Taylor & Francis.

Maslow, A. H. (1981). *Motivation and personality*. Prabhat Prakashan.

McLean, P. (2016). *Culture in networks*. John Wiley & Sons.

Messer, S. B. (1992). A critical examination of belief structures in integrative and eclectic psychotherapy. In *Handbook of psychotherapy integration*, pages 130–165. Basic Books.

Miroshnik, V. (2002). Culture and international management: a review. *Journal of management development*.

Nakata, C. and Sivakumar, K. (1996). National culture and new product development: An integrative review. *Journal of Marketing*, 60(1):61–72.

Newman, M. (2018). *Networks*. Oxford university press.

Olivier, J., Thoenig, M., and Verdier, T. (2008). Globalization and the dynamics of cultural identity. *Journal of international Economics*, 76(2):356–370.

Pepper, S. C. (1942). *World hypotheses: A study in evidence*, volume 31. University of California Press.

Pieterse, J. N. et al. (2019). *Globalization and culture: Global mélange*. Rowman & Littlefield.

Rapoport, H., Sardoschau, S., and Silve, A. (2020). Migration and cultural change. *CESifo Working Paper*.

Reckwitz, A. (2002). Toward a theory of social practices: A development in culturalist theorizing. *European journal of social theory*, 5(2):243–263.

Salant, T. and Lauderdale, D. S. (2003). Measuring culture: a critical review of acculturation and health in asian immigrant populations. *Social science & medicine*, 57(1):71–90.

Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*, volume 25, pages 1–65. Elsevier.

Schwartz, S. H. (2008). Cultural value orientations: Nature and implications of national differences. *Moscow: Publishing house of SU HSE*.

Schwartz, S. H. and Bilsky, W. (1987). Toward a universal psychological structure of human values. *Journal of personality and social psychology*, 53(3):550.

Sivakumar, K. and Nakata, C. (2001). The stampede toward Hofstede's framework: Avoiding the sample design pit in cross-cultural research. *Journal of International Business Studies*, 32(3):555–574.

Smith, A., Calder, C. A., and Browning, C. R. (2016). Empirical reference distributions for networks of different size. *Social networks*, 47:24–37.

Taras, V., Rowney, J., and Steel, P. (2009). Half a century of measuring culture: Review of approaches, challenges, and limitations based on the analysis of 121 instruments for quantifying culture. *Journal of International Management*, 15(4):357–373.

Throsby, D. (2001). *Economics and culture*. Cambridge university press.

Tomlinson, J. (2012). Cultural imperialism. *The Wiley-Blackwell Encyclopedia of Globalization*.

Towse, R. and Hernández, T. N. (2020). *Handbook of cultural economics*. Edward Elgar Publishing.

Tylor, E. B. (1871). Primitive culture: Researches into the development of mythology, philosophy, religion, language, art and custom. *NY, US: Henry Holt and Company. xii*.

Van Wijk, B. C., Stam, C. J., and Daffertshofer, A. (2010). Comparing brain networks of different size and connectivity density using graph theory. *PloS one*, 5(10):e13701.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press.

Weber, M. (1904). Die "objektivität" sozialwissenschaftlicher und sozialpolitischer erkenntnis. *Archiv für sozialwissenschaft und sozialpolitik*, 19(1):22–87.

Networks and symbols

A network in its mathematical formulation is described by a graph $G = (V, E)$, where V is the set of vertices (nodes) and E is the set of edges (links) between them. They can be organized in a matrix way and in a graphical way. In the first case, we can imagine the V set of nodes organized in rows and columns, while the values of the E set of edges provide to fill the cells. In the second case, nodes are simply represented by points, while edges by lines that connect them.

The vertices of a network can be of various kinds, from individuals, animals, economic agents, countries, genes, neurons, etc., and they can be organized in one set (one-mode networks) or two sets (bipartite networks). In our context we will only take into account one-mode networks. With regard to this, since the arguments and methods used in this thesis are various, often different nomenclatures are associated with the nodes of the networks. Basically, in Chapter 3 the number of vertices are outlined by v and the number of edges with e. In Chapter 1, the number of nodes coincides with the number of cultural traits p, while in Chapter 2 it coincides with the number of countries r, where the matrices of distances among countries are considered as networks.

What characterizes a graph are the attributes associated with edges. They, in the more classical sense, attest the *presence/absence (1/0)* of a link between nodes, but they can contemplate the *direction* of the edges (e.g., in trade networks there are exporter and importer countries), the *weights* which describe the strength of the links, or the *signs* which usually outline the positive or negative meaning of the edges. The latter together with weights, in the case of Chapter 1, will highlight the type and strength of relationship between two nodes (cultural traits). In Chapter 2 we will mostly make use of weighted networks, where the weights suggest the distance between two countries. Finally, in Chapter 3 we refer to binary networks.

Here, there are some useful definitions concerning networks:

- *Adjacency matrix*: matrix organization of a one-mode network.
- *Affiliation matrix*: matrix organization of a bipartite network.
- *Adjacent*: two vertices are adjacent if they share an edge.
- *Path*: a sequence of nodes and edges (all different one from each other).

- *Walk*: a sequence of nodes and edges (even if not different one from the others) that describes a path in the graph.
- *Cycle*: a closed path in which each edge and each node are included once and only once in a sequence, except for the origin node.
- *Neighborhood*: the set of nodes adjacent to a generic node.
- *Isolate*: a vertex not connected with any other vertex, namely without neighborhood.
- *Connected graph*: the network has no isolate nodes.
- *Bridge*: an edge that, if deleted, disconnects the graph.
- *Geodesic*: the shortest path between two nodes.
- *Dyad*: pair of nodes and the possible edge among them.
- *Triad*: triple of nodes and the possible edges among them.
- *Motif*: configuration of nodes and the possible edges among them. It is often associated with subsets of 4 or 5 nodes, but implicitly even dyads and triads are motifs.
- *Group*: generic subset of nodes and their edges, over which we measure some properties.
- *Subgraph*: a selection of nodes contained in the overall set V of nodes, and their connections.
- *Clique*: maximum complete subgraph of three or more nodes.
- *Component*: subgraph connected inside but disconnected with the other subgraphs.

As said, the composition of the set of edges is fundamental for defining the structure of the network. As follow are listed some particular kind of networks:

- *Empty graph*: the set of edges E is empty, there are no connections between the vertices, namely each node is isolate.
- *Complete graph*: the opposite situation of an *empty graph*. The set of edges E is composed by all possible edges. They are $e = v(v-1)$ for a directed network, while $e = \frac{v(v-1)}{2}$ for an undirected network. Each vertex is connected with all the other vertices of the network.

- *Star-like graph*: one of the nodes is connected with all the other nodes, while the remaining $v - 1$ nodes are only connected with that node.
- *Circle-like graph*: Each node is adjacent with other two nodes in the way to form a cycle.
- *Line-like graph*: Each node and each edge form a non-closed path.

Another requirement is the nomenclature used in the description of the Subgroup Discovery. As we will see, considering the categorical nature of the network descriptors given by the discretization of Chapter 3, we will denote - using the same symbols of the categorial/ordinal cultural traits of Chapter 1 - with p the number of the generic attributes (in our case network descriptors) and with c_i the generic categories of them.

Chapter 1

The network component of countries' cultural distances

Abstract

The cultural background plays an important role in determining the socio-economic status of a country and its characterization in terms of similarity to other countries. This Chapter makes use of data from the WVS/EVS Joint 2017 to operationalize a definition of culture that takes into account the interdependencies between cultural traits at country level, and calculates a new measure of cultural distance. Taking advantage of a recent Bayesian algorithm by Copula Gaussian graphical model, this Chapter estimates for each of 76 countries included in the WVS/EVS Joint 2017, the cultural network of interdependencies between cultural traits. After defining the distances between countries considering both cultural networks and distributions of cultural traits, it observes via DISTATIS how the addition of this component to the classic distributional one, substantially modifies the measure of cultural distance both in the case of a few cultural traits (6, 10 and 14) and in the case of larger dataset (60). It concludes that the network structure of the national culture matters for the definition of the cultural distance among worldwide countries.

1.1 Introduction

Culture and *sculpture*, two words that seem so similar but that derive from two different Latin words, the first from "colere" which means to cultivate, the second from "sculpere" which means to sculpt. It is immediate to think of culture as a piece of marble that the evolution has continuously sculpted (and continues to do so), which offers itself both to an overall view and to the examination of the most minute details. This dualism extends to the whole wide range of studies and applications inherent in the cultural sphere and, for example, it is found inside the quantitative and qualitative framework.

The quantitative analysis for the study of culture is comparable to the overview of the sculpture. It is synthetically defined in Corbetta (2014) as the study of

collected statistical data, which for their characteristics are structured and provide the necessary support to draw *general conclusions* from the research[1]. From the work of Hofstede (1984)[2], the quantitative streak, in the form of cultural measurement, has taken a very specific path well summarized in Taras et al. (2009): culture is popularly investigated at an aggregate dimension level, placing countries as synonymous with culture (then, as unit of analysis)[3] (Schaffer and Riordan, 2003) and using self-report questionnaires (Hofstede, 1984) or face-to-face interviews (Inglehart, 1997) to collect data. The approach of the World Values Survey and the work of Inglehart and Welzel in positioning countries on their Cultural Map, is also a child of this way of proceeding. In this context, the main researchers' effort was spent for unpacking the culture into dimensions and, simultaneously, to position and categorize countries with respect to them[4].

On the other hand, the qualitative analysis for the study of culture can be compared to the examination of the minute details of the sculpture. In general, a qualitative survey is less structured and aims to go deeper into the topic in question by gathering information about the impressions, opinions, points of view, motivations, thoughts and attitudes of a small number of individuals (Patton 2005, Silverman 2020, Flick 2018). Often the results are not generalizable, but they provide pills of knowledge to broaden the visual spectrum about cultural phenomena. At the same time, talking about the qualitative study of sociologists, anthropologists and psychologists, we especially refer to definitions about the global concept of culture. In this context, the cultural unity of analysis is quite abstract. Following the well-known procedure of Sampson and Laub (1995), to move from hypothetical theoretical definitions[5] to inquire into quantitative measurement and come back to the theory, is a very common practice in the literature.

The common thread of both approaches is the objective of definition/measurement the concept of culture, in its general formulation, about individuals in a definite geographical area, for small groups of people or for individual subjects[6]. In this

[1] For further details on the definition of quantitative analysis see the authoritative works of Cohen et al. 2017 and Creswell and Creswell 2017.

[2] The book was firstly published in 1980. As mentioned in (Taras et al., 2009), it represents a precursor to the quantitative study of cultural phenomena and was immediately followed by other works in this area.

[3] Based on WVS data, in Minkov and Hofstede (2012) is shown how internal regions of East and Southeast Asia, sub-Saharan Africa, Latin America, and the Anglo world overwhelmingly cluster along national lines on basic cultural values.

[4] Besides the work of Hofstede (2011) and Inglehart and Welzel (2005), an important contribution to this research area was provided by Schwartz (2008), House et al. (2004) and Trompenaars and Hampden-Turner (2011).

[5] Among the most popular definitions, one of the first wider is provided by Tylor (1871), after him Weber (1904) and Durkheim (1912) shown two different points of view, while more recent definitions have been argued by Ferrarotti (1986), Lévi-Strauss (1987), Giddens (2000) and Bourdieu (1977).

[6] Mahoney and Goertz (2006) like Brady and Collier (2010) believe that *"qualitative and quantitative scholars share the overarching goal of producing valid descriptive and causal inferences"*.

sense, inside the empirical studies on culture, measurement doesn't exist without definition and definition doesn't exist without measurement, indeed the two fundamental issues are strictly related.

The starting point of cultural measurement is the collection of data. The complexity of the concept of culture - clear from the theoretical definition of Tylor (1871), where culture is broadly defined as *"that complex whole which includes knowledge, belief, art, morals, law, custom, and any other capabilities and habits acquired by man as a member of society"* - inevitably spills over into the composition of the survey questionnaires. Regardless of their level of territorial aggregation[7], each survey usually prepares a particularly long questionnaire, aimed at investigating all the globally recognized cultural aspects[8], containing the specific cultural traits (questions). The aforementioned binomial definition/measurement (theoretical/operative) of culture, then, is triggered in the synthesis of the information coming from the collection of individual data in the surveyed countries.

As far as the theoretical definition issue is concerned, Kroeber and Kluckhohn (1952) offers a wide collection of definitions as evidence that the meaning of culture is something highly questionable in literature. They distinguish seven typologies: descriptives, historical, normative, psychological, structural, genetic and incomplete. However, this collection only contains the definitions up to the time of its writing, while recently, scholars[9] have been proposed further definitions. Taking all the definitions together, Taras et al. (2009) identify four common elements present in virtually all of them: culture is a complex multi-level construct[10]; culture is formed over a relatively long period; culture is relatively stable; culture is shared between individuals. This sharing characteristic of culture is revealed in the constitution of a system of values among the individuals of a community (in our case a country), which automatically defines its cultural status.
In the way of De Benedictis et al. (2021) - which defines culture as the set of local norms, customs, attitudes, values, and their subsets and *interactions* - our contribution to this debate is to point out and confirm the fundamental role in characterizing specific cultures (defined by their cultural system of values) that,

[7] See note 2 in the Introduction of the thesis.
[8] Identified by the sections of the questionnaire. Those of the World Values Survey are detailed in the next Section.
[9] Claude Levi-Strauss (Lévi-Strauss, 1987), Talcott Parsons and Alfred Louis Kroeber (Kroeber and Parsons, 1958), Anthony Giddens (Giddens 2000, Baecker 1997), Pierre Bourdieu (Bourdieu, 1977), Franco Ferrarotti (Ferrarotti, 1986), and so on. See the Introduction of the thesis for more details.
[10] Culture is metaphorically compared to an onion. For the "onion" diagram see Hofstede (1984).

given any set of cultural traits, is played by the interdependence among them[11].

On the operationalization side of things, namely on the cultural measurement, culture have been measured in different ways. As mentioned above, quantitative sociologists focus on construct dimensionalization. Shalom H. Schwartz proposed two main dimensions: Egalitarianism vs Hierarchy, Autonomy vs Embeddedness (Schwartz 2008, Schwartz and Bilsky 1987). Geert Hofstede's work led to six cultural dimensions: Power Distance Index, Individualism vs Collectivism, Masculinity vs Femininity, Uncertainty Avoidance Index, Long Term Orientation vs Short Term Normative Orientation, Indulgence vs Restraint (Hofstede, 2011). Even eighteen dimensions are contained within the GLOBE model, of which nine are due to cultural values and nine are due to practices. Whilst, Inglehart and Welzel, in the Cultural Map affirm the preponderance of two dimensions, which are in parallel with those of Shalom H. Schwartz: Secular vs Traditional, Survival vs Self-Expression (Inglehart and Welzel, 2005). Therefore, to operationalize the cultural definition we need to choose, firstly the source of data from the many opportunities provided by the current literature, then the cultural traits collected from it and finally an appropriate statistical methodology.

This Chapter, in the way of the above contributions, adopts a cross-country perspective: making use of data from the WVS/EVS Joint 2017[12] and follow De Benedictis et al. (2021), first, it defines and quantifies each national system of values as network, second, it measures the distance among countries' cultures considering both, their characteristics on the selected cultural traits, and the notion of interdependence among cultural traits[13]. At the data choice step, in reference to De Benedictis et al. (2021), beyond updating the source of data from the Wave 6 of WVS to the latest WVS/EVS Joint 2017[14], it extends the analysis to four subsets of the data with increasing sizes: 6, 10, 14 and 60 cultural traits. Whereas, at the methodology stage, considering country's cultural traits as nodes and their interdependence as edges of a network structure, their quantification is obtained by exploiting the possibilities offered by Copula graphical model (Lauritzen, 1996) under a newly Bayesian approach (Mohammadi et al.,

[11] In Taras et al. (2009) is explicitly indicated as *"in the context of culture measurement, the problem of faulty equivalence assumptions is not limited to generalizability across levels of measurement, but also refers to generalizations across cultures. A relationship between variables found in one culture may not be generalizable to other cultures"*.

[12] However it is even plausible that the main finding could be achieved by processing data from other surveys (this remains as future replication works).

[13] The punctual delimitation of the main purpose is fundamental. Because the vastness of the topic, Taras et al. (2009) warns of dangerous generalization when one studies cultures.

[14] EVS/WVS (2021). European Values Study and World Values Survey: Joint EVS/WVS 2017-2021 Dataset (Joint EVS/WVS). JD Systems Institute & WVSA. Dataset Version 1.1.0, doi:10.14281/18241.11.

2017), as well as De Benedictis et al. (2021) does. A new cultural distance index (like the *network index*) is finally found putting together the De Benedictis et al. (2021) procedure within the DISTATIS methodology (Abdi et al., 2005), suitable for the joint analysis of several distance measures.

The Chapter is organised as follow. Section 1.2 provides the description of the World Values Survey framework, it re-makes the Inglehart-Welzel Cultural Map considering only data from the latest WVS/EVS Joint 2017 and selecting questions for the analysis implemented in the rest of the Chapter. Section 1.3 defines the empirical definition of culture that this Chapter wants to verify, namely culture is composed by two components: one is due to the distributional content of the cultural traits, and another to the cultural traits interdependencies. Section 1.4 infers the national cultural traits interdependencies (the cultural networks) to be considered in the final cultural distance measure. After Section 1.5 defines the distance measures applied on the different theorized components of culture, Section 1.6 finally finds a new index of cultural distance between countries.

1.2 Data: WVS/EVS Joint 2017

For almost 40 years the World Values Survey Association (WSVA)[15] through the WVS has been contributing to the study of worlwide cultural changes and how they can derive or affect economic and social phenomena. The original purpose was to study how economic development is evolving the values and attitudes of individuals within national societies and it was fielded with the first Wave (1981-1984) inspired by the European Values Study (EVS)[16]. Asking to a national sample of citizens from 11 worldwide countries to answer a questionnaire about socio-demographics (gender, age, level of education, income), religion, family, life habits, politics, generical values of life, etc., and deeply analyzing the collected data, researchers found substantial differences between the values of younger and those of older connected to the economic growth experienced by society in which they live.

Understanding the extent of information that the survey may have in explaining the change in values from generation to generation based on economic and tech-

[15] As reported in the official WVS website https://www.worldvaluessurvey.org/WVSContents.jsp *"The WVSA is a non-commercial non-governmental international social research organization with the Legal Seat in Stockholm, Sweden...The mission of the WVSA is to contribute to a better understanding of global changes in values, norms and beliefs of people by the means of comparative representative national surveys worldwide – known as the World Values Survey (WVS)"*.

[16] The first Waves of WVS were Eurocentered with small representatives of Southeast Asia and African countries.

nological developments, seven Waves have been carried out approximately every five years[17]. They include societies that cover the widest range of attitudes and situations such as different economic levels, different geographical areas, different climate areas, different cultural traditions.
As the years passed, the questionnaire has been subject to changes and adaptations to cover all the aspects of culture and improving the measurement (Taras et al., 2009) by accepting the suggestions of social scientists from all over the world. Currently, the 290 questions are divided in distinct different macro-categories: social values, attitudes and stereotypes; happiness and well-being; social capital, trust and organizational membership; economic values; corruption; migration; security; postmaterialist index; science and technology; religious values; ethical values and norms; political interest and political participation; political culture and political regimes; demographics. Questionnaire is translated and adapted[18] according to the country and it is sometimes pre-tested to assess the critical issues. Most interviews take place face-to-face using paper or CAPI (Computer Assisted Personal Interview) to record answers, and sometimes via phone for remote areas. The minimum sample per country is 1200 individuals and it is representative of all inhabitants over the age of 18[19].

Since the WVS is inspired by the EVS, considerations made for the former can be extended to the latter, however, EVS is performed on a regular basis of nine years and focuses only on European countries. Since 1981 it has managed to cover almost all of the Europe territories, from Azerbaijan to Portugal, from Sweden to Greece, becoming one of the most important sources of secondary data for researchers who focus their study on this geographical area. For the last survey[20] the sample size was set to 1200 individuals for countries with population over 2 million and to 1000 individuals for countries with population below 2 million, while the sampling methods since 2008 are only probabilistic[21].

[17] Many important macro-developments of the society have been highlighted over the years by the work of the researchers of the World Values Survey. For example, the complex mechanism that regulates political, cultural and economic changes based on data of the Wave 2 is described in Inglehart (1997). The profound changes in the traditional of gender roles are examined in Inglehart and Norris (2003) through data of Wave 3. While, Inglehart and Welzel (2005), summing up from the previous surveys, shows as modernization is a process of human development, in which economic development triggers cultural changes that make individual autonomy, gender equality and democracy increasingly likely.

[18] The translation process, as Taras et al. (2009) highlights, is fundamental to keep under control the error created during data collection. This is not a small problem, in fact when the survey concerns specific regions of a country, the presence of dialects and different facets of language, especially in rural areas, can generate misunderstandings. Furthermore, some populations may be more sensitive to a question than others, this often leads to the elimination of some questions for some countries.

[19] The national representative random sample is based on multi-stage territorial stratified selection. The first stage is the selection of primary sampling units (PSUs), the second is the random sampling of individuals, e.g. the household. For details see Haerpfer et al. (2020).

[20] European Values Study 2017: Integrated Dataset (EVS 2017). GESIS Data Archive, Cologne. ZA7500 Data file Version 4.0.0, doi:10.4232/1.13560.

[21] For a description of sampling methods see the European Values Survey: method report. European Values Study (EVS). (2020). European Values Study (EVS) 2017: Method Report. (GESIS Papers, 2020/16).Köln. https://doi.org/10.21241/ssoar.70109.

Data used to point out the importance of the *interdependence* between cultural traits in defining the cultural background, which characterizes the set of individuals of a country, come from the latest survey concerning World Values Survey (WVS) jointly to European Values Study (EVS), and available publicly and freely on the WVS website. The two organizations agree to collaborate in carrying out a single, shared EVS/WVS survey[22]. Both the EVS and WVS remain and operate as independent research organisations, EVS coordinates the data collection process in Europe, while WVS does so for the rest of the world. The questionnaires are developed independently of each other, but have common traits that converge towards the joint survey.
In the four-year period 2017-2020, the WVS, in its Wave 7, has completed the survey of 70,867 individuals from 49 countries, while the EVS (Wave 5) in the same period completed the survey in 34 countries for a total of 56,491 interviewed individuals. By combining the two surveys, they cover 79 worldwide countries because surveys in Germany, Romania, Russia and Serbia have been conducted by both. A number never historically reached by the WVS.

These numbers, the free access to data and the recent last pubblication (November 2020), the way data are collected, together with the myriad of papers and books that use this data and quote[23] the theories coming from the network of scholars of the WVS project[24], make the WVS and its main result summarised by the Inglehart-Welzel Cultural Map (Inglehart and Welzel, 2005), the major social survey in the world, a reference point for many empirical cultural studies on secondary data[25], and the perfect motivation for using it for the task of this Chapter.

1.2.1 Inglehart-Welzel contribution: Cultural Map

Among the many specific and general works that Ronald Inglehart and Christian Welzel together, alone or with other scholars have carried out on the data coming from the WVS (of which Inglehart is the founder), the Cultural Map is the most striking and well-known application. In the attempt of summarizing

[22] The contract that the two associations have signed is available on the WVS website.
[23] There are more than 30,000 publications in the literature based on the WVS data and over 60,000 citations for WVS in Google Scholar.
[24] The network of social scientists and researchers working in the WVS project includes scholars from 120 world countries.
[25] The range of social disciplines in which the WVS has found application are innumerable. For example, for an effective review of the role of cultures in economics and institutions see Alesina and Giuliano (2015); in the context of cultural economics, for the use in immigrational issues Rapoport et al. (2020); for a use of WVS data in comparative political science Alemán and Woods (2016); for the application on social capital Guiso et al. (2008); for a nice application on public health Jen et al. (2009).

the complexity of the national cultural background (system of values), it places the investigated countries on a bi-dimensional plane of a reduced space (factorial). This allows to understand which countries are similar at the cultural traits level, observing how their positioning varies over time and comparing it to predefined groups. Up to WVS Wave 6 nine different groups were distinguished on the Cultural Map: African-Islamic, Orthodox, Latin America, English-speaking, Confucian, Baltic, Protestant, South Asia, Catholic Europe. Regarding Wave 7, and therefore the WVS/EVS Joint 2017, eight groups (names are little modified) are depicted as the Baltic countries have been included in other groups: African-Islamic, Orthodox Europe, Latin America, English-speaking, Confucian, Protestant Europe, West & South Asia, Catholic Europe.

From the many variables based on the WVS questionnaire, for building the Cultural Map, Inglehart and Welzel selected 10 variables[26]. Following, they are listed by their questionnaire section, subsection, the longitudinal variable id and the Wave 7 question id[27]:

[**Perceptions of life**]

- `HAPPINESS AND WELL-BEING`
  A008 - *Happiness.*
  Q46. Taking all things together, would you say you are

  1. Very happy
  2. Rather happy
  3. Not very happy
  4. Not at all happy

- `SOCIAL CAPITAL, TRUST & ORGANIZATIONAL MEMBERSHIP`
  A165 - *Trust.*
  Q57. Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?

  1. Most people can be trusted

[26] Originally, Inglehart (1997) used factor scores (Cultural Map derives from a Factorial Analysis) based on 22 variables, while Inglehart and Baker (2000) and subsequently Inglehart and Welzel (2005) reduced them to 10 variables, essentially to avoid missing data and an excessive dropping of countries from the analysis. According to the authors Inglehart and Welzel (2005), reducing the number of variables, the information scope is not affected because of the high correlation among some of the original questions considered.

[27] The questionnaire for each WVS Wave is subject to revisions, some questions are removed, other added and sometimes reviewed. For this and for the main organization of the Sections, the order of questions is always different, namely the question id varies Wave by Wave. Both questions, and variables measured by them, are uniquely defined by their longitudinal id, specified in the WVS integrated dictionary codebook.

2. Need to be very careful

[**Politics and Society**]

- SOCIAL VALUES, ATTITUDES & STEREOTYPES
  E018 - *Respect for authority.*
  Q45. I'm going to read out a list of various changes in our way of life that might take place in the near future. Please tell me for each one, if it were to happen, whether you think it would be a good thing, a bad thing, or you don't mind.
  Greater respect for authority:

  1. Good
  2. Don't mind
  3. Bad

- POLITICAL INTEREST & POLITICAL PARTICIPATION
  E025 - *Voice.*[28]
  Q209. Now I'd like you to look at this card. I'm going to read out some forms of political action that people can take, and I'd like you to tell me, for each one, whether you have done any of these things, whether you might do it or would never under any circumstances do it.
  Signing a petition:

  1. Have done
  2. Might do
  3. Would never do

[**Religion and Moral**]

- RELIGIOUS VALUES
  F063 - *Importance of God.*
  Q164. How important is God in your life? Please use this scale to indicate. 10 means "very important" and 1 means "not at all important".

- ETHICAL VALUES AND NORMS
  F118 - *Homosexuality.*
  Q182. Please tell me whether you think homosexuality can always be justified (10), never be justified (1), or something in between.

[28] In the sense of Hirschman (1970).

- ETHICAL VALUES AND NORMS
  F120 *Abortion.*
  Q184. Please tell me whether you think abortion can always be justified (10), never be justified (1), or something in between.

[**National identity**]

- POLITICAL CULTURE & POLITICAL REGIMES
  G006 - *Proud of nationality.*
  Q254. How proud are you to be [nationality]?

  1. Very proud
  2. Quite proud
  3. Not very proud
  4. Not at all proud
  5. I'm not [nationality]

From the combination of the answers received to specific questions, Inglehart (1997) constructed two derived indices.

[**Post-Materialism**[29]]

- Y002 - *Post-Materialism.* Post-Materialist 3-items index

  1. Materialist
  2. Mixed
  3. Postmaterialist

  Y002 is calculated as follows (see the Appendix of Inglehart (1997)):

  – if (E003=1 or E003=3) and (E004=1 or E004=3) then 1 (Materialist)
  – if (E003=2 or E003=4) and (E004=2 or E004=4) then 3 (Postmaterialist)
  – else if ((E003=1 or E003=3) and (E004=2 or E004=4)) or ((E003=2 or E003=4) and (E004=1 or E004=3)) then 2 (Mixed)

  where:

  – POSTMATERIALIST INDEX
    E003 - *Social values.*
    Q154. If you had to choose, which one of the things on this card would you say is most important? (first choice):

[29] In sociology, Postmaterialism is the transformation of individual values from materialist, physical, and economic to new individual values of autonomy and self-expression (Inglehart, 2018).

1. Maintaining order in the nation
2. Giving people more say in important government decisions
3. Fighting rising prices
4. Protecting freedom of speech

– `POSTMATERIALIST INDEX`
E004 - *Social values 2.*
Q155. And which would be the next most important? (second choice):

[**Autonomy**]

- Y003 - *Independence/Obedience.* Autonomy index: from Determination, Perseverance/Independence (-2) to Obedience/ Religious Faith (2)

  To calculate Y003[30] Inglehart and Welzel (2005) used Children qualities battery questions inside the `SOCIAL VALUES, ATTITUDES & STEREOTYPES` subsection: it is more important for a child to learn obedience or faith than independence and determinations? A040 - Q15. Religious Faith; A042 - Q17. Obedience; A029 - Q8. Independence; A039 - Q14. Determination, Perseverance.[31]

From these variables, they extract the final set of data over which they apply the Factorial Analysis for projecting countries in the reduced space. The procedure provides for the calculation of the average value for each variable and each country (Inglehart, 1997). Negative values which code answers "Don't know", "No answer", "Not applicable", "Not asked in survey", "Missing; unknown", are considered missings, as result any case without answers to the 10 variables may be skipped, causing in certain circumstances the deletion of countries with no scores in the Cultural Map[32].

In an effort to create a basis for the work of this Chapter, that can serve both as a starting point and as a comparison for our cultural distance index, in the way of De Benedictis et al. (2021), that has replicated the Cultural Map for the only data from Wave 6, we do the same for data from the WVS/EVS Joint 2017.
We apply a Principal Component Analysis (PCA) over the 10 cultural traits (Inglehart and Welzel, 2005) selected by Inglehart and Baker (2000) with a varimax[33]

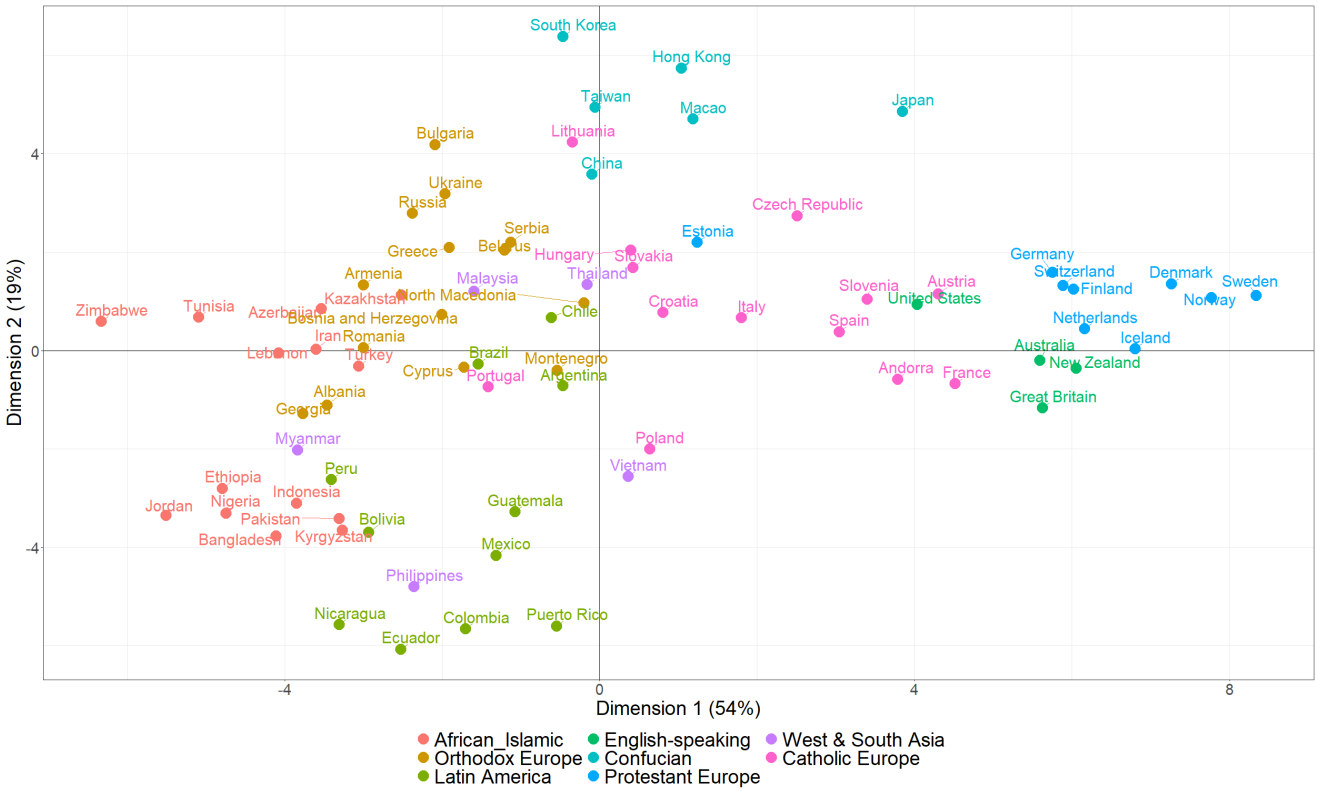
[30] The original support of the variables Y003 was [-2,2]; in our analysis we re-coded it on the support [1,5]. In our case, using longitudinal aggregate coding, 1 indicates Determination, Perseverance/Independence and 5 Obedience/Religious Faith, for the individual Wave the scale is the opposite.

[31] For further details about the autonomy index see Inglehart and Welzel (2005). Details on the variable Y003 are in the Appendix of Inglehart (1997) and on the WVS Website, where it is the described the way in which it is built.

[32] For specific situation, some questions are not asked in the questionnaire of certain countries.

[33] For the theoretical definition of the well-known PCA method see Abdi and Williams (2010) and for its application on R software Husson et al. (2017). Varimax rotation is the rotation which maximizes the sum of the squares of the loadings (Husson et al., 2017).

Figure 1.1: Inglehart-Welzel Cultural Map



Note: Our elaborations on WVS/EVS Joint 2017 data. Colors correspond to Inglehart and Welzel (2005) groups, where from this Wave Baltic is ousted (Baltic countries are included in other groups). The horizontal axis corresponds to the first principal component of the PCA on the 10 variables associated to the cultural traits considered in the analysis. The vertical axis corresponds to the second principal component.

rotation of the axes. Differently from De Benedictis et al. (2021) - where did not make use of varimax rotation - the conformation of the map is proportionally more faithful to that of Inglehart-Welzel[34] published on the WVS Website. At any rate, small inconsistencies are still found for various practical reasons: we do not use the merged WVS/EVS longitudinal data from 2005 (from the WVS Wave 5), but we use only the WVS/EVS Joint 2017; we do not use SPSS software[35] for our elaborations, but R software. These differences even return factorial axes that are more explicative of the total inertia of the average matrix cultural traits per countries (73% vs. 71% of the Cultural Map).

All in all, the interpretation of the dimensions (factorial axes) can be assimilated to that of Ronald Inglehart and Christian Welzel, and having no ambition to discuss and argue about them, we take for good the specific sociological in-

[34] See the WVS Website for the latest published including the new data from WVS and EVS.
[35] The SPSS code is available in the WVS official website.

terpretation provided by the two authors: the first factor describe the two poles *Survival values* versus *Self-expression values*, while the second *Traditional values* versus *Secular-rational*[36] (Inglehart and Welzel, 2005).

1.2.2 Selected questions/variables

In De Benedictis et al. (2021) the set of data is selected according to the Inglehart-Welzel Cultural Map, while the analysis object of this Chapter is carried out from four subsets of variables of increasing size, coming from the WVS/EVS Joint 2017 data. Keeping in mind that Inglehart (1997) have already experimented with building their Cultural Map with more than 10 variables, here, we are interested in observing any changes in the nature of the interdependencies between variables as the number of selected cultural traits (nodes) increases. Basically, we want to analyze the role of the cultural network in measuring the national system of values and the cultural distances between countries when a more complex and informative network structure is depicted preserving as much information as possible (namely, avoiding as much as possible the presence of missing values).

Cultural Map implicitly measures the cultural distance between countries. In the way of De Benedictis et al. (2021), to meet the primary purpose to compare our results with those of the Cultural Map, the first selected subset of data is composed by the same variables used to construct it, and that we listed in the paragraph 1.2.1. This inevitably leads to delete Egypt, Tajikistan and Iraq from the 79 worldwide countries of the WVS/EVS Joint 2017. Answers of the individuals contained in their samples result entirely missing for questions described by variables F118[37] (Egypt and Tajikistan) and F063[38] (Iraq).

On this basis, two further datasets of increasing size are considered. The first is composed by 14 cultural traits and includes the same of the first dataset above,

[36] According to the official WVS website the two dimensions can be interpreted as follow: 1) In the first factor the two extremes symbolize *survival* vs *self-expression values*. *"Survival values give emphasis to economic and physical security and are associated to low levels of trust and tolerance. Self-expression values give high priority to subjective well-being, individualism and quality of life (environmental protection)."* A rightward movement from survival to self-expression also represents the transition from industrial society to post-industrial society, as well as the embracing of democratic values. 2) In the second factor the two extremes symbolize *traditional* vs *secular values of societies*. *"Traditional values emphasize the importance of religion, parent-child ties, deference to authority, absolute standards and traditional family values. People who embrace these values also reject divorce, abortion, euthanasia and suicide. Societies that embrace these values have high levels of national pride and a nationalistic outlook."* Consequently, at the opposite vision there are Secular-rational values. A downward shift represents a movement away from traditional values and toward more secular-rational values.

[37] It is the unique variable ID used merging WVS and EVS, called Q182 in the WVS Wave 7 questionnaire and V203 in the WVS Wave 6 questionnaire. The formulation is: "Please tell me for each of the following actions whether you think it can always be justified, never be justified, or something in between, using this card (Homosexuality)".

[38] Called Q164 in the WVS Wave 7 questionnaire and V152 in the WVS Wave 6 questionnaire. The formulation is: "How important is God in your life? Please use this scale to indicate. 10 means very important and 1 means not at all important".

replacing the indices Y002 and Y003 with their disaggregation by the variables from which they are calculated (see paragraph 1.2.1). Taking up to this, the second integrates this latter dataset with further 46 variables, for a total of 60 variables. Finally, the last dataset contains variables derived from the first battery of questions inside the questionnaire[39]. Summing up based on increasing size, from this moment datasets will be called in this way:

- *Battery dataset*: variables from the first battery of questions.
- *IW dataset*: Inglehart-Welzel Cultural Map variables.
- *IW1 dataset*: Inglehart-Welzel Cultural Map variables (disaggregation of Y002 and Y003).
- *Large dataset*: Inglehart-Welzel Cultural Map variables (disaggregation of Y002 and Y003) plus further 46 selected variables.

While variables from Inglehart-Welzel Cultural Map (Inglehart and Welzel, 2005) are widely accepted in the literature[40] and have been extensively discussed in the previous paragraph, we have no further theoretical prior on which cultural traits are most important for the national culture definition, therefore practical reasons lie behind of the choice of the *Battery* and *Large* subsets.

Straightforwardly, variables of *Battery dataset*, as first, describe a very general and transversal construct (Priorities in Life), as second, they have an high response rate[41] because their generality and because they are the first questions of the questionnaire.

About the choice of the 46 variables that complete together with the *IW1 dataset* variables the *Large dataset*, it is strictly guided by the methods to infer the interdependencies among cultural traits and the least loss of information. A first skimming is due to the overlapping questions of WVS and EVS, that considerably reduces the number of cultural traits derived from the individual answers (189 questions). Anyway, because we have to estimate a network for each country and each dataset, the computational cost of the Bayesian inferential scheme,

[39] Priorities in Life battery questions inside `SOCIAL VALUES, ATTITUDES & STEREOTYPES` subsection: "For each of the following, indicate how important it is in your life. Would you say it is: 1. Very important; 2. Rather important; 3. Not very important; 4. Not at all important." A001 - Q1. Family; A002 - Q2. Friends; A003 - Q3. Leisure time; A004 - Q4. Politics; A005 - Q5. Work; A006 - Q6. Religion.

[40] In addition to the astonishing numbers of publications and citations already cited in note 23, this is highlighted also in the important reviews Taras et al. (2009) and Alesina and Giuliano (2015).

[41] For this reason is used also in Rapoport et al. (2020). In this case it is included in a model of cultural transmission over time to examine how migration affects cultural change in home and host countries.

that characterizes the newly algorithm of Copula Gaussian graphical models[42] (Mohammadi et al., 2017), is too high to include all of them. By adding cultural traits the combinations of graphs in which they can be arranged increases exponentially and with it also the processing time to reach the convergence of the algorithm. Similarly, the probability of having to remove more countries from the analysis also increases because more variables imply more probability to observe patterns of missing values[43]. Following the instructions of the author of the method (Abdolreza Mohammadi), in order to have good performances, we set an hypothetical limit at 100 cultural traits to be included in the analysis. Considering again that our analysis is not inherent to a single network, but estimates the cultural network of 76 countries and 4 different datasets, we try to further reduce this limit in order to have a substantial group of cultural traits, but not to burden the timing of the analysis. With this purpose, for each cultural trait we calculate the distribution of missing values per country. Then, we associate each cultural trait with the maximum of its distribution of missing values over countries[44]. From this new distribution we take the cultural traits which stand under the median, namely 46 cultural traits.
The next step is the choice of the number of iterations to run for reaching the convergence of the algorithm. There is no way to assess it without having to run the inferential procedure of the algorithm at least for o e country. For this reason we empirically use as benchmark the estimation of the Italy's cultural network that may have a fairly complex cultural system of values. We run the algorithm under different number of iterations. Table 1.11 in Annex 1 shows how a good solution that combine convergence and processing time is 300,000 iterations. Apart from the variables of the *IW1 dataset*, as evidence of the high response rate for the questions of the first battery (Priorities in Life), they are contained in the 46 selected variables, outlining the *Large dataset* as the union of the *IW1 dataset*, the *Battery dataset* and other 40 variables derived from questions with a high response rate over countries (see Table 1.13 in Annex 1 where they are listed).

1.3 Empirical definition of culture

The primary objective of the empirical analysis of Inglehart and Welzel (2005), or Hofstede (2011), or more generally of all the scholars inside the research field reviewed by Taras et al. (2009), is of unpacking the complex concept of culture

[42] It is introduced in Section 1.4 and fully described in Annex 1.
[43] This was the same reason why Inglehart and Baker (2000) decided to reduce the 22 variables to 10 of them.
[44] In other words, we describe each cultural trait with the maximum value of missing that it presents for a country. The higher is this values, the higher information we lose for one or more countries.

in a few important dimensions, on the basis of which to evaluate the aggregate national cultural peculiarities, and therefore the cultural distance between countries. The starting point of these works is the operation of averaging the national culture, that automatically implies two issues: as Taras et al. (2009) highlights, this obscures some of the information contained within the distributions of each cultural trait for each country, and at the same time, it requires that the structure of relations between cultural values be constant for all the countries, that is, not contextualized but universally generalizable.
As De Benedictis et al. (2021), in the attempt of exceeding these limits, and recognizing the complexity of the concept of culture already clear inside the main theoretical definitions (Guiso et al. 2006, Ferrarotti 1986, Tylor 1871, Bourdieu 1977), we imagine the national cultural system of values can be composed by two fundamental objects:

- The more or less polarized entity of cultural traits, represented by their distributions.
- The interdependence between cultural traits, identified by the cultural network, where cultural traits are intended as nodes and their mutual associations inside the national collective as edges.

Including the measurement of interdependencies between cultural traits can improve the accuracy of cultural distance measurements. Finding that two countries are similar, both in terms of the distribution of their cultural traits and the way they are interconnected with each other, certainly supports their measure of similarity. Just as for countries that are dissimilar, their national cultures can be dissimilar because they differ in the distribution of cultural traits and in the way their interconnections form different multifaceted relational structures. However, the interdependencies of cultural traits are crucial for cases in the midst of these two extremes, in fact, instead of corroborating their distance relationship, they can completely reverse it.

Since in this vision the empirical definition of culture is composed by the two complex listed objects, in order to calculate the cultural distance between countries, distances based on distributions and based on the cultural traits interdependence are evaluated separately.

We find the distributional part in two ways: as first, in the way of Inglehart (1997) by averaging each variable for each country, and as second, by following

the literature on Symbolic Data Analysis (Diday and Noirhomme-Fraiture, 2008), specifically by using the symbolic generalization[45] by country as shown in Tables 1.1 and 1.2.

Table 1.1: Example of *Battery dataset*

	Country_iso2	Family	Friends	Leisure time	Politics	Work	Religion
1	AL	2	1	2	3	2	2
2	AL	1	1	4	4	1	1
...	...	...	...	...	...	...	...
50,000	SE	1	1	1	3	2	3
50,001	SE	1	2	1	3	2	3
...	...	...	...	...	...	...	...
123,757	ZW	1	2	2	1	1	1
123,758	ZW	1	2	2	2	1	1

Table 1.2: Example of symbolic generalization for *Battery dataset*

Country_iso2	Family	Friends	Leisure time	Politics	Work	Religion
AL	1(), 2() 3(), 4()	1(), 2() 3(), 4()	1(), 2() 3(), 4()	1(), 2() 3(), 4()	1(), 2() 3(), 4()	1(), 2() 3(), 4()
...	...	...	...	...	...	...
SE	1(), 2() 3(), 4()	1(), 2() 3(), 4()	1(), 2() 3(), 4()	1(), 2() 3(), 4()	1(), 2() 3(), 4()	1(), 2() 3(), 4()
...	...	...	...	...	...	...
ZW	1(), 2() 3(), 4()	1(), 2() 3(), 4()	1(), 2() 3(), 4()	1(), 2() 3(), 4()	1(), 2() 3(), 4()	1(), 2() 3(), 4()

Note: Each category associated to a cultural trait has a certain relative frequencies, ideally represented by ().

The interdependencies between cultural traits are found via the Bayesian inferential scheme developed by Mohammadi et al. (2017) inside the literature of Copula Gaussian graphical modelling. The next paragraph is entirely dedicated to motivate the use of it and to describe the inferred cultural networks derived from it.

[45] Symbolic Data Analysis (SDA) fits into the context of analyzing large datasets that require a first reduction in higher level predefined concepts. Starting from a set of individuals, generalization provides the aggregation of them in 'symbolic objects' (higher level concepts, in our case worldwide countries) characterized by their 'symbolic descriptions'. Basically, a symbolic description is a vector of symbolic variables (lists, histograms, distributions, intervals, and so on) that describe a symbolic object in a boolean or probabilistic way. In our case we create boolean symbolic objects. For more details see Billard and Diday (2003), Diday and Noirhomme-Fraiture (2008), Bock and Diday (2012).

1.4 Cultural traits interdependence

As regards the developments of this Section and more generally of this Chapter, we believe it is sufficient to justify the methodological choice by mentioning some of its key characteristics, whereas for more details about the methodology, the inference and outputs of the procedure, refer to the Annex 1. More in general, we do not intend to make a direct contribution to the method, but rather to use it as a tool to reach empirical evidences.

With an effective analogy, interdependencies among cultural traits for a country are previously defined by its cultural network, where cultural traits are connected according to their relationship embedded in the complex cultural system of value. As mentioned above, in De Benedictis et al. (2021) cultural networks are estimated for each country on the basis of one single dataset, while here, four different subsets of cultural traits are considered, namely four different Graphs will have 6, 10, 14 or 60 cultural traits (nodes) and their set of edges, according to the underlying network structure. Since these connections are not directly observed, in this framework, network inference is performed via Graphical models, individually for each country and subset of data.

The latent network structure of a system described by variables can be of interest when, as said, it is not directly observed, i.e. when the presence/absence of the edges is unknown. In this context, besides the average and marginal cultural characteristics of each country, we assume that its internal cultural system of values - defined by the non-observed network of the important cultural traits - may have an important meaning for the definition of its culture and for the cultural distance between countries.

A simple example that can illustrate firstly the importance of considering the relation between cultural traits (De Benedictis et al., 2021), and secondly the importance of the conditional independence - then of the underlying network structure - for the measurement of the cultural proximity between countries is below. Imagine we live in a world with two countries (country 1 and country 2), and two cultural traits representative of the measurement of their culture. Each of them takes on two values, e.g., *belief in God* (yes/no) and *trust in others* (yes/no). Imagine that in both countries, half the people believe in God and half the people trust others. The two countries would appear to be culturally coincident. However, suppose that in the first country everyone who believes in

God also trusts others (Table 1.3), while in the second country everyone who believes in God does not trust others (Table 1.4). Then, the two countries are actually culturally different because the relation between the two cultural traits across individuals within each country is different.

Table 1.3: Country 1

		TRUST		
		Yes	No	
GOD	Yes	600	0	600
	No	0	600	600
		600	600	1200

Table 1.4: Country 2

		TRUST		
		Yes	No	
GOD	Yes	0	600	600
	No	600	0	600
		600	600	1200

Now imagine the joint distribution of *belief in God* and *trust in others* is the same for the two countries and coincide with Table 1.3. In this case country 1 and country 2 are culturally equivalent both for the distribution of the individual cultural traits and for their interdependence.

Table 1.5: Country 1

PETITION	**GOD**	**TRUST**		
		Yes	No	
Yes	Yes	600	0	600
	No	0	0	0
No	Yes	0	0	0
	No	0	600	600
		600	600	1200

Table 1.6: Country 2

PETITION	**GOD**	**TRUST**		
		Yes	No	
Yes	Yes	300	0	300
	No	0	300	300
No	Yes	300	0	300
	No	0	300	300
		600	600	1200

Consider that in our hypothetical world there is a third important cultural trait: *signing a petition*. It takes on two values (yes/no), and, as the other two, in both countries its distribution is the same: half the people have signed at least one petition in their life. The two countries seems to be culturally equivalent considering the marginal distributions of the three cultural traits and their interdependence except for *signing a petition*. The joint distributions depicted in Table 1.5 and 1.6 show that when the third cultural trait is considered, it acts on the way the traits interact one each other. Then, the two countries are actually culturally different because the pattern of interdependence between cultural traits across individuals within each country is different. Measuring the cultural distance between countries without taking into account the network structure of

cultural traits would therefore result in a systematic downward bias of potentially relevant magnitude.

In this context, Graphical models are useful methods because we are interested in the estimation of a complex structure of relations between a set of variables (cultural traits), which are symbiotically connected to each other to describe the cultural system of values of a country. The latter is defined by a graph, where the absence of interaction - conditional independence - between the cultural traits (nodes) is reflected in the graph as a missing edge between them. They are highly performing both when over a large set of variables are collected or available several observations and when the number of them is in some way limited. Obviously, they remain effective even when the set of variables is not very large, e.g. in our case for *Battery dataset*, *IW dataset* and *IW1 dataset*.

Inside the graphical modelling framework, Gaussian graphical models (Ggm) are popular models for inference of undirected graphs when variables are supposed to be Normal distributed, then when the starting point is the multivariate normal distribution of the nodes (Lauritzen, 1996). They move beyond simplistic approaches based on pairwise correlations and measure dependencies in terms of partial correlations, that is the correlation of two nodes conditional on all the others[46].

Since the responses to the selected questions of the WVS/EVS Joint 2017 are ordinal or categorical, and therefore not normally distributed, Copula Gaussian graphical models (CGgm) will be considered for our analysis. Inside the different approach to CGgm methods , of which one was proposed by Dobra and Lenkoski (2011), for the elaboration of data in our framework we make use of the Bayesian inferential procedure that has been recently developed by Mohammadi et al. (2017).

The procedure of graph inference in a CGgm, in a nutshell, is traditionally made of two tasks: *parameter estimation*, namely the estimation from the empirical marginal distributions of the precision matrix (inverse of the correlation matrix), and *model selection*, that is selecting a Graph where some edges may be missing.

[46] Considering each node as a random variable, Graphical models move their theory from the parallel among the conditional independence between nodes (variables) and sets of nodes (Markov properties) and the joint distribution of them, as demonstrated in the Hammersley-Clifford Theorem (Hammersley and Clifford 1971, see also Lauritzen 1996, theorem 3.9). As a direct implication, this allows to study the reticular structure of the conditional independence of a graph through the joint density function of the variables/nodes.

In a classical frequentist framework, these two tasks are treated separately: given a Graph, the precision matrix is estimated by constrained maximum likelihood, whereas model selection criteria based on the model likelihood and model complexity (in this case number of edges) are subsequently used to select an optimal Graph (Lauritzen, 1996). In contrast to this, a Bayesian approach allows to account simultaneously for uncertainty both at the level of Graph inference and precision matrix estimation. The computational cost that this operation requires is mitigated by the improved efficiency of the algorithms. Among these, there is the one we have chosen for our elaboration. Mohammadi et al. (2015) proposed a fast implementations based on Birth-Death Markov Chain Monte Carlo (BDMCMC) method. Authors demonstrate its statistical performance compared to other algorithms of Ggm (Mohammadi et al., 2015), and they further extended it for the case of CGgm for ordinal and categorical data (Mohammadi et al., 2017). The BDMCMC procedure[47] remarkably improves the performances and returns an innovative output, where the inferred network is described by a set of edges representing the probabilities to observe a link between nodes, we will call later *Posterior Edge Inclusion Probabilities*. Setting a cut-off 0.5 on the posterior edge inclusion probabilities, the *Binary Network* is easily found, while the *Partial Correlation Network* is implicitly provided by every graphical modelling analysis.

On the basis of these outputs, the remaining part of this Section is dedicated to the description of the features about some inferred networks.

1.4.1 Inferred countries networks of cultural values

Inside the network part, for each country, some traits could be symbiotically expressed admitting a significantly positive or negative connection; some others could be not connected because citizens of that country do not relate them directly; finally national cultural may lack of one or more cultural traits, these are values that are not significantly taken into account by the people of that country.

The approach described in Annex 1 and just mentioned above, is applied for each country and subset of variables. For investigating interdependencies among 6 (*Battery dataset*), 10 (*IW dataset*), 14 (*IW1 dataset*) and 60 (*Large dataset*) cultural traits, we set respectively 10,000, 10,000, 20,000 and 300,000 iterations for the BDMCMC algorithm. These numbers of iterations - which exponentially increase with the number of variables/nodes - has been considered sufficient to

[47] The method is implemented in the R package `BDgraph` (Mohammadi and Wit, 2019).

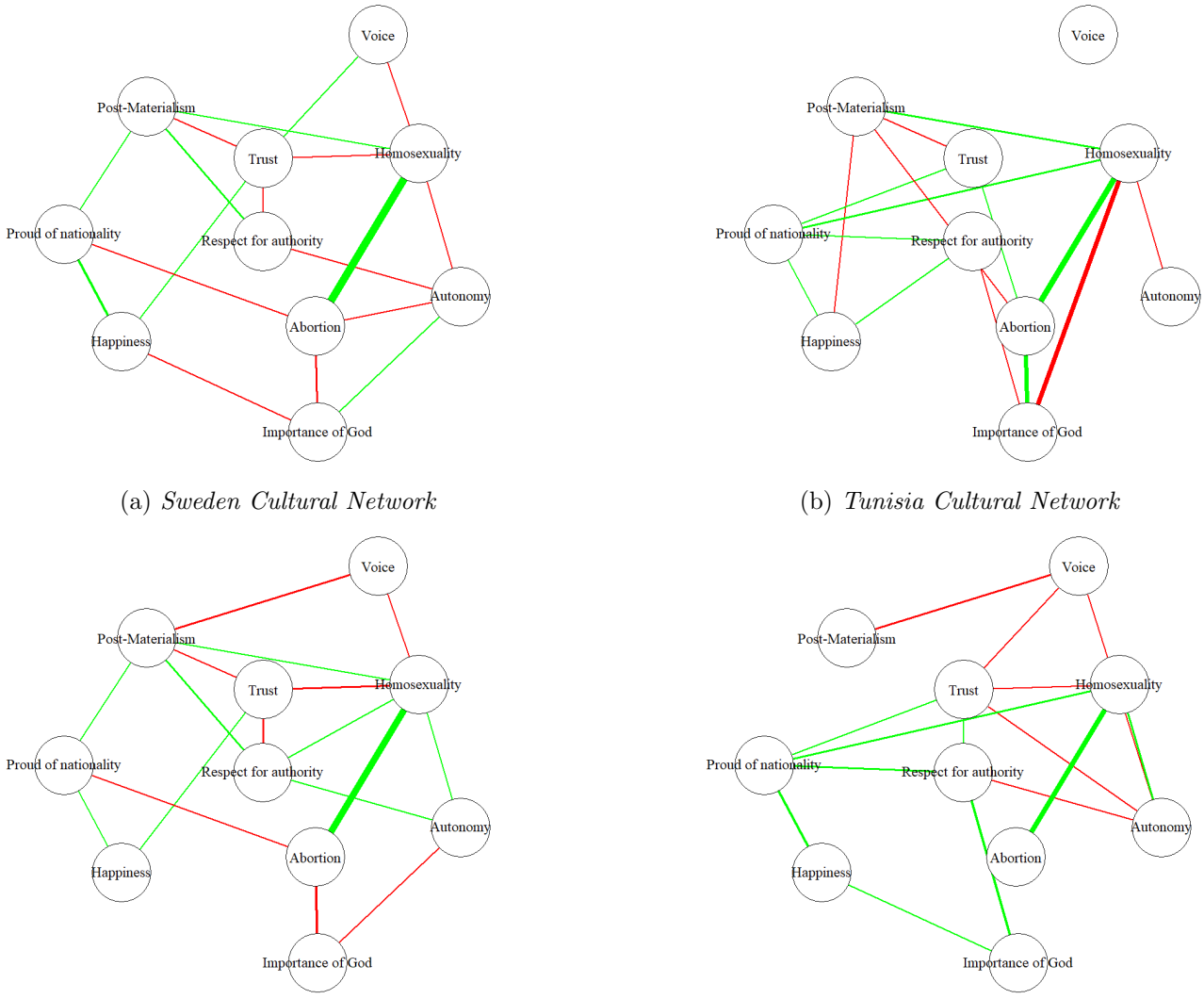
reach the convergence of the model for each analysis, even though half of them are intended as burn in, and the remaining half are used in the estimation process. The choice of the number of iterations for *Battery dataset*, *IW dataset* and *IW1 dataset* is in line with De Benedictis et al. (2021), while, as mentioned in 1.2.2 and shown in Table 1.11 of Annex 1, for the *Large dataset*, it was dictated by the number of selected variables.

For having a complete picture of the complex national system of values, the wider network of 60 variables/nodes is included in the final definition of the cultural distance index, but it turns out to be too broad to represent, describe and compare, so the networks with 6 and 10 cultural traits are depicted below. As an example, we respectively compare 4 countries in pairs: we show cultural networks for Sweden and Tunisia, which are located far apart in the Cultural Map, compared to their inferred networks in De Benedictis et al. (2021); we further add the cultural networks of Italy and Slovakia, which instead, beyond belonging to the same group (Catholic Europe), are located close each other. The presence/absence of a tie is dictated by the *binary network*, while the color and the thickness of the line by the information in the *partial correlation network*. Positive partial correlations are indicated in green and negative ones in red, while the thickness of the line is given by the absolute value of the partial correlation, namely the strength of the relationship.
From cultural networks depicted in Figure 1.2 and Figure 1.3 emerge interesting hint for thought for the interpretation of the individual national system of values and of the differences among countries.

Globally, as for Wave 6, there is not an evident difference in edge density[48] of the cultural network of Sweden (Figure 1.2a and 1.2c) and Tunisia (Figure 1.2b and 1.2d). From Figure 1.2 we notice Sweden as a more stable and "mature" culture, while Tunisia appears very unstable. As the years passed, comparing the two cultural networks of Sweden, the overall role of cultural traits seems to be very similar, at the contrary for Tunisia this is only partially verified. Provided the considerations made by De Benedictis et al. (2021), for the cultural networks of Tunisia only the role of *Proud of nationality* remains the same, stating that it represents a founding value of Tunisian culture. Peripheral cultural traits become central and central cultural traits become peripheral. The former is the case of *Post-Materialism* and *Abortion*, which centralities was one of the main differences

[48] The edge density of a graph is calculated as the ratio between the number of observed edges over the maximum number of edges in the case of a complete graph (namely, when all the nodes are connected each other).

Figure 1.2: The Network Structure of National Cultures (Sweden and Tunisia, IW variables



(a) *Sweden Cultural Network*

(b) *Tunisia Cultural Network*

(c) *Sweden Cultural Network - Wave 6*

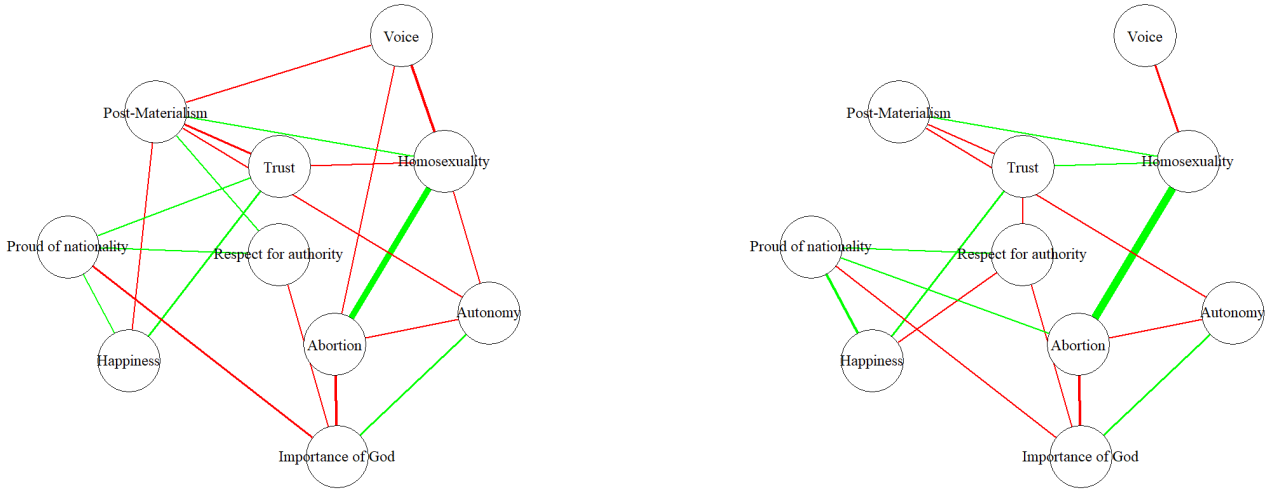
(d) *Tunisia Cultural Network - Wave 6*

Note: The optimal graphs for Sweden (a) and Tunisia (b), that are placed far away each other in the Inglehart-Welzel Cultural Map. Colouring and thickness of edges is based on the associated partial correlations (Positive: Green, Negative: Red). Node labels refer to the description of the cultural traits used for depict the Cultural Map.

between Sweden and Tunisia in Wave 6. The latter is the case of *Autonomy* and *Voice*, which disappear, in the case of *Voice*, and almost disappear, in the case of *Autonomy* (it is only connected with *Homosexuality*), from national cultural framework. The important core on the Sweden network is represented by *Homosexuality*, *Post-Materialism*, *Respect for authority* and *Trust*. This core seems to be visualized also in the new cultural network, even though now *Homosexuality* and *Respect for authority* are not linked. As for De Benedictis et al. (2021), we can confirm that this cultural traits appear to define the core structure of the country culture: tolerance, independence, post-materialism, and social trust

stand out as the leading elements in Swedish culture. Finally, we can conclude that the extreme difference between these two cultures in the Cultural Map is confirmed also by their cultural networks, now even more evident than the Wave 6.

Figure 1.3: The Network Structure of National Cultures (Italy and Slovakia, IW variables



(a) *Italy Cultural Network* (b) *Slovakia Cultural Network*

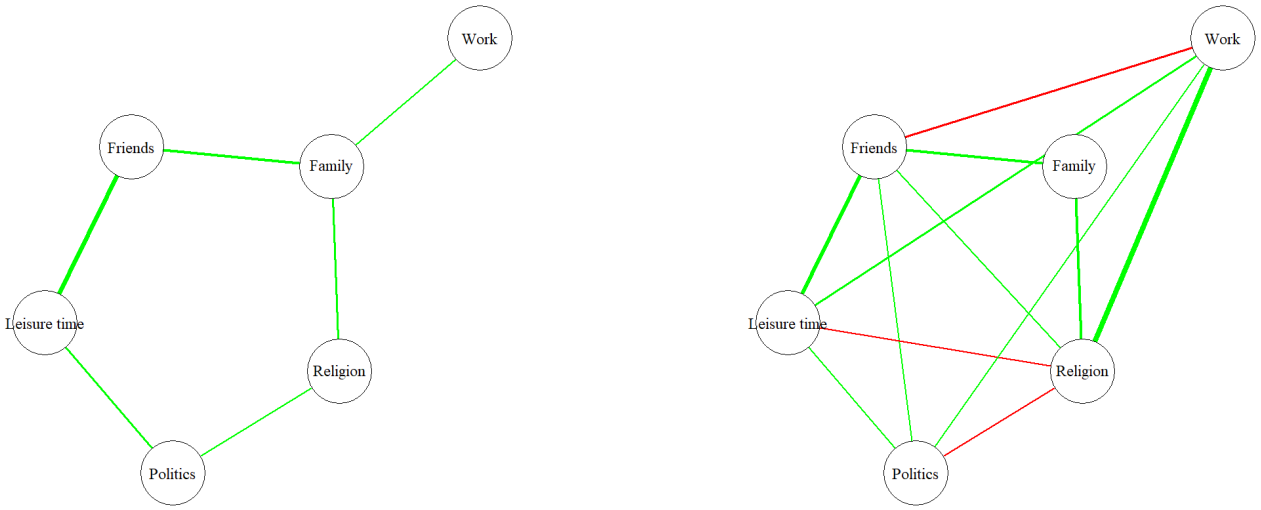
Note: The optimal graphs for Italy (a) and Slovakia (b), that are placed close each other in the Inglehart-Welzel Cultural Map. Colouring and thickness of edges is based on the associated partial correlations (Positive: Green, Negative: Red). Node labels refer to the description of the cultural traits used for depict the Cultural Map.

Proximity among Italy and Slovakia inside the Cultural Map is not completely found for their cultural networks (Figures 1.3a and 1.3b). The way in which the individual cultural traits are perceived is in contrast with the way in which they interact. These differences could be induced by many factors, such as climatic and economic conditions, historical evolution, demographic composition, urban/rural distribution of the inhabitants, and so on. Often a complex cultural identity can result in more dense networks, while more standardized ones can lead to more sparse networks. This is the case of Italy and Slovakia, in fact if we check for trait by trait, we can easily notice as a good part of the edges are in both the networks, while another consistent part is present only in the cultural network of Italy. For example, *Voice* is negatively connected only with *Homosexuality* for Slovakia, while for Italy it is negatively connected also with *Post-materialism* and *Abortion*. Another example is the role of *Post-materialism*, which is more central for Italy than Slovakia. The two countries are grouped as "Catholic Europe" by Inglehart and Welzel in their Cultural Map, in fact not surprisingly they share the same and really meaningful role of *Importance of God*. Although in the overall

vision the two networks have some points in common, the specific differences due to the complexity of the cultural identity, certainly enrich the distance relationship between them, otherwise wrongly considered very similar.

One element of interest shared by the four networks (Sweden, Tunisia, Italy and Slovakia), and more generally by all 76 cultural networks, is the positive link between *Abortion* and *Homosexuality*. They tend to be either both supported or both opposed and the thickness (intensity) becomes the measure of its importance inside a specific network.

Figure 1.4: The Network Structure of National Cultures (Sweden and Tunisia, first battery variables)



(a) *Sweden Cultural Network* (b) *Tunisia Cultural Network*

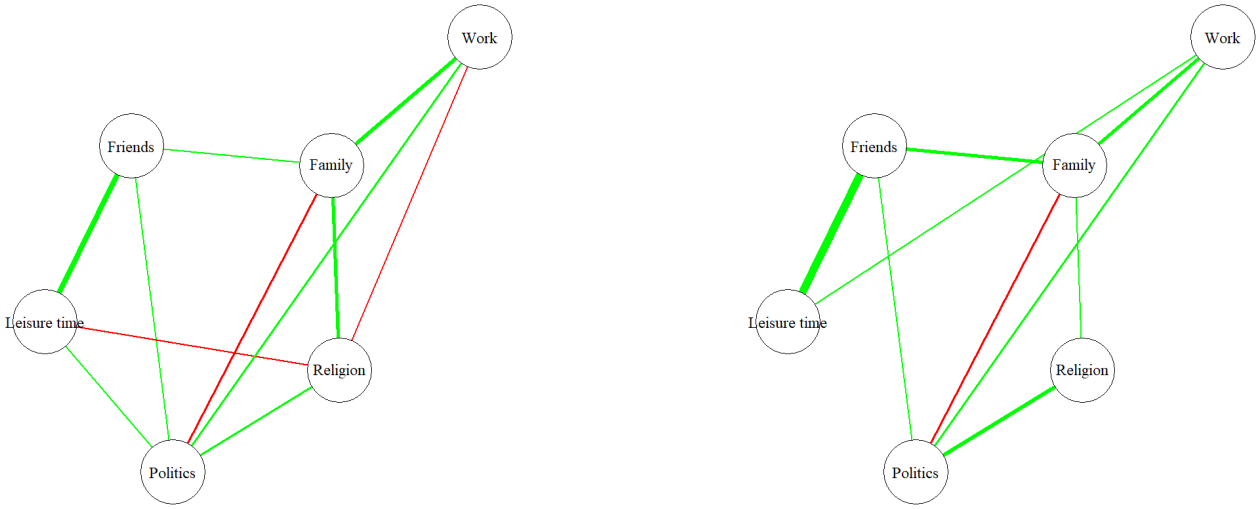
Note: The optimal graphs for Sweden (a) and Tunisia (b), that are placed far away each other in the Inglehart-Welzel Cultural Map. Colouring and thickness of edges is based on the associated partial correlations (Positive: Green, Negative: Red). Node labels refer to the description of the cultural traits of the first battery of questions.

Considering the cultural networks derived from the questions of the "Priorities in Life" battery in Figure 1.4 and 1.5, at first glance, we notice 3 main things: the prevalence of positive ties; the higher similarity between Italy and Slovakia compared to Sweden and Tunisia, confirming with their proximity on the Cultural Map; the lower edge density in the network of Sweden compared to the others. In the Sweden's cultural network (Figure 1.4a) there is not a cultural trait that is particularly central compared to the others. The network is very sparse with reference to Tunisia, Italy and Slovakia, while the two main evident things are represented by: the circle subgraph between all cultural traits except *Work*; the link between *Family* and *Work*. This schematicity in the network of priorities in life is the reflection of a society that historically can be considered to have a very

rational and schematic organization.

Tunisia's "Priority in Life" network is much denser than that of Sweden. The priority of *Work* is extremely important, but unlike Sweden it is connected with all cultural values except *Family*. The non-connection between *Work* and *Family* is certainly singular as it can be imagined that work priorities can be linked to family ones. Other interesting things concern the negative link between *Work* and *Friends*, *Religion* and *Leisure time*, *Religion* and *Politics*, which therefore, given the rest of the network, are symbiotically perceived with an inverse priority scale.

Figure 1.5: The Network Structure of National Cultures (Italy and Slovakia, first battery variables)



(a) *Italy Cultural Network* (b) *Slovakia Cultural Network*

Note: The optimal graphs for Italy (a) and Slovakia (b), that are placed close each other in the Inglehart-Welzel Cultural Map. Colouring and thickness of edges is based on the associated partial correlations (Positive: Green, Negative: Red). Node labels refer to the description of the cultural traits of the first battery of questions.

For Italy and Slovakia's "Priority in Life" networks (Figure 1.5), we see something similar to Figure 1.3. All edges observed for the network of Slovakia are found even in the network of Italy, except for that connecting *Leisure time* with *Work*. In the network of Slovakia, the link between *Leisure time* and *Friends* is striking for its importance, which is outlined by the evident thickness of the line. Furthermore, we note the negative link between *Family* and *Politics* (also in the network of Italy) whose priority is inversely perceived.

The role of the priority of *Leisure time*, *Work* and *Religion* are the things that most differentiate Italy from Slovakia. In the Italian culture, the priority in

Leisure time is also assimilated with a positive link to the priority in *Politics*, and negatively connected to the priority in *Religion*. *Work*, which is not connected with *Leisure time*, is however negatively connected with *Religion*. *Politics* is certainly a central priority in the Italian network, whether it is perceived as a high priority or not.

Table 1.7: Networks statistics

		N/P Edges		Density		Degree		Betweenness		Closeness	
Battery	Min	0	(Many)	0.4	(Many)	0.1	(Many)	0.04	(KZ)	0.08	(Many)
	Average	0.09		0.57		0.3		0.23		0.34	
	Max	0.5	(BD)	0.87	(KZ)	0.8	(MK)	0.74	(MK)	0.87	(MK)
IW	Min	0.33	(BD, KR)	0.16	(AM)	0.08	(SK)	0.08	(AT)	0.08	(VN)
	Average	0.94		0.37		0.3		0.24		0.3	
	Max	1.8	(CO)	0.58	(DE)	0.69	(BR)	0.54	(MX)	0.69	(BR)
IW1	Min	0.53	(MO)	0.19	(PH, VN)	0.12	(CN)	0.07	(AT)	0.09	(AT)
	Average	1.32		0.31		0.26		0.2		0.28	
	Max	3.5	(BR)	0.43	(ME)	0.46	(BR)	0.47	(MM)	0.5	(BR)
Large	Min	0.58	(PH)	0.14	(ID)	0.11	(KG)	0.02	(TN)	0.09	(CH)
	Average	0.76		0.22		0.19		0.04		0.17	
	Max	0.95	(AZ)	0.3	(AZ)	0.3	(AL)	0.1	(IR)	0.27	(IR)

Note: Network summary statistics calculated on the optimal binary Graph for each country. For the calculation of the ratio *N/P Edges* (Negative/Positive Edges) we consider the signed adjacency matrix, with 1 assigned to an edge with positive partial correlation and -1 for an edge with negative partial correlation, whereas 0 for missing edges. ISO 3166-1 alpha-2 code, reported in square brackets, is replacing the full name of countries: AL - *Albania*; AM - *Armenia*; AT - *Austria* AZ - *Azerbaijan*; BD - *Bangladesh*; BR - *Brazil*; CH - *Switzerland*; CN - *China*; CO - *Colombia*; DE - *Germany*; ID - *Indonesia*; IR - *Iran*; KG - *Kyrgyzstan*; KR - *South Korea*; KZ - *Kazakhstan*; ME - *Montenegro*; MK - *North Macedonia*; MM - *Myanmar*; MO - *Macao*; MX - *Mexico*; PH - *Philippines*; SK - *Slovakia*; TN - *Tunisia*; VN - *Vietnam*.

In addition, some summary measures are calculated for all the country-networks of different sizes. The five statistics concerns the sign of edges, the sparsity and the level of centralization[49] of them. Table 1.7 shows the results of these descriptive statistics (see Wasserman and Faust 1994 for their formulation). One general evidence regards the decreasing maximum value for the distribution along countries of *Density*, *Degree centralization*, *Betweenness centralization*, *Closeness centralization* when the number of cultural traits increases. Following, there are the main results measure by measure considering the *cultural networks* from the *IW dataset*, which can be compared with findings of De Benedictis et al. (2021).

N/P Edges is a statistic derived by crossing the binary network and partial correlation matrix. Presence/absence of a link is described by the former, while the sign of them by the latter. It measures the ratio of negative edges over positive edges in the network: when the ratio is greater than 1 there are more negative edges connecting cultural traits than positive ones. When combined with the

[49] For the elaboration of these measures we make use of functions provided by the R package `igraph`.

other statistics of Table 1.7 it can provide relevant information about the symbiotic circulation of values in national cultures. For example, considering two structural equivalent networks, sign of edges can be fundamental to define the meaning of connections among cultural values. Differently from De Benedictis et al. (2021), here, the ratio is *Negative/Positive Edges*[50]. As for the sample of countries of Wave 6, the average ratio shows the prevalence of positive over negative relations. The minimum value is 0.33 (Bangladesh and South Korea: with 6 and 12 positive links, and, with 2 and 4 negative links), the maximum is 1.8 (Colombia: with 5 positive links, and 9 negative links).

The *Density* of a network, as mentioned in note 47, is the ratio between observed number of edges and maximum possible number of edges[51]. In line with De Benedictis et al. (2021) values of the *Density* correspond to the level of the cultural complexity of a society. We found as extreme cases Armenia[52], which is characterized by a high sparsity/low density, and Germany which low sparsity/high density represents a culturally complex society. As De Benedictis et al. (2021), high density is observed also for United States (0.51). Since values of density are remarkably higher than the ones of De Benedictis et al. (2021), in general networks of the WVS/EVS Joint 2017 seem to describe more complex national cultures.

The *Degree* centralization summarizes for each country and network size the centralization of the overall network with respect to the degree of each node (cultural trait). It is the sum of the difference between the maximum observed degree and the node degree, standardized by the theoretical maximum of this sum[53]. For this reason this measure is in the interval [0,1], where 0 is verified for empty or complete networks and 1 is verified for a star-like network (see Networks and symbols). As De Benedictis et al. (2021) suggests, *"a high level of Degree centralization indicates that the national culture is structured around one or few traits, showing a strong core-periphery structure"*. This is the case of Brazil, where *Importance of God* is central, while at the contrary is not the case of Slovakia, which network is depicted in Figure 1.3b.

[50] For the *Battery cultural networks* some networks lack negative links.

[51] In a cultural network with $p = 6$ nodes, the maximum number of edges is $15 = \frac{6\times(6-1)}{2}$; for $p = 10$ nodes is $45 = \frac{10\times(10-1)}{2}$; for $p = 14$ nodes is $91 = \frac{14\times(14-1)}{2}$; for $p = 60$ nodes is $1770 = \frac{60\times(60-1)}{2}$.

[52] For Armenia, *Importance of God* has the higher degree (it is connected with 5 cultural traits). *Voice* and *Post-materialism* are isolates in the network, while the link between *Homosexuality* and *Abortion* represents a component (within-connected subgraph, but disconnected with the other subgraphs of the network).

[53] Analytically, the *Degree* centralization is $\frac{\sum_{i=1}^{p}(max(Degree_i)-Degree_i)}{max\sum_{i=1}^{p}(max(Degree_i)-Degree_i)}$.

The *Betweenness* centralization moves from *Degree* centralization with a similar intent, but, using the Betweenness centrality as reference point, it investigates the presence of mediating cultural traits[54], namely, one or more nodes that connect different parts of the network that would otherwise be unconnected. Values close to 1 indicate a network highly conditioned by the presence of mediators, while values close to zero show the opposite. In the network of Mexico *Importance of God* covers the role of bridging trait between different subgraphs[55].

Using Closeness centrality at the node level, we calculate the *Closeness* centralization to detect the proximity between the nodes of the country-network[56], e.g. in some sense the smallworldness of the network. Values close to 1 describe networks in which traits can be reachable each other through shortest paths (e.g., a cohesive network). The network of Vietnam is highly disconnected: *Respect for authority*, *Proud of nationality* and *Happiness* form a component of the network, *Trust* is isolate, *Abortion* is only connected with *Homosexuality* and the other traits have a degree equal to 2, except for Voice (3). Since the size of networks is limited to 10 nodes, for our sample of countries the high level of Degree Centralization of the Brazil's network coincides with an high level of Closeness centralization, in fact around *Importance of God*, nodes are quite connected between them.

1.5 Definition of distance measures

In Section 1.4 we fitted 76 network models of the cultural system of values for each country considered by WVS/EVS Joint 2017. The analysis returns for each county two main findings: the estimated partial correlation matrix and the estimated probability of inclusion for each edge.
An integral part of each country-model are the marginal distributions F_j for each cultural trait, estimated non-parametrically by the empirical cumulative distribution function (see Equation 1.3 in Annex 1).
Following the intuition explicitly mentioned in Section 1.3, the main purpose of this Chapter is to empirically demonstrate as the cultural distance between countries should account both these statistical object: distances between coun-

[54] The *Betweenness* centralization is calculated as $\frac{\sum_{i=1}^{p}(max(Betweenness_i)-Betweenness_i)}{max\sum_{i=1}^{p}(max(Betweenness_i)-Betweenness_i)}$.

[55] From one side there are *Abortion*, *Homosexuality*, *Autonomy* and *Voice*, from another *Proud of nationality* and *Happiness*, instead *Trust* and *Respect for authority* are uniquely connected with it, while *Post-materialism* is isolate.

[56] The *Closeness* centralization is defined as $\frac{\sum_{i=1}^{p}(max(Closeness_i)-Closeness_i)}{max\sum_{i=1}^{p}(max(Closeness_i)-Closeness_i)}$.

tries' cultural inferred networks and distances between countries' cultural traits marginal distributions. Different countries could have different spread of cultural traits inside their population, then different inter-connections between them; at the same time countries' attitudes to the individual cultural traits are captured by their marginal distributions. In this aim, as the two components are distinct statistical objects, for aggregating them into one cultural distance measure we need first to calculate distances for both separately.

Following De Benedictis et al. (2021), the first distance measure is the common Edge Difference Distance, defined by the Frobenius norm of the differences between two networks (Hammond et al., 2013). It measures in a simple way similarities between networks enhancing the information due to the weight and the sign of ties. Applied to cultural matrices of partial correlation results in:

$$\texttt{ParCorr}(PC^{(l)},PC^{(m)}) = ||PC^{(l)}-PC^{(m)}||_F = \sqrt{\sum_{i,j=1}^{p}(PC_{ij}^{(l)}-PC_{ij}^{(m)})^2},$$

with $PC^{(l)}$ and $PC^{(m)}$ the partial correlations of countries l and m, respectively, (with $l = 1, ..., r$ and $m = 1, ..., r$ where r is the number of countries) and p is the number of cultural traits (variables).
While applied to cultural matrices of posterior probabilities of edge inclusion:

$$\texttt{ProbEdge}(PP^{(l)},PP^{(m)}) = ||PP^{(l)}-PP^{(m)}||_F = \sqrt{\sum_{i,j=1}^{p}(PP_{ij}^{(l)}-PP_{ij}^{(m)})^2},$$

with $PP^{(l)}$ and $PP^{(m)}$ the posterior probabilities of edge inclusion of countries l and m, respectively.

A second distance measure is proposed in the way of Inglehart and Welzel (2005). In their Cultural Map, Inglehart and Welzel detect proximity between countries' cultures showing the euclidean distance between countries in the bi-dimensional reduced space from the PCA over the average-based country responses to the 8 survey questions and the two derived indices. Here, instead the reduced space, the complete one is considered.

$$\texttt{Mean}(M^{(l)},M^{(m)}) = \sqrt{\sum_{j=1}^{p}(M_j^{(l)}-M_j^{(m)})^2}$$

where, $M_j^{(l)}$ and $M_j^{(m)}$ $(j = 1, \ldots, p)$ are the average of country l and m, respectively, for cultural trait j.

The average value is one of the usual index for summarizing the distribution of each cultural trait for each country, mostly in this case in which variables'

supports are in a definite short interval (as in our case, the ordinal variables describe the cultural traits), but marginal distributions provide a more precise and complete information about each cultural trait. Given the starting point of the considered network model are the marginal distributions F_j, finally, a third distance measure on the marginal distributions is defined according to the Ichino and Yaguchi (1994) first formulation of a dissimilarity measure U_2. It is placed inside the wider field of symbolic data analysis, in fact, the vector of modal distributions for each country is seen as the description of a symbolic object. This measure is a generalized Minkowski metrics and was defined in Ichino and Yaguchi (1994) for applications to data of different nature[57]. Here we use it for calculate distances between modal distributions (marginal distributions of categorical and ordinal variables) and the specific definition is as follow:

$$\texttt{MargDistr}(F^{(l)},F^{(m)}) = \sqrt{\sum_{j=1}^{p}[\phi(F_j^{(l)},F_j^{(m)})]^2}$$

where, $F_j^{(l)}$ and $F_j^{(m)}$ $(j = 1,\ldots,p)$ are the marginal distributions of country l and m, respectively, for cultural trait j, and where the function ϕ is defined by:

$$\phi(F_j^{(l)},F_j^{(m)}) = \sum_{i=1}^{C_j}\left(f_j^{(m)}(c_i)\ln\left(\frac{f_j^{(m)}(c_i)}{f_j^{(l)}(c_i)}\right) + f_j^{(l)}(c_i)\ln\left(\frac{f_j^{(l)}(c_i)}{f_j^{(m)}(c_i)}\right)\right)\Big/2$$

where C_j is the total number of categories for the j-th survey question (categories not observed in the data are not included in the calculation), f_j denotes the probability mass function, so $f_j^{(l)}(c_i)$ gives the frequency associated to the category c_i for country l[58].

1.6 A new index of cultural distance

In respect to De Benedictis et al. (2021) the `CulturalMap` is dropped from the analysis[59], but the main objective of this section remains the same. To consider the defined four distance measures and to evaluate them in their effectiveness at quantifying cultural distances between countries, respecting the theorized dualism between the full distribution of cultural traits and their interdependencies.

[57] For example, *intervals* and *histograms* from classic discrete or continuous variables, *modal* from categorical or ordinal variables.

[58] The definition of Ichino and Yaguchi (1994) is more general, e.g. allowing distances between symbolic objects of different nature. In our case, the first term of the formula in Ichino and Yaguchi (1994) is 0 and we consider the case of $\gamma = 0.5$.

[59] The distance between countries in the Cultural Map is shown in Figure 1.1. Since it is obtained from the first two factors of the PCA over the *IW dataset*, it carries part of the information contained in it, which is implicitly included in the general average-based distance (`Mean`).

In doing so, De Benedictis et al. (2021) applies a PCA over the distances distributions, observing an orthogonality relationship between the two parts of the cultural distance definition. To confirm the goodness of the results then it shows scatter plots of the pairwise distances and analyzes them via the Social Relations Regression Model[60] (SRRM), finally compares the *network index* - obtained as the sum of the normalized `ParCorr` and `MargDistr` distances - with the euclidean distance between countries within the Inglehart-Welzel Cultural Map (*IW index*), where interdependencies among cultural traits are not considered.
Anyway, given the statistical dependencies between the elements of a distance, PCA is not usually used to study distances without using any trasformation[61]. To overcome the problems of this operation, De Benedictis et al. (2021) firstly considers the upper triangle of the distance matrices, secondly vectorizes them[62] and lastly normalizes them through the minimum and maximum of their distributions.

In an effort to improve the final cultural distance index, that takes into account both the two components of Section 1.3 like the *network index*, here, we make use of the DISTATIS approach (Abdi et al., 2005). It is an extension of the Multidimensional Scaling (MDS) (Torgerson, 1958; Cox and Cox, 2008) and unifies the PCA approach provided in De Benedictis et al. (2021) in a single standardized procedure. Basically, as reported in Abdi et al. (2005), DISTATIS is organized in seven main steps:

1. Transform each distance matrix (i.e., each study) into a between-object cross-product matrix.
2. Analyze the structure of the cross-product matrices.
3. Derive an optimal set of weights for computing the compromise.
4. Compute the compromise as a weighted sum of the individual cross-product matrices.
5. Compute the eigen-decomposition of the compromise matrix.
6. Plot the projections of the observations in the compromise space.
7. Plot the trajectory of the observations as supplementary points in the compromise space.

[60] The model is designed for network data and, in its basical formulation, accounts for the dyadic nature of distances (statistical dependencies between distances) including of random effects for each node (Hoff, 2009).
[61] This is the founding principle of Multidimensional Scaling (MDS), where a PCA is applied on a distance matrix only after its transformation in a cross-product matrix (Torgerson, 1958).
[62] The result is a matrix with rows representing the country pairs and with columns representing the distances.

The above points can be merged in two main steps: the between-distances analysis (the first three) and, the computation and analysis of the compromise distance (the last four).
The former provides the comparison between the countries cultural distance matrices. As first, because distance matrices cannot be analyzed directly via eigen-decomposition, they need to be transformed in a more convenient form, namely in cross-product matrices (see the Multidimensional Scaling approach Torgerson (1958)). In order not to risk orienting the analysis to matrices with greater inertia, they are normalized by dividing each of them by their first eigenvalue. After this, matrices are vectorized and organized by columns in a unique matrix. The between-distances analysis is performed by calculating the pairwise similarities using the *Rv coefficient*, which exploit the normalization into *cosine*. This second normalization balances the heterogeneity of information.
The latter starts computing the *non centered PCA* of the *cosine matrix*, which elements are the *Rv coefficients*. Given the non centered PCA maximizes the explained inertia of the first dimension in the way similar distances must weigh more than the dissimilar ones, elements of its re-scaled eigenvector[63] are used as weights to find the compromise distance. This is calculated as weighted average of the cross-product matrices. The analysis of the compromise matrix via its eigen-decomposition is the last step that allows to map individuals (in our case countries) in the compromise space.

Finally, we remain in the same main purpose of De Benedictis et al. (2021), namely to reach a final cultural distance index (like the *network index*). Furthermore, we evaluate how the association between the distance measures varies when the number of nodes increases and, again with DISTATIS, we compare four final proposal for the updated formulation of "*network index*" with the *IW index*.

1.6.1 The compromise cultural distance

At this point, for each initial dataset (*Battery*, *IW*, *IW1* and *Large*) we have obtained three different synthesis: the network structure of the cultural traits for each country; the dataset of symbolic objects, where each country is one of these objects described by the vector of the cultural traits distributions; the matrix of averages, in which each country-row is identified by the mean vector along the considered cultural traits. Over these we have calculated four distance matrices: `Parcorr` and `Probedge` are calculated on the cultural networks, respectively on

[63] The sum of the elements of the eigenvector needs to be equal to one.

the partial correlation networks and the posterior edge inclusion probabilities networks; `Margdistr` is calculated between the countries seen as symbolic objects; while, `Mean` is calculated between countries considering the vector of averages. Summing up, from each dataset we have extracted four measures of cultural distance among countries, two of which describe the network part and the other two the distributional one. In order to observe how similarity relations between these distances (the parts of the theoretical definition of Section 1.4) change as the number of cultural traits included in the analysis increases, we apply the DISTATIS method by observing the results of the first step, that is, the between-distance analysis.

Table 1.8: Vectorized *cosine matrices* by dataset

	`Battery dataset`	`IW dataset`	`IW1 dataset`	`Large dataset`
(ParCorr,ProbEdge)	0.71	0.83	0.77	0.91
(ParCorr,MargDistr)	0.53	0.49	0.51	0.52
(ParCorr,Mean)	0.57	0.52	0.52	0.57
(ProbEdge,MargDistr)	0.31	0.38	0.35	0.35
(ProbEdge,Mean)	0.39	0.41	0.38	0.42
(MargDistr,Mean)	0.91	0.94	0.94	0.93

In Table 1.8 similarities between cultural distances are compared based on the different datasets from which are calculated. Each column represents the vectorization of the upper triangle of the *cosine matrix* (it is simmetric). Looking at the table by rows (from left to right), it is evident how the pairwise similarities between distances remain almost constant as the variables included in the analysis increase, except for the similarity between `ParCorr` and `ProbEdge`, which increases. On the one hand, this table shows more and more clearly the contrast between the two parts of the definition described in paragraph 1.3, on the other, it shows an overlap of information between them (all except the first and the last row). Although a *cosine* varies between -1 and 1, the observed matrices have only positive values, in fact the different distances *"tend to agree on what they measure on the objects"* (Abdi et al., 2005), at most they can be unrelated[64].
Table 1.9, instead, shows the results of the non centered PCA applied over the *cosine matrices*. Observing the table, three fundamental things are of notice (deductible in part also from the *cosine matrices*). The first concerns the high value of the first eigenvalue for each reference dataset, which is a peculiarity of the non centered PCA. Since this eigenvalue expresses the explained inertia of

[64] For the *Perron-Frobenius theorem*, when the elements of a positive semi-definite matrix (the *cosine matrix* has these proporties) are all positive, the first eigenvector has all its elements with the same sign (in our case positive).

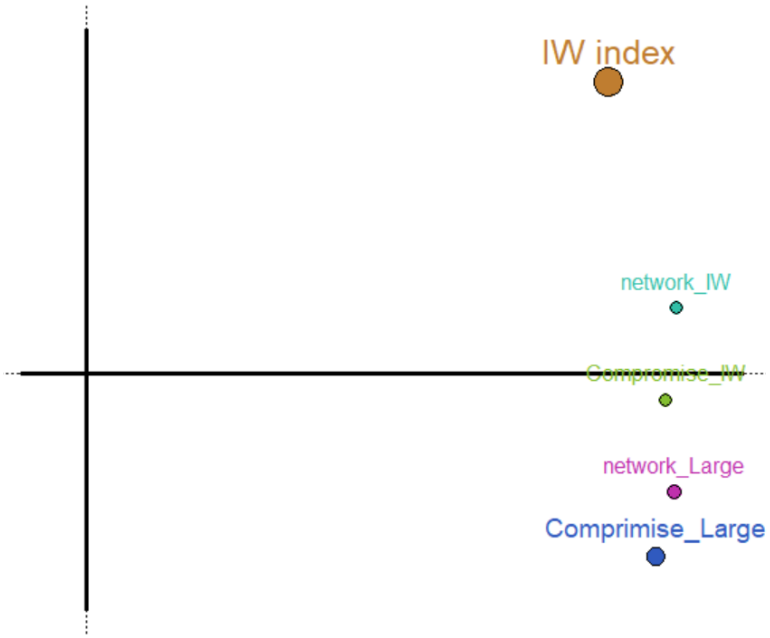
Table 1.9: Results of the non centered PCA over the *cosine matrices* associated to each dataset

			Dim 1	Dim 2	Dim 3	Dim 4
Battery	Eigenvectors	Parcorr	0.51	0.39	0.77	0.03
		ProbEdge	0.43	0.66	-0.61	-0.09
		MargDistr	0.52	-0.49	-0.07	-0.69
		Mean	0.54	-0.41	-0.17	0.71
		Eigenvalues	2.72	0.94	0.26	0.08
		Inertia (%)	68	24	6	2
IW	Eigenvectors	ParCorr	0.5	0.45	0.73	-0.03
		ProbEdge	0.46	0.57	-0.67	-0.005
		MargDistr	0.51	-0.5	-0.07	-0.7
		Mean	0.52	-0.47	-0.04	0.71
		Eigenvalues	2.79	0.99	0.16	0.06
		Inertia (%)	70	25	4	1
IW1	Eigenvectors	ParCorr	0.51	0.43	0.75	0.03
		ProbEdge	0.44	0.62	-0.65	-0.04
		MargDistr	0.52	-0.48	-0.05	-0.71
		Mean	0.53	-0.46	-0.12	0.71
		Eigenvalues	2.74	0.98	0.21	0.06
		Inertia (%)	69	25	5	1
Large	Eigenvectors	ParCorr	0.52	0.42	0.7	0.24
		ProbEdge	0.47	0.59	-0.6	-0.27
		MargDistr	0.49	-0.52	0.17	-0.67
		Mean	0.51	-0.45	-0.33	0.65
		Eigenvalues	2.86	1.00	0.08	0.07
		Inertia (%)	71	25	2	2

the first dimension which is described by the first eigenvector, the importance of the latter is relatively the same as the number of variables included in the analysis increases (the same for the importance of the second, third and fourth eigenvectors). From one point of view this shows the slight overlap of information between the two parts (network and distribution), on the other, and we come to the second point, the increase in the measure of similarity between `ParCorr` and `ProbEdge` results in a substantial balance in the contribution of distances to the first eigenvector. Finally, the third is noted in the less important dimensions (3 and 4). They show an inverse relationship between `ParCorr` and `ProbEdge` for Dimension 3, and between `MargDistr` and `Mean` for Dimension 4. To explain the case of `ParCorr` and `ProbEdge`, imagine two countries over which the network of partial correlations and the network of probabilities are calculated. Each of these networks conveys the ties between all cultural traits. Some of the probabilistic links may be high for both the countries, but this can translate, for the partial correlation networks, into high negative partial correlations for one country and positives for the other. If we think that each network contains in its adjacency matrix at least 15 potential links in the case of *Battery* and 1770 in the case of *Large*, it is immediate to understand why this component is less important when

the number of variables included in the analysis increase. Furthermore, given the nature of non centered PCA, this relationship can be partially contained in the first dimension.

Figure 1.6: Projection of the cultural distances on the first two dimensions of the non centered PCA over their *cosine matrix*



Note: The size of the points is the sum of the contribution of each distance over the two first dimensions.

From one side, the substantial invariance of the results when the number of variables involved in the analysis increases, confirms the goodness of Inglehart and Baker (2000) choice to reduce the number of cultural traits included in the construction of the Cultural Map. Anyway, the greater informative content inside the *Large dataset* - which considers 60 cultural traits - provides to emphasize slightly more the effectiveness of the two parts to empirically define the national culture, and therefore the cultural distance between countries. On the other side, we can confirm also the goodness of the deductions made by De Benedictis et al. (2021) using data from the WVS Wave 6, in fact it is clear as the two components, although some overlap of information, are both important in defining the cultural distance between countries.

Given the high level of similarity among the distances taken two by two (`ParCorr` and `ProbEdge`, `MargDistr` and `Mean`) we decide to obtain different proposal for the final cultural distance index by merging via DISTATIS only `ParCorr` and `MargDistr` (the two used by De Benedictis et al. (2021)).

Considering results for *IW dataset* and *Large dataset*, the first two proposal are calculated running two further DISTATIS including the two selected distances. Using DISTATIS on two variables is a particular case in which, whether the distances are similar or dissimilar, the weights to calculate the final compromise are 0.5. For this reason we are not interested in the results of the between-distances analysis (the first part of DISTATIS), but in the compromise measurement, which can be equated to the *network index* of De Benedictis et al. (2021), except for the normalization procedure. We therefore have a *Compromise_IW* and a *Compromise_Large*.

After that, we calculate the cultural distance index for the information of *IW dataset* and *Large dataset* in the manner of De Benedictis et al. (2021), considering the simple average and not the sum of `ParCorr` and `MargDistr`. We call these two index *network_IW* and *network_Large*.

Finally, we deduce the Euclidean distances among countries in the Inglehart-Welzel Cultural Map shown in Figure 1.1, calling it the *IW index*.

The results of the between-distances analysis of the last DISTATIS are depicted in Figure 1.6. It shows the projection of the five cultural distance indices on the first two dimensions of the non centered PCA applied on the *cosine matrix*. These two dimensions contribute to explain 95% of the total inertia of the *cosine matrix*, of which 88% due to the first dimension. This means that all distances intend to measure the same concept, however *IW index* has the smallest factor score[65] on the first dimension: 0.86 versus 0.94 of *Compromise_Large*, 0.95 of *Compromise_IW*, 0.97 of *network_Large* and 0.97 of *network_IW*. Furthermore, with a factor score of 0.48 versus -0.3 of *Compromise_Large*, *IW index* has an important role for the second dimension. Even the *cosine matrix* (Figure 1.10) shows lower values for the similarities among *IW index* and the other cultural distance indices, although they are of more than good intensity, because each measure contains the distributional part. Not surprisingly, of these, the lowest is found among *Compromise_Large* and *IW index*.
We can conclude that considering as final cultural distance index the *Compromise_Large*, we can improve the precision of the measure of cultural distance between worldwide countries analyzed by the WVS/EVS Joint 2017. Binary cul-

[65] Factor scores are obtained as transformation of the eigenvectors. Lower values for eigenvector corresponds to lower final weight for the compromise distance.

Table 1.10: *Cosine matrix* of the cultural distances

	Compromise_Large	Compromise_IW	network_Large	network_IW	IW index
Compromise_Large	1	0.86	0.99	0.85	0.69
Compromise_IW		1	0.88	0.97	0.75
network_Large			1	0.89	0.76
network_IW				1	0.85
IW index					1

tural networks from the *Large dataset* will be used in the third Chapter as a case study.

1.7 Conclusions

In the way of De Benedictis et al. (2021) and on the basis of Inglehart and Welzel (2005), the main purpose of this Chapter was to extend the current cross-country cultural distances, specifically finding a new and more complete measure of cultural distance that takes into account the interdependencies between cultural traits at the country level. Unlike De Benedictis et al. (2021), for which the data from the WVS Wave 6 were used, the basis on which the elaborations of this Chapter were grounded are the data from the new WVS/EVS Joint 2017, that, within the framework of WVS, more than ever, have included many worldwide countries. In the same way as De Benedictis et al. (2021), instead, we have benefited from graphical modelling methods to estimate the interdependencies between cultural traits. Given the ordinal and categorical nature of the data, we used the latest Bayesian implementation of Copula Gaussian graphical models. One interesting peculiarity of this approach is that it automatically takes into account the missing values often contained in data from surveys.

The contribution of this Chapter to the definition of cultural distance is in line with that of De Benedictis et al. (2021) and can be divided into two substantial parts. Firstly, each country is culturally identified by the set of marginal distributions of cultural traits, so their distance is measured as the distance between these distributions rather than the synthesis by the mean used by Inglehart and Welzel (2005). Secondly, each country internally is characterized by its own reticular structure of interdependencies between cultural traits, which being not observed is estimated by graphical modelling. The distance between the national cultural networks found in this way, is obtained thanks to a measure of distance between

networks. The DISTATIS method[66] between distances is suited to highlight how this hidden component plays a fundamental role in the complete formulation of cultural distance. Interesting results are also obtained for a larger number of cultural traits than that of the Inglehart-Welzel Cultural Map (which represents another extension of the results of De Benedictis et al. (2021)), demonstrating how the approach uses additional information in the right direction.

In general, the approach used in this Chapter can be easily extended to other data from other Waves or surveys, and to the temporal evolution of cultural distance measurement.

[66] It has been introduced by this Chapter, unlike De Benedictis et al. (2021) in which the PCA was used.

Annex 1

Copula Gaussian graphical models and the BDMCMC implementation

Here, we provide a concise overview of graphical modelling[67] and the Bayesian implementation proposed by Mohammadi et al. (2015) for the estimation of network structures in the context of Gaussian graphical modelling, and adapted in Mohammadi et al. (2017) for Copula Gaussian graphical models. Considering to be the method of recent development and having left to the main text of Section 1.4 only a summary explanation of it and motivation of use, this overview has the only purpose of showing the functioning of the method. Specifically, we remain faithful to the exposition made by Mohammadi et al. (2015, 2017) and taken up by De Benedictis et al. (2021), adding - with respect to the latter - more details about the BDMCMC algorithm.

We start the overview of the method by briefly reviewing a Gaussian graphical model (Ggm), as the copula extension relies on the more traditional Ggm in some suitably defined latent space. Let then $\boldsymbol{Z} = (Z_1, \ldots, Z_p)$ be a vector of p random variables. A Ggm assumes that

$$\boldsymbol{Z} = (Z_1, \ldots, Z_p) \ \sim \ N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\mu}$ is the vector of means and $\boldsymbol{\Sigma}$ is the $p \times p$ covariance matrix. The p-dimensional Gaussian joint distribution is given by

$$f(z|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-p/2} \ |\boldsymbol{\Sigma}^{-1}|^{1/2} \ e^{-1/2(z-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(z-\boldsymbol{\mu})}. \tag{1.1}$$

The matrix $\boldsymbol{K} = \boldsymbol{\Sigma}^{-1}$ is called the precision or concentration matrix and has a special role in Ggm, as it is uniquely associated to the underlying conditional independence graph. In particular, given the k_{ij} element of the $\boldsymbol{K}$ matrix:

$$Z_i \perp Z_j \mid Z_{-ij} \quad \text{if and only if} \quad k_{ij} = 0, \quad i \neq j,$$

that is Z_i is conditionally independent of Z_j given all other variables (Z_{-ij}) if and only if the corresponding (i, j) element of the precision matrix is zero. Equivalently, conditional independence corresponds to zero partial correlation, as

$$\gamma_{ij} = \rho(Z_i, Z_j \mid Z_{-ij}) = -\frac{k_{ij}}{\sqrt{k_{ii} k_{jj}}}, \quad i \neq j. \tag{1.2}$$

In our study, the variables are not normally distributed, as they are categorical or ordinal. Let then $\boldsymbol{Y} = (Y_1, \ldots.., Y_p)$ be the vector of responses to the p survey

[67] For a recent collection of reviews you may see Maathuis et al. (2018).

questions for a certain country. For simplicity, the notation for country is skipped, as this model is performed individually for each country. The main idea of a CGgm is to map these cultural traits in a latent space where a standard Ggm can be used. In particular, a CGgm is defined by

$$\begin{aligned} Y_{ij} &= F_j^{-1}(\Phi(Z_{ij})) \\ Z_1, \ldots, Z_p &\sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{R}), \end{aligned} \tag{1.3}$$

where F_j and F_j^{-1} are the marginal distribution and its pseudo inverse, respectively, for the Y_j variable, $\Phi(\cdot)$ is the cumulative distribution of the univariate Gaussian distribution and $\boldsymbol{R}$ is the correlation matrix of the latent multivariate Gaussian distribution. The two equations combined give the joint distribution of $\boldsymbol{Y}$:

$$P(Y_1 \leq y_1, \ldots, Y_p \leq y_p) = C(F_1(y_1), \ldots, F_p(y_p) \mid \boldsymbol{R}),$$

where $C(\cdot)$ is the Copula Gaussian (Mohammadi et al., 2017).

As mentioned in Section 1.4, inference in a CGgm is traditionally made of two tasks: *parameter estimation*, most notably estimation of the precision matrix $\boldsymbol{K} = \boldsymbol{R}^{-1}$, and *model selection*, that is selecting a Graph where some edges may be missing. In a classical frequentist framework, these two tasks are treated separately: given a Graph, the precision matrix is estimated by constrained maximum likelihood, whereas model selection criteria based on the model likelihood and model complexity (in this case number of edges) are subsequently used to select an optimal Graph (Lauritzen, 1996). In contrast to this, a Bayesian approach allows to account simultaneously for uncertainty both at the level of Graph inference and precision matrix estimation. While Bayesian approaches were computationally prohibitive until not long ago, recent fast implementations have been proposed based on Birth-Death Markov Chain Monte Carlo (BDMCMC) methods both for the case of Ggm (Mohammadi et al., 2015) and CGgm for ordinal and categorical data (Dobra and Lenkoski, 2011; Mohammadi et al., 2017).

As with Bayesian procedures, priors need to be defined on the quantities to be estimated, in this case the precision matrix $\boldsymbol{K}$ and the Graph G. These are then combined with the likelihood function to produce posterior distributions. Starting with the likelihood function, this is defined, in Copula graphical models, using the rank likelihood approach of Hoff (2007), which connects the observed space with the latent one. In particular, given the observed data $\boldsymbol{Y}$, the latent

data $\boldsymbol{Z}$ is constrained to belong to the set

$$\mathcal{D}(\boldsymbol{Y}) = \{\boldsymbol{Z} \in \mathbb{R}^{n\times p} : \; L_j^r(\boldsymbol{Z}) < z_j^{(r)} < U_j^r(\boldsymbol{Z}), \quad r = 1,\dots,n; \quad j = 1,\dots,p\}$$

where:

$$L_j^r(\boldsymbol{Z}) = \max\{Z_j^{(s)} : Y_j^{(s)} < Y_j^{(r)}\}, \qquad U_j^r(\boldsymbol{Z}) = \min\{Z_j^{(s)} : Y_j^{(r)} < Y_j^{(s)}\}$$

and n is the number of observations. For ordinal and categorical data, these sets are intervals between two consecutive y values.[68] Following Hoff (2007), the likelihood function is defined by:

$$P(\boldsymbol{Z} \in \mathcal{D}(\boldsymbol{Y}) \mid \boldsymbol{K}, G) = \int_{\mathcal{D}(\boldsymbol{Y})} P(\boldsymbol{Z}|\boldsymbol{K}, G)\; dZ$$

where $P(\boldsymbol{Z}|\boldsymbol{K}, G)$ is the profile likelihood in the Gaussian latent space, that is

$$P(\boldsymbol{Z}|\boldsymbol{K}, G) \propto |K|^{n/2} \exp\left\{-\frac{1}{2}\text{Trace}(\boldsymbol{K}\boldsymbol{U})\right\}$$

with $\boldsymbol{U} = \boldsymbol{Z}^T\boldsymbol{Z}$ the sample moment. The likelihood function is combined to prior distributions to derive the posterior distribution:

$$P(\boldsymbol{K}, G \mid \boldsymbol{Z} \in \mathcal{D}(\boldsymbol{Y})) \propto P(\boldsymbol{Z} \in \mathcal{D}(\boldsymbol{Y}) \mid \boldsymbol{K}, G)P(\boldsymbol{K}|G)P(G). \tag{1.4}$$

As for prior distributions, Mohammadi et al. (2017) sets a Bernoulli prior on each link of a graph G, leading to a prior on the Graph given by:

$$P(G) \propto \left(\frac{\theta}{1-\theta}\right)^{|E|}$$

where $|E|$ is the number of links in G and $\theta \in (0, 1)$ is the prior probability of a link. Given a Graph G, a G-Wishart distribution $W_G(b, \boldsymbol{D})$ is used to sample a positive definite precision matrix $\boldsymbol{K}$ with zeros corresponding to the missing edges in the Graph G. This conveniently leads to a G-Wishart posterior distribution in the latent space:

$$P(\boldsymbol{K}|\boldsymbol{Z}, G) = \frac{1}{I_G(b^*, \boldsymbol{D}^*)}|\boldsymbol{K}|^{(b^*-2)/2} \exp\left\{-\frac{1}{2}\text{Trace}(\boldsymbol{D}^*\boldsymbol{K})\right\}$$

with $b^* = b + n$ and $\boldsymbol{D}^* = \boldsymbol{D} + \boldsymbol{U}$ the new parameters of the G-Wishart distribution and $I_G(b^*, \boldsymbol{D}^*)$ the normalizing constant for the Graph G (Mohammadi

[68] For missing data, this interval is simply set to $(-\infty, \infty)$ so Bayesian methods can easily accommodate the case of missing data, which is rather common for survey data.

et al., 2017).

Mohammadi et al. (2017) propose an efficient continuous time BDMCMC (birth-death MCMC) process to sample from the joint posterior distribution (1.4). The algorithm is able to explore efficiently the Graph space by adding/removing a link in a birth/death event. It shows notable improving in the computational performance to reach the convergence with respect to other algorithms[69].

In brief, the procedure of the BDMCMC algorithm for Copula Gaussian graphical models concerns three main steps:

1. Data transformation using the copula to sample the latent variables, as described above.
2. The birth and death process for sampling the graph G.
3. Sample the new precision matrix $\boldsymbol{K}$ according to the type of jump between two different graphs.

Basically, given the framework described above, where the prior distributions for G and $\boldsymbol{K}$, and the joint posterior distribution of them, are defined, the algorithm starts from the observed data by sampling the latent variable obtained with the copula transformation.
Based on this sample, the birth and death rates and waiting times are calculated to select the type of jump of the process at each step. This is the core of the algorithm. It decides which edge is born or which edge dies based on the stationary distribution of the process, which determines the birth and death rates, in fact Mohammadi et al. (2015) demonstrated as the BDMCMC algorithm converges to the target joint posterior distribution of the graph and precision matrix (1.4). Each birth or death is an event that makes the process to jump to a new state, in which we have a new graph $G^{+\varepsilon}$ and precision matrix $\boldsymbol{K}^{+\varepsilon}$, where $\varepsilon \in \overline{E}$, if we observe a birth, or a new graph $G^{-\varepsilon}$ and precision matrix $\boldsymbol{K}^{-\varepsilon}$, where $\varepsilon \in E$, if we observe a death. The birth and death processes are independent Poisson processes[70], then the time between two successive events is exponentially distributed (probabilities of birth or death are proportional to their rates), then the waiting time for a new event is defined by the birth and death rates.
From this the jump (birth or death) is selected and a new precision matrix is sampled[71].

[69] The method is implemented in the R package `BDgraph` (Mohammadi and Wit, 2019).
[70] For the efficient calculation of the birth and death rates see Section 3.2 of Mohammadi et al. (2015).
[71] Mohammadi et al. (2015) describes the direct sampler from the precision matrix in Section 3.3.

The algorithm continues to complete a defined number of iterations, and the convergence of it can be evaluated after the end of them. After convergence is found, the posterior on the Graph space is returned, whereby each Graph is given a weight, corresponding to the time the process visited that Graph. This is a peculiar result of this algorithm that can be used to calculate estimates of quantities of interest by Bayesian averaging. In particular, for our analysis, we will consider the:

Posterior Edge Inclusion Probabilities: This is given by:

$$P(\varepsilon \in E \,|\boldsymbol{Y}) = \frac{\sum_{t=1}^{N} 1(\varepsilon \in G^{(t)}) W(\boldsymbol{K}^{(t)})}{\sum_{t=1}^{N} W(\boldsymbol{K}^{(t)})},$$

where E is the set of edges, N is the number of MCMC iterations and $W(\boldsymbol{K}^{(t)})$ is the waiting time for the Graph $G^{(t)}$ with the precision matrix $\boldsymbol{K}^{(t)}$.

Binary network: A single optimal Graph obtained by setting as a default the cut-off 0.5 on the posterior edge inclusion probabilities calculated above.

Partial Correlation network: A single precision matrix $\boldsymbol{K}$ can be obtained from the posterior distribution by Bayesian averaging. From this, a partial correlation matrix can be derived using Equation (1.2). This matrix is a more informative output than the precision matrix as it contains both the intensity and the sign of the relationships between cultural traits.

Choice of the number of iterations for the estimation of cultural networks with 60 cultural traits

Table 1.11: Differences between BDgraph with different number of iterations, applied to 60 cultural traits for Italy

	PARTIAL CORRELATION				PROBABILITY NETWORKS			
	mean	*variance*	*max*	*min*	*mean*	*variance*	*max*	*min*
100,000 vs 200,000	0.00009	0.00006	0.0598	-0.0588	-0.0368	0.0099	0.46	-0.56
200,000 vs 300,000	-0.00008	0.00004	0.0544	-0.0455	-0.0309	0.0068	0.35	-0.44
300,000 vs 400,000	0.00009	0.00002	0.0349	-0.0298	-0.0125	0.0045	0.28	-0.31
400,000 vs 500,000	-0.00002	0.00002	0.0342	-0.0356	-0.0088	0.0047	0.29	-0.35
300,000 vs 500,000	0.00007	0.0003	0.0464	-0.0394	-0.0213	0.0053	0.28	-0.37

Partial correlation matrix and Probability network (intended as adjacency matrix) are estimated for individual data referring to Italy considering 100.000, 200.000, 300.000, 400.000 and 500.000 iterations. From these, only the upper triangle is considered, and the difference distributions between the partial correlations and the probabilities from different number of iterations are calculated. Table 1.11 organizes the synthesis measures of the difference distributions (mean, variance, maximum and minimum). The differences stabilize after 300,000 iterations. That is, passing from 300,000 iterations to 400,000 and 500,000 there is no substantial difference as is instead found between 100,000 and 200,000, 200,000 and 300,000.

The processing time for a machine that has an i7-2600k processor and 16 gigabytes of RAM, was around 135 minutes for Italy. Summarizing:

Table 1.12: Differences between BDgraph with different number of iterations, applied to 60 cultural traits for Italy

	76 countries			
	`Battery dataset`	`IW dataset`	`IW1 dataset`	`Large dataset`
total time (hours)	5.5	7.5	14.5	106

Although for both *Battery dataset* and *IW dataset* we use 10,000 iterations, convergence is found more quickly for *Battery dataset* than *IW dataset*, because the combination of networks between 6 nodes is obviously less than 10 nodes.

List of the 40 selected variables

Table 1.13: List of the 40 selected variables

Joint EVS/WVS Variable Name	**Joint EVS/WVS Variable Label**	**WVS 7 Variable Name**	**WVS 7 Variable Label**	**Maximum missings**
A009	State of health (subjective)	Q47	State of health (subjective)	0.03784
A027	Important child qualities: good manners	Q7	Important child qualities: Good manners	0.08774
A032	Important child qualities: feeling of responsibility	Q10	Important child qualities: feeling of responsibility	0.11970
A066	Member: Belong to education, arts, music or cultural activities	Q96R	Active/Inactive membership of art, music, educational	0.14412
A071	Member: Belong to conservation, the environment, ecology, animal rights	Q99R	Active/Inactive membership of environmental organization	0.14191

Table 1.13: List of the 40 selected variables

Joint EVS/WVS Variable Name	Joint EVS/WVS Variable Label	WVS 7 Variable Name	WVS 7 Variable Label	Maximum missings
A074	Member: Belong to sports or recreation	Q95R	Active/Inactive membership of sport or recreation	0.11485
A124_03	Neighbours: Heavy drinkers	Q24	Neighbours: Heavy drinkers	0.12917
A124_08	Neighbours: Drug addicts	Q18	Neighbours: Drug addicts	0.12917
A170	Satisfaction with your life	Q49	Satisfaction with your life	0.03027
A173	How much freedom of choice and control	Q48	How much freedom of choice and control	0.04750
C001	Jobs scarce: Men should have more right to a job than women (3-point scale)	Q33R	When jobs are scarce, men should have more right to a job than women	0.05741
C002	Jobs scarce: Employers should give priority to (nation) people than immigrants (3-point scale)	Q34R	When jobs are scarce, employers should give priority to people of this country over immigrants	0.05109
C038	People who don't work turn lazy	Q39	People who don't work turn lazy	0.07982
C039	Work is a duty towards society	Q40	Work is a duty towards society	0.06055
C041	Work should come first even if it means less spare time	Q41	Work should always come first even if it means less spare time	0.04447
D001_B	How much you trust: Your family (B)	Q58	How much you trust: Your family (B)	0.02460
D026_03	Duty towards society to have children	Q37	Duty towards society to have children	0.05676
D026_05	It is childs duty to take care of ill parent	Q38	It is childs duty to take care of ill parent	0.07243
D060	University is more important for a boy than for a girl	Q30	University is more important for a boy than for a girl	0.13821
E023	Interest in politics	Q199	Interest in politics	0.03500
E035	Income equality	Q106	Income equality	0.07379
E037	Government responsibility	Q108	Government responsibility	0.07096
E039	Competition good or harmful	Q109	Competition good or harmful	0.07571
E069_06	Confidence: The Police	Q69	Confidence: The Police	0.07267
E111_01	Satisfaction with the political system	Q252	Rate political system for governing country	0.09830
E235	Importance of democracy	Q250	Importance of democracy	0.09830
E236	Democraticness in own country	Q251	How democratically is this country being governed today	0.11826
F028	How often do you attend religious services	Q171	How often do you attend religious services	0.08116

Table 1.13: List of the 40 selected variables

Joint EVS/WVS Variable Name	**Joint EVS/WVS Variable Label**	**WVS 7 Variable Name**	**WVS 7 Variable Label**	**Maximum missings**
F034	Religious person	Q173	Religious person	0.10939
F116	Justifiable: Cheating on taxes	Q180	Justifiable: Cheating on taxes	0.06051
F117	Justifiable: Someone accepting a bribe	Q181	Justifiable: Someone accepting a bribe	0.04267
F121	Justifiable: Divorce	Q185	Justifiable: Divorce	0.14119
F123	Justifiable: Suicide	Q187	Justifiable: Suicide	0.13418
G007_18_B	Trust: Your neighborhood (B)	Q59	Trust: Your neighborhood (B)	0.07243
G007_33_B	Trust: People you know personally (B)	Q60	Trust: People you know personally (B)	0.03315
G052	Evaluate the impact of immigrants on the development of [your country]	Q121	Evaluate the impact of immigrants on the development of [your country]	0.10473
G255	How close you feel: Your village, town or city	Q255	How close you feel: Your [village, town or city]	0.04654
G257	How close do you feel: to country	Q257	I see myself as citizen of the [country] nation	0.06426
H009	Government has the right: Keep people under video surveillance in public areas	Q196	Government has the right: Keep people under video surveillance in public areas	0.09534
H011	Government has the right: Collect information about anyone living in [COUNTRY] without their knowledge	Q198	Government has the right: Collect information about anyone living in [COUNTRY] without their knowledge	0.13659

References

Abdi, H., O'Toole, A. J., Valentin, D., and Edelman, B. (2005). Distatis: The analysis of multiple distance matrices. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops*, pages 42–42. IEEE.

Abdi, H. and Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459.

Alemán, J. and Woods, D. (2016). Value orientations from the world values survey: How comparable are they cross-nationally? *Comparative Political Studies*, 49(8):1039–1067.

Alesina, A. and Giuliano, P. (2015). Culture and Institutions. *Journal of Economic Literature*, 53(4):898–944.

Baecker, D. (1997). The meaning of culture. *Thesis Eleven*, 51:37–51.

Billard, L. and Diday, E. (2003). From the statistics of data to the statistics of knowledge: symbolic data analysis. *Journal of the American Statistical Association*, 98(462):470–487.

Bock, H.-H. and Diday, E. (2012). *Analysis of symbolic data: exploratory methods for extracting statistical information from complex data*. Springer Science & Business Media.

Bourdieu, P. (1977). *Outline of a Theory of Practice*, volume 16. Cambridge University Press.

Brady, H. E. and Collier, D. (2010). *Rethinking social inquiry: Diverse tools, shared standards*. Rowman & Littlefield Publishers.

Cohen, L., Manion, L., and Morrison, K. (2017). *Research methods in education*. routledge.

Corbetta, P. (2014). *Metodologia e tecniche della ricerca sociale*. il Mulino Bologna.

Cox, M. A. and Cox, T. F. (2008). Multidimensional scaling. In *Handbook of data visualization*, pages 315–347. Springer.

Creswell, J. W. and Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

De Benedictis, L., Rondinelli, R., and Vinciotti, V. (2021). The network structure of cultural distances. *arXiv:2007.02359*.

Diday, E. and Noirhomme-Fraiture, M. (2008). *Symbolic data analysis and the SODAS software*. John Wiley & Sons.

Dobra, A. and Lenkoski, A. (2011). Copula Gaussian graphical models and their application to modeling functional disability data. *The Annals of Applied Statistics*, 5(2A):969–993.

Durkheim, É. (1912). *Les formes élémentaires de la vie religieuse: le système totémique en Australie*, volume 4. Alcan.

Ferrarotti, F. (1986). *Manuale di sociologia*. Laterza.

Flick, U. (2018). *An introduction to qualitative research*. sage.

Giddens, A. (2000). *Fondamenti di sociologia*. il Mulino.

Guiso, L., Sapienza, P., and Zingales, L. (2006). Does culture affect economic outcomes? *Journal of Economic Perspectives*, 20(2):23–48.

Guiso, L., Sapienza, P., and Zingales, L. (2008). Social capital as good culture. *Journal of the European economic Association*, 6(2-3):295–320.

Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., Lagos, M., Norris, P., Ponarin, E., Puranen, B., et al. (2020). World values survey: round seven-country-pooled datafile. *Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat*.

Hammersley, J. M. and Clifford, P. (1971). Markov fields on finite graphs and lattices. *Unpublished manuscript*.

Hammond, D. K., Gur, Y., and Johnson, C. R. (2013). Graph diffusion distance: A difference measure for weighted graphs based on the graph Laplacian exponential kernel. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 419–422. IEEE.

Hirschman, A. O. (1970). *Exit, voice, and loyalty: Responses to decline in firms, organizations, and states*, volume 25. Harvard university press.

Hoff, P. D. (2007). Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics*, 1(1):265–283.

Hoff, P. D. (2009). Multiplicative latent factor models for description and prediction of social networks. *Computational and mathematical organization theory*, 15(4):261.

Hofstede, G. (1984). *Culture's consequences: International differences in work-related values*, volume 5. sage.

Hofstede, G. (2011). Dimensionalizing cultures: The Hofstede model in context. *Online readings in Psychology and Culture*, 2(1):8.

House, R. J., Hanges, P. J., Javidan, M., Dorfman, P. W., and Gupta, V. (2004). *Culture, leadership, and organizations: The GLOBE study of 62 societies*. Sage publications.

Husson, F., Lê, S., and Pagès, J. (2017). *Exploratory multivariate analysis by example using R*. CRC press.

Ichino, M. and Yaguchi, H. (1994). Generalized Minkowski metrics for mixed feature-type data analysis. *IEEE Transactions on Systems, Man, and Cybernetics*, 24(4):698–708.

Inglehart, R. (1997). *Modernization and postmodernization: Cultural, economic, and political change in 43 societies*. Princeton university press.

Inglehart, R. (2018). *Cultural evolution: people's motivations are changing, and reshaping the world*. Cambridge University Press.

Inglehart, R. and Baker, W. E. (2000). Modernization, cultural change, and the persistence of traditional values. *American Sociological Review*, pages 19–51.

Inglehart, R. and Norris, P. (2003). *Rising tide: Gender equality and cultural change around the world*. Cambridge University Press.

Inglehart, R. and Welzel, C. (2005). *Modernization, cultural change, and democracy: The human development sequence*. Cambridge University Press.

Jen, M. H., Jones, K., and Johnston, R. (2009). Global variations in health: evaluating wilkinson's income inequality hypothesis using the world values survey. *Social science & medicine*, 68(4):643–653.

Kroeber, A. L. and Kluckhohn, C. (1952). Culture: A critical review of concepts and definitions. *Papers. Peabody Museum of Archaeology & Ethnology, Harvard University*.

Kroeber, A. L. and Parsons, T. (1958). The concepts of culture and of social system. *American Sociological Review*, 23(5):582–583.

Lauritzen, S. L. (1996). *Graphical Models*. Clarendon Press.

Lévi-Strauss, C. (1987). *Introduction to the work of Marcel Mauss*. Taylor & Francis.

Maathuis, M., Drton, M., Lauritzen, S., and Wainwright, M. (2018). *Handbook of graphical models*. CRC Press.

Mahoney, J. and Goertz, G. (2006). A tale of two cultures: Contrasting quantitative and qualitative research. *Political analysis*, pages 227–249.

Minkov, M. and Hofstede, G. (2012). Hofstede's fifth dimension: New evidence from the world values survey. *Journal of cross-cultural psychology*, 43(1):3–14.

Mohammadi, A., Abegaz, F., van den Heuvel, E., and Wit, E. C. (2017). Bayesian modelling of Dupuytren disease by using Gaussian copula graphical models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(3):629–645.

Mohammadi, A. and Wit, E. C. (2019). BDgraph: An R Package for Bayesian Structure Learning in Graphical Models. *Journal of Statistical Software*, 89(3).

Mohammadi, A., Wit, E. C., et al. (2015). Bayesian structure learning in sparse Gaussian graphical models. *Bayesian Analysis*, 10(1):109–138.

Patton, M. Q. (2005). Qualitative research. *Encyclopedia of statistics in behavioral science*.

Rapoport, H., Sardoschau, S., and Silve, A. (2020). Migration and cultural change. *CESifo Working Paper*.

Sampson, R. J. and Laub, J. H. (1995). *Crime in the making: Pathways and turning points through life*. Harvard University Press.

Schaffer, B. S. and Riordan, C. M. (2003). A review of cross-cultural methodologies for organizational research: A best-practices approach. *Organizational research methods*, 6(2):169–215.

Schwartz, S. H. (2008). Cultural value orientations: Nature and implications of national differences. *Moscow: Publishing house of SU HSE*.

Schwartz, S. H. and Bilsky, W. (1987). Toward a universal psychological structure of human values. *Journal of personality and social psychology*, 53(3):550.

Silverman, D. (2020). *Qualitative research*. Sage Publications Limited.

Taras, V., Rowney, J., and Steel, P. (2009). Half a century of measuring culture: Review of approaches, challenges, and limitations based on the analysis of 121 instruments for quantifying culture. *Journal of International Management*, 15(4):357–373.

Torgerson, W. S. (1958). *Theory and methods of scaling*. Wiley.

Trompenaars, F. and Hampden-Turner, C. (2011). *Riding the waves of culture: Understanding diversity in global business*. Nicholas Brealey International.

Tylor, E. B. (1871). Primitive culture: Researches into the development of mythology, philosophy, religion, language, art and custom. *NY, US: Henry Holt and Company. xii*.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press.

Weber, M. (1904). Die "objektivität" sozialwissenschaftlicher und sozialpolitischer erkenntnis. *Archiv für sozialwissenschaft und sozialpolitik*, 19(1):22–87.

Chapter 2

Comparing cultural distances with other distances among countries

Abstract

The effect of cultural distance on the economic situation of a country or more generally of a geographically definable area, has been scoured in recent years by the economic literature. Cultural, genetic, geographical, climatic, semantic, ethnic, linguistic, political distances have often been included in econometric models as independent or control variables. This Chapter follows this literature, firstly by individually comparing three measurements of cultural distance calculated in Chapter 1 with other distances used in literature together with cultural distance or as a proxy of it, and secondly by jointly comparing them (the measurements of cultural distance and those from literature) via DISTATIS. The three cultural distances are the two new measures mentioned above (*Compromise_Large* and *Compromise_IW*) and the *IW index* obtained as Euclidean distance between countries in the Inglehart-Welzel Cultural Map, while the other distances take into consideration climatic condition, ethnicity and language, genetics and the recent phenomenon of Facebook. Finally, it considers these distance measures into a Social Relations Regression Model (SRRM) over the distance between countries in GDP per capita (year 2017). The final result shows that cultural distances are poorly correlated with the other distances, and when a compromise is found between them, usually the *Compromise_Large* distance is characterized by a slightly higher weight. The main conclusion regards the important explanatory power of the distance *Compromise_Large* on the distance in GDP per capita compared to *IW index* and the *Compromise_IW*, which has an intermediate meaning between the two.

2.1 Introduction

From Chapter 1 we conclude that in order to properly measure national culture, and therefore the cultural distance between countries, it is important to include the cultural traits interdependencies of the cultural network of each country. This is verified both by following the approach used by Ronald Inglehart and Christian Welzel, i.e. using the 10 cultural traits selected by them to construct their Cultural Map, and - with even more evident results - by adding other variables (cultural traits) for a total of 60.

Everyday real events clearly show how distances of any kind can have a decisive influence on the economic and technological development of a country, or more generally of a group of individuals. Similarly, development brings with it changes in people's attitudes, which are reflected in other areas of life and so on. Technological progress has a current effect on the climate, which can inevitably have an impact on people's way of life, their habits and their culture. From this point of view geographical, climatic, cultural, linguistic and genetic distances are closely linked to economic development. The Greeks, the Roman Empire, the advent of Christianity, the discovery of America, colonization, the industrial revolution, trade with the Indies, the world wars, up to the current pandemic. These are all examples of events that, not only have affected the geography and movement of human beings on Earth, but have also contaminated human cultures, languages and even genes. It is not the purpose of this Chapter to make an excursus of the history, but recall sometimes how everything is connected is a good practice. For example, a congestion of the traffic in the Suez Canal (March 2021) was enough to make us realize that geographical distances in the world still matters nowadays. It sufficed that one virus was more "aggressive" than others (January 2020) to point out that economic investments in research and health are not enough, and how social culture and political situation have played an important role in the spread of the pandemic country by country. A series of chain mechanisms affecting a very wide sphere of human status has been invested by this.

In this regard, Guiso et al. (2006) provides an excellent review of the point of origin of the debate on economics and culture, and opens economists to several questions. The recent (but also past) economic literature has endeavoured to highlight the importance of all these things in explaining the economic condition in which countries pour, and viceversa. For example, as mentioned in Section 1.2, the birth of the World Values Survey project (repeated over time) rests on the need to measure how technological and economical development exacerbates cultural changes. It is possible to affirm that culture determines and undergoes the human economic and technological development. Economists and other scholars have often solved this endogeneity demonstrating as some other exogenous distances can be used as proxy of the cultural distance (Fearon, 2003; Özak, 2018; Desmet et al., 2007; Obradovich et al., 2020).

On the basis of this literature, in this Chapter we try to support the measures found in Chapter 1. In the way of Alesina et al. (2003) and Spolaore and Wacziarg (2016), in Section 2.2 we compare the *Compromise_Large*, *Compromise_IW* and *IW index* with some distances used as independent or control variables within recent economic literature works. This comparison is only used to check how the cultural distance measurements calculated by us, interact with other well-known measures used in the literature. Specifically, we use the recent distance due to the facebook Social Connectedness Index (SCI), a joint distance of the ethnic and linguistic fractionalization of the countries, two measures of genetic distance and a climatic distance we built from the Koeppen-Geiger climate classification. In Section 2.3 we jointly compare all the distances in the aim of understanding the weight that each of them may have in a hypothetical compromise measure. Finally, in Section 2.4 we estimate a model for the distance in GDP per capita (year 2017) between countries, in which we highlight the central role of the cultural distance measures calculated in Chapter 1.

2.2 Proposed cultural distance versus other distance measures

In this paragraph we compare in a descriptive way cultural distances proposed in Chapter 1 with other popular distance measures used as indipendent variables or controls inside econometric models. For this comparison, the lower triangle of the distance matrices is vectorized, in the way for each scatter plot of Figures 2.1, 2.2, 2.3, 2.4 and 2.5, points depict the country pairwise distance, measured with three cultural distance indices estimated in Chapter 1 and different other distances including a measure of connectivity, ethnic-linguistic, genetic and climatic dissimilarities. As cultural distances, for each Figure we consider: (a) cultural distance *Compromise_Large* from cultural traits of the *Large dataset*; (b) cultural distance *Compromise_IW* from cultural traits of the *IW dataset*, namely variables used in the Inglehart-Welzel Cultural Map (Inglehart and Welzel, 2005); (c) cultural distance *IW index* from the Inglehart-Welzel Cultural Map[1].

The objective of this operation is to have a general idea on two important aspects concerning the cultural distance measures estimated in Chapter 1. The first is the level of correlation between cultural distance and other distances used in literature. In other words, we are interested in defining how much the cultural measures proposed in this thesis provide information other than different distance

[1] For their specific formulation see Section 1.6.

measures, which are relevant for economic problems and often used as a proxy for cultural distance. The second is to verify - by different means than those used in the first Chapter - the differences in information between the measures of cultural distance that take into account the cultural network of the countries (interdependencies among cultural traits within each country) and that obtained calculating the Euclidean distance between the projected countries in the Inglehart-Welzel Cultural Map. Our hypothesis is that the three cultural distances identify a different concept of distance (or a concept of higher level) than the other measures, i.e. we expect levels of correlation pretty low or close to zero. In a certain way, this may be at odds with current literature, but the different ways of empirically defining cultural distance may be decisive in admitting or not admitting a certain relationship. At the same time we expect there to be some differences, albeit minimal, between the three cultural distances and mainly between the two calculated by combining distributional and network parts and that due solely to the distances on the Inglehart-Welzel Cultural Map.

2.2.1 Facebook Social Connectedness Index

The first considered measure is the recent facebook Social Connectedness Index (SCI) proposed by Bailey et al. (2018). This index *"uses an anonymized snapshot of active facebook*[2] *users and their friendship networks"*. Each user is assigned to a location based on its activities, the devices connected to facebook and the location mentioned in its facebook profile, then the index is defined as:

$$SCI_{l,m} = \frac{Connections_{l,m}}{Users_l * Users_m}$$

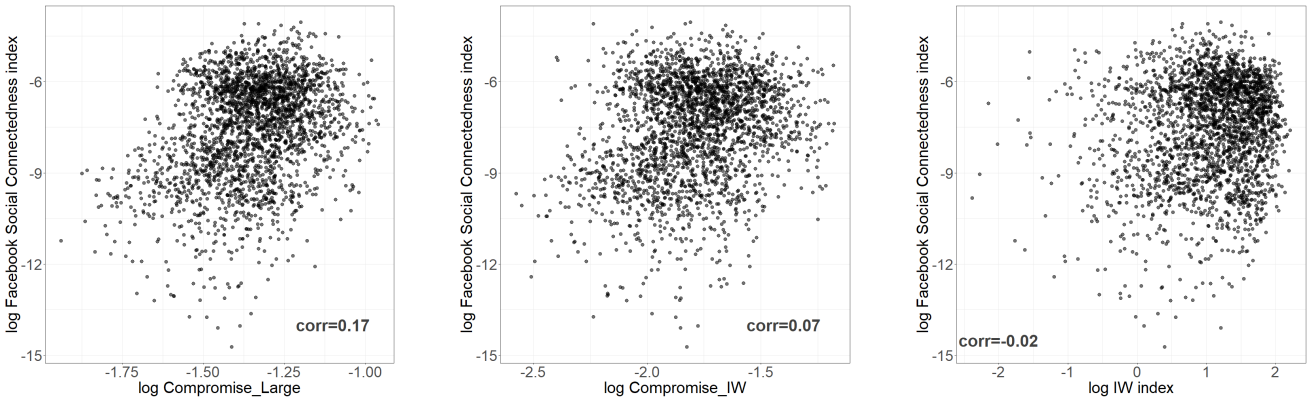
where, $Users_l$ and $Users_m$ are respectively the total number of facebook users located in the country l and country m, while $Connections_{l,m}$ is the total number of friendship connections between users of two locations[3]. This index defines the relative probability of a facebook friendship between different geographical locations. It can be considered as the strength of the connectivity between two countries and then as a proximity measure (values of the index increase when two countries are more connected). The need of comparison led us to transform it in a distance measure. For the strong asymmetry of the index distribution we excluded the conversion via the maximum value of the distribution, then we carried out the transformation by using $d_{lm} = 1/s_{lm} + \epsilon$ (the inverse operation

[2] In the last years the use of data from facebook to measure the cultural distance and the connectedness is increasing. For example, see Obradovich et al. (2020); Kosinski et al. (2015); Huang and Park (2013).

[3] For further details see https://data.humdata.org/dataset/social-connectedness-index.

showed by Wierzchoń and Kłopotek 2018), where d_{lm} is the final distance measure between countries l and m, s_{lm} is the "similarity" measure (in this case the SCI) between countries l and m, and ϵ is considered in this case as null. Finally for graphical representation purpose we calculated the logarithm of it and of all the three cultural distances (Figures 2.1a, 2.1b and 2.1c). Bailey et al. (2018) uses

Figure 2.1: Cultural distance vs. facebook Social Connectedness Index



(a) (b) (c)

Note: Facebook Social Connectedness Index is not available for countries where facebook is banned or countries with few active users. For our set of countries SCI is not available for *Andorra*, *China*, *Iran* and *Russia*, then we exclude these countries also from our cultural distances.

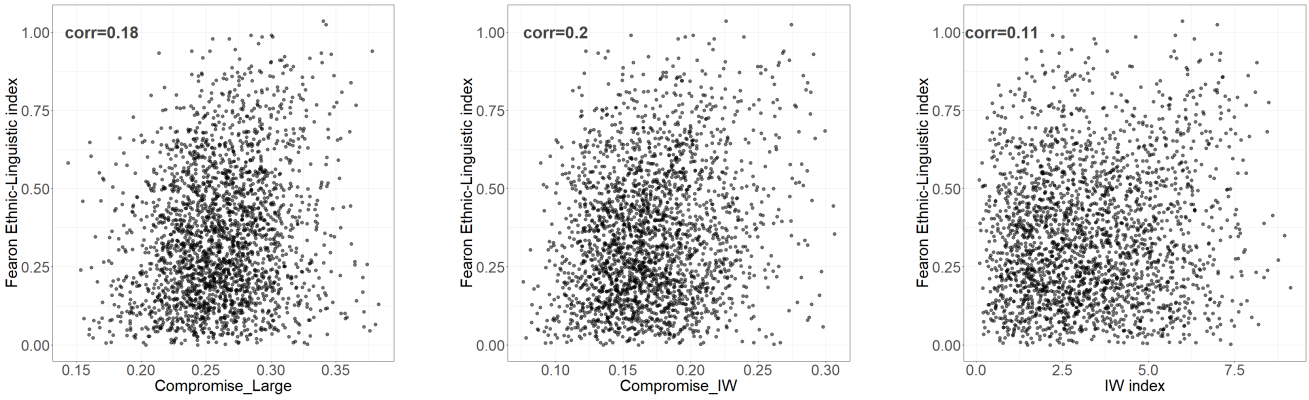
this measure in context of US countries. They set a solid foundation to study a variety of real economic and social problems by highlighting its correlation with social and economic phenomena. In Figure 2.1 we compare the measure with our cultural distances. Hypothesis stated in the first part of this Section are fully reflected in this paragraph, in fact, correlations (reported inside each figure) correspond to values close to zero, while we have a marked difference between Figures 2.1a and 2.1c. In this case, including more cultural traits in the cultural distance definition changes the relationship with the SCI, which does not seem to be completely different (at least at an aggregate level) for Figures 2.1b and 2.1c that use the Cultural Map variables.

2.2.2 Combined ethnic and linguistic distance

Fearon (2003) proposed two measures of fractionalization extensively used in the current literature within econometric models. The first is a measure of ethnic fractionalization, the second uses the linguistic distance as a proxy of the cultural distance. Several authoritative works have employed and developed distance measures from the fractionalizations proposed by James D. Fearon. For example, Özak (2018) used these measures to validate its Human Mobility In-

dex with Seafaring (HMISea)[4]; while in Spolaore and Wacziarg (2009) authors made use of data from Fearon (2003) to construct indices of linguistic distance between countries and insert them in the regression model of the absolute value of logarithm income differences (year 1995); the same authors explored in Spolaore and Wacziarg (2016) the relationships between two measures of linguistic distance calculated from Fearon (2003) data, and other distance measures, such as genetic, religious and cultural[5]; The same measures are used in Spolaore and Wacziarg (2018) in the regression model of logarithm income per capita (year 2005).

Figure 2.2: Cultural distance vs. Fearon Ethnic-linguistic index



(a) (b) (c)

Note: Ethnic and linguistic distances are not available for *Andorra*, *Puerto Rico*, *Macao*, *Iceland*, *Hong Kong*, *Myanmar*, then we exclude these countries also from our cultural distances. Since data used by Fearon (2003) refer to 1990, for *Serbia* and *Montenegro* we consider outputs for *Yugoslavia* (they were included in a unique State in that period). While, for *Germany* we use *Germany Federal Republic*, which has fractionalization values not so different from *German Democratic Republic*.

As James D. Fearon suggests - despite the "slippery" of the concept of an ethnic group - fractionalizations proposed by him have a good correlation with those proposed by other researchers, e.g. like of Alesina et al. (2003). *"Fractionalization is defined as the probability that two individuals selected at random from a country will be from different ethnic-linguistic groups"*, namely when the value of fractionalization associated to a country is low means perfect homogeneity, while when the value is close to 1 there is a perfect heterogeneity. Since the measure of ethnic fractionalization could equate countries which ultimately have substantial linguistic/cultural differences, the linguistic measure complements the level of ethnic fractionalization. For this reason, we use both measures of the Ap-

[4] It estimates the time required to cross each square kilometer on land and sea, and it was used by Omer Ozak as proxy of the distance to the technological frontier in the pre-industrial period.

[5] This cultural measure is obtained from World Values Survey data, while we will talk in the next paragraph about the genetic measure.

pendix of Fearon (2003) to obtain a combined measure of ethnic and linguistic distances between the countries we have considered. Specifically, we calculated the euclidean distance between countries over the these two variables.

The calculated distance is related to the three cultural distances in Figure 2.2. Hypothesis stated in the first part of this Section are verified also in this case. Correlations have positive low values, more structured for network-based distances (Figures 2.2a and 2.2b) than *IW index* (Figure 2.2c). Differently from Section 2.2.1, here, we observe a slightly higher correlation of the ethnic-linguistic distance with *Compromise_IW* than that with *Compromise_Large*. Including a considerable number of cultural traits does not seem to have an important effect in the information carried by the cultural distance compared to that calculated by James D. Fearon's fractionalizations.

2.2.3 Genetic distance

In the context of genetic distance there is an impressive literature headed by the work of Cavalli-Sforza et al. (1994), from which have derived many of the distance measures proposed in economic research. As an example, Spolaore and Wacziarg (2009) calculates a measure of genetic distance[6] between countries by matching the populations considered by Cavalli-Sforza et al. (1994) with the ethnicities by country listed in Alesina et al. (2003). The measure elaborated by Spolaore and Wacziarg (2009) is also used by Omer Ozak in Özak (2018), where he shows how his HMISea has a strong explanatory power towards this measure of genetic distance. Desmet et al. (2007) also uses data from Cavalli-Sforza et al. (1994), here, authors compare the genetic distance with the cultural one finding a high correlation between them, even after having added the control for the linguistic distance. The cultural distance is calculated from the 430 questions of the first four Waves of WVS. Distance between two countries is calculated as the Manhattan distance on the percentage of people for each possible answer of each question[7].

We make use of the genetic distance proposed in Spolaore and Wacziarg (2016) and firstly suggested by Wright (1949). They use two distances: one based on the dominant genetic trait in a country and another weighted on the various genetic traits found in a country. Although the latter is more specific, we choose to use

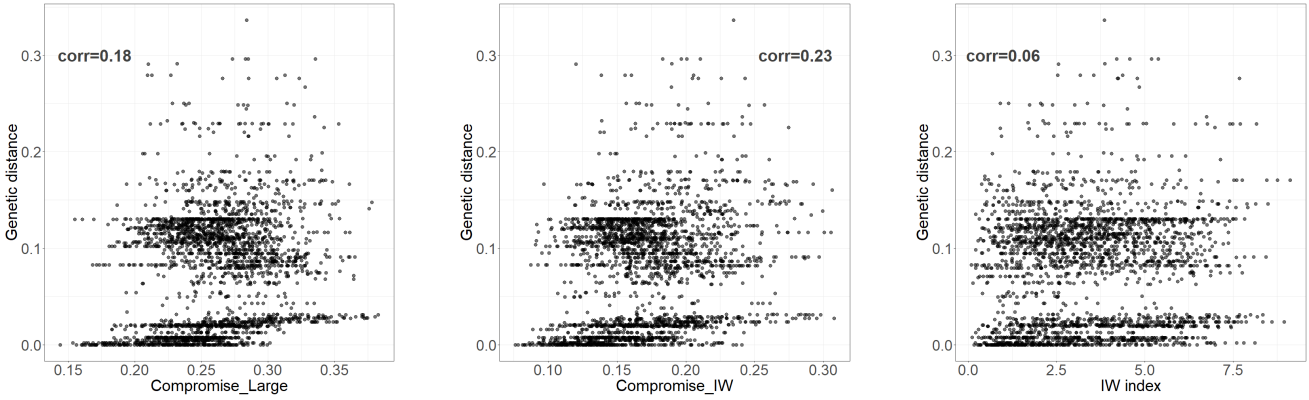
[6] More generally, Enrico Spolaore and Romain Wacziarg have opened a line of works, in which they exploit the information provided by Cavalli-Sforza et al. (1994). See for example Spolaore and Wacziarg (2013), Spolaore and Wacziarg (2016) and Spolaore and Wacziarg (2018).

[7] Results remain robust when using the Euclidean distance.

the former as strongly related to the latter (0.93) and less affected by missings values on the countries we consider.

Secondly, we use a new genetic measure proposed by Spolaore and Wacziarg (2018)[8], deriving from the work of Pemberton et al. (2013)[9]. The latter, as Cavalli-Sforza et al. (1994), proposes the genetic distance for populations, so as done by Spolaore and Wacziarg (2009), even here, the matching between the measure of Pemberton et al. (2013) and the ethnic groups by country proposed by Alesina et al. (2003) needs. Spolaore and Wacziarg (2018) overall elaborates 2 measures in line with Spolaore and Wacziarg (2016) and a third that considers the populations as they were in the AD 1500. We choose the latter to differentiate the analysis from the previous one.

Figure 2.3: Cultural distance vs. Genetic distance



(a) (b) (c)

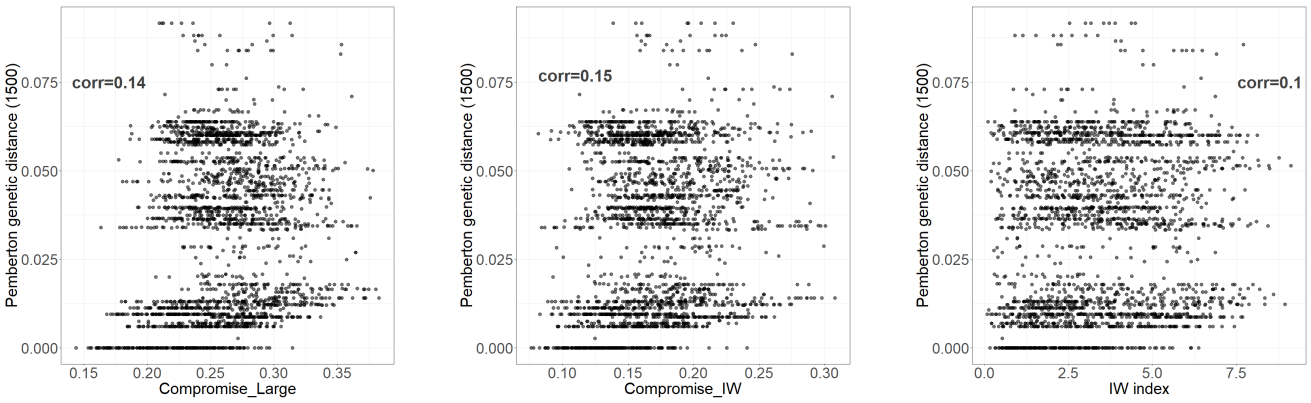
Note: Genetic distance is not available for *Andorra*, *Bosnia and Herzegovina*, *Montenegro* and *Serbia*, then we exclude these countries also from our cultural distances.

The genetic distance of Spolaore and Wacziarg (2016) is depicted with the three cultural distances in Figure 2.3. The association between the variables seems to be very similar to that found for the ethnic-linguistic distance in Section 2.2.2. The genetic distance measure is little correlated with all three measures, but much less with the *IW index*. Considering a larger number of cultural traits does not increase the correlation, indeed it seems to confuse it even more: the correlation between *Compromise_IW* and the genetic distance is 0.23 against 0.18 observed for *Compromise_Large*.

[8] Both this and Spolaore and Wacziarg (2009) (as well as Özak 2018 and Ashraf and Galor 2013) contribute to explore *"how the characteristics of a society's ancestral population exert an influence on its current level of development"*.

[9] Compared to the measure of Cavalli-Sforza et al. (1994), this has the privilege of allowing a better matching with the populations of African and Asian countries Spolaore and Wacziarg (2018).

Figure 2.4: Cultural distance vs. Pemberton genetic distance



(a) (b) (c)

Note: This genetic distance is not available for *Andorra*, *Bosnia and Herzegovina*, *Montenegro*, *Serbia*, *Puerto Rico* and *Macao* then we exclude these countries also from our cultural distances.

Moving to the genetic distance supposed for countries' composition in AD 1500 (Spolaore and Wacziarg, 2018), relations slightly change their shape. There is a flattening of the correlation along the three measures of cultural distance, where for the *IW index* it is always slightly lower than *Compromise_Large* and *Compromise_IW*, for which it is almost the same. We could imagine that the genetic differences country by country were more pronounced in AD 1500 than the current period. This inevitably merges with the current culture, which could be less divisive in the distinctions between countries than in the past[10]. This can be favored by the technological evolution and the fall of some geographical barriers.

2.2.4 Climatic distance

Finally, we compare the cultural distances with a measure of climatic dissimilarity between countries. In literature exist a variety of contributions about climatic control for analyzing economic problems, some of them making use of the Koeppen-Geiger classification of climate zones. Climatic conditions are surely fundamental for the definition of the HMISea (Özak, 2018), in fact the speed of travel of a distance by land and sea surely depends on the climatic conditions. These are demonstrated to be decisive even for the spread of agriculture (Ashraf and Galor, 2011) and in general for technology (Sachs, 2001)[11]. In this context,

[10] As stated in Spolaore and Wacziarg (2016), *"genetic distance measures relatedness between populations and is roughly proportional to time since two populations shared the same ancestors, that is, since they were the same population. Over time, ancestors transmit a large number of traits to their descendants, not only biologically (through DNA), but also culturally"*. At the same time, when two countries did not share a similar genetic population in the past, it is not immediate they have a current low level of cultural distance. The two cultures may have been mixed by some historical event.

[11] For a precise description of the use of climatic conditions in the study of economical processes, see Spolaore and Wacziarg (2013).

the important role of similarity of two countries in relation to the climatic conditions has been defined by Gallup et al. (1999) together with the latitudinal position which may define barriers to technological development. In this paper, authors have used 12 Koeppen-Geiger climate zones providing data about the climatic composition of several countries. Spolaore and Wacziarg (2009) uses these data as benchmark to develop two measures of similarity between countries. The first is calculated as *"the average absolute value difference, in the percentage of land area in each of the 12 climate zones"*[12] and the second as *"the absolute difference in the percentage of land areas in tropical climates"*.

In the way of Gallup et al. (1999) and Spolaore and Wacziarg (2009), we constructed a measure of distance based on the climatic conditions of the land area of each country according to the Koeppen-Geiger climatic classification (Kottek et al., 2006; Rubel and Kottek, 2010; Koeppen et al., 2011; Rubel et al., 2017). Here, we collect geographical coordinates for each country, then we attribut each coordinate to its observed climatic condition according to the 31 climate subzones[13] defined by the Koeppen-Geiger classification. Finally, we have a vector of sum equal to 1 for each country, containing the percentage of land (represented by coordinates) for each Koepper-Geiger climatic zone[14]. We calculated the Euclidean distance between each pair of country to obtain the final climatic distance[15].

Although the hypothesis made at the beginning of the Section tend to be satisfied, compared to the previous distance measures (Sections 2.2.1,2.2.2,2.2.3), correlations between cultural distances and the climatic distance are controversial. As first, values are overall slightly higher and, as second, correlation with *IW index* is higher than correlation with *Compromise_Large* and *Compromise_IW*. If for the previous cases the addition of the component of interdependence between cultural traits and the increase of the number of cultural traits has contributed to add to the measure of cultural distance a certain component of information

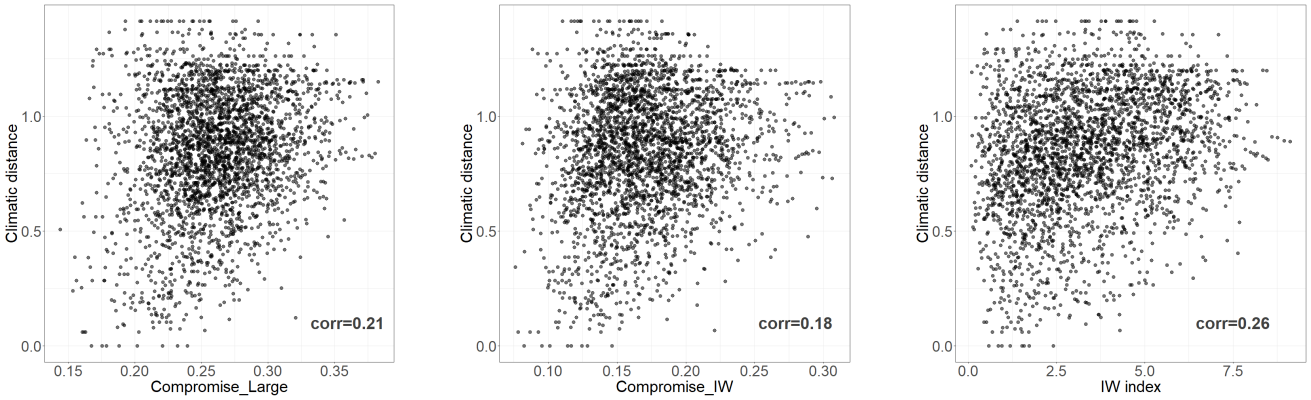
[12] This measure was used also in Spolaore and Wacziarg (2018) as control variable inside the regression over the income.

[13] Differently from Gallup et al. (1999) we do not use 12 climate zones, but all of them. The 31 climate zones are: Af, Am, As, Aw, BWk, BWh, BSk, BSh, Cfa, Cfb, Cfc, Csa, Csb, Csc, Cwa, Cwb, Cwc, Dfa, Dfb, Dfc, Dfd, Dsa, Dsb, Dsc, Dsd, Dwa, Dwb, Dwc, Dwd, EF, ET. They are defined by different characteristics. The *main climate*: equatorial (A), arid (B), warm temperate (C), snow (D), polar (E). The *precipitations*: desert (W), steppe (S), fully humid (f), summer dry (s), winter dry (w), moonsonal (m). The *temperature*: hot arid (h), cold arid (k), hot summer (a), warm summer (b), cool summer (c), extremely continental (d), polar frost (F), polar tundra (T).

[14] The association of the coordinates with the countries under analysis was done through the R function `joinCountryData2Map` contained in the R package `rworldmap`, while for the overlap of coordinates with climatic zones we make use of the option `LookupCZ` contained in the R package `kgc`.

[15] Usually data of this type are processed within the field of analysis of compositional data. Anyway, when the number of variables is quite large it is possible to use a classic approach on them.

Figure 2.5: Cultural distance vs. Climatic distance



(a) (b) (c)

superimposed on the other distances, in this case it seems that the reverse operation takes place, that is, the more precise measurement of the cultural distance clears it from a part of information related to the climatic conditions of a country.

In any case, results follow the conceptual intuition. We imagine current culture is driven by a variety of historical and current human mechanisms, of which the climatic condition, the genetic background, ethnic and linguistic differences, and more recent phenomena like facebook, are some of the foremost. In the same way, each component of this variety that defines such a broad concept as culture, cannot explain a substantial part of it. This is a small part when the measure is specified by considering interdependencies among cultural traits and becomes almost zero when interdependencies are not considered (*IW index*).

2.3 Between-distances analysis

To make a joint analysis of all distances and, at the same time, for incorporating them into a single general measure, we can use the DISTATIS method, already used successfully in the previous Chapter.

The measures we have selected try to cover roughly the wide range of measures proposed in the literature, without overloading the analysis with too many comparisons. We use as a focus the cultural measures produced in the Chapter 1, which represent the core of this thesis and the fixed point of comparison. We therefore adopt a measure of distance that can take into account current social mechanisms, namely the facebook Social Connectedness Index (SCI). The eth-

nic and linguistic measure covers the concept of ethnic fractionalisation of the inhabitants of a country. On the other side, the two genetic distances represent two exogenous variables often used as a proxy of culture. On the one hand they show the fairly recent composition of the population, on the other that of the pre-industrial age. Finally, we synthesize the climate condition in the last measure[16]. From now, the considered distance measures will be called in this way:

- *Compromise_Large*: Cultural distance with interdependencies between the 60 cultural traits of *Large dataset* (see Section 1.2.2).
- *Compromise_IW*: Cultural distance with interdependencies between the 10 cultural traits of *IW dataset* (see Section 1.2.2).
- *IW index*: Cultural distance over the first two dimensions of Inglehart-Welzel Cultural Map (remake considering only the WVS/EVS Joint 2017, see Section 1.2.1).
- *SCI*: facebook Social Connectedness Index (Bailey et al., 2018).
- *Fearon*: Combined distances of the ethnic and linguistic fractionalization (Fearon, 2003).
- *Genetic*: Genetic distance proposed by Spolaore and Wacziarg (2009).
- *Pemberton*: Genetic distance proposed by Spolaore and Wacziarg (2018).
- *Climatic*: Climatic distance proposed by us in the way of Gallup et al. (1999).

The most impressive thing highlighted by the previous Section is that not only the distances *Compromise_Large* and *Compromise_IW* are poorly correlated with the other distances, but this is verified even more for the distance due to the Inglehart-Welzel Cultural Map. Desmet et al. (2007), Desmet et al. (2011) and Spolaore and Wacziarg (2016) showed in the case of European countries and worldwide countries how a cultural distance obtained from the questions of the WVS is strongly correlated with a genetic distance measure, so much so that the latter can be regarded as a proxy for cultural distance. The powerfully discordant results that we find in our analysis even for *IW index* can be due to the way with which the cultural distance is measured. For example, authors did not take the cultural traits used for the construction of the Cultural Map, nor the average value for each country, but they have considered a large group of variables (even 430 Desmet et al. (2007)) and their frequency distribution (as we have done for

[16] We do not consider a measure of religious distance, nor a measure of geographical distance, which could be included in further analysis.

Compromise_Large and *Compromise_IW*). Furtheremore, they even aggregated the answers to the questions considering more Waves, while we consider only the latest available data (WVS/EVS Joint 2017). Even when we select more cultural traits than those used by inglehart-Welzel, our criterion is not generical, but it is closely related to the amount of missings per country and the estimation method used for cultural networks (see Section 1.2.2).

Results of the joint comparison via DISTATIS are depicted in Table 2.1 and Table 2.2. The first describes the *Rv coefficients* between the distances organized in the *cosine matrix*, while the second lists by column the weights associated to each distance in the compromise distance identified by different DISTATIS.

Table 2.1: *Cosine matrix* of the joint comparison between cultural distances and other distances

	Compr_Large	Compr_IW	IW index	SCI	Fearon	Genetic	Pemberton	Climatic
Compromise_Large	1	0.88	0.7	0.22	0.3	0.36	0.36	0.57
Compromise_IW		1	0.75	0.18	0.27	0.35	0.3	0.52
IW index			1	0.14	0.2	0.22	0.2	0.42
SCI				1	0.06	0.22	0.51	0.2
Fearon					1	0.19	0.15	0.18
Genetic						1	0.43	0.28
Pemberton							1	0.35
Climatic								1

Note: In red the *Rv coefficients* between the cultural distances, in black the *Rv coefficients* between the cultural distances and the *comparison distances*, in blue the *Rv coefficients* between the *comparison distances*. *Compr_Large* and *Compr_IW* stand for *Compromise_Large* and *Compromise_IW*. For jointly compare the 8 measures of distance, we can consider only the intersection of countries which are covered by all of them. From *Compromise_Large*, *Compromise_IW* and *IW index* we have to rule out: Andorra, China, Iran, Russia, Puerto Rico, Macao, Iceland, Hong Kong, Myanmar, Bosnia and Herzegovina, Serbia, Montenegro, Taiwan. For the other measures we have to exclude those countries we did not already avoid in the comparison of the previous Section.

The values of the *Rv coefficients* roughly confirm the correlations observed in the previous Section. The hypotheses formulated in the previous Section are widely observed through the *cosine matrix*.

The values in black represent the *Rv* between the cultural distances and the other distances, which from now we will call *comparison distances*. They are always higher when associated to the *Compromise_Large*, intermediate for the *Compromise_IW*, while the lower ones are observed for the *IW index*. The coefficients are lower and closer to each other for the *SCI*, but reading the table from left to right, as their intensity increases also the gap between *IW index* and *Compromise_Large* sharpens. Although the genetic distance measures refer to different periods, they do not present substantial differences in *Rv*, however the

Rv coefficients associated with *Compromise_Large* and *Compromise_IW* are almost identical for *Genetic*. The highest values of association are verified between cultural distance measures and climatic distance, which shows the important role of climate in creating cultural discrepancies.

As expected, the red values of *Rv coefficients* between *Compromise_Large*, *Compromise_IW* and *IW index* are slightly different from Table 1.10 because, here, we delete some countries to facilitate the joint comparison of distances. Instead, taking a look at the *Rv coefficients* among *comparison distances* (in blue), we can also come to other conclusions on the relation between these measures and those of cultural distance. The *Rv* of 0.43 between *Genetic* and *Pemberton* from one hand makes us realize that the genetic distances have changed from AD 1500 to nowadays, on the other hand this points out that the almost identical *Rv coefficients* between *Genetic*, *Pemberton* and the cultural distances may be due to "qualitative" differences. In other words, despite the aggregate result, some countries may be more similar by comparing *Genetic* with the cultural distances, but the same countries may be dissimilar when considering *Pemberton*. In the same way, other countries may be similar considering *Pemberton* against cultural distances, but dissimilar considering *Genetic*.
Another important coefficient is that between *SCI* and *Pemberton* (0.51). The relation is not excessively strong, but it is plausible that countries distant at a genetic level in AD 1500, have not even currently developed a certain contact at the level of social networks (in this case facebook). It is therefore also plausible that countries historically similar in genetic terms, have had more contacts over the years, which have been finalized with many "friendships" among their inhabitants. Anyway, the phenomena that facilitate the connection between facebook users could be various.
The ancestral component of the population of a country seems to have a minimal link also with the climatic condition (0.35). It is possible that a similar genetic composition has to do with similar climatic conditions.

As described in Section 1.6, weights for the compromise are deduced by the eigenvector associated to the first dimension of the non centered PCA over the *cosine matrix*. For construction the non centered PCA gives much importance to the first principal component, which justifies the weights for compromise. In the case of Table 2.2, when we consider all or most distances, the variability explained by the first dimension is not so high. This is certainly due to the fact that the level of

Table 2.2: Associated weights to distance measures for different DISTATIS

	1	2	3	4	5	6	7	8	9	10	11	12	13
Compromise Large	0.17	0.2			0.21	0.26	0.24	0.21	0.24	0.21	0.27	0.25	0.29
Compromise IW	0.17		0.18							0.21		0.23	0.28
IW index	0.15			0.16						0.18			
SCI	0.08	0.14	0.15	0.16	0.20	0.14	0.18	0.17				0.13	
Fearon	0.08	0.11	0.11	0.11	0.13	0.16	0.13		0.14				
Genetic	0.11	0.17	0.18	0.18	0.21	0.2		0.19	0.2	0.12	0.23		
Pemberton	0.11	0.2	0.2	0.21	0.25		0.23	0.23	0.2	0.12	0.24	0.18	0.18
Climatic	0.13	0.18	0.18	0.18		0.24	0.22	0.2	0.22	0.16	0.26	0.21	0.25
First eigenvalue (%)	45	42	41	38	43	42	44	48	46	56	55	54	64

relatedness between the distances is generally not very high, namely, they do not synthesize the same concept. This percentage increases up to 64% when we insert in the DISTATIS only the distances that have a rather high *Rv*. In fact, in DISTATIS 13 (the last column of the Table 2.2) we only consider *Compromise_Large*, *Compromise_IW*, *Pemberton* and *Climatic*. It is not surprising that the lowest percentage of variability explained by the first component (38%) coincides with the compromise 4 that contains, as cultural distance, the *IW index* together with the *comparison distances* (in fact, both Section 2.2 and Table 2.2 show how this measure is poorly linked to the others). Compromises from column 5 to column 9 consider the *Compromise_Large* as cultural distance and four of the *comparison distances* (we exclude one of them for each DISTATIS). The effectiveness of the compromise (explained percentage of the first dimension) increases when we do not include the *SCI* (column 9) and the *Fearon* (column 8), which in the compromise 1 play a minority role. Result for compromise 11 remains similar to compromise 10, in fact *Compromise_Large* is sufficient to represent the information of the *Compromise_IW* and *IW index*. In the analysis of column 12 we include the distances that are most consistent in term of *Rv* in Table 2.1, but the result does not differ much from that of 10 and 11. In general, when we consider cultural distances one by one (compromises 2-4) the weights are almost identical: obviously *Compromise_Large* weighs more than *Compromise_IW* and especially more than *IW index*.

Looking at the compromises from column 5 to 9, which all have the same number of distances, we notice that the weight of the *Pemberton* increases up to 0.25 when we delete from the analysis the *Climatic* distance, and in the same way, the highest weight for *Climatic* is verified when *Pemberton* is discarded. When the 3

cultural distances (columns 1 and 10) are considered, the weights associated with the *comparison distances* drop dramatically. The most homogeneous weights are described in compromise 11.

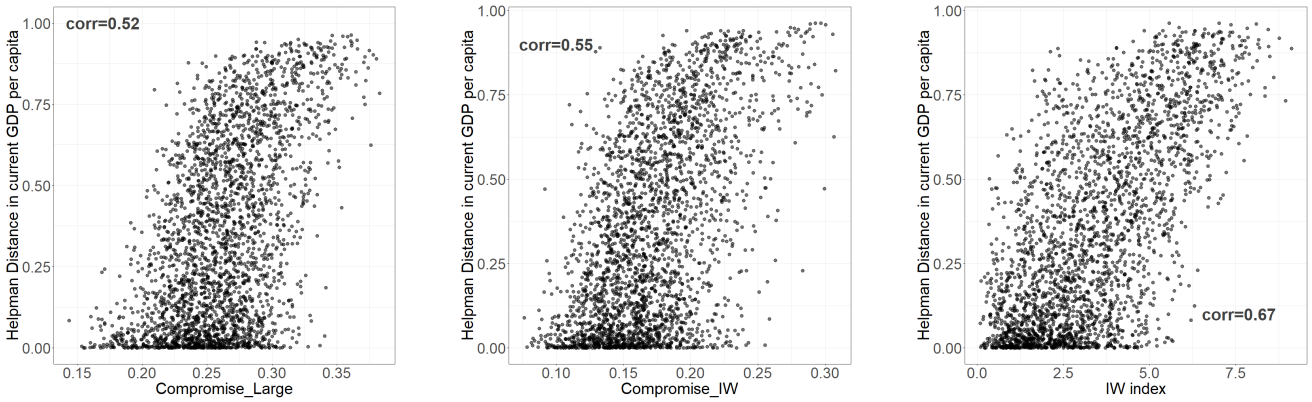
2.4 A model for the GDP per capita

In the way of Section 2.2, in this Section we firstly propose a comparison between the three cultural distances and the GDP per capita, and secondly we integrate this comparison with a model of the cultural distances over the distance in GDP per capita.

This economic measure is one of the most used in literature for describing the economic situation of a country and explore its relation with possible determinants. For example, Özak (2018) finds a U-shape relation between the current GDP per capita and the distance to the technological frontier around 1800, calculated as the distance in term of HMISea to UK. Alesina et al. (2003), following elaborations from Easterly and Levine (1997), shows the highly negative correlation between the ethnic fractionalization and the GDP per capita. Sachs (2001) describes as the economic gap of tropical countries was amplified by their climatic conditions. Desmet et al. (2011) firstly sets a model for the stability of a country or an economical union of countries (at European level) by considering its cultural heterogeneity and its economical situation by the level of GDP per capita. Measures of genetic distances are usually used in modelling trade (see for example Guiso et al. (2009); Giuliano et al. (2006, 2014)) controling for GDP per capita. At the same time this is recently done with measures obtained by social networks (facebook) that accounts for social connectedness as a proxy of cultural connectedness (Bailey et al., 2018).

In Figure 2.6 each point of the scatter plots depicts the country pairwise distance measured with the cultural index estimated in Chapter 1 and the Helpman similarity index (Helpman, 1987) in GDP per capita in current US dollars (year 2017). This measure is transformed in distance by using $d_{lm} = 1 - \frac{s_{lm}}{max(s_{lm})}$ (the inverse operation showed by Wierzchoń and Kłopotek (2018)), where d_{lm} is the final distance measure between countries l and m, s_{lm} is the similarity measure (in this case the Helpman one) between countries l and m. Clearly the hypotheses set in Section 2.2 are only partially transferable in this case. Here, we expect some differences between the three measures of cultural distance when compared

Figure 2.6: Cultural distance vs. Helpman distance in GDP per capita



(a) (b) (c)

Note: GDP per capita in current US dollars refers to the 2017 and comes from the WITS (World Integrated Trade Solution) website https://wits.worldbank.org. It is not available for *Puerto Rico* and *Taiwan*, then we exclude these countries also from our cultural distances.

with the distance in GDP, but in general higher levels of correlation in respect to *comparison distances*. In fact, compared to the low correlations observed in Section 2.2, from Figure 2.6 we see more sustained values, which are surprisingly higher for *IW index* than the other two, which basically should be more complete measures.

The correlation, as seen in Section 2.3, can be misleading when calculated between distance measures, which contain internally a certain structure of interdependencies. In this regard, considering the distances as dyadic variables, we are interested in evaluating how much each distance accounts in predicting an economic output, namely the distance in GDP per capita. Table 2.3 shows the results deriving from the Bayesian implementation (Hoff, 2009, 2021) of the SRRM[17] (Warner et al., 1979) of the distance measures on the distance measure in GDP per capita. The model accounts for statistical dependencies between distances that relate to the same node with the inclusion of random effects for each node.

Based on these calculations, there appears to be strong evidence for associations between the cultural distances and the distance in GDP per capita. However, unlike correlation, through this model we see a more evident association of distance in GDP per capita with *Compromise_Large* and *Compromise_IW*, in fact the coefficient associated with them is quite high, but pretty different one each other. Although the coefficient associated with *IW index* in SRRM1 and SRRM5

[17] Social Relations Regression Model. See the `ame` function implemented in the `amen` R package.

Table 2.3: Regression Modeling over distance in GDP per capita

	GDP per capita						
	SRRM1	SRRM2	SRRM3	SRRM4	SRRM5	SRRM6	SRRM7
Intercept	3.72*** [3.17, 4.24]	7.38*** [6.96, 7.73]	5.88*** [5.57, 6.24]	-2.75*** [-2.89, -2.63]	-0.01 [-1.54, 1.65]	4.1*** [2.58, 5.34]	-2.73** [-4.37, -1.24]
Compr_Large	2.47*** [2.03, 2.9]	6.85*** [6.61, 7.13]			2.93*** [2.5, 3.35]	7.33*** [7.05, 7.62]	
Compr_IW	1.54*** [1.23, 1.79]		4.34*** [4.17, 4.51]		1.57*** [1.3, 1.83]		
IW index	0.54*** [0.48, 0.6]			1.1*** [1.06, 1.14]	0.53*** [0.48, 0.59]		
SCI					-0.9*** [-1.24, -0.54]	-0.74*** [-1.07, -0.43]	-0.5** [-0.91, -0.18]
Fearon					0.04· [0.00, 0.07]	0.02 [-0.01, 0.06]	0.14*** [0.09, 0.18]
Genetic					-0.03 [-0.08, 0.01]	0.03 [-0.01, 0.08]	0.23*** [0.18, 0.29]
Pemberton					-0.08* [-0.15, -0.02]	-0.22*** [-0.28, -0.15]	0.05 [-0.03, 0.14]
Climatic					-0.07· [-0.13, 0.00]	-0.03 [-0.08, 0.04]	0.48*** [0.41, 0.57]
w-row variance	0.136	0.282	0.224	0.104	0.150	0.314	0.133
error variance	0.452	0.495	0.506	0.508	0.445	0.483	0.702

Note: Distances are previously transformed using the logarithm function. The regression models are fitted using a mixed effect model (Hoff, 2009) that takes into account the dependencies due to the network structure of the data. The table reports estimated regression coefficients and 89% highest density intervals in square brackets. The latter are obtained from posterior samples of regression coefficients. We discard from the model countries that we already deleted for the joint comparison of Section 2.3. *Compr_Large* and *Compr_IW* stand for *Compromise_Large* and *Compromise_IW*, while w-row variance stands for within-row variance.

(when it is included together with other distances) is highly significant, it is close to zero. On the other side, together with SRRM4 - where *IW index* is considered alone - these models have the lowest variance, which is matched by larger parameter ranges in the other models. This defines them as more precise and reliable estimations, anyway the error variance does not vary too much model by model. We can say that the result in terms of correlation is reversed when considering a random effect by row (country, or node). Cultural distance uses in the right way the information that we add firstly by including the network component and secondly by adding further cultural traits to that of the Inglehart-Welzel Cultural Map. This latter result indicates that the further information added by including more cultural traits play an important role in the effect of cultural distance on economic distance. In general, the three distances obviously measure the same concept because when they are together in the same model (SSRM1 and SSRM5) the coefficients have a lower intensity than those obtained when the distances are considered separately (SSRM2, SSRM3 and SSRM4).
In the SRRM5, SRRM6 and SRRM7 we consider also the *comparison distances*. Specifically, SRRM5 and SRRM6 highlight the fundamental effect of cultural distances, in fact the relationships between the *comparison distances* and the economic distance in SRRM7 are completely upset (here, only the *comparison*

distances are included). Basically, when we check for the node effect, there is an overlap of information between the cultural distance and all the others. Some coefficients change sign, such as *Genetic*, *Pemberton* and *Climatic*, even *Pemberton* loses significance, while *Fearon*, *Genetic* and *Climatic* become significant.
We focus on the results of SRRM6, which contain the main cultural distance and the *comparison distances*. *Fearon*, *Genetic* and *Climatic* are not significative to explain the distance in GPD per capita, while an increase of 1% in cultural distance produce an increase of almost 7.56% in the economic distance; an increase of 1% in facebook distance lead to a decrease of the 0.73% in the economic distance; and an increase of 1% in genetic distance at AD 1500 results in a decrease of the 0.22% in the economic distance. In brief, when a pair of countries is more culturally distant than another pair, also their GDP per capita is more distant, namely culturally similar countries tend to have also a similar GDP per capita.

From the conclusions of the Chapter 1 we could admit that adding variables was important to increase the informative content of the new cultural distance measure. Although this is verified even in practice, the choice of Inglehart and Baker (2000) to reduce the number of cultural traits from 22 to 10 for constructing the Cultural Map, can be useful also in our case. The reason is mostly based on the computational cost and effectiveness of the BDMCMC procedure.

2.5 Conclusions

The intent of Chapter 3 is to provide an application basis for cultural distances found in Chapter 1 on different sets of cultural traits. This has been practiced in three parts. The comparison of three cultural distances with a selection of distance measures used in economic literature as a proxy of cultural distance or in comparison to it. The joint comparison of all distance measures considered in the first part. The implementation of a simple model that verifies the association between the proposed distances and those of literature with an economic distance between countries in GDP per capita.

The most interesting results can be summarized as follow. The measures of cultural distance are poorly correlated with the distances from the literature, suggesting that they identify different concepts. In general, *Compromise_Large* and *Compromise_IW* have higher levels of association than *IW index* with distances from literature. This is also evident from the joint comparison, from which we can

deduce the importance of *Pemberton* and *Climatic* distances in the compromise of distances found via DISTATIS. Following the result of Chapter 1, *Compromise_Large* and *Compromise_IW* seem to have an excellent explanatory power on distance in GDP per capita, which is definitely higher than that of *IW index*. Adding more cultural traits helps the distance measure, but the Inglehart-Welzel selection may be already considered as a parsimonious solution (the processing time of BDMCMC to arrive at the convergence increases exponentially including more variables). Finally, considering the interdependencies between cultural traits in the definition of cultural distance offers significant explanatory benefits of the country's economic condition measured by GDP per capita.

References

Alesina, A., Devleeschauwer, A., Easterly, W., Kurlat, S., and Wacziarg, R. (2003). Fractionalization. *Journal of Economic growth*, 8(2):155–194.

Ashraf, Q. and Galor, O. (2011). Dynamics and stagnation in the malthusian epoch. *American Economic Review*, 101(5):2003–41.

Ashraf, Q. and Galor, O. (2013). Genetic diversity and the origins of cultural fragmentation. *American Economic Review*, 103(3):528–33.

Bailey, M., Cao, R., Kuchler, T., Stroebel, J., and Wong, A. (2018). Social connectedness: Measurement, determinants, and effects. *Journal of Economic Perspectives*, 32(3):259–80.

Cavalli-Sforza, L. L., Menozzi, P., and Piazza, A. (1994). *The history and geography of human genes*. Princeton university press.

Desmet, K., Le Breton, M., Ortuño-Ortín, I., and Weber, S. (2007). Stability of nations and genetic diversity. *Working paper. Universidad Carlos III*.

Desmet, K., Le Breton, M., Ortuño-Ortín, I., and Weber, S. (2011). The stability and breakup of nations: a quantitative analysis. *Journal of Economic Growth*, 16(3):183–213.

Easterly, W. and Levine, R. (1997). Africa's growth tragedy: policies and ethnic divisions. *The quarterly journal of economics*, 112(4):1203–1250.

Fearon, J. D. (2003). Ethnic and cultural diversity by country. *Journal of economic growth*, 8(2):195–222.

Gallup, J. L., Sachs, J. D., and Mellinger, A. D. (1999). Geography and economic development. *International regional science review*, 22(2):179–232.

Giuliano, P., Spilimbergo, A., and Tonon, G. (2014). Genetic distance, transportation costs, and trade. *Journal of Economic Geography*, 14(1):179–198.

Giuliano, P., Spilimbergo, A., Tonon, G., et al. (2006). *Genetic, cultural and geographical distances*, volume 2229. Centre for Economic Policy Research.

Guiso, L., Sapienza, P., and Zingales, L. (2006). Does culture affect economic outcomes? *Journal of Economic Perspectives*, 20(2):23–48.

Guiso, L., Sapienza, P., and Zingales, L. (2009). Cultural biases in economic exchange? *The quarterly journal of economics*, 124(3):1095–1131.

Helpman, E. (1987). Imperfect competition and international trade: Evidence from fourteen industrial countries. *Journal of the Japanese and international economies*, 1(1):62–81.

Hoff, P. D. (2009). Multiplicative latent factor models for description and prediction of social networks. *Computational and mathematical organization theory*, 15(4):261.

Hoff, P. D. (2021). Additive and multiplicative effects network models. *Statistical Science*, 36(1):34–50.

Huang, C.-M. and Park, D. (2013). Cultural influences on facebook photographs. *International Journal of Psychology*, 48(3):334–343.

Inglehart, R. and Baker, W. E. (2000). Modernization, cultural change, and the persistence of traditional values. *American Sociological Review*, pages 19–51.

Inglehart, R. and Welzel, C. (2005). *Modernization, cultural change, and democracy: The human development sequence*. Cambridge University Press.

Koeppen, W., Volken, E., and Brönnimann, S. (2011). The thermal zones of the earth according to the duration of hot, moderate and cold periods and to the impact of heat on the organic world (translated from: Die wärmezonen der erde, nach der dauer der heissen, gemässigten und kalten zeit und nach der wirkung der wärme auf die organische welt betrachtet, meteorol z 1884, 1, 215-226). *Meteorologische Zeitschrift*, 20(3):351–360.

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., and Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist*, 70(6):543.

Kottek, M., Grieser, J., Beck, C., Rudolf, B., and Rubel, F. (2006). World map of the koeppen-geiger climate classification updated. *Meteorologische Zeitschrift*, 15(3):259–263.

Obradovich, N., Özak, Ö., Martín, I., Ortuño-Ortín, I., Awad, E., Cebrián, M., Cuevas, R., Desmet, K., Rahwan, I., and Cuevas, Á. (2020). Expanding the measurement of culture with a sample of two billion humans. Technical report, National Bureau of Economic Research.

Özak, Ö. (2018). Distance to the pre-industrial technological frontier and economic development. *Journal of Economic Growth*, 23(2):175–221.

Pemberton, T. J., DeGiorgio, M., and Rosenberg, N. A. (2013). Population structure in a comprehensive genomic data set on human microsatellite variation. *G3: Genes, Genomes, Genetics*, 3(5):891–907.

Rubel, F., Brugger, K., Haslinger, K., and Auer, I. (2017). The climate of the european alps: Shift of very high resolution koeppen-geiger climate zones 1800–2100. *Meteorologische Zeitschrift*, 26(2):115–125.

Rubel, F. and Kottek, M. (2010). Observed and projected climate shifts 1901-2100 depicted by world maps of the koeppen-geiger climate classification. *Meteorologische Zeitschrift*, 19(2):135.

Sachs, J. D. (2001). Tropical underdevelopment. *National Bureau of Economic Research Working Paper Series*, (w8119).

Spolaore, E. and Wacziarg, R. (2009). The diffusion of development. *The Quarterly journal of economics*, 124(2):469–529.

Spolaore, E. and Wacziarg, R. (2013). How deep are the roots of economic development? *Journal of economic literature*, 51(2):325–69.

Spolaore, E. and Wacziarg, R. (2016). Ancestry, language and culture. In *The Palgrave handbook of economics and language*, pages 174–211. Springer.

Spolaore, E. and Wacziarg, R. (2018). Ancestry and development: New evidence. *Journal of Applied Econometrics*, 33(5):748–762.

Warner, R. M., Kenny, D. A., and Stoto, M. (1979). A new round robin analysis of variance for social interaction data. *Journal of Personality and Social Psychology*, 37(10):1742.

Wierzchoń, S. T. and Kłopotek, M. A. (2018). *Modern algorithms of cluster analysis*. Springer.

Wright, S. (1949). The genetical structure of populations. *Annals of eugenics*, 15(1):323–354.

Chapter 3

Mapping networks: evidences from a simulation study

Abstract

The abnormal production of data in our time has allowed the observation of large collections of networks, which refer to the same field of analysis and which can have different sizes, e.g. think of the commercial network of each product between countries. A common way to compare these networks is to map them in the space via network descriptors. This is where the problem dealt with this Chapter arises: what is the subset of descriptors that keeps the characteristics of networks as much as possible unchanged in the mapping process? Through a simulation of networks from 4 generative models (Random, Scale-free, Small-world and Stochastic block model) and the selection of a wide set of descriptors, this Chapter finds evidence of a small subset of descriptors via Subgroup Discovery. Furthermore, it evaluates the effectiveness of descriptors by applying them to the set of binary cultural networks estimated in Chapter 1 and comparing distances between networks in the space of the descriptors with popular network distances.

3.1 Introduction

With respect to the first Chapter, which empirically found a new cultural distance index between countries including the role of the cultural traits interdependencies, i.e. through the cultural networks of each country, and the second, where its effectiveness is assessed, this Chapter touches on another topic concerning network distances which may be of interest in the economic and social sciences and their linkage.

In recent years, the extent of the data available and the popularity of the methods of network analysis have inevitably made possible to observe large collections of networks within defined fields, highlighting the importance and need to study sets of networks. Clear examples emerge in several fields: sets of brain networks (Bassett et al., 2008; Eguiluz et al., 2005); sets of collaboration networks (Kro-

negger et al., 2012); set of governance networks (Bisceglia et al., 2014); sets of social networks (Faust, 2006); sets of ego networks, where egos belong to the same category (Lubbers et al., 2010); sets of economic networks (Harland et al., 2001); or sets of cultural networks like those estimated in Chapter 1.

Since networks are by nature complex objects, procedures for analyzing them when they are organized in collections must even accept a certain degree of complexity. It could be of interest to compare networks to each other, and literature has moved in this direction by providing numerous contributions[1]. One way to study sets of networks is to assess their similarity via distance measures[2], while the most used and accepted way (in literature) is to map them as points in a reference space (one, two or multi-dimensional) through network descriptors.

However, there are two main issues in mapping networks, which are closely related to each other and with the isomorphism problem[3]. The first is due to the uniqueness of descriptors, namely their discrimination power (Dehmer and Mowshowitz, 2013); as shown in Dehmer et al. (2013) each network descriptor has a certain degree of degeneracy, in fact a single index cannot distinguish between non-isomorphic graphs. The second is the statistical relation between the descriptors measure[4] (Smith et al., 2016). We suppose to overcome the first problem by using a variegated set of descriptors, able to measure the complex characteristics of networks. While for the second, we use factorial methods, which are based on the correlation between the variables[5] (in this case network descriptors).

An important effort in the use of network measures to study a set of networks has been made at the level of topological characteristics and within the chemical field. In this context sets of topological network descriptors flattened to a single dimension. This is the case of the "SuperIndices" (Bonchev et al., 1981), which

[1] For example, see Nassimbeni (1998), Pathan and Buyya (2007), Onnela et al. (2012), Barnard and Chaminade (2011), Bazzoli et al. (1999).

[2] For common network distance measures see Hamming (1950), Hammond et al. (2013), Bunke and Shearer (1998), Wilson and Zhu (2008), Fay et al. (2009); for centrality-based distances see Roy et al. (2014); while for layer similarity in multiplex networks see Bródka et al. (2018).

[3] Graph isomorphism is an equivalence relation on graphs. Given G and H be two graphs, the relation $\mathcal{F} : V(G) \longrightarrow V(H)$ is verified, such that any two vertices g_1 and g_2 of G are adjacent in G if and only if $\mathcal{F}(g_1)$ and $\mathcal{F}(g_2)$ are adjacent in H (see McKay and Piperno (2014) for further details). From our knowledge no polynomial time algorithm (see Karmarkar (1984)) exists for determining whether two arbitrary graphs are isomorphic, but Babai (2016) found a quasi-polynomial time solution. After its publication - as the same author have declared - Harald Helfgott pointed out an error in the analysis of that Graph Isomorphism test. This stimulated Babai to discover the problem (Babai, 2017). The case of sets of network is left to the elaboration of efficient heuristics.

[4] For example, Valente et al. (2008) provides a quick review of papers which examined the correlation among the main network centrality measures, and even it aims at the same purpose.

[5] We have seen in Chapter 1 as the partial correlations among variables (conditional relations between cultural traits) can play an important role in defining higher-level objects (countries). At the same time we observed that some information is overlapped with the classical correlation. Here, we decide to remain in the classical multidimensional scheme.

offer a greater discrimination power than single descriptors (Dehmer and Mowshowitz, 2013). In the context of social networks, instead, Wang and Krim (2012) empirically demonstrate how degree centrality and clustering coefficient plays a fundamental role in classifying networks into groups after having mapped them.

The main purpose pursues in this Chapter is in between. We want to select a small subset of descriptors among a set of well-known network measures available in literature, which is able to preserve the original distances between networks in the mapping operation. They have to be suited to discriminate networks generated from different models (i.e, structurally different from each other), and at the same time, to join networks that have similar structures. Furthermore, it has to map non-isomorphic networks in different points of the space. We simulate networks without taking into account their isomorphism property, then, having no preliminary information on it, we detect it in the aftermath. In short, descriptors must allow the mapping operation not to lose or confuse the individual structural characteristics of the networks and their comparison.

The proposed procedures has four main steps: *i*) simulate/generate N networks with different sizes (Van Wijk et al., 2010; Smith et al., 2016) from different models; *ii*) describe each network through a mixture of features based on its local, global and intermediate structure (Onnela et al., 2012), namely, associate each network with a numerical vector of descriptors (mapping networks); *iii*) make use of Subgroup Discovery (Atzmueller, 2015) to find the best set of descriptors among those considered, and confirm them with a PCA; *iiii*) apply descriptors on the case study of binary cultural networks inferred in Chapter 1.

Section 3.2 lists and describes the network models from which we simulate, while Section 3.3 does the same for the selected network descriptors. Section 3.4 details the procedure of selection of the best subset of descriptors via Subgroup Discovery. At the same time it shows how they improve the discrimination among different network generative models and their uniqueness (they map non-isomorphic networks in different points of the space). This subset of descriptors is finally applied in Section 3.5 to a case study involving networks inferred in Chapter 1 (the ones from the *Large dataset*).

3.2 Network simulation

Each network in the world (we know that the possible networks are limitless) or within a specific field has some properties, which we mathematically measure via network descriptors (also called measures or indicators). By definition, each non-isomorphic network has its own structure and the probability that two networks, referring to the same context, does not share their structure increases when number of nodes increase. The relational structures among real nodes obviously depends on the mechanism by which they establish relationships among themselves (the network formation process), and, as all the real events, scholars have been interested in statistically modelling them.

The parameterization of network generative processes allows the simulation of fictitious networks with different characteristics based on the way we imagine the links form between the nodes. For this, each network model implicitly defines one group of networks. Similarly, as the parameters vary, networks simulated by different network formation processes can be confused.

Here, we are interested in generate networks with different sizes from the main well-known network models in order to have a natural division among them. In addition, we vary the parameters of the network models so as to reduce the strong structural differences that they may have. Allowing some overlap between networks simulated by different models may support the discrimination power of descriptors and their goodness in identifying groups of networks according to their characteristics.

Basically, we simulate networks by considering three main distributions of size, from which we in turn extract the number of nodes for each network. Due to the obvious computational[6] problems that some descriptors may have, networks may not have thousands of nodes, then sizes are extracted according to this scheme:

- `SMALL`: networks size varies uniformly between $v = [75, 150]$ nodes.
- `MEDIUM`: networks size varies uniformly between $v = [175, 350]$ nodes.
- `LARGE`: networks size varies uniformly between $v = [375, 500]$ nodes.

[6] For the timing of the simulation see Table 3.10 in Annex 3.

3.2.1 Random networks

Random networks were theorized in Erdös and Rényi (1959, 1960); Erdös and Rényi (1964) to describe network formation processes in which links are formed independently dyad by dyad according to a random rule defined by a probability pr. Considering v nodes and a probability pr (with $0 < pr < 1$) to observe a tie between two nodes, and given the formation of links is an indipendent process dyad by dyad, Erdos-Renyi networks are described by a binomial model[7]. Without concern on the position of nodes and architecture, the probability of a network with e edges and v vertices to be formed with a random process is:

$$pr^e(1-pr)^{\frac{v(v-1)}{2}-e}$$

For the simulation of networks from this model, two parameters need to be set. The number of nodes v and one between the probability pr of forming a link among two random nodes or the desired number of edges e. Here[8], we simulate 200 `SMALL`, 200 `MEDIUM` and 200 `LARGE` networks with pr uniformly generated in the interval [0.05, 0.2].

3.2.2 Scale-free networks

Even random networks[9] are a good suggestion, the network formation process is always affected by an underlying real mechanism. The reasons why two nodes can form a link are the most disparate. In the case of individuals, they can have a relation because homophily[10]; because they share relations with one or more individuals; because there is a territorial proximity between them; because they share a third factor besides their non-homophily[11]; because they receive a material/immaterial benefit. In this last context, the spread of information inside a network depends on its structure, but, at the same time, can be the reason why one node decides to establish a link with another node in the network.

This is the case of growth and preferential attachment mechanism, where more influential nodes attract to themselves most of the links that form in the network.

[7] For the first formulation of the model see Erdös and Rényi (1959), while a punctual description of its properties is available in the main books of Network Analysis. Between them see Jackson (2010) and Newman (2018).

[8] For the simulation we make use of the `erdos.renyi.game` function of the R package `igraph`.

[9] In addition to being immediate to thought, they possess many interesting properties and are also used as a benchmark model to compare and study real networks.

[10] Jackson (2010) states that homophily *"refers to the fact that people are more prone to maintain relationships with people who are similar to themselves."* See this book also for the main references on this scope.

[11] In the context of the spread of obesity Christakis and Fowler (2007) considers one of the explanation of clustering of obese people given by *"egos and alters might share attributes or jointly experience unobserved contemporaneous events that cause their weight to vary at the same time (confounding)"*, e.g. maybe they train both in the same gym.

This mechanism was identified by Barabási and Albert (1999) from a mapping of the World Wide Web (WWW) network[12] and denoted as the Barabasi-Albert model which generates the so called Scale-free graphs. This model is characterized by a power law distribution for the distribution of the degree centrality. In other words, the probability that one node has exactly w links (*Degree*) is:

$$Pr(w) \sim w^{-\delta}$$

As for the Erdos-Renyi networks, the simulation of networks from this model requires two fundamental parameters. The number of nodes v and the δ power of the preferential attachment. We simulate 200 `SMALL`, 200 `MEDIUM` and 200 `LARGE`[13], networks with δ uniformly generated in the interval [0.2, 1.6].

3.2.3 Small-world networks

The third model we simulate from is the so called Watts-Strogatz Small-world network model (Watts and Strogatz, 1998). The formulation of this model is a mathematical analogy with the popular theory of Small-world and six degrees of separation, from which derives the name Small-world networks. A Small-world network is found between two extremes: a regular lattice and a random network. Virtually, we can imagine the process of simulation from this model starting from a regular lattice and rewiring[14] a fraction rew of its links, where if $rew = 0$ we remain in a regular lattice, if $rew = 1$ we obtain a random network (Erdos-Renyi network). Hence, the parameters of the model are the size of neighborhood w (coincides with the number of links of a node, namely the *Degree* of a node) and the rewiring probability (rew), while the two main properties are: i) the rewiring process allows to preserve the local neighborhood; ii) as a direct consequence of "Small-world" - and in the same way of random networks - the average shortest path length (diameter of the network) between two nodes increase logarithmically with the number of nodes.

The class of Small-world networks gathers different typologies of networks. As an evidence, Amaral et al. (2000) analyzes the properties of a variety of diverse real-world networks, finding three main classes which contain scale-free, broad-scale and single-scale networks. In order to include these diversities, as made for the

[12] Subsequently verified for many other real networks (Caldarelli, 2007; Dorogovtsev and Mendes, 2013).

[13] For the simulation we make use of the `sample_pa` function of the R package `igraph`, using *psumtree* as option for the algorithm of the graph generation. This algorithm handles each power and never generates multiple edges.

[14] It is a random process, in which all the edges of the current network are cut into two halves and all these half edges are reunited randomly in order to preserve the degree of each vertex.

previous models and as target of have a certain noise to overlap networks from different models, we simulate[15]:

- 200 `SMALL` networks, with w uniformly generated in the interval [4, 9] and rew in the interval [0.04, 0.08].
- 200 `MEDIUM` networks, with w uniformly generated in the interval [7, 12] and rew in the interval [0.04, 0.08].
- 200 `LARGE` networks, with w uniformly generated in the interval [10, 15] and rew in the interval [0.04, 0.08].

3.2.4 Stochastic block model networks

After having generated networks from a random process, a preferential attachment mechanism and a Small-world model, we simulate networks from stochastic blockmodel. Faust and Wasserman (1992) describe it as model that organizes nodes in partitions, called positions, which are densely connected within them and sparsely connected between them. Nodes are classified in partitions according to an equivalence relation, namely based on their role in the network. Usually, we have three definitions of equivalence: *structural equivalence* (Lorrain and White, 1971; Batagelj et al., 1992) is verified when two nodes share identical relational edges to and from all other actors in a network; the *regular equivalence* (White and Reitz, 1983; Batagelj et al., 1992) is when nodes have the same or similar connections pattern to (possibly) different neighbors; the *stochastic equivalence* (Faust and Wasserman, 1992) implies the assumption that data are generated from a stochastic process, two nodes can be considered as stochastically equivalent if the probability distribution from which the network is generated does not change when we interchange their parameters.

The partitioning of the blockmodel networks, together with the desired insertion of noise, induces a slightly more elaborate parameterization compared to the previous three network models. Specifically, three objects need: the number of nodes v; a vector of block sizes (the number of nodes for each block); the matrix of probabilities of forming an edge within and between blocks. Finally, we simulate[16] for each size (`SMALL`, `MEDIUM` and `LARGE` networks) networks with 3 and 4 blocks.

- 100 `SMALL` networks, with 3 blocks.

[15] We make use of the `sample_smallworld` function of the R package `igraph`.
[16] We make use of the `sbm.game` function of the R package `igraph`.

- 100 `SMALL` networks, with 4 blocks.
- 100 `MEDIUM` networks, with 3 blocks.
- 100 `MEDIUM` networks, with 4 blocks.
- 100 `LARGE` networks, with 3 blocks.
- 100 `LARGE` networks, with 4 blocks.

Given the number of nodes v, and v_1, v_2, v_3, v_4 respectively the number of nodes for the first, second, third and fourth block, we have:

- For three blocks: v_1 and v_2 uniformly in the interval $[\frac{v}{6}, \frac{v}{5}2]$, and $v_3 = v - v_1 - v_2$
- For four blocks: v_1, v_2 and v_3 uniformly in the interval $[\frac{v}{10}, \frac{v}{10}3]$, and $v_4 = v - v_1 - v_2 - v_3$

We consider a probability of forming a link between blocks pr_{bet} uniformly generated in the interval [0.01, 0.05], while for the probability of forming a link within the blocks pr_{wit}, we vary uniformly in these intervals:

- `SMALL` 3 blocks: [0.3, 0.45], [0.1, 0.2], [0.2, 0.3]
- `SMALL` 4 blocks: [0.1, 0.2], [0.1, 0.2], [0.25, 0.4], [0.2, 0.3]
- `MEDIUM` 3 blocks: [0.15, 0.25], [0.25, 0.5], [0.1, 0.15]
- `MEDIUM` 4 blocks: [0.25, 0.4], [0.15, 0.2], [0.3, 0.35], [0.1, 0.15]
- `LARGE` 3 blocks: [0.1, 0.2], [0.1, 0.25], [0.25, 0.35]
- `LARGE` 4 blocks: [0.2, 0.3], [0.2, 0.3], [0.1, 0.2], [0.25, 0.35]

3.3 Network descriptors

The choice of descriptors is fundamental to grasp the main characteristics of the generated networks. They need to allow us, on the one hand, to map non-isomorphic networks in not overlapping points of the space, and on the other hand, to distinguish the networks according to their generative process. In short, descriptors must contain the information needed to faithfully transform (map) the complexity of the networks structure in objects of simpler entity, therefore analyzable via the usual statistical methods.

As mentioned in the Introduction, the greatest effort in the use of network measures to map graphs, was made in the study of chemical and biological networks, using mostly topological descriptors. A topological graph index - also called molecular descriptor - is a mathematical formula that produce a number representing a chemical structure and can be applied to any graph which models some molecular structure (Bonchev, 1995). In our framework they have some limitations: as first, they are specific for the chemical field, in fact some of them lack of meaning for the analysis of networks in other topics; as second, there are many proposed measures in literature[17], often redundant each other and sometimes organized in "SuperIndices" (Dehmer and Mowshowitz, 2013); as third, they are often applied to assert graph isomorphism.

we aim to integrate these measures, with measures related to other network features and used in other areas of application. In this regard we know from Dehmer et al. (2012) that degree-based measures have a lower discrimination power than distance-based measures. Furthermore, from Wang and Krim (2012) we know that degree centrality and clustering coefficient play an important role in clustering networks. By accepting these results, but remaining in our general purpose, in agreement with Onnela et al. (2012), we choose the descriptors respecting the three levels of analysis of the networks. In decreasing resolution they are: micro-level descriptors (node level), meso-level descriptors (modules/partitions of nodes) and macro-level descriptors (the whole network).

Besides, descriptors are related to different areas of network analysis and to different network aspects. We include descriptors mainly used in the context of social networks, for their meaningfulness in the description of different structural characteristics of the networks, like centralities and partitions; in economic networks, for their capability to detect the efficiency and resilience of networks; in chemical and biological networks, for their specific measurement of entropy, i.e. information content and network complexity.

We list below the micro-level, meso-level and macro-level descriptors.

[17] Todeschini and Consonni (2008) proposed a large review of the topological graph indices present in literature. They list 1600 molecular descriptors.

3.3.1 Micro-level descriptors

For the choice of node-level descriptors we start from considerations made in Wang and Krim (2012) stating the importance of *Degree centrality* and *Local clustering coefficient* to group networks into consistent clusters. Then, we extend the coverage of the description of the micro-level analysis. We include the main social network descriptors, like *Betweenness centrality*, *Closeness centrality*, *Eigenvector centrality*, *Pagerank centrality* and *Coreness* (Wasserman and Faust, 1994); some topological and structural indices, like *Balaban index* (Balaban, 1983; Balaban and Balaban, 1991), *Eccentricity*, *Average Node Distance* and *Distance Vertex Deviation* (Skorobogatov and Dobrynin, 1988); and a measure of weaknesses, *Loss in Connectivity* (Lhomme, 2015).

For each network we extract the distribution of each descriptor along nodes and we include in the analysis the first four moments.

Table 3.1: Micro-level descriptors

Normalized Degree centrality	Normalized Betweenness centrality	Normalized Closeness centrality
Eigenvector centrality	Pagerank centrality	Clustering coefficient
Coreness	Loss in Connectivity	Balaban index
Eccentricity	Average Node Distance	Distance Vertex Deviation

3.3.2 Meso-level descriptors

As many studies (Onnela et al., 2012; Rombach et al., 2014; Zhang et al., 2015; Zhang and Thill, 2019) on real-world networks have showed, to capture the wide range of complexity contained in networks, together with micro-level descriptors and macro-level descriptors, we need to analyze the cluster composition of a network. Usually, the meso-level structure of a network is detected in different ways: community detection (Fortunato, 2010; Newman and Girvan, 2004; Porter et al., 2009), blockmodeling (Doreian et al., 2020), core-periphery structure (Borgatti and Everett, 2000; Holme, 2005; Rombach et al., 2014; Zhang et al., 2015), or motif configurations (Faust, 2006, 2008; Choi and Wu, 2009; Alon, 2007; Milo et al., 2002).

In order to observe as much complexity as possible, our main effort is to include all these aspects. We fully cover the *motifs isomorphism classes*, proposing

for each class a measure of comparison between the observed relative frequencies of each motif configuration and the theoretical probabilities under the hypothesis of a random graph. *Motif class 2* and *Motif class 3* describe configuration of triads respectively with two edges and three edges, while the remain classes describe the motif configurations of four nodes[18]. For all the possible triads and motifs we further propose two overall measures which are based on a χ^2 distance between the observed and theoretical as made for the individual configurations. For the core-periphery structure we make use of the measure proposed by Della Rossa et al. (2013) and the k-degeneracy of a graph (Alvarez-Hamelin et al., 2006), while for the community structure we propose the Frobenius norm of the modularity matrix defined by Newman and Girvan (Newman and Girvan, 2004). Finally, we include some topological indices that rely on the use of the Shannon information: *Bonchev 2* and *Bonchev 3* (Bonchev and Trinajstić, 1977), *Topological information content* (Mowshowitz, 1968), *Radial Centric Information index* (Bonchev, 1983), *Bertz Complexity index* (Bertz, 1981) and *Spanning Tree Sensitivity* (Kim and Wilhelm, 2008).

Table 3.2: Meso-level descriptors

Motif class 2	Motif class 3	Motif class 4	Motif class 6
Motif class 7	Motif class 8	Motif class 9	Motif class 10
Triadic	Motifness	Cp-centralization	Modularity
Bonchev2	Bonchev3	Topological Information Content	K-core number
Radial Centric Information index	Bertz Complexity Index	Spanning Tree Sensitivity	

3.3.3 Macro-level descriptors

A wide collection of descriptors is proposed to study the overall network. Basically, these descriptors aim at synthesizing each network with one single value considering different characteristics ranging from degree-based features to distance-based features. As for micro-level and meso-level, we select a large set of descriptors (27) trying to cover all the aspects of a network. From the Social Network Analysis, we take the concept of *Centralization* (Wasserman and Faust, 1994) and *Assortativity* (Newman, 2002), while to analyze the structure we make use of *Smallworldness* (Watts and Strogatz, 1998), *Scalefreeness* (Barabási and Albert, 1999) and *Hierachy* (Ravasz and Barabási, 2003). We consider descriptors mostly

[18] We consider the isomorphism classes as the `graph_from_isomorphism_class` function of the R package `igraph`. From it we consider only the connected ones: *Motif class 4*, *Motif class 6*, *Motif class 7*, *Motif class 8*, *Motif class 9*, *Motif class 10*.

used in economic networks, such as *Gini index* (Goswami et al., 2018) over the degree distribution and *Vulnerability* (Piccardi and Tajoli, 2018). Finally, we include a substantial list of topological indices[19], used at the level of chemical or biological networks, which analyze mostly the entropy of the networks and their distance-based features (see for examples, Balaban (1982); Nikolić et al. (2003); Randic (1975)).

Table 3.3: Macro-level descriptors

Centralization Degree-based	Centralization Betweenness-based	Centralization Closeness-based
Assortativity	Smallworldness	Scalefreeness
Hierarchy	Gini Index	Vulnerability betweenness-based
Vulnerability degree-based	Vulnerability cascading	Vulnerability random
Diameter	Balaban J index	Compactness
Modified Zagreb	Complexity Index B	Normalized Edge Complexity
Graph vertex complexity index	Medium Articulation	Efficiency Complexity
Graph Index Complexity	Randic connectivity index	Graph energy
Group Cohesion	Balaban like Information Index	Graph distance complexity

3.4 Find the best subset of descriptors

After generating the 2400 networks and then calculating on them the selected descriptors, the mapping process is virtually finished, and provided a matrix with 2400 rows (as many as there are networks, which in our vision become "individuals") and 94 columns (as many as there are descriptors). By identifying each descriptor as a dimension, mapping produce a very large space, likely allowing for information redundancies from descriptor to descriptor. Therefore, we are interested in identifying a subset of descriptors to reduce this space. They have to be sufficient to fulfil the two tasks that we had set at the beginning of this Chapter: to discriminate networks according to the different generative models from which they are simulated; to map non-isomorphic network at different points in space.

3.4.1 Subgroup Discovery

A method that can give us good performance when the dataset is quite large, which can exploit the generative model as a natural class of object (in the way of supervised classifications), and at the same time, implicitly consider all combinations of descriptors, is the Subgroup Discovery.

[19] The R functions we use to calculate them are contained in the `QuACN` R package.

In general, Subgroup discovery is *"a versatile and effective method in descriptive and exploratory data mining"* (Atzmueller 2015, see it for an extensive review and a deeply definition). It aims at identifying interesting patterns with respect to a given target property (variable) of interest and according to a specific quality measure (interestingness). The top patterns are then ranked according to the selected quality measure.

Specifically, considering a Database $D = (N, A)$, where N is the set of individuals (in our case networks) and A is the set of attributes (in our case the descriptors), for discretized attributes, a *basic pattern* $(a_j = c_i)$ is a Boolean function that is true if the value of the attribute $a_j \in A$ - with $j = 1, ..., p$ (where p is the number of attributes) - is equal to c_i for the respective individual, where c_i denotes the values (category) of each attribute[20]. The set of *basic patterns* is denoted by η.

Considering each *basic pattern* $(a_j = c_i)$ as a selector sel, the conjunction, disjunction or both of two or more selectors creates a *complex pattern* P, which is then given by a set of *basic patterns*. Generically, the *complex pattern* $P = \{sel_1, ..., sel_\Psi\}$, $sel_\psi \in \eta$, where $\psi = 1, ..., \Psi$ and with $length(P) = \Psi$, can be read as a conjunction $P(N) = sel_1 \wedge ... \wedge sel_\Psi$. For example, in Table 3.8 at the fourth row, a complex pattern is identified by the conjunction of two *basic patterns*: $Motif\ class\ 4 = (63772.1403 - 95526.704877] \wedge Degree\_st\_m2 = (497.840128 - 1958.237337]$. This is interpretable as the body of a rule, that depends on the property of interest (target variable, quality function or both). Therefore, a subgroup is the set of all individuals (networks) that are covered by the subgroup description, which formally is:

$$S_P := ext(P) := \{n \in N | P(n) = true\}$$

The set of all possible subgroup descriptions, and thus the possible search space, is then given by 2^η, that is, all combinations of the basic patterns contained in η (as said a *basic pattern* is a Boolean function). A quality function is evaluated on these descriptions, this could be simple, complex or related to the presence of a target variable. In general it is formulated as:

$$q : 2^\eta \longrightarrow R$$

which maps every *complex pattern* in the search space to a real number that reflects the interestingness of that pattern. As results, a subgroup discovery task

[20] In case of numerical attributes, intervals of attributes $a_j \in A$ are considered, in some sense discretizing them.

provides a set of k subgroups of individuals (networks in our case), each of these characterized by specific values of one or more variables (network descriptors in our case). Each subgroup has a certain degree of interestingness, results are sorted by the highest interestingness according to the selected quality function.

To perform the Subgroup Discovery in our case, a discretization is required. We make use of the *Fayyad and Irani discretization* (Fayyad and Irani, 1993) which is based on the distinction between networks from different generative models, like the task of our Subgroup Discovery. For this reason, it acts as a double-check for the identification of subgroups. It suggest us to delete 6 descriptors for which discretization returns the flattening of all values in a single category. These descriptors are: *Balaban (first moment)*, *Balaban (second moment)*, *Balaban (third moment)*, *Balaban (fourth moment)*, *Pagerank (first moment)*, *Graph distance complexity*.

Furthermore, Subgroup Discovery works with great performances when the target variable is dichotomous. For each model we create a single matrix with the same number of rows and columns (2400 and 89), where the last column corresponds with a dichotomous variable describing whether a network is simulated from that generative model or not. Consequently, we apply Subgroup Discovery for each matrix and after we merge the results denoting which descriptors are able to discriminate networks by model.

Between the different configurations of parameters we can use to set the Subgroup Discovery task[21], we decided for our elaboration a *binomial quality* function, a *top-k*[22] equal to 200 and the *Minimum Improvement (Global) postfiltering*[23].

For each network model, Subgroup Discovery returns a matrix, in which rows defines the interesting found subgroups, whereas the five columns indicates: the value of the binomial quality function; the value of the Chi-squared quality function[24]; the size of each subgroup; the probability that the networks belonging to the target network model are included in the subgroup; the description of each subgroup, namely the values of one or more descriptors (pattern) characterizing

[21] For our elaboration we make use of the function `DiscoverSubgroups` inside the R package `rsubgroup` (see http://www.rsubgroup.org/ for details).

[22] Methods of pruning are applied to the first results from the Subgroup Discovery. The set of these patterns can be chosen considering a minimal quality threshold or the top-k subgroups. Increasing the number of post-k subgroups can affect results from the post-processing procedure applied over them.

[23] It is the post-processing filter, that checks the patterns against all possible generalizations.

[24] The two quality functions agree in the interestingness of each subgroup.

that subgroup. An extract of the results is represented in Table 3.4, while the overall results are reported in Tables 3.5, 3.6, 3.7 and 3.8 (see Annex 3).

Table 3.4: Subgroup Discovery: main results

	Erdos-Renyi	Scale-free	Small-world	Stochastic block model
Clustering coefficient m1		(-inf, 0.008] pr=1, 600/600	(0.32, inf) pr=1, 600/600	
Motif class 4			(-inf, 0.64] pr=1, 600/600	(0.64, 0.96] pr=1, 568/600
Motif class 7	(0.2, 1.25] pr=1, 599/600	(-inf, 0.2] pr=1, 600/600		
Motif class 8		(-inf, 0.15] pr=1, 600/600		(1.3, inf) pr=1, 540/600
Smallworldness	(0.24, 1.2] pr=1, 597/600	(-inf, 0.24] pr=1, 600/600		(1.3, 2.2] pr=0.99, 515/600
Motif class 3	(0.21, 1.3] pr=1, 600/600	(-inf, 0.21] pr=1, 600/600		
Motif class 6	(0.98, inf] pr=1, 602/600			
Motif class 9	(0.21, 1.81] pr=1, 599/600			
Motif class 2			(-inf, 0.76] pr=1, 600/600	
Assortativity				(0.12, inf] pr=1, 556/600
Motif class 4, Degree st. m2				(0.64, 0.96], (0.005, 0.02] pr=1, 528/600
Motif class 4, Motif class 8				(0.64, 0.96], (1.3, inf) pr=1, 522/600
Motif class 4, Triadic				(0.64, 0.96], (0.0003, 0.006] pr=1, 519/600
Motif class 4, Hierarchy				(0.64, 0.96], (-1.8, -0.45] pr=1, 518/600

Note: Only patterns with the highest interestingness are depicted. Intervals for each descriptor are produced by the Fayyad-Irani discretization based on the target variable (generative model). They can not coincide with the real support of each network model on each descriptor, but discretization favors the observation of the descriptors discrimination. The fraction under the interval for each combination descriptor-model has at the numerator the size of the found subgroup and at denominator the generated networks for each network model (it is always 600). we report also the probability pr that the networks belonging to the target network model are included in the subgroup described by a simple or complex pattern. $m1$ stands for first moment, while $m2$ for second moment.

As you may see from Tables 3.5, 3.6, 3.7 and 3.8, each subgroup (row) is identified by a description, namely the *basic* or *complex patterns* that uniquely define the networks of that subgroup. For each network generative model there were found different subgroups that discriminate the networks generated from it versus the others, on the top of the Tables there are the subgroups with the high interestingness. The objective of Table 3.4 is to show the overlapping of the complex patterns characterizing each network model. In other words, it extracts the descriptions of the interesting subgroups in the way the attributes (descriptors) that define them can cover - without overlapping - as more generative models

as possible. For example, the best situation happens when one basic pattern ($a_j = c_i$) describes a subgroup with 600 networks and $pr = 1$ for each of the network generative model, with c_i different for each of them. In this context, we verify that *Smallworldness* covers an important role in the model-based discrimination, which confirms the central role of the concept of Small-world in literature. Amaral et al. (2000) shows as different networks models can be seen as classes of Small-world networks, furthermore the very way the model is theorized places it between two poles: regular lattice and random graph. As result, different intervals of values of *Smallworldness* identify with a good precision different kind of networks, except for Small-world networks which the randomness of the simulation makes them varying along a good part of the support of *Smallworldness*, in particular with values between 2 and 10, slightly overlapping with Stochastic block model. Not surprisingly we notice the importance of *Local Clustering Coefficient*, already highlighted by Wang and Krim (2012). Furthermore, 4-nodes configurations play an interesting part, specifically *Motif class 4*, *Motif class 7* and *Motif class 8*. *Motif class 4* describes a configuration in which one node is connected to the other three (see Selected network descriptors in Annex 3), which at the contrary are not connected among them. The intervals associated to Small-world and Stochastic block model include low values, denoting the lower presence of this configuration compared to a random graph. Values for Erdos-Renyi are expected to be in the range near 1, in fact they stand between 0.93 and 1.06, while for Scale-free they are between 1.05 and 4. The overlap for Erdos-Renyi and Stochastic block model, and for Erdos-Renyi and Scale-free does not allow the discretization[25] and subsequently the Subgroup Discovery to emerge this patterns, but they are a great evidence of how this configuration is verified in the preferential attachment mechanism. At the contrary, *Motif class 7* describes 3 nodes in a closed triad and one of these connected to the fourth node (see Selected network descriptors in Annex 3 for its description). It is important for Small-world (values in [1.6, 8.4]) and Stochastic block model (values in [1.3, 3.2]) networks compared to a random graph. *Motif class 8* is defined as a cycle of 4 nodes and it is relevant for Stochastic block model networks, while is not observed for Scale-free[26] networks and ranging near to 1 for Erdos-Renyi and Small-world networks. The other descriptors tend to discriminate each model versus the others three, leaving the latter in an apparent non-discriminatory overlap according to the precise search for subgroups.

[25] It is good to remember that discretization is made considering all the four models, then intervals can not coincide with the full observed support of a descriptor over one model.

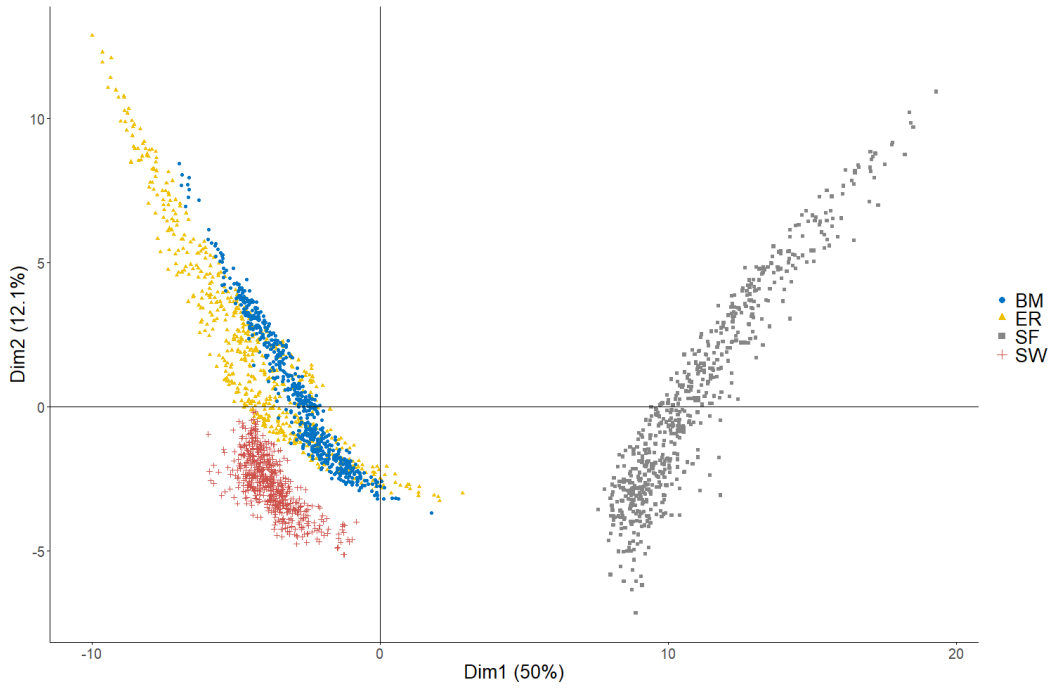
[26] The great discrimination of some descriptors for Scale-free networks is often due to the absence of one or more properties in that kind of network. For example there are no closed triads and no configuration of *Motif class 7* and *8*.

For the following elaborations we decide to use the first five descriptors listed in Table 3.4, which, in our case, are the most interesting, in fact they discriminate more than one network. We considerably reduce the space, cover the three level of analysis (micro, meso and macro) and at the same time we do not include redundancies[27].

3.4.2 PCA: mapping networks

Considering the large number of descriptors and the fact we want to observe the improvements in using only few descriptors, the simplest way to visually map networks in the space is via Principal Component Analysis (PCA). We apply it for the full dataset and the reduced one according to the results from Subgroup Discovery.

Figure 3.1: Mapping of simulated networks on the first two dimensions of the PCA using the overall set of descriptors



Note: Dataset to which is applied the PCA is composed by 2400 networks and 88 descriptors (all those used in the Subgroup Discovery). The third dimension explain 11.7% of the total variability, then it is as important as the second one. For simplicity of representation we focus on the first two dimensions. Since the number of variables is pretty high, the Correlation circle is not reported.

Three main things are evident from Figure 3.1. The first is the great percentage

[27] As mentioned in Annex 3, *Motif class 3* contains an information already included in the *Smallworldness*, which could be linked also with *Local Clustering Coefficient* if it was calculated only accounting for triads, but in our case we use the overall neighborhood of a node.

of variability explained by the Dimension 1 (50%), that suggests much information on the characteristics of networks is shared by many descriptors. The second is the clear division between Scale-free (SF) networks and the rest. While, the third is the overlapping between Erdos-Renyi (ER), Stochastic block model (BM) and Small-world (SW) networks, however the latter detaches from the other two when considering also the third dimension (11.7%).

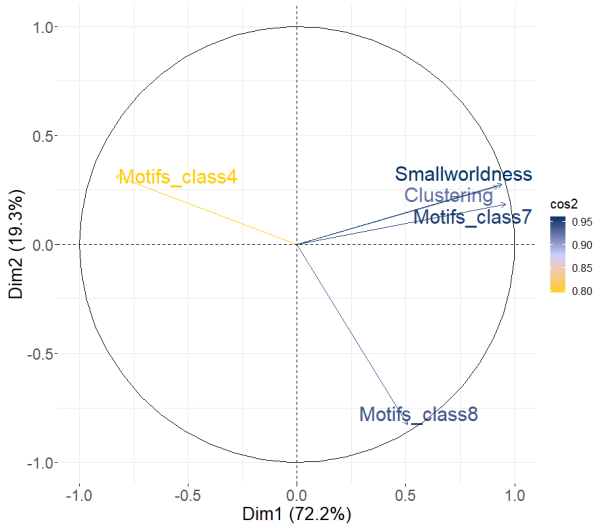
In total the first three dimensions explain almost 74% of the total variability contained in the data matrix, which is a considerable amount. This is a good result, because from one side marks the presence of latent dimensions aimed at reducing a large space and, in parallel, suggests that these dimensions could summarize the main characteristics to faithfully map and group networks. The contrast between Scale-free and the other networks on the horizontal axis is mainly dictated by descriptors of the network vulnerability and topology. The complexity of networks at meso and macro-level delineates the differences between networks along the vertical axis, while only for the third dimension we observe substantial contribution from *Smallworldness*, *Motifs* and *Clustering coefficient*[28].

Anyway, compacting the information of 88 descriptors does not guarantee an acceptable yield in distinguishing networks based on their generative process: Erdos-Renyi and Stochastic block model networks remain confused with each other even when a three dimensional space is considered.

As suggested by the Subgroup Discovery and mentioned at the end of the previous paragraph, we remake the PCA considering only a small subset of descriptors, composed by: *Local Clustering Coefficient (first moment)*, *Motif class 4*, *Motif class 7*, *Motif class 8*, *Smallworldness*. The first two dimensions explain 91.5% of the total variability, which the main part is due to Dimension 1 (72.2%). As proof of this, 4 out of 5 variables lie on the horizontal axis, while only 1 (*Motif class 8*) characterizes the vertical axis. Reading Figure 3.2 from left to right the first dimension can be interpreted as preferential attachment (*Motif class 4*) versus small-world mechanism (*Smallworldness*, *Local Clustering Coefficient (first moment)* and *Motif class 7*), namely network structure with open triads versus closed triads and cohesive networks. Looking from the bottom up, the second dimension characterizes networks organized in groups. The presence of 4-nodes cycles (*Motif class 8*) is probably due to the connection between pairs of nodes

[28] See Annex 3 for tables of descriptions contributions and squared cosines (Table 3.9).

Figure 3.2: Correlation Circle of the first two dimensions of the PCA on simulated networks using the selected subset of descriptors



Note: Dataset to which is applied the PCA is composed by 2400 networks and 5 descriptors: *Local Clustering Coefficient (first moment)*, *Motif class 4*, *Motif class 7*, *Motif class 8*, *Smallworldness*.

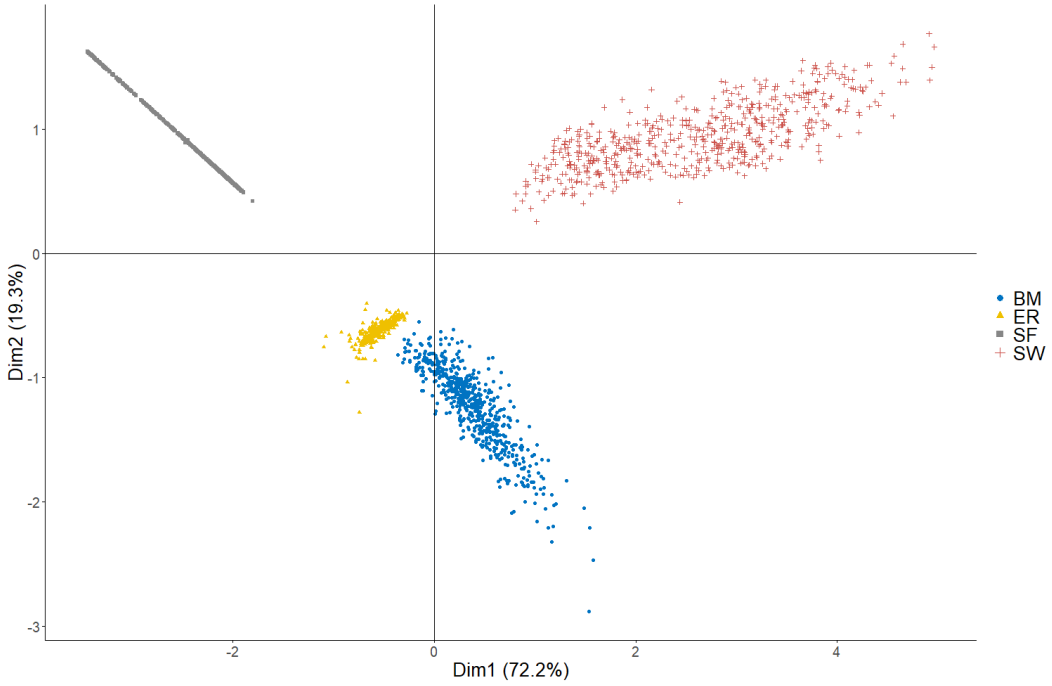
of different groups, without excluding the possibility that this configuration can also be displayed within the groups.

The layout of the networks in the space identified by the first and second dimensions of the PCA, reflects the peculiarities of generative processes. The first dimension contrasts from left to right Scale-free and Small-world networks passing through Erdos-Renyi and Stochastic block model networks that occupy a more central position. The second dimension, on the other hand, projects from the bottom upwards the Stochastic block model networks and gradually the other networks less characterized by the features described by the second principal component. As a confirmation of their benchmark function and their degree of generality, the Erdos-Renyi networks are located in the position closest to the origin of the axes, forming a compact group. But the most important result is the perfect discrimination between networks coming from different generative models through only two dimensions obtained from five descriptors. Although Erdos-Renyi and Stochastic block model networks are close in space, they are perfectly distinct and allow for no overlap[29].

By calculating the pairwise Euclidean distance between the points-networks in

[29] We did some tests including *Motif class 3*, *Assortativity*, *Degree centrality (first and second moment)*, *Hierarchy*, *Bonchev 2* and *Modified Zagreb* in the analysis (together and separately), but we did not get substantial improvements in mapping the networks in the two dimensions. Indeed, the two dimensions have turned out less explanatory power than those we have represented in Figure 3.2. Anyway, even these descriptors can be considered as good for mapping networks in further studies.

Figure 3.3: Mapping of simulated networks on the first two dimensions of the PCA using the selected subset of descriptors



Note: Dataset to which is applied the PCA is composed by 2400 networks and 5 descriptors: *Local Clustering Coefficient (first moment)*, *Motif class 4*, *Motif class 7*, *Motif class 8*, *Smallworldness*.

the spaces shown in Figure 3.1 and 3.3, we can imagine two different situations: 1) one or more different networks have a distance equal to zero, namely they lie on the same point; 2) the distance is equal to zero only in the case of networks with themselves. The first case may be dictated by two reasons: either the networks are isomorphic or the spatial reduction of the PCA erroneously projects two different networks at the same spatial point. To understand if the networks are non-isomorphic, we then calculate these hind distances over both spaces and find that the distances are zero only between the networks and themselves. We can therefore say that the networks we have generated are non-isomorphic and that the subset of selected descriptors manages to map them in different points of space, thus having a certain degree of uniqueness.

As final consideration, the objective of mapping non-isomorphic networks in not overlapping points of the space, and of distinguishing the networks according to their generative process, was reached.

3.5 Case study: binary cultural networks

It is clear that the partial correlations networks and those of probabilities are very informative of the national culture, but given the simulation study was made for binary networks, as case study, we exploit binary cultural networks inferred[30] in Chapter 1. Although the conclusions of Chapter 2 implicitly state that the informative content of the *IW dataset* is parsimonious in identifying the cultural network of a country, the lack of structure due to the size of the networks from *IW dataset*, makes us incline to use those from *Large dataset*. Furthermore, the five selected descriptors are tested for networks of different sizes, the constant size of networks estimated in the Chapter 1 should not affect their performance (indeed). Thus, we use hierarchical clustering on principal components (hcpc) in order to map the binary cultural networks in a reduced space and divide countries according to clusters of the structures of their networks.

Figure 3.4 shows how descriptors[31] are related on the first two dimensions in the context of the national cultural networks. The two dimensions explain in total 73.8% of the total variability. Here, descriptors assume different relations between them and towards the two main dimensions, in fact the set of considered networks have their own intrinsic characteristics, which may not coincide to those from the simulation study.

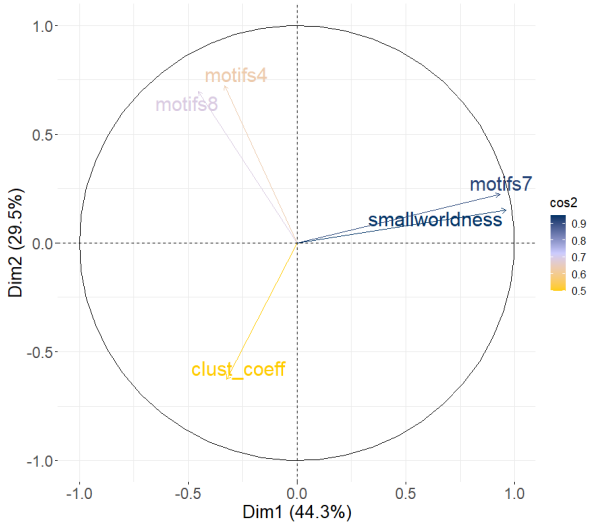
Motif class 7 and *Smallworldness* describe elements on the first axes. Cultural networks with high level of reachability between nodes stand in the right part of it. When we talk about reachability between cultural traits (nodes) usually we refer to cultures in which there are not few central cultural values, but overall many aspects could be involved directly or indirectly in the definition of their cultural attitudes. In other words, these networks are not characterized by an high value of density, neither a power law distribution of degree centrality for their nodes, but it is very likely that the path that leads from one cultural value to another is quite short.
From one side (the bottom part) the average value of the *Local Clustering Co-*

[30] It is worth to remember that is a single optimal Graph obtained by setting as a default the cut-off 0.5 on the posterior edge inclusion probabilities.

[31] We accept that the five selected descriptors may not have an universal meaning. For this reason we make some further tests adding other descriptors, resulted as lightly important from the Subgroup Discovery. Individually adding *Motif class 3*, *Assortativity*, *Normalized Degree centrality (second moment)*, *Hierarchy*, *Modified Zagreb* and *Bonchev 2* to further PCAs, we compare the distances between the national cultural networks mapped in the first two dimensions via *Rv coefficients* in the way of DISTATIS. We found values greater than 0.9, that means similar results of the mapping of cultural networks in the different PCAs, namely not a great contribution of the added descriptors. The same is verified adding simultaneously *Assortativity*, *Normalized Degree centrality (second moment)* and *Modified Zagreb*.

efficient and from the other side (the upper part) *Motif class 4* and *Motif class 8*, contributes to define the second dimension. The former describes networks characterized by clusterized nodes (connected neighborhood), while the latter, inversely, networks with less clusterized nodes and obviously few closed triads. Remaining in a general context, in terms of cultural networks, in the first case, cultural values are locally clusterized: people of a country identify cultural traits (directly or indirectly) together with small groups of other cultural traits. In the second case, there are fundamental cultural traits to which other cultural traits revolve, or there are consequentialities (cycles) between cultural traits.

Figure 3.4: Correlation Circle of the first two dimensions of the PCA on binary cultural networks using the selected subset of descriptors



Note: Dataset to which is applied the PCA is composed by 76 national cultural networks and the five selected descriptors: *Local Clustering Coefficient (first moment)*, *Motif class 4*, *Motif class 7*, *Motif class 8*, *Smallworldness*.

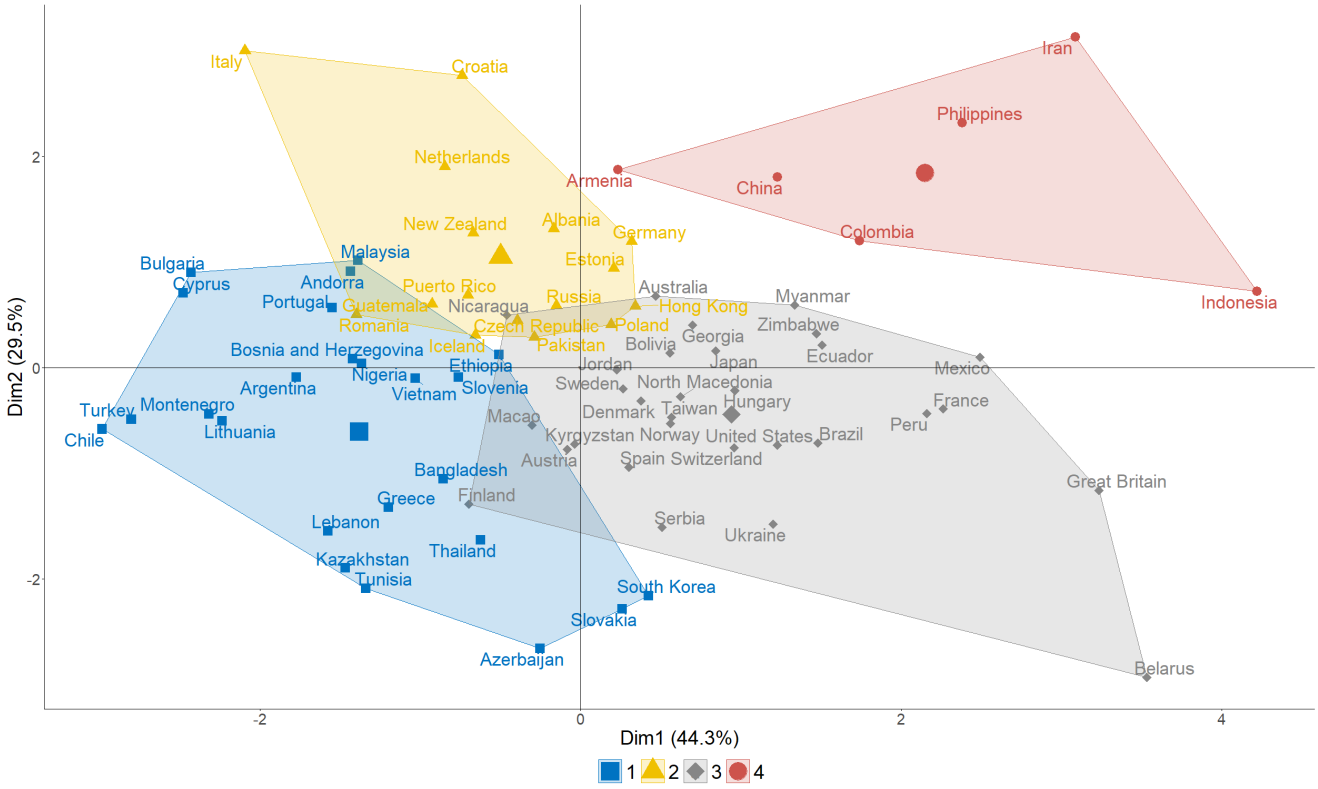
The almost orthogonality between *Smallworldness* and *Local Clustering Coefficient* gives an extreme importance to the Average Path Length[32] and the triads in defining the characteristics of the networks on the first axis. As partially mentioned above, in brief, the Average Path Length of the networks at the right of the first dimension needs to be quite low, but at the same time not due to triads and highly variable, in the way the relation between *Smallworldness* and *Local Clustering Coefficient* can result as confused. This means that cultural traits clusterize in higher order groups than triads in bottom part of the second dimension.

The "cultural map" depicted in Figure 3.5 is obviously completely different from

[32] Involved at the denominator in the calculation of the *Smallworldness*, see Annex 3.

the one proposed by Inglehart-Welzel, in fact the *Rv coefficient* between the distance matrices over their first two dimensions is close to 0. On the one hand, we do not take into account the purely distributional part, on the other, there is a substantial loss of information in using the binary cultural networks. Indeed, taking the network structures, its descriptive potential of national culture is lacking, i.e. the specific dyadic characteristics (like the sign and intensity of the relationship) remain completely hidden. In this sense it is not easy to reach interpretable conclusions or explain similarities between culture of countries, anyway as a result from the hcpc, four main clusters of the structure of national cultural networks were found.

Figure 3.5: Mapping of binary cultural networks on the first two dimensions of the PCA using the selected subset of descriptors



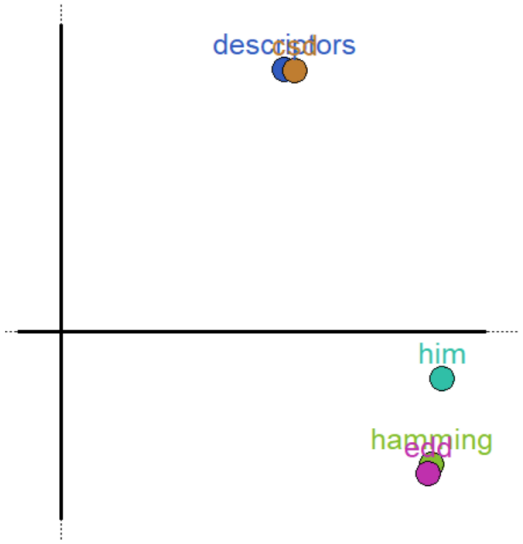
Note: Colors of polygons that identify groups are not related to those of Figure 3.1 and 3.3.

Even though some countries of the European area are concentrated in the *yellow* cluster, others scatter into the *gray* and *blue* groups. From the main important, Great Britain, France, Spain and Greece are located far away from the *yellow* group. In general only few geographical patterns are verified. Like in the case of the *blue* group, where some countries of the East Europe are located within

it as well as countries from the South-East Asia, while there is no trace of a recognizable group of Latin American countries. We can consider countries close to the origin of the axes (mainly grouped in the *gray* cluster) having a structure closer to the random one. Finally, countries in the *red* cluster have no apparent qualitative cultural relations, anyway the overall shape of their binary cultural networks seems to be similar.

There are some national cultural networks that exacerbate some characteristics identified by the Correlation circle in Figure 3.4 to the limit. Indonesia has the strongest smallworld-like structure, while for Iran this structure is contaminated by the simultaneous presence of motifs configurations that do not include closed triads. At the contrary Chile and Turkey have cultural structures far away from the small-world mechanism. Other particular cases are represented by Italy, Croatia, Great Britain and Belarus. The first two are strongly described by *Motif class 4* and *Motif class 8*, while the seconds have diametrically opposite structures in respect to the firsts.

Figure 3.6: Non centered PCA of the *cosine matrix* of the network distances



Note: First two dimensions of the non centered PCA over the *cosine matrix*, which is generated from the DISTATIS applied over the network distances listed below.

As a validation empirical result, we decide to compare via the between-distances analysis of DISTATIS[33], various distance measures calculated over the cultural binary networks. The purpose is to investigate the relationship of similarity between the distance among networks due to the characteristics detected by the

[33] See the Chapter 1 for an overall description of the method.

subset of descriptors provided by the simulation study, and other popular network distances[34]. Specifically, we consider the following network distances:

- *descriptors*: The Euclidean distance between networks based on the best subset of descriptors.
- *csd*: Distance of Continuous Spectral Densities (Ipsen, 2004).
- *edd*: Edge Difference Distance[35] (Hammond et al., 2013).
- *hamming*: Hamming distance (Hamming, 1950)
- *him*: Hamming-Ipsen-Mikhailov distance[36] (Jurman et al., 2015)

The disposition of the distance network measures in Figure 3.6 clearly shows the distinction between *edd*, *hamming* and *him* versus *descriptors* and *csd*. The action of DISTATIS tends to reward the distances that are most similar to each other, in order to find a reliable compromise, that is why the first three distances have a greater importance on the first dimension (which explains the 66% of the variability), compared to the other two, which seem to have a considerably effect on the second one (which explains the 22% of the variability) and to measure different aspects of the distances among networks. In general, it is clear how *descriptors* and *csd* - although interrelated with *Rv coefficient* not particularly high (0.53) - have a certain level of similarity. Ipsen (2004) formulates the *csd* and shows its effectiveness regarding the discrimination between networks with different sizes and coming from different generative models, in a similar way proposed in this Chapter but without simulating networks. In such a way, the parallelism with that work can serve as a validation of the goodness of the selected descriptors. The *edd*, *hamming* and *him* are very similar to each other: *him* is calculated as a combination of *hamming* and *csd*, then it is located a bit halfway between the two, tending more towards *hamming* with which it has an *Rv* equal to 0.91 (for *csd* is equal to 0.52); while *edd* seems to resemble *hamming* for binary networks ($Rv = 0.99$).

3.6 Conclusions

The focus of this Chapter differs from that of the previous two, but it maintains the thread of the study of distances and uses as a case study the binary networks

[34] All the considered distances are calculated using functions contained in the R package `NetworkDistance`.
[35] Used also for the elaborations of Chapter 1.
[36] It combines the *csd* and *hamming*.

of 60 cultural traits estimated in Chapter 1. The main objective of this Chapter is to give an empirical answer as to which subset of descriptors should be used to map in space a set of networks referring to the same field of application. The descriptors must be able to project the networks with the least loss of information on their structural content, maintaining their properties as much as possible, thus respecting their distances. In other words, they must have a good degree of discrimination and succeed in grouping together in different clusters the networks generated by different processes.

For this purpose, the Chapter generates binary networks from four models (Random, Scale-free, Small-world, Stochastic block model) and selects a wide set of descriptors that are able to summarize different aspects of micro, meso and macro analysis of networks. A subset of network descriptors is selected via Subgroup Discovery and tested via PCA against the total set initially considered. The results provide a subset of descriptors, which among those considered, respond more closely to the research question. It is composed by the first moment of the distribution of the Local Clustering Coefficient, from three configurations of Motifs and from the Smallworldness. The goodness of this subset is verified both within the mapping procedure through network simulation, and by the case study of mapping of binary cultural networks, although, in this last case, descriptors seem to measure different aspects of the distance between networks compared to popular network distance measures.

Annex 3

Subgroup Discovery: complete results

Table 3.5: Subgroup Discovery for Random networks (ER)

quality	pr	size	chi2	description
18.37	1	600	2400	Motif class 3=(21421.452429-130070.355374]
18.36	1	599	2394.67	Motif class 9=(20571.392596-180656.357546]
18.36	1	599	2394.67	Motif class 7=(20267.36282-124892.64341]
18.33	1	597	2384.03	Smallworldness=(23831.134413-123377.334918]
18.28	1	602	2378.75	Motif class 6=(97547.052794-inf)
18.07	1	586	2304.79	Motif class 6=(97547.052794-inf), Loss_conn_m1=(-inf-91764.705882]

Table 3.6: Subgroup Discovery for Scale-free networks (SF)

quality	pr	size	chi2	description
18.37	1	600	2400	Vulnerability_casbet=(73339.510957-inf)
18.37	1	600	2400	Cluster_m4=(-inf-0.434214]
18.37	1	600	2400	Norm_edge_compl=(-inf-1657.777778]
18.37	1	600	2400	Vulnerability_deg=(90011.343369-inf)
18.37	1	600	2400	Degree_st_m1=(-inf-3339.063992]
18.37	1	600	2400	Vulnerability_ran=(93538.14805-inf)
18.37	1	600	2400	Motif class 3=(-inf-21421.452429]
18.37	1	600	2400	K-core number=(-inf-200000]
18.37	1	600	2400	Vulnerability_bet=(90574.950676-inf)
18.37	1	600	2400	Complexity_B=(10474.510605-136556.391951]
18.37	1	600	2400	Eigenvector_m1=(-inf-20402.67943]
18.37	1	600	2400	Eigenvector_m2=(-inf-9132.412418]
18.37	1	600	2400	Eigenvector_m3=(-inf-5140.46247]
18.37	1	600	2400	Coreness_m3=(-inf-1023456.790123]
18.37	1	600	2400	Centr_Bet-based=(33014.082402-inf)
18.37	1	600	2400	Coreness_m2=(-inf-392592.592593]
18.37	1	600	2400	Motif class 8=(-inf-14770.494872]
18.37	1	600	2400	Motif class 7=(-inf-20267.36282]
18.37	1	600	2400	Coreness_m4=(-inf-2874074.074074]
18.37	1	600	2400	Loss_conn_m3=(81782140797037.2-inf)
18.37	1	600	2400	Loss_conn_m4=(270026961973733536-inf)
18.37	1	600	2400	Coreness_m1=(-inf-175308.641975]
18.37	1	600	2400	Cp=(54567.38256-inf)
18.37	1	600	2400	Loss_conn_m1=(9713119.14324-inf)
18.37	1	600	2400	Betweenness_st_m4=(47.484422-inf)
18.37	1	600	2400	Loss_conn_m2=(26442526088.369072-inf)
18.37	1	600	2400	Cluster_m2=(-inf-128.564817]
18.37	1	600	2400	Betweenness_st_m2=(193.672568-inf)
18.37	1	600	2400	Cluster_m1=(-inf-843.621399]
18.37	1	600	2400	Betweenness_st_m3=(86.680317-inf)
18.37	1	600	2400	Cluster_m3=(-inf-7.263733]
18.37	1	600	2400	Smallworldness=(-inf-23831.134413]

Where, values of the intervals from the discretizion of network descriptors are intended to be multiplied by 100000, and:

- *quality*: is the value of the binary quality function.
- *pr*: is the probability that the networks belonging to the target network model are included in the subgroup described by a simple or complex pattern.

- *size*: is the size of each subgroup
- *chi2*: is a further quality function
- *description*: is the subgroup description S_P.

Table 3.7: Subgroup Discovery for Small-world networks (SW)

quality	pr	size	chi2	description
18.37	1	600	2400	Cluster_m4=(3452.206547-inf)
18.37	1	600	2400	Motif class 4=(-inf-63772.1403]
18.37	1	600	2400	Motif class 2=(-inf-76497.016963]
18.37	1	600	2400	Cluster_m3=(5865.589042-inf)
18.37	1	600	2400	Cluster_m1=(32489.653057-inf)
18.37	1	600	2400	Cluster_m2=(12796.608561-inf)

Table 3.8: Subgroup Discovery for Stochastic block model networks (BM)

quality	pr	size	chi2	description
17.83	1	568	2221.85	Motif class 4=(63772.1403-95526.704877]
17.68	1	556	2170.93	Assortativity=(11636.693872-inf)
17.39	1	540	2080.01	Motif class 8=(126830.440771-inf)
17.23	1	528	2030.77	Motif class 4=(63772.1403-95526.704877], Degree_st_m2=(497.840128-1958.237337]
17.14	1	522	2001.28	Motif class 8=(126830.440771-inf), Motif class 4=(63772.1403-95526.704877]
17.09	1	519	1986.6	Motif class 4=(63772.1403-95526.704877], Triadic=(27.247355-641.896573]
17.07	1	518	1981.72	Motif class 4=(63772.1403-95526.704877], Hierarchy=(-178561.686463–44584.148882]
16.89	0.99	515	1936.67	Smallworldness=(133427.040746-221234.900902]
16.79	1	501	1899.53	Triadic=(27.247355-641.896573], Degree_st_m2=(497.840128-1958.237337]
16.75	1	499	1889.95	Motif class 4=(63772.1403-95526.704877], Zagreb_M=(469991.013257-733119.306791]
16.75	1	499	1889.95	Motif class 8=(126830.440771-inf), Degree_st_m2=(497.840128-1958.237337]
16.69	1	495	1870.87	Motif class 8=(126830.440771-inf), Triadic=(27.247355-641.896573]
16.67	1	494	1866.11	Motif class 4=(63772.1403-95526.704877], Loss_conn_m1=(-inf-91764.705882]
16.64	1	492	1856.6	Motif class 4=(63772.1403-95526.704877], Loss_conn_m3=(-inf-2191558441.5]
16.6	1	490	1847.12	Motif class 8=(126830.440771-inf), Hierarchy=(-178561.686463–44584.148882]
16.6	1	490	1847.12	Triadic=(27.247355-641.896573], Hierarchy=(-178561.686463–44584.148882]
16.53	1	486	1828.21	Motif class 4=(63772.1403-95526.704877], Cluster_m1=(12085.802877-20607.100011]
16.47	1	482	1809.38	Motif class 4=(63772.1403-95526.704877], Degree_st_m3=(50.353382-284.517633]
16.47	1	482	1809.38	Motif class 4=(63772.1403-95526.704877], Loss_conn_m4=(-inf-328733766250]
16.47	1	482	1809.38	Motif class 4=(63772.1403-95526.704877], Loss_conn_m2=(-inf-14610389.61]
16.47	1	482	1809.38	Smallworldness=(133427.040746-221234.900902], Motif class 4=(63772.1403-95526.704877]
16.45	1	481	1804.69	Smallworldness=(133427.040746-221234.900902], Bonchev2=(10340960504.7-403723694723.5]
16.41	1	479	1795.31	Motif class 4=(63772.1403-95526.704877], Degree_st_m1=(7146.531552-11942.310936]
16.38	1	477	1785.96	Triadic=(27.247355-641.896573], Cluster_m1=(12085.802877-20607.100011]
16.35	1	475	1776.62	Motif class 8=(126830.440771-inf), Zagreb_M=(469991.013257-733119.306791]
16.29	1	472	1762.66	Motif class 4=(63772.1403-95526.704877], Norm_edge_compl=(3617.770441-5938.66744]
16.29	1	472	1762.66	Smallworldness=(133427.040746-221234.900902], Degree_st_m2=(497.840128-1958.237337]
16.26	1	470	1753.37	Motif class 4=(63772.1403-95526.704877], Degree_st_m4=(5.300544-52.970125]
16.26	1	470	1753.37	Smallworldness=(133427.040746-221234.900902], Hierarchy=(-178561.686463–44584.148882]
16.24	1	469	1748.73	Motif class 4=(63772.1403-95526.704877], Motifness=(5983227.990388-23373670.803143]
16.19	1	466	1734.85	Triadic=(27.247355-641.896573], Zagreb_M=(469991.013257-733119.306791]
16.19	1	466	1734.85	Triadic=(27.247355-641.896573], Degree_st_m3=(50.353382-284.517633]

Selected network descriptors

Follow a description of the five selected descriptors.

- `Local Clustering Coefficient` (first moment): It defines the way in which each node is clusterized inside the network via the level of connectedness of its neighborhood. For binary undirected graphs:

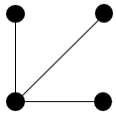
$$lc(i) = \frac{w_i^e}{w_i(w_i - 1)/2}$$

where, i is a generic node (vertex), w_i is the neighborhood of node i (it coincides with its *Degree*), w_i^e is the number of links between the w_i neighbors of v_i.

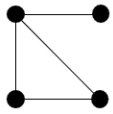
$$m_1(lc) = \frac{\sum_{i=1}^{v} lc(i)}{v}$$

where, m_1 denotes the first moment of the Local Clustering Coefficient distribution (lc) and v is total number of nodes.

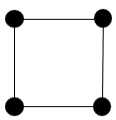
- `Motif class 4`: it is described by all 4-nodes configurations isomorphic to the follow:



- `Motif class 7`: it is described by all 4-nodes configurations isomorphic to the follow:



- `Motif class 8`: it is described by all 4-nodes configurations isomorphic to the follow:



- `Smallworldness`: it is calculated as the ratio between the *Global Clustering Coefficient* and the *Average Path Length*. The *Global Clustering Coefficient* may be obtained from the first moment of the *Local Clustering Coefficient* when the latter is calculated considering only the triads of neighbors (it is not our case).

$$SW(G) = \frac{C^g / C^g_R}{AvP / AvP_R}$$

where, C^g is the *Global Clustering Coefficient* of network denoted by graph G, C^g_R is *Global Clustering Coefficient* for a Random graph with same size (v) and *Degree* distribution; AvP is the *Average Path Length* of network denoted by graph G, AvP_R is *Average Path Length* for a Random graph with same size (v) and *Degree* distribution.

$$C^g(G) = \frac{number\ of\ closed\ triads}{number\ of\ total\ triads}$$

where, the number of closed triads is described also by *Motif class 3*.

$$AvP(G) = \frac{1}{v(v-1)/2} \sum_{i \neq j} geod(i,j)$$

where, i and j are two generic nodes and $geod(i,j)$ is the geodesic distance between them.

Overall set of descriptors: PCA results

Table 3.9: Contributes and squared cosine of the PCA on the overall set of descriptors

	Contribution			*Squared cosine*		
	Dim 1	Dim 2	Dim 3	Dim 1	Dim 2	Dim 3
Degree_st_m1	1,529282	0,370941	0,809146	0,672287	0,039451	0,08344
Degree_st_m2	0,911449	0,89348	1,366947	0,400682	0,095026	0,140962
Degree_st_m3	0,587722	1,04732	1,599937	0,258368	0,111387	0,164988
Degree_st_m4	0,370361	0,906372	1,932129	0,162814	0,096397	0,199244
Betweenness_st_m1	1,078588	1,190769	0,40048	0,474158	0,126644	0,041298
Betweenness_st_m2	1,181109	0,383205	0,473167	0,519227	0,040756	0,048794
Betweenness_st_m3	1,170448	0,433665	0,705902	0,51454	0,046122	0,072794
Betweenness_st_m4	1,119381	0,510153	0,979275	0,49209	0,054257	0,100984
Closeness_st_m1	1,978098	0,224935	0,637665	0,869591	0,023923	0,065757
Closeness_st_m2	1,791473	0,741927	0,857345	0,787549	0,078907	0,088411
Closeness_st_m3	1,562933	1,357756	1,091058	0,68708	0,144403	0,112512
Closeness_st_m4	1,347712	1,91575	1,285377	0,592467	0,203749	0,13255
Eigenvector_m1	1,783705	0,053678	0,676076	0,784134	0,005709	0,069718
Eigenvector_m2	1,5839	0,027511	1,051851	0,696297	0,002926	0,108469
Eigenvector_m3	1,451016	0,021853	1,251402	0,63788	0,002324	0,129047

Table 3.9: Contributes and squared cosine of the PCA on the overall set of descriptors

	Contribution			*Squared cosine*		
	Dim 1	Dim 2	Dim 3	Dim 1	Dim 2	Dim 3
Eigenvector_m4	1,346087	0,022738	1,356943	0,591753	0,002418	0,13993
Pagerank_m2	0,170295	2,327655	3,170039	0,074863	0,247557	0,326899
Pagerank_m3	0,114677	0,345148	1,736763	0,050413	0,036708	0,179098
Pagerank_m4	0,066884	0,153128	1,285214	0,029403	0,016286	0,132533
Cluster_m1	0,836965	1,322332	3,069872	0,367938	0,140636	0,31657
Cluster_m2	0,402255	1,897516	3,910493	0,176835	0,201809	0,403256
Cluster_m3	0,294979	2,039881	4,062198	0,129675	0,216951	0,4189
Cluster_m4	0,255948	2,064748	4,022957	0,112517	0,219595	0,414854
Eccentricity_m1	2,094726	0,090813	0,268062	0,920861	0,009658	0,027643
Eccentricity_m2	1,90271	0,442837	0,414213	0,83645	0,047098	0,042714
Eccentricity_m3	1,656468	0,788456	0,611956	0,728199	0,083856	0,063106
Eccentricity_m4	1,409422	1,024419	0,78842	0,619595	0,108952	0,081303
And_m1	2,000439	0,124296	0,474585	0,879412	0,013219	0,04894
And_m2	1,843377	0,442455	0,554992	0,810366	0,047057	0,057232
And_m3	1,637053	0,772229	0,696787	0,719664	0,08213	0,071854
And_m4	1,424246	1,020962	0,84372	0,626112	0,108584	0,087006
Dvd_m1	1,681693	1,237	0,628161	0,739288	0,13156	0,064777
Dvd_m2	1,24285	1,833702	1,040473	0,546368	0,195023	0,107295
Dvd_m3	0,966457	1,948682	1,183067	0,424864	0,207251	0,122
Dvd_m4	0,76793	1,843555	1,141726	0,337589	0,19607	0,117737
Coreness_m1	1,154855	3,314172	0,0012	0,507685	0,352477	0,000124
Coreness_m2	0,509731	4,373928	0,129881	0,224083	0,465187	0,013394
Coreness_m3	0,295431	4,025198	0,214992	0,129874	0,428098	0,02217
Coreness_m4	0,208555	3,491306	0,22053	0,091683	0,371316	0,022741
Loss_conn_m1	1,652353	1,302094	0,684622	0,72639	0,138484	0,070599
Loss_conn_m2	1,144636	1,819053	0,837107	0,503193	0,193464	0,086324
Loss_conn_m3	0,881251	1,630688	0,556882	0,387406	0,173431	0,057427
Loss_conn_m4	0,660669	1,244406	0,272083	0,290436	0,132348	0,028058
Modularity	0,955124	3,093347	0,12469	0,419882	0,328991	0,012858
Cp	2,026057	0,034334	0,64896	0,890674	0,003652	0,066922
Bonchev2	0,538105	3,478108	2,633052	0,236556	0,369912	0,271524
Bonchev3	2,089488	0,108235	0,124335	0,918559	0,011511	0,012822
Topological	0,498316	2,765703	3,293597	0,219065	0,294145	0,339641
Radial	1,93658	0,001733	0,044477	0,851339	0,000184	0,004587
Bertz	0,026592	4,432542	2,359456	0,01169	0,471421	0,243311
K-core number	1,098234	3,369793	0,01701	0,482794	0,358393	0,001754
STSD	0,061486	1,831795	0,316621	0,02703	0,19482	0,03265
Triadic	0,000673	0,530732	0,148656	0,000296	0,056446	0,01533
Motifness	0,751492	0,628672	0,03992	0,330363	0,066862	0,004117
Motif class 2	0,154362	0,004194	1,118432	0,067859	0,000446	0,115334
Motif class 3	0,558096	1,342997	4,713917	0,245344	0,142834	0,486106
Motif class 4	1,250009	0,109872	1,80523	0,549516	0,011685	0,186158
Motif class 6	0,699514	0,466322	0,000367	0,307513	0,049595	3,79E-05
Motif class 7	0,706108	1,225556	4,281786	0,310412	0,130343	0,441544
Motif class 8	1,284552	0,112657	0,092638	0,564701	0,011982	0,009553
Motif class 9	0,22613	1,168439	5,184881	0,099409	0,124269	0,534673
Motif class 10	0,122529	0,818747	4,655377	0,053865	0,087077	0,48007
Assortativity	0,515652	0,366106	2,38E-05	0,226686	0,038937	2,46E-06
Smallworldness	0,584292	1,365581	4,680376	0,25686	0,145236	0,482648
Scalefreeness	0,55634	0,075799	0,01168	0,244572	0,008062	0,001204
Hierarchy	2,13277	0,026069	0,083763	0,937586	0,002773	0,008638
Randic	0,123191	4,076322	2,584449	0,054156	0,433535	0,266512
Balaban_J	1,095812	0,789957	0,174343	0,481729	0,084016	0,017979
Compactness	2,000439	0,124296	0,474585	0,879412	0,013219	0,04894
Zagreb_M	1,750359	0,96045	0,703856	0,769474	0,102148	0,072583
Complexity_B	1,03361	3,972196	0,176087	0,454385	0,422461	0,018158
Norm_edge_compl	1,534751	0,384498	0,802672	0,674691	0,040893	0,082773
Graph_vertex_compl	2,004728	0,154955	0,218985	0,881297	0,01648	0,022582
Medium_Articulation	1,891424	0,29124	0,002902	0,831488	0,030975	0,000299
Eff_compl	1,993302	0,068934	0,104992	0,876274	0,007331	0,010827
Graph_index_compl	1,536293	0,504169	1,238556	0,675369	0,053621	0,127722
Diameter	2,1135	0,102422	0,187547	0,929115	0,010893	0,01934
Graph_energy	0,732475	4,215686	0,09929	0,322003	0,448357	0,010239
Centr_Degree-based	0,175163	0,082013	4,055977	0,077003	0,008722	0,418259
Centr_Bet-based	2,036427	0,001032	0,155333	0,895233	0,00011	0,016018

Table 3.9: Contributes and squared cosine of the PCA on the overall set of descriptors

	Contribution			*Squared cosine*		
	Dim 1	Dim 2	Dim 3	Dim 1	Dim 2	Dim 3
Centr_Clos-based	1,210933	0,740606	1,332853	0,532338	0,078767	0,137446
Group_Cohesion	0,930399	2,282346	0,103759	0,409012	0,242738	0,0107
Gini_index	1,667828	0,050072	1,210676	0,733193	0,005325	0,124847
Vulnerability_bet	2,141183	0,000221	0,10222	0,941284	2,35E-05	0,010541
Vulnerability_deg	2,129464	0,002163	0,086727	0,936132	0,00023	0,008943
Vulnerability_ran	2,091579	0,184768	0,001276	0,919478	0,019651	0,000132
Vulnerability_casbet	2,126419	0,048756	0,001045	0,934794	0,005185	0,000108
Balaban-like	0,838102	0,582856	0,50693	0,368438	0,061989	0,052275

Simulation mapping timing

The processing time for a machine that has a Xeon E5-1650v2 processor and 16 gigabytes of RAM was:

Table 3.10: Simulation mapping timing

Model	**Parameters**	**timing (hours)**
ER	n=200, v=[75-150]	1.75
	n=200, v=[175-350]	18.5
	n=200, v=[375-500]	94.5
SF	n=200, v=[75-150]	0.4
	n=200, v=[175-350]	2
	n=200, v=[375-500]	8.5
SW	n=200, v=[75-150]	0.5
	n=200, v=[175-350]	6
	n=200, v=[375-500]	55
BM	n=100, groups=3, v=[75,150]	0.6
	n=100, groups=4, v=[75,150]	0.8
	n=100, groups=3, v=[175,350]	21.3
	n=100, groups=4, v=[175,350]	15.3
	n=100, groups=3, v=[375,500]	81.5
	n=100, groups=4, v=[375,500]	39.3

References

Alon, U. (2007). Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6):450–461.

Alvarez-Hamelin, J. I., Dall'Asta, L., Barrat, A., and Vespignani, A. (2006). Large scale networks fingerprinting and visualization using the k-core decomposition. In *Advances in neural information processing systems*, pages 41–50.

Amaral, L. A. N., Scala, A., Barthelemy, M., and Stanley, H. E. (2000). Classes of small-world networks. *Proceedings of the national academy of sciences*, 97(21):11149–11152.

Atzmueller, M. (2015). Subgroup discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(1):35–49.

Babai, L. (2016). Graph isomorphism in quasipolynomial time. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 684–697.

Babai, L. (2017). Fixing the upcc case of split-or-johnson. manuscript on Babai's homepage http://people.cs.uchicago.edu/~laci/.

Balaban, A. T. (1982). Highly discriminating distance-based topological index. *Chemical physics letters*, 89(5):399–404.

Balaban, A. T. (1983). Topological indices based on topological distances in molecular graphs. *Pure and Applied Chemistry*, 55(2):199–206.

Balaban, A. T. and Balaban, T.-S. (1991). New vertex invariants and topological indices of chemical graphs based on information on distances. *Journal of Mathematical Chemistry*, 8(1):383–397.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *science*, 286(5439):509–512.

Barnard, H. and Chaminade, C. (2011). Global innovation networks: towards a taxonomy. *CIRCLE Work. Pap*, pages 1–45.

Bassett, D. S., Bullmore, E., Verchinski, B. A., Mattay, V. S., Weinberger, D. R., and Meyer-Lindenberg, A. (2008). Hierarchical organization of human cortical networks in health and schizophrenia. *Journal of Neuroscience*, 28(37):9239–9248.

Batagelj, V., Ferligoj, A., and Doreian, P. (1992). Direct and indirect methods for structural equivalence. *Social networks*, 14(1-2):63–90.

Bazzoli, G. J., Shortell, S. M., Dubbs, N., Chan, C., and Kralovec, P. (1999). A taxonomy of health networks and systems: bringing order out of chaos. *Health services research*, 33(6):1683.

Bertz, S. H. (1981). The first general index of molecular complexity. *Journal of the American Chemical Society*, 103(12):3599–3601.

Bisceglia, A., Lumino, R., and Ragozini, G. (2014). Il nuovo corso delle politiche giovanili in campania (the new course of youth policies in campania region). *Milan: Franco Angeli*.

Bonchev, D. (1983). *Information theoretic indices for characterization of chemical structures*. Number 5. Research Studies Press.

Bonchev, D. (1995). Topological order in molecules 1. molecular branching revisited. *Journal of Molecular Structure: THEOCHEM*, 336(2-3):137–156.

Bonchev, D., Mekenyan, O., and Trinajstić, N. (1981). Isomer discrimination by topological information approach. *Journal of Computational Chemistry*, 2(2):127–148.

Bonchev, D. and Trinajstić, N. (1977). Information theory, distance matrix, and molecular branching. *The Journal of Chemical Physics*, 67(10):4517–4533.

Borgatti, S. P. and Everett, M. G. (2000). Models of core/periphery structures. *Social networks*, 21(4):375–395.

Bródka, P., Chmiel, A., Magnani, M., and Ragozini, G. (2018). Quantifying layer similarity in multiplex networks: a systematic study. *Royal Society open science*, 5(8):171747.

Bunke, H. and Shearer, K. (1998). A graph distance metric based on the maximal common subgraph. *Pattern recognition letters*, 19(3-4):255–259.

Caldarelli, G. (2007). *Scale-free networks: complex webs in nature and technology*. Oxford University Press.

Choi, T. Y. and Wu, Z. (2009). Triads in supply networks: theorizing buyer–supplier–supplier relationships. *Journal of Supply Chain Management*, 45(1):8–25.

Christakis, N. A. and Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. *New England journal of medicine*, 357(4):370–379.

Dehmer, M., Grabner, M., and Furtula, B. (2012). Structural discrimination of networks by using distance, degree and eigenvalue-based measures. *PLoS One*, 7(7):e38564.

Dehmer, M., Grabner, M., Mowshowitz, A., and Emmert-Streib, F. (2013). An efficient heuristic approach to detecting graph isomorphism based on combinations of highly discriminating invariants. *Advances in Computational Mathematics*, 39(2):311–325.

Dehmer, M. and Mowshowitz, A. (2013). The discrimination power of structural superindices. *Plos one*, 8(7):e70551.

Della Rossa, F., Dercole, F., and Piccardi, C. (2013). Profiling core-periphery network structure by random walkers. *Scientific reports*, 3(1):1–8.

Doreian, P., Batagelj, V., and Ferligoj, A. (2020). *Advances in Network Clustering and Blockmodeling*. John Wiley & Sons.

Dorogovtsev, S. N. and Mendes, J. F. (2013). *Evolution of networks: From biological nets to the Internet and WWW*. OUP Oxford.

Eguiluz, V. M., Chialvo, D. R., Cecchi, G. A., Baliki, M., and Apkarian, A. V. (2005). Scale-free brain functional networks. *Physical review letters*, 94(1):018102.

Erdös, P. and Rényi, A. (1959). On random graphs. *Publicationes Mathematicae Debrecen*, 6:290–297.

Erdös, P. and Rényi, A. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60.

Erdös, P. and Rényi, A. (1964). On the strength of connectedness of a random graph. *Acta Mathematica Academiae Scientiarum Hungaricae*, 12:261–267.

Faust, K. (2006). Comparing social networks: size, density, and local structure. *Metodoloski zvezki*, 3(2):185.

Faust, K. (2008). Triadic configurations in limited choice sociometric networks: Empirical and theoretical results. *Social Networks*, 30(4):273–282.

Faust, K. and Wasserman, S. (1992). Blockmodels: Interpretation and evaluation. *Social networks*, 14(1-2):5–61.

Fay, D., Haddadi, H., Thomason, A., Moore, A. W., Mortier, R., Jamakovic, A., Uhlig, S., and Rio, M. (2009). Weighted spectral distribution for internet topology analysis: theory and applications. *IEEE/ACM Transactions on networking*, 18(1):164–176.

Fayyad, U. and Irani, K. (1993). Multi-interval discretization of continuous-valued attributes for classification learning. In *IJCAI*, pages 1022–1029.

Fortunato, S. (2010). Community detection in graphs. *Physics reports*, 486(3-5):75–174.

Goswami, S., Murthy, C., and Das, A. K. (2018). Sparsity measure of a network graph: Gini index. *Information Sciences*, 462:16–39.

Hamming, R. W. (1950). Error detecting and error correcting codes. *The Bell system technical journal*, 29(2):147–160.

Hammond, D. K., Gur, Y., and Johnson, C. R. (2013). Graph diffusion distance: A difference measure for weighted graphs based on the graph Laplacian exponential kernel. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 419–422. IEEE.

Harland, C. M., Lamming, R. C., Zheng, J., and Johnsen, T. E. (2001). A taxonomy of supply networks. *Journal of Supply Chain Management*, 37(3):21–27.

Holme, P. (2005). Core-periphery organization of complex networks. *Physical Review E*, 72(4):046111.

Ipsen, M. (2004). Evolutionary reconstruction of networks. In *Function and Regulation of Cellular Systems*, pages 241–249. Springer.

Jackson, M. O. (2010). *Social and economic networks*. Princeton university press.

Jurman, G., Visintainer, R., Filosi, M., Riccadonna, S., and Furlanello, C. (2015). The him glocal metric and kernel for network comparison and classification. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10. IEEE.

Karmarkar, N. (1984). A new polynomial-time algorithm for linear programming. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pages 302–311.

Kim, J. and Wilhelm, T. (2008). What is a complex graph? *Physica A: Statistical Mechanics and its Applications*, 387(11):2637–2652.

Kronegger, L., Mali, F., Ferligoj, A., and Doreian, P. (2012). Collaboration structures in slovenian scientific communities. *Scientometrics*, 90(2):631–647.

Lhomme, S. (2015). Analyse spatiale de la structure des réseaux techniques dans un contexte de risques. *Cybergeo: European Journal of Geography*.

Lorrain, F. and White, H. C. (1971). Structural equivalence of individuals in social networks. *The Journal of mathematical sociology*, 1(1):49–80.

Lubbers, M. J., Molina, J. L., Lerner, J., Brandes, U., Ávila, J., and McCarty, C. (2010). Longitudinal analysis of personal networks. the case of argentinean migrants in spain. *Social Networks*, 32(1):91–104.

McKay, B. D. and Piperno, A. (2014). Practical graph isomorphism, ii. *Journal of Symbolic Computation*, 60:94–112.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827.

Mowshowitz, A. (1968). Entropy and the complexity of graphs: I. an index of the relative complexity of a graph. *The bulletin of mathematical biophysics*, 30(1):175–204.

Nassimbeni, G. (1998). Network structures and co-ordination mechanisms: a taxonomy. *International journal of operations & production management*.

Newman, M. (2018). *Networks*. Oxford university press.

Newman, M. E. (2002). Assortative mixing in networks. *Physical review letters*, 89(20):208701.

Newman, M. E. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113.

Nikolić, S., Kovačević, G., Miličević, A., and Trinajstić, N. (2003). The zagreb indices 30 years after. *Croatica chemica acta*, 76(2):113–124.

Onnela, J.-P., Fenn, D. J., Reid, S., Porter, M. A., Mucha, P. J., Fricker, M. D., and Jones, N. S. (2012). Taxonomies of networks from community structure. *Physical Review E*, 86(3):036104.

Pathan, A.-M. K. and Buyya, R. (2007). A taxonomy and survey of content delivery networks. *Grid Computing and Distributed Systems Laboratory, University of Melbourne, Technical Report*, 4:70.

Piccardi, C. and Tajoli, L. (2018). Complexity, centralization, and fragility in economic networks. *PloS one*, 13(11):e0208265.

Porter, M. A., Onnela, J.-P., and Mucha, P. J. (2009). Communities in networks. *Notices of the AMS*, 56(9):1082–1097.

Randic, M. (1975). Characterization of molecular branching. *Journal of the American Chemical Society*, 97(23):6609–6615.

Ravasz, E. and Barabási, A.-L. (2003). Hierarchical organization in complex networks. *Physical review E*, 67(2):026112.

Rombach, M. P., Porter, M. A., Fowler, J. H., and Mucha, P. J. (2014). Core-periphery structure in networks. *SIAM Journal on Applied mathematics*, 74(1):167–190.

Roy, M., Schmid, S., and Tredan, G. (2014). Modeling and measuring graph similarity: The case for centrality distance. In *Proceedings of the 10th ACM international workshop on Foundations of mobile computing*, pages 47–52.

Skorobogatov, V. and Dobrynin, A. (1988). Metric analysis of graphs. *match*, pages 105–151.

Smith, A., Calder, C. A., and Browning, C. R. (2016). Empirical reference distributions for networks of different size. *Social networks*, 47:24–37.

Todeschini, R. and Consonni, V. (2008). *Handbook of molecular descriptors*, volume 11. John Wiley & Sons.

Valente, T. W., Coronges, K., Lakon, C., and Costenbader, E. (2008). How correlated are network centrality measures? *Connections (Toronto, Ont.)*, 28(1):16.

Van Wijk, B. C., Stam, C. J., and Daffertshofer, A. (2010). Comparing brain networks of different size and connectivity density using graph theory. *PloS one*, 5(10):e13701.

Wang, T. and Krim, H. (2012). Statistical classification of social networks. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3977–3980. IEEE.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge University Press.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442.

White, D. R. and Reitz, K. P. (1983). Graph and semigroup homomorphisms on networks of relations. *Social Networks*, 5(2):193–234.

Wilson, R. C. and Zhu, P. (2008). A study of graph spectra for comparing graphs and trees. *Pattern Recognition*, 41(9):2833–2841.

Zhang, W. and Thill, J.-C. (2019). Mesoscale structures in world city networks. *Annals of the American Association of Geographers*, 109(3):887–908.

Zhang, X., Martin, T., and Newman, M. E. (2015). Identification of core-periphery structure in networks. *Physical Review E*, 91(3):032803.

Conclusions

In addition to the practical conclusions set out in Sections 1.7, 2.5 and 3.6, here we summarise the main innovations contained in the three Chapters. This thesis does not reach mathematical/demonstrative conclusions, but in the full field of application of the PhD in Quantitative Methods for Economic Policy, it makes use of empirical approach to reach conclusions applicable to real problems. Moreover, in the context of the curricula of Network Analysis and Online Data Mining Methods for Economic Policy, it intends to open a window in the study of some relevant issues of network analysis for real economic problems and it is therefore open to theoretical formalizations or applicative extensions of the works contained in it.

The main innovations of this thesis are listed below:

- The reticular structure of the interdependencies between cultural traits is a new and fundamental component in the quantitative measurement of the national culture and therefore of the cultural distances between countries. Taking all together, the 10 cultural traits proposed by Inglehart-Welzel in their Cultural Map, can be considered a parsimonious solution, because they are suited to capture information about the culture of a country. However, it is undeniable that the information added considering 60 cultural traits is used in the right direction, providing even more evident results. Indeed, considering 60 variables adds interesting information in the relationship between cultural distance and other distances from literature, and much more to the impact on distance between countries in GDP per capita.
- Although cultural distances are not very correlated with distances from literature (connectedness, ethnic/linguistic, genetic and climatic), the found cultural distances are good regressors of the economic distance between countries in GDP per capita. Culture is an important determinant of economic mechanisms, especially when in its measurement is considered the network structure of interdependencies between cultural traits.
- In addition, we propose a subset of network descriptors, which is suitable of well discriminating networks according to their structure in the process of mapping them in a space. It can be useful when one studies a set of networks - even of different sizes - referring to the same field and one wants to use classical methods for objects less complex than graphs, such as vectors

of values.

The future extensions of these results are countless. Below is a list of them.

- The approach used in Chapter 1 can be easily applied to other WVS Waves, to other surveys and to the temporal evolution of culture. It can also be extended in all those situations where higher level objects that have an individual specification can be identified. From this point of view, the use in this thesis of graphical modelling is innovative.
- Measures to compare the new cultural distance index (Chapter 2) can be extended to other distance measures used in literature, such as HMISea or other genetic, climatic, religious, etc., or to new distances, like food tradition, music culture, cinema production, and so on. In the same way, the cultural distance can be included in other econometric models that vary for the investigated outputs or for the very nature of the model. For example, nodal (countries) attributes may be included, or other economic outputs may be considered.
- Finally, the work of the Chapter 3 lends itself to various extensions and specifications. As a first step, further simulations of various complexities could be implemented. Second, even more network descriptors could be included. Third, the Subgroup Discovery method can also be used with other specifications or replaced by other classification algorithms. Fourth, the work could be completed by mathematical formalizations of the properties of the found subset of descriptors. Finally, the possible case studies could be the most varied, both in the economic and other fields.