TÉCNICO LISBOA

Using Network Science to enhance the analysis of Delphi surveys' results in health settings

Carolina Campos Paes de Faria

Thesis to obtain the Master of Science Degree in

Biomedical Engineering

Supervisor: Prof. Mónica Duarte Correia de Oliveira
Co-Supervisor: Prof. Francisco João Duarte Cordeiro Correia dos Santos

Examination Committee

Chairperson: Prof. Mário Jorge Costa Gaspar da Silva
Supervisor: Prof. Mónica Duarte Correia de Oliveira
Member of the Committee: Prof. Ana Catarina Lopes Vieira Godinho de Matos

November 2021

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Preface

The work presented in this thesis was developed during the period of March 2021 to October 2021, at Instituto Superior Técnico (IST), under the supervision of Prof. Mónica Duarte de Oliveira and co-supervision of Prof. Francisco Correia dos Santos. This work was conducted in the context of FCT MEDI-VALUE (PTDC/EGE-OGE/29699/2017) and H2020 IMPACT HTA (grant agreement 779312) projects.

Acknowledgments

I would first like to thank my supervisor, Prof. Mónica Oliveira, for her guidance since day one. Her feedback and continuous motivation were fundamental throughout this work. I would also like to thank Prof. Francisco Santos for his advice and incredible insights, decisive to define the focus of this thesis.

A special thank you to my family and friends. For their patience and constant support during this stage where my priorities were not so easy to define. For always being there and helping me get here.

Resumo

A Avaliação de Tecnologias de Saúde (ATS) consiste na comparação de diferentes aspetos, de forma a avaliar tecnologias médicas e apoiar os processos de decisão referentes à sua utilização e financiamento. Aquando da determinação dos aspetos de comparação a serem considerados, o método Delphi é frequentemente usado para reunir opiniões de diversos atores-chave sobre a sua relevância. Apesar de inúmeros autores implementarem estes questionários neste contexto, a análise e a representação das opiniões ainda carecem de exploração. Este trabalho tem como objetivo investigar alternativas inovadoras para a análise de questionários Delphi de ATS. Para tal, propomos uma metodologia baseada em Ciência de Redes Complexas. Em particular, o algorítmo de Deteção de Comunidades de Louvain é utilizado para o *clustering* de atores-chave, consoante as suas respostas. Aplicámos a metodologia a dados dos projetos MEDI-VALUE e IMPACT-HTA, de forma a responder a duas principais questões: (1) verificar a adequabilidade desta abordagem para analisar questionários Delphi e (2) obter informações adicionais sobre as opiniões de atores-chave. Os resultados sugerem a adequabilidade da abordagem descrita, sendo que o modelo, quando aplicado a diferentes conjuntos de dados, origina comunidades com topologias semelhantes. A metodologia permite, ainda, obter informações sobre as opiniões dos diferentes atores-chaves. As comunidades não são caracterizadas pelo tipo de ator-chave, mas sim pelo tipo de respostas. Existe uma tendência para os critérios serem considerados relevantes. A propriedade de *Triadic Closure* também é observada nestas redes, i.e., se dois stakeholders estão em concordância com um terceiro, provavelmente irão eles também concordar no futuro.

Palavras-chave: Avaliação de Tecnologias de Saúde, Método Delphi, Proximidade de Opiniões, Análise de Atores-chave, Ciência de Redes Complexas, Deteção de Comunidades

Abstract

Health Technology Assessment (HTA) systematically compares distinct aspects to evaluate health technologies and support decision-making processes concerning their use and financing. When choosing the evaluating aspects, it is common to use the Delphi technique to gather opinions from several stakeholders regarding their relevance. Although many authors implement Delphi surveys in this context, the study and abstraction of stakeholders' views lack exploration. In this work, we investigate innovative alternatives for the analysis of HTA Delphi surveys. To do so, we propose a framework based on Network Science tools. More specifically, the Louvain Community Detection algorithm is applied to cluster stakeholders according to their answers. We implement the framework to data from MEDI-VALUE and IMPACT-HTA projects, exploring two research questions: (1) verify the suitability of this approach for the analysis of Delphi surveys and (2) obtain novel and relevant information regarding stakeholders' opinions. The results suggest that the described framework is suitable for Delphi analysis, with the model originating communities with similar topologies when applied to different datasets. Additionally, the framework allows new insights to be obtained regarding stakeholders' opinions and how they relate. Communities found are typically not characterised by stakeholders' type and instead by the kind of answers, and people tend to consider criteria as relevant. The Triadic Closure is also verified for these networks meaning that if two stakeholders both agree with a third stakeholder, there is an increased likelihood that they will agree in the future.

Keywords: Health Technology Assessment, Delphi Technique, Proximity of Views, Analysis of Stakeholders, Network Science, Community Detection

Contents

List of Tables

List of Figures

List of Acronyms

BBIVT Biomarkers-based *in vitro* Tests

BPA Buyers, Policymakers and Academics

CD Community Detection

CDC Consensus Development Conference

CN Complex Network

DBSCAN Density-Based Spatial Clustering of Applications with Noise

DNA Discourse Network Analysis

HPro Healthcare Professionals

HTA Health Technology Assessment

ICC Interclass Correlation Coefficient

IMD Implantable Medical Devices

IQR Interquartile Range

IRA Interrater Agreement

IRR Inter-rater Reliability

IST Instituto Superior Técnico

IST-ID Instituto Superior Técnico for Research and Development

LSE London School of Economics and Political Science

MANOVA Multivariate Analyses of Variance

ML Machine Learning

NGT Nominal Group Technique

NS Network Science

R1 Round 1

R2 Round 2

UL Unsupervised Learning

Chapter 1

Introduction

1.1 Topic Overview

The literature recognises that decision-making processes in the health sector are highly complex (Stratil et al., 2020). First of all, health decision-making is influenced by a careful deliberation of many normative and technical criteria, which are often in conflict with one another (Stratil et al., 2020). Secondly, the values and perceptions of different stakeholders regarding these criteria frequently vary across and within societies, leading to several disagreements (Stratil et al., 2020).

One health process particularly challenging is resource allocation (Jakab et al., 2020). When choosing to use and finance health technologies, Health Technology Assessment (HTA) is used to support decision-making processes (Street et al., 2020; Polus et al., 2019). HTA is based on a systematic evaluation of several factors, used to assess and compare the properties, effects, and impacts of health technologies or interventions (Street et al., 2020; Polus et al., 2019; Infarmed, 2021). Therefore, the main goal of HTA is to provide the most refined available information, supporting decision-makers in a transparent and informed decision (Jakab et al., 2020; Stratil et al., 2020). HTA is based on a set of criteria with a high conflict potential, such as efficacy, safety, cost-effectiveness, or fairness (Jakab et al., 2020). Thus, challenges appear concerning the measurement of the relevance of distinct aspects and the agreement of different stakeholders, each with different views and priorities. With this in mind, organisations are becoming more aware of the importance of defining and measuring the relevance of HTA aspects as well as bringing the different stakeholders into the discussion (Jakab et al., 2020; Street et al., 2020).

Due to the significant impact of positively involving stakeholders, many approaches have been studied and developed. Several participatory approaches are known to be used. This work focuses on a particular one frequently cited in the field of decision-making in health - the Delphi Technique.

In summary, the Delphi technique consists of a series of surveys based on four principles - anonymity,

iteration, controlled feedback, and statistical aggregation of responses (Geist, 2010). This approach removes geographic and time challenges, allowing every stakeholder to participate while reducing the adverse sociological effects of group interactions (Belton et al., 2019; Geist, 2010). The foundations and principles of Delphi will be discussed in more detail in chapter 2.

Delphi's characteristics are critical in the current times. Although face-to-face approaches, such as conventional meetings, can be efficient, it is frequently not possible to gather everyone in the same room (Geist, 2010). Nowadays, this is even more a reality, given the world's globalisation and society accepting and embracing a *remote* lifestyle, especially after the COVID-19 outbreak (Sintevi et al., 2021).

Not only in HTA but for health research in general, the Delphi method is commonly used for many purposes (Keeney et al., 2006; Hasson et al.; Sintevi et al., 2021). When Delphi is used to collect opinions from a vast number of stakeholders, it is typical for a smaller group of decision-makers later to discuss results through a combination of other participatory approaches (Bowling, 2014). Innovative analysis and report of the results support decision-makers, providing them with new perspectives for this later discussion, improving resource allocation processes.

1.2 Objectives

Research shows that involving stakeholders unquestionably brings positive results into the evaluation of health technologies. Additionally, several authors explore the construction and limitations of Delphi surveys and implement them in specific cases. However, the analysis and representation of stakeholders' views in the health context lack exploration. Typically, data is analysed using traditional statistical tools, and results are presented using exhaustive and complex tables. Besides not being presented straightforwardly, the analysis is frequently done considering the pre-defined types of stakeholders. Even though inter and intra-group analyses are performed, this approach can be reductive since it assumes that stakeholders' views are strongly influenced by their type.

Given this scenario and with the question "Which HTA stakeholders share more similar views?" in mind, we want to explore alternatives for analysing Delphi results in health settings. Taking into consideration the powerful applications of Network Science (NS) in several contexts, we believe that the application of NS tools to analyse Delphi results can be not only innovative but also highly promising in bringing a new perspective to Delphi analysis. Thus, we propose a groundbreaking, state-of-the-art framework, combining NS and Community Detection (CD), designed to analyse results from web-Delphi surveys in health settings. Toward this goal, Delphi surveys' results from two HTA projects - MEDI-VALUE and IMPACT-HTA - are used. These results reflect the opinion of different stakeholders on the relevance of distinct aspects considered in HTA.

In a nutshell, we want to cluster the involved stakeholders according to their opinions and priorities, identifying which ones share similar views. Starting with no pre-defined groups and understanding which

clusters emerge, this work has the main objectives of (1) investigating the suitability and performance of NS tools for the analysis of Delphi results and (2) exploring what information about stakeholders' views is revealed. This way, we hope to enhance health decision processes following HTA Delphi surveys, enabling strategic coalitions in negotiations and resolving possible conflicts. For this purpose, we propose to address two research questions, each answered in one research stage, as presented in Figure 1.1.

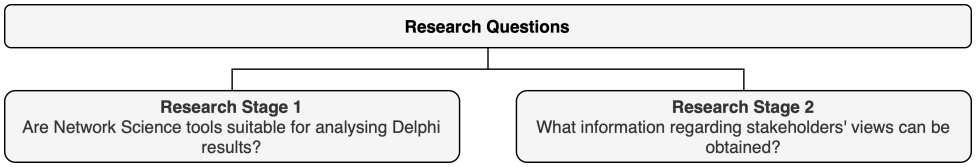


Figure 1.1: Research questions to be answered.

There is no evidence of the eligibility and performance of NS tools in the context of Delphi, health-related or not. Thus, it is the first objective of this work to verify it, answering the question "Are Network Science tools suitable for analysing Delphi results?". For that purpose, the topology of the networks and communities for different datasets is analysed. This step allows understanding the conversion of data into network-based data and verifying if a typical structure of "agreement networks" exists.

After verifying the appropriateness of the approach, it is the second objective of this work to explore the information that can be retrieved regarding HTA stakeholders' opinions and relations, answering the question "What information regarding stakeholders' views can be obtained?". At this point, we want to be able to understand what characterises the obtained communities using CD tools. Is there a match between them and the previously defined stakeholder groups, based on their types? Or are they defined by other properties? Furthermore, we want to extract new information about the participants and their views, including examining the results from groups of related aspects.

During this work, while answering the two previous questions, other contributions are intended to be made. The two projects followed different protocols. In MEDI-VALUE, there was a single panel for all stakeholders. In contrast, in IMPACT-HTA, there were parallel panels for each group of stakeholders, based on their types. Understanding the influence of these conditions, alongside other variations, including the used scales, will also be considered throughout this work.

1.3 Thesis Outline

This thesis is composed of a total of eight chapters. Chapter 2 provides a context to this work, introducing the world of HTA and Delphi. Chapter 3 focuses on the projects providing us with data. Chapter 4 contains the state-of-the-art concerning the analysis of Delphi as well as Community Detection algorithms, including the basic theory of the related topics. Chapter 5 comprises the proposed framework, and chapter 6 presents its implementation. The results are unveiled and discussed in chapter 7. Finally, chapter 8 explores the main conclusions, study limitations and possible future steps.

Chapter 2

Context

Before proceeding further, this chapter presents an overview of the importance of stakeholder participation in health decision-making processes and possible participatory approaches, focusing on Delphi.

2.1 Healthcare decision-making and stakeholder participation

As discussed in chapter 1, organisations are paying increasing attention to the methods used to evaluate technologies within the scope of HTA and adopting a more holistic view, involving every stakeholder.

Regarding the involvement of stakeholders, it is crucial first to discuss who are the stakeholders in health. According to Lübbeke et al. (2019), stakeholders can be defined as the ones - people, groups, or organisations - who have an interest or concern in an organisation, being able to affect and be affected by the organisation's actions, objectives and policies. In healthcare, "The four Ps" - patients, providers (professionals and institutions), payers, and policymakers - are considered the main stakeholders. However, a more inclusive definition should also include the industry (medical device, pharmaceutical, biotechnology), regulators, research community and media as stakeholders for their essential role.

As in any other industry, in the health sector, different stakeholders broadly differ not only on their interests and needs but also in their influence and priorities, being their identification and the analysis of their opinions and interests crucial for the success of the organisation (Lübbeke et al., 2019).

While the inclusion of providers, payers, and policymakers in HTA is explicit, the participation of patients in health decisions might not be, at least not until recent years. Patients are, however, being turned into customers of health services, creating a patient-centred healthcare, with the belief that giving patients purchasing power improves health services (Muehlbacher and Kaczynski, 2016; Street et al., 2020). This broader involvement contributes to an improvement of services and higher transparency while paying attention to the opinions of patients and carers who experience the technologies "first-hand".

2.2 Participatory Approaches in Health

Before discussing the technique highlighted in this work - Delphi - we will, in this section, introduce a broader context of the participatory approaches used in health settings. This way, we believe that the use of Delphi and its combination with other techniques will be more straightforward.

There are currently several research methods used in health investigation for addressing different research questions (Bowling, 2014). The wide variety of research methods available try to address different questions, and their use should be according to the goal and conditions of the desired study (Bowling, 2014).

Possible approaches for gathering stakeholders' opinions include individual techniques, such as surveys or interviews, or group methods, such as focus groups, workshops, or traditional meetings. In 2008, Bourrée et al. (2008) defended four main group consensus methods for healthcare use - Delphi, Nominal Group Technique (NGT), Consensus Development Conference (CDC) and RAND/UCLA, a modified version of Delphi.

Even though the term "consensus" is commonly used, nowadays, the goal of using Delphi and these other techniques is often no longer to reach consensus and instead to search for group communication and share of views and priorities to solve a complex problem (Landeta, 2006). Thus, we should interpret Bourrée et al. (2008)'s statements in a more updated way, considering these techniques suitable for gathering opinions but not necessarily common agreement. Actually, although Bourrée et al. (2008) presents these methods as "consensus methods", the authors shift their attention from consensus to the fact that in public health, these methods are used for decision-making and generation of ideas rather than total agreement. Given all, these methods can be considered important stakeholder participation methods, with Bowling (2014) in 2014 also stating Delphi method, CDC and NGT as the "three main methods of establishing consensus views".

Section 2.3 discusses the Delphi technique. Regarding the other two main approaches, briefly, CDC is a conference where a multidisciplinary panel writes guidelines on a given topic, after experts present a synthesis of the current knowledge and answer possible questions. Usually, a promoter, commonly a public institution, defines a topic, finances and holds the conference (Bowling, 2014). This method is commonly used in the healthcare sector (Bourrée et al., 2008), particularly in HTA by large health care, government organisations and medical-doctors-representing bodies (Bowling, 2014). However, it can be costly and requires high levels of organisation (Bowling, 2014). The NGT is a group meeting led by a coordinator. First, experts form their individual opinions, and then all views are presented, discussed and rated. Finally, the points of each item are counted, and the results are discussed (Bourrée et al., 2008; Bowling, 2014). This method is frequently used in health for gathering a group opinion but commonly criticised, for example, for not considering patients' and carers' preferences or for the criteria used, which are often not based on scientific evidence (Bowling, 2014).

All techniques have their pros and cons, and their use depends on the study's goal, not being easy to choose one when the work's scope is not mentioned (Bowling, 2014). For instance, although Delphi registers better results when compared to statistical groups and classic groups using direct interaction, there appear to be no better or worst results for Delphi when compared to similar techniques such as NGT (Landeta, 2006).

Considering group approaches, they have the advantage of allowing the formation of a group opinion. There are two main benefits of Delphi when compared to the other group approaches. First, anonymity, with the removal of adverse sociological effects of group interactions. Second, the possibility of involving a large group of stakeholders, since there is no number of participants restriction and there is a reduction of time, and geographic challenges (Landeta, 2006; Bourrée et al., 2008). However, the anonymity advantage can also be a downside since it does not allow participants to discuss (Hasson et al.). Thus, a mixed-use is typical (Bowling, 2014), including for Delphi to be followed by another participatory approach. For instance, one can perform a Delphi survey to gather opinions from several stakeholders and then discuss a summary of those results in a CDC. The Delphi technique and its variations and limitations will now be explored in section 2.3.

2.3 Delphi Technique

This section describes the Delphi technique, presenting its fundamental concepts, different types, and variations, along with its advantages and limitations.

2.3.1 History and fundamental concepts

The Delphi method is a widely known technique used in forecasting and decision-making, developed in the 1950s at the RAND Corporation (Landeta, 2006; Linstone and Turoff, 2011). Before we proceed, it is important to refer that the Delphi technique has "evolved dramatically" since its first application (Hasson and Keeney, 2011), resulting in a significant inconsistency around its definition, how studies should be conducted and how their results should be reported, an issue evidenced by the several definitions and variations found in the literature (Belton et al., 2019).

One proposal is the definition of the Delphi method as a consensus-building tool aiming to promote the involvement of all stakeholders during a decision-making process (Geist, 2010). A focus on consensus is, however, not needed. Through the use of a series of questionnaires in a controlled way, Delphi surveys were first used to get a reliable consensus of a group of experts (Landeta, 2006). However, a more updated definition considers Delphi a "method of structuring communication between a group of people who can provide valuable contributions to resolve a complex problem" (Landeta, 2006), with no obligatory search for consensus.

Despite all inconsistencies and variations, there are fundamental characteristics that are common to every well-performed Delphi study[1]. Regardless of being done remotely or not, the basics of Delphi point to a series of questionnaires, answered by experts on the field, in a repetitive way, where participants should answer the same question at least twice so that they can, after receiving information on the overall responses, reconsider their answer, resulting on a final extensive outcome (Landeta, 2006). Four main characteristics define Delphi studies:

i. First, the anonymity of the experts, or at least of their answers (Landeta, 2006), needs to be granted. Individuals should be able to respond without being judged by other participants (Geist, 2010). Measures to guarantee anonymity include participants taking the surveys alone at their homes or offices (Geist, 2010);
ii. Iteration is the second feature characterising this technique. A Delphi study should start with a generative round presenting the topic, and, at this stage, panellists have the chance to generate comments and ideas. Then, the researcher presents a survey based on the given inputs, which should be completed at least two times (Geist, 2010), resulting in several iterations;
iii. The third property of Delphi is the controlled feedback happening between iterations. The researcher uses inputs from the previous iteration to provide participants with feedback, allowing them to read and comment on these qualitative results and change their minds based on them (Geist, 2010);
iv. Finally, the fourth feature of these studies is the statistical group response, meaning quantitative feedback based on the ratings of each question, provided at the end of all iterations, in the format of a conclusion considering all the given opinions (Landeta, 2006; Geist, 2010).

Delphi characteristics allow an equal participation chance to all stakeholders while reducing the negative effects of group interactions (Geist, 2010). In traditional face-to-face meetings, where anonymity is not assured, there is a tendency of ”low-status members” to support ”high-status members” or dominant personality individuals (Landeta, 2006). Also, there are more opportunities for behavioural inhibition, loss of focus or pursuant of a single idea (Geist, 2010). As noted before, this method powerfully removes geographic and time boundaries, allowing stakeholders to participate from any place and most of the time, if not a real-time Delphi, at any time. This advantage is even more reinforced with the current use of online versions (Linstone and Turoff, 2011).

2.3.2 Variations of Delphi

As mentioned before, the Delphi technique has evolved drastically. One of the main consequences of this evolution is the need to constantly consider new variations and applications of the method (Hasson and Keeney, 2011).

[1]Note that a *well-performed Delphi study* is strongly difficult to define. There are, however, some available proposals of guidelines for that purpose. For instance, (Belton et al., 2019) outlines ”six crucial procedural steps of conducting a successful Delphi study”.

Different Delphi surveys are conducted for a wide variety of research purposes (Hasson and Keeney, 2011). In addition, the characteristics of surveys can differ regarding the number of rounds, the level of anonymity, the type of feedback given, or even the method of analysis (Hasson and Keeney, 2011). In fact, there is no consensual division for the types of Delphi, with the term ”Modified Delphi” being commonly used to refer to any variation. Some authors have, nevertheless, proposed some Delphi types. Hasson and Keeney (2011) reefers ten main categories: classical, modified, decision, policy, real-time, e-Delphi, technological, online, argument, and disaggregative policy.

According to Hasson and Keeney (2011), the classical Delphi design is performed to generate opinions and achieve consensus via postal. Experts are selected according to the research's goals, and three or more rounds follow an open qualitative first round. The other types arise from this first one. The modified Delphi can go through fewer rounds, and its goal varies with the nature of the project, from predicting events to reaching a consensus. The first round may also be changed to pre-selected items, and the technique can be performed in different ways, including an online format.

The main differences among other types rely on the objectives of the project and the format of the surveys. Real-time Delphi presupposes the use of computer technology where consensus on a topic is achieved in real time. E-Delphi, online Delphi, and technological Delphi, with the latest using hand-held keypads, count on online tools, via email, online web surveys, chat rooms and forums, being the key characteristics of each type challenging to identify. Pivotal aspects of the other types are related to the goals of the project. Decision Delphi focuses on gathering opinions for decision-making, policy Delphi on generating opposing views on policy and finding possible solutions, argument Delphi on elaborating arguments on a specific issue and coming up with reasons for distinct views and disaggregative policy Delphi on discussing future scenarios and probable or preferable futures.

Still, these are only some of the several Delphi variations found in the literature. Overall, new variations of Delphi are constantly being described. Online apps for conducting Delphi surveys are gaining a higher reputation, giving a name to other variance, the Web-Delphi.

2.3.3 Applications in health

In the healthcare research field, the powerful tools of Delphi are widely used for collecting the opinions of a vast number of stakeholders (Keeney et al., 2006; Hasson et al.; Sintevi et al., 2021). Bourrée et al. (2008) found, in 2008, Delphi applications in several health-related domains, including strategy orientation in public health, health-related education, prevention priorities, the definition of professional practices and their improvement, quality of care, medical practices evaluation, epidemiology and clinical research. Additional, searching the terms ”Health Delphi” in *Web of Science* leads to 10,158 results (October 2021).

More specifically, a significant number of authors employ Delphi surveys in the context of HTA. A search

of ”Health Technology Assessment Delphi” in *Web of Science* leads to 324 results (October 2021), remembering that a great majority of these surveys is likely for private use and is not published. Using these questionnaires to understand stakeholders' opinions on the relevance of HTA aspects is a common practice. Furthermore, a combination of participatory approaches can and should be done, with the report of Delphi results being commonly discussed later by a smaller group of decision-makers (Bowling, 2014).

2.3.4 Limitations

Thus far, we have discussed the basics of Delphi and its variations. Nevertheless, it is essential to explore the limitations attached to this technique. The identification and study of these constraints grant an understanding of the weaknesses that should be taken into consideration when analysing Delphi results and the recognition of possible improvements and future steps.

First of all, despite reducing the adverse effects of group interactions, this method brings other social impacts which should be considered. Literature has shown that when participants receive fabricated or distorted feedback between iterations, there is sometimes a tendency for them to change their opinions based on this incorrect information (Geist, 2010). This reaction is called the *Band Wagon Effect* and should be kept in mind when using Delphi surveys since it reflects the possibility of participants following the opinion of the majority and changing their minds in response to others' behaviour (Geist, 2010).

Second, research points out that, when not performed with the required knowledge, Delphi studies can lead to disappointing results (Landeta, 2006). In particular, for being perceived as a simple technique, non-experts commonly use it, increasing dissatisfying results (Landeta, 2006). Even though all techniques are threatened by careless execution, a poor design of questionnaires or experts' choice, a reckless feedback provision or an untrustworthy analysis of results are all factors that may jeopardise outcomes (Geist, 2010).

Third, regarding the participants, it is crucial to notice that anonymity is sometimes considered a ”quasi-anonymity” (Keeney et al., 2006). Some authors defend that true anonymity cannot be assured since participants are anonymous to each other but not to the researcher. Additionally, participants may not know who gave each response, but they might know who the participants are, and discuss the answers between rounds (Keeney et al., 2006). Also, it is essential to carefully choose the experts who compose the panel and assure the best conditions for their participation. Problems related to fatigue are less reported with more interactive and modern questionnaires. Still, participants may feel that they are being asked to do a lot or that ”they have been used and that they have received practically nothing in return” (Landeta, 2006). It is crucial to assure that the experts are involved, informed, and interested (Landeta, 2006). Finally, the anonymity advantage can also be a downside since it does not allow the topic's discussion between participants (Hasson et al.).

Forth but not least, the methodology and science behind Delphi are often discredited. Doubts regarding its reliability, validity, and trustworthiness are constantly being raised, with some authors suggesting that the Delphi technique is not a scientific method (Hasson and Keeney, 2011). The constant evolution of the method and the appearance of new variations do not help. While this flexible profile leads to several advantages, it also brings some concerns into the discussion, with a continual need for new measuring instruments for each variant and application (Hasson and Keeney, 2011). Additionally, it is challenging to measure the method's accuracy (Landeta, 2006).

Overall, regardless of its limitations, Delphi promising applications are unquestionable. Nevertheless, it is always good to state its current limitations to choose better the tools for the results' analysis. Moreover, the awareness of these limitations allows a more credible and conscious application.

2.4 Conclusion

In this chapter, we presented a context for this work. First, in section 2.1, we discussed the involvement of stakeholders in healthcare decision-making processes. In section 2.2 we presented possible approaches for this involvement and how Delphi can be used as a pre-CDC method. Finally, in 2.3 we introduced the Delphi technique since we will be using data from Delphi surveys. In chapter 3 we will present the projects which provided us with the data to be analysed and scale-originated data.

Chapter 3

Case studies using Delphi processes in HTA and scale-originated data

This chapter introduces the two HTA projects that provided us with Web-Delphi surveys' data - MEDI-VALUE and IMPACT-HTA. We start by describing the projects, including their context and goals. Following, we detail how the surveys were conducted, presenting the conditions in which they were performed. Next, we compare both case studies. Finally, we finish this chapter by discussing scale-originated data and its standard treatment and concerns, focusing on the scales used in the case studies.

We want to inform the reader that the results were provided to us after the development of the Delphi surveys, meaning that this work does not explore their construction and execution. Both projects used the decision support system *Welphi* (Welphi, 2021), an online platform for Web-Delphi surveys conduction.

3.1 MEDI-VALUE

We start with MEDI-VALUE, a project focused on advancing HTA literature. Its primary goal is to develop "innovative methodologies and tools to assess the multidimensional value of medical devices, being ultimately aligned with promoting resilient health systems that balance access to care with innovation and sustainability" (MEDI-VALUE, 2021). It is a collaboration between two academic institutions - Association of Instituto Superior Técnico for Research and Development (IST-ID) and London School of Economics and Political Science (LSE) -, the Portuguese national HTA agency (Infarmed) and hospitals (Centro Hospitalar Lisboa Norte, Hospital do Espírito Santo and Instituto Português de Oncologia de Lisboa).

MEDI-VALUE intends to design and implement methods to facilitate the involvement of several health stakeholders and their consensus in the structuring and development of Medical Devices' evaluation models. It focuses on four main questions: (A) "What contributes to medical devices value and how it

can be measured?", (B) "Do stakeholders have similar perspectives on what is medical devices value?", (C) "Which models can be used to evaluate distinct medical devices in practice by the HTA regulating agency and hospitals?" and (D) "Do evaluation models differ across medical devices, for the regulator and hospitals? Which are the policy implications?". This thesis focuses on questions (B) and (C), whose contribution to MEDI-VALUE relies on enhancing the analysis and use of the generated Delphi results.

3.1.1 Web-Delphi structure

The Web-Delphi performed in MEDI-VALUE had two rounds. On each round, there were two screens, one for Implantable Medical Devices (IMD) and a second one for Biomarkers-based *in vitro* Tests (BBIVT). For each screen, participants were asked to classify 34 aspects regarding its relevance for the evaluation of these technologies using a 4-point relevance scale plus a no-answer option ("Critical", "Fundamental", "Complementary", "Irrelevant" and "Don't know/don't want to answer"). A list of the aspects, divided into categories, is presented in appendix A.

Regarding the participants, invitations were sent using *Welphi* to 365 stakeholders who would include themselves in one of four groups of stakeholders: patients and citizens, Healthcare Professionals (HPro) (doctors, nurses, pharmacists, technicians), Buyers, Policymakers and Academics (BPA) or industry. Since participants classified themselves as a specific type of stakeholder in the first round of the Delphi, we have no information concerning the distribution of the invited stakeholders who did not attend the study. All stakeholders interacted with each other during the process, i.e. there was a single panel.

Table 3.1: Distribution of participants in MEDI-VALUE's Web-Delphi.

Stakeholder Group	Invited	%[1]	R1	%[2]	R2	%[3]
BPA	-	-	37	22.2%	31	23.1%
Healthcare Professionals	-	-	74	44.3%	60	44.8%
Industry	-	-	17	10.2%	15	11.2%
Patients and citizens	-	-	39	23.4%	28	20.9%
Total	365	100.0%	167	100.0%	134	100.0%

[1] Considering the total of participants invited (365).

[2] Considering the total of participants who participated in Round 1 (R1) (167).

[3] Considering the total of participants who participated in Round 2 (R2) (134).

In total, there were 365 invited participants, four stakeholder groups, and a 4-point relevance scale (plus a "Don't know/Don't want to answer" option) was used. The distribution of participants per group is described in Table 3.1. Although 365 stakeholders were invited, only 167 participated in the Web-Delphi, and only 134 stakeholders finished Round 2 (R2). In total, 198 invited stakeholders did not participate in the study and 33 did not continue from the first to the second round. Thus, the rate of not-answered invitations was around 45.8%, and the rate of dropouts from Round 1 (R1) to R2 was around 19.8%.

The report of results consisted of the presentation of the percentage of choice of each scale item, for each aspect, in both rounds, for IMD and BBIVT. Tables were used to present data.

3.2 IMPACT-HTA

IMPACT-HTA focuses on new and improved methods across ten areas, organised in ten ”Work Packages”. Its three main objectives are (1) understand variations in costs and health outcomes and the best practices of economic evaluation of new technologies, (2) assist HTA decision-making and health system performance measurement, developing innovative methodologies, tools and processes and (3) facilitate EU-wide cross-country collaboration between stakeholders (IMPACT-HTA, 2021).

As a collaboration of 15 international institutions, mainly universities (including Instituto Superior Técnico (IST)) and national institutes and agencies, IMPACT-HTA wants to address policy-relevant gaps in economic evaluations and selected performance measurement activities. To achieve its goal, the project relies on the combination of academic excellence in several fields and the collection of primary data on patient views and incorporates the perspectives and needs of national decision-makers and HTA bodies.

This thesis assists the Work Package 7 - ”Methodological tools using multi-criteria value methods for HTA decision-making”, developed in a collaboration between IST and the LSE. Its main goals are to create an analytical framework explaining HTA determinants, to generate and test a predictive model, evaluating results across different countries and therapeutic areas, to impact HTA decision-making processes and to evaluate new medicines with HTA agencies and relevant stakeholders. Again, our contribution relies on enhancing the potential analysis and use of the generated Web-Delphi results.

3.2.1 Web-Delphi structure

IMPACT-HTA’s Web-Delphi also had two rounds performed on *Welphi*. On each round, participants were asked to evaluate the sentence ”This aspect should be considered in the evaluation of new medicines on a common basis” regarding 24 aspects, using an agreement Likert scale (”Strongly disagree”, ”Disagree”, ”Neither agree nor disagree”, ”Agree”, ”Strongly agree”) with the additional ”Don’t know/don’t want to answer” option. A list of the aspects, divided into categories, is presented in appendix A.

Stakeholders were divided into six groups - Patients and Carers, Healthcare Professionals, Scientific Experts, Industry, HTA and Payers. Unlike MEDI-VALUE, the interaction was restricted to the same group of stakeholders. Thus, six identical Web-Delphi surveys were performed simultaneously for each group, i.e., there were six parallel panels. In total, there were 193 invited participants, six stakeholder groups and a 5-point Likert scale (plus a ”Don’t know/Don’t want to answer” option) was used.

The distribution of participants per stakeholder group is described in Table 3.2.

Table 3.2: Distribution of participants in IMPACT-HTA's Web-Delphi.

Stakeholder Group	Invited	%[1]	R1	%[2]	R2	%[3]
Healthcare Professionals	30	15.5%	25	15.0%	21	13.7%
HTA	32	16.6%	29	17.4%	28	18.3%
Industry	40	20.7%	31	18.6%	29	19.0%
Patients and Carers	26	13.5%	22	13.2%	19	12.4%
Payers	19	9.8%	18	10.8%	18	11.8%
Scientific Experts	46	23.8%	42	25.1%	38	24.8%
Total	193	100.0%	167	100.0%	153	100.0%

[1] Considering the total of participants invited (193).

[2] Considering the total of participants who participated in R1 (167).

[3] Considering the total of participants who participated in R2 (153).

Although 193 stakeholders were invited, only 167 participated in the Web-Delphi, and only 153 concluded both rounds. In total, 26 invited stakeholders did not participate in the study and 14 did not continue from the first to the second round. Thus, the rate of not answered invitations was around 13.5%, and the rate of dropouts from R1 to R2 was around 8.4%.

3.3 Comparison

In this section, we present an overview of both HTA projects' Web-Delphi. The main characteristics and differences between MEDI-VALUE and IMPACT-HTA projects are described in Table 3.3.

Table 3.3: Comparison between both projects.

	MEDI-VALUE	IMPACT-HTA
Web-Delphi platform	*Welphi*	*Welphi*
Number of rounds	2	2
Scale Used	4-point relevance scale plus a "No answer" option	5-point agreement Likert-scale plus a "No answer" option
Scale Structure	Unbalanced	Balanced
Number of stakeholders groups	4	6
Number of aspects	34	24
Interaction between stakeholders	Single panel	6 parallel panels
Number of invited participants	365	193
% of not-answered invitations	45.8%	13.5%
Number of participants in R1	167	167
% of withdraws from R1 to R2	19.8%	8.4%
Number of participants in R2	134	153

Both MEDI-VALUE and IMPACT-HTA are examples of studies using more than one collaborative approach. The Web-Delphi's were used to gather opinions from several stakeholders, and a summary of the results was then presented in a CDC. We expect to analyse these results to facilitate the bridge between the Delphi survey and CDC, better informing conference decision-makers in the future.

The recognition of the distinct surveys' conditions allows us to identify their influence and understand the origin of possible differences in results. It is possible to observe that the main differences rely on:

i. **Rating scales used** - MEDI-VALUE used a relevance rating scale, while IMPACT-HTA used an agreement rating Likert scale. Also, besides the "Don't know/Don't want to answer option", MEDI-VALUE used a 4-point scale while IMPACT-HTA used a 5-point scale, with a midpoint;
ii. **Scales' structure** - MEDI-VALUE used an unbalanced scale while IMPACT-HTA used a balanced scale. This concept will be soon explained in section 3.4;
iii. **Division of stakeholders and their interaction** - the division of participants was different, resulting in different numbers of groups (four in MEDI-VALUE and six in IMPACT-HTA). Also, MEDI-VALUE allowed all participants to interact with each other (single panel) while IMPACT-HTA performed separated Delphi's for each group (six parallel panels);
iv. **Distribution of participants** - although both Delphi surveys counted with the same number of participants on R1, the rate of not-answered invitations and withdraws was much higher (more than double) for MEDI-VALUE than for IMPACT-HTA. Additionally, the distribution of participants throughout the groups was different but not comparable since the stakeholders' division was not the same. It is also important to state that participants from MEDI-VALUE received feedback concerning the other 166 (R1) or 133 (R2) participants' views, while IMPACT-HTA stakeholders, only interacting within their group, were provided with feedback concerning fewer participants;
v. **Focus of the survey** - MEDI-VALUE focuses perspectives on criteria relevant for assessing specific technologies (Implantable Medical Devices and Biomarkers-based *in vitro* Tests) while IMPACT-HTA concerns overall new medicines.

We will now give a special focus to scales and scale-originated data. In general, research shows that the type and format of scales used in surveys can strongly influence results (Weijters et al., 2010). Additionally, we should be aware of several ways of treating and interpreting scale-originated data and concerns regarding its analysis. Thus, we finish this chapter by exploring the theory behind scales and their treatment and caveats, particularly for the scales used in MEDI-VALUE and IMPACT-HTA.

3.4 Analysis of scale-originated data

One crucial aspect to consider when developing a Delphi or any other survey is the choice of the type of response scale. Thus, this topic is largely discussed in the literature. The choice of the most appropriate scale depends on the type of question and depth, or grain, of responses required (Belton et al., 2019).

3.4.1 Theoretical concepts

There are many ways of classifying scales based on distinct criteria, with categorisation options commonly overlapping. We will here present some classifications.

Rating scales are close-ended survey questions. They are a variation of multiple-choice questions where participants are asked to rate abstract concepts, as satisfaction or importance, instead of being asked specific questions (QuestionPro; MeasuringU). In this work, we are concerned about rating scales.

According to Bowling (2014) and Kampen (2019) there are four levels of data. Given an attribute X and a measurement of the attribute M, and considering two objects i and j, a scale can be of four types:

1. Nominal, when numbers are used only for classification, for example, 'yes' = 1, 'no' = 0. In this case, if $x_i = x_j$ then $m_i = m_j$ and when $x_i \neq x_j$ then $m_i \neq m_j$;
2. Ordinal, when the items are somehow related to each other. For instance, 'very difficult' through to 'not very difficult'. These scales are at least nominal, and if $x_i < x_j$ then $m_i < m_j$;
3. Interval, when the characteristics are the ones of ordinal scales, but items' distance is of a known size. For example, the case of temperature scales. Thus, these are at least ordinal, and $x_i - x_j = \beta(m_i - m_j)$ for some $\beta > 0$;
4. Ratio, when the characteristics are similar to interval scales with the addition of a "true zero point" (in interval scales that zero is arbitrary), including weight scales. These are at least interval, and $x_i \div x_j = m_i \div m_j$.

Considering another classification of data, nominal and ordinal scales can be categorised as "qualitative" measurement scales, and interval and ratio as "quantitative" (Kampen, 2019).

There is another type of scale - attitude scales. An attitude can be defined as "an organised predisposition to think, feel, perceive, and behave toward a referent or cognitive object" (Desselle, 2005). Attitude scales arise in the context of questions in which respondents are asked to answer whether they have positive or negative feelings concerning the "referent or cognitive object". There are many types of attitude scales, including Thurstone, Likert, Guttman, and semantic-differential methods (Bowling, 2014).

Delphi surveys typically use Likert-type scales, i.e., the classical Likert scale or variations (Belton et al., 2019). The Likert scale is an ordinal scale proposed in 1932 by Likert (Sangthong, 2020) and measures the extent to which the participant agrees or disagrees with a statement, usually on a five-point scale (Wadgave and Khairnar, 2016; Jamieson, 2004). There are, though, distinct opinions regarding the more appropriate number of points, with a common belief among authors that 7-point scales are more reliable (Jamieson, 2004; Belton et al., 2019). The literature presents options from 3 to 11-points ranking scales. Debates also focus on the choice of an even or odd number of items and the midpoint option (Weijters et al., 2010). Likert scales have highly evolved, and variations are usually called Likert-type scales. In addition to several scales, there are also several ways of analysing and scoring the responses generated by them, a topic we will further discuss.

3.4.2 Case studies' scales

Following the proposed classifications, in MEDI-VALUE and IMPACT-HTA, we are dealing with rating, ordinal, qualitative, attitude, itemised scales in this work. However, although both scales fall into these categories, they are not equal, and their differences are crucial. Let us now recall the used scales.

MEDI-VALUE used a relevance rating scale, and participants were presented with a description explaining each label of the scale:

- ”Critical”: this aspect, besides being fundamental, can, by itself, forbid the evaluation of the technology's added value when compared to an alternative;
- ”Fundamental: this aspect should, without a doubt, integrate the evaluation of the technology to determine if it has an added value when compared to an alternative;
- ”Complementary”: this aspect is not fundamental, yet, can add value to the technology when compared to an alternative;
- ”Irrelevant”: this aspect should not be used to evaluate the technology; it is not applicable or does not allow to evaluate the technology's added value when compared to an alternative;
- ”Do not know/do not want to answer”.

IMPACT-HTA used an agreement Likert scale, with five points - ”Strongly disagree”, ”Disagree”, ”Neither agree nor disagree”, ”Agree”, ”Strongly agree” - plus a ”Don't know/don't want to answer” option.

There are concerns regarding both scales we will now explore.

3.4.3 Concerns regarding the used scales

There are ongoing solid debates regarding scales (DeWees et al., 2020; Kampen, 2019). Disagreements and doubts include constructing an appropriate scale, analysing the results mathematically, and psychologically interpreting them. We now present some concerning common behaviours.

Attribution of numbers to labels

According to Rasburn et al., Polisena et al., Gnatzy et al. and Weir et al., one common practice is to correspond numbers to the qualitative labels for parametric analysis. For example, ” 1 = ”Strongly disagree”, 2 = ”Disagree”, 3= ”Neither agree nor disagree”, 4 = ”Agree” and 5 = ”Strongly agree”. This behaviour is concerning, especially when the attribution is performed later and is not presented to participants.

As mentioned in Bowling (2014), corresponding numbers to labels is a problem because it tries to quantify something categorically. When done linearly, as presented in the previous example, another problem emerges - we treat ordinal data as interval data. Theoretically, no assumption of equal intervals

on ordinal scales can be made (Jamieson, 2004). For example, in the Likert scale, the distance between ”Strongly agree” and ”Agree” may be perceived differently, greater or smaller, than that between ”Agree” and ”Undecided” (Bowling, 2014). With this said, these scales can suggest an ordering of people’s opinions but not how distant these opinions are (Bowling, 2014).

To work this problem around, some authors, including Bishop and Herron (2015), propose the use of nonlinear numerical assignment, for example, ” 3 = ”Strongly disagree”, 11 = ”Disagree”, 17= ”Neither agree nor disagree”, 23 = ”Agree” and 31 = ”Strongly agree”. This way, the distance between the extremes is smaller than when compared to the middle item. However, in this case, it is tough to justify the choice of intervals and the nonlinear attribution would also affect mean and other statistical metrics.

These concerns not only affect the psychologically interpretation of responses but also originate one of the most problematic discussions around ordinal scales (Jamieson, 2004; Weijters et al., 2010) - the use of parametric tests when analysing ordinal data.

Parametric methods and ordinal data

With the assumption of attribution of numbers to the scales’ items, a more significant problem arises - whether or not to use parametric methods to analyse data. Parametric tests are meant to be used with continuous, interval data showing equality of intervals (Mircioiu and Atkinson, 2017; Wadgave and Khairnar, 2016). These methods are based on a normal or Gaussian distribution, defined by the mean and standard deviation (Mircioiu and Atkinson, 2017). Oppositely, nonparametric tests, usually using median and Interquartile Range (IQR) (Belton et al., 2019), are not based on any assumptions about the probability distribution of the data (Mircioiu and Atkinson, 2017).

According to Belton et al. (2019) and Jamieson (2004), due to the inaccuracy of considering equal intervals for consecutive items on ordinal scales, it is commonly defended that only nonparametric measures of central tendency should be used with ordinal data, with the concern of getting incorrect results otherwise. Additionally, the authors mention that parametric metrics lose meaning on ordinal responses. As exemplified by Wadgave and Khairnar, ”What does the average of ’never’ and ’rarely’ really mean?”. On the other side of the discussion, many authors including Norman (2010) and Mircioiu and Atkinson (2017) support the use of these methods with ordinal data with no fear of wrong conclusions. Supporters state that the question should be ”how important are the violations of theory assumptions?” instead of ”are the requirements met?” since empirical research demonstrate that the use of parametric tests does not affect results.

Arguments against the use of parametric methods with ordinal data are frequently based on theoretical assumptions while in favour arguments are based on evidence and simulations.

The efforts to use parametric tests are due to the advantages of these methods and the downsides of nonparametric tests. According to Norman, parametric tests are ”incredibly versatile, powerful, and

comprehensive". DeWees et al. (2020) describes that, usually, when compared to parametric tests, nonparametric tests are less powerful, and their interpretation is more complicated. Therefore, Mircioiu and Atkinson declares that using parametric statistics alongside graphical analysis, subsets analysis, and data transformation allows better and easier visualisation, interpretation, and in-depth analysis.

In summary, considering only theoretical assumptions, the use of parametric tests with ordinal data is not a good practice. However, when looking at practical and empirical studies, several pieces of evidence in the literature support this practice. One should consider and understand the risks and advantages and decide which side to stand in depending on the study, data and objectives.

Neutral/midpoint responses and the "Don't know/don't want to answer" option

Another problem emerges with the use of neutral items (Chyung et al., 2017), including the IMPACT-HTA's "Neither agree nor disagree" item and the "Don't know/don't want to answer" option.

Regarding the midpoint, this item is not mandatory. Some authors defend that there should not be a midpoint item since people should be forced to whether agree or disagree. However, sometimes, people purely do not have the information or knowledge, so it seems logical that they are not forced to do it.

With the scale chosen, the controversy does not concern using a midpoint item but how to interpret it. Research shows that people do not necessarily use it in the way it was intended. According to Chyung et al. (2017), some possible interpretations for this answer, apart from a neutral response, are that the participant is either confused, unsure, undecided, needs more information, or does not care.

To not have an opinion can be an opinion. We can assume that if two participants both choose "Neither agree nor disagree", they have a similar view, even if that view is uncertain, or at least more similar than with participants agreeing or disagreeing. However, considering equal intervals between items seems, as mentioned before, not right. This is why many researchers including Freitas et al. (2018) choose to consider two groups - (1)" Strongly Agree" +" Agree", (2)" Neither agree nor disagree" and (3) "Strongly disagree" + "Disagree", considering stronger proximity between the extreme items.

Moreover, there are few guidelines and proposals on interpreting the "Don't know/don't want to answer" option since the literature usually does not consider it. However, its purpose is also to represent a neutral position, meaning that conclusions can be equivalent to the midpoint.

Balance

Finally, we discuss the balance. If we ignore the "Don't know/don't want to answer", according to Bishop and Herron (2015), the IMPACT-HTA scale is balanced since it has an equal number of items on both sides of the neutral item, as shown in Figure 3.1 (A) and (B). Note that a neutral item is not required. A

4-point Likert scale with two items on the agreement side and two on the disagreement is also balanced.

Contrarily, as also presented in Figure 3.1 (C) and (D), the MEDI-VALUE scale is not balanced. In the MEDI-VALUE's scale, discarding the "Don't know/don't want to answer" item, there are four response options. Thus, the scale cannot be considered bipolar, meaning that there are no opposite nor balanced sides. "Strongly agree" can be considered, even if it presents some controversy, the opposite of "Strongly disagree", and the same for "Disagree" and "Agree". The same does not happen for "Critical" and "Irrelevant" and even less for "Fundamental" and "Complementary".

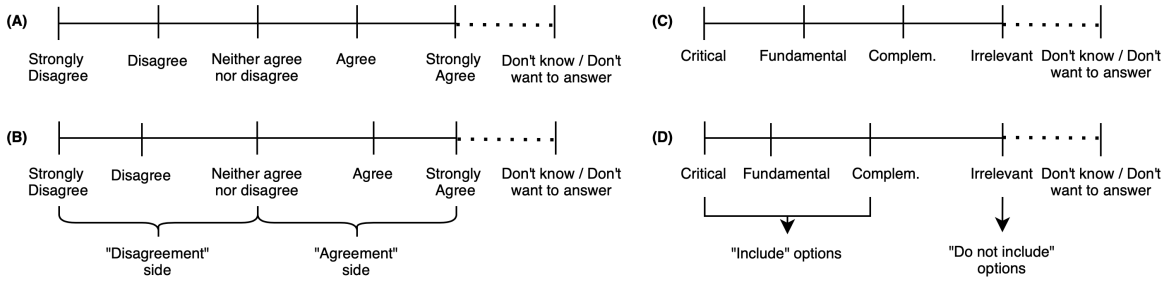


Figure 3.1: Visualization of possible distance between scales' items. (A) and (B) consider the IMPACT-HTA Likert scale. (C) and (D) refer to the MEDI-VALUE relevance scale.

Figure 3.1 shows the balance of the IMPACT-HTA scale in (A) and the unbalance of the MEDI-VALUE scale in (C), considering equal intervals for simplification and it also presents a proposal of the proximity of answers. In (B), it represented a Likert common assumption - the use of only three groups, with stronger proximity of the extremes (Chyung et al., 2017). In (D), there is a proposal of what this proximity might be for MEDI-VALUE. This relevance scale is not commonly used, meaning that there are no proposed groups. Additionally, these representations should not be considered official since these proposals of proximity were not presented to the participants. In (D), it seems reasonable to state that "Critical" and "Fundamental" are closer since they both find a given aspect essential and relevant. Another possible division is the one between the options which defend the inclusion of an aspect when evaluating a technology (the first three) and the ones that do not ("Irrelevant"). The distance of the items between in Figure 3.1 is just a representation, not made following any scale.

3.5 Conclusion

In this chapter, we introduced the projects collaborating with this work. In section 3.1 we presented MEDI-VALUE and in section 3.2 IMPACT-HTA. Next, in section 3.3 we compared both projects. Finally, in section 3.4 we paid more attention to the scales being used, presenting a brief theoretical introduction to scale-originated data and its concerns.

In chapter 4 we will explore essential topics related to the analysis and report of Delphi results as well as relevant concepts related to promising fields - Complex Networks (CNs) and CD.

Chapter 4

Literature Review

In the last chapters, we presented a general overview of this work's context and motivation. Before delving into more specific details of the proposed framework, we present, in this chapter, a review of the related work in the field. First, we investigate how Delphi results are typically analysed and reported. Next, we explore promising alternatives to analysing Delphi results - Unsupervised Learning (UL), clustering, and CN. Finally, we present applications of CD algorithms inspiring this work.

4.1 Analysis and report of Delphi results

More important than the conduction of a research study is the data generated and its analysis and report (Belton et al., 2019). In the specific case of the Delphi technique, there is a variety of ways to do it (Belton et al., 2019). In this section, we investigate the most common approaches for analysing and reporting Delphi surveys' results and we present practical examples of Delphi applications in health settings.

4.1.1 Analysis of data

Data generated in Delphi studies can be of many types, with the items' responses being commonly followed by comments. Also, these outputs can be analysed in several ways, using both qualitative and quantitative approaches (Belton et al., 2019; Evans et al., 2014; Hasson et al.). The analysis of data from Delphi surveys occurs between rounds when providing participants with feedback and, in a more detailed and exhausting way, to generate the final results at the end of the rounds.

The analysis of comments is not the focus of this work. Still, it is relevant. It is common to find authors who analyse comments manually. Nevertheless, there are also more efficient procedures. For instance, Evans et al. uses automatic strategies to identify common themes and key issues within comments.

For the measurement of singular items' responses, many authors, including Hasson et al.; Kearney et al.; Falzarano and Pinto Zipp perform statistical summaries for each item of the survey. Descriptive measures include the frequency counts for each item (Weir et al., 2006) and scores' range (Hoekstra et al., 2017). In order to measure a group opinion, meaning the collection of individual opinions, it is also possible, as presented in Freitas et al. (2018) to calculate the percentage of responses given for each scale item, for each question, on each round. This approach is found appropriate for Likert-type scales, and it is also the approach found in the MEDI-VALUE report, as stated before.

In a more holistic view, to describe group agreement, a variety of articles found in the literature including Kallio et al. (2020); Belton et al. (2019); Freitas et al. (2018); Hasson et al. frequently use descriptive analysis in Delphi studies, being the most common the measurement of:

1. Central tendency, i.e., mean, median and mode;
2. Level of dispersion, for example, standard deviation, coefficient of variation and IQR;

Particularly for the study of Inter-rater Reliability (IRR), stability and changes of opinions, and group's opinion variance, inferential statistics is often applied, as done by Freitas et al.. The authors calculate the IRR using Scott's Pi statistic, a kappa-like coefficient which is appropriate for nominal data with three or more labels. According to Weir et al. (2006), it is also possible to use Interclass Correlation Coefficients (ICCs). Regarding the stability of opinions between rounds, it is usual to use the McNemar Chi-square test, a nonparametric test used for the measurement of the degree of shift in responses between rounds (Freitas et al., 2018), or paired-samples t-tests (Weir et al., 2006).

Another important analysis is related to variations across the type of panellist or field of expertise. For instance, comparing patients' opinions with industry or healthcare professionals can be extremely valuable. To address the group's opinion variance, Multivariate Analyses of Variance (MANOVA) can be used (Freitas et al., 2018). Fisher exact test (Weir et al., 2006), and Kruskal-Wallis test (Evans et al., 2014) can also be employed. Not only inter-stakeholder group comparison is relevant but also intra-stakeholder, to explore views within each group, which can be made using IQR (Hoekstra et al., 2017). For this purpose, predefined groups of types of stakeholders are usually used.

To help with statistics analysis, software such as IBM SPSS Statistics (Hoekstra et al., 2017; Diamond et al., 2014) or R (Lange et al., 2020) are frequently used.

4.1.2 Representation of results

After the analysis, it is necessary to represent and report the results. Besides the written narrative, according to Belton et al. and Falzarano and Pinto Zipp, conventional ways of presenting results found in the literature include tables of descriptive statistics, such as central tendency values or frequency tables and graphical representations. The last includes bar graphs, plot graphs, boxplots, dendrograms, scatterplots and radar charts (as used by Freitas et al. for representing the group agreement). Belton

et al. also mention that is possible to use only written description, tables or graphical representations, or a mix, which is usually better, recommending the inclusion of graphics for easier visualisation of results.

4.1.3 Practical examples of Delphi results analysis and different approaches

We now present a review of different health-related Delphi studies, found in the literature, performed by several authors. Table 4.1 summarises the methods and visualisation tools employed in a sample of recent studies, representing the current approaches performed in Delphi surveys' results analysis.

Table 4.1: Review of different Delphi surveys' results analysis, found in the literature.

Article	Scope	Methodology	Visualisation tools
"Using a Modified Delphi Approach to Gain Consensus on Relevant Comparators in a Cost-Effectiveness Model: Application to Prostate Cancer Screening" (Keeney et al., 2021).	Implementation of a Modified Delphi to investigate stakeholders' opinions on the prostate cancer screening strategies to consider in a cost-effectiveness model.	r^*_{wg} statistic for calculating within-group Interrater Agreement (IRA). Measurement of the number of participants answering each response to questions.	Bar charts with the number of response selections for each question. A table with the final outcomes, including distribution of answers, consensus and IRA.
"Environmental responsibility in nursing in hospitals: A modified Delphi study of nurses' views" (Kallio et al., 2020).	Use of a modified Delphi to investigate nurses' views on environmentally responsible clinical practices and on each stakeholder role.	Measurement of the distribution of responses for each question.	Tables with the number with the percentage of responses for each question.
"Indicators for evaluating European population health: a Delphi selection process" (Freitas et al., 2018).	Employment of a Web-Delphi to promote agreement on indicators considered relevant to evaluate population health at the European regional level.	Inferential statistics to measure the level of agreement and opinion change. MANOVA to check if the field of expertise influenced the panellist responses.	Tables with the % of scale items chosen, MANOVA results and the comparison of responses between expertise groups. Graphics with the panellist's vote distribution curve. A radar chart with the level of agreement.
Health Technology Assessment methods guidelines for medical devices: How can we address the gaps? The international federation of medical and biological engineering perspective. (Polisena et al., 2018).	Use of a modified Delphi survey to reach a consensus among clinical and biomedical engineers on the proposed recommendations for medical devices HTA.	Median scores were calculated and IQR was used to represent the spread of the data and to assess the level of consensus.	Table with the median score for each question and the number of participants answering it.
"Identifying research priorities for effective retention strategies in clinical trials" (Kearney et al., 2017).	A Delphi survey was used to gain consensus amongst the registered CTUs on effective retention strategies.	Measurement of the distribution of responses for each question.	Tables with the number with the percentage of responses for each question.

As shown in table 4.1, authors use different methodologies for the analysis and report of the results.

However, they do not differ much. Even though it is becoming more common to use various tools, including innovative charts and different statistical measures, there is room for deeper exploration.

Although current methods are effective, usually serving their purpose, they represent few of the possible approaches that can be employed. For the particular case of HTA Delphi surveys, the results report is of great importance to support decision-makers, usually in further meetings or conferences. We believe that providing them with new and different insights and adopting a new perspective is of great interest. With this work, we do not intend to minimise the currently practised approaches. Instead, we propose exploring and capitalising data unconventionally, using unique and powerful tools already proved immensely promising in other contexts. We want to understand if different and fresh information can be obtained, with the final goal of enhancing health decision-making processes.

Considering the state-of-the-art techniques in data analysis, there is one that stands out - Machine Learning (ML). According to Markets and Markets (2021), the ML market is expected to grow from USD 1.03 Billion in 2016 to USD 8.81 Billion by 2022, at a Compound Annual Growth Rate of 44.1% during this period. According to Silva and Zhao, people from all over the world are using it for the most various purposes, from academic and scientific research to customer experience improvement and business operations. ML techniques provide effective data analysis, and their importance is currently undeniable.

Additionally, the world has never been so connected, and research reflects that. Another field that is gaining attention is NS, in particular the study of CNs, structures allowing innovative and easy representations of data connections and interactions, in highly diverse contexts (Silva and Zhao, 2016).

According to Silva and Zhao, the combination of both allows the use of ML powerful analysis tools in structured connected data represented by CNs. Again, the possibilities are huge, and the applications suit different contexts. We found no application of this strategy for HTA Delphi surveys analysis. However, we did find its use in other contexts with the goal of clustering stakeholders based on their views. For instance, Sathiyakumari and Vijaya (2016), and Chen et al. use this combination for social media networks' analysis. Others, including Buckton et al. (2019); Hilton et al. (2019); Fergie et al. (2019) focus on the clustering of stakeholders, regarding their argumentative similarities, however not in health settings. These strategies are found to be suitable and powerful for similar purposes in different contexts. We want, in this work, to understand if they are also appropriate and as robust for clustering HTA stakeholders, according to their responses in Delphi surveys and better understand how they relate.

The approach we propose is to represent our data using CNs and apply ML clustering techniques to group stakeholders according to their views. Thus, we will now explore these fields. In the next section, we will start by presenting the basic theory behind ML and clustering. After, we will describe the relevant information concerning CNs. Next, we will investigate the combination of both. Finally, we will present practical applications inspiring us.

4.2 Unsupervised Learning and clustering

The use of Machine Learning techniques with CNs is being explored for joining the power of ML with the data representation advantages provided by networks (Silva and Zhao, 2016). We will, in this section, start by discussing the basic concepts regarding clustering, an Unsupervised Learning technique. Later, we will explore CNs and how they can be possibly used along with clustering in Delphi results analysis.

According to Silva and Zhao (2016), ML can be defined as the study, design and development of algorithms providing computers with the ability to learn without being explicitly programmed. This work focus on UL, the field of ML aiming to describe behaviours or trends in the data, with no external information or labels being provided. More specifically, we are interested in clustering, the grouping of data into clusters in a way that similarity within clusters is maximised and minimised between different ones.

As mentioned by Ester et al. (1996) and Silva and Zhao (2016), there are two main types of clustering algorithms: partitioning and hierarchical algorithms. Partitioning algorithms, such as K-Means, determine all clusters at once, while hierarchical find successive clusters using previously determined clusters. Hierarchical algorithms can be agglomerative if they begin with separate clusters for each element and then merge them into larger clusters, or divisive if they begin with a large cluster which is then divided into smaller clusters. We can also consider another type of algorithms, the density-based ones, such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN), which is a spatial and partitional algorithm able to find clusters of any shape, based on the density of points within a certain ratio.

Since we want to represent our data using Complex Networks, we will now present the theory related to it to allow a complete comprehension of clustering techniques in these structures.

4.3 Complex Networks

According to Silva and Zhao, CNs are large-scale graphs with nontrivial connection patterns, which can describe a large variety of systems, allowing the capture of spatial, topological, and functional relations of the data. We start by presenting important definitions regarding the graph theory, the basics of networks.

4.3.1 Graph Theory

To properly understand the proposed framework, there are some basic concepts the reader should be familiarised with, including the basics of graph theory, the structures composing CNs.

Definition 1. ***Graph:*** *A graph G is defined as an ordered pair $\langle V, E \rangle$, where V is a finite nonempty set of vertices or nodes and E is the set of edges or links between the vertices $E \subseteq \{(u, v) | u, v \in V\}$ (Silva and Zhao, 2016).*

Note that, in the context of this work, we are dealing with directed graphs. For instance, for a social representation where users can ”follow” another user but not necessarily be followed by them, the ”following” situation would be represented by a directed graph (Raj P.M. et al., 2018).

Definition 2. ***Undirected graph:*** *A graph G is undirected when the relation E is symmetric, i.e.,* $\forall(u, v) \in E \implies (v, u) \in E$ *(Silva and Zhao, 2016).*

Definition 3. ***Weighted graph:*** *A weighted graph G is defined as* $G = \langle V, E, W \rangle$*, where* V *is the set of vertices,* E *is the set of edges and* W *is a matrix carrying the edge weights. This means that the entry* $W_{uv} = w, (u, v) \in E$ *fixes* $w > 0$ *as the weight of the edge linking vertices* u *to* v*. In the case where* $(u, v) \notin E \implies W_{uv} = 0$ *(Silva and Zhao, 2016).*

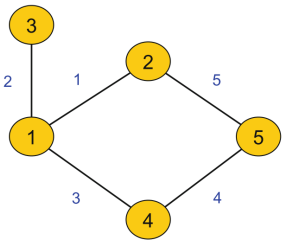


Figure 4.1: Example of an undirected weighted graph (Silva and Zhao, 2016).

Regarding the weight of the edges, it can represent, for instance, the strength of a friendship between two people in the network. Figure 4.1 shows an example of an undirected weighted graph with five vertices/nodes and the correspondent weighted edges, illustrating the concepts previously defined.

Definition 4. ***Adjacent vertices:*** *Two vertices* $u \in V$ *and* $v \in V$ *are called adjacent if they share a common edge, in which case the common edge is said to join the two vertices. In undirected graphs, if* u *is adjacent to* v*,then* v *must be adjacent to* u *as well (Silva and Zhao, 2016).*

Definition 5. ***Neighborhood of a vertex:*** *The neighborhood of a vertex* $v \in V$*, in a graph* G *is the set of vertices adjacent to* v*. The neighborhood is denoted by* $N(v)$ *and is formally given by* $N(v) = \{u : (v, u) \in E\}$ *(Silva and Zhao, 2016).*

Definition 6. ***Degree (valency or connectivity) of a vertex:*** *The degree of a vertex* v*, called* k_v*, in an undirected graph, is is the total number of vertices adjacent to* v *(Silva and Zhao, 2016).*

In the context of social representations, the neighbours of a vertex or node are the people with whom someone is connected, and the degree is the total number of neighbours.

4.3.2 Complex Networks models

As described by Silva and Zhao, aclpcn models arise to describe complex and high-scale graphs, following the same basic structure, with nodes and edges and respective relations. There is a large variety of

described Complex Networks models, depending on the structure and context of the data. Networks can present communities, meaning groups of nodes (or clusters) with many interconnecting edges. Vertices from different communities have typically relatively few edges interconnecting each other. Figure 4.2 describes a schematic network with strong community structure.

Much attention is now given to a specific type of network - social networks. Its particular property is related to the scope of the structure. According to Sathiyakumari and Vijaya (2016), in social networks, nodes represent people, organisations, or other entities, and the edges represent relationships or interactions. This type of network presents particular properties and patterns of topology. This definition can be adapted to the scope of our work since we want to explore networks where nodes represent stakeholders (people) and edges represent the distance of their views and opinions (interactions). It is then interesting to readjust this type and consider ”agreement networks”. The nodes and connections represent similar scenarios, and these structures can be compared.

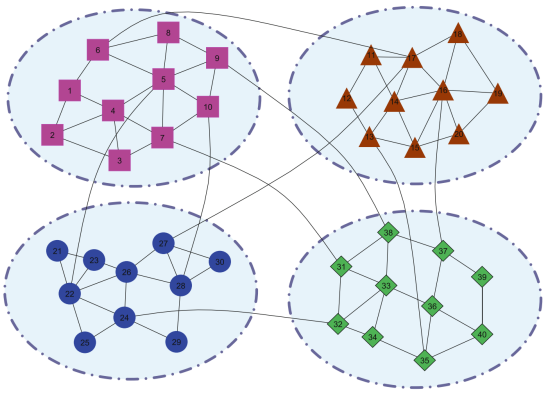


Figure 4.2: Schematic of a random clustered network. The network is divided into four communities and the nodes of each other are represented with different colors and formats (Silva and Zhao, 2016).

The detection of communities in Complex Networks, the technique we are about to present, is, according to Sathiyakumari and Vijaya, a hot topic in the literature. However, it is not an easy task, primarily due to the topology of the network and the possible overlapping of different communities.

4.4 Network-based clustering

In this section, we discuss network-based unsupervised methods, more specifically network-based clustering techniques and Community Detection.

The use of network analysis alongside clustering methods usually follows a specific framework (Silva and Zhao, 2016). First, it is necessary to construct the network from the original dataset with the definition of edges based on a given similarity measure. When the network is constructed, it is then possible to use clustering algorithms to detect communities on the structure. These communities can be defined

as subgraphs whose nodes are densely connected within them but sparsely connected with the rest of the network's nodes. Another way of putting it, according to Felfli et al. (2019), there is a "community structure" when a group of nodes present a higher probability of being linked when compared to other groups. The detection of these communities corresponds to the identification of the structure of the network, meaning its organisation, based on the interaction of nodes (Felfli et al., 2019).

Figure 4.3 represents the framework we just described, as well as the differences between data clustering and community detection. In data clustering (upper part), the algorithm finds similar groups (clusters) based on a similarity criterion. If the unstructured data is converted into network data, using a network formation technique, it is then possible to perform community detection (lower part), which allows the clustering of structured, organised network data points.

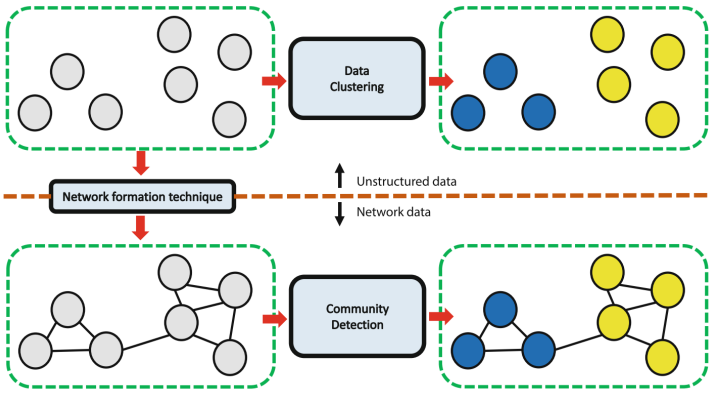


Figure 4.3: Data clustering and community detection (Silva and Zhao, 2016).

4.4.1 Community Detection Algorithms

As mentioned before, there are several available CD algorithms. According to Fortunato (2010), the earlier algorithms such as Kernighan-Lin can be considered traditional methods. These include graph partitioning, hierarchical clustering, partitional clustering and spectral clustering. However, these early methods are known for not fitting well real-world network data, including social networks (Felfli et al., 2019). In fact, the first algorithm which showed successful results for CD in this context was the one proposed by Girvan and Newman in 2002 (Felfli et al., 2019), a "non-traditional" divisive algorithm.

According to Felfli et al., the Girvan and Newman algorithm allows the detection of communities through the identification of edges lying between communities and removing them, leaving the community structures highlighted. The original method is, however, relatively slow and non-practical for networks of more than a few thousand nodes and thus, many alternatives and new approaches have been suggested, including modularity methods, Bayesian and regularised likelihood approaches.

We will, in chapter 5 discuss the theory behind some algorithms and justify our choice. For now, in section 4.5, we provide some applications which demonstrate how promising these techniques can be.

4.5 Application of Community Detection Algorithms

There are some applications of classical statistical clustering methods for stakeholder clustering, based on their opinions. For instance, in 2011, Veerappa and Letier used a weighted average linkage clustering algorithm to investigate stakeholders' similarity. This similarity was calculated using the distance between stakeholders' ratings, where smaller distances indicate higher similarity. For two ratings r_i and r_j from stakeholders i and j for the same requirement, the distance between them was defined as:

$$distance = |r_i - r_j| \tag{4.1}$$

For n requirements $R_1, R_2, ..., R_n$, the distance was computed as the Euclidean distance between the two sets of ratings for all n requirements:

$$distance = \sqrt{[(r_1 i - r_1 j)^2 + (r_2 i - r_2 j)^2 + ... + (r_n i - r_n j)^2]} \tag{4.2}$$

The case presented uses one of the earlier methods of CD, known for not fitting well real-world data. The results are a hierarchy displayed using a dendrogram. This study is an example of how stakeholder proximity can be analysed using clustering techniques. However, the algorithm can be considered outdated, and the visualisation tools are not much robust. Moreover, the study is not related either to the Delphi or health. Additionally, it considers numerical ratings, i.e., the ordinal scales' concerns come up.

With that being said, even though there is some evidence of clustering applications for stakeholders' opinions, this area lacks exploration. Few articles propose the use CD for this direct purpose and none in the context of HTA or Delphi. Still, new studies are appearing using more recent approaches. We will now present a particular example found relevant for this work, using Discourse Network Analysis (DNA), even though out of the context of Delphi and HTA.

4.5.1 Discourse Network Analysis and Community Detection

In 2019, in the UK, Hilton et al. published an article comparing stakeholder coalitions across pricing policies. In this study, for the identification of stakeholder subgroups, regarding their argumentative similarities, the Girvan-Newman edge-betweenness CD algorithm was used. Some of these authors published, in 2019, two more studies using DNA and CD. The first, Buckton et al. (2019), focused on Discourse Network Analysis to produce visual representations of stakeholders' networks and their coalitions regarding the "sugar tax" debate, before and after the announcement of the Soft Drinks Industry Levy in the UK. The other Fergie et al. (2019), following a similar methodology, used DNA for the same purpose but in the context of the minimum unit pricing for alcohol debate.

Although these studies concern discourse networks, i.e., process discourse data and are not Delphi-related, the methodology used for clustering stakeholders based on their views can be a great inspiration

for this work. Briefly, the methodology used in these studies comprises:

1. Data extraction and content analysis, where the data was gathered and properly treated. This part of the protocol has less to do with this work since the data type is different.

2. Network visualisation and analysis, where data was used to originate a network for visualisation and CD. This part of the protocol is of great interest for this work:
   (a) A weighted stakeholder × stakeholder matrix was created using the DNA software. Ties and their relative weights represented agreement or disagreement between stakeholders on individual concepts;
   (b) The ”subtract” transformation with ”average activity normalisation” was applied. This step was based on Leifeld's proposal to score stakeholders' agreement. A tie weight between two stakeholders was expressed as the number of concepts on which they have identical opinions minus the number of concepts on which these actors have diverging opinions. Then, the tie weights were divided by the average number of concepts the two mentioned for normalisation. Finally, a threshold value of 0.4 was used, where ties lower than this were replaced by 0.
   (c) The stakeholder × stakeholder network was imported into the visualisation software Visone;
   (d) Girvan-Newman edge betweenness community detection was applied to identify coalitions.

The Girvan-Newman algorithm has its limitations. It is, however, commonly used for CD in several contexts in recent works. Even though context and data type are different, the favourable results and applications of Community Detection with the purpose of understanding stakeholders relationships supports the suitability of the application of a similar framework to the HTA Delphi context.

4.6 Conclusion

This chapter focused on the main concepts of Delphi analysis and report and promising alternatives. First, in section 4.1 we discussed the commonly used methods and why other approaches should be considered. In section 4.2 we introduced clustering. In section 4.3 we analysed graph theory and CNs models. To understand how these two fields relate, in section 4.4 we explored network-based clustering. Finally, in section 4.5 we explored some relevant current uses of these techniques.

We found, among the literature, authors using standard statistical measures for analysing Delphi, both health-related or not. We also found studies using outdated clustering methods for the analysis of stakeholder views, representing their results as hierarchies. However, neither health nor Delphi related. Additionally, we found some researchers using NS for stakeholder proximity representation, based on their opinions. Still, the combination of these techniques was not found for Delphi studies nor the health context. The promising results of using this synergy of UL and networks in other areas, in other social networks, encourage great applications for Delphi analysis, including health surveys. This way, we believe we can present decision-makers with new and fresh perspectives and insights.

Chapter 5

Proposed Framework

In the previous chapters, we presented the context and objectives of this work and relevant related theory and literature review. We are now in a position where we can define and present our proposal. Thus, in this chapter, we propose a framework of an innovative approach on how to analyse and represent HTA Delphi results, using Network Science and Community Detection techniques.

5.1 Problem statement

We have already understood the importance of HTA and the complexity of informing decision-makers. As explained before, in this work, we aim to help and improve the way information is extracted from health Delphi surveys, enhancing the visualisation of the results and reports decision-makers are presented with. Before going any further, we remind the reader of the main goals of this work: (1) investigating the suitability and performance of NS tools for the analysis of HTA Delphi surveys' results and (2) exploring what information about health stakeholders' views can be extracted.

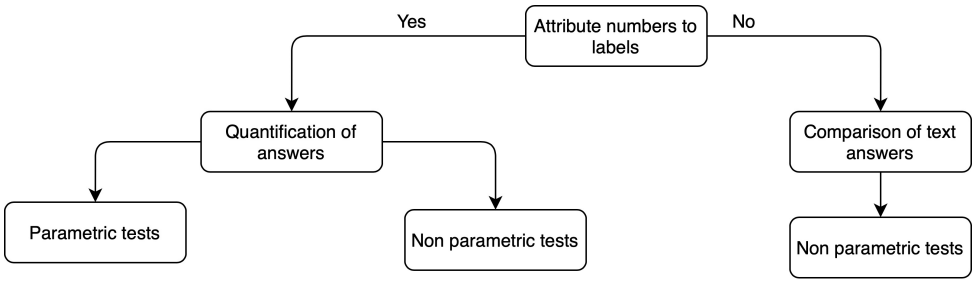


Figure 5.1: Possible approaches for data analysis.

We have already described the main concerns related to this approach. Firstly, we are dealing with the limitations of the Delphi technique itself and any inconvenience that can arise from the protocols.

Secondly, we have to be concerned about the caveats associated with the scales and ordinal data analysis. There are, as mentioned before, many options for analysing data, as presented in figure 5.1. Typical Delphi analysis often attributes numbers to labels, quantifying answers and using both parametric and nonparametric tests. If that is not done, only a comparison of answers and nonparametric tests can be applied. The pros and cons of these choices were already discussed. Parametric tests can be extremely useful. However, participants were not provided with the attribution of numbers to labels. To avoid the concerns of using parametric tests with ordinal data, we propose only performing a comparison of answers with no use of parametric methods.

Thirdly, we want to measure the proximity of answers and evaluate stakeholders' views differently than what is commonly done. Usually, when trying to understand stakeholders' views, studies evaluate the differences between pre-defined groups and then perform intra-group (within the same group) or inter-group (between different groups) measurements. In this study, we propose a different approach - evaluate results without pre-defining groups and find the clusters with similar views based on the responses. After communities are found, it is possible to analyse the distribution of health stakeholders within those clusters and perform intra and inter-group analysis.

With this in mind, using the framework we are about to present, we intend to unconventionally apply the powerful tools of Network Science, analysing health Delphi results innovatively and helpfully. For that, going back to the main research questions, we expect to (1) prove that NS tools are suitable for analysing Delphi results and (2) explore the added information which can be obtained regarding stakeholders' views. The second point includes investigating what characterises communities. If stakeholder groups do not define them, we can add value by transforming data into a network and, using CD tools, obtain new information we otherwise would not.

5.2 Proposed framework

We propose, in this work, the application of a CD algorithm in the hope of successfully clustering stakeholders based on their similarity of opinions. Also, we expect to find what characterises these clusters while also obtaining new and fresh information about the stakeholder network, using NS tools. This strategy will hopefully allow a finer information tool for decision-makers. Based on the guidelines from Silva and Zhao (2016), we propose a framework comprising five main steps:

1. Pre-processing of the vector-based dataset;
2. Measurement of proximity and similarity of answers;
3. Conversion of the vector-based dataset into network-based data;
4. Application of the Community Detection algorithm;
5. Visualisation and analysis of results in order to answer the research questions.

These five phases, briefly illustrated in figure 5.2, are described in more detail in the following sections.

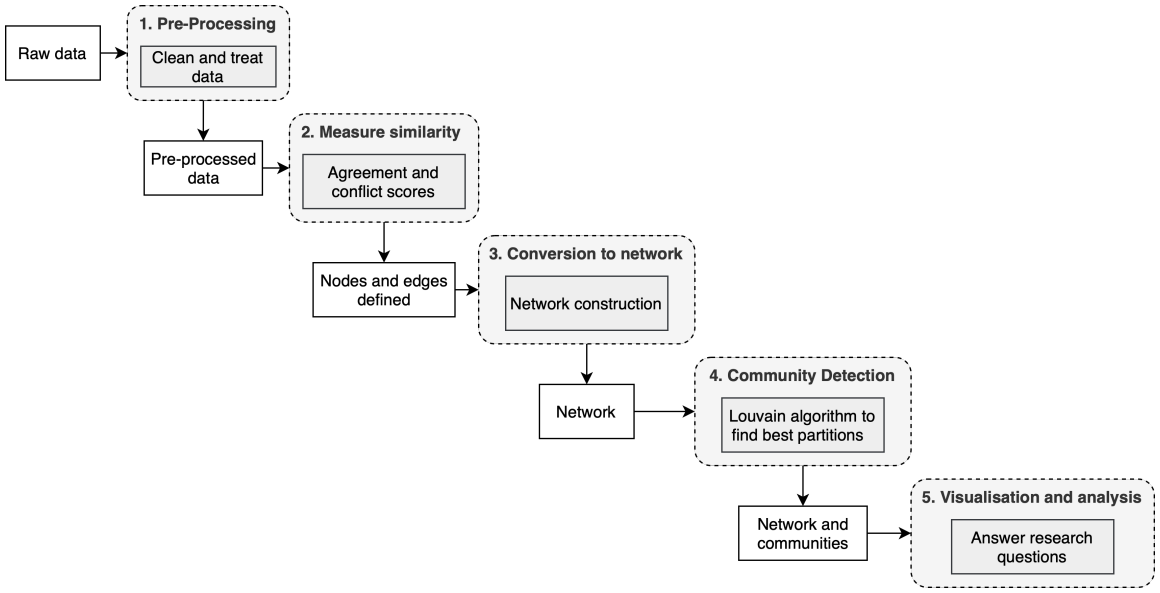


Figure 5.2: Framework diagram.

5.2.1 Pre-processing

As a starting point, we propose the application of pre-processing operations on the original data. This step is crucial and involves treating the original data according to the specific framework, including, as mentioned in Silva and Zhao, different transformations such as scaling, normalisation, or cleaning.

Pre-processing includes specific transformations related to the algorithm needs and raw data treatment concerning survey conditions and implementation. It is necessary to verify if results are consistent with the study design, i.e., that all results were saved correctly and that no information was lost, including questions or participants related data. Also, Delphi surveys' results are scale-originated data usually exported in Excel files. It is necessary to transform and organise data to be further manipulated according to the methods meant to be applied.

5.2.2 Measurement of proximity and similarity of answers

The measurement of the proximity of answers is needed to quantify the similarity of views between health stakeholders. We propose to define the proximity between a pair of stakeholders and later, when constructing the network, define the edge between their corresponding nodes based on this proximity.

To escape issues associated with ordinal data analysis, we propose an analysis that sets the relation between two stakeholders (a pair of stakeholders) in one of four scenarios:

1. "Same answer", when both stakeholders choose the same scale point for a given aspect;
2. "Same group", when the answers are not the same but belong to the same group;

3. ”Opposite group”, when the answers are not the same and belong to opposite groups;
4. ”Different group”, when neither answers belong to the same group or opposite groups.

Given this, we also propose a definition of what are the groups for each of the projects.

The MEDI-VALUE’s scale is not commonly described in the literature, and we found no guidelines on how to group items. Thus, we propose a new division. The most natural criteria we can use to divide the items into groups is if the item reflects the participant belief that a given aspect should or not be considered when evaluating a health technology. However, it also seems natural to assume that the answers ”Critical” and ”Fundamental” are closer than ”Fundamental” and ”Complementary”. For this reason, we propose the following groups:

1. ”Include with higher importance” group, which combines the ”Critical” and ”Fundamental” items;
2. ”Include with lower importance” group, corresponding to the ”Complementary” item;
3. ”Do not include” group, containing the ”Irrelevant” item;
4. ”No answer” group, covering the ”Don’t know/don’t want to answer” item.

For IMPACT-HTA, we propose a division based on the one commonly presented in other studies and used by Freitas et al.:

1. ”Agreement” group, which comprises the ”Strongly agree” and ”Agree” items;
2. ”Neutral” group, incorporating the”Neither agree nor disagree” item;
3. ”Disagreement” group, which joins the ”Strongly disagree” and ”Disagree” items;
4. ”No answer” group, carrying the ”Don’t know/don’t want to answer” item.

We propose to consider ”Include with higher importance” and ”Do not include” and ”Agreement” and ”Disagreement” as opposite groups.

Agreement and conflict scores

Based on the responses from a pair of stakeholders, an agreement score is then calculated. To define this score, an adaptation of the approach described by Leifeld and used in Buckton et al. (2019), and Fergie et al. (2019) is proposed. Leifeld suggests that it is possible only to consider an agreement network, i.e., not considering conflict or to compute both an agreement and a conflict network and subtract the conflict edge weights from the corresponding agreement edge weights. For the latter, the resulting weighted network has positive ties when there is more agreement than conflict and negative ties for the opposite case.

Buckton et al. (2019) and Fergie et al. (2019) compute both the agreement and conflict network. The similarity between two stakeholders a and b is defined by the number of concepts both have in common, and the conflict network is calculated by counting the number of concepts where both stakeholders present opposing agreement patterns.

This method needs, however, to be adapted for the context of this work. We propose considering two variables for each stakeholder pair: an agreement variable A and a conflict variable C. On the one hand, for each aspect, if the relationship is considered an agreement, A adds up 1. On the other hand, if the relationship is considered a conflict, C adds 1. In the end, it is possible only to consider the variable A or subtract C from A, as performed by Buckton et al. (2019) and Fergie et al. (2019).

Regarding agreement and conflict, they are defined based on the group division previously presented. For a pair of stakeholders, we propose to consider an agreement (add 1 to variable A) when the scenario is "Same answer" or "Same group" and to consider a conflict (add 1 to variable C) when facing the "Opposite group" one. In the case of "Different group", no change is made.

Sensitivity analysis

Considering that no similar frameworks for Delphi surveys' results were found in the literature, no particular approach is known to be more appropriate. To cover more than one option and analyse their influence on the results, we propose a sensitivity analysis for some scenarios and parameters - the calculation of the final similarity, what is considered agreement or conflict, and the threshold for the edge definition. The third item is related to the network formation, and it is discussed in section 5.2.3. The first and second items are now explored, varying for both projects.

For both MEDI-VALUE and IMPACT-HTA, we propose to explore the influence of considering only agreement (variable A) and then agreement and conflict (subtracting C from A).

For IMPACT-HTA, we propose an additional experiment - first to consider both "Neutral" and "No answer" as separate groups and then as one group only. This way, for a given aspect, if a stakeholder a answers "Neither agree nor disagree" and b "Don't know/don't want to answer", their agreement variable A adds 1. The foundation for this experiment is that both these groups represent people with no strong opinion, and it might be interesting to understand how different these options are.

Finally, we propose applying this approach for single aspects (aspect-level), groups of aspects and all aspects simultaneously. For one item, the score for two stakeholders can only be -1 ("Opposite group"), 0 ("Different group), or 1 ("Same answer" or "Same group"). Extrapolating to a group of aspects (or all aspects), we propose summing these scores and normalising them, dividing by the number of aspects considered. A more detailed explanation of how the similarity is calculated is presented in appendix B.

5.2.3 Conversion into a network-based dataset

In order to apply a CD algorithm, it is first necessary to transform the vector-based data into network-based data, i.e., transform each stakeholder ID into a node and define their interaction, the edges. In other words, we propose to represent stakeholders as nodes and for edges to represent their agreement.

As mentioned by Silva and Zhao, in general terms, given a non-networked formatted set of N data points $X = \{x_1, ..., x_n\}$, it is possible to transform it into a network G, with the vertex set $V = \{v_1, ..., v_V\}$ and the edge set E, a subset of VxV. For that purpose, a mapping procedure is needed:

$$g : X \to G = \langle V, E \rangle \tag{5.1}$$

Usually, as for this work, $X = V$, i.e., each original set's data item corresponds to a vertex in the network. For N data items in X and V=$|V|$ vertices, since no data reduction is conducted, V=N. Finally, to obtain the set of edges E, a similarity function and a network formation technique are used.

Similarity function

According to Silva and Zhao, a similarity function $s : V$ x $V \mapsto \mathbb{R}$ allows the quantification of how similar two data items are, based on their attributes, transforming a pair of data items into a scalar value . When applied to all pairs of vertices, this function allows constructing the similarity matrix S. Note that:

$$S_{ij} = s(v_i, v_j), v_i, v_j \in V, \tag{5.2}$$

We propose to use the previously described agreement score for the similarity function.

Network formation technique

With the similarity method chosen, we can decide when to add an edge between v_i and v_j. When the agreement score between two stakeholders is higher than a given threshold value, an edge is defined.

For normalisation, aspects' scores are summed and the result divided by the number of aspects being considered, meaning that the score is always equal or smaller than 1. We propose a sensitivity analysis to evaluate different thresholds for the similarity value. Buckton et al. and Fergie et al. used a value of 0.4. However, we want to explore higher values. This way, we can obtain different connections between stakeholders and investigate the influence of the threshold value. Considering the normalised scores, we propose to perform tests for the following set of thresholds - 0.4, 0.5, 0.6, 0.7, and 0.8.

5.2.4 Community Detection algorithm

Recalling the vast available options of CD algorithms, we will now introduce some of the most used ones and walk the reader through the choice of the proposed algorithm - the Louvain algorithm.

As discussed by Smith et al. (2020), the use of CD algorithms can be potent and helpful, including in many public health research areas. However, their application can be complicated due to the choice of the method to use. Several reviews already present guidance for choosing a method based on com-

putation time, community structure or other technical aspects. For example, specific methods are more suitable if the study's goals include finding overlapping communities or representing directed networks. Other researchers including (Lee et al., 2020) compare algorithm performance using known networks.

In 2020, Smith et al. presented "A Guide for Choosing Community Detection Algorithms in Social Network Studies". The authors approached the selection differently, focusing on a topic they consider the most relevant - how the communities will be used and with what purpose. The guide presents a review of 6 Community Detection methods: Walktrap, Edge-Betweenness, Infomap, Louvain, Label Propagation, and Spinglass, divided in divisive, agglomerative, and optimisation-based algorithms.

The six algorithms support edge weights and undirected networks, the ones we are interested in, and are among the most chosen ones. Considering not only these findings from Smith et al. but also other authors comparing algorithms, including Lee et al.; Mkhitaryan et al.; Yang et al., we choose the Louvain algorithm to incorporate our framework. However, before delving into it, we describe the other possibilities and why they were not so promising.

Divisive algorithms

Divisive algorithms start with a complete network and divide it into smaller communities. The best-known method is the Girvan-Newman edge betweenness CD algorithm, which iteratively removes the edges with a high likelihood of linking different clusters (Smith et al., 2020). It is particularly useful for interrupting transmission within a network, for example, in biology contexts for identifying risk connections in a sexual transmission disease network. However, this application does not match the goal of this work.

Agglomerative algorithms

Agglomerative algorithms start by considering each node as a community and then iteratively combine them into larger clusters. We present two relevant methods - Walktrap and Label Propagation.

According to Smith et al., the Walktrap is based on the assumption that nodes within communities are likely to be connected by shorter random walks. This approach is also helpful for the transmission of information. Briefly, for the Label Propagation method, each node is given a label, and randomly selected nodes adopt the label of the majority of neighbours. It is a good option for modelling the adoption of social norms. Both methods most appropriate use do not match this works' goal.

Optimization-based algorithms

Optimisation-based algorithms work in the hope of finding the optimal solution for a specified objective function. There are three main algorithms - Infomap, Spinglass and Louvain.

Once again, Infomap is preferred for optimising the flow of information, especially when a clearly defined stopping rule is needed (Smith et al., 2020), being out of the scope of this work.

According to Mkhitaryan et al. (2019) and Smith et al. (2020), the Spinglass is an approach from statistical mechanics, where the total number of spin states represents the number of communities. The algorithm identifies which edges are in the same spin state, optimising a function where intra-community edges are rewarded and between communities are penalised. One important caveat is that it cannot be used in unconnected graphs, i.e., not complete graphs, which is the case of our datasets. In the case of a loose network, it is possible to decompose the network into connected components and cluster those parts. However, it is a clear limitation.

Smith et al. states that the Louvain algorithm merges nodes into communities in order to maximise modularity. It stops when no merging results in a modularity increase. We will next examine this method. First, we want to inform the reader that both Louvain and Spinglass have similar applications, being better suited when Community Detection results are meant to be used later, as part of a broader analysis, which matches the goal of this work.

The only algorithms whose most suitable purpose matches this works' goals are Spinglass and Louvain. However, among other advantages, the Louvain algorithm is easier to implement, not having the unconnected network constraint, justifying our choice. Several authors mention the Louvain algorithm as powerful and, many times, as the best approach, for CD. In 2016, Yang et al. performed a comparison of the accuracy and computing time of eight CD algorithms, Edge Betweenness, Fast Greedy, Infomap, Label Propagation, Leading Eigenvector, Louvain, Spinglass and Walktrap. Louvain outperformed all of them. In 2019, Mkhitaryan et al. analysed several CD algorithms and found the Louvain algorithm to be "consistently the best across both the measures and the networks tested", including eight real world and a variety of synthetic networks. In 2020, Lee et al. also compared network clustering algorithms performance, and the results once again revealed Louvain as "the best performance in terms of modularity and processing time". Thus, we propose the use of the Louvain algorithm in our framework.

Louvain Algorithm, the chosen approach to incorporate our framework

The Louvain algorithm was first described in 2008 in "Fast unfolding of communities in large networks" (Blondel et al., 2008). The primary goal of this method is to maximise modularity, a measure commonly used to evaluate the quality of the division of a network into communities, i.e., the quality of a partition.

Modularity ranges from -1 to 1, measuring the density of intra-community edges when compared to inter-community edges (Blondel et al., 2008; Smith et al., 2020). If e_{ij} is the fraction of edges connecting nodes from community i to nodes in community j, and $\alpha_i = \sum_j e_{ij}$, the modularity Q is defined as (Mukherjee et al., 2013):

$$Q = \sum_i (e_{ii} - \alpha_i^2) \tag{5.3}$$

It can also be interpreted as the difference between the fraction of intra-community edges and the expected value in a network with the same community division, but with random edges (Mukherjee et al., 2013). In this case, if the number of intra-community edges is less than the expected number of edges for a random graph, $Q = 0$. The maximum value is $Q = 1$, and values close to 1 indicate a strong community structure (Mukherjee et al., 2013). In real networks, it is rare to obtain high values, with typical ones ranging from 0.3 to 0.7.

When the network is a single community, the two terms are equal and cancel, thus $Q = 0$. In contrast, when each node constitutes a single community, the modularity is negative (Mukherjee et al., 2013).

Note that, for a given network divided into communities, being c_i the community assigned to node i, the fraction of intra-community edges is defined as (Mukherjee et al., 2013):

$$\frac{\sum_{i,j} W_{ij}\delta(c_i, c_j)}{\sum_{i,j} W_{ij}} = \frac{1}{2m}\sum_{i,j} W_{ij}\delta(c_i, c_j) \tag{5.4}$$

where W_{ij} represents the weight of the edge between i and j, the function $\delta(u, v)$ is 1 if $u = v$ and 0 otherwise and m is the total number of edges. Also, $m = \frac{1}{2}\sum_{i,j} W_{ij}$ (Mukherjee et al., 2013).

Additionally, the expected number of edges between nodes i and j, for randomly drawn edges, is $\frac{k_i k_j}{2m}$, where $k_i = \sum_j W_{ij}$ is the weighted degree of the node i, meaning the sum of the weights of the edges linked to node i (Mukherjee et al., 2013). Thus, for weighted networks, modularity Q is defined as (Blondel et al., 2008):

$$Q = \frac{1}{2m}\sum_{i,j}(W_{ij} - \frac{k_i k_j}{2m})\delta(c_i, c_j) \tag{5.5}$$

The Louvain algorithm method is divided into two phases, repeated iteratively until no modularity increase is possible (Blondel et al., 2008), as shown in figure 5.3.

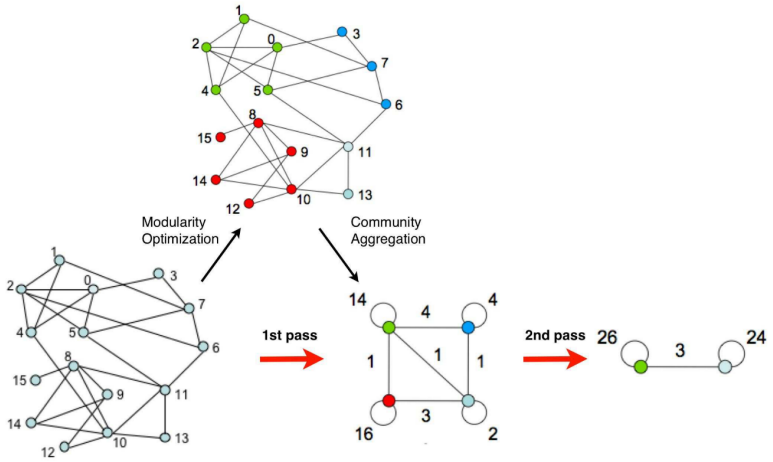


Figure 5.3: Louvain algorithm method. There are two phases, repeated iteratively, until no modularity increase is possible. First, there is the optimisation of modularity by allowing only local changes of communities, and then there is the aggregation of communities to build a network of communities (Blondel et al., 2008).

First, for a weighted network of N nodes, a distinct community is assigned to each node, meaning that

initially, there are N communities. Then, for each node i, the modularity gain of removing the node i from its community and assigning it the community of its neighbour j is calculated. The node i is allocated to the community for which this (positive) gain is maximum. In case of a tie, a breaking rule is used and, if no positive gain is possible, i remains in its original community. This phase is repeated for all nodes, knowing that a node can be analysed several times, until local maxima of modularity are achieved (Blondel et al., 2008).

With the first phase completed, it is possible to build a new network whose nodes are the communities. For this purpose, the weight of an edge between two new nodes is obtained by summing the weights of the edges between nodes in the corresponding two communities. After this second phase, it is possible to repeat the first phase. In figure 5.3, a ”pass” corresponds to both phases (Blondel et al., 2008).

Besides the already stated advantages, according to Blondel et al., the Louvain algorithm is intuitive and of easy implementation, with a low computational time, and linear complexity on typical and sparse data.

It is incredibly relevant to state that, according to Blondel et al., the output of the algorithm depends on the order in which the nodes are evaluated, i.e., the algorithm is stochastic. Although it appears that this order does not significantly influence the modularity, it can influence the computation time. It is, however, possible to choose and fix the order of node evaluation, which also facilitates the analysis.

5.2.5 Visualisation and analysis of the results

After the construction of the network and application of the CD algorithm, we now describe the proposed visualisation and analysis of results.

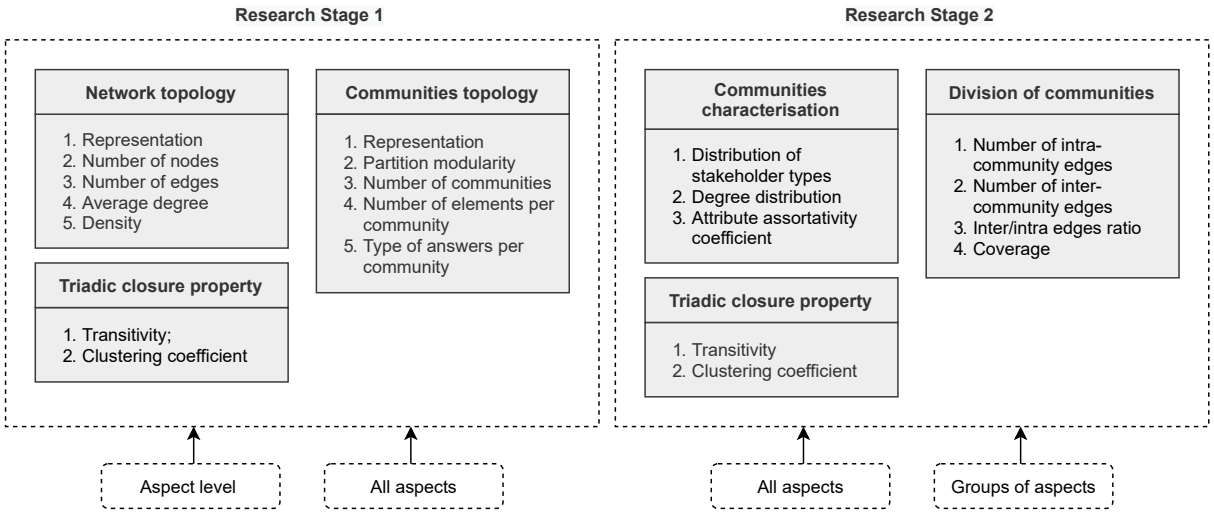


Figure 5.4: Diagram describing the procedure for analysing results.

Visualisation and analysis focus on the research questions described in section 1.2, each answered in one research stage. In the first stage, to answer the question ”Are NS tools suitable for analysing Delphi

results?", the results are analysed to understand how Delphi data is converted into network-based data and to find if there is a pattern of results for different datasets. This way, it is possible to verify the appropriateness of the framework. At this point, results from individual aspects and all aspects together are used. In the second stage, to answer the question "What information regarding stakeholders' views can be obtained?", we analyse the networks and communities in a more detailed way, for groups of aspects and all aspects together. The tools used throughout the analysis are shown in figure 5.4.

In research stage 1, we can compare results from the three datasets. Since there is no previous report of using network analysis for Delphi, it is crucial to investigate if there is a typical structure of these networks. At this stage, we propose to use measures such as average degree and density. As mentioned by Silva and Zhao, in the context of social representations, the degree is the total number of people with whom someone is connected. In this work, degree corresponds to the number of stakeholders an individual agrees with, with their agreement score overcoming the threshold. We now define density.

Definition 7. ***Density:*** *According to NetworkX and Baeldung, the density of a graph is 0 when there are zero edges and 1 for a complete graph. The density d of an undirected graph G, with n nodes and m edges, is given by:*

$$Density = d = \frac{Number\ of\ Edges}{Maximum\ Number\ of\ Edges} = \frac{m}{\frac{n(n-1)}{2}} = \frac{2m}{n(n-1)} \tag{5.6}$$

The expression corresponding to the maximum number of edges is divided by 2 since, otherwise, it would be the maximum number of edges for a directed graph, where two nodes can establish two connections.

In the context of these networks, a density of 1 means all stakeholders agree with each other. The lower the density, the fewer stakeholders are connected, meaning less overall agreement. We believe it is crucial to ensure that the obtained networks present similar topologies when this framework is applied to different datasets from similar contexts.

We also propose to investigate how measures, such as transitivity and clustering coefficient, are calculated using simple examples.

In the second research stage, to extract new information on stakeholders' views and relationships, we propose to answer three primary questions:

1. "What characterises the obtained communities?"
2. "Is there a clear division between communities?"
3. "Is the triadic closure property, from social networks, also verified in "agreement networks"?

"What characterises the obtained communities?"

In this study, we want to justify the promising application of NS and CD in the context of Delphi surveys, starting with no pre-defined groups, as well as to understand data better. We would expect stakeholders

from the same type to behave similarly and communities to be defined by these types. However, if that is not the case, we can infer an added value from not using these groups and rather understanding what clusters emerge. Furthermore, if communities do not match the stakeholder groups, we want to understand what defines these clusters. For that purpose, we propose four principal analyses:

1. Calculate the degree distribution per stakeholder type. According to Mukherjee et al. (2013) this is "the most fundamental characteristic of a network". In this context, degree distribution represents the variation in the number of stakeholders a given agrees with across the network. If we can verify that there is a various degree distribution, per stakeholder topology, we can state that there is a diversity of opinions within each type, i.e., similar stakeholders agree with different individuals, meaning that their type does not determine their opinions;
2. Explore the assortativity (homophily in social networks). According to Cinelli et al. (2020), it is the tendency for nodes presenting similar attributes to be connected. For example, in online social networks, connected individuals are usually from close locations and have similar ages (Mukherjee et al., 2013). There can also be a negative assortativity (heterophily) (Pelechrinis and Wei, 2016), which is the tendency for linked nodes to have different properties. Briefly, the assortativity coefficient is based on the comparison of the number of edges linking nodes of a similar type and the expected number of those connections in the case of a random model (Pelechrinis and Wei, 2016). It ranges between -1 and 1 (Cinelli et al., 2020), where a value closer to 1 corresponds to strong homophily, a value close to 0 means that there is no relevant association between the edges and the considered property, and a value closer to -1 represents heterophily (Shi, 2019). We propose the evaluation of assortativity regarding the stakeholder type to investigate if there is a tendency for similar stakeholders to agree with each other;
3. Measure the heterogeneity of communities, regarding stakeholders' type. This means both measuring the composition of each community and the distribution of each type per community. If communities have diversity in stakeholder types, and stakeholders are spread across different communities, this suggests that stakeholders' types do not dictate their opinions;
4. Measure the distribution of answers per community. If communities are indeed not defined by stakeholders' type, we want to understand the properties characterising them. The distribution of answers is also proposed to verify if the type of answers given defines them.

"Is there a clear division between communities?"

Many inter-community edges, meaning several stakeholders agreeing with others from different communities, can be associated with higher susceptibility for opinion changes. We propose to analyse intra and inter-community edges and the external ratio. According to Buckton et al. (2019), the external ratio is the number of inter-community edges as a proportion of total edges. This measure can indicate the agreement of stakeholders of different communities and how likely they are of changing opinion, which is pivotal for decision-making processes.

”Is the triadic closure property, from social networks, also verified in ”agreement networks”?

Social Network theory shows that if node A is connected to B and if B is connected to C, it is highly likely that A and C are also connected. This is called the triadic closure principle, which can be translated in ”*If two people in a social network have a friend in common, then there is an increased likelihood that they will become friends themselves at some point in the future.*” (Aalabaf-Sabaghi, 2012). According to Mukherjee et al. (2013), a high level of transitivity or clustering coefficient, associated with the triadic closure, is possibly ”the most important feature distinguishing social networks from other types of networks”. We want to investigate if this also happens in these ”agreement networks”. Can we state that if A agrees with B and B agrees with C, A and C are more likely to agree with each other eventually? We propose the use of transitivity and clustering coefficient to answer this question. This way, we can learn more about HTA stakeholders' networks and compare them to typical social networks.

5.3 Conclusion

In this chapter, we introduced the proposed framework. In section 5.1 we reviewed the main concerns and research questions. In section 5.2 we described the framework, going through its main five steps. Table 5.1 summarises the reasoning behind the major decisions proposed.

Table 5.1: Key points from the proposed framework and the reasoning behind them.

Proposed decision or path	Reasoning
Not to attribute numbers to labels	Avoid the concerns of analysing ordinal data as interval and the use of parametric tests.
Community Detection	It is promising in similar studies on other contexts and allows the identification of which groups emerge without using the predefined types.
Louvain algorithm	Its applications fit our needs. It is possible to use it with undirected weighted graphs, the computational time is not high, and its implementation is straightforward. It is possible to make it non-stochastic. Its shows better performance results when compared to other Community Detection algorithms.
Sensitivity analysis	Allows us to compare different scenarios and test variables when we do not know which are the most appropriate ones.

Chapter 6

Framework implementation

This chapter presents the implementation of the proposed framework and the followed research steps.

6.1 Dataset

Datasets used during this work correspond to the data presented in chapter 3. Since we are not exploring the change of opinion between rounds, only R2 results, meaning the final results were used.

In MEDI-VALUE, raw Web-Delphi data is separated into two Excel files, one for each round. Each document has two sheets, one for IMD and the other for BBIVT. The first line of each file regards the 34 aspects, and each following line presents the answers and comments of each participant, identified by an ID. Additionally, the Excel from R2 has a third sheet associating a stakeholder type to each ID.

In IMPACT-HTA, raw data from the Web-Delphi is divided into 12 Excel files, one for each round, for each stakeholder group. Each document has three sheets, one for the results, a second one for statistics and the third one for comments. The first line of the first sheet corresponds to the 24 aspects, and each other line corresponds to each participant's answers, identified by an ID.

One final note is related to the already described fact that MEDI-VALUE, in reality, consisted of two Delphi surveys, one for IMD and another for BBIVT. For this reason, we used, in practice, three datasets.

6.2 Implementation environment

The implementation of this work was performed using Excel and Python. Excel was used for the pre-processing stage since the raw data was provided as *.xlsx* files. All code developed for this work was

implemented using Python. The code was implemented from scratch, using available useful libraries.

The choice of Python comprehends two main reasons. First, it is one of the more used programming languages, making it an updated and complete tool. Thus, its implementation is simpler due to the large community of users and the several available resources, including literature and related projects. Secondly, Python also presents several relevant libraries and packages. Considering the specific scope of the work, Python offers powerful packages which facilitate the construction and analysis of CNs and implementation of CD algorithms. The *NetworkX* library allows the study of graphs and networks, and the *Community API* library implements the Louvain CD algorithm and allows further analysis.

The code was run on a regular CPU (Intel Dual-Core i5 @ 2.7 GHz), and relevant extracts are publicly available in de Faria (2021).[1]

6.3 Pre-processing

The first pre-processing stage was the same for both projects and consisted in assuring files corresponded to the expected. We verified that no data corresponding to questions or participants was missing.

The second pre-processing stage consisted of creating Excel files with R2 items' responses and deleting the lines corresponding to stakeholders who did not complete both rounds. We only considered individuals who completed R2. Also, we added a column with each participant type.

Finally, we created files with only some of the answers, corresponding to the groups of items presented in the appendix A and for the individual aspects. These files were converted to *.csv* files.

6.4 Measurement of proximity and similarity of answers

The measurement of proximity and the definition of edges were implemented in Python and are available in the file *write_edges.py*. The implementation was adapted for each project and sensitivity analysis. Still, the reasoning was always the same. The implementation for IMPACT-HTA and MEDI-VALUE for the considering of agreement and conflict, and separate groups for IMPACT-HTA, are presented in de Faria (2021). Taking advantage of *cycles*, each pre-processed *.csv* file is open, the data is retrieved, and a variable containing the list of all stakeholders is defined. We implemented the function *evaluate_group (answer1,answer2)* that receives the answers from a pair of stakeholders and classifies it into one of the proposed groups - "Same answer", "Same group", "Opposite group" or "Different group".

[1] MEDI-VALUE and IMPACT-HTA datasets are currently not public. Thus we did not add them to the repository. For this reason, it is currently not possible to replicate the work and obtain the results since the original data is needed.

Again with a *cycle*, the relationship between every pair of stakeholders is evaluated, with the proposed thresholds. If the final score is higher than the threshold, a line containing the ID of both stakeholders and their score is added to a new file of the name *Edges_[aspects]_[threshold].csv*. If lower, no line is added, meaning that no edge will exist between those stakeholders in the network.

6.5 Conversion into a network-based dataset

The conversion of data into a network-based dataset and the construction of networks was implemented using the *NetworkX* library. The same logic of automatic generation was used in *netowork_and_CD.py* files, where *cycles* were used to generate networks and analysis on individual, groups and all aspects.

The original pre-processed files were opened and read to generate a network to provide the stakeholders' information, meaning the nodes. Also, the attribute "stakeholder_group" was added to each node using a *NetworkX* function - *NetworkX.set_node_attributes*. For the edges, the files generated in section 6.4 were used and the edges were added using the function *NetworkX.add_weighted _edges_from*, using the agreement scores as weights.

6.6 Community Detection algorithm

With the network constructed, we used the *Community API* library to implement the Louvain Community Detection algorithm. This package allows to find the best network's partition, optimising modulation, using the method described in "Fast unfolding of communities in large networks" (Blondel et al., 2008; Networkx, 2021). We specified the parameter "randomize=False" to assure that the implementation of the algorithm was not stochastic. Otherwise, we would get different results each time the code ran.

6.7 Network visualisation and analysis

6.7.1 Visualisation

The visualisation and analysis of the complete network and detected communities were performed simultaneously, but we will describe them separately.

The visualisation of the network and communities was implemented using one of the most used *Python* libraries - *matplotlib.pyplot*. In the case of the networks, nodes from different stakeholder groups were represented in different colours. For the communities, each community's node had the same colour. The figures were saved in *.pdf* format in order not to lose detail and quality.

6.7.2 Analysis

After the network was constructed and the communities were detected, results' visualisation and further analysis were possible. At this stage, we took advantage of available analysis tools from *NetworkX* and *Community API* libraries. When needed, other functions were implemented.

The information obtained followed the proposed framework. This means that we analysed the networks' topology and extracted the information needed to answer the two research questions.

All results were saved in *.txt* files.

Network Topology

For the analysis of the networks' topology we used:

1. The *NetworkX.info (G)* function which returns a summary of information for the given graph G. This summary includes the number of nodes, the number of edges and the average degree;
2. The *NetworkX.density(G)* function which returns the density of the provided graph G.

General Measures

General measures were used to gain insights into the communities. Information, namely the number of communities found and their size, types of answers in each community, and partitions' modularity, were obtained. For this phase, functions were implemented, and already existing ones from the *Community API* library were used, including the *CommunityAPI.modularity(partition, G))* function. In this modularity function, the weights are considered by default.

"What characterises the obtained communities?"

To first verify what defines the obtained communities, we used several measures, according to the proposed framework presented in chapter 5. Part of the analysis was performed on the original network to understand the connections between stakeholders. The evaluation of the heterogeneity of communities used the obtained clusters. Thus, among others, the primary measures used are:

1. The *G.degree()* function allowed to calculate degree distribution, for all stakeholders and per group. Histograms were used to represent results.
2. For the exploration of the assortativity, we used the function *NetworkX.attribute_assortativity_coefficient (G,'stakeholder_group')*. It returns the assortativity of the graph G for a given attribute, in this case regarding the stakeholder type group.

3. After the application of the Louvain Community Detection algorithm, we implemented a function to measure the heterogeneity of communities by obtaining information on the percentage of each type of stakeholder for each community.
4. Finally, we implemented a program to visualise multipartite networks, using the *Bipartite* library from *NetworkX*. This library allows the implementation of a bipartite graph. We used the same reasoning of affiliation networks, mentioned in Aalabaf-Sabaghi (2012), where individuals are affiliated with groups or activities, to represent the distribution of answers per community and health stakeholders groups. This particular implementation is shown in folder *Multipartite Networks* in de Faria (2021). This tool helps to visualise the possible match of communities and groups and analyse the distribution of answers per cluster and, therefore, verify a possible characterisation of communities. This step was performed manually, adapted for each group and all aspects.

"Is there a clear division between communities?"

A function comparing the number of intra and inter-community edges was implemented to understand if there is a clear division between communities. This function returns the number of intra and inter-community edges for each community, the ratio of inter/intra edges for each community, and each community's coverage. The coverage of a community is the ratio of the number of intra-community edges of that cluster to the total number of the graph's edges. Finally, the function returns the total number of intra and inter-community edges.

"Is the triadic closure property, from social networks, also verified in "agreement networks"?

Finally, to test the concept of triadic closure present in social networks, meaning the measure of the tendency for edges in a graph to form triangles (Necromuralist, 2021), we used two functions. The *NetworkX.transitivity(G)* function, which returns the transitivity of the graph G and, for the network's average (local) clustering coefficient, we used the *NetworkX.average_clustering(G)* function. We now provide more detailed definitions of transitivity and (local) clustering coefficient and how *NetworkX* functions calculate them.

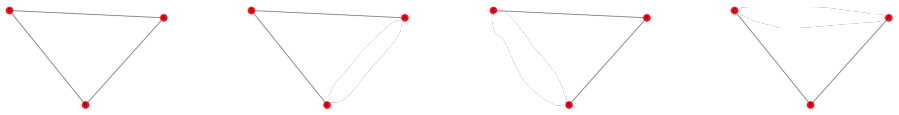
Definition 8. ***Transitivity:*** *(or global clustering coefficient) measures the fraction of triads in a network, which are triangles. A triad is defined as three nodes with only two edges between them. A triangle is defined as a set of three nodes connected by three edges. Each triangle has three open triads in it, as shown in figure 6.1 (Necromuralist, 2021). The transitivity T, of a graph G is expressed as follows:*

$$T = \frac{3 \times number\ of\ triangles\ in\ the\ network}{number\ of\ open\ triads\ in\ the\ network} \tag{6.1}$$

Note that factor 3 is needed because each triangle contributes with three "triads" (Necromuralist, 2021).

Transitivity ranges from 0, and 1 (di Udine, 2021), where 1 corresponds to a graph where all open

triangles are triangles. In figure 6.2 this concept is better illustrated. In this example from di Udine, there are five triads: one centred at x, one centred at y and three centred at z. Three of them are part of triangles. The other two, the ones centred on z containing w, are not. Thus, the transitivity T for the given graph is $T = \frac{3}{5} = 0.6$. If edges were connecting y and w and z and z, transitivity would be 1.



(a) Triangle ABC (b) Open triad 1 (c) Open triad 2 (d) Open triad 3

Figure 6.1: Triangles and open triads. Example of a triangle ABC and its three open triads (Necromuralist, 2021).

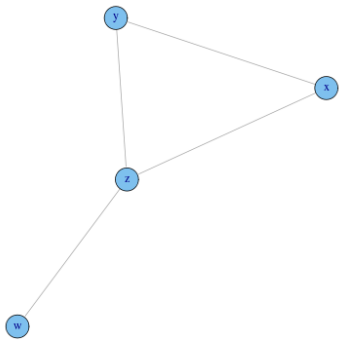


Figure 6.2: Network with triangles and open triads, for transitivity calculation (di Udine, 2021).

Definition 9. ***Clustering coefficient*** **(C):** *(or local clustering coefficient) is a measure for a single node. It measures how close the neighbors of a node are to being a complete graph (Necromuralist, 2021). It is defined as:*

$$C = \frac{Number\ of\ pairs\ of\ a\ node's\ friends\ that\ are\ friends\ (PTAF)}{Number\ of\ pairs\ of\ the\ node's\ friends\ (POF)} \tag{6.2}$$

with the number of pairs of friends (POF) being, with for a node's degree k (Necromuralist, 2021):

$$POF = \frac{k(k-1)}{2} \tag{6.3}$$

The clustering coefficient ranges from 0 to 1. A node with no pairs of friends is given a value of 0 (Necromuralist, 2021).

The average clustering coefficient of all nodes is used to calculate the clustering coefficient for a complete network (Necromuralist, 2021).

Definition 10. ***Average clustering coefficient*** **(C):** *is the average of the clustering coefficients of all nodes of the graph. It is expressed, for a graph G with n nodes, where c_v is the clustering coefficient of node v, as (NetworkX, 2021):*

$$C = \frac{1}{n} \sum_{v \in G} c_v \tag{6.4}$$

6.8 Conclusion

In this chapter, we guided the reader throughout the implementation of the framework proposed in chapter 5. First, we described the dataset in section 6.1 and the implementation environment in section 6.2, justifying the choice of Python for the implementation. Then, from section 6.3 to 6.7 we described how the framework was implemented, including a brief description of the code.

With the implementation described, in the next chapter, we will present and discuss the obtained results.

Chapter 7

Results presentation and discussion

At this stage, we were able to explore the areas of HTA, Delphi, NS and CD and how these can relate. Additionally, we proposed a framework for the analysis of Delphi results using NS and CD and presented a suggestion of implementation. This chapter presents a compilation of the results obtained from the simulations performed. First, section 7.1 contains an explanation of the Delphi data conversion to networks and how they should be interpreted, justifying the suitability of this framework, corresponding to the results from research stage 1. In section 7.2 the choice of parameters resulting from the sensitivity analysis is presented. Section 7.3 presents the results regarding research stage 2, for groups of aspects and 7.4 for all aspects.

7.1 Research stage 1 - Interpretation and suitability of the method

7.1.1 Aspect-level

To understand the transformation of Delphi results into network-type data, we start by presenting some examples of how data is represented in this approach, using an aspect level analysis. Since the results are similar, there is no added value in presenting them all. For the sake of simplicity, we now present some results from six IMPACT-HTA aspects. Datasets from MEDI-VALUE resulted in similar conclusions. The aspects being presented are:

1. *Aspect 4 - "Medicine's impact on mortality"*. Here, all stakeholders answered either "Strongly agree" or "Agree". Thus, the relationship between stakeholders was either "Same answer" or "Same group", meaning that all scores were 1. Speaking in "network terms", this means that all stakeholders were connected;
2. *Aspect 5 - "Medicine's impact on morbidity"*. Concerning this aspect, all stakeholders showed to either "Strongly agree" or "Agree" to include it in the evaluation of new medicines, except one

stakeholder, in particular a Scientific Expert, which answered, ”Neither agree nor disagree”. For this reason, the network obtained is similar to the previous scenario, but with an isolated node;

3. *Aspect 6 - ”Medicine’s impact on health-related quality of life”*. In this case, all stakeholders showed to ”Strongly agree” or ”Agree” to include the aspect in the evaluation of new medicines, except two stakeholders, one HTA professional and one payer, who answered, ”Neither agree nor disagree”. For this reason, the network obtained is similar to the previous scenario, but with an isolated pair;
4. *Aspect 1 - ”Severity of the disease”*. In this case, answers from three distinct groups were given but none from the ”No answer group”. For this reason, the results do not vary with the difference of analysing” Neutral group” and ”No answer” groups separately or as one;
5. *Aspect 11 - ”Medicine’s mechanism of action”*. For this aspect, more diverse answers were obtained. Again, no answers from the ”No answer” group were given;
6. *Aspect 2 - ”Unmet need of the disease”*. Here, there were answers from all categories. The detected communities vary depending on how we analyse these ”Neutral” and ”No answer” groups.

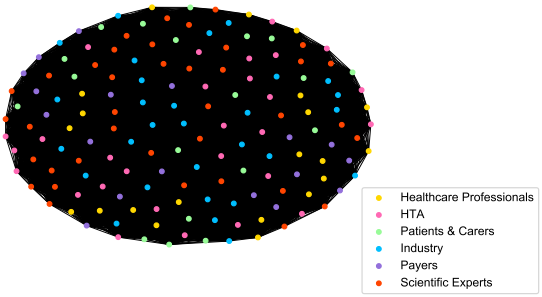
Note that a 0 threshold value was used when analysing only one aspect since scores are -1, 0 or 1, meaning that edges weights are always 1. The results do not vary considering only agreement (A) or agreement and conflict ($A-C$). For agreement only, when two stakeholders give answers from opposite groups, their score is zero. For both agreement and conflict, even though the” raw” score is -1, after applying the threshold value of 0, their score becomes 0. Furthermore, the difference in the analysis of considering ”Neutral” and ”No answer” groups as separate or one group is only noticed when there are answers from both these groups. If there are no answers from one (or both) of the groups, no changes occur.

Network Topology

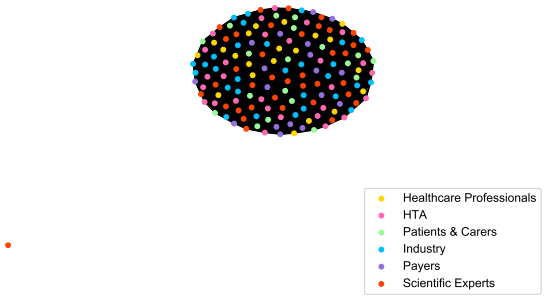
In this subsection, we discuss the topology of the networks regarding these aspects. Figure 7.1 presents the networks’ representation. The representations match the descriptions just made for each aspect. For aspect 4, we observe a full network, where all stakeholders agree with each other. In the aspect 5 network, there is an isolated orange node corresponding to the scientific expert. For aspect 6, we can visualise a pair of stakeholders, one HTA professional and one payer, who got separated from the rest of the participants. For aspects 1 and 11, three groups are observed, the first with less diversity in the distribution of views. Finally, in aspect 2, we begin to understand what happens when answers from the four groups are given, with two isolated stakeholders in 7.1 (f) and none in (g).

Information and measures concerning the topology of the networks are presented in table 7.1. Note that the number of nodes is always 153 since there are always 153 stakeholders. In fact, this number is only presented as a ”control” variable to verify if any errors occurred during simulations.

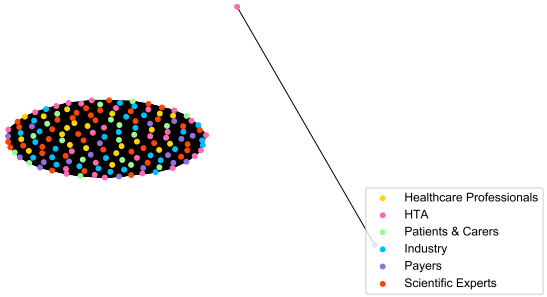
Regarding the number of edges and density, it is important to recall equation 5.6. The maximum number of edges is $\frac{N(N-1)}{2} = \frac{153(152)}{2} = 11628$.



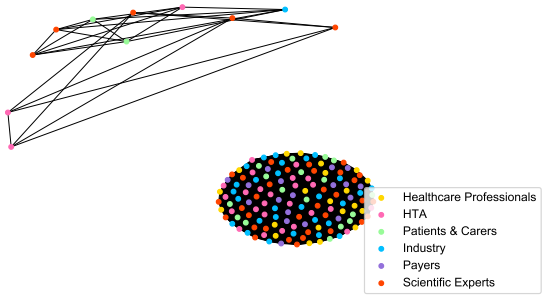
(a) Aspect 4.



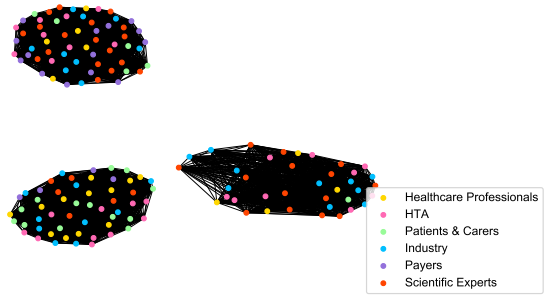
(b) Aspect 5.



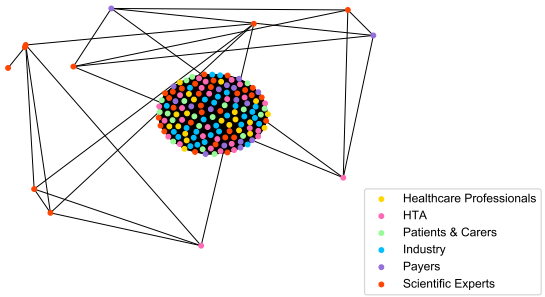
(c) Aspect 6.



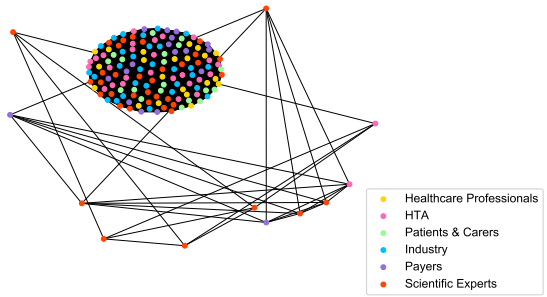
(d) Aspect 1.



(e) Aspect 11.



(f) Aspect 2, with ”neutral” and ”no answer” groups considered separately.



(g) Aspect 2, with ”neutral” and ”no answer” groups considered as one.

Figure 7.1: Network representation regarding IMPACT-HTA’s aspects 4, 5, 6, 1, 11 and 2.

Table 7.1: Network's topology analysis regarding IMPACT-HTA's aspects 4, 5, 6, 1, 11 and 2.

Aspect	# Nodes	# Edges	Avg Deg	Density
Aspect 4	153	11628	152.000 0	1.000 0
Aspect 5	153	11476	150.013 1	0.986 9
Aspect 6	153	11326	148.052 3	0.974 0
Aspect 1	153	10036	131.189 5	0.863 1
Aspect 11	153	3958	51.738 6	0.340 4
Aspect 2 (Different groups) [1]	153	9891	129.294 1	0.850 6
Aspect 2 (One group) [2]	153	9901	129.424 8	0.851 5

[1] "Neutral" and "no answer" groups as separate groups.

[2] "Neutral" and "no answer" groups as one group.

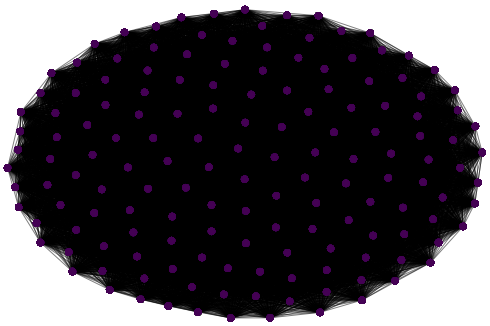
For aspect 5, the edges concerning the isolated stakeholder (152) need to be subtracted from the maximum, resulting in 11476 edges. Moreover, since 152 are connected to the other 151 and one is connected to no one, the average degree is lower since $Average\ degree = \frac{152 \times 151 + 1 \times 0}{153} = 150.013$. Finally, density is also lower - $density = \frac{11476}{11628} = 0.987$, since less stakeholders are connected to each other, when compared to the maximum number of possible connections.

When it comes to aspect 6, from the maximum number of edges (11628), the edges concerning the two isolated stakeholders with the remaining ($151 \times 2 = 302$) need to be subtracted, resulting in 11326 edges. Since 151 are connected to the other 150 and two are connected to one (degree is one), the average degree is even lower ($Average\ degree = \frac{151 \times 150 + 2 \times 1}{153} = 148.052$). Again, the density is lower: $density = \frac{11326}{11628} = 0.974$.

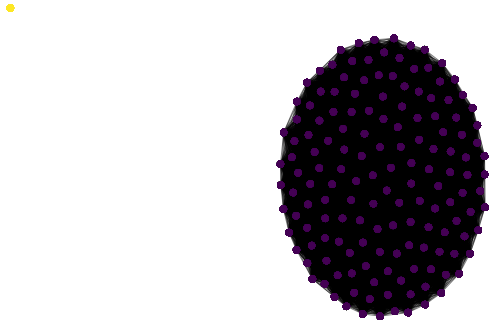
For the remaining aspects, the reasoning is the same. The more diverse the stakeholders' opinions, the fewer connections there are, and therefore, the lower the number of edges, the average degree, and the density. One crucial detail is that the logic is always the same concerning the analysis of the "Neutral" and "No answer" groups. When these groups are considered as one, individuals who give "Neither agree nor disagree" and "Don't know/don't want to answer" answers are considered to agree, and an edge is established. That does not happen when separate groups are admitted. For this reason, when only one group is recognised, the number of edges, the average degree, and density are slightly higher.

Communities topology

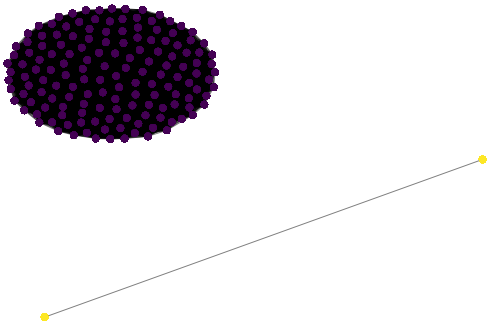
The representations of the detected communities for each considered aspect are presented in figure 7.2. Different nodes' colours represent different communities. More detailed information on the communities and distribution of answers can be found in table 7.2.



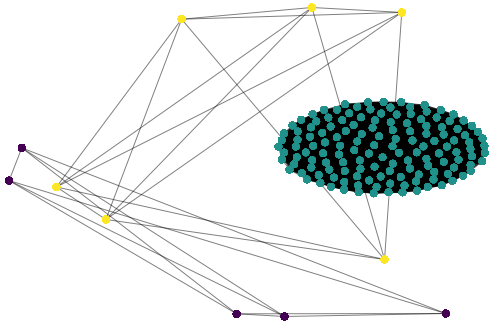
(a) Aspect 4.



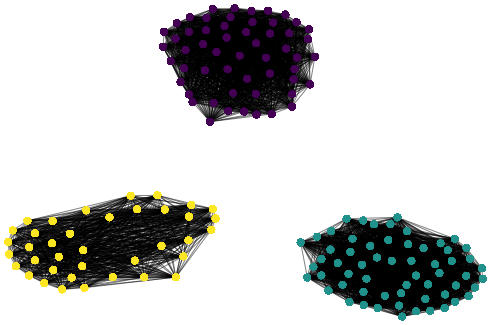
(b) Aspect 5.



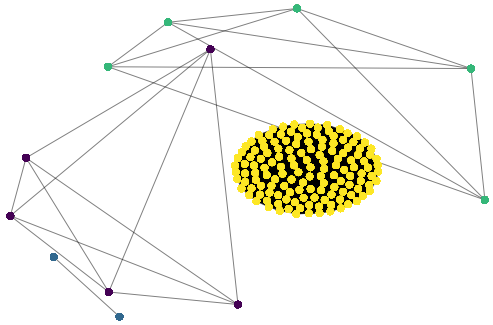
(c) Aspect 6.



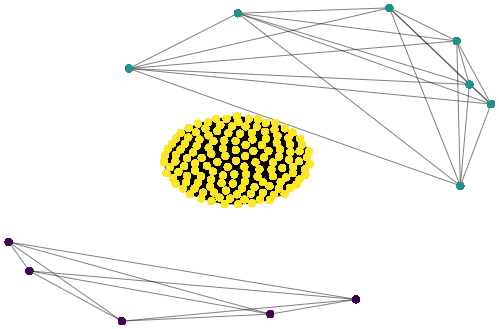
(d) Aspect 1.



(e) Aspect 11.



(f) Aspect 2, with ”neutral” and ”no answer” groups considered separately.



(g) Aspect 2, with ”neutral” and ”no answer” groups considered as one.

Figure 7.2: Communities’ representation regarding IMPACT-HTA’s aspects 4, 5, 6, 1, 11 and 2. Different nodes’ colours represent different communities.

For aspect 4, since all stakeholders agree with each other, only one community is detected, with its nodes in purple. Concerning the distribution of answers per group, per community, shown in table 7.2, since all stakeholders answered" Strongly agree" or" Agree", 100% of the answers in the community are from the agreement group. For aspects 5 and 6, the algorithm finds two communities. The first one, represented in purple, corresponds to the stakeholders who gave "Agree" answers. The second, represented in yellow, contains the isolated nodes from the "Neutral" group, as indicated in table 7.2.

Regarding aspects 1 and 11, three communities (purple, yellow, and blue) were found since no answers from the "No answer" group were given. However, for aspect 1, most stakeholders belong to the same community, the "Agree" group. Contrarily, for aspect 11, the distribution per community is much more balanced among communities, with the three groups presenting similar numbers of members. This means that opinions were more diverse for the aspect of "Medicine's mechanism of action", and stakeholders were more divided. Interestingly, the majority of answers (C1), was from the "Neutral" group, supporting the indecision regarding this aspect, as shown in table 7.2.

Table 7.2: Topology analysis of communities and distribution of answers, per group, per community regarding IMPACT-HTA's aspects 4, 5, 6, 1, 11 and 2.

Aspect	# Communities	Part. Modularity	Community	# Elements	% A	% D	% Ne	% NA
Aspect 4	1	0.000 0	C1	153	100	0	0	0
Aspect 5	2	0.000 0	C1	152	100	0	0	0
			C2	1	0	0	100	0
Aspect 6	2	0.000 2	C1	151	100	0	0	0
			C2	2	0	0	100	0
Aspect 1	3	0.005 0	C1	142	100	0	0	0
			C2	6	0	0	100	0
			C3	5	0	100	0	0
Aspect 11	3	0.627 7	C1	60	0	0	100	0
			C2	55	100	0	0	0
			C3	38	0	100	0	0
Aspect 2 (Different groups) [1]	4	0.004 2	C1	141	100	0	0	0
			C2	5	0	0	100	0
			C3	5	0	100	0	0
			C4	2	0	0	0	100
Aspect 2 (One group) [2]	3	0.006 2	C1	141	100	0	0	
			C2	7	0	0	100	
			C3	5	0	100	0	

[1] "Neutral" and "no answer" groups as separate groups.

[2] "Neutral" and "no answer" groups as one group.

[3] The columns correspond to the percentage of answers belonging to the "Agree" group (A), "Disagree" group (D), "Neutral" group (Ne) and "No answer" group (NA).

Finally, for item 2, three or four communities can be obtained, depending on how "Neutral" and "No answer" groups are evaluated. For separate groups, one community for each group is obtained, meaning

four communities. If ”Neutral” and ”No answer” groups are considered as one, individuals answering as those groups are combined into one community, meaning that there are only 3.

About modularity, we can observe that the obtained values are all close to zero, except for aspect 11. Regarding aspects 4 and 5, as explained in chapter 5, when there is a single community, the two parcels of equation 5.3 cancel each other, and its value is zero. A value of 0 and the other values close to 0 represent the lack of community structure of the network, where stakeholders tend not to group in defined and robust communities.

In contrast, for aspect 11, we observe that communities are well defined and have a balanced number of stakeholders. The higher modularity of 0.6277, an excellent result for a real network, represents this strong community structure.

Hence, in the context of Delphi networks, a lower value of modularity does not mean a poor Community Detection algorithm performance. If all stakeholders agree, they should all be in the same community, and therefore modularity is 0. That does not mean that the division does not represent reality. It just means that all stakeholders gave similar answers, and there is a poor community structure. In summary, in the context of this work, modularity reflects the variety of answers and how stakeholders tend to group and is not a good measure for partition quality. Therefore, it should not be used to compare performance results, including in the sensitivity analysis.

Network and communities measures

One of the properties we are interested in exploring in this type of network is the triadic closure property from social networks. For that purpose, transitivity and clustering coefficient were measured. We will now understand how those measures can be interpreted in this context, where edges represent agreement. Results are presented in table 7.3.

Table 7.3: Transitivity and average clustering coefficient for IMPACT-HTA's aspects 4, 5, 6, 1, 11 and 2.

Aspect	Transitivity	Average Clustering. Coefficient
Aspect 4	1.000 0	1.000 0
Aspect 5	1.000 0	0.993 5
Aspect 6	1.000 0	0.986 9
Aspect 1	1.000 0	1.000 0
Aspect 11	1.000 0	1.000 0
Aspect 2 (Different groups) [1]	1.000 0	0.986 9
Aspect 2 (One group) [2]	1.000 0	1.000 0

[1] ”Neutral” and ”no answer” groups as separate groups.

[2] ”Neutral” and ”no answer” groups as one group.

It is important to state that for the analysis of only one item, the transitivity is always 1. This happens

because, for all communities with three or more individuals, all stakeholders share the same views and thus will be connected, meaning that all triads are triangles. There are no triads for communities with one or two stakeholders, so they do not enter the equation.

For the clustering coefficient, regarding aspect 4, since all stakeholders are connected, all the pairs of neighbours are themselves, neighbours, justifying the value of 1. As stated in chapter 6, an isolated node has a clustering coefficient of zero. This means that, for aspect 5, the average clustering coefficient is given by $\frac{(152\times1)+(1\times0)}{153} = 0.9935$ and for aspect 6 $\frac{(151\times1)+(2\times0)}{153} = 0.9869$. In fact, the clustering coefficient in the case of aspect-level analysis is different from 1 only if there is a community with less than three elements, as it happens for aspects 5, 6, and 2 (when considering the four groups).

This property is likely verified for the aspect level. Thus, this analysis is not much insightful. Since all stakeholders with similar views are in the same community and all linked, this property is always valid unless the community has less than three individuals. Thus, here, we only want to understand how these measures are calculated, and later we will explore it when considering more aspects.

In general, we were able to understand how network-based data from Delphi can be interpreted and, so far, the aspect-level results are in concordance with what was expected and support the appropriateness of NS and CD in the context of HTA Delphi. We will now present results considering all aspects. Since we have data from three surveys, we can explore the suitability of this framework by assuring that the general topology of networks is preserved for different datasets.

7.1.2 All aspects

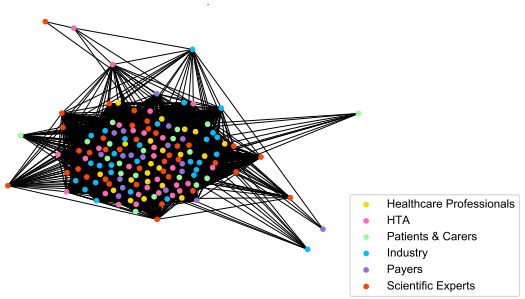
Now that the representation of Delphi results using networks has been introduced and explained, we are in an excellent position to explore results regarding complete networks for the three datasets - IMPACT-HTA and MEDI-VALUE IMD and BBIVT. For the full network, the scores are not -1, 0 or 1 anymore. As described before, the scores for each item are summed and divided by the number of items. Thus, different threshold values are considered. Additionally, now, the results differ if we consider agreement and conflict or only agreement.

Network topology

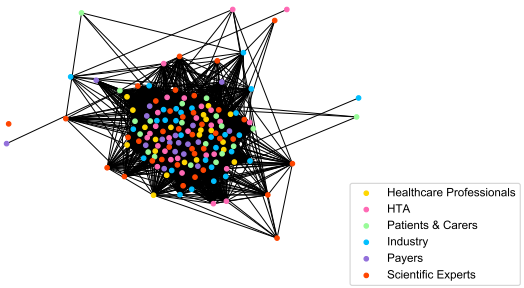
When analysing full networks' visualisation, it is not as easy to understand the relations between stakeholders, as before, at an aspect level. Starting with IMPACT-HTA, in the case where "Neutral" and "No answer" groups are considered as separate, the topology analysis in table 7.4. The network representation, when considering agreement and disagreement, is presented in figure 7.3. The results are similar when only agreement is considered, but with more edges being formed.

Table 7.4: Topology analysis for the IMPACT-HTA network, with neutral and non-answer groups as separate groups.

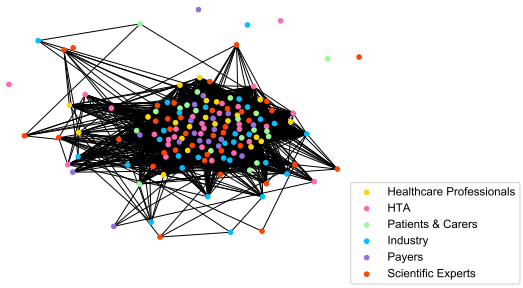
Threshold value	Agreement and conflict				Only agreement			
	# Nodes	# Edges	Avg Deg	Density	# Nodes	# Edges	Avg Deg	Density
0.4	153	9161	119.752	0.788	153	11102	145.124	0.955
0.5	153	7758	101.412	0.667	153	10177	133.033	0.875
0.6	153	6047	79.046	0.520	153	8459	110.575	0.727
0.7	153	2934	38.353	0.252	153	4342	56.758	0.373
0.8	153	1336	17.464	0.115	153	1789	23.386	0.154



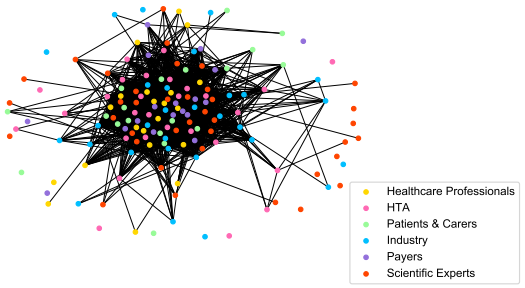
(a) Threshold = 0.4.



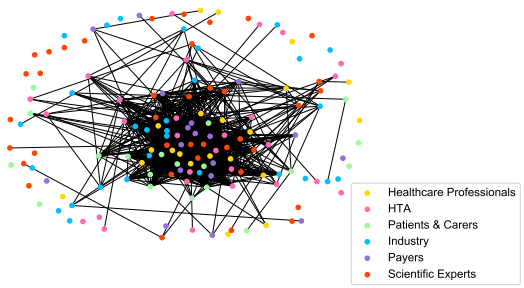
(b) Threshold = 0.5.



(c) Threshold = 0.6.



(d) Threshold = 0.7.



(e) Threshold = 0.8.

Figure 7.3: Full representation of the IMPACT-HTA network, when the ”Neutral” and ”No answer” are considered as separate groups, considering both agreement and conflict, for the five threshold values.

What is possible to observe is that, as expected, the higher the threshold value, the fewer edges are established. In other words, the higher we define the minimum value for agreement, the hardest it is to achieve agreement and to connect two stakeholders. For this reason, the higher the value, the lower the number of edges, the average degree, and the density. In figure 7.3 we can observe that, for higher values, more individuals are isolated, not connected to any other individuals.

Another important aspect is related to the consideration of agreement and conflict or only agreement. When only agreement is considered, stakeholders giving opposite answers in certain aspects are not penalised. For that reason, only the agreement is summed, and the final score is higher. That means that it is easier to reach and stay above the threshold. Once again, the higher the threshold, the lower the number of edges, the average degree, and the density. However, these values are higher when compared to the case where conflict is subtracted from the agreement.

This logic also applies to the results considering ”Neutral” and ”No answer” groups as one. However, when these groups are considered one, a stakeholder answering ”Neither agree nor disagree” agrees with a stakeholder saying ”Don’t know/don’t want to answer”. Consequently, there are more scenarios where individuals agree with each other, and therefore the scores are higher. Thus, the results for IMPACT-HTA, when considering these groups as one, reflect the previously rational. Once again, higher threshold values lead to a lower number of edges, average degree, and density. The logic for agreement and conflict consideration repeats, and, as expected, the overall values are slightly higher than the previous ones, where ”Neutral” and ”No answer” groups were considered separately.

Results for both MEDI-VALUE’s datasets reflect the same rationale, as shown for the IMD network in table 7.5 and for the BBIVT network in table 7.6 . Keep in mind that for MEDI-VALUE, no sensitivity analysis for ”Neutral” and ”No answer” group was performed, since there is no ”Neutral” group.

Table 7.5: Topology analysis regarding the Implantable Medical Devices network, from MEDI-VALUE.

Threshold value	Agreement and conflict				Only agreement			
	# Nodes	# Edges	Avg Deg	Density	# Nodes	# Edges	Avg Deg	Density
0.4	134	7334	109.463	0.823	134	7660	114.328	0.860
0.5	134	6068	90.567	0.681	134	6793	101.388	0.762
0.6	134	4523	67.508	0.508	134	5144	76.776	0.577
0.7	134	2930	43.731	0.329	134	3245	48.433	0.364
0.8	134	1081	16.134	0.121	134	1163	17.358	0.131

Table 7.6: Topology analysis regarding the Biomarkers-based *in vitro* Tests network, from MEDI-VALUE.

Threshold value	Agreement and conflict				Only agreement			
	# Nodes	# Edges	Avg Deg	Density	# Nodes	# Edges	Avg Deg	Density
0.4	134	6017	89.806	0.675	134	6785	101.269	0.761
0.5	134	4997	74.582	0.561	134	5717	85.328	0.642
0.6	134	3891	58.075	0.437	134	4491	67.030	0.504
0.7	134	2780	41.493	0.312	134	3098	46.239	0.348
0.8	134	1316	19.642	0.148	134	1430	21.343	0.160

By analysing results from the three datasets, we understand two things. First, the results vary across datasets. Second, these variations are minor.

On the one hand, paying close attention to tables 7.4, 7.5 and 7.6, it is possible to understand that there

are small changes in the values obtained. In general, the results are higher for the IMPACT-HTA network and lower for MEDI-VALUE. Additionally, for both MEDI-VALUE datasets, the results also vary. Not only the results are slightly different for different questionnaire setups (IMPACT-HTA and MEDI-VALUE), but the same people answering the same survey for different health technologies gave different answers. This finding can either mean that the context of the project influences the results or that people did not answer the surveys carefully, easily changing responses.

On the other hand, it is possible to understand that, even though some variations occur, the results regarding topology are much similar. For the three datasets, where a similar number of stakeholders took part, the number of edges, the average degree, and density are, in general, similar and vary in the same way, regarding the variations of threshold values and considering conflict or not. This discovery matters because it shows that, for similar setups but different participants in the surveys, different questions, and different scales, the topology of networks is similar, supporting the appropriate use of NS in this context. Otherwise, we could be dealing with arbitrary results.

Communities topology

We now explore the topology of communities. Results regarding IMPACT-HTA are presented in table 7.7, Implantable Medical Devices in table 7.8 and Biomarkers-based *in vitro* Tests in table 7.9.

Table 7.7: Communities' topology for IMPACT-HTA, with neutral and non-answer groups as separate groups.

Threshold value	Agreement and conflict		Only agreement	
	# Communities	Partition Modularity	# Communities	Partition Modularity
0.4	3	0.046 3	3	0.021 4
0.5	4	0.057 5	3	0.028 1
0.6	10	0.071 1	3	0.039 9
0.7	25	0.129 1	14	0.091 7
0.8	48	0.165 2	31	0.159 6

Table 7.8: Communities' topology regarding the IMD network, from MEDI-VALUE.

Threshold value	Agreement and conflict		Only agreement	
	# Communities	Partition Modularity	# Communities	Partition Modularity
0.4	3	0.047 1	3	0.036 5
0.5	3	0.067 4	3	0.049 7
0.6	6	0.092 3	5	0.071 2
0.7	14	0.116 6	10	0.105 3
0.8	40	0.180 6	39	0.152 4

The results obtained are in harmony with the ones from the networks' topology. The higher the threshold value used, the higher the number of communities found. Again, this is justified because, for higher threshold values, a pair of stakeholders need to concordance in more items (and disagree in fewer items if the conflict is considered) for a connection to be made. Moreover, when only agreement is considered,

it is easier to establish edges since disagreement is not penalised. For this reason, fewer communities are found when compared to the case where the conflict is penalized.

Table 7.9: Topology analysis regarding the BBIVT network, from MEDI-VALUE.

Threshold value	Agreement and conflict		Only agreement	
	# Communities	Partition Modularity	# Communities	Partition Modularity
0.4	3	0.077 8	3	0.064 5
0.5	8	0.092 1	3	0.079 8
0.6	10	0.115 3	7	0.103 5
0.7	17	0.131 5	13	0.131 9
0.8	46	0.223 5	43	0.203 4

As mentioned before, for the IMPACT-HTA project, when the ”Neutral” and ”No answer” groups are analysed as one, it is also easier to achieve a higher agreement score. For higher threshold values, few isolated stakeholders were obtained, and thus slightly fewer communities were found.

One notable result is that the number of communities for a threshold value of 0.4 is always the same - 3. This outcome suggests that there are three main communities and that others emerge for higher threshold values when more constraints for agreement are applied. The higher the threshold value, the more different the results are. For high threshold values, the results are volatile.

Regarding modularity, as mentioned before, it does not necessarily provide a quality measure but instead informs us on how strong communities are. For that reason, we can observe the modularity increasing with the increase of threshold values since more communities are being formed. However, the modularity values are relatively close to zero. It is not easy to achieve high modularity values for real networks. When analysing the quality of a partition, more exciting measures can be used, such as the number of communities found and the average degree.

In general, with some minor variations, different datasets from the same type lead to similar results. This finding is relevant since that it supports the hypothesis that there is a typical structure for ”agreement networks” and that NS and CD tools are appropriate for this context. After validating our approach, we are in a position where we can analyse the results aiming to obtain fresh information about stakeholders’ opinions and relationships.

7.2 Choice of conditions

Before going any further, we discuss the sensitivity analysis results and choose further conditions.

Since there are no significant differences or benefits for considering ”Neutral” and ”No answer” groups as one, and no evidence of the similarity of these answers in the literature, further analysis will be made considering those groups as separate, maintaining their original differences.

Regarding the consideration of conflict or not, we observe that the results change. When conflict is subtracted from the agreement, the situation of stakeholders having opposed views is penalised. This strategy reduces the number of edges and, therefore, the average degree and density, reducing the chances for establishing agreement, but it also seems more reasonable. This approach was the one chosen by Buckton et al. (2019) and, since the overall topology is not changed, we believe that it is more appropriate to reward agreement and penalise disagreement/conflict.

The threshold value decision is not straightforward. (Buckton et al., 2019) proposed a threshold value of 0.4 to assure that only robust ties reflecting agreement would be considered. However, the context of this work and the type of data are not the same. In Buckton et al.'s study, even though many newspaper articles were analysed, the agreement was always regarding one main topic. In the end, the authors expected to find two communities - one for the sceptics and one for supporters. In this work, communities do not necessarily match one type of answer. In a given community, stakeholders can agree but find some aspects relevant and others irrelevant. What is relevant is if a pattern of agreement is shared with other individuals and whom. For that reason, more diversity is expected to be found.

In stage 2, we analyse results from groups of related aspects and all aspects. For both cases, it is observed that, for high threshold values, several stakeholders become isolated or take part in communities with few people because they cannot reach the threshold and agree with anybody. This can be observed in tables C.4, C.5 and C.6, for the groups of aspects and in tables 7.7, 7.8 and 7.9 for all aspects. However, for low threshold values, the average degree is too high (close to the total number of participants), meaning that people are too connected, and the graph is close to a fully connected network. Notice tables C.1, C.2 and C.3 for the groups of aspects and 7.4, 7.5 and 7.6 for all aspects. Thus, we want to find a balance between too many and too few communities. This balance is associated with a reasonable average degree number. We want it not to be too close to the total number of participants (when they are all connected) but also not too low (almost no one is connected).

We consider that a threshold of 0.6 provides a reasonable average degree number, not too low nor too high, for all groups of aspects and all aspects and a reasonable number of communities.

7.3 Research stage 2 - Information extracted for groups of aspects

We now explore which additional information concerning stakeholders' views can be extracted. Remember that at this stage, the questions to be answered are "What characterises the obtained communities?", "Is there a clear division between communities?" and "Is the triadic closure property also verified in this context?". We first consider the groups of related aspects, found in appendix A, and then all aspects.

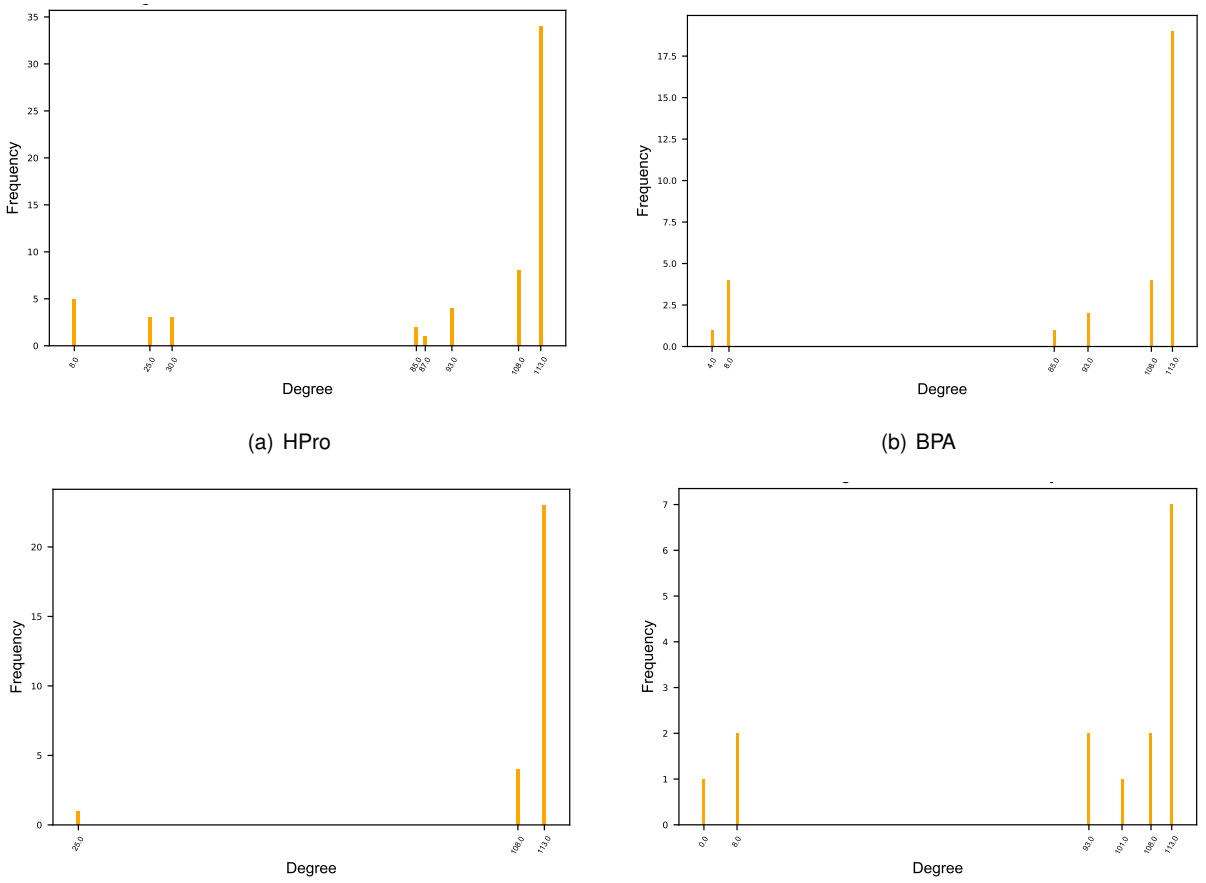
When applying CD tools, we are interested in strong communities. Although the algorithm finds them, we are not concerned about clusters with few individuals. For further analysis, we only consider the main communities with a considerable number of stakeholders.

For the sake of simplicity, because it would be too exhaustive to present results from the three datasets, the discussion will now focus on MEDI-VALUE's results. IMPACT-HTA analysis led to similar conclusions. We invite the reader to consult appendix C for the complete results regarding groups of aspects.

7.3.1 "What characterises the obtained communities?"

Initially, we want to understand what characterises communities. It is normally expected that similar stakeholders have similar opinions, so Delphi results are usually analysed based on groups of stakeholders' types. Thus, we want to investigate if the communities match these original groups. If not, our framework adds value to Delphi surveys' analysis since, contrarily to what is usually done, it starts with no pre-defined groups and instead investigates which clusters emerge. Furthermore, if the type of stakeholders does not characterise communities, we want to understand what characterises them.

We start by examining the degree distribution. In general, the results suggest that similar stakeholders do not necessarily share similar views. As an example, results from MEDI-VALUE, IMD, Group A are shown in figure 7.4. Similar distributions were observed for other groups of aspects.



(a) HPro (b) BPA

(c) Patients & Citizens (d) Industry

Figure 7.4: Degree distribution regarding IMD from MEDI-VALUE, for Group A of aspects - "Value for the patient".

Note that the edges' weight is not considered. Therefore, all weights are set to 1, with degree represent-

ing the number of a node's neighbours, i.e., other stakeholders sharing similar views.

Overall, histograms have several bars representing a heterogeneity degree within groups. This finding suggests that stakeholders from the same type are connected to different people, i.e., agree with people in different patterns. There are indeed some variations, but they are small and justified. For example, patients and citizens' histograms have fewer bars, meaning they have more similar views within their group. This outcome is probably related to the scope of the aspect - "Value for the patient". However, in general, there seems to be a diversity of degree distribution per stakeholder type for all aspects.

Considering the attribute assortativity coefficient, as shown in table C.7 in appendix C, for all groups, the values are close to zero. Hence, there is no higher likelihood for similar stakeholders, concerning their type, to be connected, meaning agree with each, than they would randomly. These findings support the degree distribution conclusions - opinions do not seem to be defined by stakeholder type.

Regarding the distribution of stakeholders per community, looking at tables C.8 and C.9 in appendix C, in general, stakeholders are widely distributed across communities. In addition, when a given type of stakeholder is concentrated in a community, it is usually because all stakeholders are and not because there is a specific match between stakeholder groups and the obtained communities. That would be the cause if, for instance, stakeholders from the same type were all in the same community.

Nevertheless, there are some differences for both groups of aspects and types of stakeholders. For instance, in MEDI-VALUE, for group F, most stakeholders are in community 1, and for group C, opinions are more diverse, and stakeholders are more diffused. For both cases, this distribution is independent of the type. However, the content being discussed in the aspects influences opinions. Aspects reflecting more straightforward topics generate less diversity of opinions, and thus stakeholders are, overall, placed in the same community. Aspects concerning more conflictive topics lead to more mixed results with a broader distribution of stakeholders per community.

There are also some differences between the two MEDI-VALUE datasets. For example, for aspect I, BBIVT industry stakeholders were more condensed in community 1 than in IMD, and healthcare professionals were more dispersed. These differences are not colossal, but they exist. We want to remind the reader of two main aspects which can be behind these variances. First, it is normal that the context of the survey influences results. As mentioned by Polisena et al. (2018), there are differences between drug therapies and medical devices impacting HTA. Therefore it is normal that differences exist for BBIVT and IMD. Second, although we count with a considerable number of participants in the survey, we are dealing with dozens of individuals who are not equally distributed between stakeholders groups. Thus, minor variations in opinions in one group corresponding to a few individuals might be perceived as greater.

We will now analyse groups of similar aspects in more detail, using tools that allow us to characterise each community based on the answers given. However, there seems to be true that there is no significant match between stakeholder types and communities. Indeed, occasionally similar people think more similarly, but that is not the general rule. Sometimes a type of stakeholder is condensed in one

community, but that is usually because all stakeholders are more located in that community, regardless of their type than stakeholders having opinions influenced by their type.

MEDI-VALUE, Group A - ”Value for the patient”

Figure 7.5 presents a multipartite network for the MEDI-VALUE's group A, ”Value for the patient”, showing the distribution of answers across the main communities and stakeholder groups. Figures 7.6 (a) and (b) present the composition of each community for both types of stakeholders and answers.

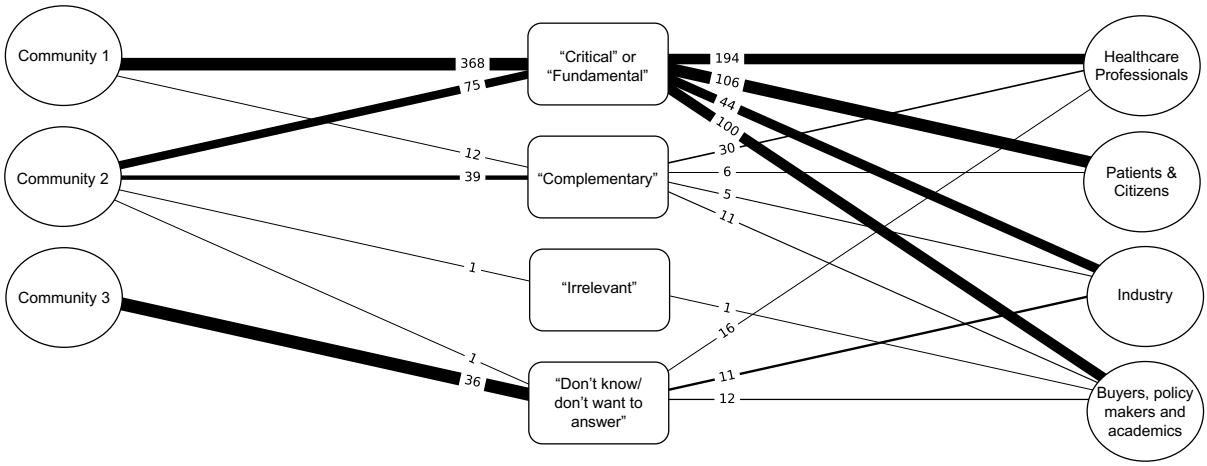
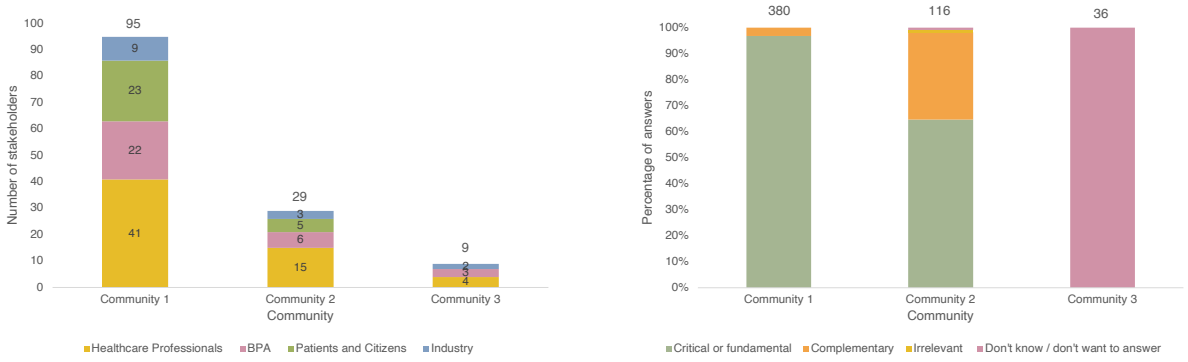


Figure 7.5: Multipartite network, regarding MEDI-VALUE's Group A - ”Value for the patient” (IMD), showing the distribution of answers across communities and stakeholder groups. The value in each line corresponds to the total number of answers, not stakeholders. Note that Group A comprises 4 aspects. The communities do not match stakeholders groups.



(a) Stakeholder types' distribution.

(b) Type of answers' distribution

Figure 7.6: Distribution of stakeholders' type and answers, per community, regarding MEDI-VALUE's Group A - ”Value for the patient” (IMD). In (a) numbers regard stakeholders and in (b) answers. Stakeholders are widely distributed across the three communities. Community 1 is characterised by ”Critical” or ”Fundamental” answers. Community 2 is more diverse but still characterised by ”Critical” or ”Fundamental”. Community 3 only has ”Don't know/don't want to answer” responses.

These visualisation tools allow the reader to verify that no match exists between the communities and stakeholder groups. That can be observed by comparing the lines in the left and right side in figure 7.5 or looking at figure 7.6 (a). The majority of stakeholders belong to community 1, meaning that similar

stakeholders gave similar answers, not because of their type but because the majority of all stakeholders gave similar answers (”Critical” or ”Fundamental”).

In figures 7.5 and 7.6 (b) we observe that community 1 has mainly individuals who answered ”Critical” or ”Fundamental”, and a few ”Complementary”. Community 2 is more diverse but still characterised by ”Critical” or ”Fundamental”, and community 3 only has ”Don’t know/don’t want to answer” responses.

In general, considering MEDI-VALUE’s Group A (IMD), communities are not characterised by stakeholders’ type. Instead, the three main communities that emerge can be divided into types of answers. Communities 1 and 2 are characterised by ”Critical” or ”Fundamental” answers, with 2 being more diverse, and community 3 by ”Don’t know/don’t want to answer ”. Remember that aspects from Group A are related to the value of the technology to the patient. Thus, it is interesting to notice that, as expected, the majority of patients and citizens answered ”Critical” or ”Fundamental”, a few ”Complementary” answers were given, and none ”Irrelevant” or ”Don’t know/don’t want to answer”.

These representations help us understand if stakeholders’ types influence opinions, but they can also help decision-makers. This new visual representation of stakeholders’ views allows to better visualise each stakeholder’s opinions and how groups of opinions are defined. For example, for this specific group of aspects, decision-makers can easily observe that, overall, most of the participants believe that the aspects should be included in IMD evaluation. Additionally, they can easily understand who is more likely not to answer (mainly healthcare professionals, industry and BPA) and who has a stronger opinion for the inclusion (patients and citizens). This way, by making these representations part of the report used for the discussion of Delphi results, the following steps may be easier to define and implement.

As previously discussed, IMD and BBIVT are extremely different, and the relevance of the aspects to consider for both is, sometimes, different. Since participants are the same and answered both surveys at close points in time, we now compare the results from Implantable Medical Devices with the ones from the Biomarkers-based *in vitro* Tests. Results regarding the latter are presented in figures C.1 and 7.8. This information can inform us on the relevance of the type of health technology.

The influence of the technology being evaluated is indeed significant, and results vary across IMD and BBIVT. For BBIVT, stakeholders were found to be more scattered across communities, and there were patients or citizens classifying aspects as ”Irrelevant” and ”Don’t know/don’t want to answer”. For this reason, as shown in figure 7.8 (b), a fourth community appears, characterised by the ”Irrelevant” answer. Again, communities are not characterised by the type of stakeholders. Similar individuals are scattered, and there is no match between groups and communities. Instead, as in IMD, communities are defined by the type of answers. However, these communities are different for both datasets.

These differences are likely due to the importance of the context of the questions. However, it can also be associated with people not having strong opinions and changing them from context to context or purely mistakes when answering the survey. The latter could explain why there were patients or citizens classifying aspects as ”Irrelevant” and ”Don’t know/don’t want to answer”, answers that were

less expected since this group is related to the value for the patient.

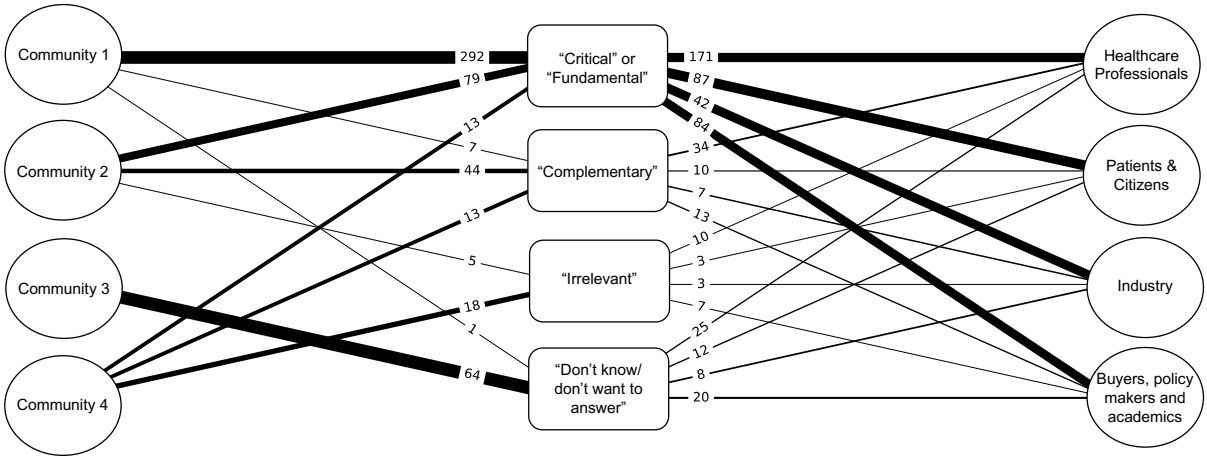
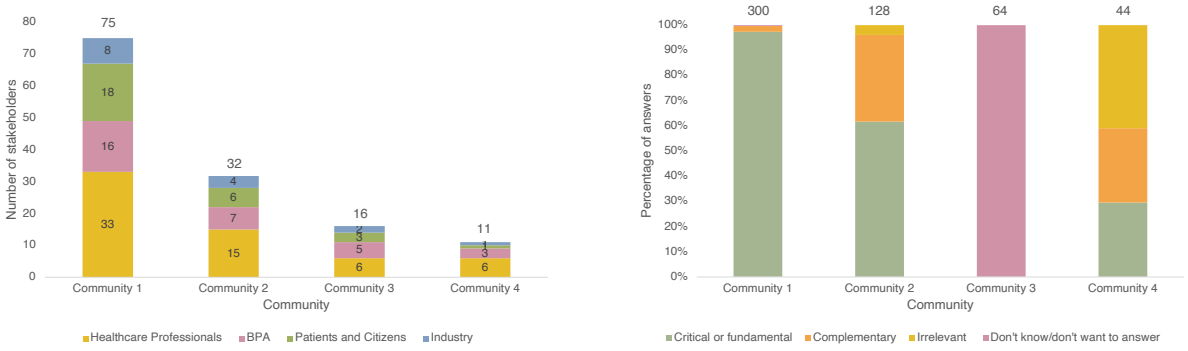


Figure 7.7: Multipartite network, regarding MEDI-VALUE's Group A - "Value for the patient" (BBIVT), showing the distribution of answers across communities and stakeholder groups. The value in each line corresponds to the total number of answers, not stakeholders. Note that Group A comprises 4 aspects. The communities do not match stakeholders groups.



(a) Stakeholder types' distribution.

(b) Type of answers' distribution

Figure 7.8: Distribution of stakeholders' type and answers, per community, regarding MEDI-VALUE's Group A - "Value for the patient" (BBIVT). In (a) numbers regard stakeholders and in (b) answers. Stakeholders are widely distributed across the four communities. Community 1 is characterised by "Critical" or "Fundamental" answers. Community 2 is more diverse but still characterised by "Critical" or "Fundamental". Community 3 is characterised by "Don't know/don't want to answer" answers and community 4, although more diversified, by "Irrelevant" responses.

Finally, once again, people tend to find the aspects as "Critical" or "Fundamental". Even though there were some variations, there is a tendency for stakeholders to find the aspects highly relevant. This can be observed by the width of the lines on the right side of the figure C.1.

MEDI-VALUE, Group H - "Societal context of the adoption of the medical device"

We now explore a less specific group related to the "Societal context of the adoption of the medical device". When the discussion concerns value for the patient, opinions are somewhat straight, and people express similar views even if with some differences. However, for less simple aspects, as Group H's, the opinions are more diverse. Notice figures 7.9 and 7.10. Diversity can be observed by the smaller difference between the lines' width in figure 7.9 and by communities' composition, in figure 7.10 (b).

Therefore, these representations can easily and quickly inform decision-makers on how dispersed and varied opinions are and how easy it can be to achieve a consensus on the considered aspects' inclusion.

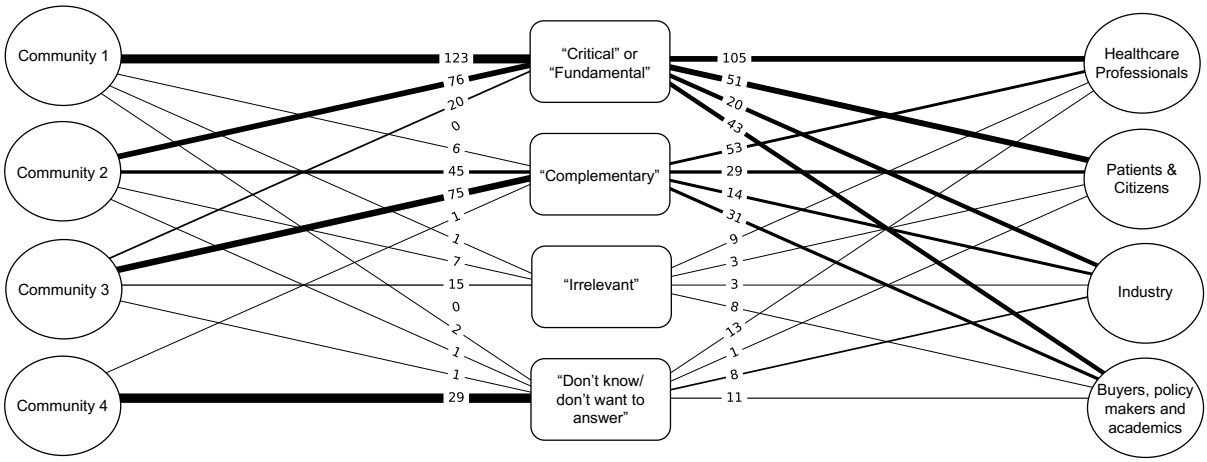
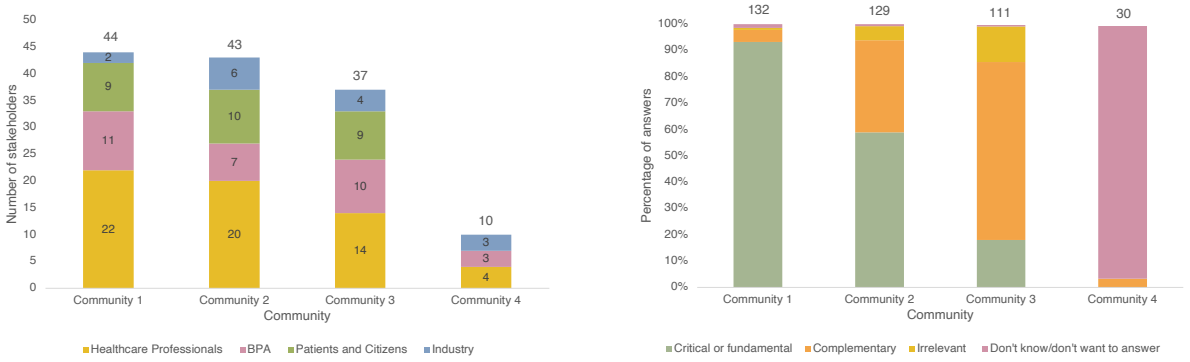


Figure 7.9: Multipartite network, regarding MEDI-VALUE's Group H - "Societal context of the adoption of the medical device" (IMD), showing the distribution of answers across communities and stakeholder groups. The value in each line corresponds to the total number of answers, not stakeholders. Note that Group H comprises 3 aspects. The communities do not match stakeholders groups.



(a) Stakeholder types' distribution.

(b) Type of answers' distribution

Figure 7.10: Distribution of stakeholders' type and answers, per community, regarding MEDI-VALUE's Group H - "Societal context of the adoption of the medical device" (IMD). In (a) numbers regard stakeholders and in (b) answers. Stakeholders are widely distributed across the four communities. Communities 1 and 2 are characterised by "Critical" or "Fundamental" answers, with the second presenting more diversity of answers. Community 3 is characterised by "Complementary" answers and community 4 by "Don't know/don't want to answer".

It is interesting to notice that communities are once again not defined by stakeholders' type. Additionally, communities can be considered characterised by the type of answers given, but with more diversity within clusters, indicating different opinions for related aspects. However, a new community is found for this group, characterised by the "Complementary" answer.

In conclusion, the results support the hypothesis that communities are not defined by stakeholders' type, suggesting that, although with some exceptions, when analysing Delphi results, decision-makers should address stakeholders not by type but as a whole.

Furthermore, opinions vary depending on the groups of aspects being evaluated. Some are more conflict-free and lead to less diversity of opinions, and others are more controversial. These findings are corroborated when looking at results related to the patient and healthcare professional's safety (Group

B) shown in figures C.1 and C.2 and the ones involving involving the aspects related to costs (Group D) displayed in figures C.3 and C.4. For a simple topic such as safety, low heterogeneity of opinions is observed. In contrast, probably because cost-related aspects' relevance is not so straightforward, more opinions are diverse. These differences can be easily recognised with these representations.

Communities seem to be characterised by types of answers, but depending on the groups. Thus, decision-makers should discuss aspects group by group. Finally, individuals tend to be located in one leading community, characterised by ”Critical” or ”Fundamental”, and the health technology being evaluated strongly influences results.

7.3.2 ”Is there a clear division between communities?”

We now discuss the division of communities. When we talk about a clear division of communities, we mean if stakeholders have explicit, distinct opinions or agree with stakeholders from other communities.

One major problem is that the answer heavily depends on the threshold value used and the number of evaluated aspects. First, the smaller the threshold value, the easier it is to consider an agreement and the higher the number of inter-community edges, meaning that communities are less defined. Second, when few aspects are considered, communities are more defined because there are fewer possible scores. However, when there are more aspects in the game, there is an increased probability of opinions being more diverse and more inter-community edges appearing. For that reason, as shown in tables C.10, C.11 and C.12, the groups which have only 2 aspects (C and D for MEDI-VALUE) have no inter-community edges, and therefore the inter/intra ration is always zero. It is possible that threshold values should be adapted to the number of aspects or that all groups should have the same number of aspects.

Aside from these caveats, generally, communities seem not to be well defined. On the one hand, when the inter/intra ratio is higher than 0 but smaller than 1, it is because there is a considerable number of inter-community edges. This means that many stakeholders agree with people from other communities, indicating a likelihood of a change of opinion or reaching a consensus. These are good news for the goal of achieving consensus in HTA. On the other hand, there are communities presenting ratio values much higher than 1. This can be good news, indicating that those communities are ”weak” and their individuals are much more likely to change opinions and move to another community, but unfortunately, it can also indicate that the partition is not so good. Note that the weight of the edges should not be forgotten. Although we do not explore it deeply, not all edges represent the same strength of agreement.

7.3.3 ”Is the triadic closure property also verified in this context?”

Looking at table C.13, in appendix C, it is possible to notice that, for the majority of the groups of aspects, both transitivity and clustering coefficient values are close to 1.

Results from Group E (MEDI-VALUE) outstand from others, with values around 0.6. For this group, a higher diversity of answers and a considerable number of extra-community edges were observed. Despite some exceptions, since they do not represent the groups' majority, we can conclude that, in general, the triadic closure property of social networks is also verified in these ”agreement networks”.

This finding is notable for two main reasons. First, it defines a property of this type of network, supporting the hypothesis that it is similar to social networks. Second, it describes a social behaviour highly relevant for the achievement of consensus. One way of interpreting this property in social networks is, quoting Aalabaf-Sabaghi, ”if two people in a social network have a friend in common, then there is an increased likelihood that they will become friends themselves at some point in the future”. Being also verified in ”agreement networks”, we can infer that, *if two stakeholders both agree with a third stakeholder, there is an increased likelihood that they will agree in the future.* This property can be pivotal to describe the importance of common agreements and the change of opinion based on other peoples' views.

7.4 Research stage 2 - Information extracted for all aspects

Up until now, we have been exploring groups of related aspects. We are now engaged in understanding if conclusions are preserved when all aspects' responses are used. As before, only relevant communities are considered for the analysis, i.e., communities composed by a few individuals are not discussed.

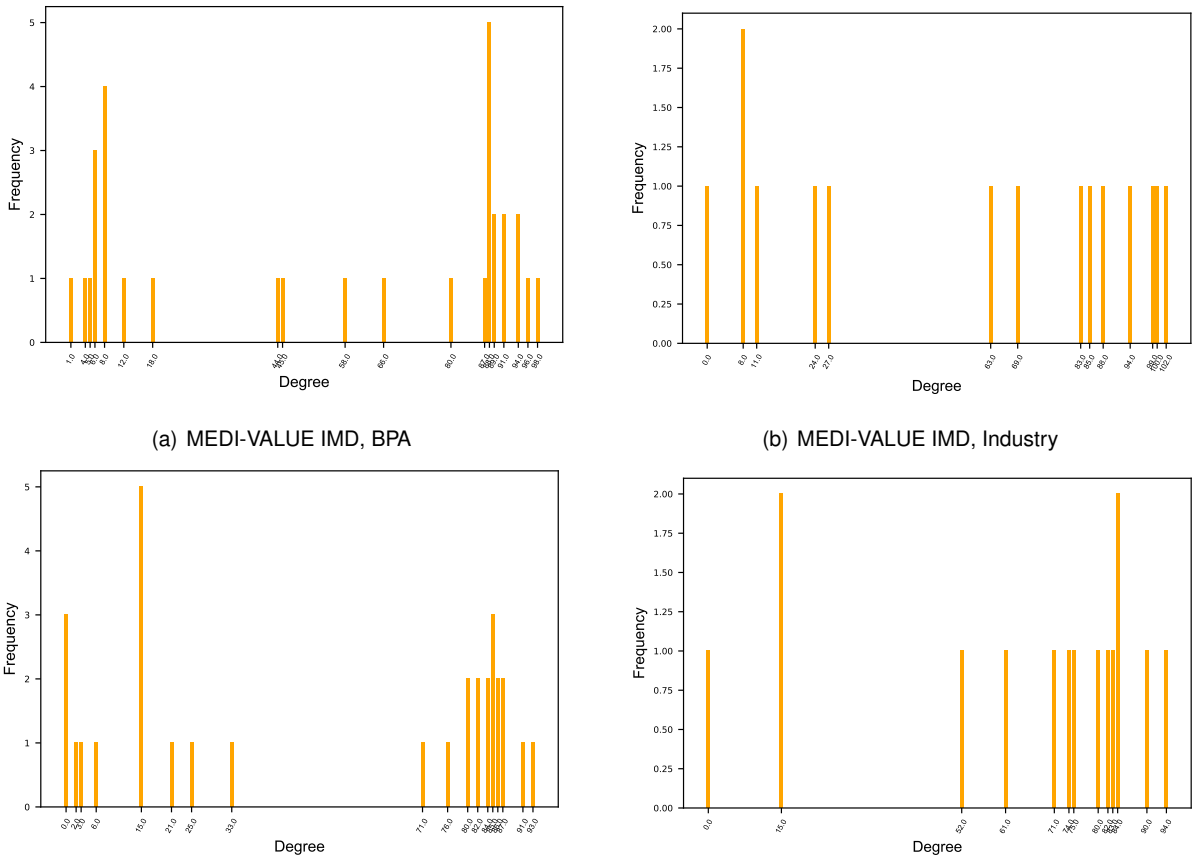
7.4.1 ”What characterises the obtained communities?”

Figure 7.11 compares the degree distribution for different types of stakeholders and table 7.10 informs the reader about the distribution of stakeholders across the main communities.

Overall, with some expected differences between BBIVT and IMD, all types present a heterogeneous degree distribution, showing the different agreement relations established within groups. For instance, there is a wider distribution per community for BPA compared to the industry. This is reflected in histograms in figure 7.11. The more varied the distribution per community, the more different degree values there are and, therefore, the more the bars in histograms. However, in general, similarly to the groups of aspects' observations, there is a diverse distribution of degrees. Similar stakeholders do not seem to agree with the same people. These results were similar for the other groups and IMPACT-HTA, considering its groups.

Moreover, as a general rule, stakeholders are spread across different communities. Once more, one can observe that results are significantly different for IMD and BBIVT, suggesting that the context of the Delphi survey strongly influences the results. One can also observe that when all aspects are considered, similar stakeholders seem to show more similar agreement patterns, especially looking at

the results of patients and citizens from IMD and industry from BBIVT. However, one should be careful to notice that four principal communities were found for IMD and only three for BBIVT. In addition, there are fewer industry and patient and citizen stakeholders, which can lead to more different results for these groups. Since the total number of participants is relatively low, few changes of opinion can seem higher. In conclusion, these differences represent few individuals.



(a) MEDI-VALUE IMD, BPA

(b) MEDI-VALUE IMD, Industry

(c) MEDI-VALUE BBIVT, BPA

(d) MEDI-VALUE BBIVT, Industry

Figure 7.11: Degree distribution, regarding MEDI-VALUE (IMD and BBIVT networks). Degree distributions of Buyers, Policymakers and Academics and Industry (I) were compared. A diverse degree distribution is observed.

Table 7.10: Distribution of answers, per group, per community (C), regarding MEDI-VALUE's full networks.[1]

Dataset	C	% HPro	% BPA	% PC	% I
MEDI-VALUE - Implantable Medical Devices	C1	45.0	38.7	53.6	26.7
	C2	30.0	25.8	21.4	26.7
	C3	16.7	25.8	25.0	26.7
	C4	6.7	9.7	0	13.3
MEDI-VALUE - Biomarkers-based *in vitro* Tests	C1	50.0	35.5	39.3	73.3
	C2	36.7	38.7	46.4	6.7
	C3	10.0	16.1	10.7	13.3

[1] The columns represent the percentage, in each community, of each stakeholder type - Healthcare Professionals, Buyers, Policymakers and Academics, Patients and citizens (PC) and Industry (I).

Table 7.11: Attribute assortativity coefficient regarding MEDI-VALUE's full networks.

Dataset	Attribute assortativity coefficient
MEDI-VALUE - Implantable Medical Devices	−0.005 574 452
MEDI-VALUE - Biomarkers-based *in vitro* Tests	−0.014 867 320

Regarding the assortativity coefficient, the values are close to zero, as shown in table 7.11. These results suggest, once again, that there is no particular tendency for similar stakeholders to agree more with each other. The relation is instead similar to random. Similar results were found for IMPACT-HTA.

We now focus on results from the main communities of the MEDI-VALUE IMD dataset in more detail. The distribution of answers across comunities and groups is shown in figure 7.12 and communities' composition in figure 7.13.

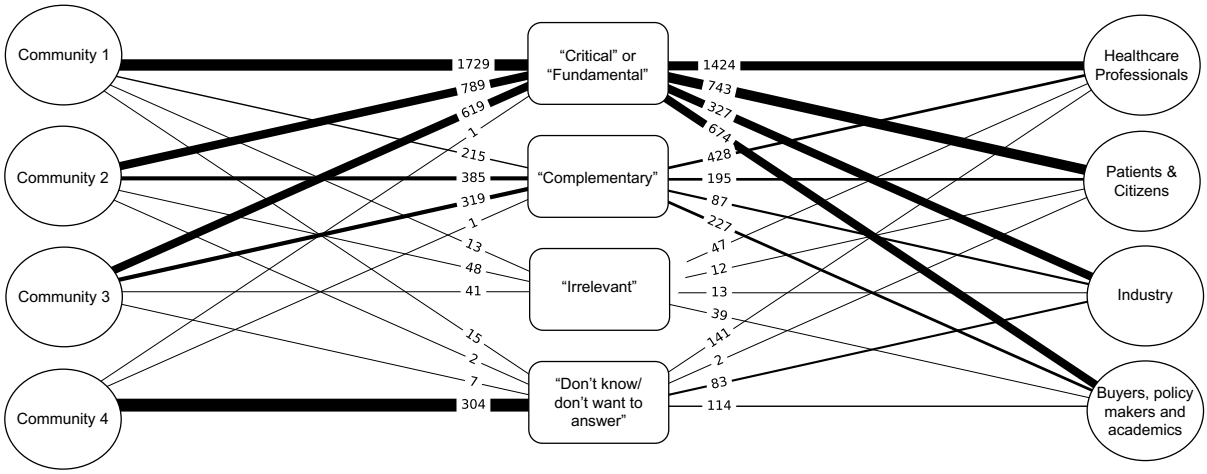
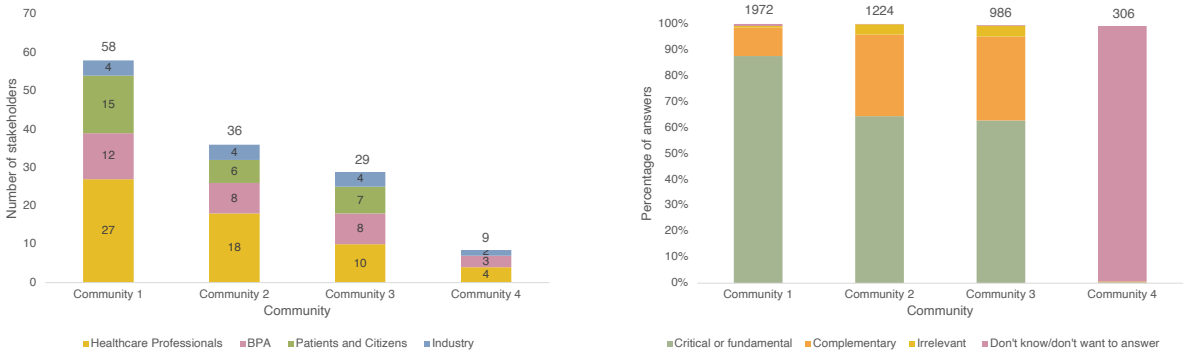


Figure 7.12: Multipartite network, regarding all MEDI-VALUE's aspects (IMD), showing the distribution of answers across communities and stakeholder groups. The value in each line corresponds to the total number of answers, not stakeholders. Note that 34 aspects are considered. The communities do not match stakeholders groups.



(a) Stakeholder types' distribution.

(b) Type of answers' distribution

Figure 7.13: Distribution of stakeholders' type and answers, per community, regarding all MEDI-VALUE's aspects (IMD). In (a) numbers regard stakeholders and in (b) answers. Stakeholders are widely distributed across the four communities. Communities 1, 2 and 3 are characterised by "Critical" or "Fundamental" answers. Community 4 is characterised by "Don't know/don't want to answer" responses.

These representations were found relevant when analysing groups of aspects and also add value here. Opinions are quite diverse, as shown in figure 7.12 and communities composed by a diversity of stake-

holders, as shown in figure 7.13 (a). These representations allow us to understand that stakeholders are, once again, quite spread across communities and the patient and citizen behaviour previously noted in 7.10 is not particularly observed.

Again, looking at the width of the lines on the right side of the figure 7.12 and also by the fact that three of the four central communities are characterised by ”Critical” or ”Fundamental” answers, we can verify that people tend to find aspects relevant. Thus, the distribution of stakeholders per community usually does not depend on their type. The majority have similar opinions, regardless of the type.

This diversity of composition of communities is also observed for the other two datasets. However, there is no direct relation between the main communities. For example, for both MEDI-VALUE datasets the results are slight different. We invite the reader to compare IMD and BBIVT results, the latter presented in figures 7.14 and 7.15.

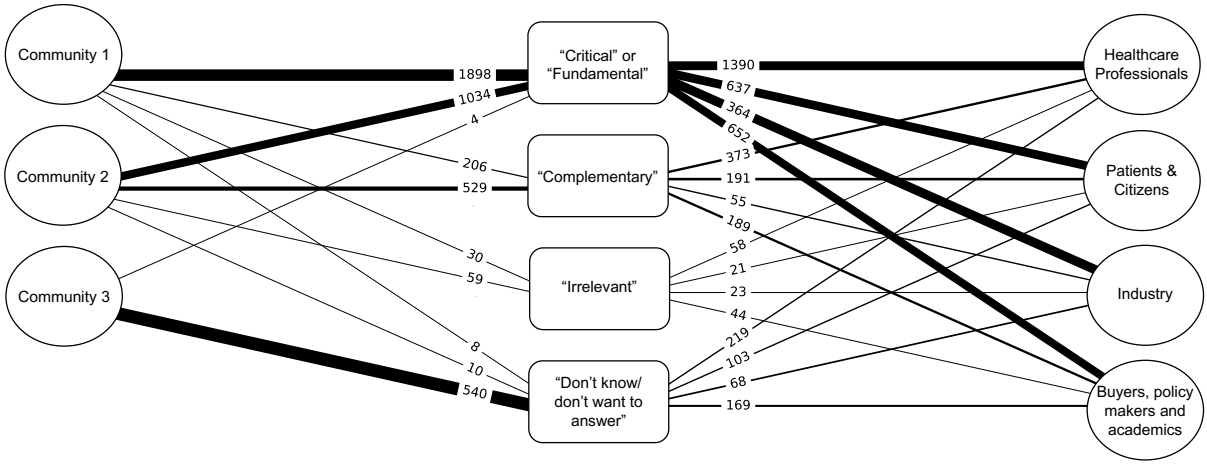
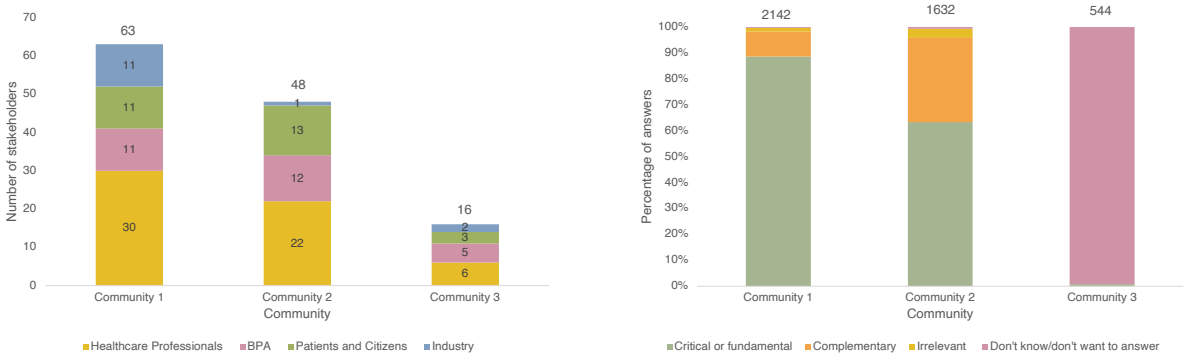


Figure 7.14: Multipartite network, regarding all MEDI-VALUE’s aspects (BBIVT), showing the distribution of answers across communities and stakeholder groups. The value in each line corresponds to the total number of answers, not stakeholders. Note that 34 aspects are considered. The communities do not match stakeholders groups.



(a) Stakeholder types’ distribution.

(b) Type of answers’ distribution

Figure 7.15: Distribution of stakeholders’ type and answers, per community, regarding all MEDI-VALUE’s aspects (BBIVT). In (a) numbers regard stakeholders and in (b) answers. Stakeholders are widely distributed across the three communities. Communities 1 and 2 are characterised by ”Critical” or ”Fundamental” answers. Community 3 is characterised by ”Don’t know/don’t want to answer” responses.

The main findings are the same, but for BBIVT there are only three main communities. If a parallelism is made, communities 2 and 3 from IMD probably correspond to BBIVT’s community 2. However, it is

possible to understand that even though stakeholders' type does not characterise communities, they are also not characterised by specific answers. They depend on the aspects and dataset.

We want to remind the reader of the nature of the projects. On the one hand, for groups of related aspects, it is expected that people have similar opinions regarding those aspects. For example, if one stakeholder finds one aspect "Irrelevant", it is expected for him/her to share a similar opinion for the other aspects of the group. Thus, opinions will be likely "polarised" for the types of answers for a group of aspects. On the other hand, for all aspects, people can have more diverse opinions. Stakeholders from the same community sharing similar views may find some aspects relevant, but not others, meaning that clusters are more likely to be characterised by patterns of answers. Overall, what matters is not only the type of answer but also which aspect the response regards.

In conclusion, observations suggest that it is more beneficial for decision-makers to analyse specific aspects or groups of related aspects and not all at once. The analysis of the responses regarding all aspects simultaneously is not as representative and insightful as for groups.

Altogether, the results from all three measures, i.e., the distribution of stakeholders, the degree distribution and the assortativity coefficient, suggest no particular match between the obtained communities and stakeholders' groups. Communities seem to be defined by patterns of answers.

7.4.2 "Is there a clear division between communities?"

The same measures used for aspects' groups were employed to understand if there is a clear division between communities for all aspects. If this division is unclear, many stakeholders agree with individuals from other clusters, suggesting a higher susceptibility for changes of opinion and possible achieving consensus. Table 7.12 presents a detailed analysis of the main MEDI-VALUE's communities' edges.

For the IMD network, there is, in general, a much higher number of edges between different communities (inter) than within (intra). Only community 4, characterised by the "Don't know/don't want to answer" option, is well defined. This suggests that communities 1, 2 and 3 are not firmly defined, and their individuals largely agree with each other. For BBIVT, not only the "Don't know/don't want to answer" community is well defined (community 3), but community 1, characterised by "Critical" or "Fundamental" answers, seems to be more clearly defined but still having connections with other communities.

It is not surprising that the "Don't know/don't want to answer" communities are more isolated since they concern people with no defined opinions. Overall, the external ratios are mainly higher than 1, which may not be ideal since there are more edges between communities than within. However, it can also suggest that stakeholders easily and commonly agree at some point in the surveys.

These findings suggest that communities are not clearly defined and that several stakeholders agree with individuals from other communities. This means that, in general, people tend to have similar opinions

at a given point, which can be good news for the achievement of consensus. However, this conclusion should be made carefully since numerous aspects are being considered and, therefore, there are several combinations of relationships between stakeholders. Analysis of groups of aspects appears to be more insightful. Results for IMPACT-HTA were similar.

Table 7.12: Intra and inter-community edges evaluation regarding MEDI-VALUE's datasets, for each community C.[1]

Dataset	C	# Intra C E	# Inter C E	Ratio	Coverage (%)
MEDI-VALUE - IMD	Total	2215	4616	2.084	-
	C1	1555	1945	1.251	34.38
	C2	363	1515	4.174	8.03
	C3	261	1156	4.429	5.77
	C4	36	0	0	0.80
MEDI-VALUE - BBIVT	Total	2426	2930	1.208	-
	C1	1802	1465	0.8129	46.31
	C2	504	1465	2.907	12.95
	C3	120	0	0	3.084

[1] The columns represent, in the total network and each community, the number of intra-community edges (# Intra C E), the number of inter-community edges (# Inter C E), the ratio of inter/intra community edges (Ratio) and, finally, the coverage (Coverage) of each community, in percentage. The latter is only calculated for communities.[1]

7.4.3 "Is the triadic closure property also verified in this context?"

Finally, we want to double-check if the triadic closure principle applies to this type of network. The average degree values are presented as a control group since these results are only valid if the average degree is not too high, which is the case. Results are shown in table 7.13.

Looking at results regarding transitivity and average clustering coefficient, it is possible to notice that the values are close to one. As explained before, this suggests that the triadic closure principle is verified.

Table 7.13: Transitivity, average clustering coefficient and average degree for MEDI-VALUE projects.

Dataset	Transitivity	Avg Clustering Coefficient	Average Degree
MEDI-VALUE - Implantable Medical Devices	0.849 0	0.827 0	67.507 5
MEDI-VALUE - Biomarkers-based *in vitro* Tests	0.880 9	0.837 4	58.074 6

As mentioned throughout this chapter, results for IMPACT-HTA were similar, according to its groups of aspects and scale used. Although the design of the surveys was distinct, the results were somewhat similar and comparable, corroborating our framework. One particular difference between the two projects is the use of one panel *versus* six panels, one for each stakeholder group. This significant difference did not seem to produce different results. However, its major influence is probably related to a change of opinions between rounds since it allows participants to interact with more people.

Chapter 8

Conclusions and Future Work

8.1 Conclusions

We started by approaching the question, ”Are NS tools suitable for analysing Delphi results?”. We verified that the conversion of Delphi-originated data into network-based data is in line with the expected interpretation of the results. Additionally, we observed that networks and communities' topology were similar for both MEDI-VALUE and IMPACT-HTA projects. These findings suggest that the results for similar surveys with different participants, questions, scales, and slightly distinct protocols follow a similar structure. This way, we believe we were able to verify that NS and CD are indeed proper and promising to be used to analyse HTA Delphi results, where nodes represent stakeholders and edges their agreement. We implemented our framework for two HTA projects, yet its employment is promising for other datasets and health contexts to explore stakeholders agreement networks.

Next, we explored the question ”What information regarding stakeholders' views can be obtained?”. Analysis of Delphi results is usually based on the analysis of stakeholder groups. Still, our framework allowed us to verify that there is no significant match between stakeholder types and communities, mainly characterised by the type and patterns of answers given. Usually, when a group of stakeholders concentrated in one community, it was not because of their type but because the majority agreed. One interesting detail is that there is a strong tendency for people to find all aspects highly relevant.

We were also able to use different tools for analysing and representing the results and understand better the division and composition of communities, which are often not well defined. Results suggest a poor strength of communities which might facilitate consensus in the future. The results and used representations can serve decision-makers in conferences when discussing the inclusion of aspects. Furthermore, we concluded that the triadic closure property of social networks is also verified in these ”agreement networks”, meaning that if two stakeholders both agree with a third stakeholder, there is an increased likelihood that they will also agree in the future.

An interesting, specific finding we were able to obtain concerns the context of the surveys. Although the results were similar for the three datasets, there were differences between IMD and BBIVT results. This finding suggests that the technology being evaluated influences the results. This is not unlikely since IMD and BBIVT technologies are much different. It can also reveal that people do not participate carefully and thus give different answers in different surveys for the same aspects.

8.2 Study Limitations

During this work, we validated our framework and obtained interesting information on stakeholders' views and interactions. However, there are some important limitations associated with our work:

- We should not forget that the Delphi technique itself presents some caveats. The validity and accuracy of this technique are often called into question, and we can never be sure that the answers genuinely reflect participants' opinions. For instance, people can make mistakes when selecting an option or even get tired and answer without careful deliberation;
- Even though we did not use parametric tests, avoiding some setbacks, there are other scale-related concerns. It is worrying that we can not be sure about the interpretation of neutral items. Additionally, although aggregating items into groups facilitates the analysis of results, there is no substantial evidence in the literature of the suitability of this grouping. The comparison of answers and calculation of agreement scores is also poorly supported due to little related research. Overall, the definition of the proximity of answers and agreement should be better defined;
- While exploring a new approach can be exciting, it can also be limited due to scarce related research. Since there are no similar studies, our framework is limited in that sense, being harder to evaluate its accuracy. Some results are challenging to analyse due to non-existent comparisons. This issue also affects the analysis of the MEDI-VALUE scale, which is not found in the literature;
- The threshold value used strongly influences results. A poor choice can lead to poor results;
- Several tasks were performed manually, which makes the replication of the results more difficult and exhaustive than it would be with a more automatic approach;
- Since we were provided with three datasets and performed several analyses, we ended up with a vast number of results not being possible to present them all.

8.3 Future Work

As discussed throughout this work, the analysis of stakeholders' views and the use of NS and CD for the analysis of HTA Delphi results are not much explored in the literature. For that reason, there are several ways of (1) improving the framework here proposed and (2) exploring different approaches.

(1) Improving the proposed framework:

- Use more datasets and projects to compare results and better validate the framework;
- Take advantage of other network measures to define this type of ”agreement networks better”;
- Better explore and define the concept of ”agreement”, how it is measured and develop a guideline to choose a threshold value according to the research goal. Possibly adapt the threshold value to the number of aspects being evaluated, in the case of groups of aspects;
- Apply this framework to the results from every survey round and analyse the change of opinions. Analyse, for instance, the correlation between the change of views and the triadic closure property. It is also possible to analyse the influence of one full panel and parallel panels for the types of stakeholders in changing opinions between rounds.

(2) Other approaches:

- Use different Community Detection algorithms and compare the results to validate the choice of the Louvain algorithm or find a better one;
- Explore algorithms accepting overlapping communities to find stakeholders who belong to more than one cluster, i.e., key individuals to achieve consensus. This path can be exciting since we found that communities are not clearly defined, and many inter-community edges were detected;
- Investigate the use of signed networks, where the edges are attributed with a (+) for agreement and (-) for disagreement. These networks allow the analysis of social balance and the frequency of different triangles and their comparison with the patterns of classic social networks;
- Develop software in order to convert the process into a more automatic and less manual procedure. It would be interesting to have a platform where a user could upload survey's raw data and visualise and manipulate the results;
- Explore a framework combining the proposed analysis with the analysis of comments. It can be a possibility to combine this framework with Discourse Network Analysis;
- Understand, alongside decision-makers, how these results can be easily presented and reported to support decision-making processes.

As a final remark, the possibilities are immense when discussing data analysis, including Delphi-originated data. There are currently several robust methodologies employed to study Delphi surveys' results that inform and support decision-makers. However, new and exciting approaches suitable in other contexts are constantly appearing. Thus, it was fascinating to find out at what level those can also be applied to HTA and Delphi surveys. With a great space for improvement, we believe we could complement the current Delphi analysis and provide decision-makers with fresh insights and visualisation tools, enhancing HTA processes. Finally, we hope this work inspires others to improve this framework and explore innovative approaches which that not be so obvious.

Bibliography

J M Stratil, R Baltussen, I Scheel, A Nacken, and E A Rehfuess. Development of the WHO-INTEGRATE evidence-to-decision framework: an overview of systematic reviews of decision criteria for health decision-making. *Cost Effectiveness and Resource Allocation*, 18(1), feb 2020. doi: 10.1186/s12962-020-0203-6.

Ivett Jakab, Bertalan Németh, Baher Elezbawy, Melis Almula Karadayı, Hakan Tozan, Sabahattin Aydın, Jie Shen, and Zoltán Kaló. Potential Criteria for Frameworks to Support the Evaluation of Innovative Medicines in Upper Middle-Income Countries—A Systematic Literature Review on Value Frameworks and Multi-Criteria Decision Analyses. *Frontiers in Pharmacology*, 11:1203, 2020. doi: 10.3389/fphar.2020.01203.

Jackie Street, Tania Stafinski, Edilene Lopes, and Devidas Menon. Defining the role of the public in Health Technology Assessment (HTA) and HTA-informed decision-making processes. *International Journal of Technology Assessment in Health Care*, 36(2):87–95, 2020. doi: 10.1017/S0266462320000094.

Stephanie Polus, Tim Mathes, Corinna Klingler, Melanie Messer, Ansgar Gerhardus, Constance Stegbauer, Gerald Willms, Heidi Ehrenreich, Georg Marckmann, and Dawid Pieper. Health Technology Assessment of Public Health Interventions Published 2012 to 2016: An Analysis of Characteristics and Comparison of Methods. *International Journal of Technology Assessment in Health Care*, 35(4): 280–290, 2019. doi: 10.1017/S0266462319000515.

Infarmed. Serviço Nacional de Saúde (SNS), Infarmed - Autoridade nacional do medicamento e produtos de saúde. *Avaliação de tecnologias de saúde*. 2021. https://www.infarmed.pt/web/infarmed/entidades/medicamentos-uso-humano/avaliacao-tecnologias-de-saude, accessed: 2021-10-23.

Monica R Geist. Using the Delphi method to engage stakeholders: A comparison of two studies. *Evaluation and Program Planning*, 33(2):147–154, 2010. doi: 10.1016/j.evalprogplan.2009.06.006.

Ian Belton, Alice MacDonald, George Wright, and Iain Hamlin. Improving the practical application of the Delphi method in group-based judgment: A six-step prescription for a well-founded and defensible process. *Technological Forecasting and Social Change*, 147:72–82, jul 2019. doi: 10.1016/j.techfore.2019.07.002.

Grupo Sintevi, Jose D Martinez-Ezquerro, Sonia Maria Ruiz-Cejudo, Alejandra Bustamante-Fuentes, Alvaro Diaz-Badillo, Esperanza M Garcia-Oropesa, Elena B Lopez-Sosa, Yoscelina E Martinez-Lopez, Oscar O Moctezuma-Chavez, Edna J Nava-Gonzalez, Adriana L Perales-Torres, Lucia M Perez-Navarro, Marisol Rosas-Diaz, and Juan C Lopez-Alvarenga. Expert consensus in times of COVID-19: health applications of the Delphi method. *Cirugia y Cirujanos*, 89(1):120–129, 2021. doi: 10.24875/CIRU.20000936.

Sinead Keeney, Felicity Hasson, and Hugh McKenna. Consulting the oracle: ten lessons from using the Delphi technique in nursing research. *Journal of advanced nursing*, 53(2):205–212, jan 2006. doi: 10.1111/j.1365-2648.2006.03716.x.

F Hasson, S Keeney, and H McKenna. *Journal of advanced nursing*, (4), oct . doi: 10.1046/j.1365-2648.2000.t01-1-01567.x.

Ann Bowling. *Research Methods in Health Investigating Health and Health Services*. Open University Press, 2014. ISBN 9780335262748.

Anne Lübbeke, Andrew J Carr, and Pierre Hoffmeyer. Registry stakeholders. *EFORT open reviews*, 4 (6):330–336, jun 2019. doi: 10.1302/2058-5241.4.180077.

Axel C Muehlbacher and Anika Kaczynski. Making Good Decisions in Healthcare with Multi-Criteria Decision Analysis: The Use, Current Research and Future Development of MCDA. *Applied Health Economics and Health Policy*, 14(1):29–40, feb 2016. doi: 10.1007/s40258-015-0203-4.

F Bourrée, P Michel, and L R Salmi. Consensus methods: review of original methods and their main alternatives used in public health TT - Méthodes de consensus: revue des méthodes originales et de leurs grandes variantes utilisées en santé publique. *Revue d'epidemiologie et de sante publique*, 56 (6):415–423, dec 2008. doi: 10.1016/j.respe.2008.09.006.

Jon Landeta. Current validity of the Delphi method in social sciences. *Technological Forecasting and Social Change*, 73(5):467–482, 2006. doi: https://doi.org/10.1016/j.techfore.2005.09.002.

Harold A Linstone and Murray Turoff. Delphi: A brief look backward and forward. *Technological Forecasting and Social Change*, 78(9):1712–1719, 2011. doi: 10.1016/j.techfore.2011.04.005.

Felicity Hasson and Sinead Keeney. Enhancing rigour in the Delphi technique research. *Technological Forecasting and Social Change*, 78(9):1695–1704, 2011. doi: https://doi.org/10.1016/j.techfore.2011.04.005.

Welphi. Welphi - The Survey App To Build Consensus. 2021. https://www.welphi.com/en/Home.html.

MEDI-VALUE. Developing HTA tools to consensualise MEDIcal devices' VALUE through multicriteria decision analysis, 2021. http://medivalue.tecnico.ulisboa.pt, accessed: 2021-08-31.

IMPACT-HTA. Work Package 7: Methodological tools using multi-criteria value methods for HTA decision-making, 2021. https://www.impact-hta.eu/work-package-7, accessed: 2021-08-31.

Bert Weijters, Elke Cabooter, and Niels Schillewaert. The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, 27(3):236–247, 2010. doi: 10.1016/j.ijresmar.2010.02.004.

QuestionPro. Rating Scale: Definition, Survey Question Types and Examples, 2021. https://www.questionpro.com/blog/rating-scale/, accessed: 2021-05-12.

MeasuringU. 15 Common Rating Scales Explained, 2021. https://measuringu.com/rating-scales/, accessed: 2021-05-12.

Jarl K Kampen. Reflections on and test of the metrological properties of summated rating, Likert, and other scales based on sums of ordinal variables. *Measurement*, 137:428–434, 2019. doi: 10.1016/j.measurement.2019.01.083.

Shane P. Desselle. Construction, implementation, and analysis of summated rating attitude scales. *American Journal of Pharmaceutical Education*, 69(5), 2005. doi: 10.5688/aj690597.

Montri Sangthong. The Effect of the Likert Point Scale and Sample Size on the Efficiency of Parametric and Nonparametric Tests. *Thailand Statistician*, 18(1):55–64, jan 2020. Available: https://ph02.tci-thaijo.org/index.php/thaistat/article/view/228886.

Umesh Wadgave and Mahesh R Khairnar. Parametric tests for Likert scale: For and against. *Asian Journal of Psychiatry*, 24:67–68, 2016. doi: 10.1016/j.ajp.2016.08.016.

Susan Jamieson. Likert scales: how to (ab)use them. *Medical Education*, 38(12):1217–1218, 2004. doi: 10.1111/j.1365-2929.2004.02012.x.

Todd A DeWees, Gina L Mazza, Michael A Golafshar, and Amylou C Dueck. Investigation Into the Effects of Using Normal Distribution Theory Methodology for Likert Scale Patient-Reported Outcome Data From Varying Underlying Distributions Including Floor/Ceiling Effects. *VALUE IN HEALTH*, 23 (5):625–631, may 2020. doi: 10.1016/j.jval.2020.01.007.

Mark Rasburn, Heidi Livingstone, and Sarah E Scott. Strengthening patient outcome evidence in health technology assessment: a coproduction approach. *International Journal of Technology Assessment in Health Care*, 37:e12, 2021. doi: DOI:10.1017/S0266462320002202.

Julie Polisena, Rossana Castaldo, Oriana Ciani, Carlo Federici, Simone Borsci, Matteo Ritrovato, Daniel Clark, and Leandro Pecchia. Health technology assessment methods guidelines for medical devices: how can we address the gaps? The international federation of medical and biological engineering perspective. *International journal of technology assessment in Health care*, 34(3):276–289, 2018. doi: 10.1017/S0266462318000314.

Tobias Gnatzy, Johannes Warth, Heiko von der Gracht, and Inga-Lena Darkow. Validating an innovative real-time Delphi approach - A methodological comparison between real-time and conventional Delphi studies. *Technological Forecasting and Social Change*, 78(9):1681–1694, 2011. doi: https://doi.org/10.1016/j.techfore.2011.04.006.

Charlene R Weir, Bret L Hicken, Hank Steven Rappaport, and Jonathan R Nebeker. Crossing the quality chasm: the role of information technology departments. *American journal of medical quality : the official journal of the American College of Medical Quality*, 21(6):382–393, 2006. doi: 10.1177/1062860606293150.

Phillip A Bishop and Robert L Herron. Use and Misuse of the Likert Item Responses and Other Ordinal Measures. *International journal of exercise science*, 8(3):297–302, 2015. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4833473/.

Constantin Mircioiu and Jeffrey Atkinson. A Comparison of Parametric and Non-Parametric Methods Applied to a Likert Scale. *Pharmacy*, 5(2), jun 2017. doi: 10.3390/pharmacy5020026.

Geoff Norman. Likert scales, levels of measurement and the “laws” of statistics. *Advances in Health Sciences Education*, 15(5):625–632, 2010. doi: 10.1007/s10459-010-9222-y.

Seung Youn Yonnie Chyung, Katherine Roberts, Ieva Swanson, and Andrea Hankinson. Evidence-Based Survey Design: The Use of a Midpoint on the Likert Scale. *Performance Improvement*, 56(10): 15–23, 2017. doi: 10.1002/pfi.21727.

Ângela Freitas, Paula Santana, Mónica D Oliveira, Ricardo Almendra, João C Bana e Costa, and Car-

los A Bana e Costa. Indicators for evaluating European population health: a Delphi selection process. *BMC Public Health*, 18(1):557, 2018. doi: 10.1186/s12889-018-5463-0.

Jenna M Evans, G Ross Baker, Whitney Berta, and Jan Barnsley. A cognitive perspective on health systems integration: results of a Canadian Delphi study. *BMC Health Services Research*, 14(1):222, 2014. doi: 10.1186/1472-6963-14-222.

Anna Kearney, Anne Daykin, Alison R G Shaw, Athene J Lane, Jane M Blazeby, Mike Clarke, Paula Williamson, and Carrol Gamble. Identifying research priorities for effective retention strategies in clinical trials. *TRIALS*, 18, aug 2017. doi: 10.1186/s13063-017-2132-z.

Mary Falzarano and Genevieve Pinto Zipp. Seeking consensus through the use of the Delphi technique in health sciences research. *Journal of allied health*, 42(2):99–105, 2013. Available: https://pubmed.ncbi.nlm.nih.gov/23752237/.

Dyon Hoekstra, Margot Mütsch, Christina Kien, Ansgar Gerhardus, and Stefan K Lhachimi. Identifying and prioritising systematic review topics with public health stakeholders: A protocol for a modified Delphi study in Switzerland to inform future research agendas. *BMJ Open*, 7(8):e015500, aug 2017. doi: 10.1136/bmjopen-2016-015500.

Hanna Kallio, Pietila Anna-Maija, and Mari Kangasniemi. Environmental responsibility in nursing in hospitals: A modified Delphi study of nurses' views. *JOURNAL OF CLINICAL NURSING*, 29(21-22): 4045–4056, nov 2020. doi: 10.1111/jocn.15429.

Ivan R Diamond, Robert C Grant, Brian M Feldman, Paul B Pencharz, Simon C Ling, Aideen M Moore, and Paul W Wales. Defining consensus: A systematic review recommends methodologic criteria for reporting of Delphi studies. *Journal of Clinical Epidemiology*, 67(4):401–409, 2014. doi: https://doi.org/10.1016/j.jclinepi.2013.12.002.

Toni Lange, Christian Kopkow, Jörg Lützner, Klaus-Peter Günther, Sascha Gravius, Hanns-Peter Scharf, Johannes Stöve, Richard Wagner, and Jochen Schmitt. Comparison of different rating scales for the use in Delphi studies: different scales lead to different consensus and show different test-retest reliability. *BMC Medical Research Methodology*, 20(1):28, 2020. doi: 10.1186/s12874-020-0912-8.

Edna Keeney, Howard Thom, Emma Turner, Richard M Martin, and Sabina Sanghera. Using a Modified Delphi Approach to Gain Consensus on Relevant Comparators in a Cost-Effectiveness Model: Application to Prostate Cancer Screening. *Pharmacoeconomics*, 2021. doi: 10.1007/s40273-021-01009-6.

Markets and Markets. Markets and Markets - Machine Learning Market. 2021. https://www.marketsandmarkets.com/Market-Reports/machine-learning-market-263397704.html, accessed: 2021-10-24.

Thiago Christiano Silva and Liang Zhao. *Machine Learning in Complex Networks*. Springer International Publishing, 1 edition, 2016. doi: 10.1007/978-3-319-17290-3.

K Sathiyakumari and M S Vijaya. Community Detection Based on Girvan Newman Algorithm and Link Analysis of Social Media. In *DIGITAL CONNECTIVITY - SOCIAL IMPACT*, volume 679 of *Communications in Computer and Information Science*, pages 223–234. SPRINGER-VERLAG SINGAPORE PTE LTD, 2016. doi: 10.1007/978-981-10-3274-5_18.

Hongxu Chen, Hongzhi Yin, Xue Li, Meng Wang, Weitong Chen, and Tong Chen. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 1353–1359, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee. doi: 10.1145/3041021.3051159.

Christina H. Buckton, Gillian Fergie, Philip Leifeld, and Shona Hilton. A discourse network analysis of UK newspaper coverage of the ”sugar tax” debate before and after the announcement of the Soft Drinks Industry Levy. *BMC Public Health*, 19(1):1–15, 2019. doi: 10.1186/s12889-019-6799-9.

Shona Hilton, Christina Buckton, Gillian Fergie, Tim Henrichsen, and Philip Leifeld. Comparison of stakeholder coalitions across pricing policies designed to improve public health, as represented in UK newspapers: a discourse network analysis study. *The Lancet*, 394:S52, 2019. doi: 10.1016/S0140-6736(19)32849-1.

Gillian Fergie, Philip Leifeld, Ben Hawkins, and Shona Hilton. Mapping discourse coalitions in the minimum unit pricing for alcohol debate: a discourse network analysis of UK newspaper coverage. *Addiction*, 114(4):741–753, 2019. doi: 10.1111/add.14514.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pages 226–231. AAAI Press, 1996. Available: https://dl.acm.org/doi/10.5555/3001460.3001507.

Krishna Raj P.M., Mohan Ankith, and Srinivasa K.G. *Practical Social Network Analysis with Python*. SPRINGER, 2018. doi: 10.1007/978-3-319-96746-2.

Zineb Felfli, Roy George, Khalil Shujaee, and Mohamed Kerwat. Community detection and unveiling of hierarchy in networks: a density-based clustering approach. *Applied Network Science*, 4(1), 2019. doi: 10.1007/s41109-019-0216-2.

Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010. doi: 10.1016/j.physrep.2009.11.002.

Varsha Veerappa and Emmanuel Letier. Clustering Stakeholders for Requirements Decision Making. In Daniel Berry and Xavier Franch, editors, *Requirements Engineering: Foundation for Software Quality*, pages 202–208. Springer Berlin Heidelberg, 2011. doi: 10.1007/978-3-642-19858-8_20.

Philip Leifeld. *The Oxford Handbook of Political Networks*.

Natalie R. Smith, Paul N. Zivich, Leah M. Frerichs, James Moody, and Allison E. Aiello. A Guide for Choosing Community Detection Algorithms in Social Network Studies: The Question Alignment Approach. *American Journal of Preventive Medicine*, 59(4):597–605, 2020. doi: 10.1016/j.amepre.2020.04.015.

Youngho Lee, Yubin Lee, Jeong Seong, Ana Stanescu, and Chul Sue Hwang. A comparison of network clustering algorithms in keyword network analysis: A case study with geography conference presentations. *International Journal of Geospatial and Environmental Research*, 7(3):1–16, 2020. Available: https://dc.uwm.edu/cgi/viewcontent.cgi?article=1130&context=ijger.

Karen Mkhitaryan, Josiane Mothe, and Mariam Haroutunian. Detecting communities from networks

: comparison of algorithms on real and synthetic networks. 26(3):231–267, 2019. Available: http://www.foibg.com/ijita/vol26/ijita26-03-p03.pdf.

Zhao Yang, René Algesheimer, and Claudio J. Tessone. A comparative analysis of community detection algorithms on artificial networks. *Scientific Reports*, 6(August), 2016. doi: 10.1038/srep30750.

Vincent D. Blondel, Jean Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10): 1–12, 2008. doi: 10.1088/1742-5468/2008/10/P10008.

Animesh Mukherjee, Monojit Choudhury, Fernando Peruani, Niloy Ganguly, and Bivas Mitra. *Dynamics On and Of Complex Networks*, volume 2. 2013.

NetworkX. Network analysis in Python, *networkx.classes.function.density*. 2021. https://networkx.org/documentation/stable/reference/generated/networkx.classes.function.density.html, accessed: 2021-08-31.

Baeldung. Determine Maximum Number of Edges in a Directed Graph, 2021. https://www.baeldung.com/cs/graphs-max-number-of-edges, accessed: 2021-08-27.

Matteo Cinelli, Leto Peel, Antonio Iovanella, and Jean Charles Delvenne. Network constraints on the mixing patterns of binary node metadata. *Physical Review E*, 102(6):1–22, 2020. doi: 10.1103/PhysRevE.102.062310.

Konstantinos Pelechrinis and Dong Wei. VA-index: Quantifying assortativity patterns in networks with multidimensional nodal attributes. *PLoS ONE*, 11(1):1–13, 2016. doi: 10.1371/journal.pone.0146188.

Feng Shi. Learn About Assortativity Coefficient in Python With Data From UK Faculty Dataset (2008). *Learn About Assortativity Coefficient in Python With Data From UK Faculty Dataset (2008)*, (2008), 2019. doi: 10.4135/9781526499059.

Morteza Aalabaf-Sabaghi. Networks, Crowds and Markets: Reasoning about a Highly Connected World. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 175(4):1073–1073, 2012. doi: 10.1111/j.1467-985x.2012.01069_4.x.

Carolina Paes de Faria. GitHub, *Thesis*, 2021. https://github.com/carolinapdfaria/Thesis.

Louvain Networkx. Community detection for NetworkX's documentation, *Louvain method* , 2021. https://python-louvain.readthedocs.io/en/latest, accessed: 2021-09-07.

Necromuralist. Data Science with Python, *Triadic Closure (Clustering)*, 2021. https://necromuralist.github.io/data_science/posts/triadic-closure/, accessed: 2021-08-31.

Universita' Degli Studi di Udine. Transitivity, 2021. http://users.dimi.uniud.it/~massimo.franceschet/teaching/datascience/network/transitivity.html, accessed: 2021-08-31.

NetworkX. NetworkX - Network analysis in Python, *networkx.algorithms.cluster.average_clustering*. 2021. https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.cluster.average_clustering.html, accessed: 2021-08-31.

Appendix A

Projects aspects

This appendix includes the aspects considered for both projects - MEDI-VALUE and IMPACT-HTA. The aspects are presented by category. In the case of MEDI-VALUE, a previous division into categories was adapted for the context of this work. For IMPACT-HTA, the division was defined from scratch.

A.1 MEDI-VALUE

Group A - Value for the patient

Aspect 11 - Comfort for the patient

Aspect 16 - Quality of the available scientific evidence

Aspect 19 - Patient-reported outcomes

Aspect 20 - Quality of life for the patient

Group B - Safety for the patient and/or healthcare professional

Aspect 9 - Exposure of the healthcare professional to physical or chemical agents

Aspect 14 - Risk analysis

Aspect 15 - Adverse events for the patient

Aspect 28 - Medical or technical complications for the patient

Group C - Impact of the use of the medical device in the healthcare organization

Aspect 10 - Workload for the healthcare professional

Aspect 23 - Financing

Group D - Costs with the use of the medical device

Aspect 34 - Cost of procedure without the cost of the medical device

Aspect 33 - Cost of the medical device (including complementary equipment)

Group E - Usability for the healthcare professional

Aspect 5 - User-friendliness for the healthcare professional

Aspect 6 - Time between procedure and results

Aspect 7 - Need for training of the healthcare professional

Aspect 8 - Learning curve of the healthcare professional

Aspect 12 - Connectivity

Group F - Technical performance of the medical device

Aspect 2 - Technical performance of the medical device

Aspect 1 - Specific features of the medical device

Aspect 4 - Sensitivity and Specificity

Aspect 13 - Clinical efficacy and/or effectiveness

Group G - Interest in the adoption of the medical device for the health system

Aspect 3 - Regulatory status of the medical device

Aspect 17 - Target population

Aspect 22 - Clinical guidelines

Aspect 24 - Public health interest

Aspect 27 - Market competitiveness

Aspect 32 - Capacity of the health system

Group H - Societal context of the adoption of the medical device

Aspect 26 - Equity

Aspect 29 - Stakeholders agreement on the adoption of the medical Aspect device

Aspect 30 - Environmental impact of the production and use of the Aspect medical device

Group I - Impact of the adoption of the medical device for the health system

Aspect 18 - Impact of the disease

Aspect 21 - Space for innovation for the healthcare organization

Aspect 25 - Budget impact to the health system

Aspect 31 - Efficiency

A.2 IMPACT-HTA

Group A – Disease related

Aspect 1 - Severity of the disease

Aspect 2 - Unmet need of the disease

Aspect 3 - Disease frequency (e.g. rarity)

Aspect 19 - Disease duration

Aspect 20 - Disease age of onset

Group B – Impact on society

Aspect 4 - Medicine's impact on mortality

Aspect 5 - Medicine's impact on morbidity

Aspect 14 - Medicine's impact on wider public health in terms of disease risk reduction in the community

Aspect 21 - Impact of the medicine's adoption on the health care system's organisation and delivery of care

Aspect 22 - Impact of the medicine's adoption on equity and ethical issues

Group C – Impact on patients

Aspect 6 - Medicine's impact on health-related quality of life

Aspect 13 - Medicine's ease and convenience for patients

Group D – Risks and precautions

Aspect 7 - Medicine's adverse events profile

Aspect 8 - Medicine's tolerability to patients

Aspect 9 - Medicine's contraindications of use

Aspect 10 - Medicine's special warnings and precautions

Group E – Medicine related

Aspect 11 - Medicine's mechanism of action

Aspect 17 - Medicine's efficiency

Aspect 24 - Medicine's therapeutic positioning

Group F – Economical and political related

Aspect 12 - Medicine's spill-over effects

Aspect 15 - Medicine's economic impact

Aspect 16 - Medicine's affordability

Aspect 23 - Alignment of the medicine-indication pair with leadership goals and governance requirements, including political support

Appendix B

Stakeholder proximity measure

This appendix includes a more detailed explanation of the proposal for measurement of stakeholders' answers proximity. This measurement is considered for a pair of stakeholders.

Remember that there are two scenarios for both projects: (1) only considering agreement and (2) subtracting the conflict variable from the agreement. Furthermore, for the case of IMPACT-HTA, we proposed other two variances: (I) "Neutral" and "No answer" groups as separate and then (II) as one group.

B.1 MEDI-VALUE

Starting with the MEDI-VALUE project, table B.1 presents how variables A (agreement) and C (conflict) change when comparing answers from one aspect for a pair o stakeholders.

Table B.1: Calculation of proximity of stakeholders answers in MEDI-VALUE. The proximity is calculated for a pair of stakeholders - Stakeholder A (S_A) and Stakeholder B (S_B).[1][2]

S_A \ S_B	Critical	Fundamental	Complementary	Irrelevant	DNDWTA[3]
Critical	SA (A=A+1)	SG (A=A+1)	DG	OG (C=C+1)	DG
Fundamental	SG (A=A+1)	SA (A=A+1)	DG	OG (C=C+1)	DG
Complementary	DG	DG	SA (A=A+1)	DG	DG
Irrelevant	OG (C=C+1)	OG (C=C+1)	DG	SA (A=A+1)	DG
DNDWTA[3]	DG	DG	DG	DG	SA (A=A+1)

[1] SA = *Same answer*; SG = *Same group*; DG = *Different group* and OG = *Opposite group.*

[2] (A=A+1) means that one point is added to the agreement variable A and (C=C+1) means that one point is added to the conflict variable C.

[3] DNDWTA = "Don't know/don't want to answer".

This comparison is made for every aspect to be considered. In the end, in scenario (1), the final value

of A should be normalized by performing the division of A by the number of aspects. In scenario (2), the score is given by A-C, and then the normalization is performed.

B.2 IMPACT-HTA

For IMPACT-HTA, the reasoning is similar, but scenarios (I) and (II) should be taken into consideration.

B.2.1 I - *Neutral* and *No answer* groups as separate groups

When *Neutral* and *No answer* groups are analysed as separate groups, the combination of these answers will be ”different group”. Table B.2 presents how variables A (agreement) and C (conflict) change when comparing answers from one aspect for a pair o stakeholders.

Table B.2: Calculation of proximity of stakeholders answers in IMPACT-HTA. The proximity is calculated for a pair of stakeholders - Stakeholder A (S_A) and Stakeholder B (S_B).[1][2]

S_A \ S_B	S Agree	Agree	NAND	Disagree	S Disagree	DNDWTA[3]
S Agree	SA (A=A+1)	SG (A=A+1)	DG	OG (C=C+1)	OG (C=C+1)	DG
Agree	SG (A=A+1)	SA (A=A+1)	DG	OG (C=C+1)	OG (C=C+1)	DG
NAND	DG	DG	SA (A=A+1)	DG	DG	DG
Disagree	OG (C=C+1)	OG (C=C+1)	DG	SA (A=A+1)	SG (A=A+1)	DG
S Disagree	OG (C=C+1)	OG (C=C+1)	DG	SG (A=A+1)	SA (A=A+1)	DG
DNDWTA[3]	DG	DG	DG	DG	DG	SA (A=A+1)

[1] SA = *Same answer*; SG = *Same group*; DG = *Different group* and OG = *Opposite group*.

[2] (A=A+1) means that one point is added to the agreement variable A and (C=C+1) means that one point is added to the conflict variable C.

[3] S Agree = ”Strongly Agree”, NAND = ”Neither agree nor disagree”, S Disagree = ”Strongly Disagree” and DNDWTA = ”Don’t know/don’t want to answer”.

Once again, this comparison is made for every aspect to be considered. In the end, in scenario (1), the final value of A should be normalised by performing the division of A by the number of aspects. In scenario (2), the score is given by A-C, and then the normalisation is performed.

B.2.2 II - *Neutral* and *No answer* groups as only one group

When ”Neutral” and ”No answer” groups are analysed as one group, the combination of these answers will be ”Same group”. Table B.2 presents how variables A (agreement) and C (conflict) change when comparing answers from one aspect for a pair o stakeholders.

Once again, this comparison is made for every aspect to be considered. In the end, in scenario (1), the final value of A should be normalised by performing the division of A by the number of aspects. In scenario (2), the score is given by A-C, and then the normalisation is performed.

Table B.3: Calculation of proximity of stakeholders answers in IMPACT-HTA. The proximity is calculated for a pair of stakeholders - Stakeholder A (S_A) and Stakeholder B (S_B). The difference from scenario (I) is highlighted in bolt.[1][2]

S_A \ S_B	S Agree	Agree	NAND	Disagree	S Disagree	DNDWTA[3]
S Agree	SA (A=A+1)	SG (A=A+1)	DG	OG (C=C+1)	OG (C=C+1)	DG
Agree	SG (A=A+1)	SA (A=A+1)	DG	OG (C=C+1)	OG (C=C+1)	DG
NAND	DG	DG	SA (A=A+1)	DG	DG	**SG (A=A+1)**
Disagree	OG (C=C+1)	OG (C=C+1)	DG	SA (A=A+1)	SG (A=A+1)	DG
S Disagree	OG (C=C+1)	OG (C=C+1)	DG	SG (A=A+1)	SA (A=A+1)	DG
DNDWTA[3]	DG	DG	**SG (A=A+1)**	DG	DG	SA (A=A+1)

[1] SA = *Same answer*; SG = *Same group*; DG = *Different group* and OG = *Opposite group*.

[2] (A=A+1) means that one point is added to the agreement variable A and (C=C+1) means that one point is added to the conflict variable C.

[3] S Agree = "Strongly Agree", NAND = "Neither agree nor disagree", S Disagree = "Strongly Disagree" and DNDWTA = "Don't know/don't want to answer".

Appendix C

Detailed results for groups of aspects

This appendix presents more detailed results regarding the analysis of groups of similar aspects.

C.1 Choice of the threshold value

This section includes a more exhaustive explanation of the choice of the threshold value when analysing the results for a group of aspects.

Looking at results from tables C.4, C.5 and C.6, it is possible to notice that, for the majority of the groups, there is an ”in-between” regarding the number of communities found. This midterm is found, most of the time, for values of 0.5, 0.6 and 0.7. It reflects a balance between not too few communities and not too many isolated stakeholders. The same balance is found for the average degree, as shown in tables C.1, C.2 and C.3

Interestingly, this midterm is not found for groups C, D and H from MEDI-VALUE and C and E from IMPACT-HTA. These groups correspond to the groups with fewer aspects and therefore are less possible arrangements of scores.

The only value that always contains the midterm result, if existing, is 0.6. Also, for the other items where this middle result does not occur, 0.6 still provides a suitable partition, with few isolated stakeholders and a relatively strong community structure. Therefore, the chosen value was 0.6.

Note that results from the same groups can only be compared for MEDI-VALUE projects since IMPACT-HTA has its own group division.

Table C.1: Number of edges (#E) and average degree (AD), for each group and threshold value (T), for the MEDI-VALUE Implantable Medical Devices dataset.

T	A		B		C		D		E		F		G		H		I	
	#E	AD	#E	AD	#E	AD	#E	AD	#E	AD	#E	AD	#E	AD	#E	AD	#E	AD
0.4	7348	109.67	7566	112.93	5235	78.13	5458	81.46	4013	59.90	7777	116.07	6331	94.49	3624	54.09	6050	90.30
0.5	6396	95.46	6777	101.15	1793	26.76	4023	60.04	4013	59.90	7302	108.99	4957	73.99	3624	54.09	3678	54.90
0.6	6396	95.46	6777	101.15	1793	26.76	4023	60.04	1959	29.24	7302	108.99	3431	51.21	3624	54.09	3678	54.90
0.7	6396	95.46	6777	101.15	1793	26.76	4023	60.04	1959	29.24	7302	108.99	3431	51.21	1181	17.63	3678	54.90
0.8	3633	54.22	4492	67.04	1793	26.76	4023	60.04	554	8.27	6265	93.51	1794	26.78	1181	17.63	1179	17.60

Table C.2: Number of edges (#E) and average degree (AD), for each group and threshold value (T), for the MEDI-VALUE Biomarkers-based *in vitro* Tests dataset.

T	A		B		C		D		E		F		G		H		I	
	#E	AD	#E	AD	#E	AD	#E	AD	#E	AD	#E	AD	#E	AD	#E	AD	#E	AD
0.4	5476	81.73	5928	88.48	4995	74.55	4755	70.97	3810	56.87	7119	106.25	4920	73.43	3484	52.00	5711	85.24
0.5	4223	63.03	5443	81.24	1901	28.37	4097	61.15	3810	56.87	6438	96.09	3533	52.73	3484	52.00	3711	55.39
0.6	4223	63.03	5443	81.24	1901	28.37	4097	61.15	2056	30.69	6438	96.09	3533	52.73	3484	52.00	3711	55.39
0.7	4223	63.03	5443	81.24	1901	28.37	4097	61.15	2056	30.69	6438	96.09	2113	31.54	1301	19.42	3711	55.39
0.8	2531	37.78	3846	57.40	1901	28.37	4097	61.15	729	10.88	5676	84.72	2113	31.54	1301	19.42	1505	22.46

Table C.3: Number of edges (#E) and average degree (AD), for each group and threshold value (T), for the IMPACT-HTA dataset.

T	A		B		C		D		E		F	
	#E	AD	#E	AD	#E	AD	#E	AD	#E	AD	#E	AD
0.4	6659	87.05	10182	133.10	10557	138.00	8583	112.20	5035	65.82	6450	84.31
0.5	6659	87.05	10182	133.10	6875	89.87	6380	83.40	5035	65.82	3241	42.37
0.6	3938	51.48	8167	106.76	6875	89.87	6380	83.40	5035	65.82	3241	42.37
0.7	3938	51.48	8167	106.76	6875	89.87	6380	83.40	1514	19.79	3241	42.37
0.8	1551	20.27	5202	68.00	6875	89.87	4041	52.82	1514	19.79	776	10.14

Table C.4: Number of obtained communities (#C), partition modularity (PM) and number of communities with only one element (#1), for the MEDI-VALUE IMD dataset.

T	A			B			C			D			E			F			G			H			I		
	#C	PM	# 1	#C	PM	# 1	#C	PM	# 1	#C	PM	# 1	#C	PM	# 1	#C	PM	# 1	#C	PM	# 1	#C	PM	# 1	#C	PM	# 1
0.4	3	0.048	0	3	0.041	0	4	0.258	0	3	0.141	0	4	0.206	0	3	0.012	0	4	0.091	0	4	0.262	0	4	0.141	0
0.5	4	0.062	1	3	0.055	0	11	0.642	2	9	0.133	4	4	0.206	0	3	0.015	0	4	0.121	0	4	0.262	0	5	0.247	2
0.6	4	0.062	1	3	0.055	0	11	0.642	2	9	0.133	4	6	0.336	1	3	0.015	0	7	0.161	2	4	0.262	0	5	0.247	2
0.7	4	0.062	1	3	0.055	0	11	0.642	2	9	0.133	4	6	0.336	1	3	0.015	0	7	0.161	2	24	0.622	10	5	0.247	2
0.8	12	0.121	4	12	0.091	7	11	0.642	2	9	0.133	4	40	0.632	18	7	0.016	2	13	0.278	7	24	0.622	10	23	0.608	10

Table C.5: Number of obtained communities (#C), partition modularity (PM) and number of communities with only one element (#1), for the MEDI-VALUE BBIVT dataset.

T	A			B			C			D			E			F			G			H			I		
	#C	PM	# 1	#C	PM	# 1	#C	PM	# 1	#C	PM	# 1	#C	PM	# 1	#C	PM	# 1	#C	PM	# 1	#C	PM	# 1	#C	PM	# 1
0.4	3	0.108	0	3	0.062	0	3	0.265	0	3	0.133	0	3	0.254	0	3	0.034	0	3	0.140	0	3	0.279	0	5	0.145	1
0.5	4	0.143	0	10	0.063	5	12	0.585	5	8	0.125	1	3	0.254	0	3	0.036	0	5	0.194	1	3	0.279	0	6	0.247	2
0.6	4	0.143	0	10	0.063	5	12	0.585	5	8	0.125	1	8	0.374	4	3	0.036	0	5	0.194	1	3	0.279	0	6	0.247	2
0.7	4	0.143	0	10	0.063	5	12	0.585	5	8	0.125	1	8	0.374	4	3	0.036	0	10	0.283	4	22	0.555	9	6	0.247	2
0.8	25	0.186	17	19	0.096	13	12	0.585	5	8	0.125	1	36	0.650	20	12	0.038	9	10	0.283	4	22	0.555	9	22	0.531	10

Table C.6: Number of obtained communities (#C), partition modularity (PM) and number of communities with only one element (#1), for the IMPACT-HTA dataset.

T	A			B			C			D			E			F		
	#C	PM	# 1	#C	PM	# 1	#C	PM	# 1	#C	PM	# 1	#C	PM	# 1	#C	PM	# 1
0.4	4	0.135	0	2	0.046	0	2	0.079	0	3	0.102	0	5	0.298	0	4	0.187	0
0.5	4	0.135	0	2	0.046	0	5	0.119	2	5	0.128	1	5	0.298	0	8	0.323	3
0.6	10	0.213	6	3	0.058	0	5	0.119	2	5	0.128	1	5	0.298	0	8	0.323	3
0.7	10	0.213	6	3	0.058	0	5	0.119	2	5	0.128	1	18	0.725	7	8	0.323	3
0.8	42	0.317	26	17	0.057	8	5	0.119	2	18	0.141	7	18	0.725	7	39	0.793	20

C.2 Results for a threshold of 0.6

This section presents full results regarding the groups of aspects, for a threshold of 0.6.

Table C.7: Attribute assortativity coefficient for IMPACT-HTA and MEDI-VALUE groups of aspects.

Group	MEDI-VALUE IMD	MEDI-VALUE BBIVT *in vitro* tests	IMPACT-HTA
A	−0.009 081 623	−0.014 183 277	−0.007 691 685
B	−0.008 387 004	−0.011 462 678	−0.007 148 206
C	−0.011 360 801	−0.009 433 565	−0.000 962 900
D	−0.011 562 277	−0.009 319 680	−0.007 020 703
E	0.008 578 288	−0.022 345 915	0.002 246 644
F	−0.008 113 199	−0.010 572 786	−0.002 425 718
G	−0.012 255 481	−0.009 010 410	-
H	−0.010 609 212	−0.012 615 181	-
I	−0.008 623 316	−0.011 456 965	-

MEDI-VALUE, Group B - ”Safety for the patient and/or healthcare professional”

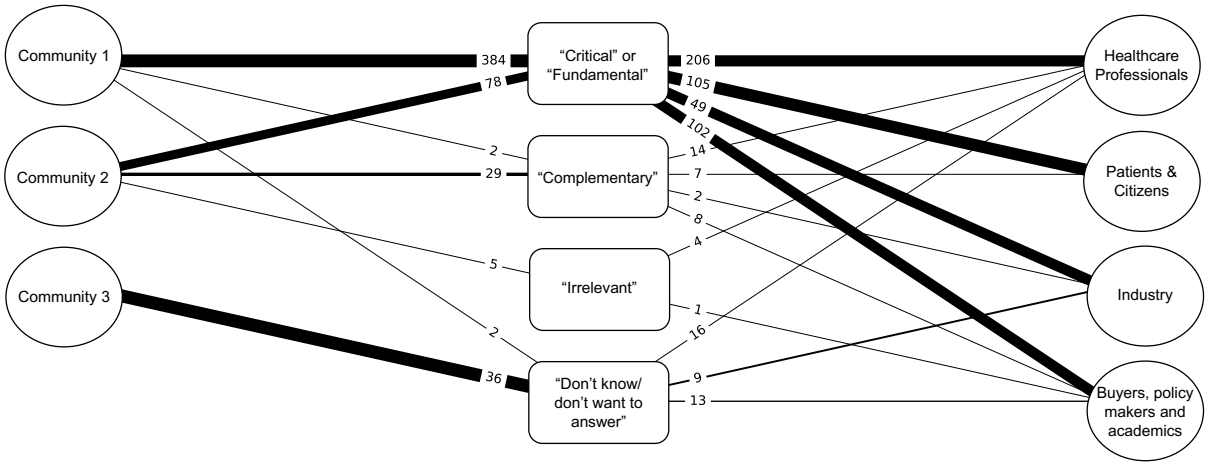
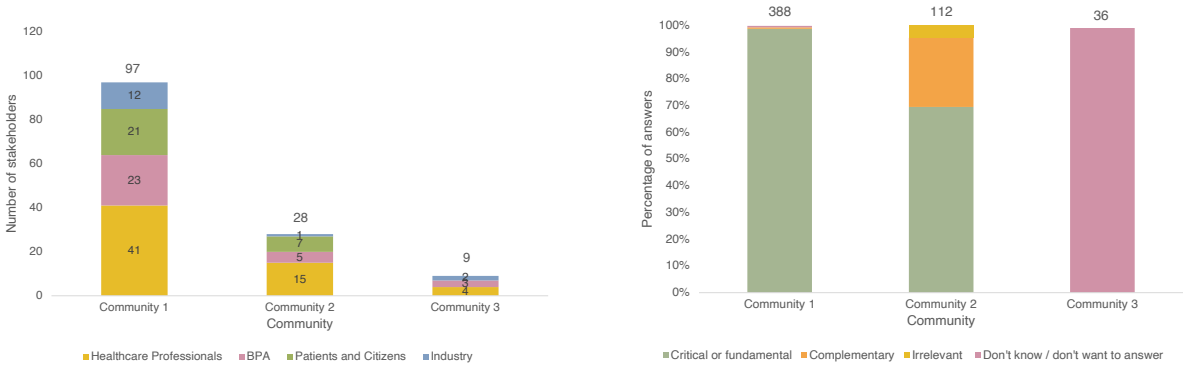


Figure C.1: Multipartite network, regarding MEDI-VALUE’s Group B - ”Safety for the patient and/or healthcare professional” (IMD), showing the distribution of answers across communities and stakeholders groups. The value in each line corresponds to the total number of answers, not stakeholders. Note that Group B comprises 4 aspects.



(a) Stakeholder types' distribution.

(b) Type of answers' distribution

Figure C.2: Distribution of stakeholders' type and answers, per community, regarding MEDI-VALUE's Group B - "Safety for the patient and/or healthcare professional" (IMD). In (a) numbers regard stakeholders and in (b) answers.

MEDI-VALUE, Group D - "Costs with the use of the medical device"

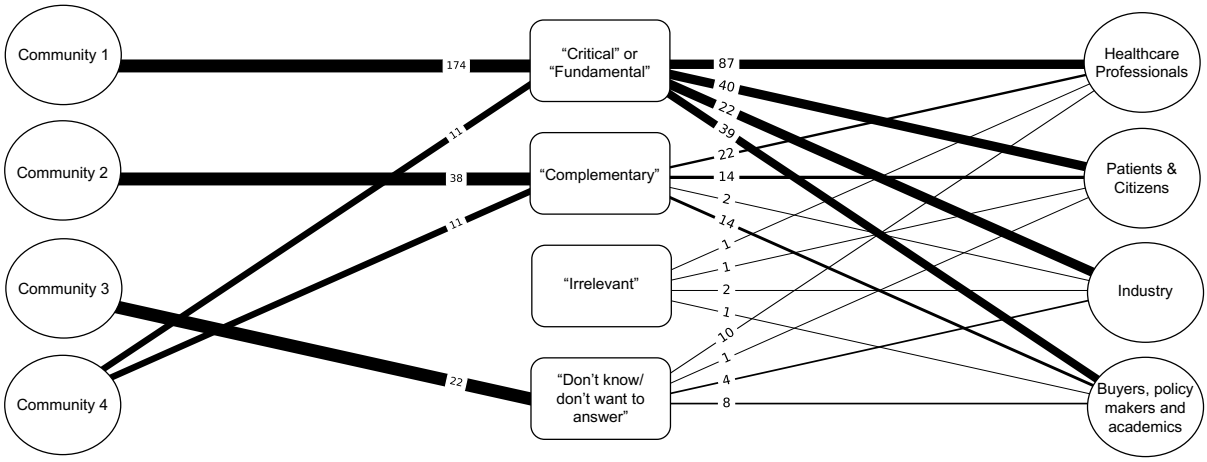
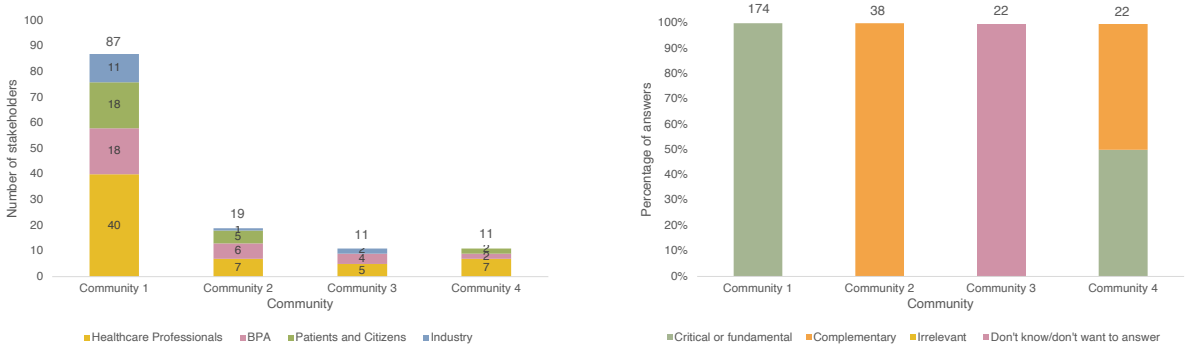


Figure C.3: Multipartite network, regarding MEDI-VALUE's Group D - "Costs with the use of the medical device" (IMD), showing the distribution of answers across communities and stakeholder groups. The value in each line corresponds to the total number of answers, not stakeholders. Note that Group D comprises 2 aspects.



(a) Stakeholder types' distribution.

(b) Type of answers' distribution

Figure C.4: Distribution of stakeholders' type and answers, per community, regarding MEDI-VALUE's Group D - "Costs with the use of the medical device" (IMD). In (a) numbers regard stakeholders and in (b) answers.

Table C.8: Distribution of stakeholders per community (C), per type, for both projects, for groups of aspects A to E.

G	MEDI-VALUE IMD					MEDI-VALUE BBIVT *in vitro* tests					IMPACT-HTA						
	C	% HPro	% BPA	% PC	% I	C	% HPro	% BPA	% PC	% I	C	% HTA	% HPro	% PC	% I	% P	% SE
A	C1	68.3	71.0	82.1	60.0	C1	55.0	51.6	64.3	53.3	C1	46.4	47.6	31.6	27.6	55.6	39.5
	C2	25.0	19.4	17.9	20.0	C2	25,0	22,6	21,4	26,7	C2	28.6	33.3	57.9	55.2	27.8	21.1
	C3	6.7	9.7	0.0	13.3	C3	10.0	16.1	10.7	13.3	C3	14.3	14.3	5.3	6.9	11.1	21.1
	C4	0.0	0.0	0.0	6.7	C4	10.0	9.7	3.6	6.7	C4	7.1	4.8	5.3	10.3	5.6	5.3
											C5	0.0	0.0	0.0	0.0	0.0	2.6
											C6	0.0	0.0	0.0	0.0	0.0	2.6
											C7	3.6	0.0	0.0	0.0	0.0	0.0
											C8	0.0	0.0	0.0	0.0	0.0	2.6
											C9	0.0	0.0	0.0	0.0	0.0	2.6
											C10	0.0	0.0	0.0	0.0	0.0	2.6
B	C1	68.3	74.2	75.0	80.0	C1	70.0	58.1	71.4	86.7	C1	71.4	76.2	63.2	69.0	66.7	73.7
	C2	25.0	16.1	25.0	6.7	C2	11.7	16.1	10.7	13.3	C2	21.4	19.0	21.1	27.6	33.3	18.4
	C3	6.7	9.7	0.0	13.3	C3	11.7	16.1	10.7	0.0	C3	7.1	4.8	15.8	3.4	0.0	7.9
						C4	1.7	0.0	3.6	0.0							
						C5	3.3	0.0	0.0	0.0							
						C6	0.0	3.2	0.0	0.0							
						C7	1.7	0.0	0.0	0.0							
						C8	0.0	0.0	3.6	0.0							
						C9	0.0	3.2	0.0	0.0							
						C10	0.0	3.2	0.0	0.0							
C	C1	26.7	29.0	46.4	40.0	C1	33.3	29.0	42.9	46.7	C1	53.6	85.7	78.9	69.0	83.3	81.6
	C2	26.7	16.1	25.0	13.3	C2	28.3	22.6	10.7	20.0	C2	39.3	9.5	21.1	31.0	0.0	7.9
	C3	15.0	19.4	14.3	6.7	C3	8.3	16.1	21.4	6.7	C3	3.6	4.8	0.0	0.0	11.1	10.5
	C4	18.3	6.5	14.3	13.3	C4	10.0	16.1	10.7	13.3	C4	3.6	0.0	0.0	0.0	0.0	0.0
	C5	6.7	9.7	0.0	20.0	C5	13.3	3.2	10.7	6.7	C5	0.0	0.0	0.0	0.0	5.6	0.0
	C6	1.7	6.5	0.0	0.0	C6	1.7	6.5	0.0	0.0							
	C7	0.0	3.2	0.0	6.7	C7	1.7	0.0	3.6	0.0							
	C8	3.3	0.0	0.0	0.0	C8	0.0	3.2	0.0	0.0							
	C9	1.7	3.2	0.0	0.0	C9	1.7	0.0	0.0	0.0							
	C10	0.0	3.2	0.0	0.0	C10	0.0	0.0	0.0	6.7							
	C11	0.0	3.2	0.0	0.0	C11	1.7	0.0	0.0	0.0							
						C12	0.0	3.2	0.0	0.0							
D	C1	66.7	58.1	64.3	73.3	C1	68.3	58.1	64.3	73.3	C1	53.6	66.7	73.7	41.4	55.6	57.9
	C2	11.7	19.4	17.9	6.7	C2	11.7	16.1	10.7	13.3	C2	32.1	28.6	21.1	37.9	27.8	23.7
	C3	8.3	12.9	0.0	13.3	C3	6.7	16.1	21.4	6.7	C3	7.1	4.8	0.0	17.2	0.0	10.5
	C4	11.7	6.5	7.1	0.0	C4	6.7	0.0	0.0	0.0	C4	7.1	0.0	5.3	3.4	16.7	5.3
	C5	1.7	0.0	3.6	0.0	C5	0.0	6.5	0.0	6.7	C5	0.0	0.0	0.0	0.0	0.0	2.6
	C6	0.0	0.0	3.6	0.0	C6	3.3	3.2	0.0	0.0							
	C7	0.0	0.0	0.0	6.7	C7	3.3	0.0	0.0	0.0							
	C8	0.0	0.0	3.6	0.0	C8	0.0	0.0	3.6	0.0							
	C9	0.0	3.2	0.0	0.0												
E	C1	41.7	35.5	32.1	40.0	C1	43.3	25.8	50.0	40.0	C1	28.6	14.3	15.8	44.8	66.7	28.9
	C2	40.0	19.4	35.7	26.7	C2	33.3	22.6	28.6	33.3	C2	32.1	28.6	31.6	17.2	16.7	55.3
	C3	11.7	29.0	32.1	13.3	C3	11.7	29.0	10.7	6.7	C3	35.7	57.1	42.1	27.6	16.7	15.8
	C4	6.7	9.7	0.0	13.3	C4	10.0	16.1	10.7	13.3	C4	3.6	0.0	5.3	6.9	0.0	0.0
	C5	0.0	6,5	0.0	0.0	C5	0.0	0.0	0.0	6,7	C5	0.0	0.0	5.3	3.4	0.0	0.0
	C6	0.0	0.0	0.0	6.7	C6	0.0	3.2	0.0	0.0							
						C7	1.7	0.0	0.0	0.0							
						C8	0.0	3.2	0.0	0.0							

Table C.9: Distribution of stakeholders per community (C), per type, for both projects, for groups of aspects F to I.

G	MEDI-VALUE IMD					MEDI-VALUE BBIVT *in vitro* tests					IMPACT-HTA						
	C	% HPro	% BPA	% PC	% I	C	% HPro	% BPA	% PC	% I	C	% HTA	% HPro	% PC	% I	% P	% SE
F	C1	85.0	87.1	96.4	73.3	C1	85.0	77.4	82.1	86.7	C1	46.4	19.0	36.8	20.7	44.4	31.6
	C2	8.3	3.2	3.6	13.3	C2	10.0	12.9	10.7	13.3	C2	25.0	47.6	26.3	20.7	22.2	39.5
	C3	6.7	9.7	0.0	13.3	C3	5.0	9.7	7.1	0.0	C3	21.4	23.8	21.1	31.0	22.2	18.4
											C4	3.6	9.5	5.3	27.6	5.6	2.6
											C5	0.0	0.0	5.3	0.0	0.0	7.9
											C6	0.0	0.0	5.3	0.0	0.0	0.0
											C7	3.6	0.0	0.0	0.0	0.0	0.0
											C8	0.0	0.0	0.0	0.0	5.6	0.0
G	C1	38.3	25.8	32.1	33.3	C1	40.0	29.0	32.1	13.3							
	C2	23.3	29.0	35.7	26.7	C2	30.0	32.3	25.0	60.0							
	C3	20.0	25.8	14.3	20.0	C3	20.0	22.6	28.6	13.3							
	C4	11.7	6.5	17.9	0.0	C4	10.0	16.1	10.7	13.3							
	C5	6.7	9.7	0.0	13.3	C5	0.0	0.0	3.6	0.0							
	C6	0.0	0.0	0.0	6.7												
	C7	0.0	3.2	0.0	0.0												
H	C1	36.7	35.5	32.1	13.3	C1	50.0	48.4	50.0	26.7							
	C2	33.3	22.6	35.7	40.0	C2	40.0	35.5	39.3	60.0							
	C3	23.3	32.3	32.1	26.7	C3	10.0	16.1	10.7	13.3							
	C4	6.7	9.7	0.0	20.0												
I	C1	56.7	45.2	53.6	53.3	C1	41.7	51.6	39.3	60.0							
	C2	35.0	41.9	46.4	33.3	C2	41.7	29.0	50.0	26.7							
	C3	6.7	9.7	0.0	13.3	C3	10.0	16.1	10.7	13.3							
	C4	0.0	3.2	0.0	0.0	C4	5.0	0.0	0.0	0.0							
	C5	1.7	0.0	0.0	0.0	C5	0.0	3.2	0.0	0.0							
						C6	1.7	0.0	0.0	0.0							

Table C.10: Number of intra (# Intra CE) and inter-community (# Inter CE) edges, ratio and coverage, for both projects, for each community (C) for groups of aspects A to C.

Group	MEDI-VALUE IMD					MEDI-VALUE BBIVT *in vitro* tests					IMPACT-HTA				
	C	# Intra CE	# Inter CE	Ratio	Coverage %	C	# Intra CE	# Inter CE	Ratio	Coverage %	C	# Intra CE	# Inter CE	Ratio	Coverage %
A	Total	4791	3540	0.739	-	Total	3174	2098	0.661	-	Total	2573	2730	1.061	-
	C1	4430	1605	0.362	69.26	C1	2696	1046	0.388	63.84	C1	1753	1215	0.693	44.51
	C2	325	1605	4.938	5.08	C2	343	1049	3.058	8.12	C2	715	874	1.222	18.16
	C3	36	0	0.000	0.56	C3	120	0	0.000	2.84	C3	91	634	6.967	2.31
	C4	0	0	-	0	C4	15	3	0.200	0.36	C4	14	7	0.5	0.36
											C5	0	0	-	0
											C6	0	0	-	0
											C7	0	0	-	0
											C8	0	0	-	0
											C9	0	0	-	0
											C10	0	0	-	0
B	Total	5007	3540	0.707	-	Total	4408	2070	0.470	-	Total	6062	4210	0.694	-
	C1	4652	1770	0.380	68.64	C1	4177	1035	0.248	76.74	C1	5666	2036	0.359	69.38
	C2	319	1770	5.549	4.71	C2	136	0	0.000	2.50	C2	359	1386	3.861	4.40
	C3	36	0	0.000	0.53	C3	93	1032	11.097	1.71	C3	37	788	21.297	0.45
						C4	1	0	0.000	0.02					
						C5	1	3	3	0.02					
						C6	0	0	-	0					
						C7	0	0	-	0					
						C8	0	0	-	0					
						C9	0	0	-	0					
						C10	0	0	-	0					
C	Total	1793	0	0.000	-	Total	1901	0	0.000	-	Total	6875	0	0.000	-
	C1	946	0	0.000	52.76	C1	1128	0	0.000	59.34	C1	6441	0	0.000	93.69
	C2	435	0	0.000	24.26	C2	435	0	0.000	22.88	C2	406	0	0.000	5.91
	C3	190	0	0.000	10.60	C3	136	0	0.000	7.15	C3	28	0	0.000	0.41
	C4	171	0	0.000	9.54	C4	120	0	0.000	6.31	C4	0	0	-	0
	C5	45	0	0.000	2.51	C5	78	0	0.000	4.10	C5	0	0	-	0
	C6	3	0	0.000	0.17	C6	3	0	0.000	0.16					
	C7	1	0	0.000	0.06	C7	1	0	0.000	0.05					
	C8	1	0	0.000	0.06	C8	0	0	-	0					
	C9	1	0	0.000	0.06	C9	0	0	-	0					
	C10	0	0	-	0	C10	0	0	-	0					
	C11	0	0	-	0	C11	0	0	-	0					
						C12	0	0	-	0					

Table C.11: Number of intra (# Intra CE) and inter-community (# Inter CE) edges, ratio and coverage, for both projects, for each community (C) for groups of aspects D to G.

Group	MEDI-VALUE IMD					MEDI-VALUE BBIVT *in vitro* tests					IMPACT-HTA				
	C	# Intra CE	# Inter CE	Ratio	Coverage %	C	# Intra CE	# Inter CE	Ratio	Coverage %	C	# Intra CE	# Inter CE	Ratio	Coverage %
D	Total	4023	0	0.000	-	Total	4097	0	0.000	-	Total	4459	3842	0.862	-
	C1	3741	0	0.000	92.99	C1	3828	0	0.000	93.43	C1	3741	1827	0.488	58.64
	C2	171	0	0.000	4.25	C2	136	0	0.000	3.32	C2	645	1573	2.439	10.11
	C3	55	0	0.000	1.37	C3	120	0	0.000	2.93	C3	50	26	0.52	0.78
	C4	55	0	0.000	1.37	C4	6	0	0.000	0.15	C4	23	416	18.087	0.36
	C5	1	0	0.000	0.02	C5	3	0	0.000	0.07	C5	1	0	0.000	0
	C6	0	0	-	0	C6	3	0	0.000	0.07					
	C7	0	0	-	0	C7	1	0	0.000	0.02					
	C8	0	0	-	0	C8	0	0	-	0					
	C9	0	0	-	0										
E	Total	1381	1156	0.837	-	Total	1409	1294	0.918	-	Total	3012	4046	1.343	-
	C1	341	264	0.774	17.41	C1	422	266	0.630	20.53	C1	1209	1781	1.473	24.01
	C2	813	454	0.558	41.5	C2	708	495	0.699	34.44	C2	729	897	1.230	14.48
	C3	190	438	2.305	9.70	C3	159	533	3.352	7.73	C3	1069	1365	1.277	21.23
	C4	36	0	0.000	1.84	C4	120	0	0.000	5.84	C4	4	3	0.75	0.08
	C5	1	0	0.000	0.05	C5	0	0	-	0	C5	1	0	0.000	0.02
	C6	0	0	-	0	C6	0	0	-	0					
						C7	0	0	-	0					
						C8	0	0	-	0					
F	Total	6630	1344	0.203	-	Total	6012	852	0.142	-	Total	2043	2396	1.173	-
	C1	6558	672	0.102	89.81	C1	5885	426	0.072	91.41	C1	615	623	1.013	18.98
	C2	36	672	18.667	0.49	C2	105	0	0.000	1.63	C2	860	984	1.144	26.54
	C3	36	0	0.000	0.49	C3	22	426	19.364	0.34	C3	535	747	1.396	16.51
											C4	27	42	1.556	0.83
											C5	6	0	0.000	0.19
											C6	0	0	-	0
											C7	0	0	-	0
											C8	0	0	-	0
G	Total	1729	3404	1.969	-	Total	1960	3146	1.605	-					
	C1	878	1432	1.631	25.59	C1	817	1430	1.750	23.12					
	C2	630	1230	1.952	18.36	C2	893	1314	1.471	25.28					
	C3	113	209	1.850	3.29	C3	130	402	3.092	3.68					
	C4	72	533	7.403	2.10	C4	120	0	0.000	3.40					
	C5	36	0	0.000	1.05	C5	0	0	-	0					
	C6	0	0	-	0										
	C7	0	0	-	0										

Table C.12: Number of intra (# Intra CE) and inter-community (# Inter CE) edges, ratio and coverage, for both projects, for each community (C) for groups of aspects H to I.

Group	MEDI-VALUE IMD					MEDI-VALUE BBIVT *in vitro* tests					IMPACT-HTA				
	C	# Intra CE	# Inter CE	Ratio	Coverage %	C	# Intra CE	# Inter CE	Ratio	Coverage %	C	# Intra CE	# Inter CE	Ratio	Coverage %
H	Total	2066	3116	1.508	-	Total	2450	2068	0.844	-					
	C1	888	1121	1.262	24.50	C1	975	1034	1.061	27.99					
	C2	777	1503	1.934	21.44	C2	1355	1034	0.763	38.89					
	C3	356	491	1.379	9.82	C3	120	0	0.000	3.44					
	C4	45	1	0.022	1.24										
I	Total	2603	2150	0.826	-	Total	2573	2276	0.885	-					
	C1	1357	1075	0.792	36.90	C1	1537	1134	0.738	41.42					
	C2	1210	1075	0.888	32.90	C2	913	1136	1.244	24.60					
	C3	36	0	0.000	0.98	C3	120	0	0.000	3.23					
	C4	0	0	-	0	C4	3	6	2.0	0.08					
	C5	0	0	-	0	C5	0	0	-	0					
						C6	0	0	-	0					

Table C.13: Transitivity, average clustering coefficient and average degree for IMPACT-HTA and MEDI-VALUE groups of aspects.

Group	MEDI-VALUE IMD			MEDI-VALUE BBIVT			IMPACT-HTA		
	Transitivity	Average Clustering Coef.	Avg Degree	Transitivity	Average Clustering Coef.	Avg Degree	Transitivity	Average Clustering Coef.	Avg Degree
A	0.947 69	0.924 56	95.462 70	0.931 92	0.844 95	63.029 90	0.809 60	0.68170	51.477 10
B	0.972 94	0.966 90	101.149 30	0.981 95	0.917 96	81.238 80	0.949 64	0.897 00	106.758 20
C	1.000 00	0.940 30	26.760 00	1.000 00	0.947 76	28.370 00	1.000 00	0.986 93	89.870 00
D	1.000 00	0.955 22	60.040 00	1.000 00	0.977 61	61.150 00	0.937 84	0.872 40	89.870 00
E	0.68601	0.60065	29.238 80	0.741 36	0.68129	30.686 60	0.730 66	0.701 90	65.820 00
F	0.994 94	0.995 40	108.985 10	0.994 67	0.974 61	96.089 60	0.66501	0.61782	42.366 00
G	0.812 97	0.777 71	51.209 00	0.839 51	0.808 88	52.731 30	-	-	-
H	0.782 71	0.762 34	54.090 00	0.804 79	0.778 46	52.000 00	-	-	-
I	0.748 37	0.729 25	54.895 50	0.812 62	0.789 49	55.388 10	-	-	-