Department of Management Science and Technology
Master of Science in Business Analytics

Part Time Class of 2016

Diploma Thesis

Data Analytics and Research Assessment: A Case Study in Health and the Diseases of the Circulatory System

Author: Markou Eleni (BAPT1604)

Project team: Markou Eleni (BAPT1604)
Klironomou Maria-Anna (BAPT1608)

Supervisor: Papageorgiou Haris,
Research Director at ATHENA RC /
Institute for Language & Speech Processing

Athens, January 2019

Table of Contents

List of Tables

List of Figures

1. Introduction

This diploma thesis was developed under the auspices of Data for Impact project, which received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 770531 ("Home - Data4Impact" n.d.). The project responded to European Commission's Health, Demographic Change and Wellbeing Societal Challenge ("Health, Demographic Change and Wellbeing - Horizon 2020 - European Commission" 2018).

Data4Impact aims to capitalise on recent technological developments in data mining, data treatment and data analysis that offer new dimensions and opportunities for performance analytics. With the introduction of new technologies and initiatives, such as open access mechanisms and social media/online media, increasing volumes of new data on the research domain are being generated and, thus, big data approaches can be utilised to improve the monitoring of research and innovation performance and assessment of the societal impact. The main objectives of the project are ("About - Data4Impact" n.d.):

- Define, develop, analyse and disseminate new indicators for assessing the performance of EU and national research and innovation systems;
- Explore and collect "big" data on health-related societal challenges at input, throughput, output/result and impact levels;
- Employ big data approaches to yield more data on the societal impact of national and EU funding on tackling health-related societal challenges;
- Engage stakeholders in the project activities, validate the project results and develop new indicators and tools using a hands-on approach.

In this context, this diploma thesis attempts to employ a network approach on the matter and develop a multi-layer graph infrastructure in order to assess the societal impact of health-related research in Europe for the past 10 years. For this purpose, information has been harnessed from a variety of projects funded under the European Union's Research and Innovation funding programmes FP7 ("Home Page - FP7 - Research - Europa" n.d.) and Horizon 2020 ("Horizon 2020 - European Commission" 2018).

An overview of the structure and the contents of this document can be found below:

- Chapter "2. Graphs and Networks" contains a high level review of the basic definitions and terminologies related to Graph theory and Network analysis. A generic classification of different types of networks, as well as commonly used metrics and computational algorithms are also presented.

- Chapter "3. Software Overview" provides a synopsis of the main software tools that were employed during development, by briefly listing their key features, capabilities and limitations.

- Chapter "4. Dataset Description" is dedicated to the description of all collection steps and extraction techniques that were followed in order to construct the corpus of the final dataset to be analyzed. For this task, alternative sources of information available per project, such as summary, metadata and related documents, have been used.

- Chapter "5. Scenario Definition and Dataset Segmentation" introduces the business scenario and the set of fundamental assumptions that dictated all aspects of the performed analysis. Moreover, an explanation of the dataset breakdown into subsets according to the International Statistical Classification of Diseases and Related Health Problems ("WHO | International Classification of Diseases, 11th Revision (ICD-11)" 2018) takes place.

- Chapter "6. Technical Implementation" describes in detail the complete process of core code development and manipulation of produced results, from dataset parsing and graph creation up to network metrics computation and generation of visualizations.

- Chapter "7. Presentation of Results" comprises of the derived outcome and corresponding interpretation of the aforementioned implementation, i.e. an ensemble of tables and illustrations, along with illuminative commentary, that depict the performed analysis in aggregate.

- Chapter "8. Conclusions and Future Work" summarizes the drawn conclusions of this diploma thesis, along with suggestions for enhancements and future work.

2. Graphs and Networks

2.1 Definitions

When representing actual networks as graphs, the results are sets of points joined with lines, showing patterns of interconnections among entities. In other words, a **graph G = (N, E)** is a representation of relationships among a collection of items, consisting of two distinct sets of objects: a set of objects (**N**) called **nodes** and a set of objects (**E**) called **edges**.

Nodes (**N = {1, … , n}**) are also referred to as **vertices**, **individuals**, **agents** or **players** depending on the context of the network that is being represented. It is important to note that nodes might be all kinds of entities such as individual people, firms, countries, or other organizations; or a node might even be something like a web page or a document belonging to some person or organization (Easley and Kleinberg 2010a).

Edges (also referred to as **arcs**, **bonds** or **ties**) are links connecting pairs of nodes, thus indicating the existence of a certain type of relationship between them, such as friendship, partnership, alliance, affiliation, communication, information exchange, cross-reference and many others. Hence, two nodes are called **neighbors** if they are connected by an edge (Easley and Kleinberg 2010a).

The relationship between two neighbors, i.e. the two ends of an edge, can be either symmetric (reciprocal) or asymmetric. In the first case, the edge simply connects them to each other and the orientation is unimportant - these are called **undirected** graphs. In the second case, the direction of the edge is important, for example node A points to node B but not vice versa - these are called **directed** graphs. It depends on the nature of the network that is being represented whether directed or undirected edges should be used (Jackson 2010).

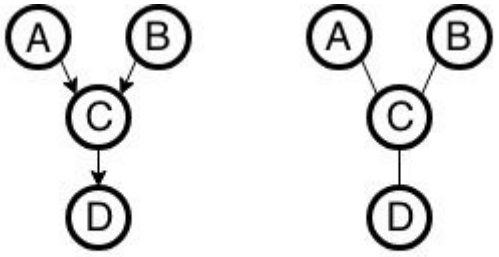


Image 1: Examples of directed and undirected graphs of 4 nodes

A graph can be modelled via a $n \times n$ matrix **A**, where element Aij represents the (possibly weighted and/or directed) relation between node i and node j. This matrix is called **adjacency matrix**, as it displays which nodes are adjacent to one another, or in other words which nodes are neighbors.

In the case where the entries of A take on various values from -1 to 1, thus indicating the intensity of relationships, the graph is referred to as a **weighted** graph. Otherwise,

values of either 0 or 1 are used, and the graph is **unweighted**. It must be noted that in case that Aij is not necessarily equal to Aji then the graph is directed, while it is undirected when it is required that Aij = Aji for all nodes i and j (Easley and Kleinberg 2010a).

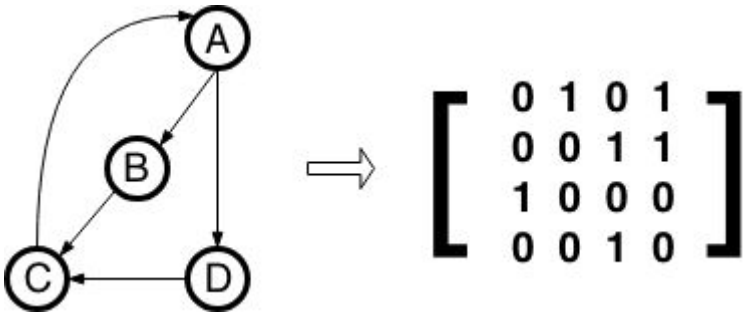


Image 2: Example of directed graph and its adjacency matrix

A **walk** can be defined as a sequence of nodes where each consecutive pair in the sequence is connected by an edge, i.e. the walk contains as well the sequence of edges connecting the sequence of nodes.

A walk as defined above can contain repeated nodes, while a **path** is a special kind of walk where each node is distinct, i.e. appears at most once in the sequence. Another important kind of walk is a **cycle**, which is a "ring" structure, i.e. a walk in which the first and the last node are the same, while all other remaining nodes are distinct (Jackson 2010).

A **geodesic** between two nodes is the shortest path between them, that is a path with no more edges than any other path between these nodes (also referred to simply as **shortest path**). It is important to mention that there may exist more that one equivalent shortest paths between the same pair of nodes.

The length of the shortest path between two nodes (defined as the number of edges for unweighted graphs or the sum of edge weights for weighted graphs) is usually called the **distance** between these nodes. Therefore, the **diameter** of a network is the largest distance between any two nodes, i.e. the largest shortest path in the network (Jackson 2010).

It should be mentioned that a network may not be fully connected, thus breaking apart into a number of separate **components**, i.e. connected groups/subsets of nodes so that no two groups overlap (Easley and Kleinberg 2010b).

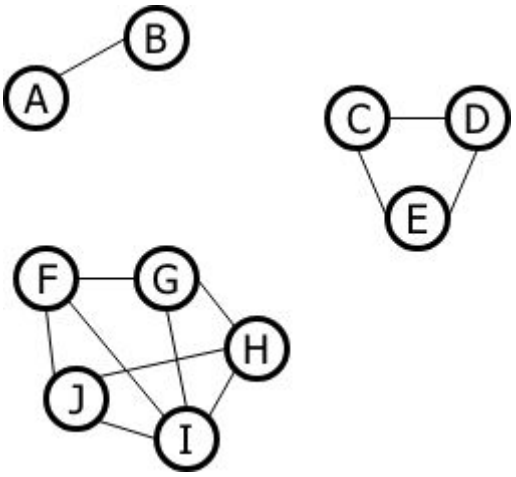


Image 3: Example of network with 3 separate components

2.2 Classification of networks

The study of complex phenomena through network modeling is a methodology that have been successfully applied in many areas, such as disease spreading to gene circuits (De Domenico et al. 2016), over the past few decades.

However, only recently it became evident that in order to go far beyond the current understanding, the fact that many real-world systems do not operate into isolation needs to be taken into consideration. Instead, they are interconnected and what happens on a single level of interaction can significantly affect another interconnected layer (Kivelä et al. 2014).

This realization led to the development of two main network subclasses: monoplex and multilayer networks. The first class represents more traditional networks that have been widely used until now while the second class aspires to accurately model more realistic phenomena.

2.2.1 Monoplex Networks

Monoplex networks is the most traditional type of networks which are completely described by just a set of entities (nodes) and their interactions (edges). In the most general form a monoplex network is mathematically represented as a graph **G = (N, E)**, following the definitions of the previous chapter (De Domenico et al. 2013).

As presented in 2.1 Definitions, this finite graph can be mathematically represented as an adjacency matrix, i.e. a square matrix the elements of which indicate whether pairs of vertices are adjacent or not in the graph. However, the adjacency matrix should be distinguished from the incidence matrix for a graph, a different matrix representation whose elements indicate whether vertex–edge pairs are incident or not, and degree matrix which contains information about the degree of each vertex (Jackson 2010).

Having defined the adjacency matrix of a monoplex network various concepts can be applied including node degree computations, centrality measures, diffusion processes and community detection algorithms in order to extract useful information from the structural properties of the network.

However, in many cases the initial assumption, i.e. the fact that the nodes are connected to each other by a single type of links while other types of interactions are not considered, can be an oversimplification of reality leading to erroneous results. Thus, in order to be able to represent systems consisting of networks with multiple types of links, the notion of multilayer structures was introduced.

2.2.2 Multilayer Networks

As multilayer networks are defined structures that are characterized by a number of layers apart from nodes and links. In the general case, nodes are organized into layers and any node present on any of the layers can be connected with any other node of any layer through a link (edge).

Each layer has a semantic that is associated with certain aspects or features that the nodes or the links belonging to it, exhibit. Thus, the links present in a multilayer network can be assorted in two categories: inter-layer edges and intra-layer edges.

As intra-layer edges are characterized those links that connect nodes belonging to the same layer. On the other hand, inter-layer notion refers to links developed between nodes that belong to different layers.

To formally define a multilayer network (Kivela et al. 2013) the definition of monoplex networks as **G = (N, E)** need to be extended to also include a set of **L = {L1,L2,...,Ln}** layers. Given that nodes are allowed to be absent in some layers, for each choice of a node and layers, an indication of whether the node is present in that layer is needed. To do so, the construction of a set **V x L1 x ... x Ln** of all these combinations is required so that a subset $V_M \subseteq V x L1 x .. x Ln$ is defined that contains only the node-layer combinations in which a node is present in the corresponding layer. This is often called node-layer tuple to indicate a node that exists on a specific layer.

Connections between pairs of node-layer tuples also need to be defined. All possible edges between any pair of node-layers - including ones in which a node is adjacent to a copy of itself in some other layer. Extending the definition of the set of edges in monoplex networks, when working with multilayer networks the starting and ending layers for edge need to be specified as well. An edge set $E_M \subseteq V_M x V_M$ of a multilayer network is thus defined as a set of pairs of possible combinations of nodes and layers.

Using the above components, a multilayer network can be defined a quadruplet $M = (V_M, E_M, V, L)$.

Regarding the mathematical representation of networks under the above definition, two main approaches have been proposed. According to the first approach developed by Domenico et al (De Domenico et al. 2013), multilayer networks can be mathematically

represented using intra- and inter-layer tensors, as a generalization of adjacency matrices used when working with monoplex networks. The main benefits in this case are two-fold: it gives concise mathematical representation, and it leads to natural generalizations of numerous network diagnostics from monoplex networks to multilayer networks (De Domenico et al. 2013).

An alternative approach (De Domenico et al. 2013, 2015), suggests the extension of the adjacency matrix representation to encode multilayer networks by building a block matrix where each diagonal block corresponds to the adjacency matrix of each layer while the non-diagonal blocks encode the corresponding couplings. This extended adjacency matrix is called supra-adjacency matrix and is constructed by basically flattening the high dimensional tensor into a rank-2 tensor of size NL x NL, where L indicates the number of layers present and the N the set of vertices.

Supra-adjacency matrices are better studied than tensors, and they are natural for representing multilayer networks that do not contain all nodes in all layers (De Domenico et al. 2013).

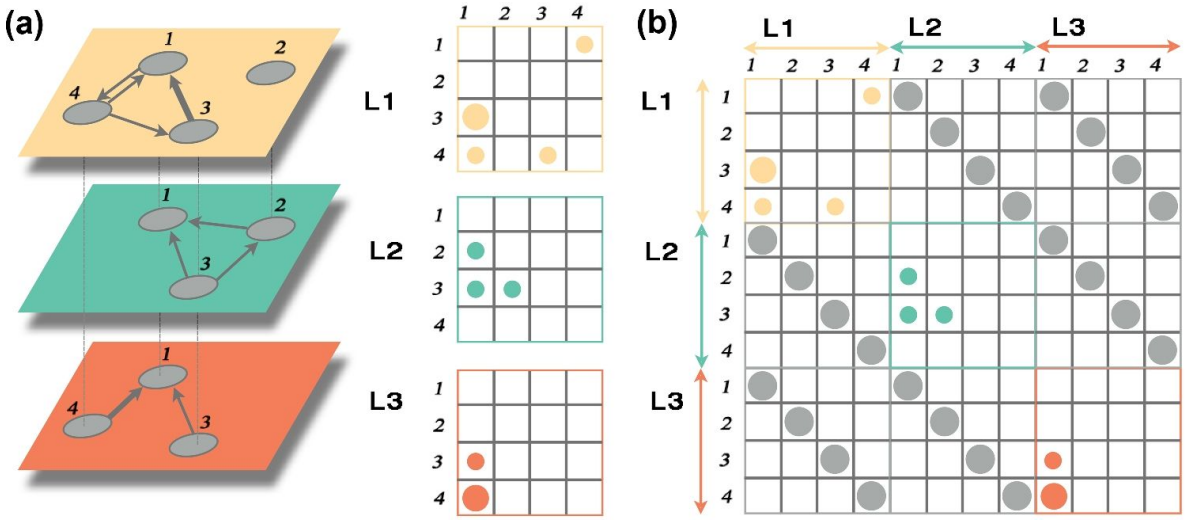


Image 4: Example of multilayer network and its supra-adjacency matrix

2.3 Metrics

2.3.1 Degree Centrality

Degree is a simple centrality measure that counts how many neighbors a node has. If the network is directed, two versions of the measure can be computed: in-degree is the number of incoming links, or the number of predecessor nodes; out-degree is the number of outgoing links, or the number of successor nodes. Typically, we are interested in in-degree, since in-links are given by other nodes in the network, while out-links are determined by the node itself ("Degree Centrality" n.d.), (Jackson 2010).

Thus, a node is important if it has many neighbors, or, in the directed case, if there are many other nodes that link to it, or if it links to many other nodes.

2.3.2 Betweenness Centrality

Betweenness centrality measures the extent to which a vertex lies on paths between other vertices. Vertices with high betweenness may have considerable influence within a network by virtue of their control over information passing between others. They are also the ones whose removal from the network will most disrupt communications between other vertices because they lie on the largest number of paths taken by messages ("Closeness Centrality" n.d., "Betweenness Centrality" n.d.).

Betweenness centrality differs from the other centrality measures. A vertex can have quite low degree, be connected to others that have low degree, even be a long way from others on average, and still have high betweenness.

2.3.3 Closeness Centrality

Closeness centrality measures the mean distance from a vertex to other vertices using the geodesic path, i.e. the shortest path through a network between two vertices. This quantity takes low values for vertices that are separated from others by only a short geodesic distance on average. Such vertices might have better access to information at other vertices or more direct influence on other vertices ("Closeness Centrality" n.d.).

In other words, closeness centrality is a measure of how easily a node can reach other nodes (i.e. how close the node is to the center of the network).

2.3.4 Farness Centrality

Farness centrality is very closely related to the closeness centrality that was previously presented. Farness centrality captures the variation of the shortest path distances of a vertex to every other vertex ("Farness Centrality" n.d.).

The farness centrality, or peripherality, of a node can be defined as the sum of its distances to all other nodes. In other words, farness centrality is the reciprocal of the closeness centrality.

2.3.5 Eigenvector Centrality

Eigenvector centrality, or EigenCentrality, is a metric that takes into consideration the number of links that a node has to other nodes within the same network. Compared to degree centrality, EigenCentrality goes a step further by also taking into account how well connected a node is, and how many links their connections have, and so on through the network ("Eigenvector Centrality" n.d.).

A high EigenCentrality score indicates a strong influence over other nodes in the network. It is useful because it indicates not just direct influence, but also implies influence over nodes more than one 'hop' away.

So, a node may have a high degree score (i.e. many connections) but a relatively low EigenCentrality score if many of those connections are with similarly low-scored nodes.

Also, a node may have a high betweenness score (indicating it connects disparate parts of a network) but a low EigenCentrality score because it is still some distance from the centers of power in the network.

2.3.6 PageRank

Pagerank is a centrality metric that takes into consideration the number and quality of the links of a given vertex in order to estimate how important this vertex is. The underlying assumption is that more important nodes are likely to receive more links from other nodes ("Facts about Google and Competition" n.d.).

Although the original implementation assumes a directed network, the algorithm can be modified not to consider link direction. Overall, PageRank can help uncover influential or important nodes whose reach extends beyond just their direct connections.

2.3.7 HITS

Similarly to PageRank, the Hyperlink-Induced Topic Search (HITS) algorithm is a link analysis algorithm which assigns a hub and an authority score to each node in the network. Nodes with high hub score serve as large directories that redirect to other important nodes (namely authorities), while nodes with high authority score are considered to be the ones holding important information and thus they are pointed by many hubs ("Hubs, Authorities, and Communities" n.d.).

2.3.8 Node Eccentricity

The eccentricity of a node in a graph is computed as the reciprocal of the maximum distance from this node to all other nodes in the network.

By doing that, an eccentricity with higher value assumes a positive meaning in term of node proximity. In contrast, if the eccentricity is low, this means that there is at least one node (and all its neighbors) that is far from node under examination. Of course, this does not exclude that several other nodes are much closer to the initial node.

Thus, eccentricity is a more meaningful parameter if it is high. Notably, "high" and "low" values are more significant when compared to the average eccentricity of the whole network calculated by averaging the eccentricity values of all nodes in the graph ("Eccentricity" n.d.).

2.4 Community Detection

In the study of complex networks, a network is said to have community structure if the nodes of the network can be easily grouped into sets of nodes, either overlapping or not, such that each set of nodes is densely connected internally. In the particular case of non-overlapping community finding, this implies that the network divides naturally into groups of nodes with dense connections internally and sparser connections between groups (Contributors to Wikimedia projects 2006).

A popular method, particularly suitable for large networks with millions of vertices and tens of millions of edges, is the Clauset - Newman - Moore, a hierarchical agglomerative algorithm which works by greedily optimizing the modularity. This algorithm runs in time O(md log n) for a network with n vertices and m edges where d is the depth of the dendrogram. For networks that are hierarchical, in the sense that there are communities at many scales and the dendrogram is roughly balanced, we have d ~ log n. If the network is also sparse, m ~ n, then the running time is essentially linear, O(n log2 n). This is considerably faster than most general algorithms (Clauset, Newman, and Moore 2004).

For assessing the quality of the generated communities, a structural measure named modularity is used. This metric is designed to measure the strength of division of a network into modules. Networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes in different modules.

More specifically, modularity is the fraction of the edges that fall within the given groups minus the expected fraction if edges were distributed at random. The value of the modularity lies in the range [-1, 1] (Li and Schuurmans 2011) and it is positive if the number of edges within groups exceeds the number expected on the basis of chance. For a given division of the network's vertices into some modules, modularity reflects the concentration of edges within modules compared with random distribution of links between all nodes regardless of modules (Contributors to Wikimedia projects 2008).

3. Software Overview

3.1 SNAP.py

The Stanford Network Analysis Platform (SNAP) is a general purpose network analysis and graph mining library. It is an open source library, written in C++ that easily scales to massive networks with hundreds of millions of nodes, and billions of edges. It efficiently manipulates large graphs, calculates structural properties, generates regular and random graphs, and supports attributes on nodes and edges (Leskovec and Sosič 2016).

SNAP has been developed for single big-memory machines and it balances the trade-off between maximum performance, compact in-memory graph representation, and the ability to handle dynamic graphs where nodes and edges are being added or removed over time.

In more detail, the key design concept of SNAP was that data structures need to be flexible in allowing for efficient manipulation of the underlying graph structure, i.e. adding or deleting nodes and edges must be reasonably fast and not prohibitively expensive. This requirement is needed, for example, for the processing of dynamic graphs, where graph structure is not known in advance, and nodes and edges get added and deleted over time. A related use scenario is motivated by on-line graph algorithms, where an algorithm incrementally modifies existing graphs as new input becomes available.

For the representation of graph and networks in SNAP, a middle ground was chosen between all-hash table and all-vector graph representations. A graph in SNAP is represented by a hash table of nodes in the graph. Each node consists of a unique identifier and one or two vectors of adjacent nodes, listing nodes that are connected to it. Only one vector is used in undirected graphs, while two vectors, one for outgoing and another one for incoming nodes/edges, are used in directed graphs.

The values in adjacency vectors are sorted for faster access. Since most of the real-world networks are sparse with node degrees significantly smaller than the number of nodes in the network, while at the same time exhibiting a power law distribution of node degrees, the benefits of maintaining the vectors in a sorted order significantly outweigh the overhead of sorting. Sorted vectors also allow for fast and ordered traversal and selection of node's neighbors, which are common operations in graph algorithms.

SNAP offers over 140 different graph algorithms. Among these, commonly used traditional algorithms for graph and network analysis are included, as well as recent algorithms that employ machine learning techniques on graph problems, such as community detection, statistical modeling of networks, network link and missing node prediction, random walks, network structure inference (Leskovec and Sosič 2016).

For the purpose of this diploma thesis Snap.py, a Python interface for SNAP, was used. Snap.py provides performance benefits of SNAP, combined with flexibility of Python. Most of the SNAP C++ functionality is also available via Snap.py ("Snap.py - SNAP for Python" n.d.).

3.2 Gephi

Gephi is an open-source network analysis and visualization software package written in Java on the NetBeans platform ("Gephi - The Open Graph Viz Platform" n.d.).

Initially developed in 2008 by UTC in France, Gephi has been successfully used in a number of research projects in academia, journalism and elsewhere. Gephi intuitively reveals patterns and trends and highlights outliers. It uses a 3D render engine to display large graphs in real-time and to speed up the exploration. This technique uses the computer graphic card, as video games do, and leaves the CPU free for other computing. It can deal with large network (i.e. over 20,000 nodes) and, because it is built on a multi-task model, it takes advantage of multi-core processors. Node design can be personalized, instead of a classical shape it can be a texture, a panel or a photo. Highly configurable layout algorithms can be run in real-time on the graph window ("Bastian" n.d.).

The user interface is structured into Workspaces, where separate work can be done, and a powerful plugin system is currently developed. Great attention has been taken to the extensibility of the software. An algorithm, filter or tool can be easily added to the program, with little programming experience. Sets of nodes or edges can be obtained manually or by using the filter system. Filters can select nodes or edges with thresholds, range and other properties.

Overall, Gephi supports loading and saving graphs in a number of traditional formats while other more comprehensive file formats which can store node and edge attributes, together with layout and presentation information (e.g. position, size, colour etc) can also be chosen.

4. Dataset Description

The construction of the final dataset corpus that is analyzed in this diploma thesis was part of Data4Impact Work Package 5 “Assessment of the societal impact of research related to the Health, Demographic Change and Wellbeing Societal Challenge” deliverables, and more specifically part of Task 5.2 “Analysis of monitoring data of finalised EU projects”.

This chapter contains a summary presentation of all collection steps and extraction techniques that were followed by the researchers. For more details, please refer to Data4Impact Deliverable 5.1 original document (“Project Deliverables - Data4Impact” n.d.).

4.1 Scope

The main goal of Task 5.2 was to monitor the EU funded projects, record the innovations that were produced in their context and measure the overall impact of the EU funding in respect of the EU policies, while attempting to reach to conclusions about how much money was spent by the EU on specific topics (in this case Health, Demographic Change and Wellbeing), and what was the impact that this funding had not only on the research and academic landscape but mainly on society.

More specifically, Task 5.2 was focused on two major EU funding programmes, namely FP7 and H2020, and examined different types of documents (such as the call, project description, final reports, results in brief, publications, patents) concerning the funded finalized and ongoing projects. Additionally, project related metadata (such as financial data about the cost of each project and the EU contribution along with the budget distribution per participant) were also taken into account, since they could lead to valuable conclusions. Moreover, data relevant to each organisation participating in the project (such as the country it is based and its type, i.e. whether it is a research organisation, a university or a company) were gathered and leveraged.

With an aspiring target to create the links between all different types of data and thus discover the actual impact of the EU-funded projects in the Health domain, the main objectives of this particular task were to:

- Analyse the content of available datasets in order to discover entities related to EU-funded projects, as well as use text analysis to discover terminology and concepts that constitute the aspirations of the EU.
- Analyse the content of available datasets in order to discover concepts depicting innovation related to Health, Demographic Change and Wellbeing and attribute them to certain types or categories like “method, technique etc.” (hereafter called insights).

- Discover clusters and communities concerning entities involved in EU-funded projects.
- Discover links and relations between projects, entities and insights.

In this context, the task was built upon the following three major stages:

1. Define impact and its major aspects and isolate particular sections within the projects data.
2. Automatically extract and track the respective insights, making use of text mining/Information Extraction modules built and/or modified for the specific purposes.
3. Build different graphs to depict the correlations between entities and insights and explore spatial/temporal trends and patterns delineating impact. This stage was directly related with the purpose of this diploma thesis.

Finally, it is important to note that disease mentions along with MeSH terms and ontology and other established disease taxonomies, like the International Statistical Classification of Diseases and Related Health Problems ("WHO | International Classification of Diseases, 11th Revision (ICD-11)" 2018), were leveraged across all stages to determine the subject issues referred in the official EU documents, the projects final reports and the subsequent publications.

4.2 Data Collection

As already mentioned, analysis was performed at project level, hence the analysis unit is the project. The data gathered and organized for each finalized and ongoing EU project in the field of Health, Demographic Change and Wellbeing (funded within either of the FP7 and H2020 frameworks) were:

- **Call document**, i.e. the text of the FP7/H2020 Call and its metadata
- **Project description,** i.e. the original text along with metadata concerning participants, publications e.tc. (in html and json format)
- Final or periodic **project reports** (depending on the stage of the project), i.e. the content of the report summary (in html/pdf format and the corresponding txt)
- **Results in Brief**, i.e. the results of the project (in html format)
- **Scholarly publications** deriving from the project, i.e. publications in PubMed ("Home - PMC - NCBI" n.d.) or other online archives
- **Patents** that resulted from the project directly or indirectly, i.e. patent metadata and the original text from EPO (European Patent Office n.d.)

The CORDIS platform ("CORDIS | European Commission" n.d.) was used as the main source for the data collection process. Targeted crawlers were developed in order to

gather data concerning the call documents and all other data related to the targeted projects, videlicet the reports, the objectives and several types of metadata. The publications that were produced as part of each project, were collected leveraging PubMed and OpenAIRE.

4.3 Data Analysis Framework

Upon collection of the aforementioned targeted data, several lexicalisations of the different aspects/facets of impact were determined. More specifically, the aim of the data analysis was to extract information about:

- **Projects** funded by EU and related elements (name, acronym, duration, budget etc.)
- **People** and **organisations** that participated in the projects
- Several **insights** mentioned within the projects, such as **methods** and **techniques** used, **products** and **technologies** produced, **diseases** and **drugs** related
- **MeSH** terms and indicative **keywords**

As far as the insights are concerned, a framework depicting innovation was built based on intuition resulting from the data and opinions of experts in the field of conceptual frameworks. The framework consists a taxonomy of several types of insights that are considered to be capturing the impact of the work performed in the context of the targeted projects. Thus, three major categories of insights were defined:

- **Domain-independent insights** that contribute to measuring academic and economic impact, i.e. innovation indicators irrespectively of the domain, such as number of publications, companies (start-ups, spin-offs) founded as a result of the project and how many people were employed etc.
- **Domain-related insights**, i.e. innovations produced by the projects that can be generic as concepts but are best defined and instantiated within each domain, in the case of the Data4Impact project, the Health domain. This kind of insights include devices, methods, software applications and patents.
- **Domain-dependent insights**, that is types of innovation that are closely related to the specific domain under examination, such as treatments, drugs, clinical trials and diagnostic tools.

4.4 Information Extraction

The information extraction task was performed by automatically processing the collected data with the use of Natural Language Processing workflows.

At first, the project reports were segmented and only certain sections related to the results and the impact of the projects were isolated for further analysis. This was considered necessary for two reasons. First, the final report texts were in many cases

quite extensive, rendering their processing in large scale data intensive in terms of computational resources. Second, in various sections of the reports, information is repeated, so analyzing all sections would lead to many duplicates of the same information types. Therefore, all project reports were segmented in the sections presented below, according to their corresponding EU funding programme:

Table 1: Project's segmentation in two sections according EU funding programme

FP7	**H2020**
Executive Summary	Summary of Context and Objectives
Context and Objectives	Work performed and main Results
Results	Progress beyond the state of the art and expected potential impact
Potential Impact	

Out of these sections, "Executive Summary" and "Potential Impact" were eventually chosen to be examined for FP7 projects, while "Work performed and main results" and "Progress beyond the state of the art and expected potential impact" for H2020 projects.

In this context, Athena RC developed text mining and targeted Information Extraction modules to run against the whole dataset, in order to produce annotations of the following types: **named entities**, **insights** depicting innovation, **diseases** and **key terms**.

The first type of annotations, i.e. named entities of general interest like Locations, People, Organisations, Companies as well as the Projects themselves (usually their acronyms), was derived using existing technologies. This information allowed for the detection of people, universities and institutes involved in EU projects, also defining spatially the impact in later stages of the analysis.

The second type of annotations, i.e. insights related to the innovation and thus the impact of the projects, were examined following the rationale of the framework described in 4.3 Data Analysis Framework. Thus, considering the Domain-independent category, a valuable innovation indicator that could also constitute an impact indicator was the insight of type "Company", which stands for the startup companies that were founded as a result of the work performed within a project.

With regard to the Domain-relevant category, methods used in the context of the projects were located. This type of information could assist in tracking whether a project had established a methodology in the health domain targeting a specific disease or medical procedure (which was also used by other projects later in time) or if that method led to the production of a commercial product or drug (which would indicate potential societal impact of the EU funding).

Finally, the Domain-specific category was associated with the detection of actual outputs of the projects (i.e. drugs) and hence these annotations represent the actual societal impact of the EU funding in the health domain.

The third type of annotations, i.e detection of diseases, aimed to provide several useful intuitions and lead to valuable conclusions regarding the impact of the projects in society. Moreover, classifying diseases according to established taxonomies such as ICD ("WHO | International Classification of Diseases, 11th Revision (ICD-11)" 2018), could result in discovering the main research areas and cluster the projects as per that. In a next step, indicators related to the principal areas in the health domain and the budget they received, could be approached.

The last type of extracted annotations from the targeted data was **key terms**, in other words indicative keywords for the work performed in the projects. These terms would also contribute to defining the subjects of the projects and the research areas they were involved into. Moreover, they could assist in identifying links and deriving correlations regarding innovations.

It should be noted that the aforementioned text analysis was carried out not only on project reports but also on all other different types of data gathered, as described in section 4.2 Data Collection.

4.5 Final Dataset

Eventually, JSON files containing the outcome of the textual analysis described in this chapter were constructed. In more detail, each project was associated with one JSON file comprising of the main attributes displayed in the picture below.

Image 5: Structure of JSON file representing a single project

As already mentioned in section 4.1 Scope, the purpose of this diploma thesis was to produce several network graphs considering social, semantic and socio-semantic information derived from the text and content analysis workflows, with nodes representing named entities, while arcs linking nodes depict relations between them.

In this context, information from the JSON attribute "administrative_data" was leveraged, as described in chapters 5. Scenario Definition and Dataset Segmentation and 6. Technical Implementation.

The exact structure of the particular JSON attribute can be found in the following illustration.

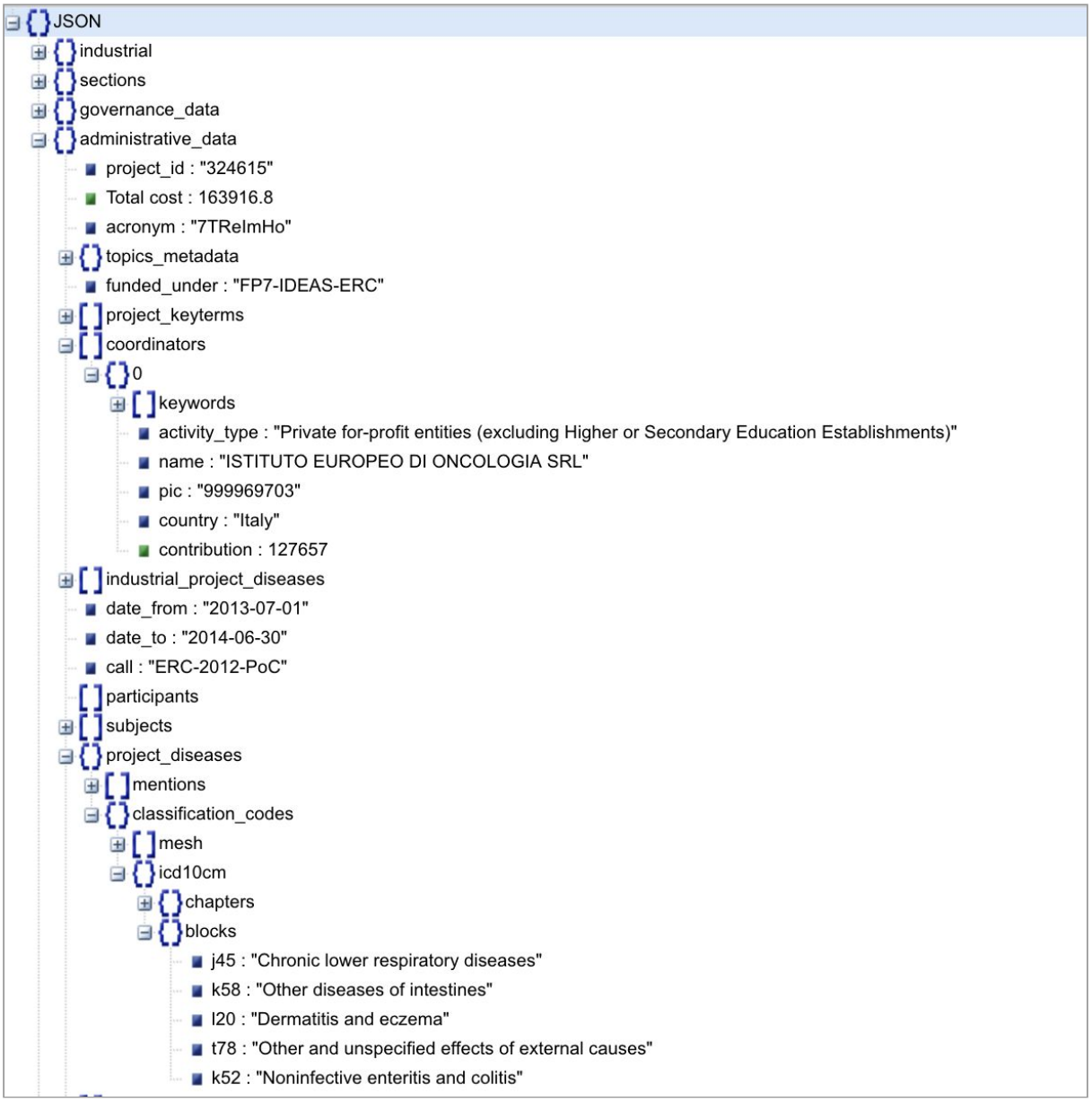


Image 6: Detailed structure of JSON file representing a single project

5. Scenario Definition and Dataset Segmentation

As already stated in chapter 4. Dataset Description, the purpose of this thesis was to leverage appropriate information from the dataset corpus in order to produce and study several graphs, where nodes would represent entities and edges would depict the relations between the nodes. Network analytics measures should be employed to interpret these relations and examine clusters and communities within the graphs. Additionally, the constructed networks should be examined in space and time in order to reveal and explore spatial and temporal trends and patterns.

In this context, a concrete and accurate business scenario was constructed, so as to base the performed analysis on solid and consistent assumptions. On top of that, a segmentation of the dataset corpus with reference to the ICD taxonomy ("WHO | International Classification of Diseases, 11th Revision (ICD-11)" 2018) was conducted, in an attempt to divide the projects as per the main research areas.

5.1 Collaboration Networks

The objective of the chosen business scenario was to identify and study the evolution of affiliations between all organisations involved in the targeted EU projects, either as coordinators or as participants. The key idea was to construct suitable graphs that would depict the collaborations among these named entities across time, and then perform cluster analysis on these graphs in order to extract the dominant research groups in Europe.

In addition, each group should be further examined with regard to the determinant attributes of its components, namely the organisation country, activity type and prevalent research (sub)area. Time intervals should also be used in order to detect any temporal norms, as described in 6. Technical Implementation.

Hence, for the aforementioned representation, each organisation (coordinator or participant) constitutes a node, while an undirected edge between two nodes indicates that these organisations collaborated at least in one project. It should be also noted that a collaboration link was assigned in case the corresponding organisations were part of the coordinators and/or participants of the project, with no further investigation of whether these organisations actually worked together to produce a project deliverable.

Thus, the set of undirected affiliation edges under consideration can be easily defined through the Cartesian product of the project nodes, with no self links allowed. For example, assuming that an EU project has (Organisation A, Organisation B) as coordinators and (Organisation X, Organisation Y, Organisation Z) as participants, the following nodes and edges should be created based on the business scenario:

- **Nodes:** A, B, X, Y, Z
- **Edges:** A – B, A – X, A – Y, A – Z, B – X, B – Y, B – Z, X – Y, X – Z, Y – Z

5.2 Diseases of the Circulatory System

As already mentioned in 4.4 Information Extraction, diseases were the third type of extracted annotations per project. These annotations were further classified using the ICD taxonomy ("WHO | International Classification of Diseases, 11th Revision (ICD-11)" 2018).

The International Statistical Classification of Diseases and Related Health Problems (ICD) is the foundation for the identification of health trends and statistics globally, and the international standard for reporting diseases and health conditions. It is the diagnostic classification standard for all clinical and research purposes, and defines the universe of diseases, disorders, injuries and other related health conditions, listed in a comprehensive, hierarchical fashion.

More specifically, ICD separates the health domain into several main disease classes (namely "chapters"), with each class containing a more detailed categorisation (namely "blocks"). For the purposes of Data4Impact, the 10th Revision of ICD was used ("ICD-10 Version:2016" n.d.), and during information extraction process each project was associated with all relative blocks (for more details refer to 4.5 Final Dataset).

For the purposes of this diploma thesis, a decision was made to study only projects associated with ICD **Chapter IX: Diseases of the Circulatory System**, so as to construct more concise graphs that would be easier to interpret. The chapter blocks (I00 - I99) were further organized into "subclasses", as shown in the following table.

Table 2: ICD-10 Chapter IX blocks

Block range	Subclass
I00 - I02	Acute rheumatic fever
I05 - I09	Chronic rheumatic heart diseases
I10 - I15	Hypertensive diseases
I20 - I25	Ischaemic heart diseases
I26 - I28	Pulmonary heart disease and diseases of pulmonary circulation
I30 - I52	Other forms of heart disease
I60 - I69	Cerebrovascular diseases
I70 - I79	Diseases of arteries, arterioles and capillaries
I80 - I89	Diseases of veins, lymphatic vessels and lymph nodes, not elsewhere classified
I95 - I99	Other and unspecified disorders of the circulatory system

6. Technical Implementation

In this chapter, all actions that were performed in order to generate the appropriate graphs and further analyze them are being described.

To begin with, the initial 10-year time interval, i.e. 1/1/2008 to 31/12/2018, was further split into 5 smaller non-overlapping intervals, each covering a 2-year period. For each one of them a marginal graph was constructed, including only projects that were active during the whole or part of the interval. These constructed graphs were used as a starting point over which various algorithms were employed for centrality metrics computation and community detection.

6.1 Data Parsing and Preprocessing

In order to transform the initial JSON files into a format suitable for the rest of the analysis, the process below was followed:

- Extraction of the chapter-relative subclass per project, as described in 5.2 Diseases of the Circulatory System
- Extraction of the name, country and activity type of both coordinator and participant entities per project.
- Extraction and filtering of the start and end date of the project, with regard to the time interval under examination.
- Construction of a dictionary that represents the links (collaborations) among entities.
- Construction of a dictionary that holds the dominant chapter-relative subclass per organisation.

The retrieval of the name, country and activity type of both coordinators and participants, was performed using a Python's built-in package named **JSON** which is suitable for encoding and decoding JSON data. Upon completion of the deserialization process of each JSON file, suitable dictionary keys representing affiliation pairs were created through the concatenation of the extracted information.

The values corresponding to each key were calculated as the percentage of the months during which the project was active compared to the whole time period under examination. For example, the value 12/24 = 0.5 would be assigned to a collaboration (an organization pair)for a project which was active for 12 months during the 2-year period.

This heuristic was considered to be an appropriate choice since the assigned value will reflect the weight of the edges connecting the affiliated nodes in the graphs that are going to be constructed. It becomes evident, that cooperations between entities that

lasted for an extended time period are considered to be more significant, i.e. to have a higher weight, than others that lasted less.

6.2 Graph Creation

As a prerequisite for the construction of the graph representing the collaboration network, the aforementioned affiliation dictionary was suitably parsed. Upon completion of this process, the SNAP.py library was used. Among the supported graph classes, the TUNGraph was selected and the New() method was used for the creation of a new undirected network.

Due to SNAP.py inability to support edge weights, a workaround was implemented by applying cut-off thresholds of 0, 0.3 and 0.6 to the computed weights. As a result, for each 2-year period the following three graphs were constructed:

- Marginal graph containing all edges regardless of their weight (threshold 0)
- Marginal graph containing only edges representing affiliations that were active for at least 30% of the time period (threshold 0.3)
- Marginal graph containing only edges representing affiliations that were active for at least 60% of the time period (threshold 0.6)

In the constructed graphs, each node was associated with a non-negative integer ID. Additionally, both nodes and edges had no attributes or data associated with them and there was at most one undirected edge between a pair of nodes. Internally, SNAP implemented the undirected graph data structures using sorted adjacency lists.

TUNGraph also provides iterators for fast traversal of nodes and edges. Iterator classes are TUNGraphNodeI for iterating over nodes and TUNGraphEdgeI for iterating over edges, providing a convenient way to retrieve the degree of any node in the graphs.

6.3 Centralities and Connected Components Computation

By leveraging the large variety of graph manipulation and analytics methods that SNAP.py offers, a number of centrality metrics was computed in an attempt to identify the most important vertices in the constructed graph.

An exhaustive list of the centralities computed along with implementation details is presented in the following table:

Table 3: List of network metrics computed with SNAP

Method	**Description**
GetDegreeCentr	Returns degree centrality of a given node NId in Graph. Degree centrality of a node is defined as its degree/(N-1), where N is the number of nodes in the network.

GetBetweennessCentr	Computes (approximate) Node and Edge Betweenness Centrality based on a sample of NodeFrac nodes.
GetClosenessCentr	Returns closeness centrality of a given node NId in Graph. Closeness centrality is equal to 1/farness centrality.
GetFarnessCentr	Returns farness centrality of a given node NId in Graph. Farness centrality of a node is the average shortest path length to all other nodes that reside in the same connected component as the given node.
GetPageRank	Computes the PageRank score of every node in Graph. The scores are stored in PRankH.
GetHits	Computes the Hubs and Authorities score of every node in Graph. The scores are stored in NIdHubH and NIdAuthH.
GetNodeEcc	Returns node eccentricity, the largest shortest-path distance from the node NId to any other node in the Graph.
GetEigenVectorCentr	Computes eigenvector centrality of all nodes in Graph and stores it in NIdEigenH. Eigenvector Centrality of a node N is defined recursively as the average of centrality values of N's neighbors in the network.

As part of the topological investigation of the constructed graph, the connected components were also computed. In the context of an undirected graph, a connected component is a subgraph in which any two vertices are connected to each other by paths, and which is connected to no additional vertices in the supergraph. The computation was facilitated by SNAP's GetSccs() method that returns all strongly connected components in graph which is provided as input.

In addition, all of the above metrics were manually normalized, in case this was not already part of the SNAP implementation.

6.4 Community Detection

Upon completion of the centrality metrics computation and the detection of the connected components, the analysis focused on the revelation of the underlying community structure by employing suitable community detection algorithms.

The identification of these communities is considered to be of great importance as:

- Quite common in real networks, communities correspond to functional units of the system and thus being able to identify these sub-structures within a network can provide insight into how network function and topology affect each other.
- Very often the underlying communities exhibit properties very different than the average properties of the network to which they belong. Thus, only concentrating on the average properties leads to neglection of many important and interesting featured inside the networks.

For this task, the CNM community detection algorithm proposed by Clauset, Newman and Moore was used (Clauset, Newman, and Moore 2004). The algorithm is based on an hierarchical agglomerative approach for detecting community structure faster than many competing algorithms thus allowing execution of very large networks with running time O(md log n) where n is the number of vertice, m the number of edges and d the depth of the dendrogram describing the community structure.

With regard to the specific implementation, the CommunityCNM() SNAP's method was used which implements the aforementioned community detection method for large networks. At every execution step of the algorithm, two communities that contribute maximum positive value to global modularity are merged.

7. Presentation of Results

In this chapter, the output of the implementation described in 6. Technical Implementation is displayed, i.e. an ensemble of tables and illustrations, along with illuminative commentary, that depict the performed analysis in aggregate.

7.1 Overview

The following table contains some summary statistics for all generated graphs, across all thresholds and margins:

Table 4: Summary statistics of generated graphs across margins

Date from	Date to	Affiliations threshold	Number of nodes	Number of edges	Number of communities > 1	Modularity	Avg degree centrality	Density
2007-01-01	2008-12-31	0	436	4847	13	0,60695	0,05111	-
2009-01-01	2010-12-31	0	864	10649	20	0,48886	0,02856	-
2011-01-01	2012-12-31	0	1200	15605	23	0,40339	0,02169	-
2013-01-01	2014-12-31	0	1310	21950	25	0,50981	0,02560	-
2015-01-01	2016-12-31	0	1169	20215	18	0,48488	0,02961	-
2017-01-01	2018-12-31	0	939	16177	16	0,48220	0,03673	-
2007-01-01	2008-12-31	0,3	436	1937	12	0,65531	0,02043	0,08700
2009-01-01	2010-12-31	0,3	864	9925	19	0,50296	0,02662	0,03200
2011-01-01	2012-12-31	0,3	1200	14910	20	0,42182	0,02073	0,02400
2013-01-01	2014-12-31	0,3	1310	21125	28	0,46838	0,02464	0,02700
2015-01-01	2016-12-31	0,3	1169	18328	22	0,50527	0,02685	0,03200
2017-01-01	2018-12-31	0,3	939	14073	13	0,47050	0,03196	0,02100
2007-01-01	2008-12-31	0.6	436	36	7	0,60031	0,00038	-
2009-01-01	2010-12-31	0.6	864	7506	15	0,55346	0,02013	-
2011-01-01	2012-12-31	0.6	1200	9720	16	0,46963	0,01351	-

2013-01-01	2014-12-31	0.6	1310	10340	21	0,49031	0,01206	-
2015-01-01	2016-12-31	0.6	1169	15555	20	0,50662	0,02278	-
2017-01-01	2018-12-31	0.6	939	11998	12	0,48990	0,02724	-

As expected, the total number of nodes and edges drops with the increase of the cut-off threshold, resulting also in a decrease in the number of formed communities. Additionally, it is evident that during the 3rd, 4th and 5th period (2011 - 2016) more collaborations were established compared with the rest of the time windows. This is quite easy to interpret, since this time period indicates the peak of the FP7 funding programme, with many active EU projects throughout.

In terms of network/cluster density of connections among the different thresholds, the three computed metrics (modularity, density, average degree centrality) also seem to follow the same pattern across periods. Specifically, the 1st and 2nd period (2007 - 2010) are the most concise, which is sensible when taking into account that during this period the FP7 programme was at its start, hence less projects were active and thus any formed collaboration was stronger then than in the later years.

At this point, it should also be noted that a database holding the characteristics of the nodes with the top 5 measure scores per CNM cluster has been constructed, across all thresholds and periods. This information could be used as a starting point for further analysis regarding the top EU players in the health domain. The database consists of **13689** records and contains the following fields:

- **Date from:** start date of the time period
- **Date to:** end date of the time period
- **Link (collaboration) threshold:** cut-off threshold on edge weight (allowed values: 0, 0.3, 0.6)
- **CNM id:** id of the node CNM community
- **CNM size:** size of the node CNM community
- **Rank:** order of the node in terms of the measure under examination (allowed values: 1, 2, 3, 4, 5)
- **Measure:** name of the measure under examination (allowed values: Degree centrality, Betweenness centrality, Closeness centrality, Farness centrality, Eigenvector centrality, Pagerank, HITS authority, HITS hub, Eccentricity)
- **Value:** value of the node for a specific measure type
- **Country:** country of the organisation represented by the specific node

- **Activity type:** activity type of the organisation represented by the specific node (allowed values: research organisations, higher or secondary education establishments, private for-profit entities (excluding higher or secondary education establishments), public bodies (excluding research organisations and secondary or higher education establishments), other)
- **Name:** name of the organisation represented by the specific node
- **Subclass:** dominant chapter-relative ICD subclass of the organisation represented by the specific node (please also refer to 5.2 Diseases of the Circulatory System and 6.1 Data Parsing and Preprocessing)

7.2 Indicative barcharts

In this section, sample barcharts for threshold 0.3 are presented and explained. The complete set of generated charts across all margins can be found in Appendix.

Each bar in the following charts represents a detected community, based on the results of the CNM algorithm that was applied for the specific margin and threshold. The height of each bar indicates the size of each community, i.e. the number of participating organizations..

Within each community, the distribution of the participating countries, activity types and chapter subclasses is encoded in different colors, as depicted in the accompanying legends. The relative height of each color indicates the observed frequency of the corresponding feature under analysis.

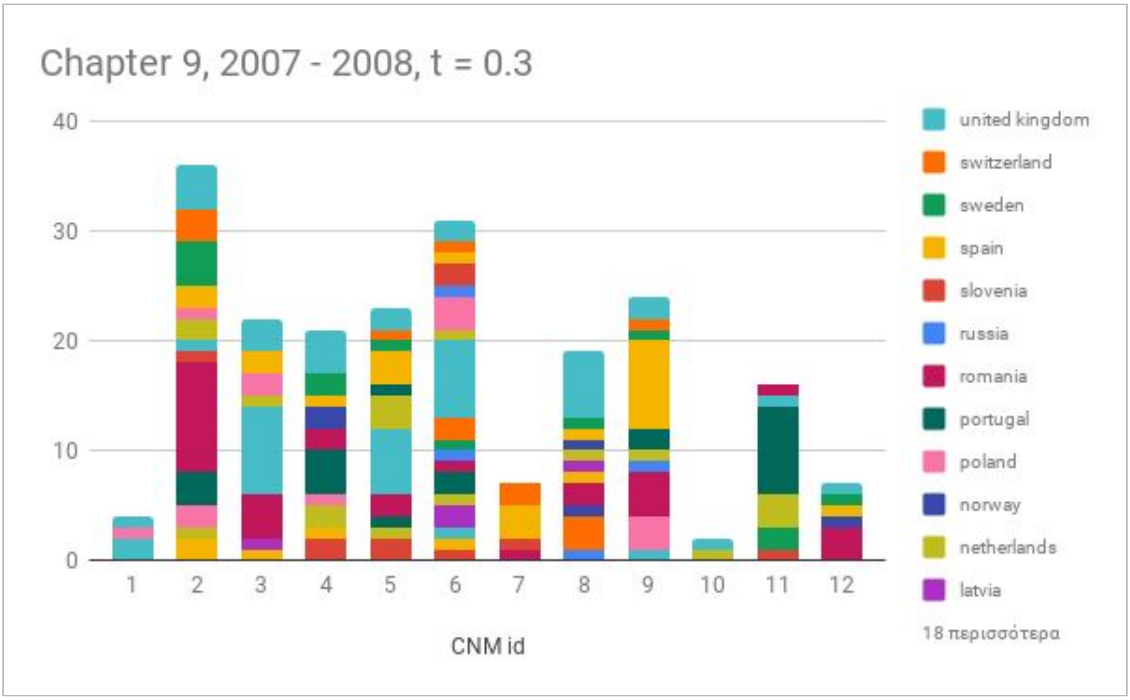


Image 7: Indicative country distribution per CNM community for time margin 2007-2008 and weight threshold 0.3

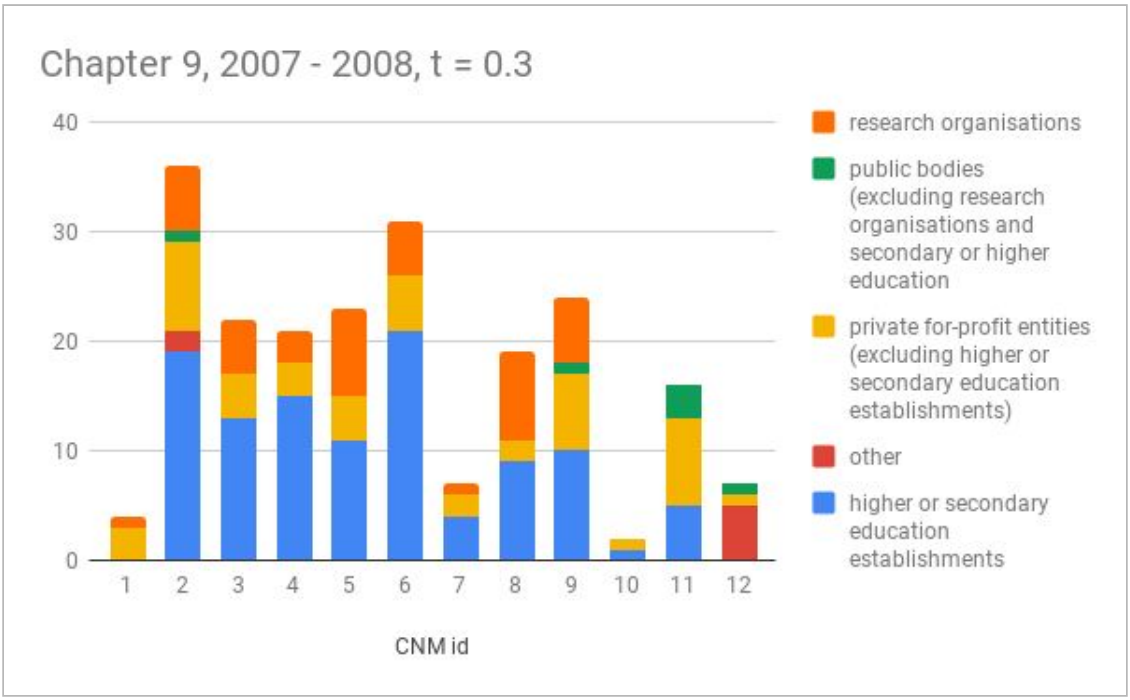


Image 8: Indicative activity type distribution per CNM community for time margin 2007-2008 and weight threshold 0.3

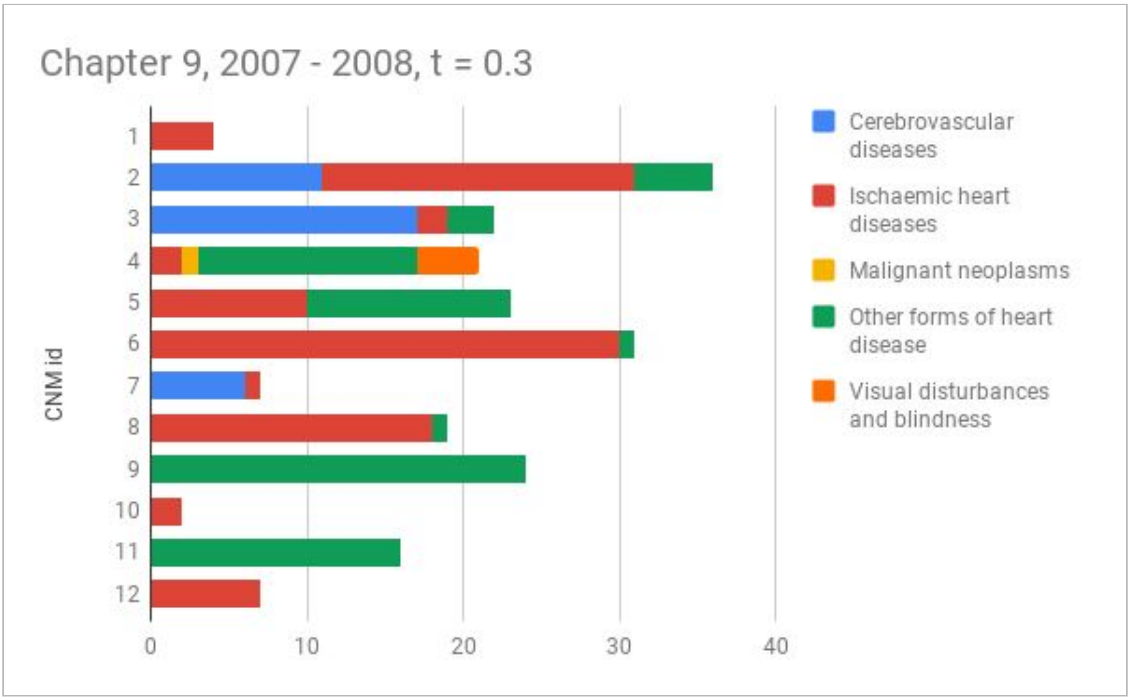


Image 9: Indicative subclass distribution per CNM community for time margin 2007-2008 and weight threshold 0.3

7.3 Graph illustrations

In this section, sample visualizations of graphs for the 0.3 threshold are presented and explained. The complete set of generated visualizations across all margins can be found in Appendix.

The size of the nodes varies in a scale of 0 to 100, analogously to the calculated betweenness centrality of the specific node. Thus, a vertex that exhibited high betweenness centrality will be represented with increased size. Regarding the rest of the plot elements, the color of the nodes represents the chapter-relative subclass to which the specific vertex belongs to. Regarding the label of each vertex, they act as an indication with reference to the community to which the node was assigned according to the applied CNM algorithm.

It must be noted that for the applied layout the attractive force is distributed along outbound links, thus pushing the hubs at the periphery while authorities remain more central.

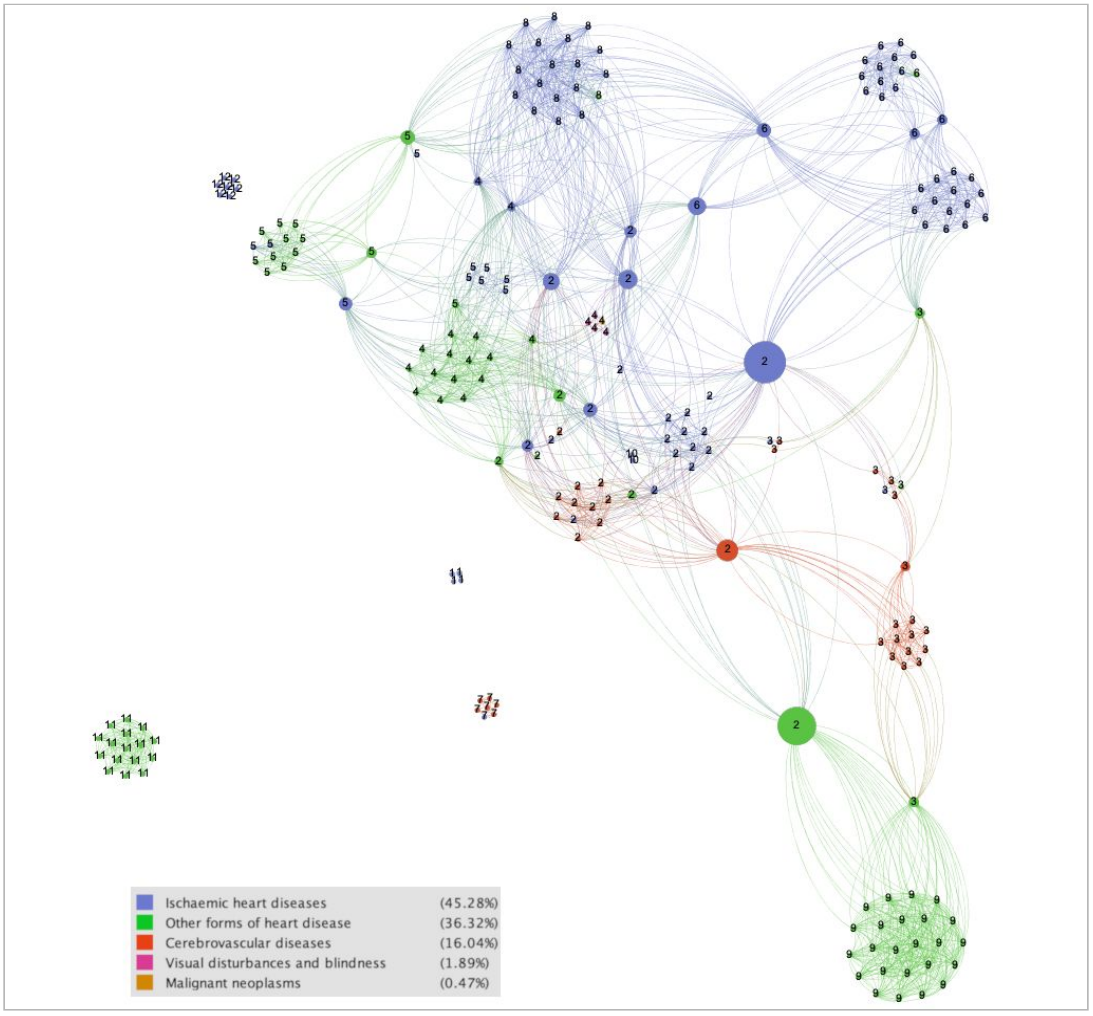


Image 10: Indicative collaboration network for time margin 2007-2008 and weight threshold 0.3

The accompanying scatterplot presents the distribution of the node degree among the vertices of the collaboration network. The degree value is plotted along the x-axis, while the number of nodes exhibiting this degree along the y-axis.

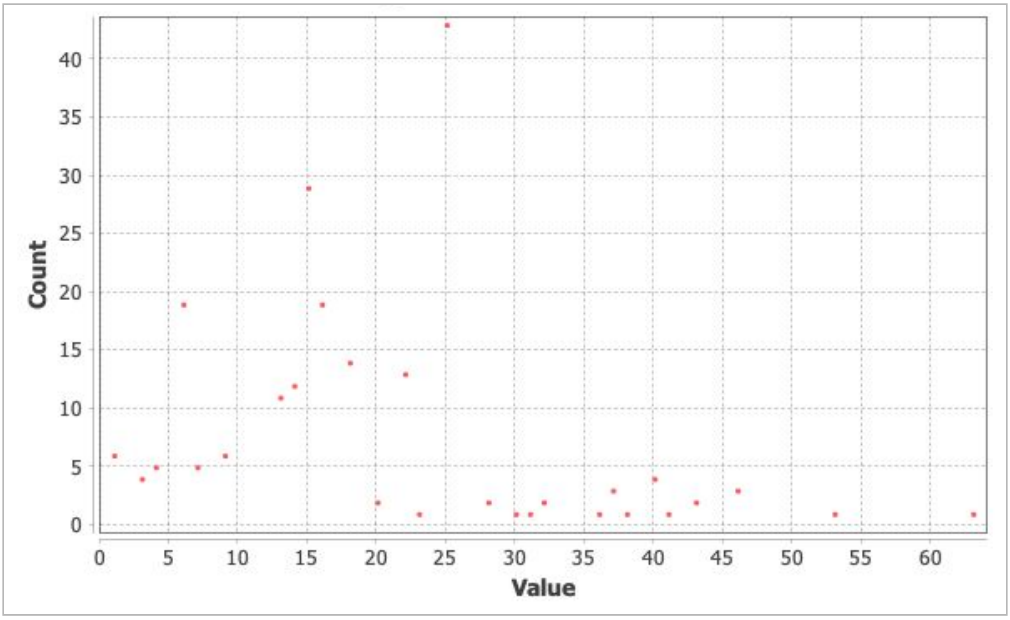


Image 11: Degree distribution for time margin 2007-2008 and weight threshold 0.3

7.4 Analysis of results

2007 - 2008 margin

For the first time margin, it is evident from the constructed barcharts that communities with IDs 2, 6 and 9 are the most prevalent. Within these communities, certain subclasses seem to stand out, namely Ischaemic Heart Diseases (IHD), Cerebrovascular Diseases (CD) and Other forms of Heart Diseases (OHD). Specifically, the following table summarizes the observed percentages per community.

Table 5: Observed percentages of dominant subclasses within the most prevalent communities for time margin 2007-2008

CNM ID	IHD	CD	OHD
2	55,56%	30,56%	13,89%
6	96,77%	-	3,23%
9	-	-	100,00%

Furthermore, the corresponding graph illustration also indicates that these subclasses are the preponderant ones, since there exist three extremely important nodes, acting as “bridges” for knowledge transfer between the nodes of the specific subclasses.

In addition, the following table outlines the distribution of the organizations’ activity type within each community.

Table 6: Observed percentages of dominant activity types within the most prevalent communities for time margin 2007-2008

CNM ID	Higher or secondary education establishments	Private for-profit entities	Research organizations
2	52,78%	22,22%	16,67%
6	67,74%	16,13%	16,13%
9	41,67%	29,17%	25,00%

High percentages of participating private for-profit entities potentially imply intensive research on consumer products that aim to directly tackle critical health-related challenges, thus producing high societal impact.

Upon stemming information from the aforementioned database (that holds the characteristics of the top 5 nodes per CNM community), the presence of nodes

belonging to the private sector can be observed in communities 6 and 9. Specifically, STMicroelectronics Agrate (Italy) for community 6 and Clothing Plus Oy (Finland) for community 9 are identified among the key players for this time margin, with regards to degree, betweenness and pagerank centralities.

Drilling down to the private sector entities alone, the following key stakeholders prevail per community across the above three centrality metrics:

Table 7: Key private sector entities within the most prevalent communities for time margin 2007-2008

CNM ID	Rank	Degree Centrality	Betweenness Centrality	Pagerank
2	1	Boehriger Ingelheim International GMBH (Germany)	Boehriger Ingelheim International GMBH (Germany)	Boehriger Ingelheim International GMBH (Germany)
2	2	GABO:mi Gesellschaft für Ablauforganisation:milliarium MBH & Co. KG (Germany)	Dando Weiss & Colucci Limited (United Kingdom)	Solvo Biotechnology ZRT (Hungary)
2	3	Nsgene As (Denmark)	GABO:mi Gesellschaft für Ablauforganisation:milliarium MBH & Co. KG (Germany)	GABO:mi Gesellschaft für Ablauforganisation:milliarium MBH & Co. KG (Germany)
2	4	Paion Deutschland GMBH (Germany)	Nsgene As (Denmark)	Nsgene As (Denmark)
2	5	Quick Cool AB (Sweden)	Paion Deutschland GMBH (Germany)	Paion Deutschland GMBH (Germany)
6	1	Stmicroelectronics SRL (Italy)	Stmicroelectronics SRL (Italy)	Stmicroelectronics SRL (Italy)
6	2	Pharnext SAS (France)	Pharnext SAS (France)	Pharnext SAS (France)
6	3	Reform E.C., Druzba Za Mednarodno Trgovino DOO (Slovenia)	Reform E.C., Druzba Za Mednarodno Trgovino DOO (Slovenia)	Reform E.C., Druzba Za Mednarodno Trgovino DOO (Slovenia)
6	4	Softeco Sismat SRL (Italy)	Softeco Sismat SRL (Italy)	Softeco Sismat SRL (Italy)
6	5	IBM Israel - Science and Technology LTD (Israel)	IBM Israel - Science and Technology LTD (Israel)	IBM Israel - Science and Technology LTD (Israel)
9	1	Clothing Plus MBU Oy (Finland)	Clothing Plus MBU Oy (Finland)	Clothing Plus MBU Oy (Finland)
9	2	Clothing Plus Oy (Finland)	Clothing Plus Oy (Finland)	Clothing Plus Oy (Finland)
9	3	Empirica Gesellschaft für Kommunikations- und Technologieforschung MBH (Germany)	Empirica Gesellschaft für Kommunikations- und Technologieforschung MBH (Germany)	Empirica Gesellschaft für Kommunikations- und Technologieforschung MBH (Germany)
9	4	Medtronic Iberica SA	Medtronic Iberica SA	Medtronic Iberica SA

		(Spain)	(Spain)	(Spain)
9	5	Philips Electronics Nederland BV (Netherlands)	Philips Electronics Nederland BV (Netherlands)	Philips Electronics Nederland BV (Netherlands)

It must be noted that the majority of the countries associated with one or more central nodes of the above table, are those which appear to be more active in terms of participation in European research projects. This observation is especially manifested in community 2, in which entities from Germany, the UK and Sweden constitute the 50% of the total participants.

2009 - 2010 margin

For the second time margin, the most prevalent communities are the ones with IDs 2, 4 and 5. Within these communities, the same subclasses seem to stand out again, namely Ischaemic Heart Diseases (IHD), Cerebrovascular Diseases (CD) and Other forms of Heart Diseases (OHD). Specifically, the following table summarizes the observed percentages per community.

Table 8: Observed percentages of dominant subclasses within the most prevalent communities for time margin 2009-2010

CNM ID	IHD	CD	OHD
2	56,35%	14,72%	27,41%
4	35,66%	20,16%	38,76%
5	23,68%	42,98%	32,46%

Furthermore, the corresponding graph illustration also indicates that the IHD subclass is amongst the preponderant ones, since there exist four relatively important nodes belonging to the this subclass, acting as “bridges” and thus being responsible for knowledge transfer between different communities.

In addition, the following table outlines the distribution of the organizations’ activity type within each community.

Table 9: Observed percentages of dominant activity types within the most prevalent communities for time margin 2009-2010

CNM ID	Higher or secondary education establishments	Private for-profit entities	Research organizations
2	46,19%	24,87%	21,83%

4	37,21%	33,33%	23,26%
5	30,70%	35,96%	25,44%

It is observed that, in comparison to the previous margin, higher percentages of participating private for-profit entities were measured, potentially implying increase in research on consumer products that aim to directly tackle critical health-related challenges, thus leading to higher societal impact.

Upon stemming information from the aforementioned database (that holds the characteristics of the top 5 nodes per CNM community), the presence of nodes belonging to the private sector can be observed in community 5. Specifically, CFC SRL (Italy) is identified among the key players for this time margin, with regards to degree, betweenness and pagerank centralities.

Drilling down to the private sector entities alone, the following key stakeholders prevail per community across the above three centrality metrics:

Table 10: Key private sector entities within the most prevalent communities for time margin 2009-2010

CNM ID	Rank	Degree Centrality	Betweenness Centrality	Pagerank
2	1	Astrazeneca AB (Sweden)	Glaxosmithkline Research and Development LTD (United Kingdom)	Astrazeneca AB (Sweden)
2	2	F. Hoffmann-La Roche AG (Switzerland)	Astrazeneca AB (Sweden)	F. Hoffmann-La Roche AG (Switzerland)
2	3	Glaxosmithkline Research and Development LTD (United Kingdom)	GABO:mi Gesellschaft für Ablauforganisation:milliarium MBH & Co. KG (Germany)	Glaxosmithkline Research and Development LTD (United Kingdom)
2	4	Pfizer Limited (United Kingdom)	Novartis Pharma AG (Switzerland)	Novartis Pharma AG (Switzerland)
2	5	Eli Lilly and Company Limited (United KIngdom)	F. Hoffmann-La Roche AG (Switzerland)	Pfizer Limited (United Kingdom)
4	1	Philips Medical Systems Nederland BV (Netherlands)	Philips Medical Systems Nederland BV (Netherlands)	Philips Medical Systems Nederland BV (Netherlands)
4	2	Philips Iberica SA (Spain)	Arttic (France)	Philips Iberica SA (Spain)
4	3	Fondazione Centro San Raffaele (Italy)	Philips Iberica SA (Spain)	Qiagen GMBH (Germany)
4	4	Baxter Innovations GMBH (Austria)	Qiagen GMBH (Germany)	Arttic (France)
4	5	Bayer Innovation GMBH	Kuros Biosurgery AG	Smith & Nephew UK

		(Germany)	(Switzerland)	Limited (United Kingdom)
5	1	CFC SRL (Italy)	CFC SRL (Italy)	CFC SRL (Italy)
5	2	Philips Electronics Nederland BV (Netherlands)	Cedrat Technologies SA (France)	Philips Electronics Nederland BV (Netherlands)
5	3	Philips Technologie GMBH (Germany)	Philips Electronics Nederland BV (Netherlands)	Philips Technologie GMBH (Germany)
5	4	Empirica Gesellschaft für Kommunikations- und Technologieforschung MBH (Germany)	Philips Technologie GMBH (Germany)	Siemens Aktiengesellschaft (Germany)
5	5	T-Systems ITC Iberia SA (Spain)	Siemens Aktiengesellschaft (Germany)	Guger Technologies OG (Austria)

It must be noted that the majority of the countries associated with one or more central nodes of the above table, are those which appear to be more active in terms of participation in European research projects. This observation is manifested across all three communities, with Germany, Italy, Spain, France and the UK claiming most of the participants.

2011 - 2012 margin

For the third time margin, the most prevalent communities are the ones with IDs 1, 2 and 3. Within these communities, the same subclasses seem to stand out yet again, namely Ischaemic Heart Diseases (IHD), Cerebrovascular Diseases (CD) and Other forms of Heart Diseases (OHD). Specifically, the following table summarizes the observed percentages per community.

Table 11: Observed percentages of dominant subclasses within the most prevalent communities for time margin 2011-2012

CNM ID	IHD	CD	OHD
1	29,77%	16,74%	40,93%
2	67,24%	15,95%	12,07%
3	35,09%	47,37%	12,28%

Furthermore, the corresponding graph illustration also indicates that the nodes present are densely connected with each other and thus very few participants exhibit relatively high betweenness centrality. Among these, the one that stands out the most belongs to the IHD subclass of community 2 and is highly connected with other nodes from the

OHD subclass, consequently facilitating knowledge transfer between the aforementioned groups.

In addition, the following table outlines the distribution of the organizations' activity type within each community.

Table 12: Observed percentages of dominant activity types within the most prevalent communities for time margin 2011-2012

CNM ID	Higher or secondary education establishments	Private for-profit entities	Research organizations
1	35,35%	32,09%	24,19%
2	47,41%	25,00%	23,28%
3	25,44%	46,49%	20,18%

For the first time it can be observed that there exists an over 45% participation of private for-profit entities in one of the dominant communities, potentially implying that the corresponding community (i.e. community 3) is heavily focusing on consumer product research that would directly benefit the European society.

Upon stemming information from the aforementioned database (that holds the characteristics of the top 5 nodes per CNM community), the presence of nodes belonging to higher or secondary education establishments, public bodies and research organizations can be observed in all communities, while private for-profit entities play a less central role with regards to degree, betweenness and pagerank centralities.

Drilling down to the private sector entities alone, the following key stakeholders prevail per community across the above three centrality metrics:

Table 13: Key private sector entities within the most prevalent communities for time margin 2011-2012

CNM ID	Rank	Degree Centrality	Betweenness Centrality	Pagerank
1	1	Philips Medical Systems Nederland BV (Netherlands)	Stmicroelectronics SRL (Italy)	Philips Medical Systems Nederland BV (Netherlands)
1	2	Stmicroelectronics SRL (Italy)	Philips Medical Systems Nederland BV (Netherlands)	Stmicroelectronics SRL (Italy)
1	3	Arttic (France)	Arttic (France)	Arttic (France)
1	4	Philips Electronics Nederland BV (Netherlands)	Philips Electronics Nederland BV (Netherlands)	Philips Electronics Nederland BV (Netherlands)

1	5	Philips Iberica SA (Spain)	T-Systems ITC Iberia SA (Spain)	Philips Iberica SA (Spain)
2	1	GABO:mi Gesellschaft für Ablauforganisation:milliarium MBH & Co. KG (Germany)	GABO:mi Gesellschaft für Ablauforganisation:milliarium MBH & Co. KG (Germany)	GABO:mi Gesellschaft für Ablauforganisation:milliarium MBH & Co. KG (Germany)
2	2	Astrazeneca AB (Sweden)	Astrazeneca AB (Sweden)	Astrazeneca AB (Sweden)
2	3	F. Hoffmann-La Roche AG (Switzerland)	Tataa Biocenter AB (Sweden)	Pfizer Limited (United Kingdom)
2	4	Eli Lilly and Company Limited (United KIngdom)	Pfizer Limited (United Kingdom)	F. Hoffmann-La Roche AG (Switzerland)
2	5	Pfizer Limited (United Kingdom)	F. Hoffmann-La Roche AG (Switzerland)	Eli Lilly and Company Limited (United KIngdom)
3	1	Glaxosmithkline Research and Development LTD (United Kingdom)	Glaxosmithkline Research and Development LTD (United Kingdom)	Glaxosmithkline Research and Development LTD (United Kingdom)
3	2	Novartis Pharma AG (Switzerland)	San Raffaele SPA (Italy).	Novartis Pharma AG (Switzerland)
3	3	Dando Weiss & Colucci Limited (United Kingdom)	Novartis Pharma AG (Switzerland)	Dando Weiss & Colucci Limited (United Kingdom)
3	4	Intercytex LTD (United Kingdom)	Ossur HF (Iceland)	Qiagen GMBH (Germany)
3	5	Baxter Innovations GMBH (Austria)	Qiagen GMBH (Germany)	Kuros Biosurgery AG (Switzerland)

It must be noted that the majority of the countries associated with one or more central nodes of the above table, are those which appear to be more active in terms of participation in European research projects. Specifically, Germany, Italy, Spain and the UK, which have the highest numbers of participants, appear to be represented from at least one node among the top 5 more central vertices.

2013 - 2014 margin

For the fourth time margin, it is evident from the constructed barcharts that there exist four prevalent communities, i.e. the ones with IDs 1, 2, 3 and 4. Within these communities, the same three subclasses stand out once again, namely Ischaemic Heart Diseases (IHD), Cerebrovascular Diseases (CD) and Other forms of Heart Diseases (OHD). Specifically, the following table summarizes the observed percentages per community.

Table 14: Observed percentages of dominant subclasses within the most prevalent communities for time margin 2013-2014

CNM ID	IHD	CD	OHD
1	49,42%	17,44%	25,19%
2	26,96%	1,74%	46,96%
3	28,82%	41,18%	22,94%
4	-	98,29%	1,71%

Interestingly enough, from the graph illustrations it appears that there exists a large community (i.e. community 4) focusing almost exclusively on CD, whose members are densely connected which each other. Existence of nodes that exhibit relatively high betweenness centrality, thus acting as "bridges", cannot be observed.

In addition, the following table outlines the distribution of the organizations' activity type within each community.

Table 15: Observed percentages of dominant activity types within the most prevalent communities for time margin 2013-2014

CNM ID	Higher or secondary education establishments	Private for-profit entities	Research organizations
1	39,73%	34,88%	21,71%
2	25,22%	35,65%	25,22%
3	27,06%	45,88%	20,59%
4	27,35%	49,57%	20,51%

As with the previous margin, it can yet again be observed that there exist communities with very high participation of private for-profit entities, namely communities 3 and 4. Especially for community 4, it appears that almost half of the participants are private for-profit entities. Taking into consideration that the specific community is exclusively focused on CD, as previously mentioned, the emergence of increased interest in the delivery of health products related to the specific type of diseases, can be derived.

Upon stemming information from the aforementioned database (that holds the characteristics of the top 5 nodes per CNM community), the presence of nodes belonging to the private sector can be observed in all four dominant communities. Specifically, Novartis Pharma AG (Switzerland) for community 2 is identified among the

key players for this time margin, with regards to degree, betweenness and pagerank centralities.

Drilling down to the private sector entities alone, the following key stakeholders prevail per community across the above three centrality metrics:

Table 16: Key private sector entities within the most prevalent communities for time margin 2013-2014

CNM ID	Rank	Degree Centrality	Betweenness Centrality	Pagerank
1	1	GABO:mi Gesellschaft für Ablauforganisation:milliarium MBH & Co. KG (Germany)	GABO:mi Gesellschaft für Ablauforganisation:milliarium MBH & Co. KG (Germany)	GABO:mi Gesellschaft für Ablauforganisation:milliarium MBH & Co. KG (Germany)
1	2	Pfizer Limited (United Kingdom)	Pfizer Limited (United Kingdom)	Pfizer Limited (United Kingdom)
1	3	F. Hoffmann-La Roche AG (Switzerland)	Basf SE (Germany)	F. Hoffmann-La Roche AG (Switzerland)
1	4	Astrazeneca AB (Sweden)	Stmicroelectronics SRL (Italy)	Astrazeneca AB (Sweden)
1	5	Randox Clinics Limited (United Kingdom)	Contipro Biotech SRO (Czech Republic)	Randox Clinics Limited (United Kingdom)
2	1	Novartis Pharma AG (Switzerland)	Novartis Pharma AG (Switzerland)	Novartis Pharma AG (Switzerland)
2	2	Bayer Pharma AG (Germany)	Synapse Research Management Partners SL (Spain)	Synapse Research Management Partners SL (Spain)
2	3	Synapse Research Management Partners SL (Spain)	Glaxosmithkline Research and Development LTD (United Kingdom)	Bayer Pharma AG (Germany)
2	4	Glaxosmithkline Research and Development LTD (United Kingdom)	Bayer Pharma AG (Germany)	Glaxosmithkline Research and Development LTD (United Kingdom)
2	5	Takeda Development Centre Europe LTD (United Kingdom)	Acies Consulting Group SAS (France)	Amgen NV (Belgium)
3	1	Sorin Biomedica Cardio SRL (Italy)	T-Systems ITC Iberia SA (Spain)	Sorin Biomedica Cardio SRL (Italy)
3	2	Philips Medical Systems Nederland BV (Netherlands)	Sorin Biomedica Cardio SRL (Italy)	Philips Medical Systems Nederland BV (Netherlands)
3	3	Philips Electronics Nederland BV (Netherlands)	Bit & Brain Technologies SL (Spain)	T-Systems ITC Iberia SA (Spain)
3	4	Kopint-Tarki Konjunkturakutato Intezet ZRT (Hungary)	Philips Medical Systems Nederland BV (Netherlands)	Philips Electronics Nederland BV (Netherlands)

3	5	T-Systems ITC Iberia SA (Spain)	San Raffaele SPA (Italy).	Guger Technologies OG (Austria)
4	1	R.U.Robots Limited (United Kingdom)	R.U.Robots Limited (United Kingdom)	R.U.Robots Limited (United Kingdom)
4	2	The Shadow Robot Company Limited (United Kingdom)	The Shadow Robot Company Limited (United Kingdom)	The Shadow Robot Company Limited (United Kingdom)
4	3	Accel (France)	Accel (France)	Accel (France)
4	4	Idrogenet SRL (Italy)	RFDN Technologies AB (Sweden)	Idrogenet SRL (Italy)
4	5	Imer International SPA (Italy)	Imer International SPA (Italy)	Imer International SPA (Italy)

It must be noted that the majority of the countries associated with one or more central nodes of the above table, are those which appear to be more active in terms of participation in European research projects. This observation is especially manifested by the UK, which is represented by at least two participants in the top 5 key players across almost all communities.

2015 - 2016 margin

For the fifth time margin, the communities that prevail are the ones with IDs 1, 3 and 4. Within these communities, yet again the same subclasses stand out, namely Ischaemic Heart Diseases (IHD), Cerebrovascular Diseases (CD) and Other forms of Heart Diseases (OHD). Specifically, the following table summarizes the observed percentages per community.

Table 17: Observed percentages of dominant subclasses within the most prevalent communities for time margin 2015-2016

CNM ID	IHD	CD	OHD
1	30,41%	29,05%	28,38%
3	43,97%	21,36%	30,40%
4	5,37%	82,55%	8,05%

From the corresponding graph illustrations, it appears that the large community focusing almost exclusively on CD exists for this margin as well (i.e. community 4). Similarly, existence of nodes acting as “bridges” cannot be observed.

In addition, the following table outlines the distribution of the organizations’ activity type within each community.

Table 18: Observed percentages of dominant activity types within the most prevalent communities for time margin 2015-2016

CNM ID	Higher or secondary education establishments	Private for-profit entities	Research organizations
1	27,03%	42,57%	22,97%
3	38,69%	34,17%	20,60%
4	30,87%	48,32%	17,45%

As already mentioned, high percentages of participating private for-profit entities potentially imply intensive research on consumer products that aim to directly tackle critical health-related challenges, thus producing high societal impact. Especially for community 4, it appears once again that almost half of the participants are private for-profit entities. Hence, the specific community seems to be exclusively focused on research and delivery of health products related to CD for two consecutive time margins (2013 - 2014 & 2015 - 2016).

Upon stemming information from the aforementioned database (that holds the characteristics of the top 5 nodes per CNM community), the presence of nodes belonging to higher or secondary education establishments and research organizations can be observed in all communities while private for-profit entities play a less central role with regards to degree, betweenness and pagerank centralities.

Drilling down to the private sector entities alone, the following key stakeholders prevail per community across the above three centrality metrics:

Table 19: Key private sector entities within the most prevalent communities for time margin 2015-2016

CNM ID	Rank	Degree Centrality	Betweenness Centrality	Pagerank
1	1	Genedata AG (Switzerland)	Genedata AG (Switzerland)	Genedata AG (Switzerland)
1	2	Etss AG (Switzerland)	Philips GMBH (Germany)	Philips GMBH (Germany)
1	3	Plastic Components and Modules Automotive SPA (Italy)	European Screeningport GMBH (Germany)	Plastic Components and Modules Automotive SPA (Italy)
1	4	Plasmachem Produktions- und Handel GMBH (Germany)	3H Biomedical AB (Sweden)	Etss AG (Switzerland)
1	5	Malsch Neelina Hermina (Netherlands)	Malsch Neelina Hermina (Netherlands)	Plasmachem Produktions- und Handel GMBH (Germany)

3	1	GABO:mi Gesellschaft für Ablauforganisation:milliarium MBH & Co. KG (Germany)	GABO:mi Gesellschaft für Ablauforganisation:milliarium MBH & Co. KG (Germany)	GABO:mi Gesellschaft für Ablauforganisation:milliarium MBH & Co. KG (Germany)
3	2	Arttic (France)	Bracco Imaging SPA (Italy)	Arttic (France)
3	3	Randox Clinics Limited (United Kingdom)	Arttic (France)	Randox Clinics Limited (United Kingdom)
3	4	Biocrates Life Sciences AG (Austria)	Pfizer Limited (United Kingdom)	Biocrates Life Sciences AG (Austria)
3	5	Pfizer Limited (United Kingdom)	Biocrates Life Sciences AG (Austria)	Pfizer Limited (United Kingdom)
4	1	The Shadow Robot Company Limited (United Kingdom)	The Shadow Robot Company Limited (United Kingdom)	The Shadow Robot Company Limited (United Kingdom)
4	2	Accel (France)	Corlife OHG (Germany)	Accel (France)
4	3	Pilz GMBH & Co. KG (Germany)	Accel (France)	Pilz GMBH & Co. KG (Germany)
4	4	AEA SRL (Italy).	Marsi Bionics SL (Spain)	AEA SRL (Italy).
4	5	Ibak Helmut Hunger GMBH & Co. KG (Germany)	Pre Gel SPA (Italy)	Ibak Helmut Hunger GMBH & Co. KG (Germany)

It must be noted that the majority of the countries associated with one or more central nodes of the above table, are those which appear to be more active in terms of participation in European research projects. Specifically, Germany, Italy, Spain and the UK, which have the highest numbers of participants in total, appear to be represented from at minimum one node among the top 5 more central vertices in at least one community.

2017 - 2018 margin

For the last time margin, the communities with IDs 2, 3 & 4 are the most prevalent. Within these communities, the same subclasses stand out as with all previous margins, namely Ischaemic Heart Diseases (IHD), Cerebrovascular Diseases (CD) and Other forms of Heart Diseases (OHD). Specifically, the following table summarizes the observed percentages per community.

Table 20: Observed percentages of dominant subclasses within the most prevalent communities for time margin 2017-2018

CNM ID	IHD	CD	OHD
2	17,37%	35,33%	40,12%

3	31,64%	31,04%	36,12%
4	6,99%	79,72%	4,20%

The graph illustrations indicate that, for the third consecutive time margin, a large community focusing almost exclusively on CD can be detected,, whose members are densely connected which each other. It must be noted that, compared to the two previous time margins (2013 - 2014 & 2015 - 2016), this community has grown in size, enclosing 40,12% of the total graph nodes. Existence of nodes that exhibit remarkably high betweenness centrality cannot be observed.

In addition, the following table outlines the distribution of the organizations' activity type within each community.

Table 21: Observed percentages of dominant activity types within the most prevalent communities for time margin 2017-2018

CNM ID	**Higher or secondary education establishments**	**Private for-profit entities**	**Research organizations**
2	34,73%	41,92%	17,96%
3	37,01%	29,85%	20,90%
4	22,38%	58,74%	16,08%

For the first time it can be observed that there exists an over 50% participation of private for-profit entities in one of the dominant communities (i.e. community 4). Taking into account that the specific community has been almost exclusively focused on CD since 2013, the formation of a European group of organizations with a continuous and increased interest in the delivery of health products related to the specific type of diseases can be derived. It remains to be verified whether this collaboration resulted in producing high societal impact across Europe.

Upon stemming information from the aforementioned database (that holds the characteristics of the top 5 nodes per CNM community), the presence of nodes belonging to higher or secondary education establishments, public bodies and research organizations can be observed in all communities while private for-profit entities play a less central role with regards to degree, betweenness and pagerank centralities. This comes as a surprise, especially for community 4, where the private sector is represented by almost 60% of the participating nodes.

Drilling down to the private sector entities alone, the following key stakeholders prevail per community across the above three centrality metrics:

Table 22: Key private sector entities within the most prevalent communities for time margin 2017-2018

CNM ID	Rank	Degree Centrality	Betweenness Centrality	Pagerank
2	1	Novartis Pharma AG (Switzerland)	Novartis Pharma AG (Switzerland)	Novartis Pharma AG (Switzerland)
2	2	Genedata AG (Switzerland)	Alta Ricerca e Sviluppo in Biotecnologie SRLU (Italy)	Genedata AG (Switzerland)
2	3	Alta Ricerca e Sviluppo in Biotecnologie SRLU (Italy)	Genedata AG (Switzerland)	Alta Ricerca e Sviluppo in Biotecnologie SRLU (Italy)
2	4	Lifetec Group BV (Netherlands)	Lifetec Group BV (Netherlands)	Lifetec Group BV (Netherlands)
2	5	Cbk Sci Con Limited (United Kingdom)	Polygene AG (Switzerland)	Cbk Sci Con Limited (United Kingdom)
3	1	GABO:mi Gesellschaft für Ablauforganisation:milliarium MBH & Co. KG (Germany)	Bracco Imaging SPA (Italy)	GABO:mi Gesellschaft für Ablauforganisation:milliarium MBH & Co. KG (Germany)
3	2	Randox Clinics Limited (United Kingdom)	GABO:mi Gesellschaft für Ablauforganisation:milliarium MBH & Co. KG (Germany)	Arttic (France)
3	3	Arttic (France)	Accelopment AG (Switzerland)	Randox Clinics Limited (United Kingdom)
3	4	ACS Biomarker BV (Netherlands)	Celyad (Belgium)	ACS Biomarker BV (Netherlands)
3	5	Inserm - Transfert SA (France)	Arttic (France)	Institut de Recherches Internationales Servier (France)
4	1	Humanware SRL (Italy)	Contipro AS (Czech Republic)	Humanware SRL (Italy)
4	2	Inloc Robotics SLU (Spain)	Contipro Biotech SRO (Czech Republic)	Inloc Robotics SLU (Spain)
4	3	Simtech Design SL (Spain)	Ab.Acus SRL (Italy)	Simtech Design SL (Spain)
4	4	Flexible Robotic Solutions (Belgium)	Istraživačko-Razvojni Institut RT-RK DOO za Sisteme Zasnovane na Računarima Novi Sad (Serbia)	Flexible Robotic Solutions (Belgium)
4	5	Fomento de Construcciones y Contratas SA (Spain)	Moog Controls Limited (United Kingdom)	Fomento de Construcciones y Contratas SA (Spain)

It must be noted that the majority of the countries associated with one or more central nodes of the above table, are those which appear to be more active in terms of participation in European research projects. Interestingly, 3 out of the top 5 private for-profit entities participating in community 4, i.e. the one focused almost exclusively on CD, are from Spain.

Overall Evolution

With reference to the performed analysis so far, it is evident that between 2007 and 2018 significant amount of research has been conducted in the area of Cardiovascular Diseases. This conclusion could be justified by the fact that according to European Cardiovascular Disease Statistics 2017 ("Annual Reports" n.d.), Diseases of the Heart and Circulatory System (also known as CVD) are the leading cause of mortality in Europe as a whole, responsible for over 3.9 million deaths a year, or 45% of all deaths.

Among these, the main forms of CVD are Ischaemic Heart Disease (IHD) and Cerebrovascular Diseases (CD, also known as Stroke), which have also been identified as two of the dominant ICD-10 subclasses across all time margins. IHD is the leading single cause of mortality in Europe, responsible for 862,000 deaths a year (19% of all deaths) among men and 877,000 deaths (20%) among women each year. Stroke is the second most common single cause of death in Europe, accounting for 405,000 deaths (9%) in men and 583,000 (13%) deaths in women each year .

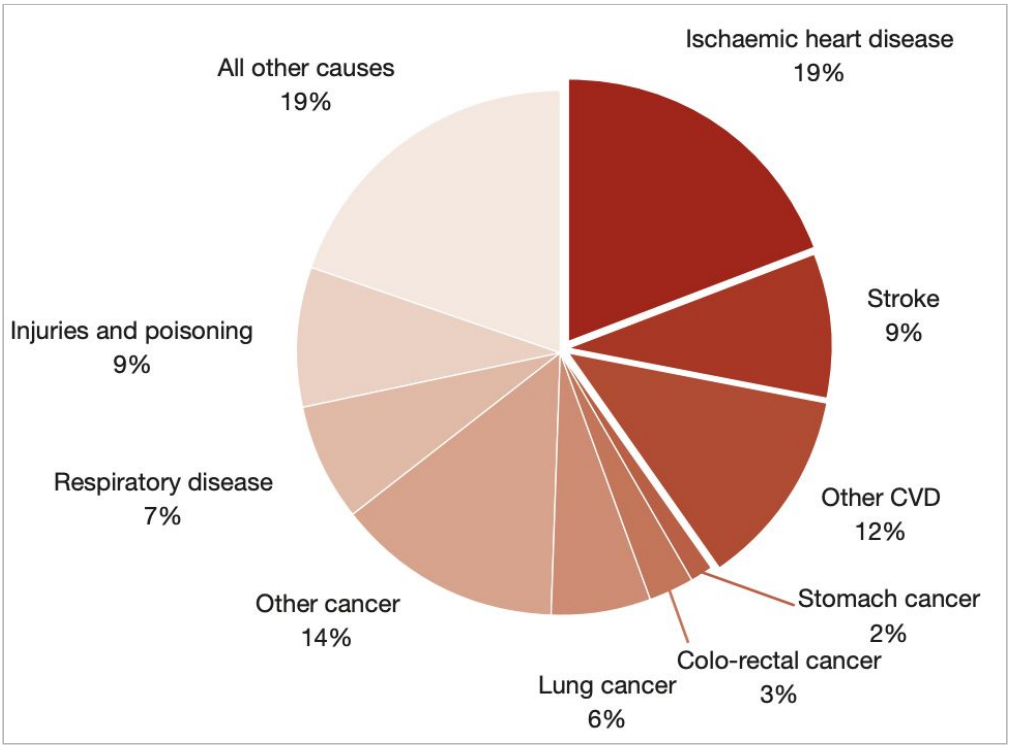


Image 12: Death rates by cause in Europe (Source: European CVD Statistics 2017)

Under these circumstances, research on different aspects of CVD has been funded since the beginning of the European Commission's Framework Programmes for Research,

Technological Development and Innovation. These transnational funding programmes have focused on the causes, diagnosis, treatment, and prevention of CVD.

With regards to the progress of research interest, IHD accounted for the largest community during years 2007 - 2015 (with 42,33% of organizations on average conducting research in this area yearly), while from 2016 and onwards, the research interest seems to focus more on CD. This may be relevant to the fact that standardised death rates for IHD followed a downward path (30.3% reduction for men and 44.3%) between 2005 and 2015, according to official reports of Eurostat presented below ("Statistics Explained" n.d.), thus probably indicating significant impact of the produced research in the well-being improvement of the European citizens.

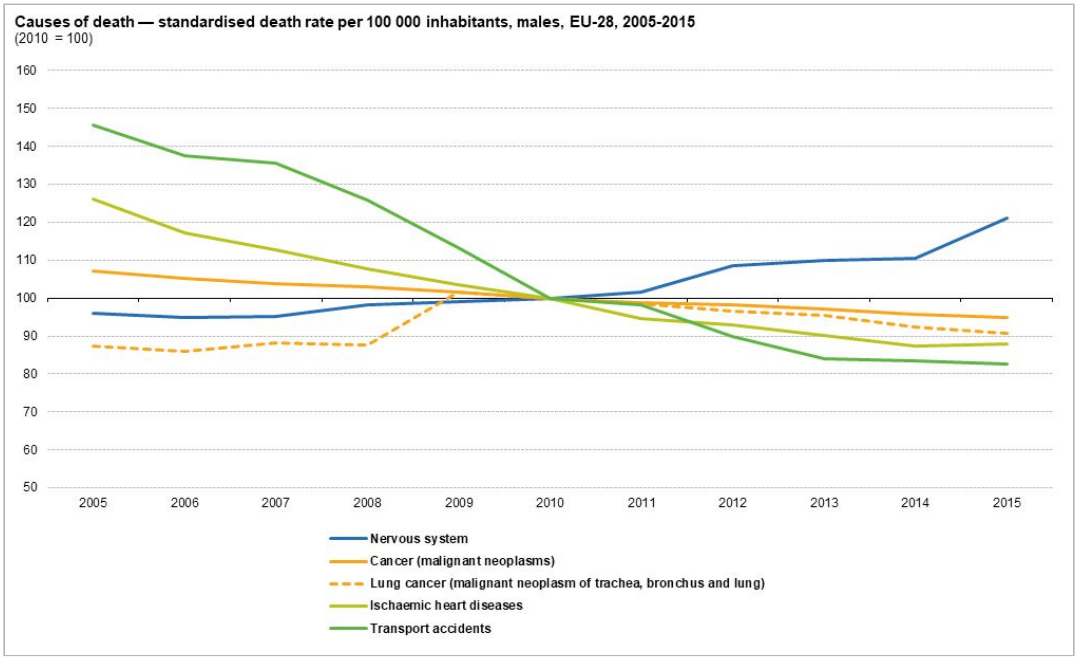


Image 13: Development of standardised death rate among males per cause between 2005-2015 (source: Eurostat)

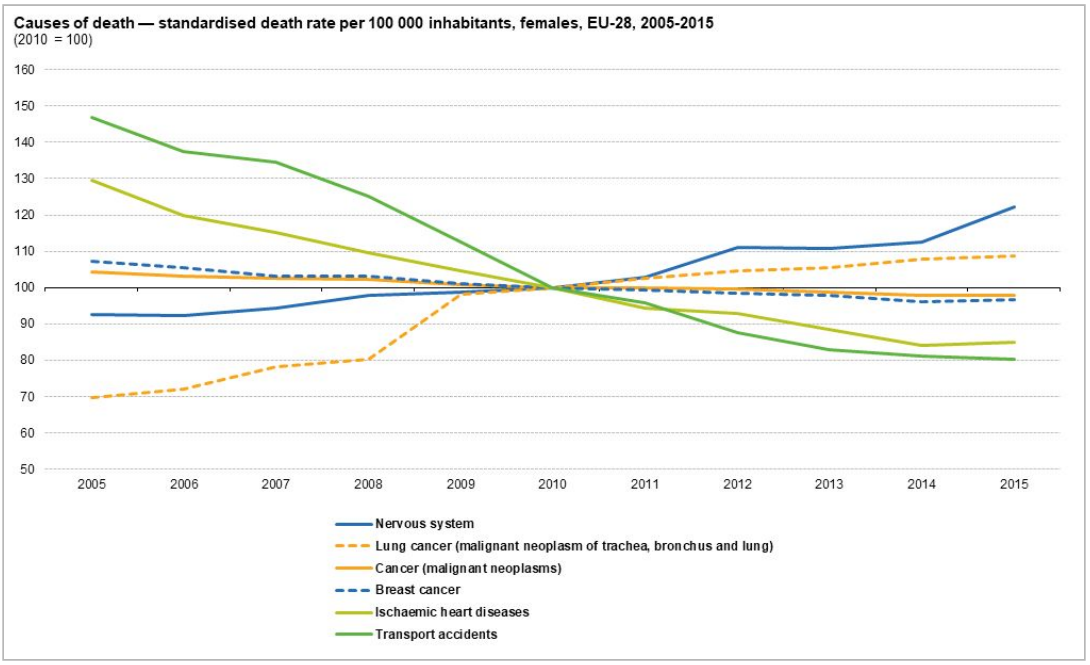


Image 14: Development of standardised death rate among males per cause between 2005-2015 (source: Eurostat)

The above conclusion can also be verified through the construction of appropriate cumulative graph visualizations that include the corresponding time margins. Hence, for the first four time margins that are associated with FP7 programme (i.e. 2007 - 2008, 2009 - 2010, 2011 - 2012 and 2013 - 2014), it is obvious that the IHD subclass is claiming the majority of participating nodes (coloured in blue).

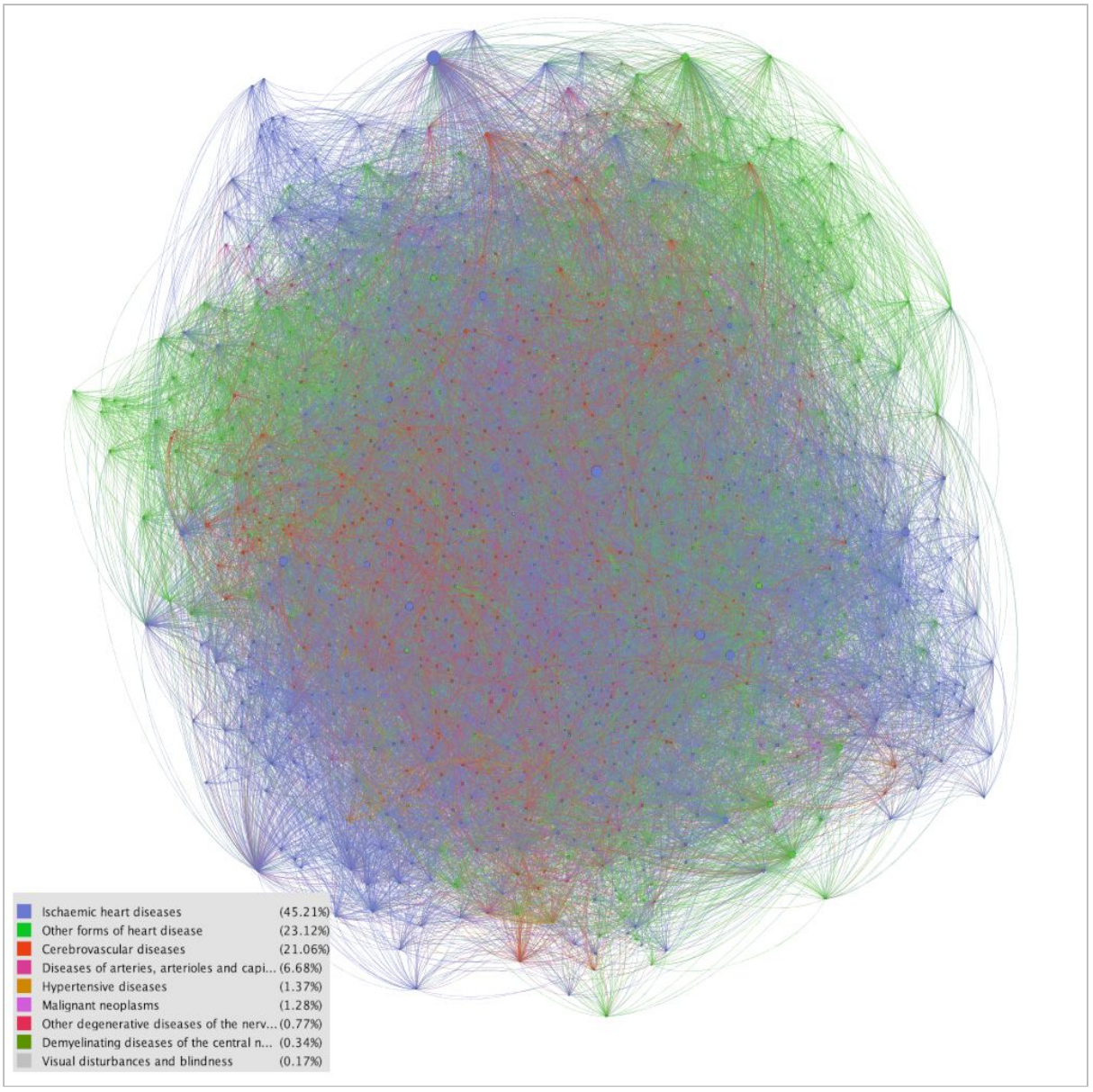


Image 15: Cumulative graph visualization for FP7 programme, time margin 2007-2014 and weight threshold 0.3

On the contrary, for the last two time margins that are associated with H2020 programme (i.e. 2015 - 2016 and 2017 - 2018), it is easy to detect the formation of the aforementioned CD community on the upper area of the relative graph (coloured in blue). As already stated, the members of this cluster seem to be densely connected which each other and compose a very cohesive group, with no evident existence of "bridges", i.e. nodes that exhibit relatively high betweenness centrality.

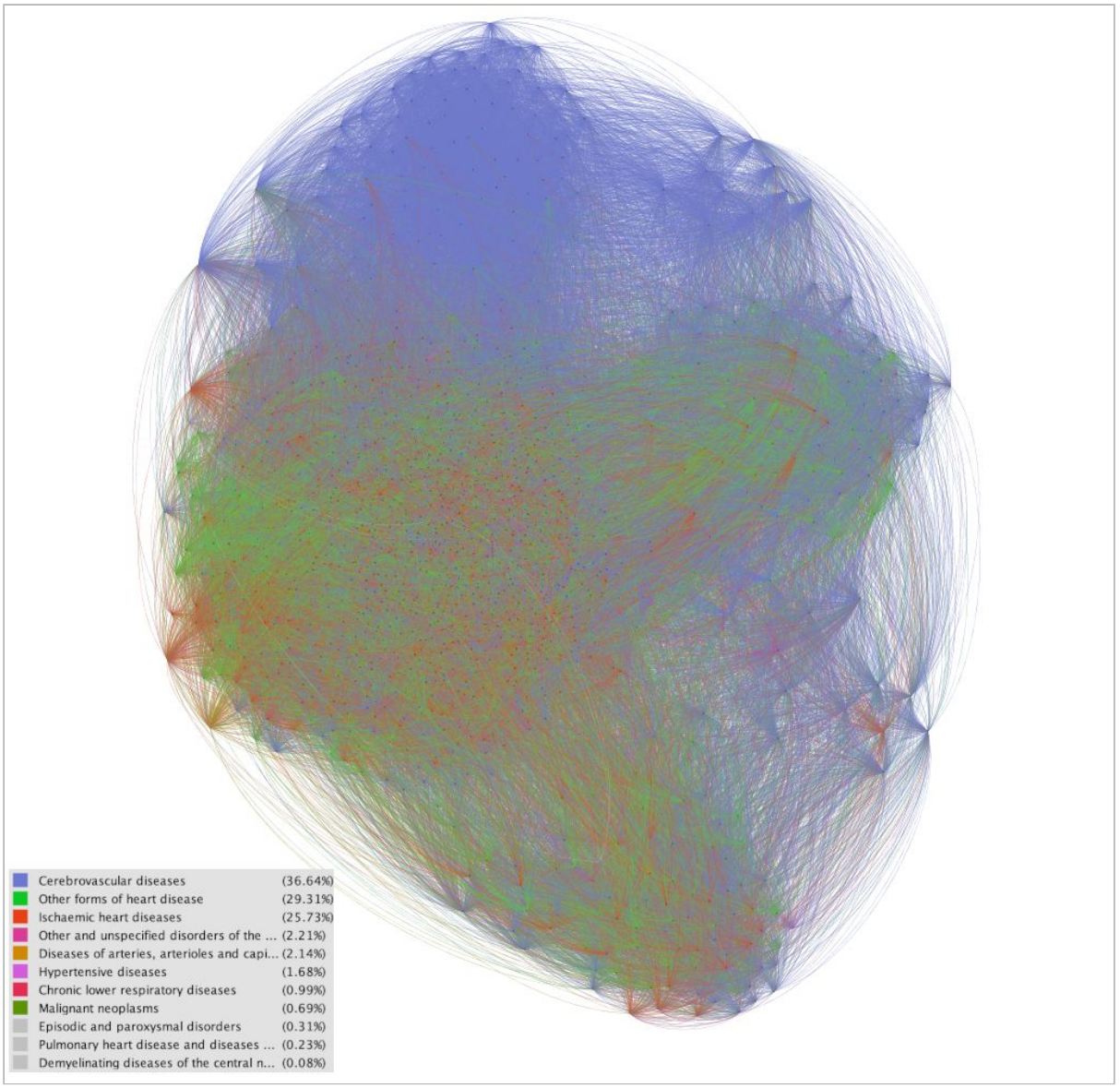


Image 16: Cumulative graph visualization for H2020 programme, time margin 2015-2018 and weight threshold 0.3

With regards to the formation and evolution of CNM communities across all time margins, the majority of countries associated with the most central nodes are those which appear to be more active in terms of participation in European research projects overall, namely Germany, the UK, Italy, France, Spain, Sweden, Switzerland, Austria and the Netherlands. However, organizations located on less active countries (such as Denmark, Hungary, Slovenia, Israel, Finland, Czech Republic, Belgium and Serbia), also seem to play an important role for specific time margins.

Furthermore, patterns of region-oriented collaborations cannot be observed, since the formatted clusters may consist of various country combinations that shift over time. Nonetheless, with the increasing infiltration of entities belonging to the private sector, a

different norm can be assumed, that organizations tend to collaborate more frequently with others coming from the same business sector.

More specifically, it can be observed that Pharmaceutical and Clinical Services companies (i.e. Pfizer, GlaxoSmithKline, Bayer, Novartis, Astrazeneca, Randox, BASF etc.) are usually present within the same communities. Similarly, companies with focus on Biotechnology, Robotics and Microelectronics (i.e. Philips, Siemens, STmicroelectronics, Guger Technologies, T-Systems ITC Iberia, Accel, The Shadow Robot Company, Contipro etc.) tend to stay together and form a cluster oriented in the discovery of new technological patents.

Perhaps these types of collaborations could result in market launching of health-related consumer products, that aim to directly tackle critical health-related challenges and thus producing high societal impact.

8. Conclusions and Future Work

In this diploma thesis, an attempt was made to assess the societal impact of health-related research in Europe for the past 10 years. In this context, a network approach was adopted and appropriate information was harnessed from a variety of projects funded under the European Union's Research and Innovation funding programmes, namely FP7 and Horizon 2020.

More specifically, the dataset corpus was leveraged in order to construct several graphs (where nodes represented named entities and edges depicted the relations between nodes), following the guidelines of the defined business scenario. Network analytics measures were employed so as to interpret these relations and investigate the formed clusters and communities within the graphs. Additionally, the constructed networks were examined across time in order to reveal and explore temporal trends and patterns.

The performed analysis produced a great volume of results, from simple tables and barcharts up to more advanced graph visualizations, out of which significant conclusions were drawn. Moreover, a multiline database holding the characteristics of nodes with the top measure scores per cluster was constructed, across all thresholds and margins. This information could be used as a starting point for further analysis regarding the top EU players in the health domain.

In terms of enhancements and points of improvement, some guidelines for future work are suggested below:

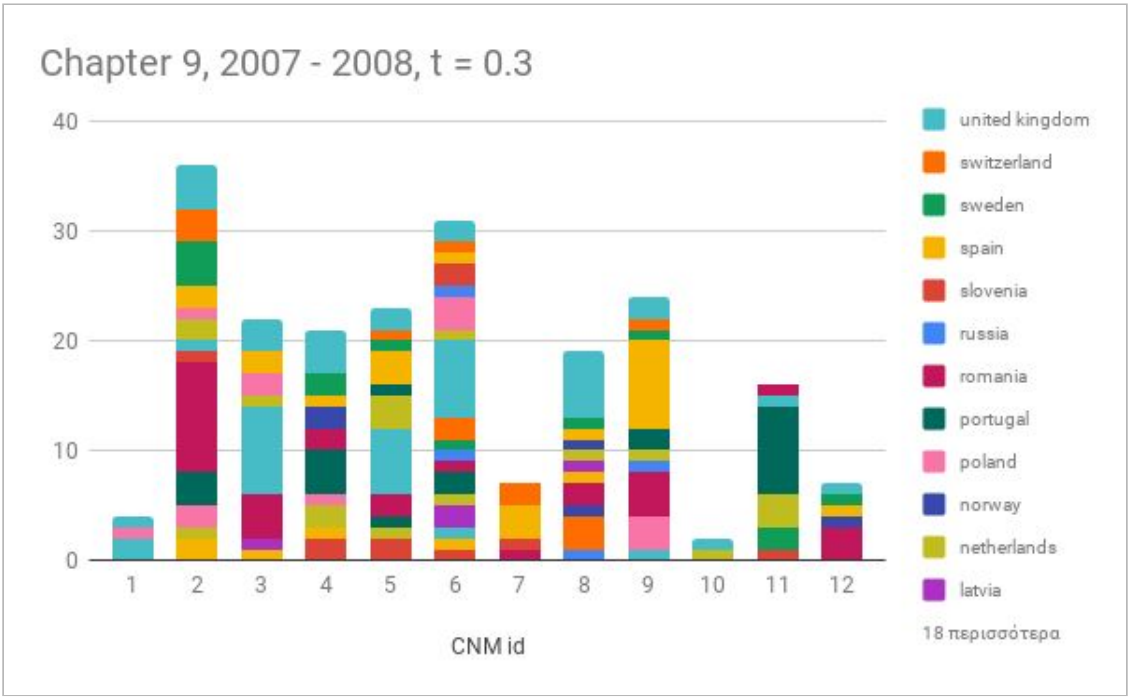
- Refinements on data parsing and preprocessing, in terms of improving the subclass association process with projects and/or organisations.
- Refinements on the definition of collaboration links, for example an edge should only be created when the corresponding organisations have actually worked together to produce a project deliverable.
- Development of alternative business scenarios in order to study different aspects of the domain, such as the evolution of research areas or the distribution of EU budget in space and time.
- Utilization of multilayer network theory during graph creation and analysis.

Appendix

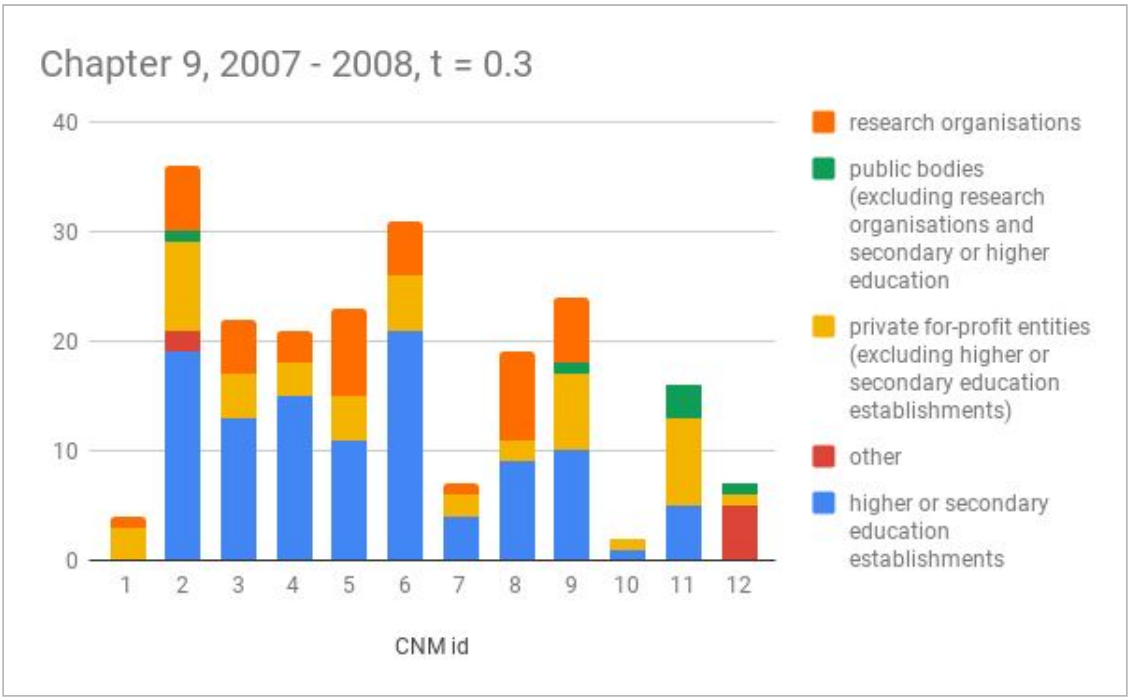
Ensemble of barcharts

2007 - 2008 margin, threshold 0.3

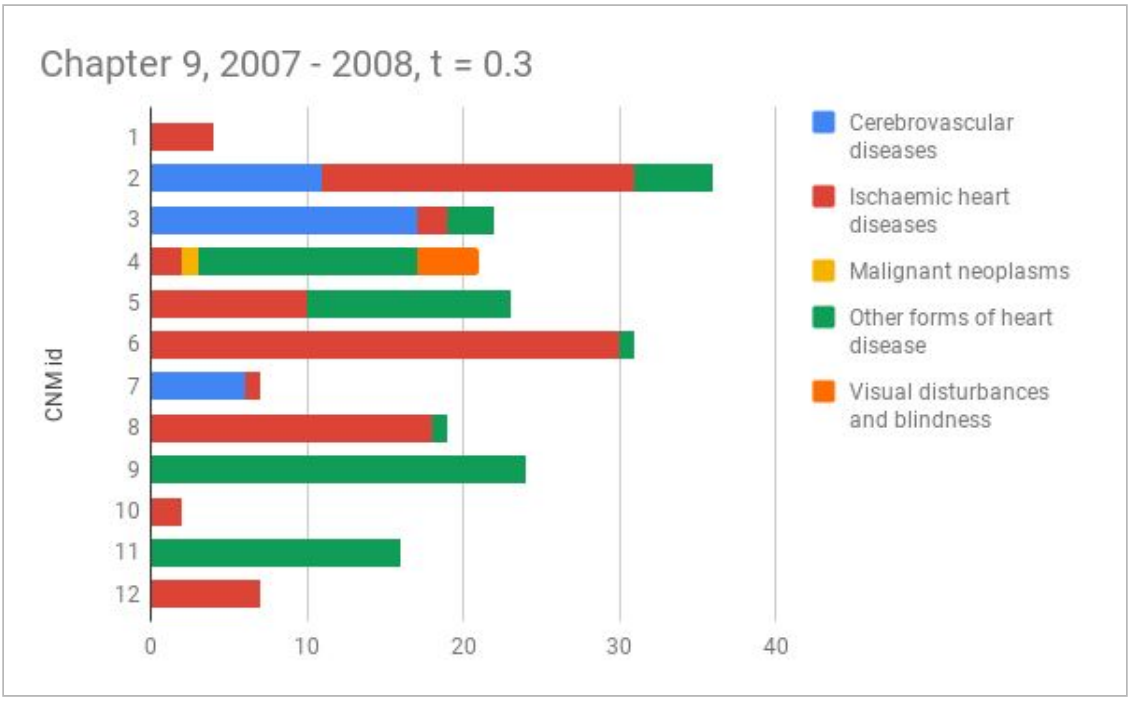
Country distribution per CNM community



Activity type distribution per CNM community

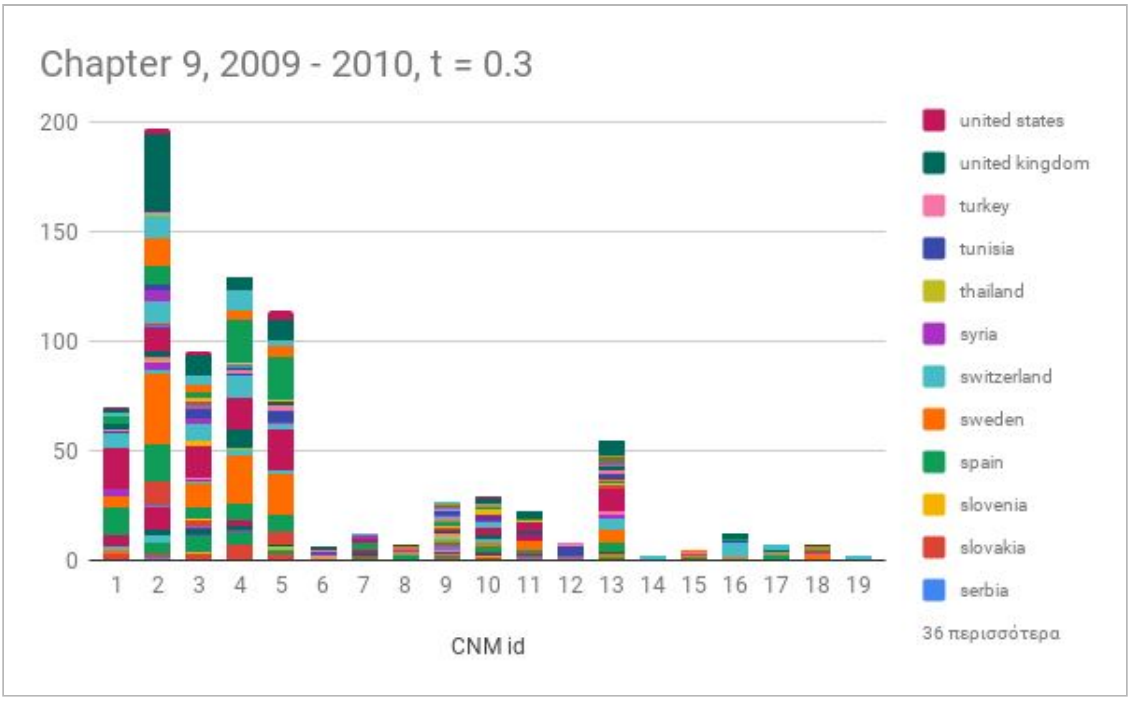


Subclass distribution per CNM community

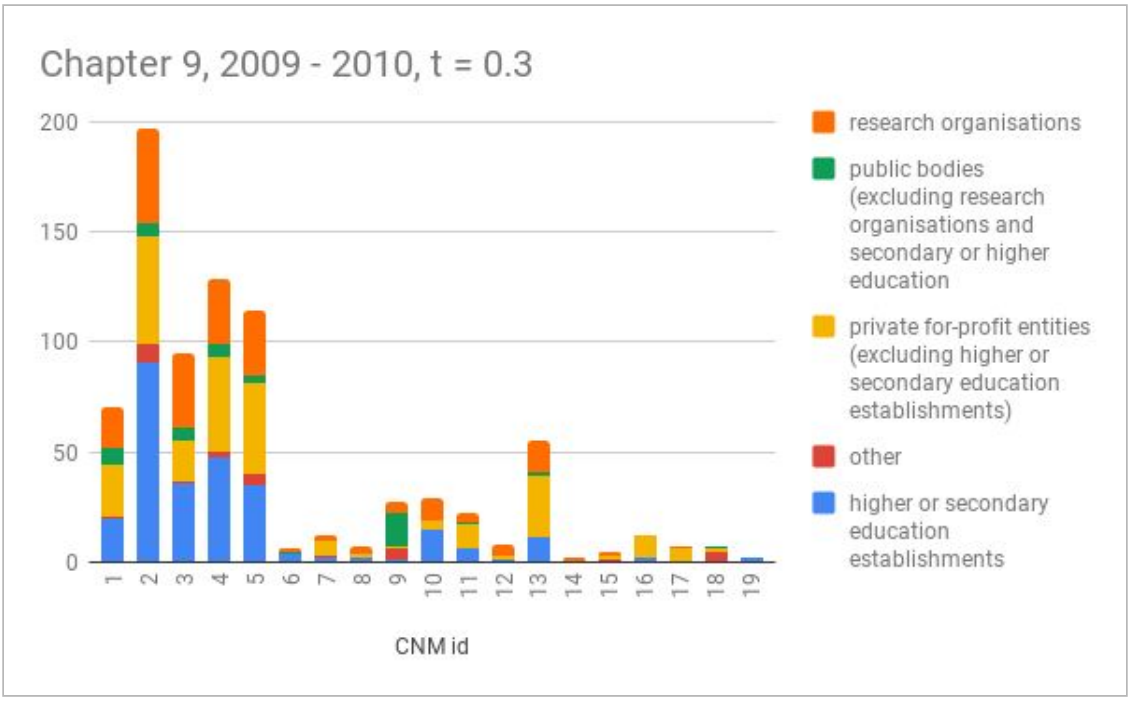


2009 - 2010 margin, threshold 0.3

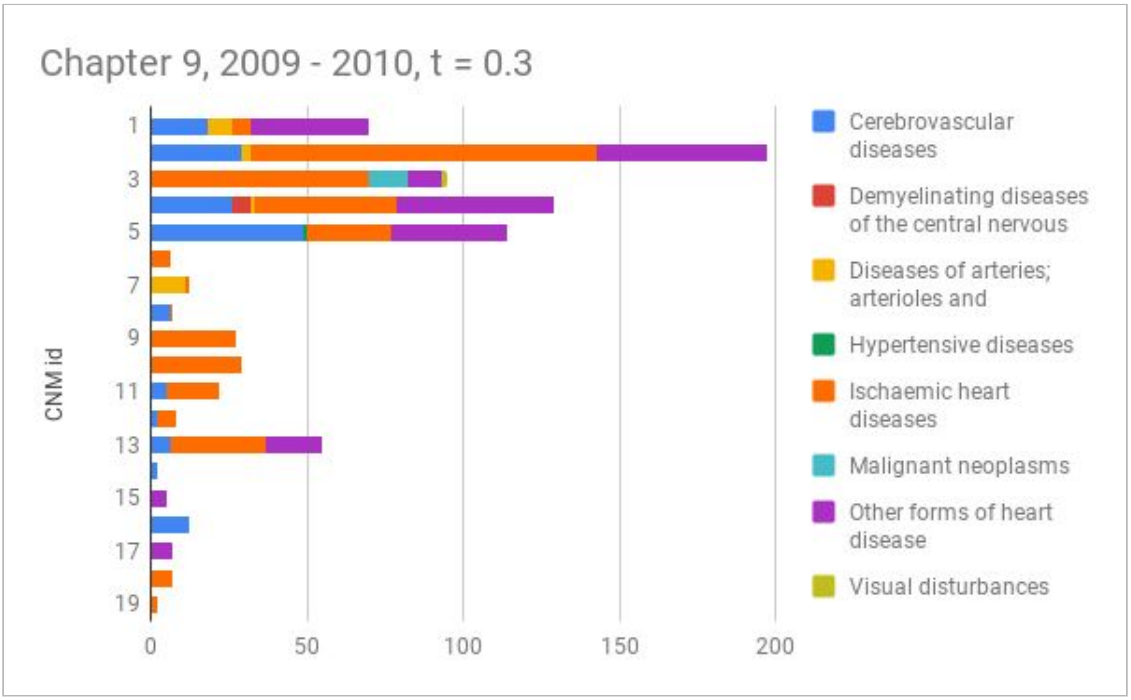
Country distribution per CNM community



Activity type distribution per CNM community

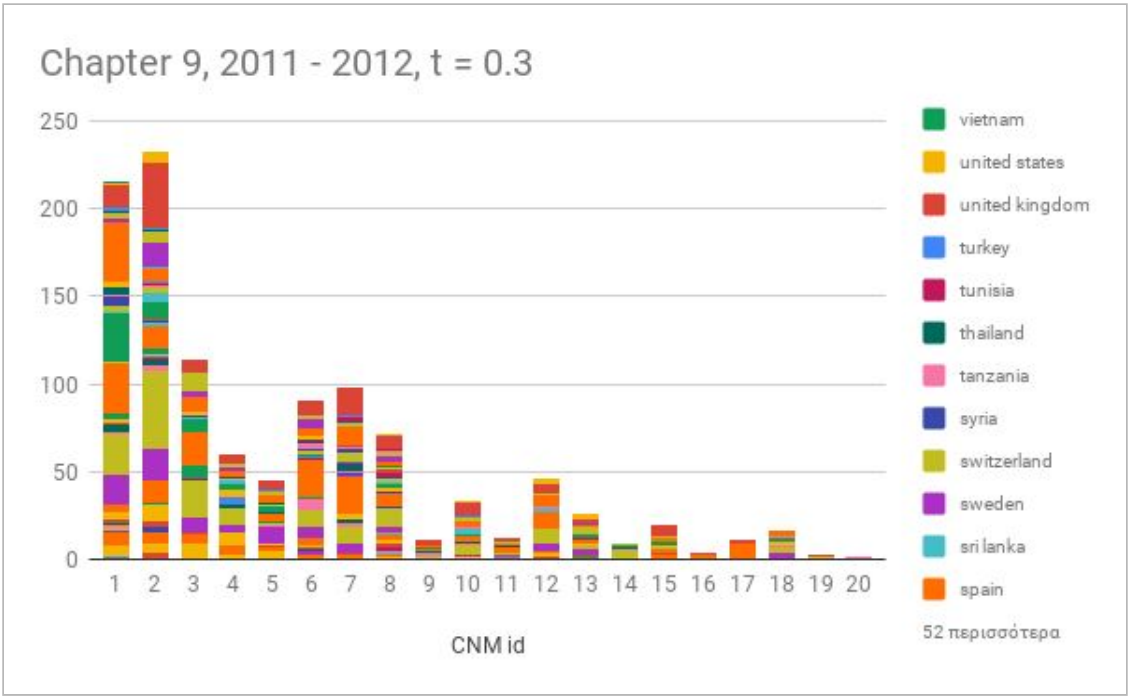


Subclass distribution per CNM community

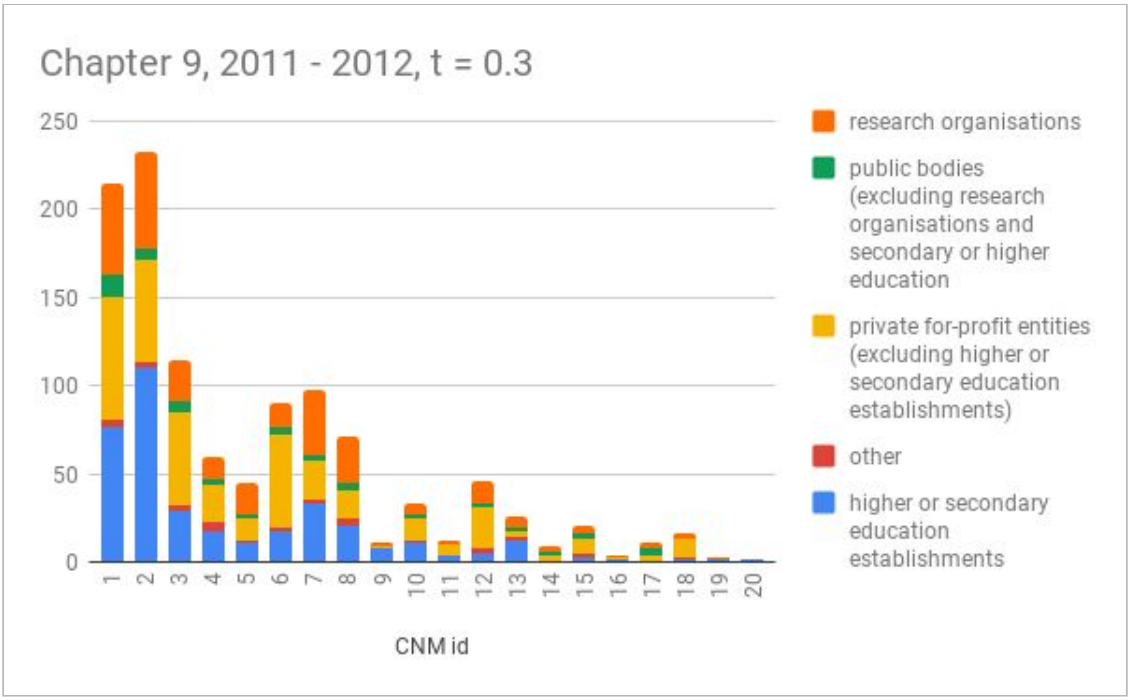


2011 - 2012 margin, threshold 0.3

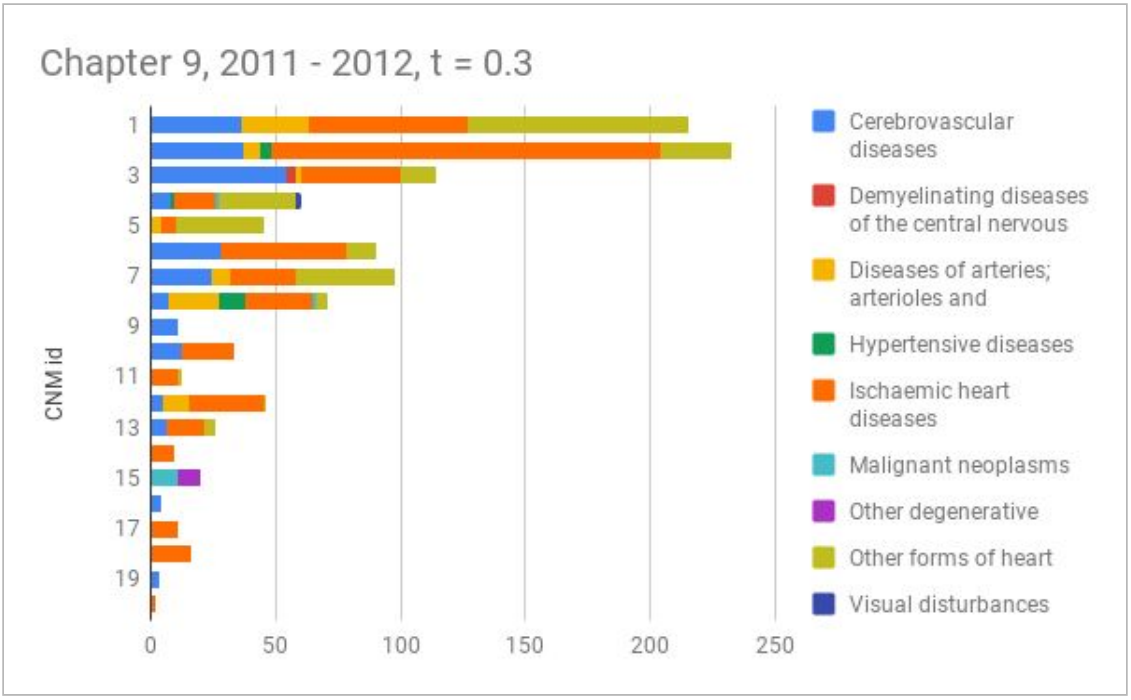
Country distribution per CNM community



Activity type distribution per CNM community

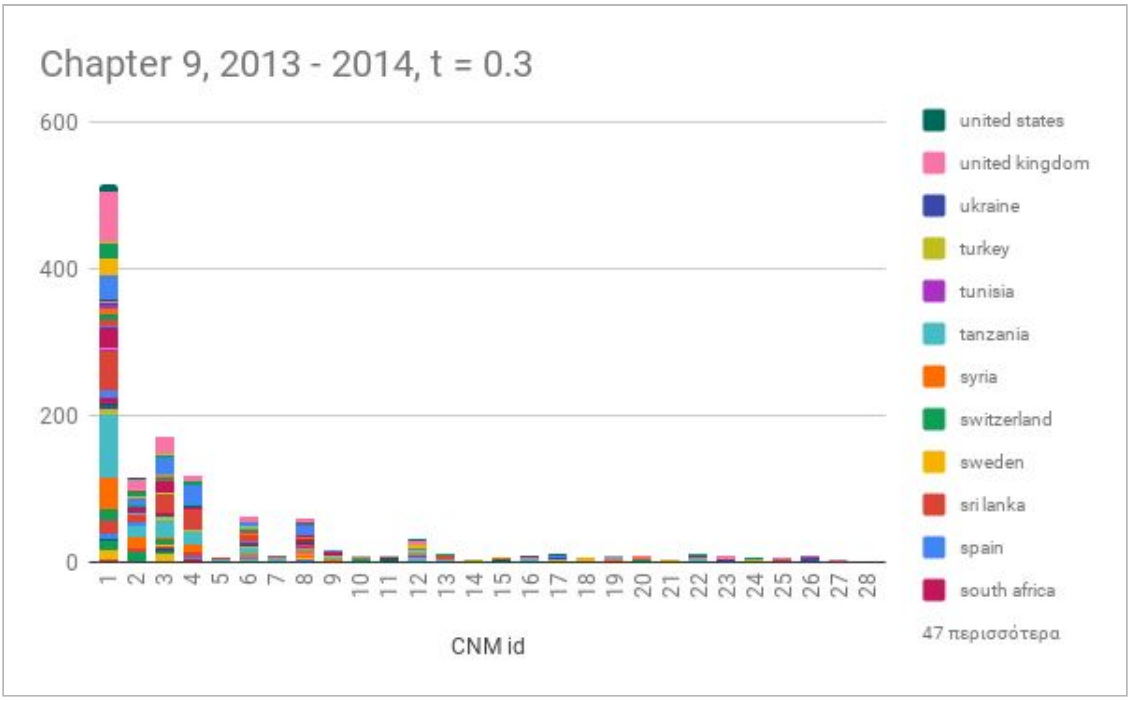


Subclass distribution per CNM community

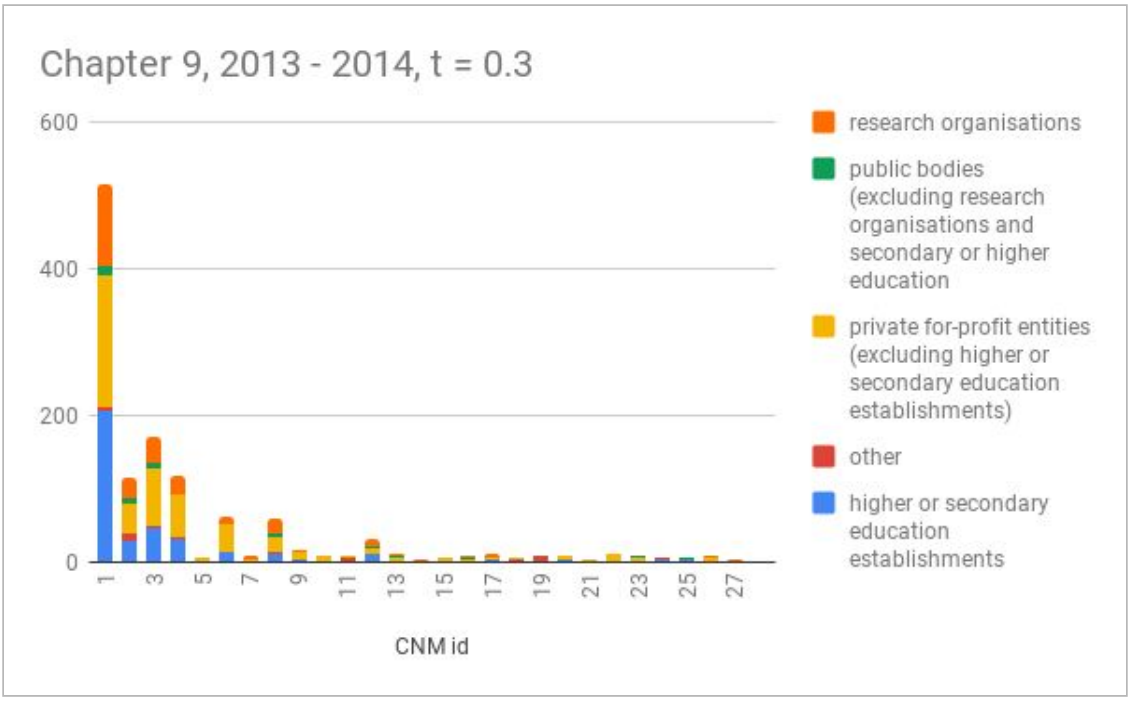


2013 - 2014 margin, threshold 0.3

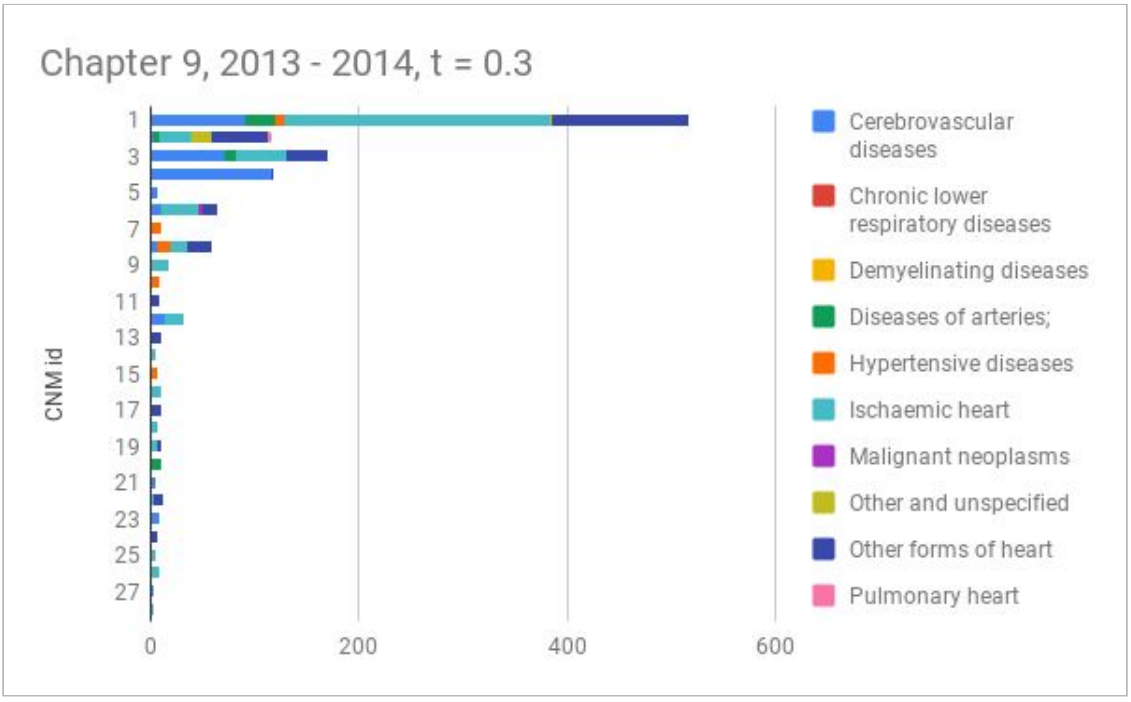
Country distribution per CNM community



Activity type distribution per CNM community

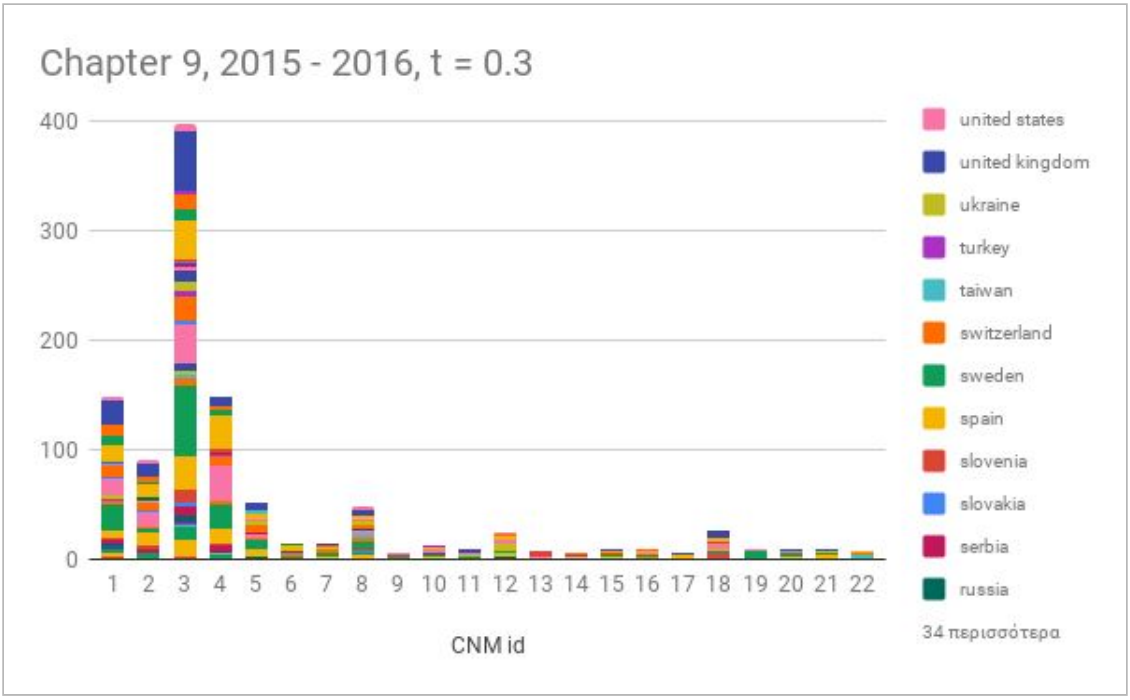


Subclass distribution per CNM community

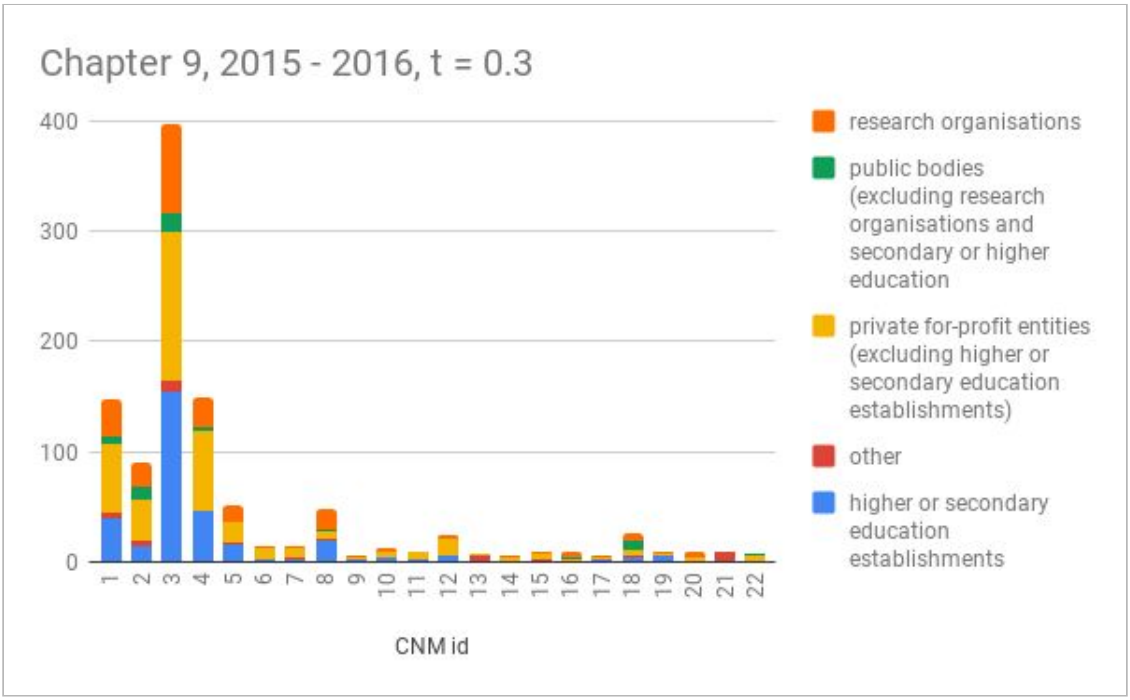


2015 - 2016 margin, threshold 0.3

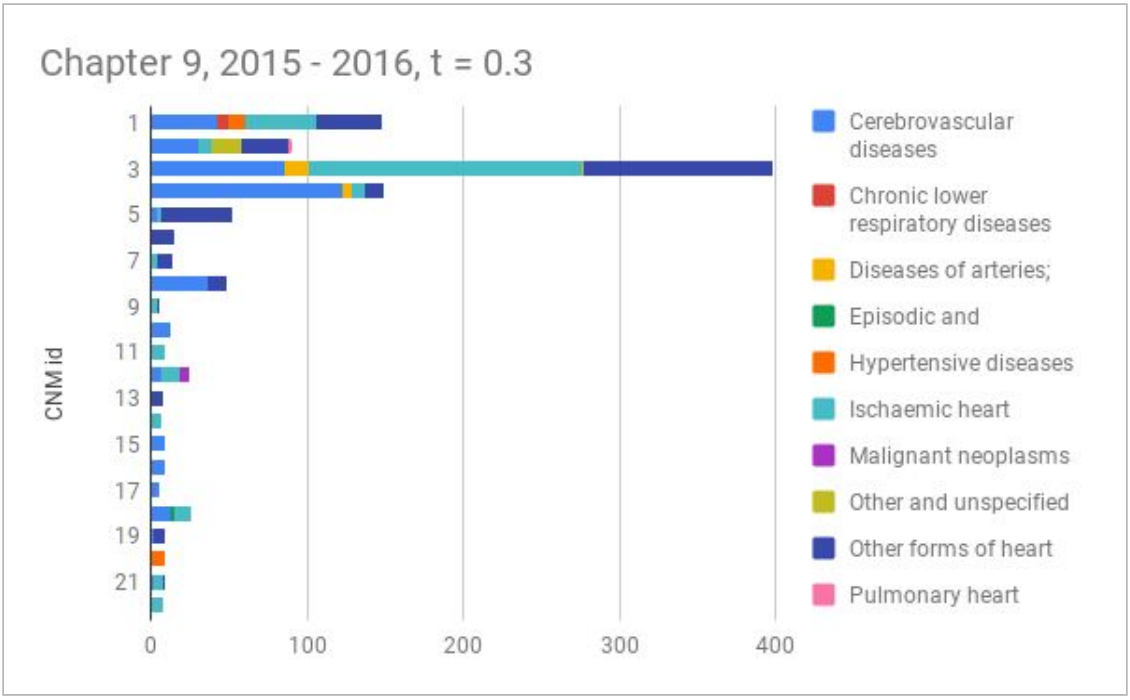
Country distribution per CNM community



Activity type distribution per CNM community

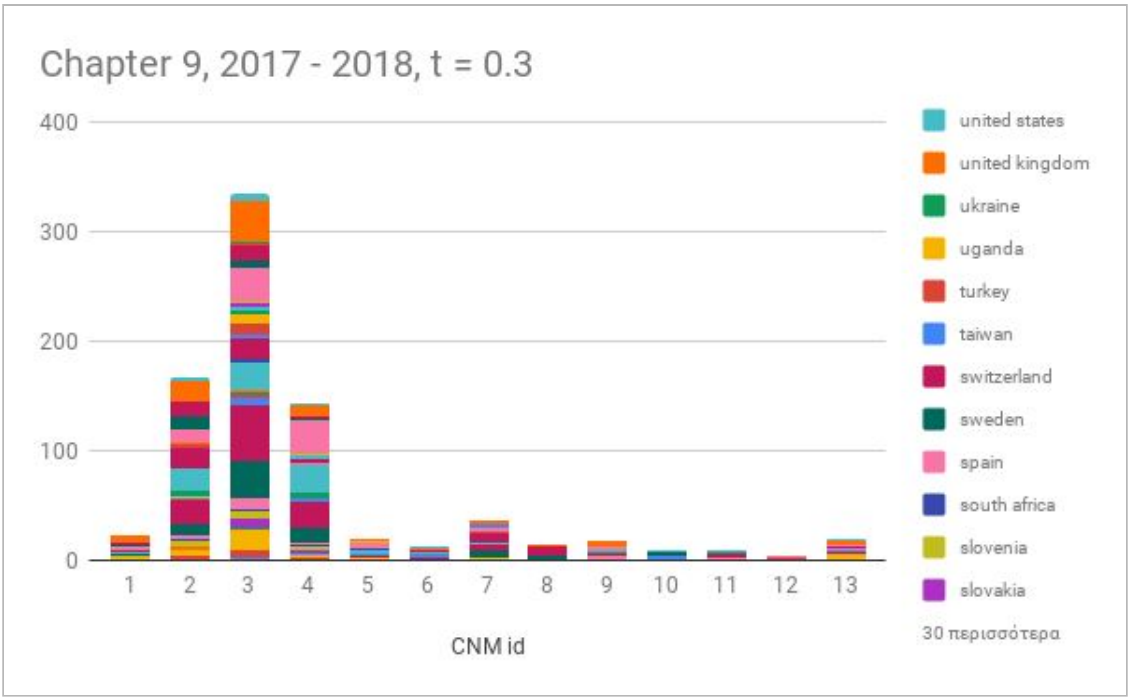


Subclass distribution per CNM community

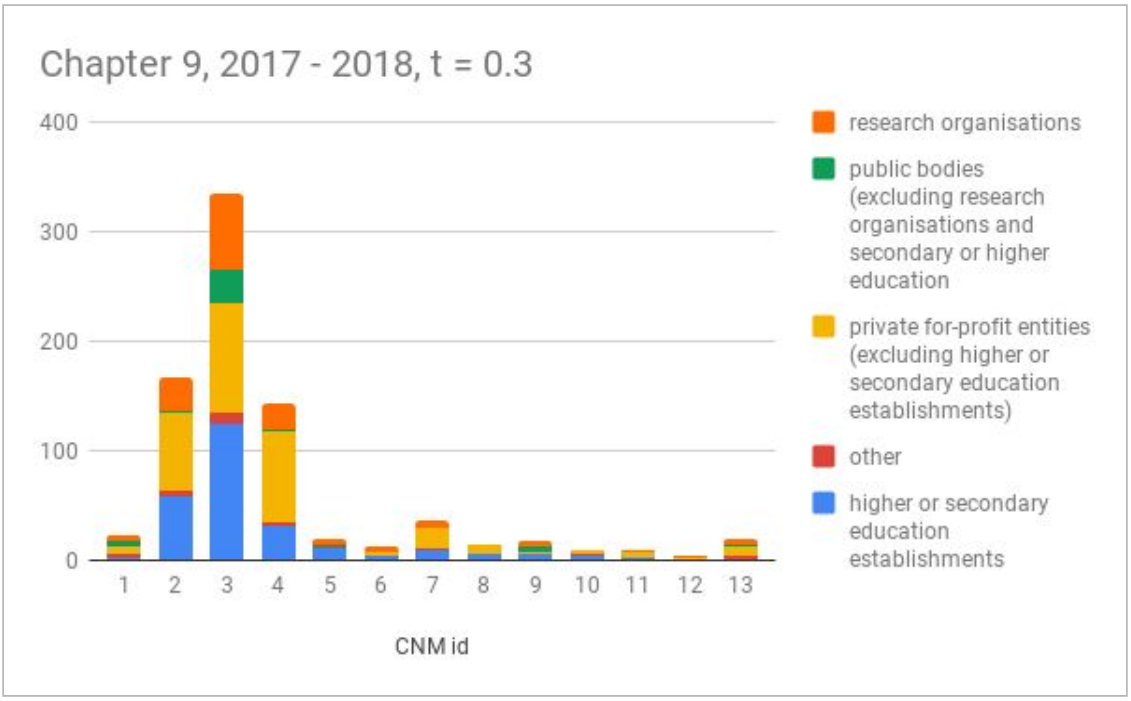


2017 - 2018 margin, threshold 0.3

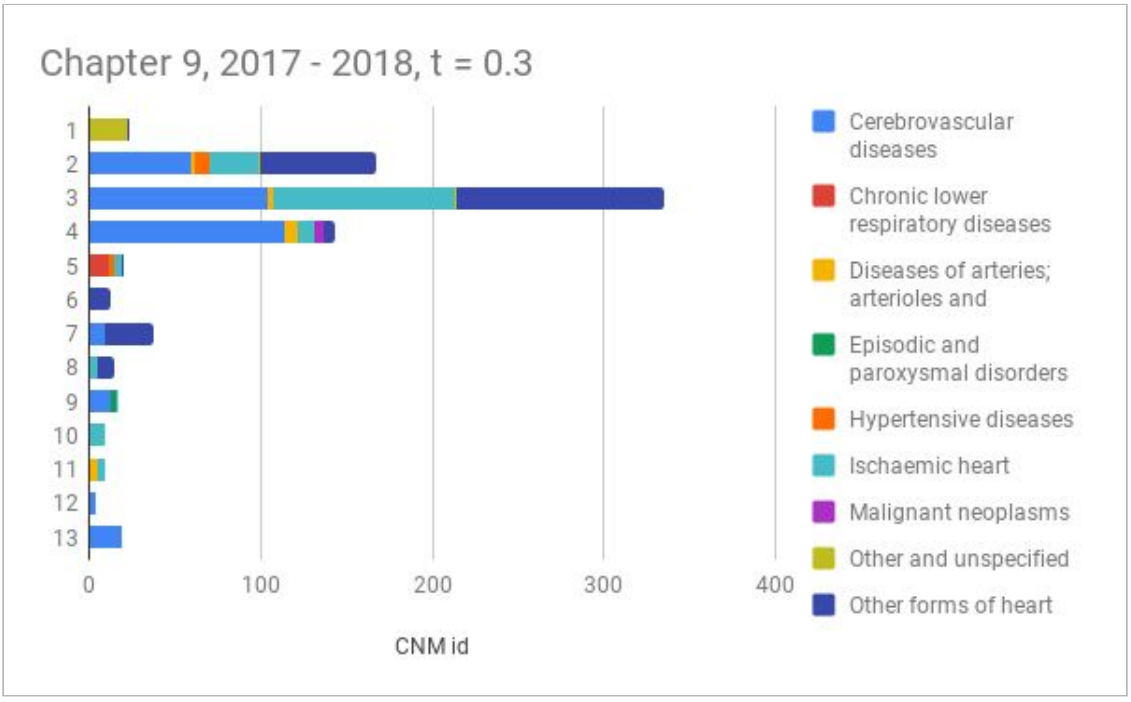
Country distribution per CNM community



Activity type distribution per CNM community



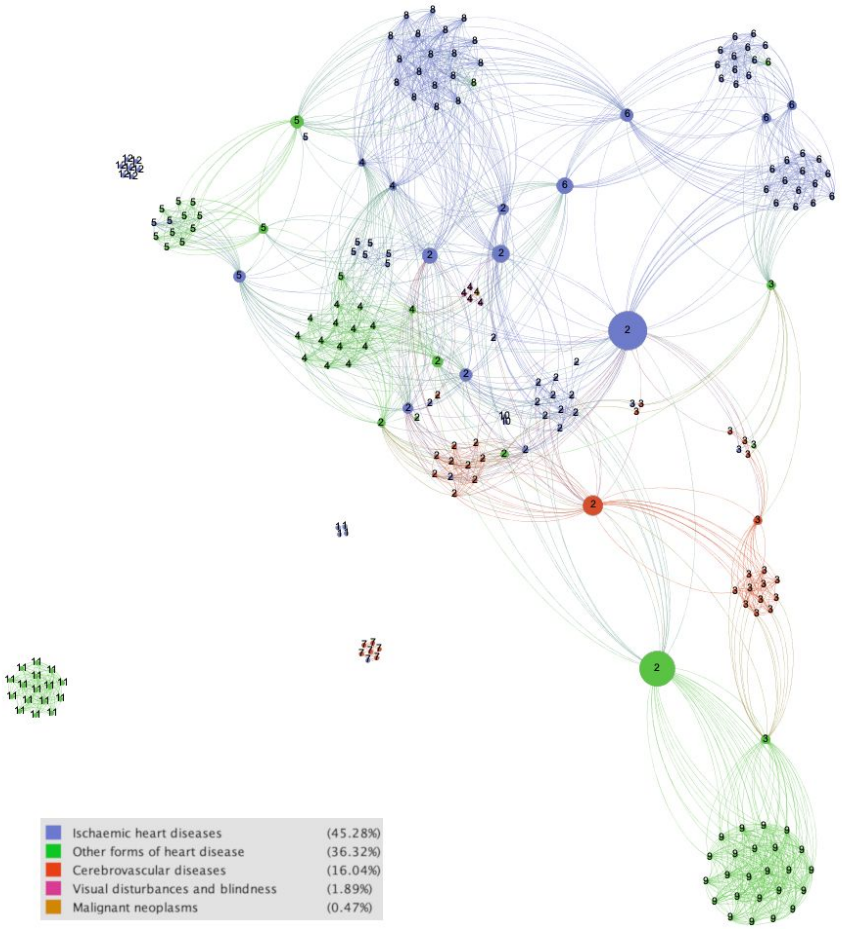
Subclass distribution per CNM community



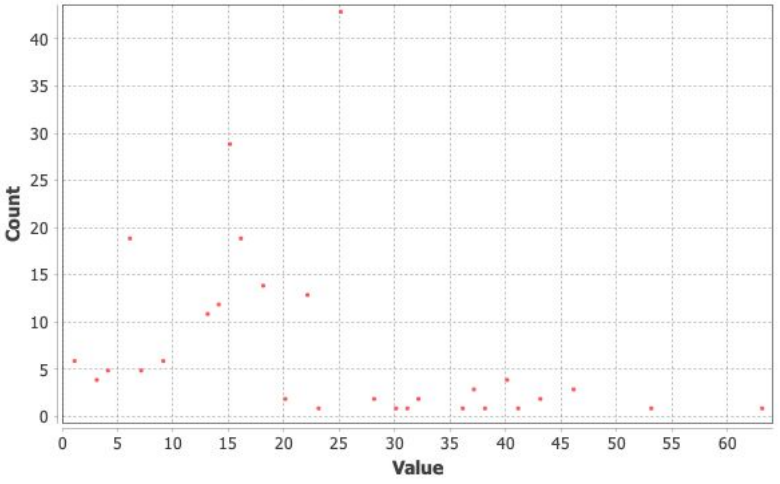
Ensemble of graph illustrations

2007 - 2008 margin, threshold 0.3

Collaboration network

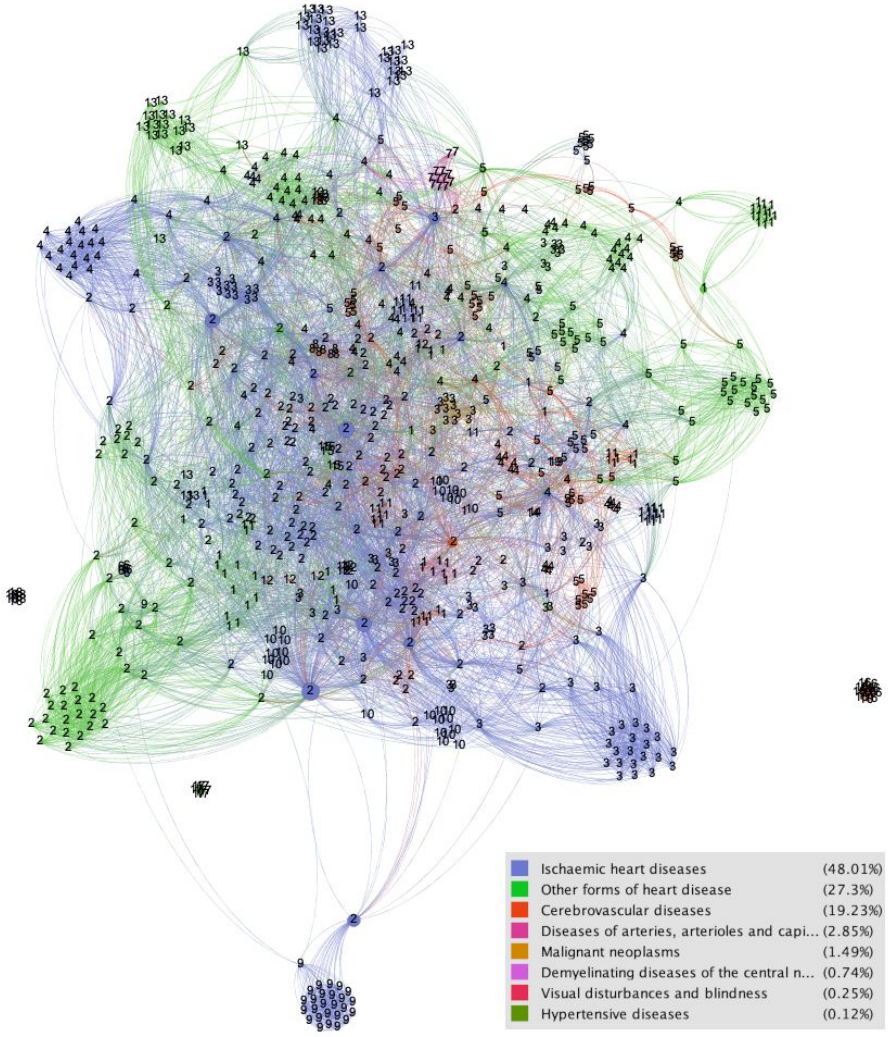


Degree distribution

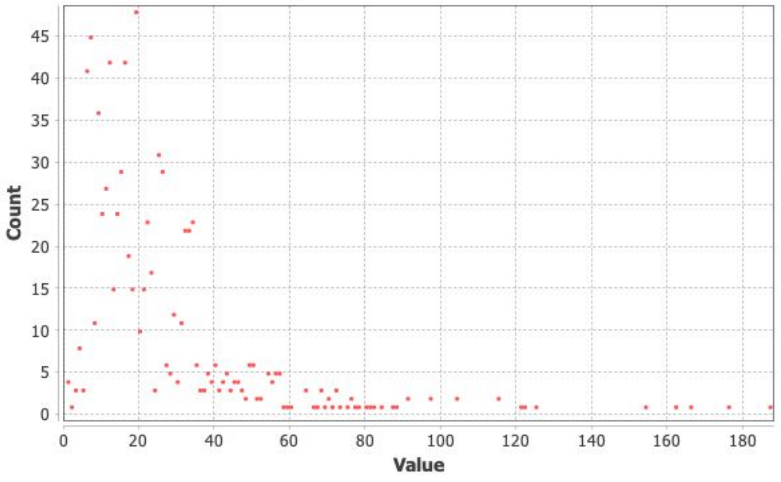


2009 - 2010 margin, threshold 0.3

Collaboration network

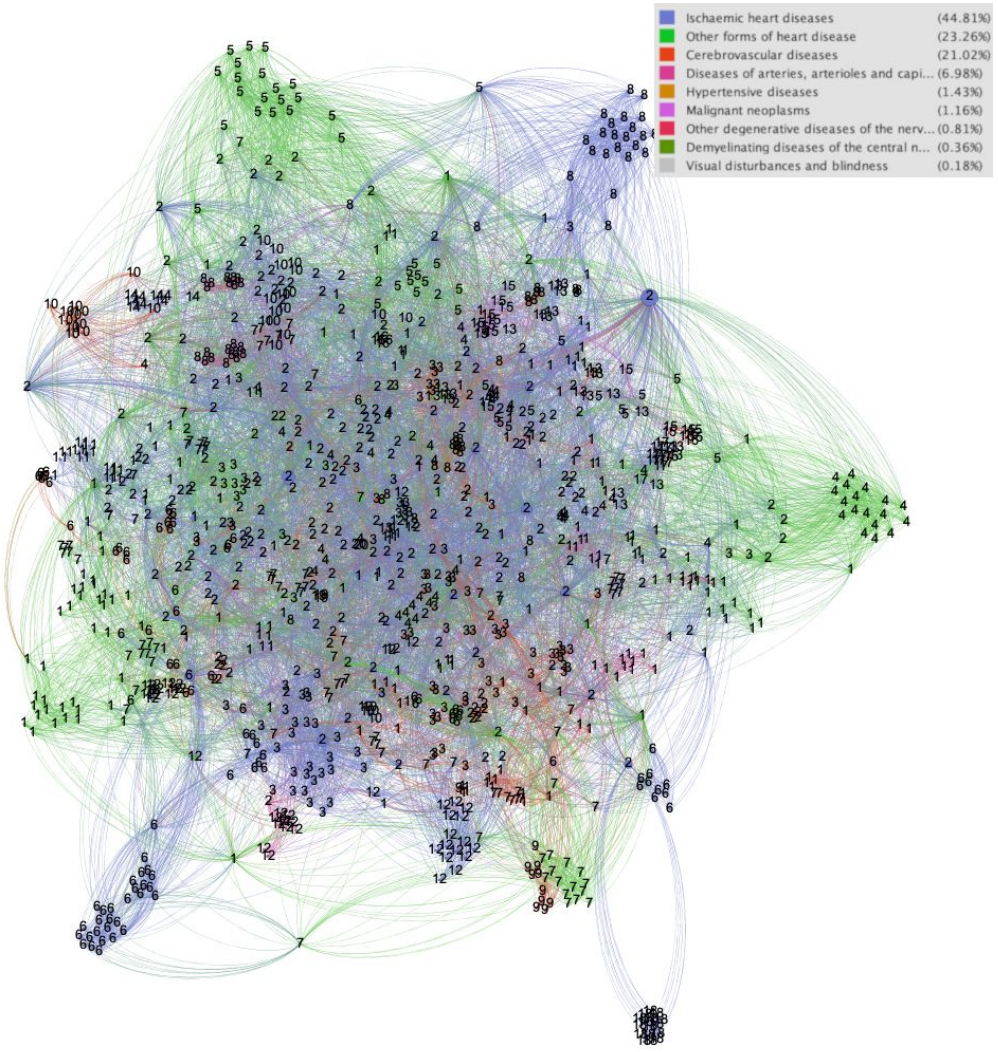


Degree distribution

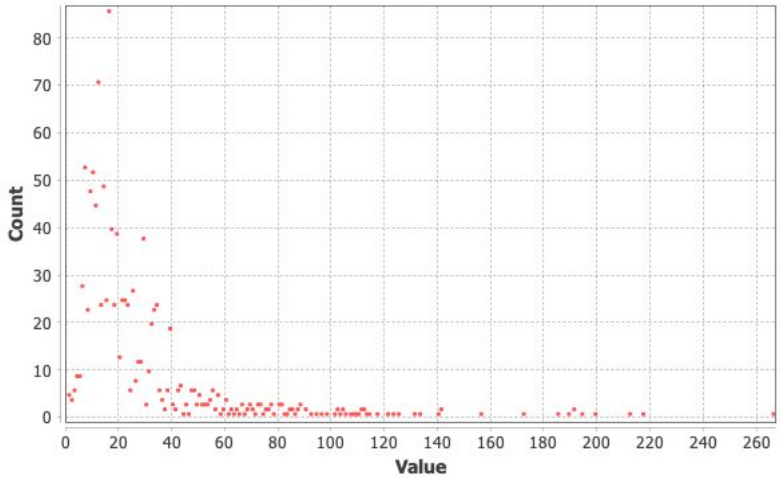


2011 - 2012 margin, threshold 0.3

Collaboration network

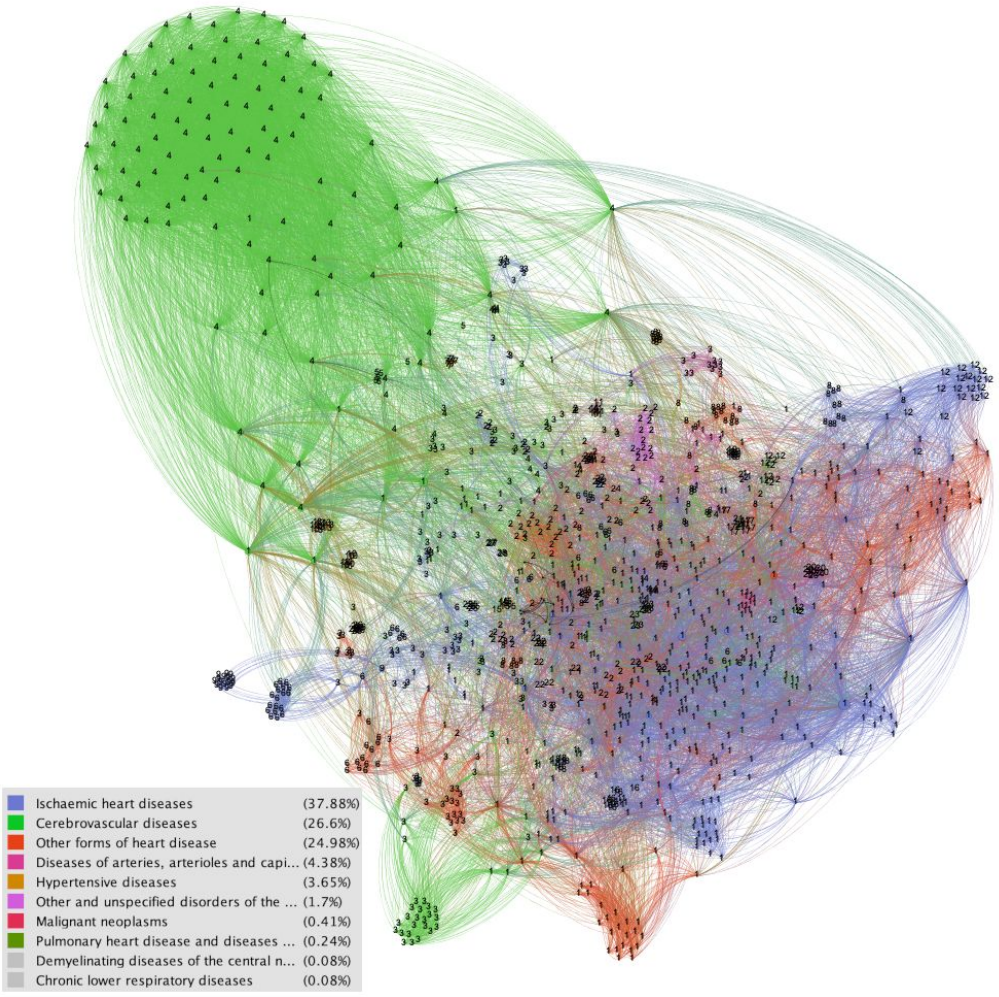


Degree distribution

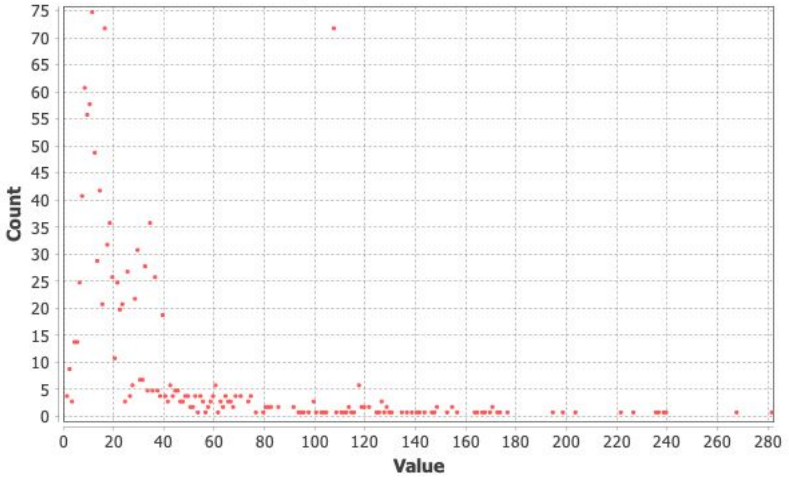


2013 - 2014 margin, threshold 0.3

Collaboration network

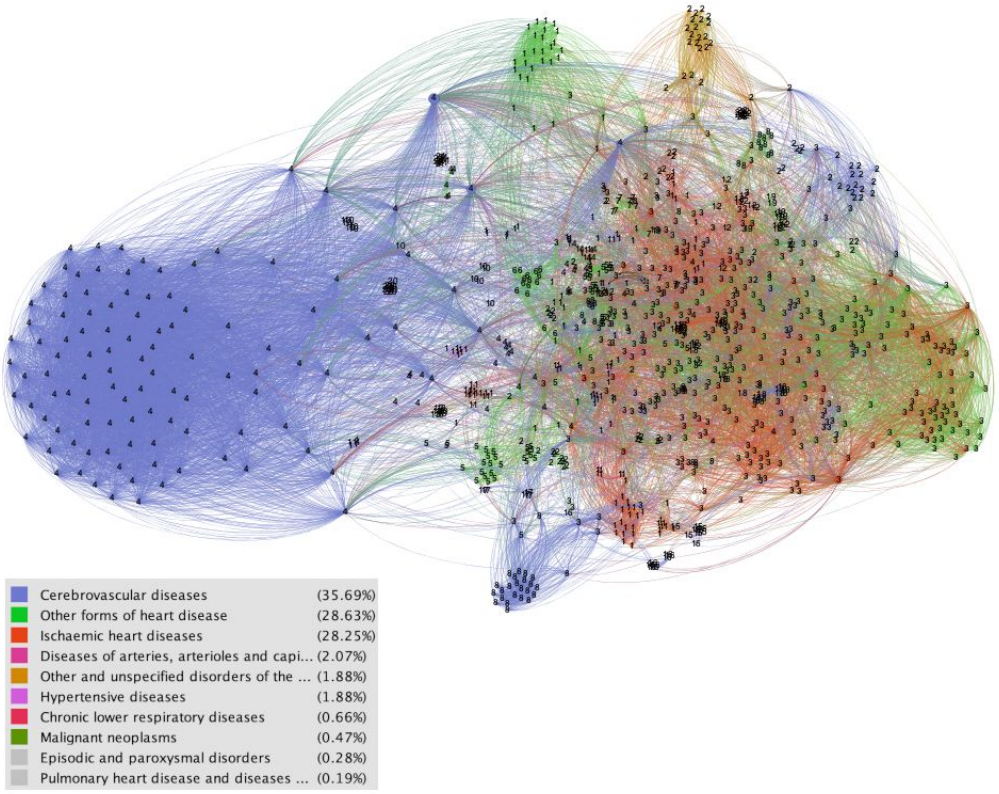


Degree distribution

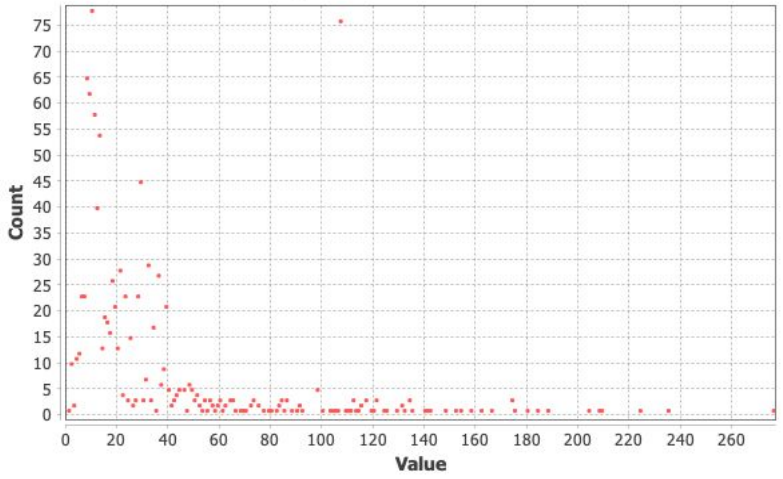


2015 - 2016 margin, threshold 0.3

Collaboration network

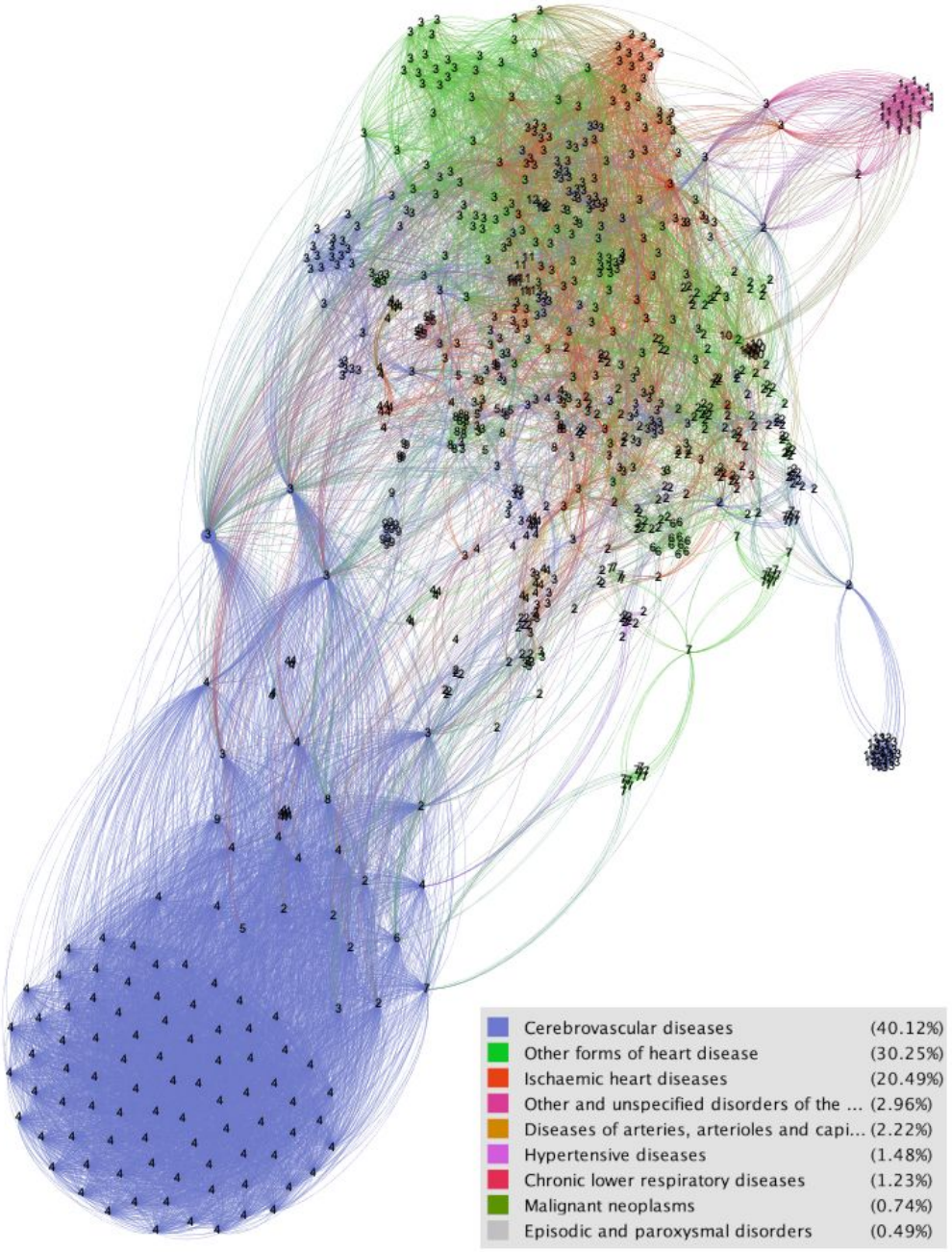


Degree distribution

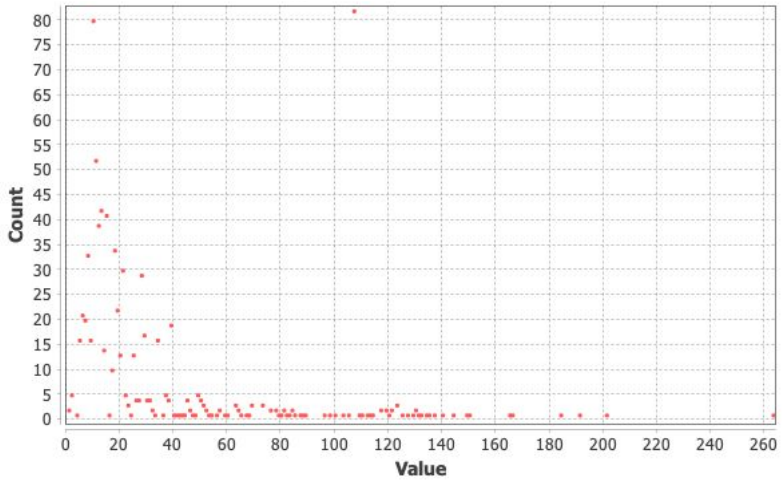


2017 - 2018 margin, threshold 0.3

Collaboration network



Degree distribution



References

"About - Data4Impact." n.d. Data4Impact. Accessed October 6, 2018. http://www.data4impact.eu/about/.

"Annual Reports." n.d. Accessed January 19, 2019. http://www.ehnheart.org/annual-reports.html.

"Bastian." n.d. Accessed November 30, 2018. https://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

"Betweenness Centrality." n.d. Accessed November 25, 2018. https://www.sci.unich.it/~francesc/teaching/network/betweeness.html.

"Cancer Statistics - Specific Cancers - Statistics Explained." n.d. Accessed January 30, 2019. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Cancer_statistics_-_specific_cancers#Colorectal_cancer.

Clauset, Aaron, M. E. J. Newman, and Cristopher Moore. 2004. "Finding Community Structure in Very Large Networks." *Physical Review E* 70 (6). https://doi.org/10.1103/physreve.70.066111.

"Closeness Centrality." n.d. Accessed November 25, 2018. https://www.sci.unich.it/~francesc/teaching/network/closeness.html.

Contributors to Wikimedia projects. 2006. "Community Structure - Wikipedia." Wikimedia Foundation, Inc. December 5, 2006. https://en.wikipedia.org/wiki/Community_structure.

———. 2008. "Modularity (networks) - Wikipedia." Wikimedia Foundation, Inc. July 11, 2008. https://en.wikipedia.org/wiki/Modularity_(networks).

"CORDIS | European Commission." n.d. Publication Office/CORDIS. Accessed November 28, 2018. https://cordis.europa.eu/.

"Database - Eurostat." n.d. Accessed January 30, 2019. https://ec.europa.eu/eurostat/web/health/causes-death/data/database.

De Domenico, Manlio, Clara Granell, Mason A. Porter, and Alex Arenas. 2016. "The Physics of Spreading Processes in Multilayer Networks." *Nature Physics* 12 (10): 901–6.

De Domenico, Manlio, Albert Solé-Ribalta, Emanuele Cozzo, Mikko Kivelä, Yamir Moreno, Mason A. Porter, Sergio Gómez, and Alex Arenas. 2013. "Mathematical Formulation of Multilayer Networks." *Physical Review* X 3 (4). https://doi.org/10.1103/physrevx.3.041022.

De Domenico, Manlio, Albert Solé-Ribalta, Elisa Omodei, Sergio Gómez, and Alex Arenas. 2015. "Ranking in Interconnected Multilayer Networks Reveals Versatile Nodes." *Nature Communications* 6 (April): 6868.

"Degree Centrality." n.d. Accessed November 25, 2018. https://www.sci.unich.it/~francesc/teaching/network/degree.html.

Easley, David, and Jon Kleinberg. 2010a. *Networks, Crowds, and Markets by David Easley*. Cambridge University Press.

———. 2010b. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press.

"Eccentricity." n.d. Accessed February 3, 2019. http://www.cbmc.it/fastcent/doc/Eccentricity.htm.

"Eigenvector Centrality." n.d. Accessed November 25, 2018. https://www.sci.unich.it/~francesc/teaching/network/eigenvector.html.

European Patent Office. n.d. "Home." Accessed November 28, 2018. http://www.epo.org.

"Facts about Google and Competition." n.d. Accessed November 25, 2018. https://web.archive.org/web/20111104131332/https://www.google.com/competition/howgooglesearchworks.html.

Farthing, Michael, Stephen E. Roberts, David G. Samuel, John G. Williams, Kymberley Thorne, Sian Morrison-Rees, Ann John, Ashley Akbari, and Judy C. Williams. 2014. "Survey of Digestive Health across Europe: Final Report. Part 1: The Burden of Gastrointestinal Diseases and the Organisation and Delivery of Gastroenterology Services across Europe." *United European Gastroenterology Journal* 2 (6): 539.

"Gephi - The Open Graph Viz Platform." n.d. Accessed November 30, 2018. https://gephi.org/.

"Health, Demographic Change and Wellbeing - Horizon 2020 - European Commission." 2018. Horizon 2020. June 10, 2018. http://ec.europa.eu/programmes/horizon2020/en/h2020-section/health-demographic-

change-and-wellbeing#Article.
"Home - Data4Impact." n.d. Data4Impact. Accessed October 6, 2018. http://www.data4impact.eu/.
"Home Page - FP7 - Research - Europa." n.d. Accessed October 6, 2018. http://ec.europa.eu/research/fp7/index_en.cfm.
"Home - PMC - NCBI." n.d. Accessed November 28, 2018. https://www.ncbi.nlm.nih.gov/pmc/.
"Horizon 2020 - European Commission." 2018. Horizon 2020. June 10, 2018. https://ec.europa.eu/programmes/horizon2020/en/.
"Hubs, Authorities, and Communities." n.d. Accessed February 3, 2019. http://cs.brown.edu/memex/ACM_HypertextTestbed/papers/10.html.
"ICD-10 Version:2016." n.d. Accessed December 15, 2018. https://icd.who.int/browse10/2016/en#/IX.
Jackson, Matthew O. 2010. *Social and Economic Networks*. Princeton University Press.
Kivelä, Mikko, Alex Arenas, Marc Barthelemy, James P. Gleeson, Yamir Moreno, and Mason A. Porter. 2014. "Multilayer Networks." *Journal of Complex Networks* 2 (3): 203–71.
Kivela, Mikko, Alex Arenas, Marc Barthelemy, James P. Gleeson, Yamir Moreno, and Mason Alexander Porter. 2013. "Multilayer Networks." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2341334.
Leskovec, Jure, and Rok Sosič. 2016. "SNAP: A General Purpose Network Analysis and Graph Mining Library." ACM *Transactions on Intelligent Systems and Technology* 8 (1). https://doi.org/10.1145/2898361.
Li, Wenye, and Dale Schuurmans. 2011. "Modular Community Detection in Networks." In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two*, 1366–71. AAAI Press.
"Farness Centrality." n.d. Accessed November 25, 2018. http://www.jsums.edu/nmeghanathan/files/2015/08/CSC641-Fall2015-Module-2-Centrality-Measures.pdf?x61976.
"Project Deliverables - Data4Impact." n.d. Data4Impact. Accessed November 28, 2018. http://www.data4impact.eu/project-deliverables/.
"Snap.py - SNAP for Python." n.d. Accessed November 25, 2018. https://snap.stanford.edu/snappy/index.html.
"Statistics Explained." n.d. Accessed January 19, 2019. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=File:Major_causes_of_death,_EU-28,_2015_(standardised_death_rates_per_100_000_inhabitants)_HLTH18.png.
"Burden of liver disease in Europe" n.d. Accessed February 3, 2019. https://www.journal-of-hepatology.eu/article/S0168-8278(18)32057-9/pdf.
"WHO | International Classification of Diseases, 11th Revision (ICD-11)." 2018, November. http://www.who.int/classifications/icd/en/.