University of Turin

Doctoral School of Sciences and Innovative Technologies
PhD Program in Computer Science
XXXV cycle

PhD Dissertation
Alessandra Urbinati

Analysis of socio-technical systems: from data to complex networks models

Advisor
Prof. Giancarlo Ruffo
Università degli Studi del Piemonte Orientale "A. Avogadro", Italy

PhD Coordinator
Prof. Viviana Patti

Academic Year 2022-2023

This thesis has been revised and positively evaluated, considering it admissible for the defense, by Prof. Michele Coscia and Dr. Luca Pappalardo.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Nature offers countless examples of how initially chaotic situations develop over time in more ordered states. "Why" and "how" this happens are key questions to a lot of disciplines, ranging from Social Science to Biology, Physics, or Artificial Intelligence [1]. To answer these questions we need to access the information stored in the system we want to study. For many natural systems, this information is the result of millions, if not billions, of years of evolution, and the structures which contain them are not simple, even if we decompose them in elementary parts we cannot predict the collective phenomena which arise from their interplay. These systems are complex.

Complex systems are rooted in almost every aspect of life, and social science has offered in the last years ample possibilities for computational applications. However, it is becoming increasingly clear that many of the systems on which the health, wealth, and security of our society depend are neither purely social nor purely computational. They are socio-technical systems. Workplace relationships, economic trade, media markets, health delivery systems, or even criminal justice organizations are all increasingly characterized by a complex mixture of human actors and institutions on the one hand, or digital platforms and algorithms on the other hand. For example, on Twitter every kind of interaction must follow some ground rules, there are specific types of engagement and also content sharing is governed by preferential logic implemented by an algorithm whose development we are not even familiar with. So the human aspect is embedded in a technical plot of guidelines. The Twitter platform is a very recent example, but throughout history, we can identify many instances where human interactions have been mixed with technical rules, just think of all commercial exchange relationships, where bureaucracy is a key component of relationships that are nonetheless the sum of multiple interactions between physical individuals. To study socio-technical systems, as for many other research fields, we need data and data sources. In some cases, data sources are straightforward to individuate, and we cannot change their nature, but only analyze them meticulously to make the interpretations as transparent and adherent

to reality as possible. In other cases, data sources are more distant from the research question and therefore need to be adapted and systems built that can best exploit them. Nevertheless, limits in the choice of research direction remain linked to the data availability. In general, *open science* and *open data* in particular remain a central key point that every research community should champion. On the other hand, data are often noisy, heterogeneous in the gathering, and partial with regard to a specific point of view. We need techniques that are "agnostic" about their biases, or we need to incorporate some of them in the model itself [2]. In this thesis, we deal with three different socio-technical systems, whose data sources are very different as well, and an important part of the effort was made to data pre-processing with an awareness that is a key step to avoid random results or misinterpretations of the phenomena. The outline of the thesis will follow the analysis of these three socio-technical systems.

In Chapter 2, I will first introduce the research questions and the different application fields in which we will exploit them.

In Chapter 3, I will focus on providing a theoretical framework for complex networks model, and how to define them according to wanted features that best adhere to the system they propose to codify. I also will frame some well-established methods to tackle some classical empirical aspects: community detection and measures of centrality. Then I will introduce the second main framework used in the analysis, the wide and varied research field of Natural Language Process.

Chapter 4 is then devoted to the empirical exploration of the first socio-technical system, the ecosystem of migrations of scientists around the world from 2000 to 2021. In this case, the data source was not designed to answer our research question, so we will show how to fit the data to extract the desired information. The interplay between the technical and the human parts in this case is very dense, and to unveil even a small portion of it we will try to identify the countries that stand out as principal drivers of the stream, to understand what topological features allow a country to stand out as the best attractor or the best provider of researchers.

Chapter 5 will attempt to disentangle the complex evolution of the online debate around the Covid-19 pandemic, on the Twitter platform, regarding a specific group of users and a given aspect of the debate.

Chapter 6 shows how to model very ancient traces of a social system that no longer exists, through the cooperation of many different experts, from archaeologists to historians to data analysts, to achieve a framework that permits to analyze of the social structure of Late Bronze Age Western Asia kingdoms, particularly Hittites and Kassites.

Finally, in Chapter 7 we will attempt to conclude each previously open discussion with final remarks and some hypotheses for future possible directions.

Chapter 2

Research Problem

Although the socio-technical systems introduced in the previous Chapter are very different from each other, they have some common traits that will allow us to delineate a narrower field of research than one might imagine. In all cases, it is possible to show how systems are characterized by a static part, namely the technical infrastructure, and a dynamic part, generated by human interactions. The technical infrastructure is what governs how the system evolves; it may be the rules of engagement of an online platform, the rules for organizing data in outlining a researcher's career, or the rules for organizing an archive to keep track of transactions in effect in a Babylonian kingdom nearly 4,000 years ago. These rules do not change, or rather make maximum sense in their constancy, they preserve the systems in a sense. However, these structures are themselves artificial traces, or data, of more complex real systems that have constraints and manifestations that cannot always be summarized or described by listing their component parts. In scientific migration, the flows are governed by multiple drives, which are academic, and economic, but also cultural and social in nature; there are specific career modules, such as Ph.D. or short postdoctoral contracts, and competition mechanisms that break the balance. In organizing an archive, administrations, food rationalizations, and seasonal availability of supplies come into play, even before the human component. It is the latter, however, that generates dynamism. Relationships between individuals change the system and shape its form, based on their importance individuals accumulate mass and distort the balance. We do not always have the same level of aggregation between relationships, for scientific migration the interacting units are countries and not individuals, but despite this, the evolutionary spark remains an interaction, the presence, for a certain period of time of a relationship between two units, the nature of which has been established *a priori*, people, countries, users.

This thesis will look at how to create models to study these systems, and how to transcribe into measurable paradigms the only traces we have of the general behavior of these networks of interactions. Specifically, having

created the models, we will investigate how to interpret their plots and the most common pattern extrapolated from them, how the strength or weakness of certain interactions shape the whole overall figure, or how the particular activity of some protagonists can cascade the behavior of all other units.

The following sections outline the existing literature relative to the three different research contexts.

2.1 Scientific Migration

Human migration has been modeled in terms of complex networks in [3]. Similarly to our case, they define the international migration network as a temporal weighted directed network having countries as nodes and volumes of migrants as edges. Differently from our work, the study by Fagiolo et al. mostly focuses on the identification of community structures and dis-assortativity; moreover, it considers the general human migration that has fundamentally different characteristics than the scientific one. Following up this seminal work, many other approaches are proposed with similar purposes, studying for example human migration from a multi-layer perspective using data gathered from social media platforms [4].

The mobility of scientists is a topic of broad interest that has been investigated in a series of works both from a data-driven perspective and a model-driven approach [5]. By means of a survey, Franzoni and co-authors [6] tackle the problem with the intent of providing consistent data about cross-country research. This study highlights that Switzerland has the largest percentage of immigrant scientists working in the country (56.7), while India has the lowest and that the most likely reason to come to a country for postdoctoral study or work is professional. On the other hand, a lot of analyses utilize bibliometric data. The study documented in [7] explores how Scopus[1] can be exploited as a data source to understand international scientific mobility for countries with high adoption of the platform. In the study, the authors show quantitative metrics and general trends about the observed countries and researchers. Comparing these indicators with OECD statistics they conclude that a bibliometric study of scientific migration using Scopus is feasible and provides significant outcomes. Another recent study, by Verginer et al. [8], describes a method to extract mobility networks from a collection of four bibliographic data sources, not including ORCID , to characterize the mobility of scientists at city granularity, finding evidence that global cities attract highly productive scientists early in their careers. In the model-driven approach, the main frameworks that have been used to study human migration, in general, are the gravity model [9], whose main formula is similar to the gravity equation, and in which the number of migrations is assumed to be related to the population at the origin or destination and to

[1] https://www.scopus.com

decrease with distance. The network model in which the nodes model the countries, the links the presence of a path among two countries, and the link weights usually stand for the amount of migration flow, that is the number of researchers that migrate from one nation to another. Robinson et. al. [10] propose a machine-learning approach to predict long-term human mobility. An extensive survey about deep learning and human mobility can be found at [11].

Other works employ different data sources, such as [12], that employ the network structure to unfold information about human mobility from GPS [13] and GSM data, or [14] in which mobile phone data has enabled the timely and fine-grained study human mobility.

Additionally, there are many works in which various exogenous features have been employed to predict the origins and destinations of human migration flows. As in the work by Cerqueti et al [15], where they have explored how to build complex networks from worldwide migration flows to identify a socioeconomic indicator that explains the reasons behind the phenomenon. Or in [16], where the authors mixed an economic point of view the authors with the traditional sociology of science.

Many works tackle research questions that are part of the migration phenomenon or tried to capture more specified aspects. Human mobility has often been related to socio-economic development, as in [17]. Another complementary work [18] correlates per-capita income and labor productivity with human migration and network centrality. Saxenian [19] and Agrawal et al. [20] discuss the concept of *brain-drain* and argue that connections between migrant scientists and their home countries are persistent in time and might ease knowledge transfer backward. For these reasons, they call this phenomenon brain "circulation" or "brain bank". In [21], the authors frame the brain-drain problem in the Russian context. Linked research topics deal with the analysis of scientific careers: the dynamics of faculty hiring [22], the gender imbalance in contributions to science projects [23], or the understanding of the onset of hot streaks across artistic, cultural, and scientific careers [24]. Moreover, the relationship between knowledge and human migration, used as as proxy for its diffusion mechanism is well studied across many multidisciplinary projects, even without limiting the data sources to researchers, as in [25] and [26] which focuses in particular on the consequences of migrants' returns to their homeland.

Finally, an area of study takes on the shaping of different types of maps of science, either geographical, as in[27] where the authors try to understand the growth of regional knowledge networks within international research collaboration, either semantic, where the scope is to quantify the rise and fall of scientific fields [28], [29].

In general, reliable data sources about the topic are often problematic to find, and as we can detect even from these two works, results may be not stable across different data sources, in particular, due to biases. On

the other hand, also in model-driven approaches, it is necessary a validation with real data to measure the performance of the models. To tackle the wide phenomenon of researchers' migration a comprehensive framework would be optimal, to pinpoint a source of data, analyze it in an empirical way to detect patterns and trends, to finally model the dynamics suggested by the data.

2.2 Ancient Civilizations

In the past decade, the study of ancient civilizations has increasingly focused on data science, relying on data analysis techniques to exploit written sources from the past to examine the social and cultural aspects of these civilizations. In particular, a line of research has addressed the use of network analysis for studying the social and political structures of the past by leveraging the mention of entities such as locations and personages in texts, with proof of concepts ranging in time and space [30, 31, 32, 33, 34].

In parallel with this trend, the advent of Linked Data has made semantic resources available for archaeological and historical research, with notable examples such as CRM Archaeo [35] and FPO [36]. Today, the use of semantic representation techniques and network analysis methods can provide an integrated approach that combines the shared, unambiguous definition of entities and relationships in the historical.

2.2.1 Social network Analysis and Ancient Civilizations

Network models conceptualize interactions between entities, leveraging the concept of factoid. We define a network as $G = (V, E)$, where V is a set of nodes, and E is the set of links that encode the relationship structure between nodes (see Chapter 3 for more details on network-based models). This definition can be enriched to improve the adherence of the model to reality. In particular, a network can be defined as directed, weighted, temporal [37], dynamic [38], bipartite, if its nodes can be divided into two not overlapping sets [39]. Since links often exhibit heterogeneous features, new structures were theorized, with layers [40] or multiedges in addition to nodes and links [38]. This flexibility allows applying the framework to multiple tasks. In the state of the art, the use of network analysis techniques in the study of cuneiform archives relies on two main approaches, which consist, respectively, of using the social network to infer new information about the nodes, and inspecting the properties of the network on a global scale to discover and confirm the working hypothesis about social structures.

A key work in laying the groundwork for social network analysis in the context of historical data, with emphasis on exploiting all possible information resources encoded by the relationships between social and material elements, was done by Brughmans in [41] and [42].

Content exploration. Entity co-occurrence in text analysis has been extensively explored when dealing with ancient texts. A fundamental research question regards annotation comparison, and the aim to merge nodes representing the same individual [43], [44]. Another line of research channels efforts into building tools for content exploration for example, Bornhofen et al. [45] employs a corpus of digitized resources about European integration since 1945 to generate a visualization tool that allows interest-driven navigation, exploiting the framework of a dynamic multilayer network that represents different kinds of named entities appearing and co-appearing in the collections.

Society structure. Studying social roles could also deeply impact the understanding of how society and in particular ancient ones worked, and analyzing the interplay of political and economic relationships can shed light on hierarchy and power dynamics. Breigher at al. [46] famously set the basis for many other works coupling the marriage and the economic trading between fifteenth-century Florence families.

Spatial network. Network model has been used to analyze hierarchical predominance in cultural practices across different regions and the evolution of cultural trends. Mizoguchi et al. [47] establish links between ten regional entities whenever the author found archaeologically recognizable similarities in pottery styles and mortuary traditions. The work by Schich et al. [48] aims to understand which processes shape and drive the geopolitical aspect of cultural history by a birth-to-death network, where nodes are countries and links represent the migration of notable individuals over time from birth to death places. Spatial and social entities could interact and discover interesting motifs in the network could lead to new insights into the life of ancient empires [49].

2.3 Online debate

The Covid-19 pandemic has attracted a great deal of attention from scientists, spanning a wide area of research interest, from medicine to economy, but also over the opinion landscape dynamics [50]. Opinions around all the possible topics related to the pandemic were tracked on social platforms exploiting different points of view and many perspectives: from medical to political [51], from social issues to habit changing, until fake news diffusion and conspiracy theories [52], despite of all the challenges of predicting microscopic dynamics of online conversations [53]. In general evidence of time-varying dynamic were found in the discussion correlated to the evolution of the pandemic [54]. The following is a partial review of the available literature around the pandemic of Covid-19 that does not purport to be definitive.

Vaccines. Hesitancy toward vaccines, the rejection of the traditional

types of medicine, and the use of alternative practices is a complex phenomenon that has occupied a portion of the public debate since well before the pandemic [55]. However, with the media attention of recent years, it has had a spike in engagement that has overflowed from previous boundaries. Johnson et al. [56] have to provide a system-level analysis of the multi-sided ecology of nearly 100 million individuals expressing views regarding vaccination. They built a cluster network to study the entanglement (recommendation or mention) between, anti-, pro-, and undecided clusters. The work presented in [57] continues this line of research by proposing a variation of the SIS model, where the undecided position is considered an indifferent position, including users not interested in the discussion. Anti-vaccination individuals form fewer but more than twice cluster the pro- and offer a more diversified range of opinions, from safety to conspiracy theories, and alternative medicine opinions. They seem to increase while outbreaks, like measles in 2019, are either located within the cities or remain global. Undecided clusters are very active. Important results highlight also how socioeconomically disadvantaged groups were more likely to hold polarized opinions on the coronavirus vaccine [58]. Other works try to grasp the more difficult dimension of the emotional sphere, Semeraro et al. [59] expose crucial aspects of the emotional narratives around COVID-19 vaccines adopted by the press, highlighting how vaccines have been consistently portrayed with significantly more trust and anticipation in mainstream news, with no significantly emotional language displayed in alternative news. This general feeling towards the vaccines in alternative news was not the same reserved for the AstraZeneca vaccine which, overall, carries significantly more sadness.

Bot. Ferrara [60] employs 43.3M (1% of the overall conversation) English tweets about Covid-19 (January to April) to tackle the bot problem concerning social media debate. The study suggests that there is evidence of bots fueling the debate, being this topic a factor shared by accounts having a high bot Botometer score [61].

Dataset availability. Chen et al. [62] collect 72 million tweets in March 2020, constituting roughly 600 GB of raw data. They track any tweet containing the keyword(s) in the text of the tweet, as well as in its metadata. Hashtags with the sub-string "coronavirus" consistently remain a more heavily used hashtag in the data set, spiking on the day the WHO declared COVID-19 a global public health emergency and the day the United States announced the first COVID-19-related death. "covid" started being used on February 11, 2020, when the WHO announced "COVID-19" as the official name for the novel coronavirus disease. The keyword "Wuhan" steadily declined.

Conspiracy theories. Gruzd et al. [63] study the propagation of the hashtag #FilmYourHospital, used to promote a conspiracy theory about Covid-19 being an invention. They collected 99,039 posts contributed by 43,461 unique users. They represent the Twitter data as networks with

nodes as users and directed posts (either replies, retweets, or mentions) as edges to study they evolve over time: influential conservative politicians and activists were behind the initial stage, then the majority of users who posted a tweet using this hashtag self-described themselves as Trump supporters. They also found more human activity than bots. Ahmed et al. [64] try to understand what kind of drivers has shaped the 5G Covid-19 conspiracy theory and which strategies could have been used to deal with such misinformation. They perform a network analysis from Twitter data collected through a 7-day period (from 27-03-2020 to 04-04-2020) in which the #5GCoronavirus hashtag was trending in the United Kingdom. They identified two large network structures: one made of isolated groups and the other of a broadcast group. There was a lack of an authority figure who was actively combating such misinformation. Of 233 sample tweets, 65.2% (n=152) of tweets were derived from non-conspiracy theory supporters, which suggests that, although the topic attracted high volume, only a handful of users genuinely believed the conspiracy. Fake news websites were the most popular web source shared by users, but also YouTube videos were shared.

Political debate. Jiang [65] studies how political characteristics of the locations affect the evolution of online discussions about COVID-19 in each US State. Users from liberal-leaning states frequently tweet content critical of political elites, whereas users from conservative-leaning states persistently utilize hashtags in support of the President. Comparing users who tweet right-leaning hashtags with users who tweet left-leaning/neutral hashtags, the former group is less likely to interact with users who tweet health and prevention hashtags. this propensity has led to two segregated communities, largely divided by their political ideology. Sha et al. [66] analyze the Twitter narratives around the political decision-making in the United States, by applying a dynamic topic model to Covid-19-related tweets by U.S. Governors and Presidential cabinet members. They employ a binomial topic model to track evolving subtopics around risk, testing, and treatment. The clusters show roughly four phases in time: 1) outbreak in China, monitoring the situation, reporting confirmed cases, stating that the risk was low; 2)-3)the government started to take action to protect the American citizens, topics as distancing policies also emerged; 4) tested, confirmed positive and death cases in different states, but also a disaster, quarantine, stay home were encouraged and reiterated on Twitter, and the sacrifices of health workers were acknowledged(thank). Green et al. [67] examine polarization in tweets of current members of the U.S. House and Senate during the onset of the COVID-19 pandemic, measuring it as the ability to correctly classify the partisanship of tweets' authors based solely on the text and the dates they were sent. They exploit a random forest trained on a randomly sampled majority of the tweets, using the text features. Democrats discussed the crisis more frequently, emphasizing threats to public health and American workers, while Republicans placed greater emphasis on China and businesses. Polar-

ization peaked in mid-February after the first confirmed case in the United States and continued into March.

Public debate. Gligorić et al. [68] study Twitter accounts that posted at least one COVID-19-related tweet that received at least 10 retweets likes during the week of 6–12 May 2020 (14,200 accounts). They create a sample of these accounts, categorize them into 13 categories and collect their entire Twitter timelines from 1 January to 31 May 2020. While accounts in all categories on average increased their tweet volume, accounts related to Science, Healthcare, and Government & Politics saw the largest boosts in engagement. According to the authors, these findings imply that users selectively promote information from structurally relevant sources during the crisis. Gallagher et al. [69] show how Covid-19 elites vary considerably across demographic groups, in terms of racial attributes, and geographic and political similarities. With this variation in mind, they discuss the potential for using the disproportionate online voice of crowd-sourced Covid-19 elites to equitably promote timely public health information and mitigate rampant misinformation.

Infodemic. Cinelli et al. [70] perform a comparative analysis of information spreading dynamics around the same argument in different platforms: Twitter, Instagram, YouTube, Reddit, and Gab. They analyze interactions between topic and users engagement, around the 20th of January, the day WHO declared the official start of the pandemic. The authors have modeled the growth of the number of people publishing a post on a subject as an infective process, extracting for each platform the R_0. They reported a high correlation between the cumulative number of posts and reactions related to questionable sources versus the cumulative number of posts and interactions referring to reliable sources. With a similar scope, the Complex Multilayer Networks (CoMuNe) Lab with the Harvard's Berkman Center for Internet & Society and IULM University of Milan develop an online platform called "Covid19 Infodemic Observatory" [71, 72]. They collect about 100M public Twitter messages to understand the digital response in online social media to the Covid-19 outbreak. They used machine learning techniques to quantify the collective sentiment and psychology, the fraction of activities of social bots, and the news reliability measured as the fraction of URLs pointing to reliable news and scientific sources.

Topic analysis. Muller et al. [73] released "COVID-Twitter-BERT", a transformer-based model, trained on a large corpus of Twitter messages on the topic of Covid-19. Ordun et al. [74] use topic modeling techniques to generate twenty different topics regarding case spread, healthcare workers, and personal protective equipment. Then, they investigated the user activity around them: the median time-to-retweet for their corpus was 2.87 hours, using directed graphs they plotted the networks of Covid-19 retweeting communities from rapid to longer retweeting times, describing how the density of each network increased over time as the number of nodes generally

decreased. Park et al. [75] generated four networks in terms of key issues regarding COVID-19 in Korea, to investigate how Covid-19-related issues have circulated on Twitter through network analysis. They classified top news channels shared via tweets and conducted a content analysis of news frames used in the top-shared sources. Wicke et al. [76], present an analysis of the discourse around the hashtag #Covid-19, based on a corpus of 200k tweets posted on Twitter during March and April of 2020. Using topic modeling they showed that war framing is used to talk about specific topics, such as the virus treatment, but not others, such as the effects of social distancing on the population.

Language. Chen et al. [77] use sentiment feature analysis (LIWC) and topic modeling (LDA) to reveal substantial differences between the use of controversial terms such as "Chinese virus" against "Covid-19": tweets using controversial terms contain a higher percentage of anger as well as negative emotions. They also point to China more frequently. The usage of the "Chinese virus" on social media leans strongly towards racism. Schild et al. [78] have collected two large-scale datasets from Twitter and 4chan over a time period of approximately five months to investigate whether there is a rise or important differences with regard to the dissemination of Sinophobic content.

Editorials. Finally, a lot of editorials have tried to imagine future research topics that should be addressed regarding Covid-19 [79].

Chapter 3

Background on Complex Networks and Natural Language Process

In the following Sections, we will detail the theoretical background of the two main conceptual and methodological frameworks used in the analyses.

3.1 Complex Networks

We need a model to investigate how simple units connect in a complex system to form an entity that is more than their sum. In the last decades, the interdisciplinary interest turned toward the understanding and the forecasting of real-world complex phenomena has created a new science field, named "network science" [80, 81].

This new discipline, made by the combination of already well-established mathematical models, such as graphs, with a data-driven approach, has been demonstrated to be the most successful representation that allows researchers to uncover non-trivial structural patterns, which, has been shown, emerge from the natural self-organizing dynamics of most natural systems. These patterns can be detected in a wide range of domains: from socio-economy to biology, brain or technology, empowering everything, from Google to Facebook and Twitter. But still, among systems belonging to the different domains we can detect a common architecture and network science aims to encode the dynamic that leads to the creation of this architecture [80].

We can define the simplest type of network as the graph $G = (V, E, \varpi)$, where V is a set of nodes (or vertexes) of N entities, and $\varpi : V \times V \rightarrow \mathbb{Y}$ is a function, defined for each pair of nodes $i, j \in V$, that maps the links (or edges) $e_{ij} = (i, j)$ encoding pairwise interactions. According to ϖ and $\mathbb{Y}$ we have different types of networks, which best model different types of systems, in fact, links can be directed, and have different strengths

(i.e. "weights"), exist only between nodes that belong to different sets (e.g. bipartite networks), or be active only at certain times.

A lot of effort has been made to establish a strong theoretical framework, mostly inherited from graph theory. The discontinuities, or structural holes, in the link pattern are responsible for the network playing a powerful role in many fields, especially just to be a bridge between the local and the global view of a system. One of the researchers' efforts, during these years, has been aimed to explain how crossing this bridge, that is how simple processes at the level of individual nodes and links can have complex effects that ripple through a system as a whole. Efforts include understanding the drivers beyond the link formation, and the resulting connectedness structure. The ability to detect such groups could be of significant practical importance. For instance, groups within the worldwide web might correspond to sets of web pages on related topics [82].

A lot of studies have tackled the problem to detect special agglomerations of nodes, which were not present in the random counterpart model, exploiting, for example, the concept of "modularity" [83], and explaining the driving forces that generate them, like "homophily", "triadic closure" [84] and "aggregation" (like belonging to the same school, the same city neighborhood, or the same sports club). Another popular line of research investigated ways of defining the centrality of a node in the system, looking for points of view that were local and global.

A field in which these principles find a natural application is Social Science since humans share behaviors, opinions, interests, and skills. Over history, humanity has fought over a certain idea of the world, or have tried to find salvation in it, using it to model countries, cities, religions, and arts, to move border and knowledge, to forge alliances and friendships, partition reality into manageable subparts. What has changed drastically in recent years was the proliferation of technology, for the first time researchers, could do something impossible before: they could "measure" these dynamics. For example, new studies have been published about information spreading [85], a new way of dating [86], how happy we are [87], where we tend to migrate [3], the effect that echoes chambers have on public opinion [88].

3.1.1 Multilayer network

Often real systems, interact in ways that could not be approximated only through a single type of relationship. To consider links exhibiting heterogeneous features, new structures were theorized, called in its most general version "multilayer network", with layers in addition to nodes and links. In the more general multilayer framework, a node i in layer α can be connected to any node j in any layer β. Layers will represent aspects or features that characterize the nodes or the links that belong to that layer. A multilayer network can have any number of aspects d. Each of them can be represented

by a sequence of elementary layers, $\mathbf{L} = \{L_a\}_{a=1}^{d}$, or we can construct a set of layers in a by assembling a set of all of the combinations of elementary layers using a Cartesian product $L_1 \times \cdots \times L_d$. Nodes can be absent in some of the layers, so, for each choice of a node and layer, we need to indicate whether the node is present in that layer. To do so, we construct a subset $V_M \subseteq V \times L_1 \times \cdots \times L_d$ that contains only the node-layer combinations in which a node is present in the corresponding layer. We now can define a *multilayer network* as a quadruplet $M = (V_M, E_M, V, \mathbf{L})$ [40], [89].

As a consequence, the set of links can be partitioned into "intra-layer edges" $E_A = \{((u, \boldsymbol{\alpha}), (v, \boldsymbol{\beta})) \in E_M \mid \boldsymbol{\alpha} = \boldsymbol{\beta}\}$, that is links that connect nodes set in the same layer, and "inter-layer edges" $E_C = E_M \setminus E_A$ which are those that connect nodes set in different layers. More specific definitions include node-colored networks, interconnected networks, interdependent networks, networks of networks, multiplex networks, multirelational networks, k-partite graphs, or even hypergraphs.

Open questions still in need of investigation about multilayer structure regard the problem of reducibility, i.e., defining the number of layers needed to accurately represent the system or the topic of robustness, that is the network's ability to preserve the structure when it is subject to failures or attack, but also there has been considerable interest in generalizing concepts from monoplex networks to multilayer networks, from community detection [90], to link prediction [91], the definition of new centrality measures [92], or how to extend the concept of distance [93].

The development of representations and models for multilayer networks contributes to a better understanding of the structure and function of multilayer systems and enables the discovery of new phenomena that cannot be explained by a single-layer network. However, in order to understand how real-world multilayer networks behave and are organized, it is also crucial to collect and study empirical data for which such frameworks are appropriate. It is also helpful to develop new visualization tools, data structures, and computational methods. One of the main contributions of this thesis follows this direction, we will try different data models and exploit multilayer networks in an empirical way, trying to develop frameworks that better adhere to real systems.

3.1.2 Temporal Networks

One can represent a temporal network as a set of events or an ordered sequence of graphs [94], [95], each of which arises from the element of the event set $e = (u, v, t)$, where $u, v \in V$ are nodes and $t \in T$ is a timestamp of an event. When using the general multilayer-network framework, two identical nodes from different layers are adjacent via an inter-layer edge only if the layers are next to each other in the sequence. Furthermore, the time progression can be incorporated into the network structure by using directed

edges between corresponding nodes in different layers. One can also allow a generalized ordinal coupling that includes a time horizon h by considering not only neighboring layers but all layers that are within h steps. A very broad field of research dealing with temporal networks aims to reconstruct the dynamic of contact processes [96].

3.1.3 Axioms for Centrality

A classic approach to assess the importance of a node in a network is to measure the value of "centrality" that node has with respect to all the other nodes and consequently with the global link structure [97]. There are of course many possible definitions of importance and so many centrality measures. Note that, in this work, we will often employ (and so we will define in this Section) the weighted and temporal version of many centrality measures. The non-weighted and non-temporal scores are defined in exactly the same way with some changes in notations, for example replacing W_t with the unweighted adjacency matrix A_t in defining equation, or by omitting timestamp in vectors' subscripts.

The simplest measure of centrality is the *degree centrality*, defined as the number of edges connected to a node. With respect to the type of network, it could be also directed or could take into account the weights of the links.

Another class of measures related to the distances, like *closeness centrality*, is defined in the simplest form as the mean geodesic distance from i to j, averaged over all vertices j in the network:

$$l_i = \frac{1}{n} \sum_j d_{ij} \tag{3.1}$$

In all our settings these measures will not hold since we will deal with relatively small networks.

A natural extension of the simple degree centrality is eigenvector centrality. We can think of degree centrality as assigning a "centroid" to each working neighbor of the network. However, not all neighbors are the same. In many cases, the importance of a vertex in the network is increased by connections to other vertices that are important in their own right. Instead of giving nodes just one point for each neighbor, eigenvector centrality gives each node a score proportional to the sum of the scores of its neighbors.

Further development of this measure has produced *Katz Centrality* but also *PageRank*, the trade centrality used by the Google web search corporation, at the beginning of the development of their web ranking technology [98], [99].

For the most part, the following analysis will Let R_t be the PageRank matrix of $G = (V, T, \varpi)$ at time $t \in T$, defined as

$$r_{ij,t} = d \frac{w_{ij,t}}{\sum_{j \in V} w_{ij,t}} + (1 - d) \frac{1}{|V|}, \tag{3.2}$$

where $d = 0.85$ is the dumpling factor. Note that, in this work, we consider the edge weights in the definition of R_t. The PageRank vector $\vec{r}_t = (r_{1,t}, \ldots, r_{|V|,t})^\intercal$ is obtained by repeating the iteration

$$\vec{r}_t(x+1) = R_t^\intercal \vec{r}_t(x) \tag{3.3}$$

until convergence, with initial conditions $r_{i,t}(0) = \frac{1}{|V|}$. $\vec{r}_t$ is computed for each timestamp, i.e., year, $t \in T$. In the following, we often refer to the PageRank vector as $\vec{r}$ neglecting the subscript.

In some networks it is appropriate also to accord a vertex high centrality if it points to others with high centrality, this is particularly useful for directed networks. One can imagine defining two different types of centrality for directed networks, the authority centrality and the hub centrality, which quantify vertices' prominence in the two roles, as a "receiver" or as a "provider" of information. We identify the *hyperlink-induced topic search* algorithm (also known as HITS or *hubs and authorities*) [100] as the main measure to study our network. The HITS hub vector $\vec{h}_t = (h_{1,t}, \ldots, h_{|V|,t})^\intercal$ and the HITS authority vector $\vec{a}_t = (a_{1,t}, \ldots, a_{|V|,t})^\intercal$ in $t \in T$ of $G = (V, T, \varpi)$ are defined by the limit of the following set of iterations:

$$\vec{h}_t(x+1) = c_t(x) W_t \vec{a}_t(x+1) \tag{3.4}$$

and

$$\vec{a}_t(x+1) = d_t(x) W_t^\intercal \vec{h}_t(x), \tag{3.5}$$

where $c_t(x)$ and $d_t(x)$ are normalization factors to make the sums of all elements become unity, i.e., $\sum_{i=1}^{|V|} h_{i,t}(x+1) = 1$ and $\sum_{i=1}^{|V|} a_{i,t}(x+1) = 1$. The initial HITS values of the scores are $h_{i,t}(0) = \frac{1}{|V|}$ and $a_{i,t}(0) = \frac{1}{|V|}$ for all $i \in V$.

A very different concept of centrality is *betweenness centrality* [101], which measures the extent to which a node lies on paths between other vertices. We define the betweenness centrality of a node $i \in V$ at time $t \in T$ as

$$c_b(i,t) = \sum_{\substack{s,e \in V \\ i \neq s \neq e}} \frac{\sigma_{se,t}(i)}{\sigma_{se,t}}, \tag{3.6}$$

where $\sigma_{se,t}$ is the total number of shortest paths from node s to node e at time t, and $\sigma_{se,t}(i)$ is the number of such paths passing through node i.

3.1.4 Mesoscale Structure

A lot of studies have tackled the problem to detect special agglomerations of nodes, which were not present in the random counterpart, defining the tasks of community detection and clustering. This procedure starts from a fundamental hypothesis, that the existence of the communities is rooted in who connects to whom, so they cannot be explained based on the degree

distribution alone, it is encoded in the complete wiring diagram. Following the definition of Barabasi and Albert [102], a *community* is a locally dense connected subgraph in a network, such that the nodes that belong to the same community have a higher probability to be linked than the nodes in different communities. Also, we expect that the nodes in the community are connected, i.e. all members of the community can be reached by the other members of the same community.

To be sure that a dense subgraph is really a community and not just a dense pattern that emerged by chance we have to compare the density of the same group of nodes after rewiring the network since randomly rewired networks lack an internal community structure. We will present the two main methods used throughout the dissertation.

The systematic deviations from a random configuration allow us to define a quantity called *modularity* [83], which measures, given a partition, its quality in terms of community structure, and has been shown to be a powerful instrument for communities detention.
Consider a network partition made of n_c communities, each community having N_c nodes connected to each other by L_c links, where c=1,...,n_c. If L_c is larger than the expected number of links between the N_c nodes given the network's degree sequence, then the nodes of the subgraph could indeed be part of a real community since its wiring structure is dense. Therefore, a simple procedure to discover valid communities in a network is measuring the difference between the network's real wiring diagram A_{ij} and the expected number of links between i and j if the network is randomly wired,

$$M_c = \frac{1}{2L} \sum_{(i,j)\in C_c} (A_{i,j} - p_{i,j}) \tag{3.7}$$

where C_c is the c-th community, and $p_{i,j} = \frac{k_i k_j}{2L}$ is the expected number of links between i and j in the rewired model. If M_c is positive, then in the subgraph C_c there are more links than we expect by chance, so it could represent a community. If M_c is zero then the connectivity between the N_c nodes is random, and fully explained by the degree distribution. Finally, if M_c is negative, then the nodes of C_c do not form a community, since its wiring structure is not dense enough. We can approximate

$$M_c \approx \frac{L_c}{L} - \left(\frac{k_c}{2L}\right)^2 \tag{3.8}$$

where k_c is the total degree of C_c, in order to obtain a global value for the modularity as

$$M = \sum_{c=1}^{n_c} \left[\frac{L_c}{L} - \left(\frac{k_c}{2L} \right)^2 \right] \tag{3.9}$$

If M is less than 0 then there is no community structure in the network, if it is 0 then there is just one community, if it is more the 0, then the higher M is for a partition, the better the corresponding community structure. A partition with lower modularity clearly deviates from these communities. M is the key instrument for a lot of community detention greedy algorithms since for a given network the partition with maximum modularity corresponds to the optimal community.

Another canonical model that has been used widely for community detection is the stochastic block model (SBM) [103]. It is a generative model for the data, that benefits from a ground truth for the communities, which allows considering in a formal context the question about the actual presence of a specific map into which the network can be partitioned that does not arise randomly. The core SBM is defined as follows. For positive integers n, k, a probability vector p of dimension k, and a matrix W of dimension $k \times k$ with entries in $[0, 1]$, the model $SBM(n, p, W)$ defines an *n-vertex* random graph with labeled vertices, where each vertex is assigned a block label in $b_1, ..., b_k$ independently under the community prior $p(\mathbf{b})$, and pairs of vertices with labels i and j connect independently with probability $W_{i,j}$. That means we wish to obtain a $p(W_i j|\mathbf{b})$ that satisfies the condition that nodes that belong to the same group are statistically indistinguishable, or that the ensemble of networks should be fully characterized by the number of edges that connects nodes of two groups r and s:

$$e_{r,s} = \sum_{[} ij] W_{i,j} \delta_{b_i,r} \delta_{b_j,s} \tag{3.10}$$

or twice that number if r = s. If we take these as conserved quantities, the ensemble that reflects our maximal indifference towards any other aspect is the one that maximizes the entropy. Given that, and leveraging the so-called Bayes' rule, we can obtain the probability $P(\mathbf{b}|\mathbf{W})$ that a node partition $\mathbf{b}$ was responsible for a network $\mathbf{W}$.

3.2 Natural Language Processing

Natural Language Processing (NLP) is an area of research that explores how computers can understand and manipulate human language in the form of text or spoken utterances [104]. NLP is a powerful field that has a wide range of applications, including automatic information retrieval, data mining and text mining, question-answering systems, and machine translation. These main frameworks can be applied to a plethora of tasks from sentiment

or stance analysis, to language generation itself. The technical modules that assemble an NLP pipeline and the tools necessary to build them are multiple and have evolved over time. Although NLP research has existed since the second half of the 20th century, the field became very popular thanks to the gradual proliferation of big data and the creation of statistical or corpus-based data sets that provided a large empirical resource for training and testing learning models. The models themselves have evolved continuously over the last decades. Many established models are based on traditional machine learning (ML) techniques: a fundamental achievement in this field has been the development of the PoS (part of speech) tagging database, which, when fed in large numbers into ML algorithms, is useful for training models that can evaluate text sequences and recognize the grammatical structure of sentences (i.e., distinguish between nouns, verbs, adverbs, etc.). Until six years ago, supervised learning techniques were dominant: most of the proposed models were based on support vector machines, Bayesian classifiers, decision trees, and other similar methods [105]. That changed with the rising of Deep Learning techniques, i.e., methods that exploit deep neural networks. In 2018 the researchers of the Google AI Language introduced Bert [106], "Bidirectional Encoder Representations from Transformers", a language representation model based on a transformer neural network, that aims to retain an extreme versatility paired with the power of learning contextual words embeddings. Its versatility is due to the fact that the model can be fine-tuned for many types of tasks, as long as a sufficient training set is available. The model is pre-trained in several languages, so the user does not have to perform a very computationally intensive task. The fine-tuning procedure does not require a lot of computational power and can therefore be used by most users. Since the creation of BERT, hundreds of variants have appeared, both customized versions of BERT for different languages, new models for specific NLP tasks, like Cicero (Meta) for the strategic game, or general new languages models, like GPT-2 [107] and GPT-3 [108]. However, traditional (non-deep) methods are still very popular because they are both easy and interpretative, which can be very important depending on the field of application. In particular, the main concerns raised regarding very large language models revolve around the unfathomable training data and how these models repeat and manifest the issues in the data [109]. In the text analysis carried out in Chapter 6, we will need a word representation, trained on unlabeled corpora to quantify, interpret and relate texts produced by different groups of people by trying to abstract their opinions. To do so we will exploit "fastText" [1], a module that allows training word embeddings from a training corpus with the additional ability to obtain word vectors for out-of-vocabulary words [110]. As a training corpus, we have used a collection of 12.8k million Italian original tweets (no retweets, quotes, or replies), with

[1] https://radimrehurek.com/gensim/models/fasttext.html

no keyword limitation, downloaded in March 2020 via Twita [2], a collection of Italian Twitter datasets [111].

[2] http://twita.di.unito.it/

Chapter 4

Scientific Migration

4.1 Overview

Human migration has been a very important phenomenon throughout history and has changed significantly over time as a result of historical and economic events. It is known for shaping each layer of a nation's society. Migration influences both the origin and the destination countries, in particular, it acts on the demographic aspect of nations, changing the composition of its populations, on its economical status, bringing or stealing resources in ways that seem at the same time desirable and undesirable, on the cultural mixing, increasing the different points of view on reality [112], [113]. The definitive outcome of human migration is subtle and extremely unpredictable, especially in the long term, due to the need for addressing different borders: geographical, political, and even cultural [114]. For all these reasons migration is perceived in many different ways, and subsequently, it is treated by the various countries of the world with opposite aims, sometimes it is encouraged, other times discouraged. For these reasons, human migration is perceived in many different manners and, consequently, treated by local states with opposite aims: it is sometimes encouraged, rather discouraged [115]. We can say for sure that it is a very difficult subject to analyze. One of the more complex parts concerning this type of study, migrations and overall phenomenons characterized by human behaviors, is the data collection. First of all, it is challenging to coordinate a global effort to obtain homogeneous data, collection often has not the same time range, and neither are the collection procedures are standardized. More recently, however, important steps have been made to improve the analysis of migrations. For example, in order to address the lack of coordinated migration data, the United Nations Population Division, with the help of the United Nations Statistics Division, the World Bank, and the University of Sussex, using the United Nations Global Migration Database, created a dataset on bilateral international migration covering most of the world's countries from

1960 to 2000 [1]. The proposed results regard a specific subdomain of the general phenomenon of human migration, the scientific researcher migration. In particular, knowledge, ideas, and information are considered to be among the most relevant assets in today's economy and are naturally embedded in researchers, scientists, and academics who, through their permanent or temporary mobility paths, move such goods from a location to another [7]. In the long term, international scientific mobility could impact fundamental social and economic aspects of the countries, such as scientific, technological, and productive assets [116]. Please, observe that hereinafter the terms "mobility" and "migration" will be used interchangeably to indicate the event of a researcher moving from one country to another, without differentiating permanent or long stays from short stays such as scholarships or post-doc periods. Albeit, most of the time, this phenomenon lacks the urgency of survival, it is highly competitive in terms of choice of the destination countries, as pointed out in [117]. We want to explore scientific migration as a global and inherently interdependent phenomenon. We analyze different frameworks to detect those countries that better attract or repel researchers, to characterize different roles, and to understand how mobility dynamics change over time: the so-called "brain drain" phenomenon. We rely our analysis on ORCID , a growing platform that collects public profiles of researchers. Given its nature, the (scientific) migration system can be modeled using a network that we define to be temporal, weighted, and directed: it turns out that a complex network perspective is very useful to define relationships between actors involved in this ecosystem, and it also provides a solid ground to define measures and parameters that can be used to study efficiently the mobility phenomenon. In this domain, nodes represent world countries and edges account for a migratory flow from one country to another. Edge weights stand for the size of the migratory flow in terms of migrants, while timestamps represent years.

Finally, considering that researcher migration is very important for the scientific world because it moves, beyond the people, and also a lot of ideas, knowledge, and information, we will try to tackle the migration of the different research fields encoded in the work of the researchers.

The structure of this Chapter follows the work *Measuring scientific brain drain with hubs and authorities: A dual perspective* (Online Social Networks and Media, 2021) [118]. However, the data reported there have been updated and the analysis extended.

4.2 Research Questions

Scientific migration is a very complex and broad phenomenon, it can be studied from many perspectives, with multiple data sources, and building

[1] http://databank.worldbank.org/data/reports.aspx?source=global-bilateral-migration

several frameworks of analysis. In the following Section, we will try to answer these questions.

- **How to model the brain drain phenomenon?** According to the "2021 UNESCO Science Report [119] brain drain is a chronic problem for many countries, whose principal causes are many and difficult to pinpoint, stagnating research expenditure to an aging researcher population, or intrinsic competition of the system. But in order to fully understand the main causes of this phenomenon, it is necessary to establish metrics by which to measure it. A metric that provides some stability compatible with a time horizon that is suitable for the system, especially given that research migrations are not immediate.

- **Beyond the drain phenomenon.** Looking at the brain drain phenomenon as a very localized event, may lead to missing the opportunity to assess which is the role of a global and heterogeneous structure of the migration network. So moving beyond the concept of drain strictly seen as the difference between the incoming and ongoing flow of researchers, could allow obtaining a more insightful picture of the brain drain, in which other aspects emerge, like how very attractive countries shape their neighbor, or which countries have the power to call back researchers.

- **Beyond the brain phenomenon.** Researchers can be more than a single unit since they move ideas, knowledge, and skills, they shape the very map of science. It could be interesting to try to disentangle physical people from their luggage of expertise, to capture how research fields and topics have moved and evolved over the years.

4.3 Data

ORCID is a nonprofit organization that collects contributions, affiliations, and personal information of the subscribed researchers. ORCID shares a Public Data File annually on the anniversary of its initial launch, in October 2012. The output folder for each file varies depending on the year in which the file was generated and the version of the XSD [2]. Specifically:

- **2013-2017**: Within the generated folder there are several folders, e.g. JSON and XML. Inside each folder is one file for each ORCID record (unique for a given user with all the information shown in the user's public ORCID page) in the specified format and XSD version.

- **2018+:**

[2] https://support.orcid.org/hc/en-us/articles/360006897394-How-do-I-get-the-public-data-file-

- *Records file:* in the folder shared, you will find the folder "summary", which contains several folders with individual ORCID records in XML format. These records are aggregated into subfolders based on the last three digits of the shared ORCID id. Each ORCID record contains the overall information of a ORCID user, name, bio, keywords, etc.

- *Activity file:* Inside the generated folder you will find multiple folders for each ORCID record. Each folder will include the full activities on each record in XML format, separated by activity subsection: education, employment, works, funding, and more.

The first attempt to use ORCID data in order to extract meaningful information about the migration of the scientific population has been carried out by Bohannon and Doran [120] through the gathering of 2.8 million ORCID public profiles from 1950 to 2016. In their work, Bohannon and Doran [121] highlight that ORCID was not designed with the specific aim of tracking researchers' mobility. Therefore, the data we consider has structural limitations as well as biases. First of all, as already observed, much of the information created by the members is retroactive since it refers to periods preceding ORCID 's launch in 2012. Therefore, some of the countries that nowadays have changed their political-geographical characteristics, are present in the dataset, making the set of considered countries highly variable year after year. Secondly, since its appearance, ORCID has always focused mainly on younger researchers. In fact, new subscriptions are often referred to researchers that pursued their Ph.D. recently, creating an overestimation of this category in the dataset, and reflecting the fact that younger researchers sign-up to ORCID more frequently than elder ones. Finally, countries are not equally represented, namely, the distribution of the number of researchers per country does not follow the distribution of the overall population. Bohannon and Doran compare ORCID data in 2013 about scientific migrations to the UNESCO Science Report[3] to discover which countries are misrepresented; e.g., China, Russia, and Japan result to be an underrepresented while, e.g, Spain, and Portugal are over-represented. For these reasons, we cannot regard the dataset as a definitive picture of the scientific migrations. Most of the data is concentrated in the 21st century, with peaks that progressively move over time. The decay of recorded migrations after each peak might be due to temporal bias given by the time when the dataset was gathered. Even if ORCID was founded in 2012, members are allowed to insert information about their previous occupations and their planned ones; as a consequence, we have data about migrations that happened before 2012 and occurred after 2021.

[3] https://en.unesco.org/node/252273

4.3.1 Data processing

ORCID database is a collection of files, one for each user that has decided to utilize the platform. As shown in Figure 4.1, we scan every user file and collect all the affiliation's changes at a yearly level, gathering both education and employment movements. A scientific migration happens if the country of one of these two affiliations changes. Nevertheless, extrapolating from the data only the transfer from one country to another would mean not exploiting the full potentiality of all the information. Researchers can change roles on top or instead of countries, or affiliations changes can be less linear. So we decide not to look just at the countries of two successive affiliations but to assign a status to each affiliation. From year to year, ORCID releases a

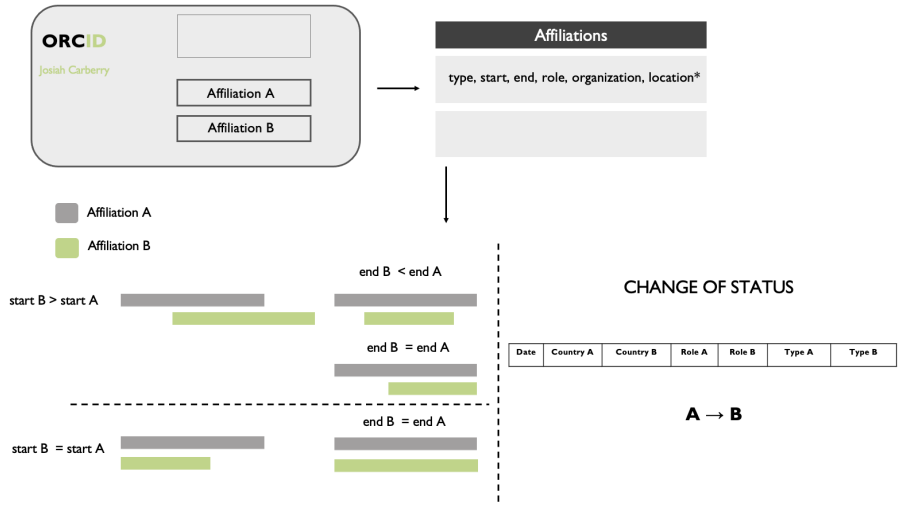


Figure 4.1: Pipeline of data preparation: from row data to network data: from the public data to change of status database. Josiah Carberry is a fictitious person, his account is used as a demonstration account by ORCID .

public screenshot of all the public data, whereby the information is repeated. When building the final database it is necessary to take this into account and aggregate the data appropriately to avoid worthless repetitions. Information is distributed over the affiliation history of each member. Through affiliations is possible to identify changes in what we previously called career status for each researcher in ORCID . Evaluate how the affiliations positioned in the timeline allow building a change of status space. Figure 4.1 shows the broader classification. All the possible cases can be divided into two main groups: the first collects all the possible changes that happen when the start of the next affiliation is greater in time with respect to the start of the current affiliation for a ORCID member, and the second the cases in which two affiliation start at the same time. Either group could be then further par-

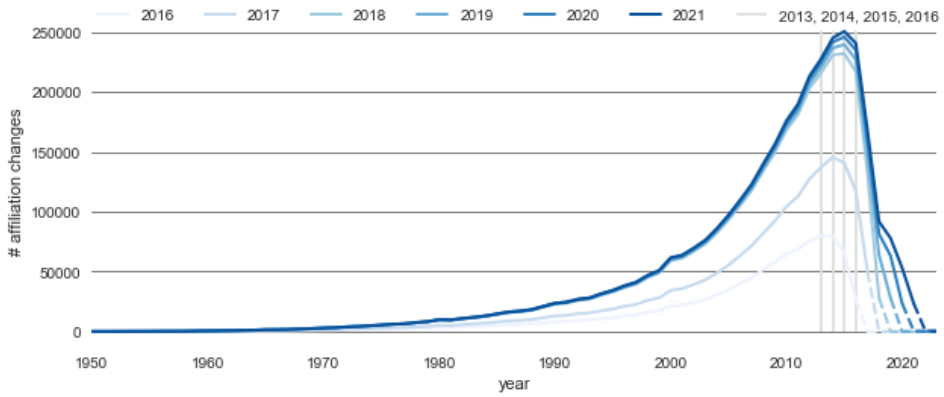


Figure 4.2: Distribution of the number of members affiliations changes per year from 1950 to 2023. Plot lines refer to the yearly ORCID public releases, they are named after the year of release. Gray vertical lines depict the increase in the maximum data availability. The dashed lines stand for projection data, that is changes in affiliation that users plan to do in later years than data collection.

titioned by looking at the end of the second affiliation. For each coupled affiliation we extrapolate the country, the role information, and the "Type". The label "Type" retains information about the nature of the migration. It combines two domains, "education" ("ed") and "employment" ("em"), giving rise to four possible combinations: from education to education, from education to employment, from employment to employment, and even if rare from employment to education. All ORCID fields are user-based, and there is no closed taxonomy, but we have evidence that 89.33% of the roles containing the *phd* string were labeled as "education", while the 9.68% labeled as "employment" needs further linguistic analysis since it encompasses both roles attributable to an actual Ph.D. and more complex acronyms such as "Professor, Ph.D., MD" or roles that are linked to the Doctorate as "Ph.D. Program Coordinator". "employment" is generally related to an academic position higher than a Ph.D. In the current work, we have not employed further meta-data available, like the Institution name or a more fine-grained location setting, like cities, but we plan to investigate the matter further in future works. Figure 4.2 shows the distribution of the number of affiliation changes, from 1950 to 2023, for each ORCID data release (named after the year they were issued). It depicts the increase in the number of affiliations per year, the pick slowly shifts toward more recent years, for 2016 release is in 2013 while for 2021 is in 2015. It increases with a time lag whose causes are probably the result of many factors and their interplay on many different levels, a systematic time-scale proper to the scientific career that can vary greatly but is rarely less than a year, the evolving flow of funding that can

shape again and again the map of scientific studies, global events, like Brexit or the Covid-19 pandemic, may already be partly present in the system but which we will have to wait to assess in their entirety.

All the possible career paths generate elementary movement, that can also combine in a more complex way. Some examples are shown in Figure 4.3. For example, in the first case, we can suppose that at t_1 the user changes their affiliation from A to B, maybe interrupting A beforehand. In the second case, U temporally becomes affiliated with B but then returns to A. It is more difficult to handle subsequent cases because they are more complex. In the third U from two possible affiliations change their status in C, but for example, if A and B are in a different country then we may have 2 different routes that take place at the same moment, from the country of A to the country of C and from the country of B to the country of C. Finally, U can start two memberships at the same time, or could have two co-affiliations. as well as for affiliations, each derived dataset increases in years

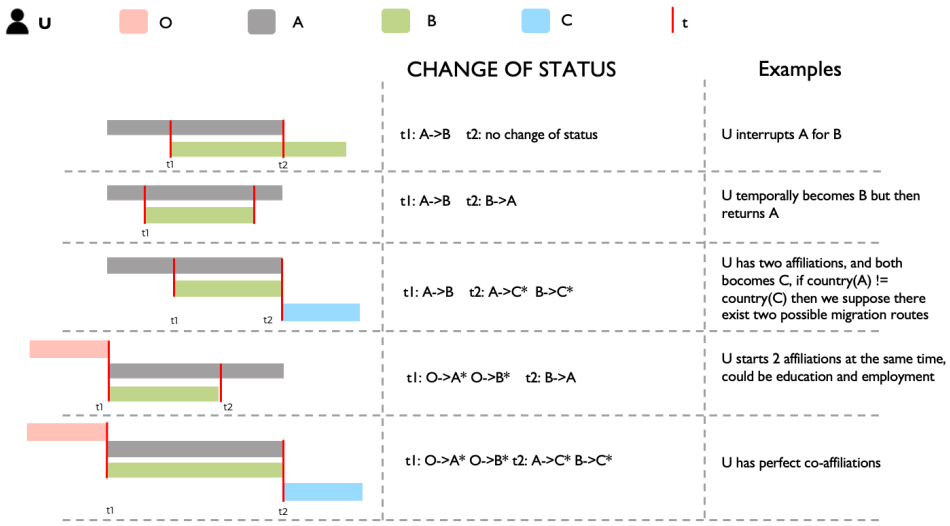


Figure 4.3: Possible cases of changes in career status of ORCID members. O-A-B-C are affiliations status of the user U, the vertical red bar identifies a change of status that happen at a certain time t. In the last column of the Table, we give real examples of what specific changes in the researcher's career could actually mean.

after being properly aggregated to eliminate repetition in the information, Figure 4.4 shows the increase in migration over the years. From now on we will consider data the most recent update of all ORCID public profiles, including appropriate arrangements to avoid redundancies in information from year to year.

It is possible to frame the reasons for the change in affiliation status into two main scenarios: either a researcher changes their role or the country of

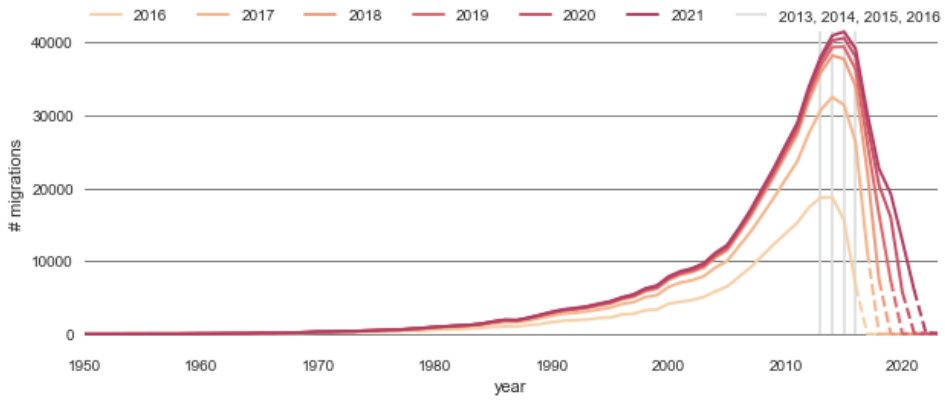


Figure 4.4: Distribution of the number of country migrations per year from 1950 to 2023. Plot lines refer to the yearly ORCID public releases, they are named after the year of release. Gray vertical lines depict the increase in the maximum data availability. The dashed lines stand for projection data, that is changes in affiliation that users plan to do in later years than data collection.

the institution sponsoring the affiliation. As shown in Figure 4.5 the majority falls in the first scenario, while almost every time it shifts the country of the institution sponsoring the affiliation then it evolves also the role.

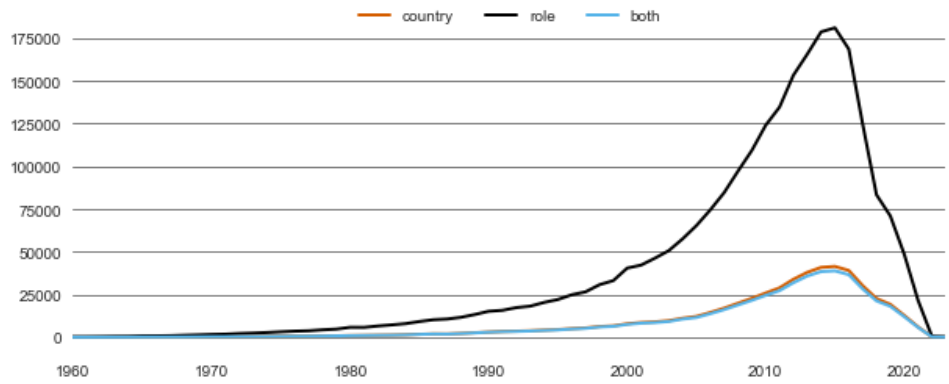


Figure 4.5: Distribution of the number of changes in affiliation status regarding two possible different scenarios: either a researcher changes their role or the country of the institution sponsoring the affiliation. The time range includes the years since 1950.

Another possible partition of the space of reason behind a change in status is generated through the type of affiliation, either of type education or of type employment. By observing Figure 4.6, it is interesting that the spike in "emem"-type changes is the fastest moving, it could be a consequence

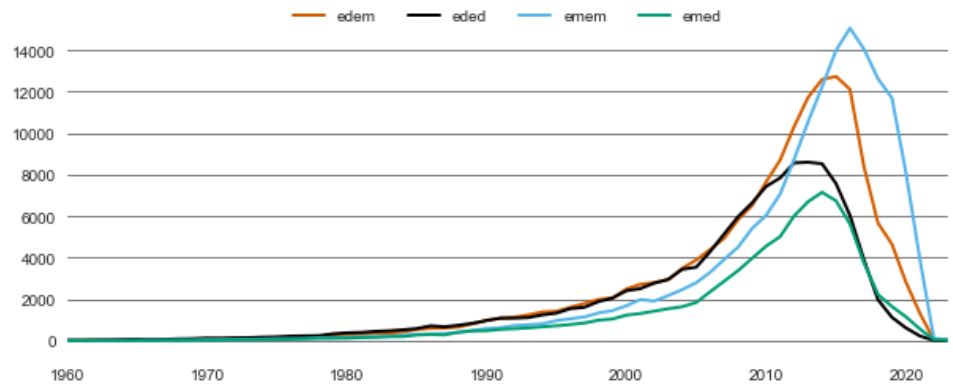


Figure 4.6: Distribution of the number of changes in country affiliation status regarding the type of affiliation. The label "Type" retains information about the nature of the migration. It combines two domains, "education" ("ed") and "employment" ("ed"), giving rise to four possible combinations: from education to education ("eded"), from education to employment ("edem"), from employment to employment ("emem"), and even if rare from employment to education ("emed").

of the fact that ORCID is skewed toward younger researchers who were among the first to enroll and who are maturing their careers as the years go by.

Finally, among the changes of status that regard countries we reach the conclusion that different routes happened, as previously discussed while presenting Figure 4.3. Different changes in status could impact the creation of different paths among countries, and with time different flow-dynamic of researchers' movement all around the world. Possible instances are many and we do not possess an *a priori* set of rules. We empirically try to shape the more reasonable system possible, describe in Figure 4.7. On the right we classified all the possible routes we are considering for this work and on the left their distribution over the years. Some of the possible routes are often coupled. Some examples are: 2 and 3 or 4 and 5 create the case of a temporary change in affiliation, and 4 and 8 a perfect co-affiliation. 6 seems a redundant case but stands for complete the bridging a co-affiliation with a new affiliation if the co-affiliation embraces institution in two different countries then potentially we have two routes that reach the country of the new affiliation, together with 6 the other path will be encoded by 1. In shaping this double pathway, we realize that the year was not enough to fully represent all the information. To build the model we will retain as timestamp the year range but for migration extrapolation, the full data is needed, Figure 4.8 exhibits one case study that led to this decision. In 4.8a to build the respective paths we consider only the year, we consider what happened for the researcher in Italy and South Africa a perfect co-affiliation, however, if we keep into consideration the full date we realize the reality is far better

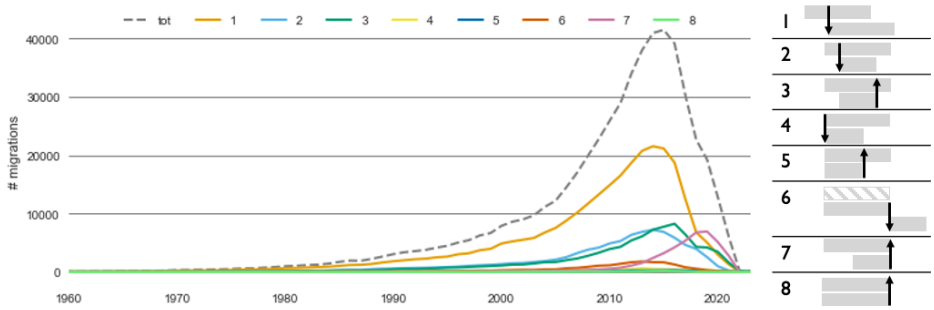
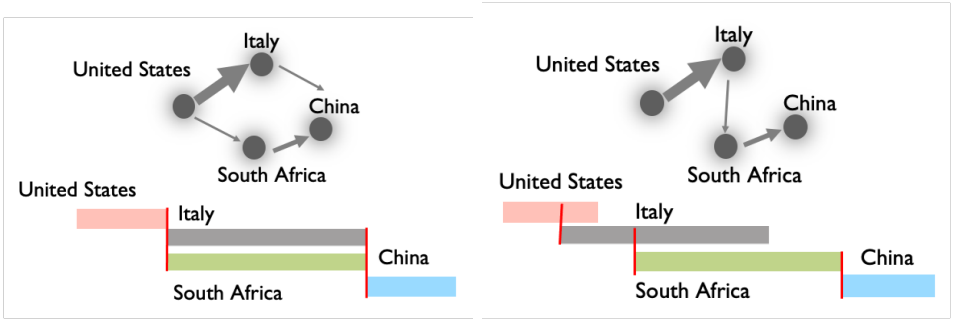


Figure 4.7: On the right we classified all the possible routes of migration. On the left is their distribution over the years.

approximate by case 4.8b, the researcher starts an affiliation in Italy and the moves to South Africa. These two scenarios will generate two different topologies in the networks, so it is worth refining the shaping of the model as thoroughly as possible.

We can also try to match routes and change of status to the keywords each researcher uses to describe their knowledge, changing the point of view of the migrations, which may represent instead of people moving streams of fields of expertise. There are some limitations regarding this line of research, being that keywords are not mandatory so not all the ORCID users decide to employ them, moreover, it is difficult to track for each researcher the evolution of the set of keywords they utilize, and every keyword is user-based which mean there is no common lexicon that uniformly divides different subject areas, so making comparisons can generate errors. Nevertheless, analyzing the evolution of topics of research could reveal new patterns in the evolution of discoveries. We extrapolate keywords per year for each ORCID profile, Figure 4.9 reveals the number of profiles with keywords over the years, the darkest part of each bar stands for the portion of the same profiles the current year shares with the previous one, the red line is the number of unique keywords.

All data described are empirically built, assessing their bias is both very difficult and very important because it bound all the possible interpretations, forcing them to adhere more closely to reality. Since comparing different data sources that have been gathered with different methods is difficult and lacks a universally established approach, we limit our consideration to the "2021 UNESCO Science Report [119]. Overall it attests that between 2014 and 2018, the researcher pool grew three times faster (13.7%) than the global population (4.6%), with China being a major player in this surge (without China, the surge in researcher numbers (11.5%) would have been only doubled the rate of population growth (5.2%). This translates into 8.854 million full-time equivalents researchers. In the 2018 ORCID release we have



(a) Path building considering just the year of the affiliation.

(b) Path building considering just the full date of the affiliation.

Figure 4.8: Different pathways formation with respect to the type of date we consider when extrapolating the change of status and in particular the migrations from one country to another.

evidence of 1.636.641 unique researcher profiles, while in 2021 there were 3.758.675 unique profiles. For more complete and rich data we compare the percentage of incoming researchers in 2018 by continent according to the UNESCO Report and the 2021 ORCID release in Table 4.1: the biggest gaps lie between Europe and Asia, according to the UNESCO Report Europe researcher are overestimated among all the ORCID profiles while Asian are underrepresented. According to the UNESCO Report, low-income economies have witnessed the fastest growth (+36%) in researcher density since 2014 but still account for only 0.2% of the world's researchers. Some of the greatest percentage changes are occurring in developing countries such as Jordan, Mauritius, Iran, and Ethiopia. Most of the limitations described by Bohannon and Doran remains also in this new data engineering pipeline, nevertheless, we can exploit it to detect regularities and patterns by the construction of a network model, useful in the understanding of the global perspective of the phenomenon, suggesting that experiments and estimations should be re-executed periodically to better monitor the phenomenon, tune previously introduced errors due to misrepresentation, and update information with fresh new data inserted/modified by researchers.

4.4 Models

Given its nature, the scientific migration can be modeled by means of a network. Nodes represent world countries and edges account for a migratory flow from one country to another. Edge weights stand for the size of the migratory flow in terms of researchers that move from one country to another, while timestamps represent years from 2000 to 2021. We use interchangeably the terminology *{network, node, link}* and *{graph, vertex, edge}*, knowing

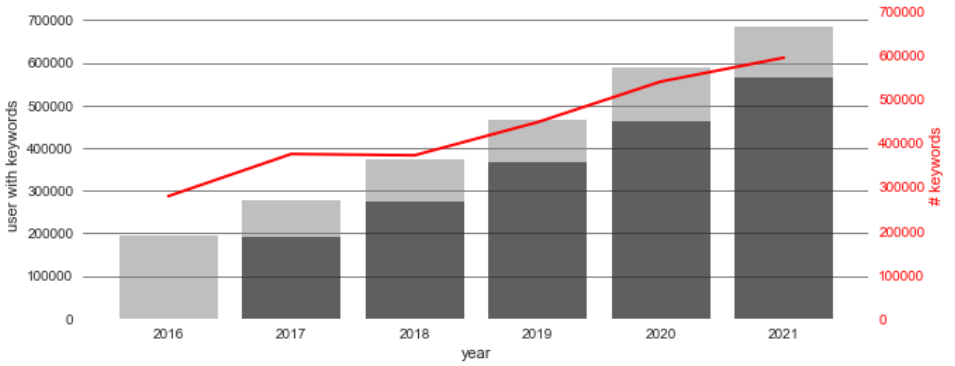


Figure 4.9: The Figure reveals the number of profiles with keywords over the years, the darkest part of each bar stands for the portion of the same profiles the current year shares with the previous one, and the red line is the number of unique keywords.

continent	UNESCO Science Report	ORCID
Europe	31%	41.09%
Americas	21.6%	26.36%
Asia	46.3%	19.17%
Africa	0.7%	6.84%
Oceania	0.3%	6.50%

Table 4.1: In-coming researchers in 2018 by continent, according to UNESCO Science Report and ORCID 2021 release. 2018 is the more recent year whose percentage appears in the UNESCO report.

that there is a subtle difference between the two: the first one indicates the real system, while the second one the mathematical representation.

Following the literature on network analysis for human migration, mainly the article by Fagiolo and Mastrorillo [3] about the International Migration Network, we have decided to build different models.

4.4.1 Overall Network

We consider a weighted directed network $G_T = (V, \varpi)$ where V is a set of nodes and $\varpi : V \times V \to \mathbb{N}$ is a function defining for each pair of nodes $i, j \in V$ the weight of edge (i, j).

In our application domain, we identify the nodes of the network as the countries involved in the scientific migration process; an edge between two countries represents a migration route. Each edge between two nodes $i, j \in V$ is attributed with a weight w: a triplet (i, j, w) represents the migration of w researchers from country i to the country j at time t. Since most of the

data is concentrated after 2000, and the geopolitical configuration of the countries is quite stable after 2000, we will encode in this network model all the researchers' movement in time domain $[2000, 2021]$. This model does not retain information about immediate returns, represented by self-loops. It has 243 nodes, and 9673 edges ($density = 0.164$), its strongly connected component retains 239 nodes, the reciprocity is 0.85, and its diameter of 4. Its strength distribution is shown in Figure 4.10.

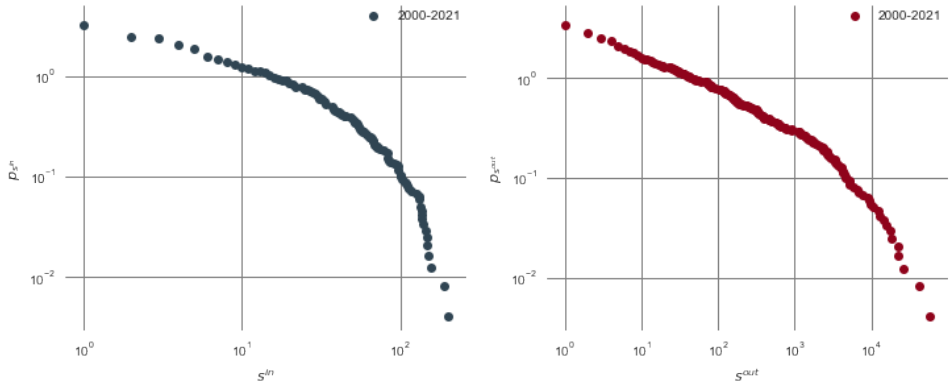


Figure 4.10: In-strength (left) and the out-strength (right) distributions in the overall network model G_T.

4.4.2 Temporal Network

We consider a weighted directed temporal network $G = (V, T, \varpi)$, where V is a set of nodes, $T = [t_0, t_1, \ldots, t_{max}] \subseteq \mathbb{N}$ is a discrete time domain, and $\varpi : V \times V \times T \to \mathbb{N}$ is a function defining for each pair of nodes $i, j \in V$ and each timestamp $t \in T$ the weight of edge (i, j) at time t. In the following, we refer to the weight of edge (i, j) at time t as $w_{ij,t}$, and we consider it missing if $w_{ij,t} = 0$. Let $s^{in}_{i,t} = \sum_{j \in V} w_{ji,t}$ and $s^{out}_{i,t} = \sum_{j \in V} w_{ij,t}$ represent the in-strength and the out-strength of node $i \in V$ at time $t \in T$, respectively. We also denote by $E_t = \{(i, j) \mid \varpi(i, j, t) > 0\}$ the set of edges existing at time $t \in T$. Finally, let W_t be the weighted adjacency matrix of G at time $t \in T$.

As in the overall network model, we identify the nodes of the network as the countries involved in the scientific migration process (231 in total); an edge between two countries represents a migration route. Each edge between two nodes $i, j \in V$ is attributed with a time $t \in T$ and weight w: a quartet (i, j, t, w) represents the migration of w researchers from country i to the country j at time t. The time domain of the scientific migration network is $T = [2000, 2001, \ldots, 2021]$, composed of 21 years, since most of the data is concentrated between 2000 and 2021, and the geopolitical configuration of the countries is quite stable after 2000. 2015 is the year for which the dataset records the largest amount of information. Figure 4.11 shows the evolution

of network dimension over the year, from 2000 to 2021, in terms of nodes, edges, and resulting density. We show in Figure 4.12 the in-strength and the out-strength distributions in the scientific migration network in 2000, 2010, and 2020. Other years are not reported here, but they show comparable behavior: the shapes of the distributions are very similar to each other. Also, there are no notable differences between in-strength and out-strength. Such distributions will come in handy in the following, as input of configuration models that create random graphs preserving in-strength and out-strength sequences. A more complete summary of basic network metrics is reported in Appendix A.2.

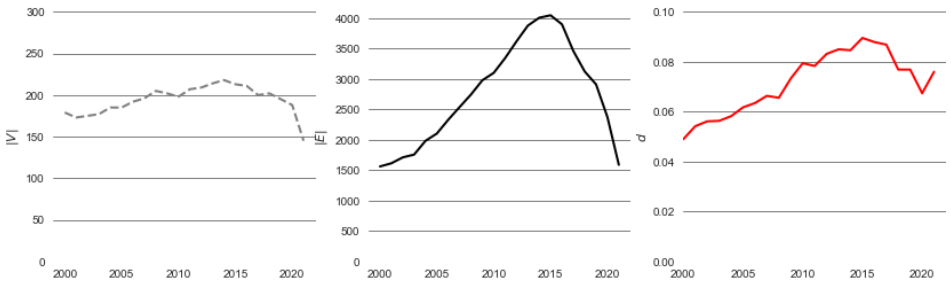


Figure 4.11: The evolution of network dimension over the year, from 2000 to 2021, in terms of nodes, edges, and resulting density.

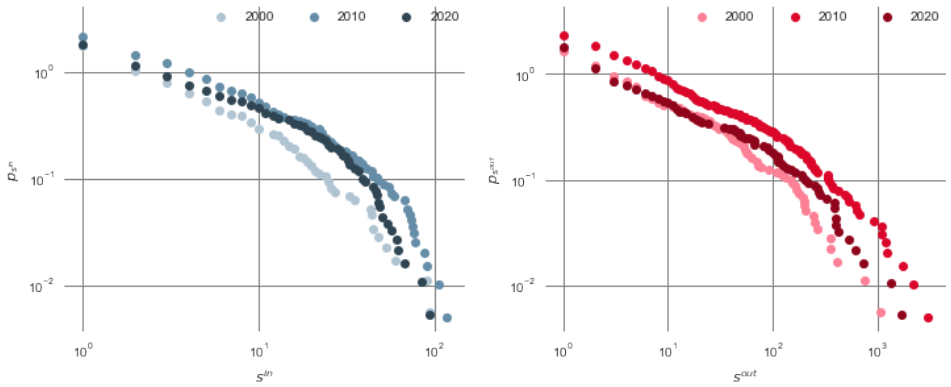


Figure 4.12: Cumulative in-strength (left) and the out-strength (right) distributions in the scientific migration network in 2000, 2010, and 2020.

4.4.3 Null Model

We employ the *weighted configuration model* [122, 123] as a null model to test whether findings and measure evaluations are non-trivial features of the scientific migration network or if they are expected by the strength distribution of the nodes. The configuration model generates a random directed graph

(with parallel edges and self loops) by randomly assigning edges to match the given degree sequences. In the resulting network parallel unweighted edges connecting two nodes have been combined to form a weighted edge. In Appendix A.5 some of the same analysis has been repeated discarding the self-loop, but in this case the strength sequence is only roughly preserved (around the 5% of edges has been removed in each round). In the following results, we consider ten different configurations of the null model.

4.5 Measuring the Brain Drain

4.5.1 From methods to measures

A strength-based approach can be considered a straightforward attempt to numerically quantify the role of a country in the scientific migration network .

We can intuitively define the *drain index* of a country $i \in V$ at time $t \in T$ as

$$\beta(i,t) = \frac{s^{out}_{i,t} - s^{in}_{i,t}}{s^{out}_{i,t} + s^{in}_{i,t}}, \tag{4.1}$$

namely the number of outgoing researchers (i.e., out-strength) minus the number of incoming researchers (i.e., in-strength) normalized by their sum. It ranges from -1 to 1, where 1 indicates maximum brain drain (the country is a pure provider) while -1 means maximum brain gain (the country is a pure receiver). Values close to 0 are adopted by those countries having balanced values of out-strength and in-strength.

Figure 4.13 graphically shows the drain index for the year 2020, while Table 4.2 reports the ranking for specific countries: the five countries of highest β, the five countries of lowest β, and the ten countries of highest out-strength. The countries standing out in Figure 4.13 are mainly located in Africa, the Middle East, Greenland, and Antarctica, while Europe and North America have milder colors. Extreme values of β are assigned when the number of migrations of a country is poor and completely unbalanced. For example, Maldives has only one outgoing migration, resulting in $\beta = 1$, while Liberia has two incoming migrations and no outgoing researchers, then its β is -1. On the other hand, those countries playing a central role in the migration network have usually β close to 0 due to the high number of both outgoing and incoming researchers. This is the case of, e.g., the United Kingdom and the United States. Drain index for 2000 can be found in Appendix A.3.

Of course, we would like to focus on countries whose number of moving scientists is not neglectable. In order to favor the identification of the central countries in the migration process, we lift the network by removing the links having weights lower than a certain threshold tr. This operation has the

ranking	country	β	s^{out}	s^{in}
1	Cayman Islands	1.0	1	0
2	Maldives	1.0	1	0
3	Greenland	1.0	1	0
4	Guinea	1.0	1	0
5	Guadeloupe	1.0	1	0
50	**Hungary**	0.**090909**	**42**	**35**
51	**Qatar**	0.**090909**	**24**	**20**
52	**United Kingdom**	**0.088535**	**1334**	**1117**
53	**Mexico**	**0.085470**	**127**	**107**
54	**United States**	**0.078100**	**1691**	**1446**
55	**Japan**	**0.077228**	**272**	**233**
56	**Portugal**	**0.075051**	**265**	**228**
57	**Brazil**	**0.071823**	**388**	**336**
58	**Russia**	**0.071429**	**90**	**78**
59	**New Zealand**	**0.057471**	**92**	**82**
183	Liberia	-1.0	0	2
184	Palestinian Territories	-1.0	0	1
185	Sint Maarten	-1.0	0	1
186	Timor-Leste	-1.0	0	1
187	Laos	-1.0	0	1

Table 4.2: Ranking (partial) of the countries by drain index β in 2020. For each country, out-strength and in-strength measured every year are also reported. Countries highlighted in bold have the highest out-strength in 2020.

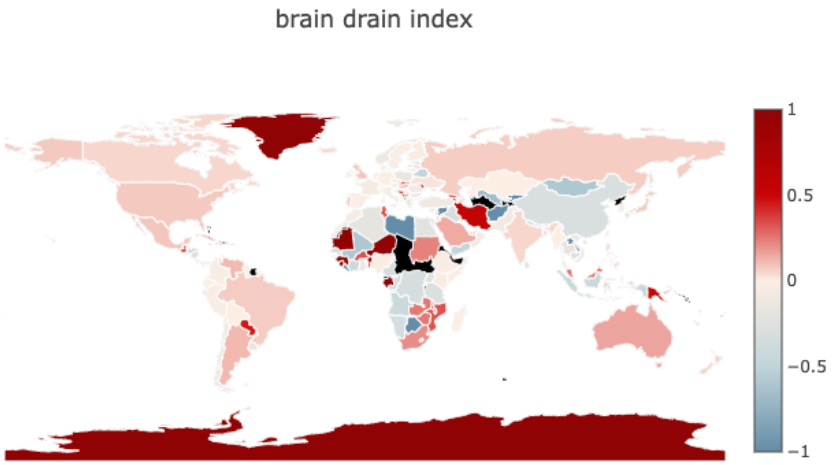


Figure 4.13: Drain index β in 2020. Positive (negative) values of β are color coded with different shades of red (blue). Countries without data have been colored black.

aim of discarding weak and not meaningful interactions between countries. We experimentally verify $tr \in [1, 2, \ldots, 10]$, and we report part of the 2014 ranking in Table 4.4 for threshold values of 1 (original network), 2, and 3. Two important aspects have to be considered: (i) the extremes of the ranking are not robust with respect to the threshold (the rankings shown in Table 4.4 considerably differ for small variations of tr); (ii) even for low values of tr, a large portion of the network is neglected by the analysis (47% and 62% for $tr = 2$ and $tr = 3$, respectively). Therefore, we cannot consider this approach a reliable and fair analysis of the scientific migration network .

In order to let emerge the *network backbone*, we apply the link filtering strategy that is proposed in [124]. This operation aims to focus on countries that have a leading role in the scientific migration flows while preserving the structural characteristics of the network as a whole. Figure 4.14 shows the fraction of nodes, links, and weights retained by the filters according to different significance levels α.

From the rankings displayed in Table 4.4, and calculated on the network backbones, we intuitively observe that a high instability emerges in such rankings at varying values of α. The ranking analysis is an open and very broad subject of interest, but a recent work [125] has shown a pattern throughout its dynamics, and how for example the top part of multiple rankings shared a certain degree of stability. Also in the $\beta(i, t)$ ranking, there are certain positions that carry out specific roles inside the migration system,

ranking	$tr = 1$ (original)	$tr = 2$	$tr = 3$
1	Cayman Islands	Somalia	Puerto Rico
2	Maldives	Cuba	Tunisia
3	Greenland	Cape Verde	Lithuania
4	Guinea	Paraguay	Papua NG
5	Guadeloupe	French Pol.	Sudan
50	Hungary	Zimbabwe	Taiwan
51	Qatar	Germany	Macau
52	United Kingdom	Hong Kong	Switzerland
53	Mexico	Sweden	Turkey
54	United States	Spain	Netherlands
55	Japan	Belgium	Chile
56	Portugal	Finland	Pakistan
57	Brazil	Italy	Norway
58	Russia	Denmark	Philippines
59	New Zealand	Colombia	Israel
182	Libya	Maldives	New Caledonia
183	Palestinian Territories	New Caledonia	Samoa
184	Laos	Samoa	Timor-Leste
185	Sint Maarten	Timor-Leste	Sierra Leone
186	Bermuda	Sierra Leone	Tonga
187	Liberia	Tonga	Liberia

Table 4.3: Ranking (partial) of the countries by drain index β in 2020, varying the threshold tr. The five countries of highest β (ties broken by out-strength) and the five countries of lowest β (ties broken by in-strength) are reported.

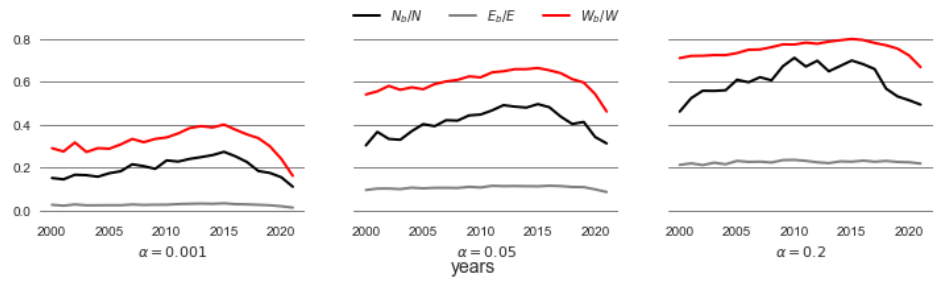


Figure 4.14: Focus on the network backbone: figures above show the percentages of retained nodes (N_b/N), edges (E_b/E), and weights (W_b/W) after the application of the filtering strategy. Each plot shows the application of the filter with increasing significance levels ($\alpha = \{0.001, 0.05, 0.2\}$).

$alpha = 0.001$	$alpha = 0.05$	$alpha = 0.2$
1 Saudi Arabia	1 Iran	1 Puerto Rico
2 Argentina	2 Puerto Rico	2 Paraguay
3 United Kingdom	3 Estonia	3 Sudan
4 Singapore	4 Jordan	4 Lebanon
5 Hong Kong	5 Nepal	5 Tunisia
...	...	...
25 South Africa	60 Algeria	92 Latvia
26 Israel	61 Thailand	93 El Salvador
27 Turkey	62 Ethiopia	94 Morocco
28 Taiwan	63 Philippines	95 Panama
29 Colombia	64 Uruguay	96 Cameroon

Table 4.4: Countries (partial) rankings by drain index β calculated on three different network backbones in 2014. Each backbone is extracted after the application of a filter with increasing significance levels ($\alpha = \{0.001, 0.05, 0.2\}$). The five countries of highest β (ties broken by out-strength) and the five countries of lowest β (ties broken by in-strength) are reported.

and we would like to estimate how stable they are over the years. To quantify it we define the Normalized Similarity s between two different partial rankings $\tilde{r_t}$ and $r_{\tilde{t+k}}$ as:

$$s(\tilde{r_t}, r_{\tilde{t+k}}) = 1 - \frac{1}{N(N+1)} \sum_{i \in \tilde{V}_t} |r_t(i) - r_{t+k}(i)| \tag{4.2}$$

where $t \in T$, $k \in [1, T-t]$, and V is the set of countries that takes part in the migration network at time t and occupy the chosen portion of the ranking. If at time $t + K$ a country is not the partial ranking anymore we place it

in the last position of the partial ranking. The term $\frac{1}{N(t)(N(t)+1)}$ is an upper bound for the sum of all the possible fluctuations, in particular, it would happen when all the countries at time t would downgrade at position $N+1$ while all new countries occupy the N position at time $t+k$, and $\frac{(N+1)(N+1-1)}{2} + \frac{(N+1)(N+1-1)}{2} = N(N+1)$. In Figures 4.15(a-c) the Normalized Similarity has been computed for the key positions of subsequent rankings based on β, calculated on the network backbone with level $\alpha = 0.2$, from the year 2000 to the year 2021. As key positions, we consider the top twenties (Fig. 4.15a), the bottom twenties (Fig. 4.15b), and the twenties in the middle (Fig. 4.15c) of each ranking, that should represent respectively the top providers, the top receivers, and the most 'balanced' countries. We can easily observe that, even with a fixed value of $\alpha = 0.2$, rankings differ significantly from one year to another; in fact, $s(r_i, r_{i+1})$ fluctuates around 0.6, meaning that the ranking calculated at year i changes dramatically the following year. The lack of stability over a not-so-fast phenomenon may prevent us to spot any significant patterns or dynamics.

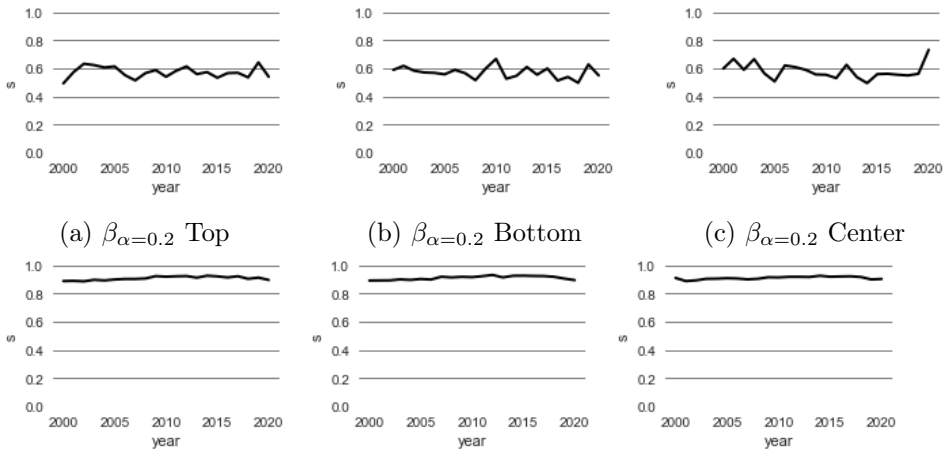
Additionally, we evaluated other strategies for normalizing the drain index by considering external data, such as the size of the overall population and the number of researchers in a country. Given the biases in the collected dataset, any normalization deriving from external sources would be inappropriate because it would misrepresent the results. Moreover, external data have to be temporal, at least of yearly granularity from 2000 to 2016, and available for all the countries included in the dataset. This is the case of the general population, but we cannot discover complete and coherent datasets about the size of the research population of all the studied states. However, we think that Eq. 4.1 fails mainly because it does not properly represent the complexity of the phenomenon itself: the brain index focuses on spotting 'pure receivers' and 'pure providers' in the network, whereas each country may behave accordingly mixed streams made of scientists moving in and out. As a consequence, such a measure would suffer from a myopic view of the migration ecosystem, because it is a function of local properties only: we miss the opportunity to assess which is the role of a global and heterogeneous structure of the migration network. This is the reason why we propose the application of eigenvector centrality-based algorithms to produce rankings more adequate to comparisons [126].

4.5.2 A global approach

A classic approach to assess the importance of a node in a network taking into account the global link structure is the well-known *PageRank* [99] described by Equation 3.3. In Figure 4.16 we graphically show the PageRank in 2020, while Table 4.5 reports the rank of the 20 countries having the highest

ranking	2000	2010 $s(r_{2000}, r_{2010}) = 0.94$	2021 $s(r_{2010}, r_{2020}) = 0.97$
1	United States	United States	United States
2	United Kingdom	United Kingdom	United Kingdom
3	Germany	Germany	Germany
4	France	Spain	China
5	Spain	France	Spain
6	Australia	Australia	Canada
7	Italy	China	Australia
8	Canada	Canada	Italy
9	Netherlands	Italy	India
10	Japan	Portugal	France
11	Brazil	Netherlands	Netherlands
12	Portugal	Switzerland	Switzerland
13	Switzerland	Sweden	Sweden
14	Sweden	Brazil	South Korea
15	South Korea	Japan	Belgium
16	Mexico	India	Japan
17	China	Malaysia	Portugal
18	Indonesia	South Korea	Brazil
19	Malaysia	Colombia	Norway

Table 4.5: Top-20 ranking by PageRank in 2000, 2010, and 2020.



(a) $\beta_{\alpha=0.2}$ Top (b) $\beta_{\alpha=0.2}$ Bottom (c) $\beta_{\alpha=0.2}$ Center

(d) Page Rank (e) Authority (f) Hub

Figure 4.15: s estimates the similarity between the rankings in two successive years. Plots in the first row represent similarities between the top twenties (a), the bottom twenties (b), and the middle twenties (c) in two successive years if we use the brain index defined in Eq. 4.1. Plots in the bottom row represent respectively the similarities between the top 20th countries in each ranking by page rank (d), authority score (e), and hub score (f).

PageRank in 2000, 2010, and 2020. As stated above, the drain index does not privilege nodes having high both in-strength and out-strength and does not account for the importance of the origin/destination of the connections. PageRank is instead able to picture such aspects; in particular, the United States and the United Kingdom place at the first and the second position of the ranking, respectively.

On the whole, PageRank is confirmed to be a powerful method to rank the nodes of a network, more stable according to the similarity measure s, as shown in Figure 4.15d and in Table 4.5. However, it assigns to each node a unique score that is not desirable in our setting, since we are instead interested in understanding the interplay between attraction and provision of researchers. Therefore, our analysis is required to rely on more refined and specific metrics that highlight such duality.

4.5.3 A dual approach: hubs and authorities

We identify the *hyperlink-induced topic search* algorithm (also known as HITS or *hubs and authorities*) [100] as the main measure to study our network (Equation 3.5).

By definitions, a node $i \in V$ has large value of h_i if it has many largely weighted links towards successor nodes $j \in V$ with high a_j; similarly, node i has large value of a_i if it is reached by predecessor nodes $j \in V$ with high h_j

ranking	2000	2010 $s(r_{2000}, r_{2010}) = 0.95$	2021 $s(r_{2010}, r_{2020}) = 0.91$
1	United States	United States	China
2	United Kingdom	United Kingdom	United States
3	Germany	Germany	United Kingdom
4	France	Australia	Germany
5	Canada	France	Canada
6	Australia	Spain	Spain
7	Spain	Canada	Italy
8	Italy	China	South Korea
9	Japan	Japan	India
10	Brazil	Portugal	France
11	Netherlands	Italy	Netherlands
12	South Korea	Switzerland	Australia
13	Portugal	Singapore	Switzerland
14	Switzerland	Sweden	Japan
15	Sweden	South Korea	Brazil
16	Mexico	Netherlands	Sweden
17	China	Hong Kong	Belgium
18	Malaysia	Brazil	Norway
19	Singapore	India	Portugal

Table 4.6: Best attractors of scientist: top-20 ranking by **authority** score in 2000, 2010, and 2021.

ranking	2000	2010 $s(r_{2000}, r_{2010}) = 0.93$	2021 $s(r_{2010}, r_{2020}) = 0.95$
1	China	China	United States
2	United Kingdom	India	United Kingdom
3	Canada	United Kingdom	Germany
4	South Korea	United States	Australia
5	India	Germany	Spain
6	Germany	Canada	Canada
7	United States	Italy	India
8	France	Spain	Brazil
9	Japan	France	Switzerland
10	Italy	South Korea	China
11	Brazil	Brazil	Italy
12	Spain	Colombia	France
13	Russia	Japan	Japan
14	Mexico	Portugal	Hong Kong
15	Australia	Australia	Sweden
16	Turkey	Netherlands	Netherlands
17	Colombia	Turkey	South Korea
18	Switzerland	Switzerland	Belgium
19	Netherlands	Iran	Portugal

Table 4.7: Best providers of scientist: top-20 ranking by **hub** score in 2000, 2010, and 2021.

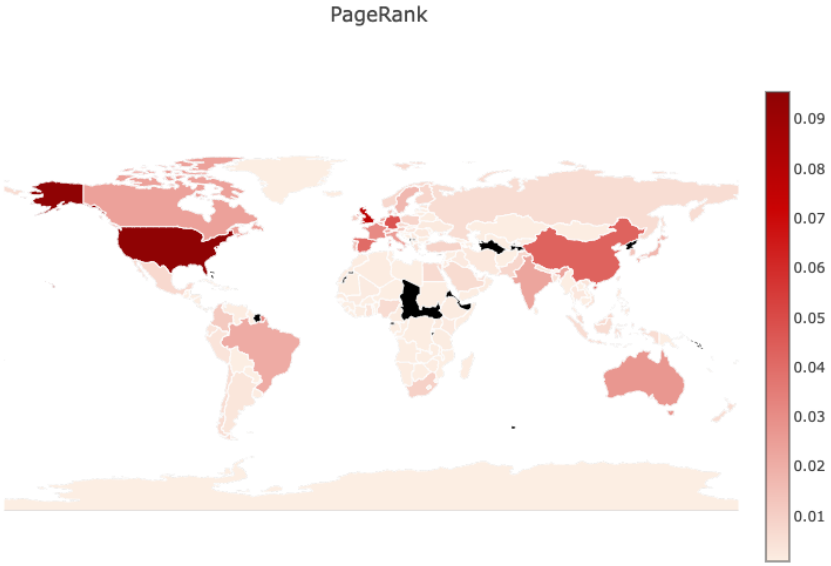


Figure 4.16: PageRank $\vec{r}_{2020}$ is color coded with different shades of red. Darker (lighter) red is used for countries with higher (lower) page rank values. Countries without data have been colored black.

throughout largely weighted links. In our specific scenario, $\vec{h}$ provides an indication of which are the countries playing the role of *providers*, that export many researchers in direction of the most attractive countries; while $\vec{a}$ indicates which are the *attractors*, whose institutions hire researchers from highly ranked providers. Tables 4.7 and 4.6 show the first twenty countries ordered by hub score and authority score, respectively, in 2000, 2010, and 2020, and the similarity score s between those years, whose consistency allows us some further analysis. The complete ranking can in found in Appendix A.4.

To provide a more in-depth understanding of the scientific migration patterns all over the world, we focus on which are the major players that rule it, and how their positions have changed over time in the ranking and inside the network structure, with the aim to detect important insight on which are the drivers that control the migration flows.

Figure 4.17 depicts the evolution of hub and authority scores of the nodes of the scientific migration network in time, by means of scatter plots. In all the years, most of the countries clump in the lower-left corner, where both scores are close to 0.

Most of the countries have comparable hubs and authority scores, meaning that if a country has a given role in the network as a scientists' provider,

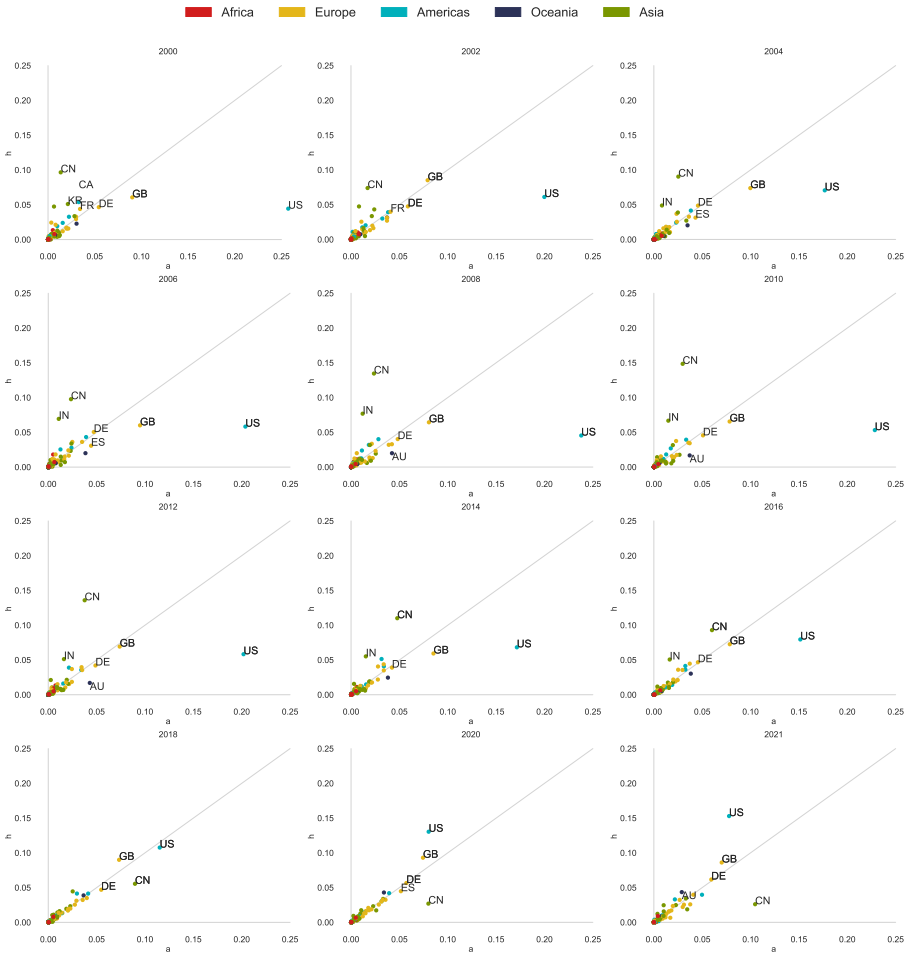


Figure 4.17: Evolution of hub and authority scores of the nodes of the scientific migration network in time. ISO 3166-1 alpha-2 codes are reported for selected countries: Australia (AU), China (CN), Germany (DE), India (IN), Italy (IT), Spain (ES), the United Kingdom (GB), and the United States (US).

then it is likely that it has a similar role as a scientists' receiver; in fact, as expected, the Pearson correlation between the two hubs/authority variables is quite high, with p-value always $< 1.5e-05$. However, when we calculate the Pearson correlation between $\vec{h}$ and $\vec{a}$ as a function of the year, and compare it to a null model, we find different trends (Figure 4.18). The correlation in the original network is strong during the whole time domain, constantly greater than 0.85. The null model has an even stronger correlation in all years, with small variations between the different configurations. The correlation between $\vec{h}$ and $\vec{a}$ and the evolution of such correlation is an interesting aspect to take into account. This means that we should expect

more countries of high (low) hub scores having also high (low) authority scores, and vice versa, in the scientific migration network , or that we have some outliers that buck the trends that could not have been expected with the null hypothesis, and that therefore are useful to characterize this peculiar ecosystem. The observed behavior should then rely on different factors, e.g., local patterns than the strength distribution.

Focusing on these outliers, we have that the United States performs significantly better as an authority than as a hub, even if the corresponding hub values are always among the highest, this scenario unexpectedly reverses in 2020 and 2021. On the other hand, the United Kingdom moves from being an equal hub and authority in the early '00 to being more authority by the end of the observed period. It is also easy to notice how China, which is constantly among the top hubs, slowly increases its authority score, with a tendency to the balance between the scores that is graphically represented by the diagonal and finally becoming predominantly an authority in 2020, showing almost an opposite trend with respect to the United States. Data in 2020 and 2021 are still less stable since many careers are still in the making being their time domain closer to the present, so any interpretation should not be considered final but still in the making. Such dynamics are particularly interesting, and they deserve further analysis. Details on these rankings for 2000, 2010, and 2020 are provided in A.4.

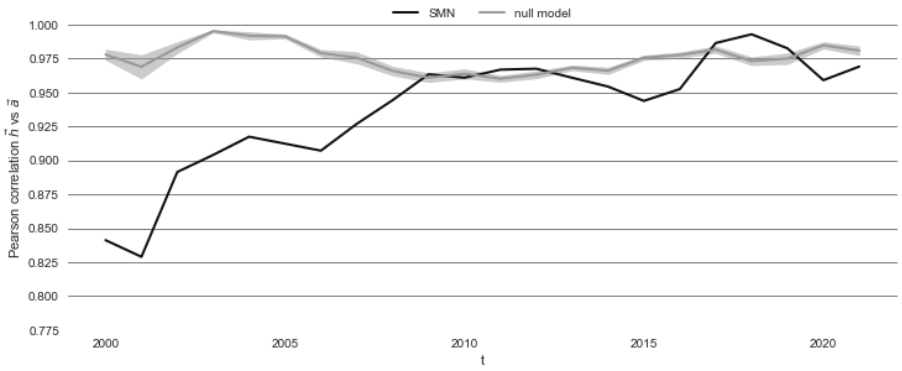


Figure 4.18: Person correlation between $\vec{h}$ and $\vec{a}$ of the scientific migration network and of the null model, for which we report mean and 95% confidence interval. p-values are smaller than $1.5e-05$ in all cases.

In order to compare the HITS and the PageRank results, in Figure 4.19 we also visualize the Pearson correlation between $\vec{h}$ and $\vec{a}$, and $\vec{r}$. Interestingly, both $\vec{h}$ and $\vec{a}$ are highly correlated to $\vec{r}$. $\vec{a}$, in particular, has correlation greater than 0.95 in all years. This validates the results obtained by the HITS algorithm that has the advantage of depicting two different aspects of the world countries, providing then more accurate indications.

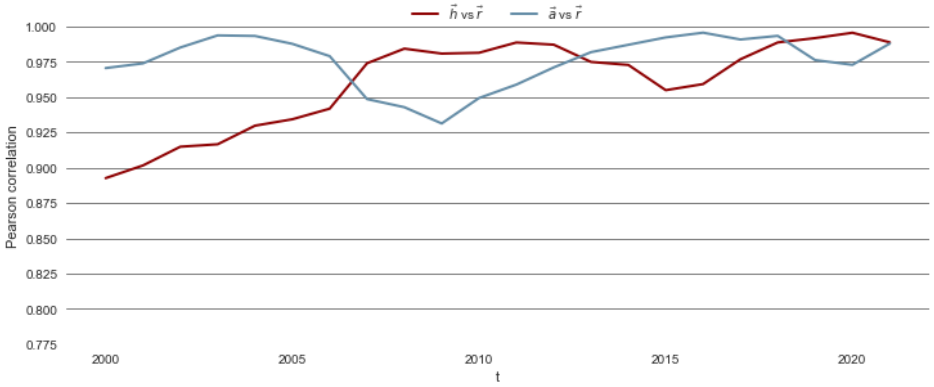


Figure 4.19: Person correlation between $\vec{h}$ and $\vec{a}$, and $\vec{r}$ of the scientific migration network .

4.6 Beyond the drain

To dive deeper into the factors that contribute to establishing a country as a leading hub or authority in the scientific migration network , we will tackle three main research tracks:

- How is the neighborhood of a principal hub or authority shaped by its characteristics?
- Which countries recall their researchers the most after a certain amount of time?
- Case studies of countries that have emerged in the evolution of one or both hits ranking.

4.6.1 Analyzing local patterns with predecessors and successors

In the next section, we rely on the study of statistical dispersion of incoming and outgoing edge weights to provide a better understanding of such local patterns, exploiting the homogeneity of the edge weights of the neighborhood of the nodes. Specifically, we want to understand how the researchers leaving (reaching) a country with a high hub (authority) score is distributed over the outgoing (incoming) routes. In order to do so, we employ the *Gini coefficient*, which measures the degree of inequality of a distribution [127]. Given a population $\mathbf{W} = \{w_o, w_1, \ldots, w_n\}$ of n values, we define the Gini coefficient as

$$G = \frac{\sum_{w_i, w_j \in \mathbf{W}} |w_i - w_j|}{2n \sum_{w_i \in \mathbf{W}} w_i}. \tag{4.3}$$

G varies between 0 and 1, where 1 expresses maximal inequality among values while 0 indicates the case in which all the values in **W** are equal.

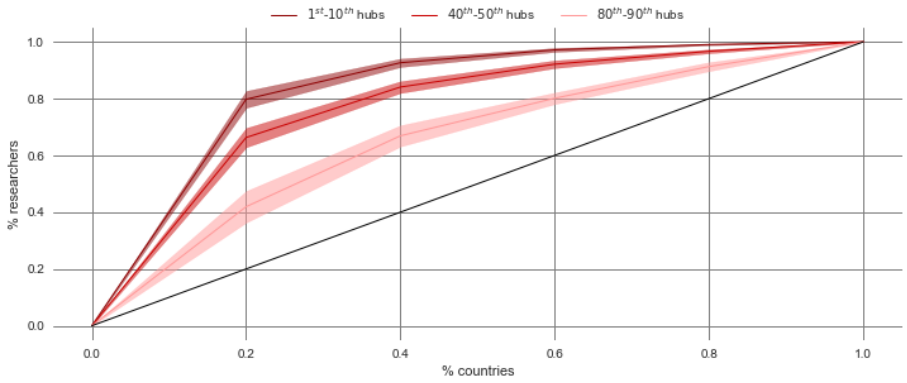


Figure 4.20: Lorenz curves and 95% confidence intervals for three classes of hubs in 2015. The population **W** is represented by the edge weights of outgoing edges.

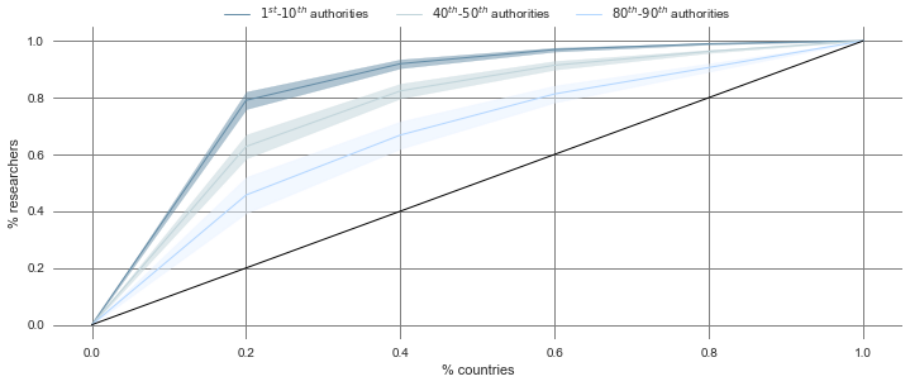


Figure 4.21: Lorenz curves and 95% confidence intervals for three classes of authorities in 2015. The population **W** is represented by the edge weights of incoming edges.

By means of Lorenz curves, it is possible to identify the population **W** as the edge weights of outgoing edges or the edge weights of incoming edges when considering a node as hub or authority, respectively. Therefore, we aim at investigating how (un)balanced the migration flows from/towards a country are and how such aspect correlates to $\vec{h}$ and $\vec{a}$. Figures 4.20 and 4.21 compare the mean Lorenz curves, along with 95% confidence intervals, of three different classes of hubs and authorities, respectively. It is immediately noticed that a high hub/authority score is associated with a high Gini coefficient. The Gini coefficient decreases progressively as we move down

with the hub and authority rankings. Then, to obtain an important position in the scientific migration network , a country is required to have strongly differentiated migratory flows from/towards its neighbors.

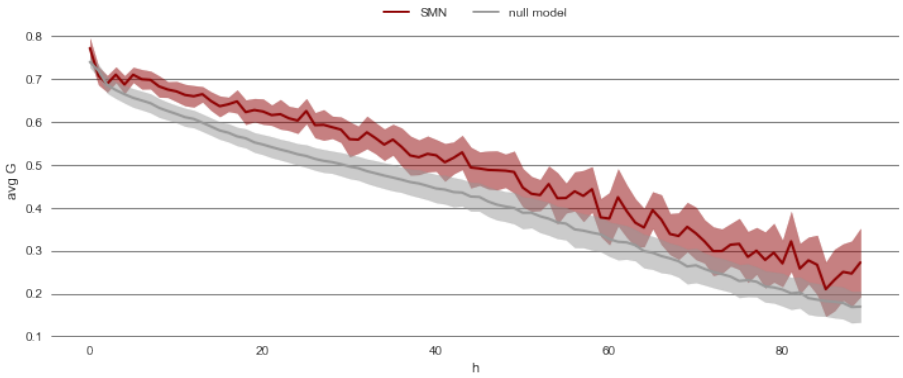


Figure 4.22: Average Gini coefficient (and 95% confidence interval) as a function of the hub ranking of the scientific migration network and of the null model. The population **W** is represented by the edge weights of outgoing edges and the average is computed over the time domain T.

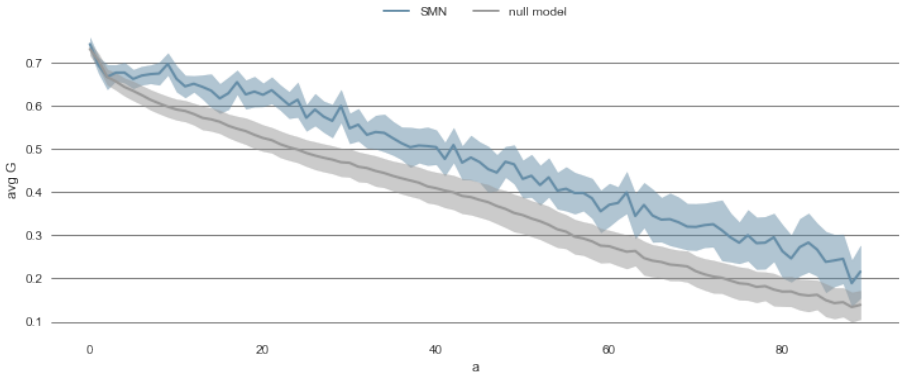


Figure 4.23: Average Gini coefficient (and 95% confidence interval) as a function of the authority ranking of the scientific migration network and of the null model. The population **W** is represented by the edge weights of outgoing edges and the average is computed over the time domain T.

The behavior of the missing classes is consistent as shown in Figures 4.22 and 4.23 which report the average (over the time domain T) of the Gini coefficient (and the 95% confidence interval) as a function of the hub/authority ranking. Such curves are compared with the null model considering the average of the ten different configurations we generated. The Gini coefficient decreases as h and a drop, both in the scientific migration network and in the null model, and the curves have very similar functional shapes. The con-

fidence intervals are quite limited in all cases, however, they become larger for the lowest positions of the ranking in the scientific migration network where data become more sparse and less significant. The Gini coefficient of the scientific migration network is higher than the null model; this means that a node occupying the first position in the hub/authority ranking also shows a high disparity in the weights of the connections from/to its predecessors/successors by the intrinsic characteristics of the network. Refer to Appendix A.5 a for comparisons with null models without self-loops, results are consistent with or without self loops.

The main hubs and authorities of the SMN tend to be central in the migration paths traversing the network, shaping their local neighborhood, and capitalizing the connections towards them. So besides the role that a country has in the overall scientific migration network , it is of our interest to understand better the interplay between their predecessors and successors.

We define the *betweenness centrality* of a node $i \in V$ at time $t \in T$ as in Equation 3.6. In the computation of the scores, we consider the reciprocal of the edge weights of the scientific migration network , since the more a path is favorable (i.e., shorter) the more researchers move through such a path. Therefore, c_b is an indication of how much a country is central in the crossing of the network by the researchers.

We also compute the *clustering coefficient* of a node $i \in V$ at time $t \in T$ as

$$cc(i,t) = \frac{|(j,k) \mid j,k \in N_{i,t} \wedge (j,k) \in E_t|}{|N_{i,t}|(|N_{i,t}| - 1)}, \tag{4.4}$$

where $N_{i,t}$ identifies the neighbour set of node i at time t. In this case, we neglect the edge weights. In our context, we consider the cc of a country i as a measure of how many possible origins or destinations the researchers residing in neighboring countries have rather than i.

Betweenness and clustering coefficients are known to measure quite opposite node's behaviors: if a node shows high betweenness, we can easily place it at the edges of different network clusters, while a node with a high clustering coefficient is usually very well embedded in a tightly connected community. At the same time, nodes of high betweenness centrality and low clustering coefficient are placed at the borders of their local clusters and have direct ties to other clusters. Therefore, we can suppose that a country with such characteristics is one of the two endpoints of a *bridge*, or more likely of a *local bridge* [128] (the local bridge is a relaxed definition of a bridge, i.e., if we delete a local bridge the two endpoints would lie further away and not in two different components of the network). The endpoints of a (local) bridge regulate the access toward different clusters of nodes and are crossroads of the flows within the network. Hence, since scientific migration moves also ideas and information in addition to people: these countries may have early access to knowledge and to new research results, possibly produced in multiple and non-interacting places of the world. So we presume that this position, i.e.,

at the endpoint of a (local) bridge, could be a potential goal for the majority of the countries.

Figure 4.24 reports the trajectories of the countries that in 2021 occupies one of the top position either in the authority or in the hub ranking. Each line of the plot is associated with a country: the point represents the country coordinates in 2000, while the star shows the same country coordinates in 2021. We cannot observe a global pattern, common to most of the countries, leading toward the lower-right corner: some of the nodes move towards the upper-left corner, and others to more central positions. Even among the top hub and authority, very polarized dynamics emerge. Some countries retain betweenness centrality values almost close to 0 while having high values of clustering coefficient, for example, South Korea in the authority ranking or Switzerland in the hub ranking. They do not occupy a position along the most favorable migration paths, but nevertheless, they are in close proximity to many other potential sources of researchers, and their institutions may be strongly invested in attracting researchers and, at the same time, limiting further brain drain as much as possible. For sure, they are in a very dense and so competitive area of the network. Some European countries emerge following different paths, Italy and France seem to tighten their cluster structure, while Spain and Germany move toward also higher values of betweenness centrality. The United States and Great Britain are the most prominent outlier, they remain stable on the high value of betweenness and low value of the clustering coefficient. We previously observed in Figure 4.17, how the United States and China seem to move according to opposite trends: China started being predominately a hub to become more an authority while the United States started as the most predominant authority to become more a hub. This scenario seems to be mirrored in Figure 4.24c where China seems interested to move toward a more bridge-like position, while the United States loses some dominance.

4.6.2 Returns

To see if there is a specific pattern between returns and migration we study the propensity of countries to recall researchers who had left in the past, even after several years. *Returns* can be of two types:

A A return can be part of a current career stage of a researcher. It could happen if they have more than one affiliation at the same time. We can detect those cases in Figure 4.7, following routes 3, 5, 7, and 8. In these situations, a researcher does not leave definitely the first country but they retain a working link.

B We define a "proper" return as the event that occurs when a researcher for at least one year stops having any official link (affiliation) with the original country.

A return is characterized by a time span δ_t, which we define in terms of the number of years from when a researcher left a certain country to when they finally returned. Overall, of the 704.177 users migrating, 669.690 never return following a routes of type B, as described above. Figure 4.25 show the δ_t distribution for all returns that happened between 2000 and 2021. The bulk of the returns happens around between 4 and 6 years, the distribution increases fast toward the median and decreases more slowly after reaching the peak. Without further analysis, this scenario does not hold any clear explanation, we can only formulate hypotheses that may become future research questions, for example, we might hypothesize that education drives initial growth, one, two, three, or five years are compatible with training paths, master's or doctoral degrees, which have a specific duration, after six years the scenario becomes more complex and less discrete, that smooths any signal in a slower decrease.

Table 4.8 shows the countries' rank by the number of returns according to δ_t. Spain dominates the ranking for $\delta_t = 1$, and some countries emerge that were not present in previous HITS rankings, at least in the top-twenty position, like Pakistan, which enters the ranking in position 19^{th} for $\delta_t = 3$ but also Ecuador, Bangladesh, and Egypt, all three in 20^{th} position for $\delta_t = 2$, $\delta_t = 5$ and $\delta_t = 6$. (See A.6 for ranking values.)

Returns are still part of migrations, so their absolute values could be dominated by the flow of incoming researchers. In order to acquire different information we define for each country i the Normalized Return Index:

$$\hat{r_i}(\delta_t) = \frac{r_i(\delta_t)}{s_i^{in}} \tag{4.5}$$

where $r_i(\delta_t)$, and s_i^{in} is defined over the network model G_T describe in Section 4.4.1 that encodes the aggregate migration data from 2000 to 2021. The domain index is $[0, 1]$. Table 4.9 summarizes the first twenty positions of each ranking. Different patterns can be noticed, some countries maintain relatively stable positions over time, while others emerge just for some values of δ_t. India, for example, is almost always positioned around the 10^{th} position, while Croatia, from 4^{th} position for $\delta_t = 1$, after decreasing in 15^{th} for $\delta_t = 2$, disappears from the top-ranking positions. This index prefers countries whose largest component in incoming migrations is in fact a return, yet we see occupying high positions in the ranking especially countries that we can imagine are not the focus of much traffic. This index has also structural limitations, for example in the aggregate network, we add arcs and units to streams without removing them when a researcher is no longer in a country. In the future, we aim to refine this measure to obtain interpretations of the phenomenon even closer to reality.

	1	2	3	4	5
1	Spain	United States	United Kingdom	United States	United States
2	United Kingdom	United Kingdom	Spain	United Kingdom	United Kingdom
3	United States	Spain	United States	Spain	Spain
4	Germany	Germany	India	India	France
5	Portugal	France	France	Italy	China
6	Italy	India	Germany	France	India
7	India	Italy	Italy	Germany	Italy
8	France	Portugal	China	China	Germany
9	Brazil	China	Portugal	Portugal	Brazil
10	China	Canada	Brazil	Brazil	Portugal
11	Colombia	Brazil	Australia	Colombia	Australia
12	Sweden	Netherlands	Colombia	Denmark	Colombia
13	Australia	Colombia	Sweden	Canada	Canada
14	Canada	Australia	Netherlands	Sweden	South Korea
15	Netherlands	Denmark	Canada	Pakistan	Pakistan
16	Singapore	South Korea	Japan	South Korea	Japan
17	Austria	Belgium	South Korea	Australia	Switzerland
18	Ireland	Japan	Switzerland	Japan	Netherlands
19	Russia	Switzerland	Pakistan	Netherlands	Sweden
20	Turkey	Ecuador	Belgium	Switzerland	Bangladesh
ranking	6	7	8	9	10
1	United States	United States	United States	United States	United States
2	United Kingdom	United Kingdom	United Kingdom	United Kingdom	United Kingdom
3	Spain	Spain	China	China	China
4	China	China	Spain	Spain	Spain
5	India	Germany	India	India	India
6	Italy	France	Germany	Germany	Germany
7	Germany	India	France	Portugal	France
8	France	Portugal	Italy	Italy	Colombia
9	Portugal	Italy	Colombia	France	Portugal
10	Colombia	Brazil	Brazil	Brazil	Italy
11	Brazil	Colombia	Australia	Colombia	Australia
12	Japan	Australia	Portugal	South Korea	Brazil
13	Australia	South Korea	South Korea	Japan	Japan
14	Switzerland	Canada	Canada	Australia	South Korea
15	South Korea	Netherlands	Japan	Turkey	Sweden
16	Netherlands	Japan	Turkey	Mexico	Canada
17	Canada	Pakistan	Sweden	Sweden	Pakistan
18	Pakistan	Turkey	Mexico	Canada	Turkey
19	Mexico	Switzerland	Pakistan	Egypt	Egypt
20	Egypt	Greece	Switzerland	Netherlands	Mexico

Table 4.8: Ranking of countries by the total number of returns and according to time difference δ_t. The table shows the first twenty positions of the ranking and values of δ_t between 0 and 10.

	1	2	3	4	5
1	Burundi	Antigua and Barbuda	Niger	Vanuatu	Svalbard
2	Seychelles	French Polynesia	Eritrea	Uzbekistan	Madagascar
3	Liberia	Mali	Tajikistan	Papua New Guinea	French Guiana
4	Uzbekistan	Burkina Faso	Uzbekistan	Myanmar	Haiti
5	Croatia	Palestinian Territories	Lesotho	Greenland	Guadeloupe
6	Afghanistan	Ukraine	Greenland	French Polynesia	Myanmar
7	Cameroon	Malta	Maldives	Georgia	Uzbekistan
8	Palestinian Terr.	Cambodia	Trinidad and Tobago	Libya	Togo
9	Hungary	Tunisia	Cape Verde	Armenia	Montenegro
10	Estonia	Honduras	Malawi	Montenegro	South Sudan
11	Sudan	India	India	India	Macedonia
12	Senegal	Uzbekistan	Nepal	Afghanistan	Serbia
13	Cambodia	Armenia	Syria	Pakistan	Belarus
14	India	Sri Lanka	Cameroon	Cambodia	Guyana
15	Portugal	Croatia	Ghana	Benin	Somalia
16	Puerto Rico	Belarus	Bolivia	Palestinian Terr.	Ghana
17	Lithuania	Afghanistan	Madagascar	Ethiopia	Bangladesh
18	Rwanda	Algeria	Rwanda	Bangladesh	Pakistan
19	Italy	Somalia	Greece	Iceland	Palestinian Terr.
20	Romania	Benin	Honduras	Sri Lanka	India
	6	7	8	9	10
1	Kyrgyzstan	Burundi	Burundi	Samoa	Guadeloupe
2	Tajikistan	Belize	Guinea	Uzbekistan	Jordan
3	Seychelles	Tajikistan	Laos	Bolivia	Papua NG
4	Liechtenstein	Bolivia	Fiji	Bhutan	Paraguay
5	Mauritius	Gambia	Azerbaijan	Mali	Sudan
6	Madagascar	New Caledonia	Montenegro	Jamaica	Swaziland
7	Niger	Malta	Bulgaria	Georgia	Palestinian Terr.
8	Cape Verde	Laos	India	Cameroon	Nepal
9	Serbia	Barbados	Zimbabwe	Nepal	Libya
10	Nepal	French Guiana	Greece	Palestinian Terr.	Mongolia
11	Papua New Guinea	Zimbabwe	Mauritius	India	Tunisia
12	Pakistan	Armenia	Gambia	Lesotho	Myanmar
13	Bangladesh	Uzbekistan	Madagascar	Libya	India
14	India	Belarus	Mongolia	Madagascar	Azerbaijan
15	Bhutan	Guyana	Uruguay	Cote dIvoire	Pakistan
16	Syria	Somalia	Pakistan	Angola	Syria
17	Jamaica	Libya	Honduras	Myanmar	Peru
18	Greece	Papua NG	Dominican Rep.	Bangladesh	Macedonia
19	Bolivia	Ghana	Zambia	Greece	Lithuania
20	Zambia	Serbia	Bangladesh	El Salvador	Lebanon

Table 4.9: Ranking of countries by Normalized Return Index 4.5 and according to time difference δ_t. The table shows the first twenty positions of the ranking and values of δ_t between 0 and 10.

4.6.3 Spotting the heterogeneity in case studies

ORCID data are constantly evolving so instead of focusing on specific case studies we propose a method for extracting interesting case studies that can be updated without structural changes at each public data dump. To pinpoint the country that leads a cumulative interesting pattern we define the following index:

$$\Delta_i(R(r,p,t)) = \sum_{t\in[1,T-1]} \frac{p(t+1)-p(t)}{r(t+1)-r(t)} \tag{4.6}$$

where $R(r,p)$ is a evolving ranking, that is a function $R : p \times t \to \mathbb{R}$, that assign to the coupled variable p,t, respectively position and time, a value r. In our setting, R could represent either the authority or the hub ranking at a specific year t with $T = [2000, 2021]$, for each year a country that occupies a position p will have a certain value of either authority or hub. $\Delta_i(R)$ allows selecting those countries that change distinctly their positions in the hub and authority rankings from 2000 to 2021. Figure 4.26 illustrates the countries with higher $\Delta_i(R)$ with respect to the authority ranking over the year (a) and with respect to the hub ranking over the year (b). We filter the countries that in the last year of the time domain (2021) have reached at least the 50^{th} position. In the authority rank, for example, Qatar patterns emerge significantly: Qatar gain 50 positions between 2000 (90th) and 2014 (32nd). In the hub ranking, we can depict the fall of Finland in 2005. Overall within the fifty top-position countries, HITS values seem very stable across the years. Filtering by the final ranking, hide the countries that worsen their position considerably, but the index $\Delta_i(R)$ is versatile enough to be adapted to the wanted research question. It could also be computed just for the first and last year giving a stark view of how a country has changed its role over the years.

4.7 Beyond the brain

To outline the evolution of the research map in term of which countries invest in different domain of science, we leverage routes and change of status to the keywords each researcher use to describe their own knowledge and wealth of skills, changing the point of view of the migrations, we stop following the flow of people and start tracking fields of expertise instead.

We consider a weighted temporal bipartite network $G_k = (\{V, K\}, T \times S, \varphi)$, where V, K are not overlapping set of nodes, $T = [t_0, t_1, \ldots, t_{max}] \subseteq \mathbb{N}$ and $S = [s_0, s_1]$ are respectively a discrete time domain and a status domain. and $\varphi : V \times K \times T \times S \to \mathbb{N}$ is a function defining for each pair of nodes i, j such that $i \in V$ and $j \in K$ and each timestamp $t \in T$ the weight of edge (i, j) at time t for the setting s. In the following, we refer to the weight of

edge (i, j) at time t and the status s as $w_{ij,t,s}$, and we consider it missing if $w_{ij,t,s} = 0$.

We also denote by $E_t = \{(i, j) \mid \varpi(i, j, t, s) > 0\}$ the set of edges existing at time $t \in T$. Finally, let W_t be the weighted adjacency matrix of G at time $t \in T$ and for $s \in S$. The network model describes is a two-layer bipartite network, in which the belonging of a coupled pair of nodes to a certain layer is encoded by the variable S. Each bipartite network is undirected. In this framework, each coupled bipartite network G_k is defined by between the set nodes keywords and countries for each time stamp $t = [2000, \ldots, 2021]$, and where each layer s represents either the status before the migration or after the migration, which we will refer to as "from" and "to", to better focus on the context of the problem, the set of keywords K flows from the set of countries $V(t, s_0)$ to the set of countries $V(t, s_1)$.

To build these models from the data described in Figure 4.9 we follow the schema 4.27: in the migration database, we replace the researcher ORCID id with the list of keywords which he himself included in the platform to summarize their work and research topic of interest.

Table 4.11 shows the ranking of the most migrating keywords in 2000, 2010, 2020, and 2021. The rise of "machine learning" as a research field was almost absent in 2000, but also for example "climate change" occupied the highest positions event in 2010.

The UNESCO Report highlights how many countries starting in 2016, adopted dedicated strategies for AI, striving to assume a leadership role in the international conversation. This purpose is endorsed by many investments, for example, China has launched many programs in science and engineering to 2030 that include quantum computing and brain science. Also, the United State government's 2020 research budget proposal for 2021 included major increases for quantum information science and AI. A the same time, the converging phenomena of strong economic growth, heightened dependence on technology, and rising temperatures are driving up energy needs. Countries are keenly aware that their future economic competitiveness will depend upon how quickly they enact the green transition. This intention emerges also in the rising of researcher keywords like bioinformatics, renewable energy, and biotechnology. 2020 was deeply impacted by the Covid-19 pandemic, and it emerges in 4.11, so much so keywords like "public health" and "epidemiology" climb the ranking positions. It will be interesting to understand the depth of change as the years go by.

Table 4.11 might show a bias toward the most used keywords per year, so, to analyze a different point of view, we normalized each migrating keyword by its total usage, obtained by extrapolating the keywords of these users who have an active affiliation during y, for $yin[2000, 2021]$. The total number of users is represented by Figure A.1. The rankings are completely different and contain words in languages other than English, which suggests a larger effort to create a single dictionary to address the problem, which we intend

to create in the future development of the project.

	2000	2010	2020	2021
1	cancer	bioinformatics	machine learning	machine learning
2	bioinformatics	machine learning	bioinformatics	qualitative research
3	immunology	climate change	qualitative research	bioinformatics
4	genetics	nanotechnology	public health	remote sensing
5	genomics	neuroscience	epidemiology	electrochemistry
6	gis	molecular biology	neuroscience	cancer
7	remote sensing	immunology	ecology	genomics
8	education	microbiology	artificial intelligence	climate change
9	epidemiology	public health	microbiology	biotechnology
10	computer science	epidemiology	molecular biology	archaeology
11	architecture	remote sensing	remote sensing	molecular biology
12	malaria	sustainability	cancer	microbiology
13	nanotechnology	genomics	climate change	nanotechnology
14	innovation	cancer	evolution	biomaterials
15	ecology	genetics	catalysis	renewable energy
16	biotechnology	nanomaterials	innovation	graphene
17	neuroscience	ecology	genomics	immunology
18	sustainability	artificial intelligence	biochemistry	additive manufacturing
19	organic chemistry	biochemistry	nanomaterials	public health
20	biochemistry	education	electrochemistry	nanomaterials

Table 4.10: Ranking of the most migrating keywords in 2000, 2010, 2020, and 2021.

	2000	2010	2020	2021
1	political behavior	political behavior	raman spectroscopy	immersive media
2	nuclear fusion	nuclear fusion	general management	health service research
3	adhesion	adhesion	world heritage	migration and refugee studies
4	cancer cell biology	cancer cell biology	latin	monsoon dynamics
5	systems ecology	oncogenes	electricity markets	plant stress physiology
6	adolescent health	embryogenesis	raman	bi-/multilingual education
7	sexual and rep. health	molecular chaperones	ict	biosensing
8	social identity	alzheimer's disease	materials physics	nuts
9	oncogenes	social identity	ict in education	homologous recombination
10	molecular chaperones	sexual and rep. health	investigación científica	protein interaction
11	embryogenesis	adolescent health	climate modeling	climate modeling
12	alzheimer's disease	systems ecology	coastal protection	international studies
13	economia	economia	solar power	neurodegenerative disorders
14	technology assessment	estrogen receptor	carlos	systematic botany
15	evidence based practice	seismic hazard	seismic attributes	nano-particles nanotechnology
16	operations strategy	debris	oct	general management
17	enfermero	chirurgie cardiaque	disease models	candida
18	historia económica	microbiome research	multiscale biology	aspergillus
19	computational topology	exocytosis	x-rays	supramolecular polymers
20	signalling	wildlife conservation	transhumanism	china

Table 4.11: Ranking of the most migrating keywords in 2000, 2010, 2020, and 2021, normalized by the yearly usage.

To better understand the different distribution of skills on a global scale, we project the coupled bipartite network over both the countries and the keyword space. Each bipartite network generates two projections. Figure 4.28

presents the network framework for 2021, with both statuses of the keywords' redistribution. On the top of the Figure we have the projection over the countries' space, a link between two countries i and j depicts a similarity over the composition of expertise shared among the researchers the leave either i or j, for the network regarding the status "form" or reaching them, for the network regarding the status "to". That similarity is reinforced by the weight of the link. What surfaces from the Figure is the appearance of a new community in the "to", which includes Italy, Norway, Finland, but also New Zealand among others. Both projections over the keyword have been represented in their reduced version, filtering out links with a weight of less than two for ease of interpretation. Node dimension scale over the strength, while colors follow the belonging to the same community, extrapolating by means of modularity [129]. It seems that a major break happens in the "machine learning" community, with respect to the status "from" after the migration is still a big player but it does not strongly belong to any community, maybe because it relates too much with too many other skill sets.

Finally, the last analysis possible is to follow the evolution of a specific set of keywords as shown in Figure 4.29 for "machine learning" and "climate change" for 2015. In this case, to build the network we simply filter the temporal network analysis in Section 4.4.2 by the wanted set of keywords. A clear downside of the data is the lack of a unified lexicon that would define a specific topic or area of research. ORCID is a growing platform, but at this point, the resulting network regarding a specific domain is not up to the real phenomenon, so we can regard these results as a guide to deeper analysis or new research ideas, but still not a perfectly adherent representation of reality.

4.8 Final Remarks

In this work, we study international migrations of researchers, scientists, and academics using a complex network-based approach. This is a data-driven study that due to the dataset bias cannot be considered definitive. We mainly focus on proposing a methodology to be applied to data extracted from the ORCID platform to find a measure to quantify the phenomenon of the brain drain, to move beyond the idea of drain as a unique interpretation of the phenomenon, and to try to change the point of view from people to set of skills.

First of all, we discarded the adoption of very localized measures that take into account only the number of scientists moving in or out because they lead to rankings that change dramatically from one year to another. As a consequence, we propose to preserve the complexity of the migration ecosystem with adequate measures, that also maintain the dual nature of a country as both an importer and an exporter of researchers. Therefore,

we model the scientific migration by means of a temporal weighted directed network and employ the HITS algorithm with the intent of catching the interplay between streams of incoming and outgoing researchers from a global perspective. We also investigate the local characteristics of successors of hubs and predecessors of authorities to dive deeper into the motivations that establish hubs and authorities, and how within the time a country is able to call back some of the researchers that used to work there.

Our findings identify different positions occupied by the main player in the scientific migration network , as shown in Tables 4.6, Tables 4.7. China, the United States, and the United Kingdom are identified as the leading provider countries during the whole time domain: they never fall below the fifth position. India and Canada, followed by various European countries, i.e., Germany, Italy, Spain, and France, consistently position after the three leading countries with few fluctuations during the years. South Korea and Russia follow instead negative trends. South Korea for example occupied the fourth position in the hub ranking in the scientific migration network during 2000, then loses thirteen positions by 2021. Also, China from 2020 looses its first position in the hub ranking in favor of the United States. Regarding the authority score, the United States have the best performance until 2019, to be surpassed by China in 2020. Germany generally occupies the 3$^{\text{rd}}$ position in early 2000, before the growth of China. Similarly to the hub score, after the top-4 positions, there is a series of European countries such as Spain, France, and Italy, together with Canada and Australia, and South Korea in 2021. Interestingly, among the best receivers, there are countries that are not identified as good hubs, e.g., Qatar that emerge from the $\Delta_i(R(r,p,t))$ index, suggesting important efforts in attracting researchers from all over the world and investing for the return of whom left the countries. These dynamics deserve to be further analyzed for uncovering latent causes and factors through the inclusion of complementary sources, e.g., local regulations, political alliances, and investments in research, development, and education.

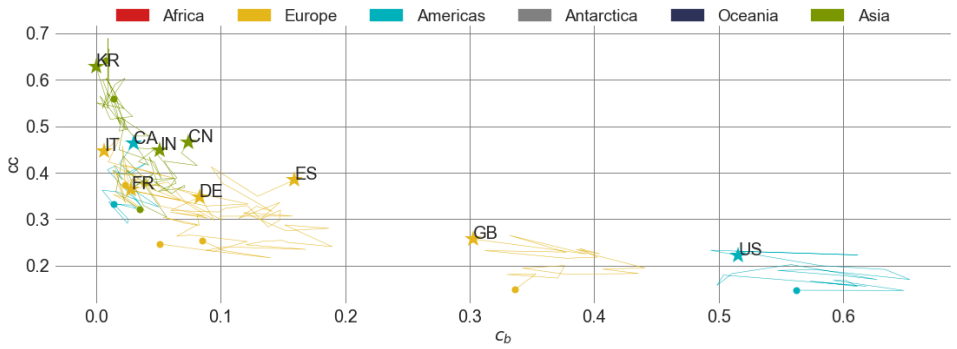
At the same time, the evolution of hubs and authorities' scores over time, alongside their relative discrepancy, and other network measures, suggests that local policies can buck the trend, as testified by the Gini coefficient. Gini coefficient decreases as h and a decrease, as Figures 4.22 and 4.23 attest. Complexity in terms of migration patterns seems to co-exist in the best positions of the hub and the authority rankings, in analogy with the economic framework, so that successful countries are extremely diversified in product export [130], [81]. Another interesting finding lies in the return pattern, and how the rankings 4.9 does not necessarily follow the same evolution as any centrality ranking in terms of principal actors involved. That happens for the Normalized Return Index 4.5, and also the evolution within the ranking according to the Normalized Similarity is different, lower, as attested by Figure 4.30. Many layers of factors may be involved in this dynamics, or maybe the same one but with different priorities, cultural and

social aspects may impact decisions more strongly the work-related opportunities, and more analysis is needed to make proper assumptions about this aspect of the phenomenon.

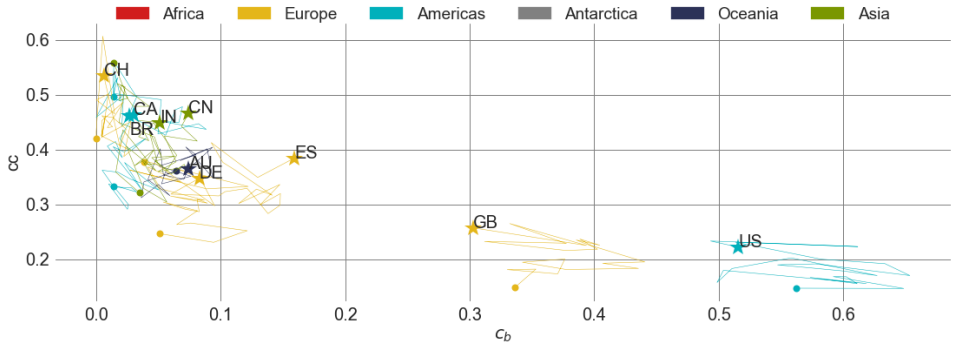
After all, ranking by means of hubs and authorities scores is insightful, but just a preliminary step toward a more refined analysis. In future work, we plan to expand the study carried toward different directions, for example, analyzing the geographic impact of migration routes, and testing the measures we have with classical models used for migration analysis, such as the gravity model. On other hand, we plan to tackle the correlation between hub and authority scores with respect to exogenous metrics, extrapolate from different data sources then ORCID to evaluate for example the research/academic success or economic indicators, as in [131]; even though not very high due to the presence of countries of high GDP showing poor performances in terms of hub or authority ranking, as also discussed in [132], where the relationship between science and investments shows complex behaviors. Furthermore, we plan to restrict the analysis to a specific geographical region (e.g., Europe) to study migrations at smaller granularity (e.g., cities), according to specific scientific fields in order to understand where skills actually move, and by different career stages, education or employment, it would be interesting to frame how the first step into a proper scientific career, just after the educational training, relate with international science exchange. Different migration routes can be exploited to better understand how the very topology of the networks changes with regard to the different paths the researchers decided to take. Moreover, co-affiliations, represented by the combination of routes 4 and 8 in Figure 4.7, may reveal unexpected patterns in worldwide collaborations. Also, the analysis of return can dive deeper into finding correlations with other factors, political, historical, or cultural.

Finally, we need to point out some issues that arise from working with real data. Every year ORCID publishes an open-access dataset that captures all the public records, so every year it should be possible to update the picture of the scientific migration. However, ORCID is a user-based platform, and we can not control how each user chooses to update their information. We present a methodology applicable to evolving datasets that grow over time, aiming to deliver a more precise picture as the information increases. Moreover, it is important to mention that the main part of the proposed methodology, after the data processing, is completely data-agnostic, meaning that it can be applied to other biblio-metric datasets obtained from different sources. In particular, testing the intrinsic differences with data extrapolated in scientific publications, where the component of successes could mask the beginning year in a researcher's career.

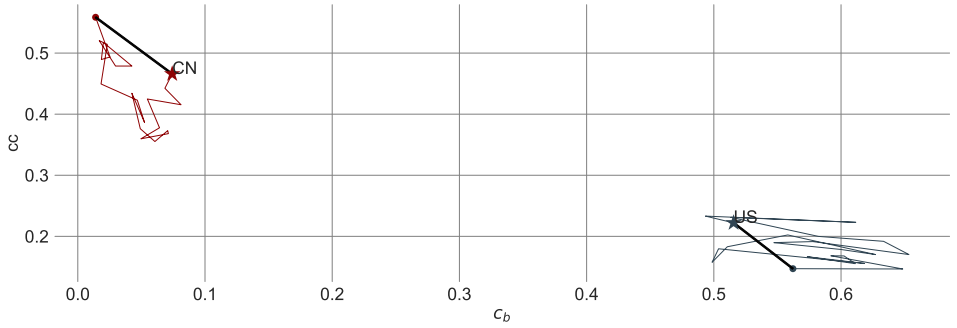
In conclusion, we want to emphasize how there are many possible new directions to tackle the analysis of ever-evolving global and local scenarios of scientific migration.



(a) Top 10 countries in Authority Ranking



(b) Top 10 countries in Hub Ranking



(c) United States and China

Figure 4.24: Betweenness centrality (c_b)-clustering coefficient (cc) trajectories from 2000 to 2022 of countries occupying the top ten position of Authority Ranking (a) and Hub Ranking (b) in 2021. (c) depicts the trajectories of China and the United States from 2000 to 2022. 2000 is depicted with a round shape, and 2021 with a star shape. ISO 3166-1 alpha-2 codes are reported for the selected countries: Australia (AU), China (CN), Germany (DE), India (IN), Italy (IT), Spain (ES), United Kingdom (GB), United States (US), Canada (CA), South Korea (KR), France (FR), Brazil (BR), and Switzerland (CH).

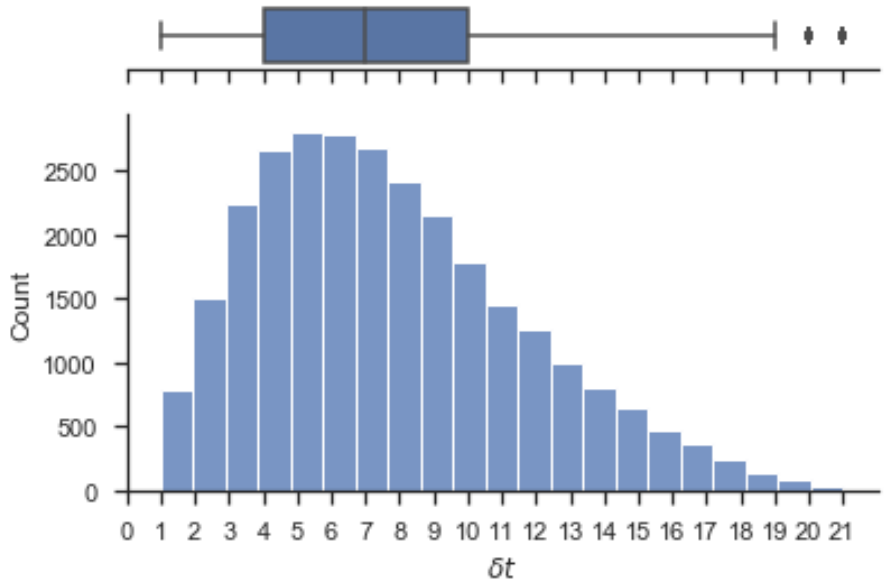
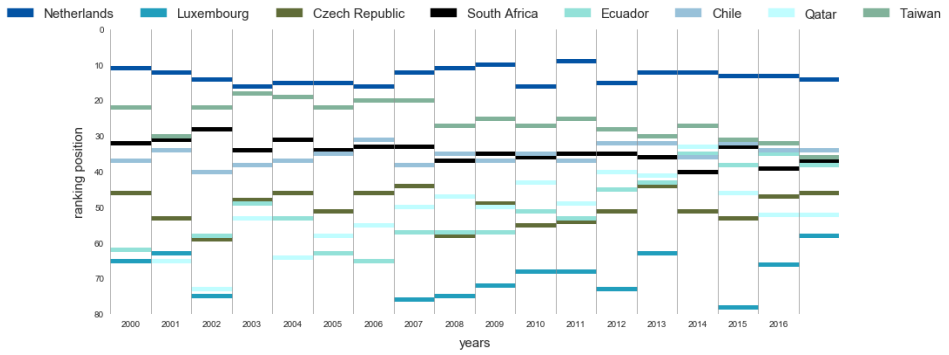
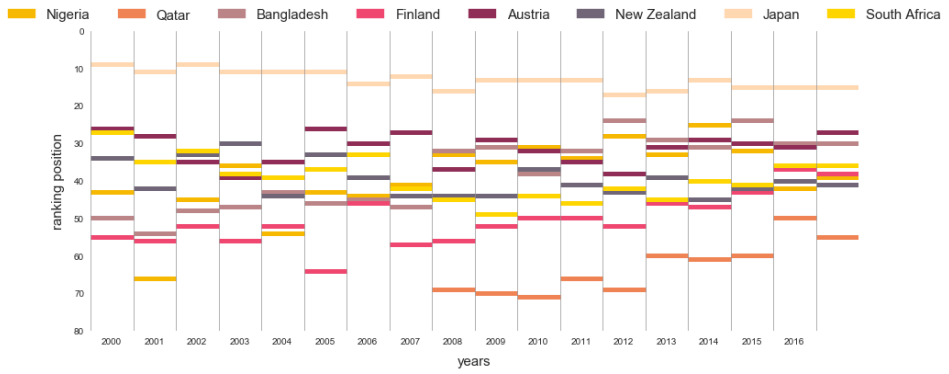


Figure 4.25: Distribution of returns by time difference with respect to the starting year, δ_t.



(a) Authority Ranking



(b) Hub Ranking

Figure 4.26: Ranking according to increase or decrease of position in the time span 2000-2021 for authorities (a) and hubs (b), computing by mean of the $\Delta_i(R)$ among the countries that in the last year of the time domain (2021) have reached at least the 50^{th} position.

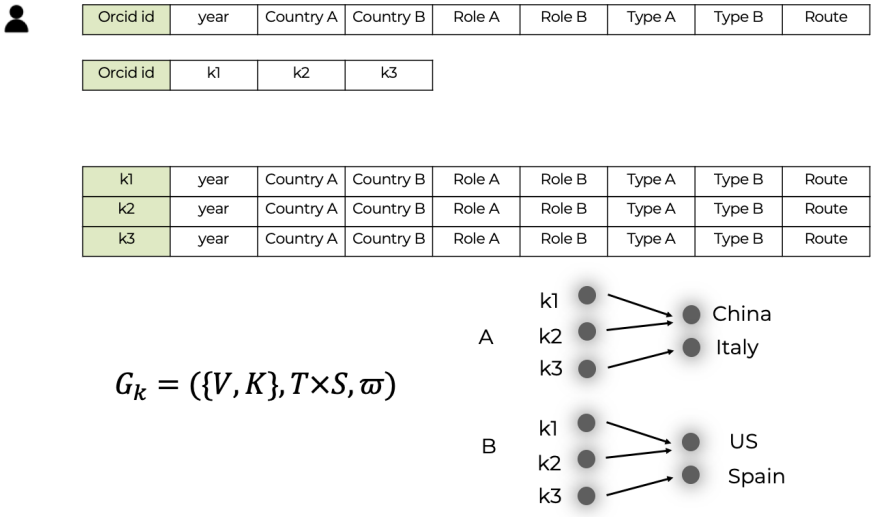


Figure 4.27: Building process to obtain $G_k(V, K, t, S)$, where V is the set of countries, K the set of keywords the moves in year t, and S represents two sets of status, A and B, that encode the before and after all the migration happening in t.

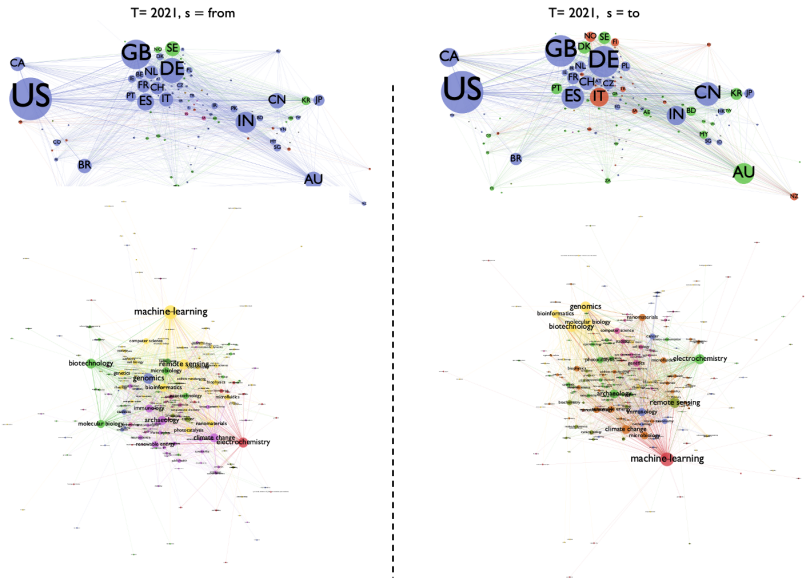
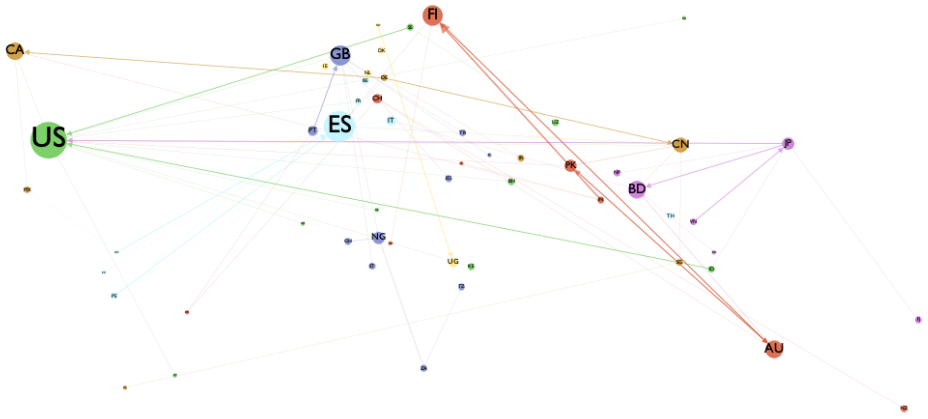
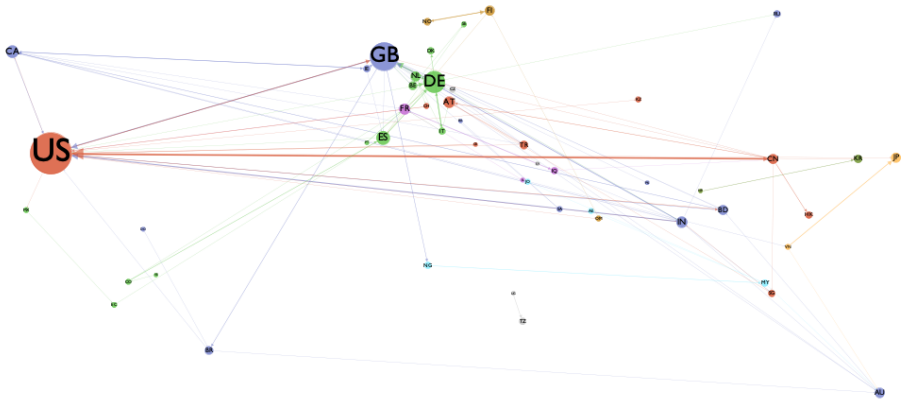


Figure 4.28: Projections of $G_k(\{V, K\}, 2021, S)$, over the set of countries (top) and over the set of keywords (bottom), for the status before $s = from$ and after the migration $s = to$. Node dimension scales over the strength, while colors follow the belonging to the same community, extrapolating by means of modularity.



(a) Climate change($|N| = 55$, $|E| = 80$)



(b) Machine learning network ($|N| = 51$, $|E| = 105$)

51

Figure 4.29: Migrations network of the keywords “machine learning” and “climate change” in 2015.

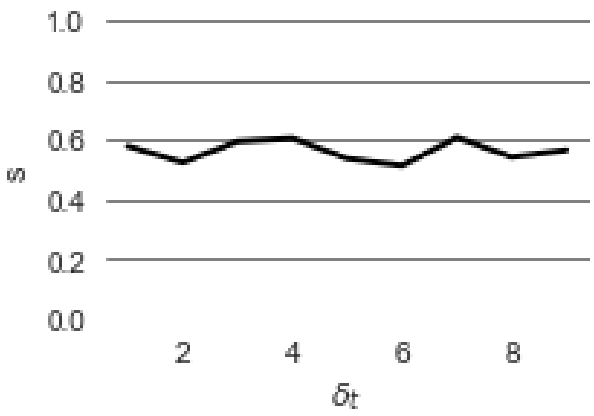


Figure 4.30: s estimates the similarity between the rankings in two successive years between the top twenties positions of the return index $r_i(\hat{\delta}_t)$ defined in Eq. 4.5.

Chapter 5

Online debate

5.1 Overview

This Chapter collects work begun during a visiting period abroad at Northeastern Network Institute, as a fellow of the program Accelnet-Multinet [1], under the supervision of Alessandro Vespignani and Matteo Chinazzi, with the collaboration of Marco Ajelli of Indiana University. The principal scope of this work is to try to untangle what happened in 2020 in the debate around Covid-19. We modeled a framework, combining different tools, from network science to natural language processing (NLP, see Section 3.2), to extrapolate opinion standpoints around specific topics shared among a specific group of participants. To test the models we need data. The multiplication of media and the personalizing of media use in the last years has fostered profound changes in the ways in which entertainment content is enjoyed and the information is accessed, disrupting traditional hierarchy, and creating diffusion paths worth investigating. In particular, the spread of personal media has spurred a process of identity construction, both individual and collective. That is, they influence the acquisition of a sense of belonging to one's community, the formation of political beliefs, and even the generation of expectations for the future, thanks to the ability of personal media and social networks to create virtual worlds. How these new worlds relate to the surrounding reality is still very difficult to assess. One reason is that online systems are socio-technical systems, the social component is embedded in a technical schema, and most of the time we do not know the rules exactly, examples are recommendation algorithms through which social media like Facebook or Twitter suggest content and possible new links among users. Clearly distinguishing one from the other should be if not a research goal a necessary premise for any analysis. In the work presented in the following Chapter, we collect data from the online world, aware that we are looking at a specific projection of reality but that it can still give us interesting clues about human behavior and how the debate surrounding such an overwhelm-

[1] https://www.accelnet-multinet.org/

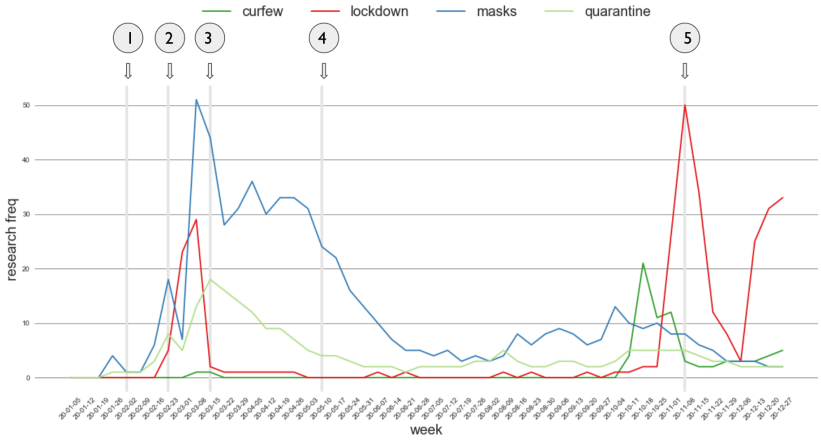


Figure 5.1: Comparison between Google Trends of different research keywords: curfew, lockdown, masks, quarantine. Google queries data was available on a weekly basis only. The spikes in the data are due to the following events: (1) the first two cases of COVID-19 detected on a couple of Chinese tourists (20-01-30), (2) the first official case of secondary transmission occurred in Codogno (20-02-18), (3) national lockdown (20-03-09), (4) national lockdown ends (20-05-03), (5) national curfew (20-11-06).

ing aspect has evolved over time.

5.2 Research questions

The Covid-19 pandemic disrupts the lives of everyone starting from the beginning of 2020. It has changed profoundly social interactions for the time being with consequences in the future years still unknown. To curb the spread of the virus many governments around the world have implemented non-pharmaceutical interventions (NPI) like lockdowns, mandatory use of masks, quarantine, curfews, travel restrictions, and so on. Of course, public opinion has absorbed these interventions in different ways, generating a huge debate over the last year and a half.

Figure 5.1 depicts the Google research frequencies of some strategic keywords in Italy. It shows a partial picture of how a country heavily subjected to the virus Italy has reacted. In particular, different phases emerge, characterized by different spikes, near important decision key points. The picture is partial since it lacks many layers of information, it can only suggest trends of interest, which, however, become important drivers for refining the research questions we propose to investigate.

- How has the public debate over non-pharmaceutical intervention in regard to COVID-19 been unrolled in 2020?

- Who have been the main protagonists of the debate and how their online activities have shaped the final map of stances regarding a specific topic of the discussion?
- How much has the debate polarized over time? In particular, our overall hypothesis is that at the beginning of the crisis, there was a more cohesive standpoint, "everything will be alright" was a major slogan utilized by almost everyone, but as time passed it seems that more clusters emerged, and the debate polarized.

Answering these questions could help on one hand the community to better understand which have been the weakest point in the interventions campaign, which could have limited the government policy effectiveness, but also, on the other hand, to build a framework to study from the very beginning the dynamics of a debate which has impacted so deeply the life of many. The debate around the NPI is a complex phenomenon, characterized by many aspects which interact in non-linear ways. We plan to exploit the said research questions employing a network science framework, in particular, every aspect that interacts in the debate could be represented by means of a layer in a multilayer network. As a data source, we choose Twitter, a well-known social network, used by a rich audience, from politicians to scientists, from influencer users to the general public. In particular, we will employ Academic Twitter API to retrieve tweets. The terminology used throughout the whole Chapter is summarized in Table 5.1.

The opinion map for us will unfold through users' activity in the online debate. Nodes will represent users while edges will be interactions. We adopt a 2-layer structure, as shown in Figure 5.2, each layer standing for a different type of user interplay:

- *Engagement Layer*: this layer encodes the way users engage each other in the conversation. It reflects both the technical possibilities of interactions on a certain social platform and the dynamics that arise from them. The result is the emergence of different phenomena: the establishment over time of a certain set of conventions, that is the meaning we generally associate with some type of action, and the unrolling of these over different topics of discourse [133]. This layer is built through "retweeting". Even if retweeting can simply be seen as the act of copying and rebroadcasting a message of another user, it is a powerful way to build connections on Twitter, and promote ideas, and stances on a certain matter contributing to the shaping of a rich and diversified conversational context, in which many voices find a position and a role in the overall network [134].
- *Text Similarity Layer*: engagement alone gives us an outline of the topology of the discussion, but can limit the understanding of the actual message that spreads in different branches of the threads, more

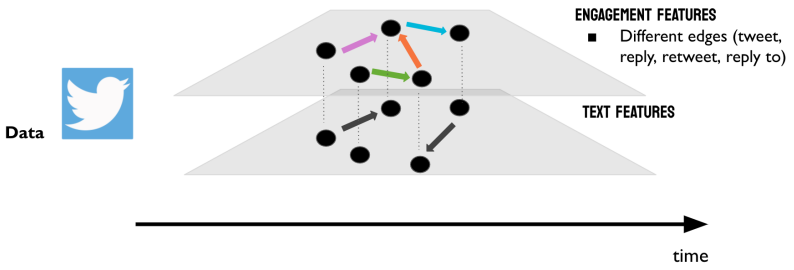


Figure 5.2: Structure of the 2-layer network employed to study the online Twitter debate, around COVID-19 and NPI.

so when the topic is characterized by many points of view. This layer will try to frame the similarity between the original content of each user. Instead of sharing others' messages, "tweets" encode the original message posted by a user. We will extrapolate some relevant features by means of Natural Language Process frameworks (see Section 3.2).

term	Meaning
Tweet	Original content posted by a user. A tweet can also include a mention to other users, or an URL pointing to an external website.
Mention	The act of mentioning another user, by citing their nickname after the '@' symbol.
Retweet (RT)	Another user's post, shared by the author of the retweet.
Quote	A retweet with additional comment by the author of the retweet.
Reply	The act of replying to another user by clicking on "*Reply*" under the original tweet. The addressed user is mentioned with the '@' symbol, at the beginning of the reply message.

Table 5.1: Terminology used throughout the chapter.

5.3 Engagement Layer

5.3.1 Experiments and pipeline

According to 17° Communication Report [135] drafted by the Research Foundation Censis[2] Italians love politics, they follow the debate between the parties, and they get passionate about the sides, from strongly asserting their opinion to insulting each other on social networks. The pandemic, with its contribution to fears, has increased the need for information, with a specific focus on scientific, medical, and technological news. Albeit with a decline

[2]https://www.censis.it/

(-2.7% between 2019 and 2021), the most seeking topic remains national politics (for 39.7% of Italians), but the desire to delve deeper into Covid-19 information was reflected in the growing interest in science and medicine, which rose from the preferences of 27.7% of the population in 2019 to 33.4% in 2021 (+5.7). In this new scenario, the presence in the media of experts in various fields of medicine, like epidemiologists and virologists, has multiplied in the last two years. For more than half of the Italians (54.2%), they were indispensable in order to have guidance on the correct behavior to adopt (15.5%) or because they were useful in understanding what was happening (38.7%). On the other hand, the ratings are negative for 45.8%: because they also contributed to creating confusion and disorientation (34.4%) or were even harmful because they caused alarm (11.4%). To tackle this complex landscape we need data. According to Audioweb, a super "partes" body that collects and distributes internet audience data in Italy, $10,8$ million people used Twitter on average in 2020. Even though it does not reach the numbers of Facebook it still has a good penetration rate, with almost 4 million users that, according to Twitter, could be reached with adverts in January 2020 [136]. Also in 2020, Twitter released the API for academic research API [3]. The endpoint allows access to the full history of public Tweets and delivers data in a request-response model with pagination, supporting query filtering by language, keyword, type of posts (either reply or retweet for example), and much more within a monthly cup of ten million posts. We used this endpoint to collect posts with the aim of understanding longitudinal trends linked to evolving topics of interest. In particular, we build a pipeline to collect posts in a meaningful way to build engagement layer networks. Mostly by means of a trade-off between snowball sampling among the most central nodes that emerges in the networks and bounding box techniques according to some keywords.

The pipeline is described by the Procedure 1: for each month in 2020 starting from a group of selected users we collect their retweets to build the networks, $G_E = (V, M, \varpi)$, where V is a set of nodes, $M = [1, 2, \ldots, 12]$ is the time domain made by each month of 2020, and $\varpi : V \times V \times M \to \mathbb{N}$ is a function defining for each pair of nodes $i, j \in V$ and each timestamp $t \in M$ the weight of edge (i, j) at time t. We are interested in the topology of the network, so for the rest of the analysis, we will work mainly with the degree and the connection map instead of strength and flows. Then we compute the authority and hub ranking through the HITS algorithm (see Section 3.1.3), update the list of relevant users and collect their retweets to update the network. We stop when could not find new relevant users within a wanted threshold in the HITS rankings, where $T = 100$. At each step, we filter both the posts by a chosen set of keywords and users, keeping out newspaper accounts, by filtering them according to the description camp.

[3] https://developer.twitter.com/en/products/twitter-api/academic-research

Retaining a news account would shift the conversation toward a broader topic of discussion. In the end, for the final networks, we employ the Stochastic Block Model framework (see Chapter 3) to extrapolate the best network partition.

procedure PIPELINE($\mathbf{U}, q, t, s = 0$) ▷ Example: [1] in Table 5.2
  $\mathbf{N} \leftarrow group[1]$
  while N do
    $RT \leftarrow collect$ ▷ From Twitter API
    $G_{RT}(t, \mathbf{U}, s) \leftarrow network(RT)$
    $H \leftarrow hits(G_{RT}(t, \mathbf{U}, s)$
    $\mathbf{N} \leftarrow set(H[0:T]) - set(\mathbf{U})$ ▷ If n in **N** not already in **U**
    $\mathbf{U} \leftarrow \mathbf{U} + \mathbf{N}$
    $s \leftarrow s + 1$
    PIPELINE(**N**,q,t,s)
  end while
  return $G_{RT}(t, \mathbf{U})$ ▷ The final network
end procedure

Algorithm 1: The algorithm describes the pipeline that builds the engagement layer networks. For example, if we choose the experiment [1] in Table 5.2, we would have as input: **U** as the influencers group, q as the query construct, retweet of all language containing all the COVID-19 keyword, and a given month t in 2020.

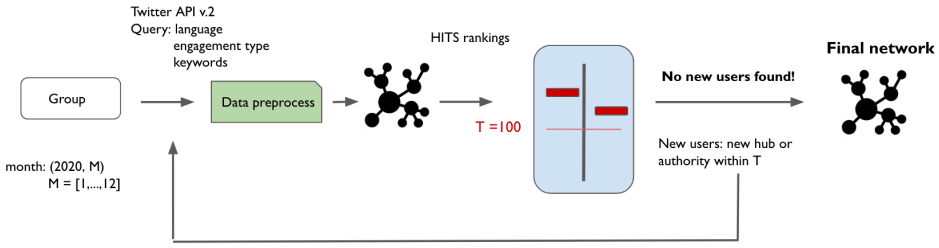


Figure 5.3: Engagement layer network pipeline.

5.3.2 Discussion

Table 5.2 there are summarized all the experiments carried out. We selected two different starting groups:

- Influencers: people that we know were active in the debate around Covid-19 in 2020. We include epidemiologists, virologists, doctors, international scientists, international scientific organizations, journalists, and so on.

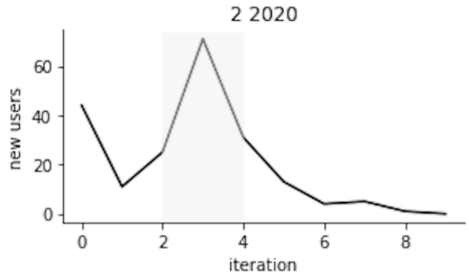
- Politics: all the official accounts of the Parliament members in office

Experiments	query
influencers	[1] all language + retweet + COVID-19
$n = 44$	[2] all language + retweet + lockdown
	[3] all language + retweet + mask
politics (A)	[4] all language + retweet + COVID-19
$n = 862$	[5] Italian + retweet + COVID-19
	[6] Italian + retweet + lockdown
Topic	keywords
COVID-19	covid, virus, mask, lockdown, sars, covid19, coronavirus, red zone, pandemic, epidemic, death, contagion, hospital, self-certification (B), doctor, quarantine
lockdown	lockdown, red zone, yellow zone, white zone (C)
mask	mask,nomask

Table 5.2: Summary of all the experiments done regarding the engagement layer network [1-6], and the specific keywords chosen for the API requests. (A) The politics group includes all the accounts of the members of the government that were in office during 2020, the parties' official accounts, and the official accounts of the Chambers. (B) self-certification was a mandatory document to travel in different Italian areas. (C) As of fall 2020, a class system was implemented for each region of Italy depending on the level of virus spread, the three areas were distinguished by severity and consequent limitations in permitted activities, from white to red, of increasing severity.

during 2020, plus the official parties account and the official accounts of the Chambers, which are the two groups that made the bi-partisan structure of the Italian government.

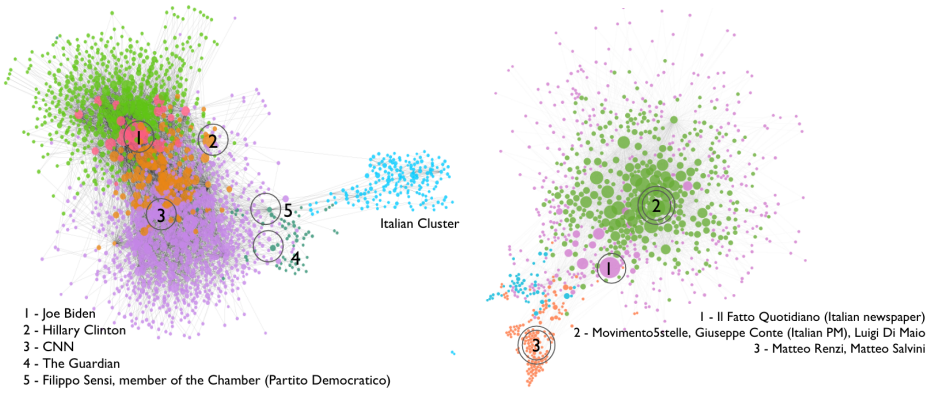
We cover three topics of debate: a general one around Covid-19 and two more focus on specific non-pharmaceutical interventions, lockdowns, and masks. To confine the debate around the lockdown and musk, we used the keyword in Table 5.2 for the requests toward the Twitter API. The first question we will address is about convergence. *Does the described procedure Converge?* We have an empirical answer, and much more theoretical that we plan to investigate in the future. Empirically the convergence holds, with some explainable exceptions. In Appendix B.1 the Figure B.3 show the full-year convergences for every engagement layer network for experiment [1] (influencers, retweets, Covid-19 keyword). In max 8 steps (September 2020) there were no new users that feed the pipeline, and the procedure converges. However, in some specific runs, anomalies emerge. Figure B.3 depicts the convergence for February 2020 in the experiment [3] (influencers, all language, retweets, mask), the curve spikes at step 3, and decreases after, if we look at the new user gathered, we notice mostly Chinese accounts, we enter a



H

Figure 5.4: Convergence steps in new user gathering of the Engagement Layer Network pipeline describe in Algorithm 11, given the experiment [3] from Table 5.2, for February 2020.

different "chamber" of the debate, which it is coherent with the unfolding of the pandemic event, since in February most cases were still detected mostly in China, and masks were still not an issue of debate in Italy. A similar result happens with the politics group if we do not query posts according to the language, we can see in Figure 5.5 the final networks obtained with and without a selection over the language for experiment [5] for May 2020. Figure 5.5a clearly displays that during the procedure the Twitter posts collection moved toward the United States debate, having as principal drivers of the conversation Joe Biden[4], Hillary Clinton, and the CNN account.



(a) No language selection. (b) Italian language.

Figure 5.5: Final Engagement Layer Networks in the experiment [5] in Table 5.2 for the month of May, obtained without any constraint over the language (a) and selected only retweets in Italian.

The convergence of the procedure coincides with the engagement networks. In Figure B.2 we have a complete view of all the in- and out-degree

[4]The account mentioned is the one prior to his appointment as president on January 20, 2021, the handle is *@joebiden*.

distribution for all the engagement networks in the experiment [3]. In-degree progressively moves toward a more scale-free distribution, following the preferential attachment mechanism of creating links, and more popular users become hubs in the debate. But to get a more in-depth view of the evolution, we try to partition the network into different areas, or "communities". An interesting result emerges when comparing March 2020 and May 2020.

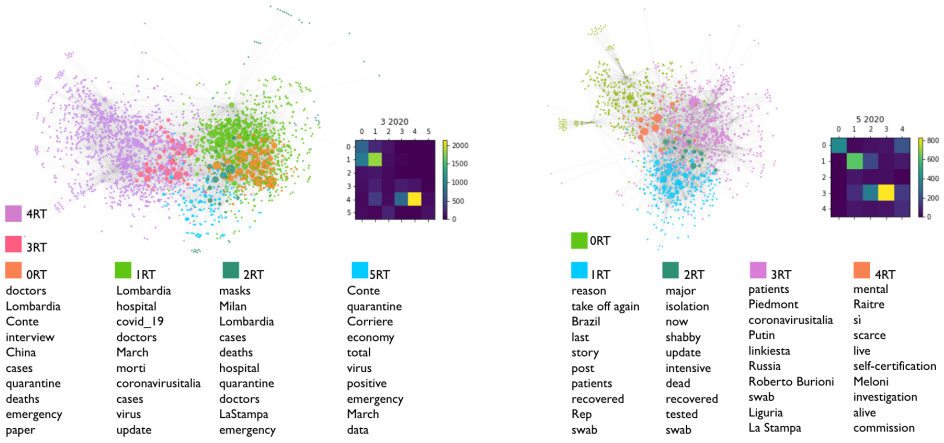


Figure 5.6: Engagement network and blocks in the experiment [1] in Table 5.2, March 2020 on the left, May on the right. For the selected blocks, we extrapolate the most relevant words. For each month the matrices of edge counts between groups are displayed. The dimension of the scale of the node follows over the in-degree.

Through the Procedure 1 we build the networks, and for reasonable analysis, we filter out nodes with an in-degree less than 2. The final networks have the following characteristics:

- March: directed, $|N| = 2221$ (27% of the original number of nodes), $|E| = 8612$
- May: directed, $|N| = 1053$ (25% of the original number of nodes), $|E| = 3381$

In the following discussion, we refer to Figure 5.6. In March 2020 the network seems to be divided into two main parts: blocks 3 and 4 relate to international debate, having among them, for example, the World Health Organization account and the New York Times account, while the other blocks cover the Italian debate. Since we can track the initial users, we can describe block 0 as composed of economists, block 1 and 2 by epidemiologists and virologists, and block 5 by news account. The matrices of edge counts between groups attest many relationships between block 1 toward block 0, a reason could be that among the economist block we can find also

a virologist that was an important driver of the conversation, and so very retweeted, virologist Roberto Burioni is very active on Twitter. Blocks 1 and 3 were the more connected. For each block we build a corpus of texts, assigning to each block the concatenated collection of all the texts the users in that block have retweeted. Then, for each block we have exploited the *Term frequency-inverse document frequency* formula to extrapolate the most relevant words [137]:

$$td - idf_{i,j} = tf_{i,j} \times idf_i \tag{5.1}$$

where,

$$tf_{i,j} = \frac{n_{i,j}}{|d_j|} \tag{5.2}$$

is the number of occurrences of the term i in the document j divided by the dimensions of the document j and

$$idf_i = \log_{10} \frac{|D|}{\{d : i \in d\}} \tag{5.3}$$

with D as the number of documents, which for us is the number of texts (one for each block), divided by the number of documents with the term i.

The most relevant words that emerge in the different blocks refer in a similar way to the main topic of the pandemic, they do not suggest different opinions or values. In May we still found an international module, the block labeled 0, and 4 blocks that cover the Italian online conversation. But, unlike in March, there seems to be not so much interaction among groups, or at least more separate communities. Block 1 retains the majority of the economists, block 3 the majority of the scientific users, while block 4 is a mix of economists and epidemiologists. Interesting are some words that emerge as most relevant in their respective collection of retweets: "take off again" was a phrase used to convey the necessity of returning to a normal life, as before the pandemic. Of course, these conclusions have many limits, structural since retweets can be very repetitive inside a block and not directly authored by who shared them, as well as lexical since words alone do not embrace the context of a more complex stream of exchange. We try to tackle these issues in the next section, diving deeper into the textual analysis.

5.4 Text Similarity Layer

5.4.1 Micro-frame analysis

Every debate, question, or aspect of a discussion could have many standpoints, more so if the argument impacts deeply the life of many, even if it is carried out on a social platform where entertainment keep remaining part of the picture, such as Twitter. Our goal is to pinpoint these standpoints and analyze how they change over time. Framing is the process of emphasizing

one aspect of a question and making the reader or listeners take a different position on an issue, without even putting forward pre-conceived arguments. To investigate the *Text Similarity Layer* we will adopt this idea, in particular, we will adopt the "FrameAxis" method [138], an unsupervised approach that can be applied to large datasets because it does not require manual annotations. It is designed to quantitatively configure how semantic axes position in the text. The semantic axes model opposite faces of an aspect in a word vector space [139]. Given a pair of antonyms w^+,w^-, for example, *open-closed*, the *Semantic Axis Vector* is $v_f = v_w^- - v_w^+$, where f is a the micro-frame, and v_w^- and v_w^+ are the corresponding word vectors. For each ax, we want to capture how biased the text is on a certain micro-frame, and how actively a certain micro-frame is used. Micro-frame bias and intensity computation are based on the contribution of each word to a micro-frame. Formally, we define the contribution of a word w to a micro-frame f as the cosine similarity between the word vector v_w and the micro-frame vector v_f:

$$c_f^w = \frac{v_w \cdot v_f}{\|v_w\|\|v_f\|} \tag{5.4}$$

In FrameAxis, a corpus is represented as a bag of words, and each word is considered an attribute of the corpus, so each word contributes to the micro-frame, and its frequency can be considered as its salience. Accordingly, the weighted average of the word's contribution to the micro-frame f for all the words in the text maps the individual's attitude toward the ax of antonyms, we called this weighted average **Micro-frame Bias**:

$$B_f^t = \frac{\sum_{w \in t} n_w c_f^w}{\sum_{w \in t} n_w} \tag{5.5}$$

where n_w is the frequency of w in the text t. **Micro-frame intensity** quantifies how strongly a given micro-frame is used in the document. Namely, given corpus t and a micro-frame f, it measures the second moment of the contributions c_f^w for all the words in t. Formally, it is calculated as follows:

$$I_f^t = \frac{\sum_{w \in t} n_w (c_f^w - B_f^T)^2}{\sum_{w \in t} n_w} \tag{5.6}$$

where B_f^T is the baseline micro-frame bias of the entire text corpus T. For instance, if a given text is emotionally charged with many words that strongly express either happiness or sadness, we can say that the happy–sad micro frame is heavily used in the document regardless of the bias, which can lean toward a specific pole or neither, being the whole text considered happier, sadder, or balanced regarding these standpoints. We can see the interplay between Micro-frame Bias and Intensity in Figure 5.7.

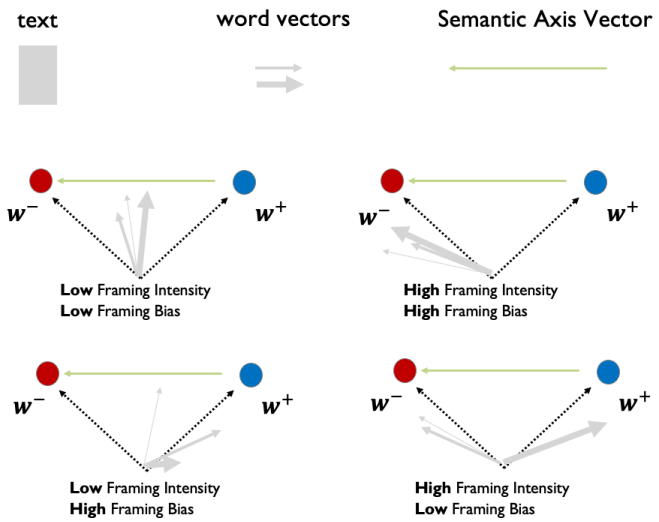


Figure 5.7: Illustrations of micro-frame intensity and bias. Red and blue circles represent two pole word vectors, which define the semantic axis vector, and gray arrows represent the vector of words that appeared in a given corpus. The width of the arrows indicates the weight (i.e., frequency of appearances) of the corresponding words. The figure shows when micro-frame intensity and bias can be high or low.

5.4.2 Networks building

The first method we exploited to build the text similarity layer was to define a complete network $G*_T$ for each month m and measure for each pair of users the cosine similarity between the collection of their original tweets, assigning it as the edge weights. To proceed with further analysis, we attempt to filter out edges according to the weight distribution. Despite many endeavors in this direction, we realize that texts were indeed quite similar. Figure 5.8 shows the edge weight $w(e_p)$ linked to a specific percentile in the weight distribution, for $G *_T (m = 10)$ of in experiment [6] in Table 5.2. We notice that we rapidly reach a 0.8 value of cosine similarity, making the decision of a cutting threshold for the weight very hard. The micro-frame approach gives

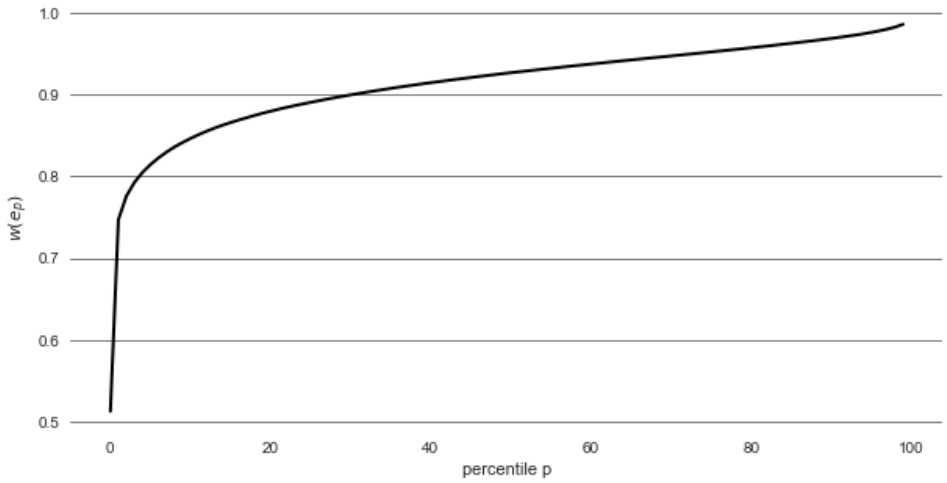


Figure 5.8: Edge weight $w(e_p)$ linked to a specific percentile in the weight distribution, for $G *_T (m = 10)$ of in experiment [6] in Table 5.2.

us the tools to extrapolate from the text its standpoint regarding opposite point-of-views of multiple aspects. However, we need to model an interacting point-of-views. Users engage in an online debate increasing their bias and intensity over a certain topic. We tried to exploit a similar micro-frame approach to build networks. We characterize each user in the engagement networks with the concatenation of all their original post in a given month of 2020. We choose an aspect and consider its vector, v_a. Then we define for each pair of users A, B a Difference Axis:

$$d^{A,B} = v^A_{text} - v^B_{text} \tag{5.7}$$

where v^A_{text} is the vector representation of the tweet production of user A in month m and measure the cosine similarity between $d^{A,B}$ and v_a:

$$c^{A,B}_a = \frac{d^{A,B} \cdot v_a}{\|d^{A,B}\|\|v_a\|} \tag{5.8}$$

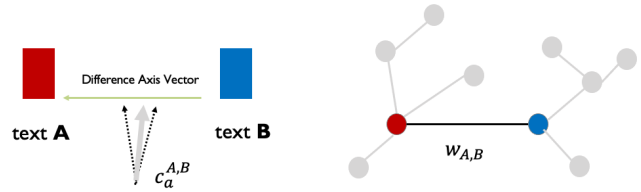
The resulting network is defined as $G_T = (V_{text}, M, \varrho)$, where V_{text} is a set of nodes, representing the text production of each user, $M = [1, 2, \ldots, 12]$ is the time domain made by each month of 2020, and $\varrho : V_{text} \times V_{text} \times M \to \mathbb{R}$ is a function defining for each pair of node-text $i_{text}, j_{text} \in V$ and each timestamp $m \in M$ the weight of edge (i_{text}, j_{text}) at month m, which consist of the absolute value of the cosine similarity between i_{text} and j_{text}. For each m, $G_T(m)$ is still complete, undirected, and weighted. To solve the completeness issue, we have tried to solve the questions of the real meaning of an edge between two nodes, what does a relationship mean in this context? More so, with the aim to find community afterward, what does it stand for as a subpart of the network very connected? Since our research question aims to find the stance of a community of users regarding an aspect of a certain topic (*With respect to the lockdown, how do the communities in the online debate position themselves to the idea of openness?*), a relationship is supposed to mean the two users have the same "position", which means the same distance. As shown in Figure 5.9a a link exists between A and B if $c_a^{A,B}$ is close to 0. The last model we conceptualize pursues the goal to break the cosine similarity symmetry which leads to the indirectness of the network. We propose a network model defined as G_T but where edges computed with ϱ maintain the positive or negative sign that is taken up by the direction of a link. As shown in Figure 5.9b a negative sign in the cosine similarity will change the direction from node A to node C in $C \to A$. This model will be used to extrapolate a nodes ranking based on a *Aspect Polarization Index*, defined as:

$$\rho_a(i, m) = \frac{s_{i,m}^{in} - s_{i,m}^{out}}{s_{i,m}^{in} + s_{i,m}^{out}}, \tag{5.9}$$

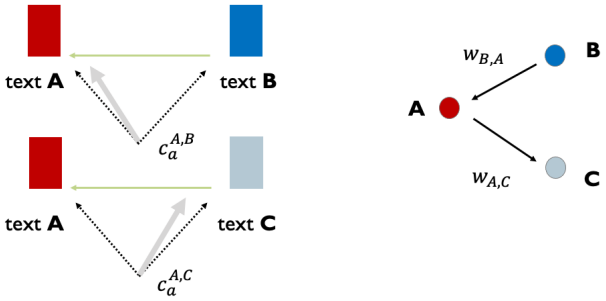
A node with a high value of ρ will have a stance close to the positive pole of the aspect, on the other hand, a node with a low value of ρ will have a stance close to the negative pole of the aspect.

5.4.3 Discussion

In this Section, we will comment on the analysis processed around the framework and model previously described. The final goal is of course ambitious, and the tools limiting due to the difficulty of the task. A conversation is indeed a very complex system with many layers, and drivers, that evolve over time in ways that can follow many dynamics and a good portion of randomness. So we need boundaries, at least to guide the experiments to translate the main research question - *How do online participants in a debate converse about a certain topic?* - in measurable quantities. The first thing that we did was to simplify this question in *How do a specific group of online participants in a debate converse about Covid-19 in March 2020?* Of course a topic, in this case, Covid-19, embrace more aspects and each aspect has more point of view. We try to slit and conquer this multitude projecting the



(a) Similarity Network.



(b) Contrasting network.

Figure 5.9: Text-similarity network layers generated by micro-frame approach.

topic over a system of aspects, each of them treated with the micro-frame approach.

Starting from the Engagement Networks we collect for each month m the tweets of the users that appear in the final network, the ones that emerge from the Procedure 1. In all the following analyses we vectorize the text using the language model describer in Section 3.2. Given a topic, to capture nuanced framing, it is crucial to cover a variety of antonym pairs, following the work by Jing et. al. [140], we tested the following:

- Open-Close
- Life-Death
- Individual-Group
- Freedom-Imprisonment
- Hope-Despair
- Easy-Difficult
- Fast-Slow
- Exaggerated-Reduced

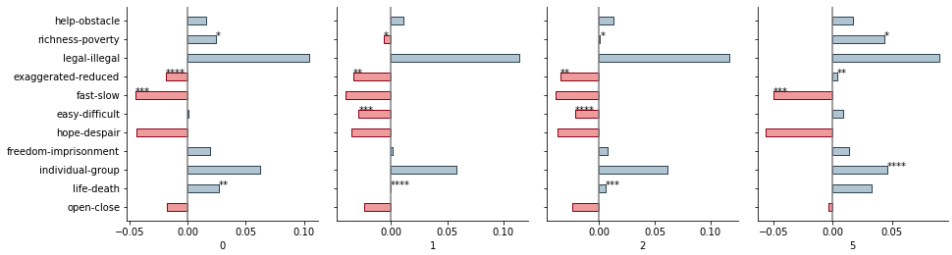
- Legal-Illegal
- Richness-Poverty
- Help-Obstacle

In particular, we are interested in discovering the shifting of the frame analysis happening between March and May 2020, describe in Section 5.3.2. Figure 5.10 depicts the micro-frame bias for the aspects tested, stars over the bars identify the four aspects that contribute more in each group. In March (Figure 5.10a) we could recognize block 0 mainly with economists and block 1 mainly with scientists, epidemiologists, and virologists. For both blocks 0 and 1 in March, the most intense aspect lies in the semantic axis "richness-poverty" but with opposite bias, for the two groups: block 0 is directed toward "richness" while block 1 is toward "poverty". With respect to these groups, the other aspects follow the same direction schema with different intensities, for example, the antonym "life-death" contributes more to the discussion in block 0 while the antonym "exaggerate-reduced" prevails in block 1. On the engagement layer in May the group of the economists (0) slit between 1 and 4, with 4 being a mixed group, with both economists and scientists, while group 3 remains predominantly made of scientists. In the micro-frame analysis what we notice is that the pair "fast-slow" seems to be the aspect more present in the text with a bias toward the pole "fast" for all the groups, followed by "richness-poverty" for groups 1,2, and 4 at least. The third aspect changes a bit across the different groups: for 1 "life", for 3 "exaggerated" and for 4 "difficult".

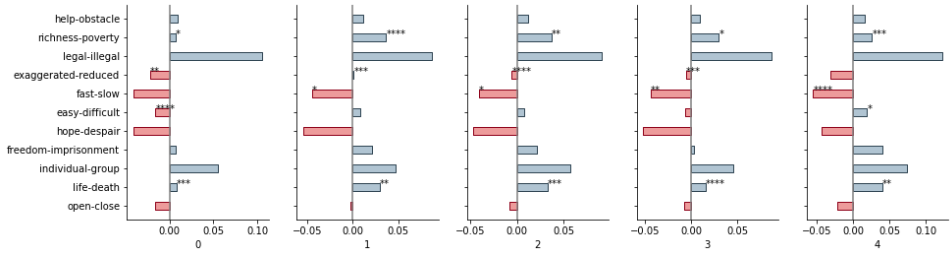
This experiment collects the online activity of a very heterogeneous bunch of individuals, and although this enriches the analysis it also introduces elements of incoherence: we have many languages for example, which means the conversation is not strictly geolocalized in Italy, and so the topic and the aspects may not be comparable in a straightforward way. So we try to transform the original research question into the more specific: *How do a specific group of online participants in a debate converse about lockdown/masks in October 2020 in Italy?* We consider a more coherent group of people, all the politicians that were in office during 2020, and a more limited topic, lockdown. The engagement network regarding experiment [6] in Table 5.2, has the following characteristics (in this case we did not operate any filter on the in-degree distribution):

- October: directed, $|N| = 670$, $|E| = 1042$

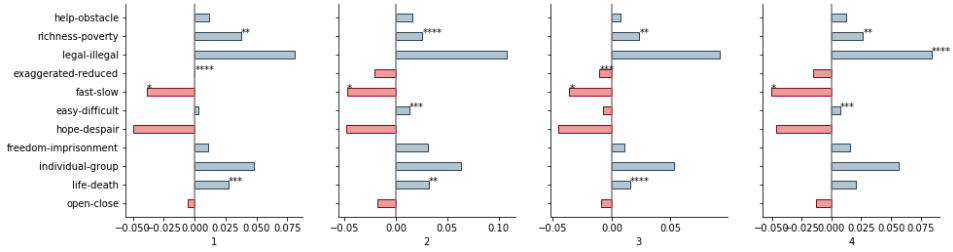
The first attempt to extract information from the experiment [6] was to exploit the Aspect Polarization Index 5.9. We have built the network following the procedure describer in Figure 5.9b, choosing the aspect "to open" ($|N| = 563$, $|E| = 158.203$, directed). Then, we generated the node ranking computing for each node i $\rho_a(i, m)$ with m equal to October 2020. In the



(a) Micro-frame bias and intensity for March 2020.



(b) Micro-frame bias and intensity for April 2020.



(c) Micro-frame bias and intensity for May 2020.

Figure 5.10: microframe analysis for experiment [1] in Table 5.2. Bars stand for Micro-frame bias, (*)-(****) for the four more intense aspects among the ones analyzed, being (*) the largest.

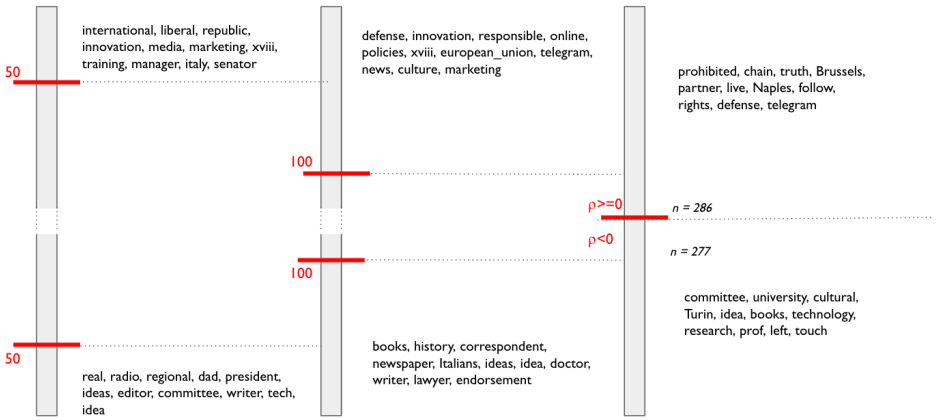


Figure 5.11: Characterizations of the users that occupy the extremes of the ranking generated by ρ_a in the experiment [6]. By means of *td-idf*, we have extrapolated the most relevant words shared by a given portion of the ranking.

top position of the ranking, the method places users whose content tends to shift toward the idea of openness with respect to their neighbors. On the bottom, are users whose content moves away from the idea of openness with respect to their neighbors. To better characterize the users that occupy extreme positions in the ranking we took advantage of the "description" field every user can write as a self-presentation and that appears under the screen name. Through Figure 5.11 we can compare characterizations of the users that occupy the extremes of the ranking generated by ρ_a. By means of *td-idf*, we have extrapolated the most relevant words shared by a given portion of the ranking. In the first attempt, we fix a threshold at the 50^{th} position from the top and bottom, in the second fix a threshold at the 100^{th} position, and in the last, we group users according to the sign of their ρ_a. Interestingly, we detect a mild pattern: culture, writing professions, university, and research appear at the bottom, while innovation, defense, marketing, and liberalism are on the top. At this point we have a half-complete network, each couple of nodes has a link between them, that can go in one direction or the other; the overall density is 0.5. To make better use of the network framework we need to apply a link-filtering strategy to better handle the information of the model. We will embark on this route in the next Section when comparing the networks of engagement and the network of textual similarity generated by the aspect.

5.5 Network layers interplay

Until now we have composed a dual structure, a two-layer system that encodes information about the online debate from two different standpoints. We began by trying to capture the engaging structure of the debate, then we exploited it to measure the context of the debate regarding a specific aspect of a certain topic. In this last part, we attempt to add to the work one last step: to look at how these two parts relate to each other. We leverage the experiment [6] and the "openness" aspect. In particular, we build an undirected network $G_T = (V_{text}, M = 10, \varrho)$, following the steps described in Figure 5.9a. Figure 5.12 summarizes all the tests. On the top of the picture, we have the engagement layer, built over the retweet of the politicians collected through the Procedure 1 on October 2020, with the partition obtained from the Stochastic Block Model. G_T shapes the conversation around an aspect by means of cosine similarity between said aspect and the difference of a pair of users' texts (their tweet collection), with the idea that the user A and B shares a link if their text is balanced with respect to the aspect. We search for the balance in values of ϱ close to 0. Figure 5.12 (a) and (b) show the different results in the network partition according to different values of the threshold T for ϱ, where choosing T equal to 0.001 means retaining all the edges such that $-0.001 < |\varrho(i,j)| < 0.001$. Different T give rise to different networks, more connected for higher values (see $T = 0.005$), or more scattered for lower values (see $T = 0.0001$). For the rest of the analysis, we choose $T = 0.001$, which generates 28 unique blocks. For interpretability reasons we aggregate these blocks by looking at the counting edges matrix between groups, we divide the overall layer network into 3 parts named A, B, and C. From one layer to the other we detect a block shifting, in the engagement network the majority of the users lie in block 0, while they distribute more evenly in the text similarity network, as shown in Figure 5.12 (d). Finally, looking at Figure 5.12 (e) we compare the most frequent words in both the posts (retweet for the engagement layer and tweet for the text similarity layer) and the users' description. In block 0 of the engagement layer, emerges the word "closeness", which thanks to the fact the said block is the biggest one seems to occupy a large portion of the debate. The most relevant word in block 0 is "area", which is consistent with the policies that were in place at that time. Based on the number of positives the different Italian regions would have been able to enter three particular administrative areas, within which increasingly stringent and limiting rules of social life were in effect, having as their goal to minimize contagions and not overburden the hospitals. According to the most relevant words in the users' descriptions, this group appears linked to media and information. In the text similarity layer, news media and information accounts seem to be present in each group. We find the least equivocal clue in Group C, where emerge the word "nolockdown". Otherwise, it is difficult to interpret how these words project

into the discourse, and what aspect they portray of the problem. To do so we need a guide or a system that somehow takes context into account. We apply the micro-frame analysis previously describe, the results are shown in Figure 5.13. On the top, we have the micro-frame bias and intensities linked to the retweet text for each block in the engagement layer, on the bottom we have the micro-frame bias and intensities linked to the tweet text for each block in the text similarity layer. What we gather is that the axes that appear the most in the engagement layer are "richness-poverty" and "open-close", the second is consistent with the data collection. In the third position, we find "fast-slow", in the fourth "individual-group" for block 0 "life-death" for block 1, and "easy-difficult" for block 2. Biases of these pair of antonyms agree with each other. Even if not among the most intense micro-frame, there are two major differences in the pole toward the biases lean, which are "freedom" for group 0 against "imprisonment" for group 2, and "exaggerated" for group 0 against "reduced" for group 2. Group 1 for these two aspects is almost perfectly balanced. What happens in the text similarity layer is that every block agrees almost perfectly and positions almost in the same way over the semantic axis, which may suggest an oversimplification of the partitioning of the network or some redundancies in the procedure used to build the network, so further analyses will be necessary.

5.6 Final remarks

In this Chapter, we present an empirical data-driven analysis of the online debate around the pandemic of Covid-19 that unfolded in 2020. We employ Twitter data to drive the analysis, using them to build a two-layer structure whose interplay aims to reveal how regarding a topic a specific group of people converse. We propose a pipeline that starting from data collection finalizes the construction of the first layer, the engagement layer, then starting from there allows the assembly of the second layer, the text similarity layer, which derives from the first layer but aims to independently shed light over a slightly different point of view. We measure the difference between re-posting content and producing the original.

Within the engagement layer, we notice predominantly the split between a group that was made largely from economics-related profiles. In the face of the pandemic situation, an unsolvable trade-off has emerged between health and economic choices, whose priority has long been discussed. However, the most relevant words that emerge in the different blocks refer in a similar way to the main topic of the pandemic, they do not explain the difference in opinions or values. To outline different standpoints we slit the general debate into different topics, then within each topic, we investigate specific semantic antithesis by means of the FrameAxis procedure. We exploit the ideas of the semantic axis to build the text similarity layer network, shaping

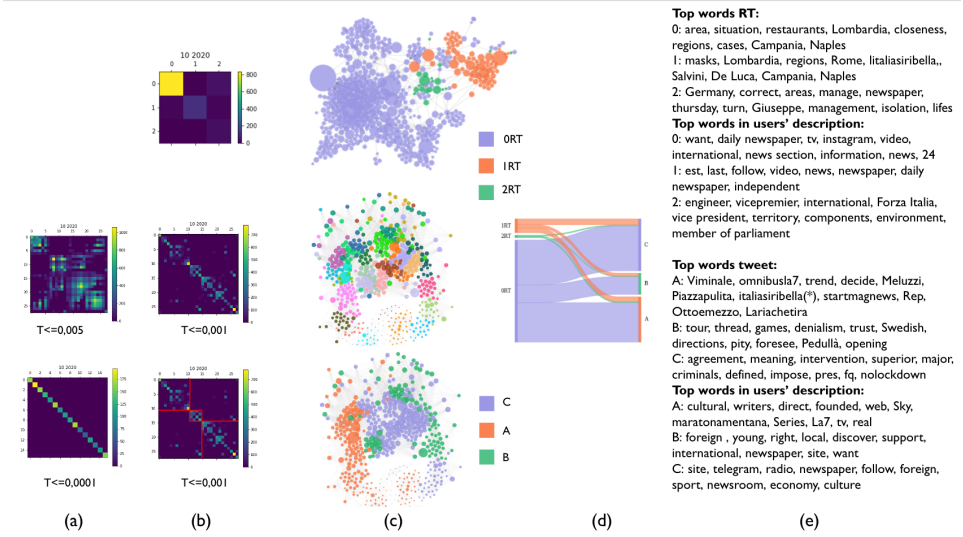
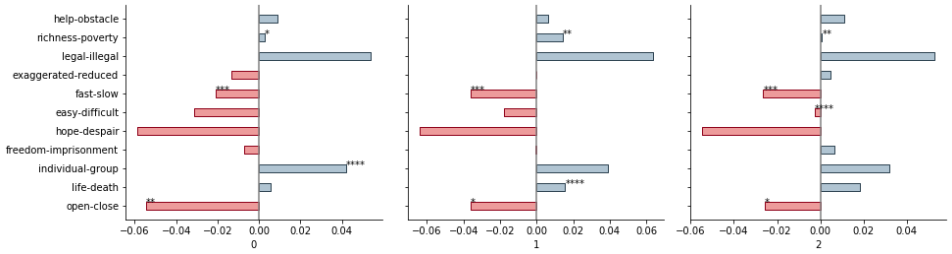
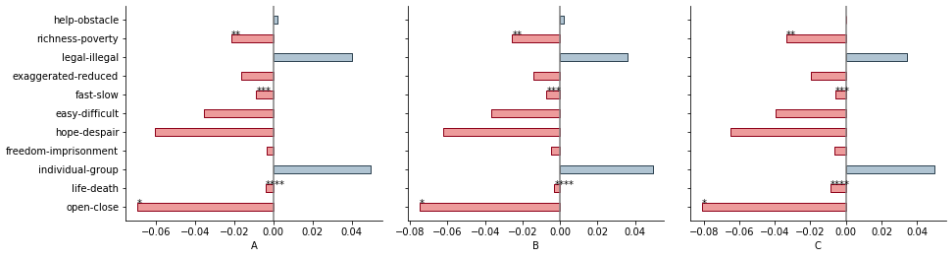


Figure 5.12: On the top of the picture, we have the engagement layer, built over the retweet of the politicians collected through the Procedure 1 on October 2020, with the partition obtained from the Stochastic Block Model; (a) and (b) show the different results in the network partition according to different values of the threshold T for ϱ; (c) exhibits the resulting networks; (d) portrays the block shifting from one layer to the other; (e) compares the most frequent words in both the posts (retweet for the engagement layer and tweet for the text similarity layer) and the users' descriptions. "italiasiribella" was a common hashtag translatable as "italyrebels", "maratonamentana" refers to a famous news commentary program.



(a) Engagement network layer.



(b) Text similarity network layer build through the aspect "open".

Figure 5.13: micro-frame analysis for experiment [6] in Table 5.2. Bars stand for Micro-frame bias, (*)-(****) for the four more intense aspects among the ones analyzed, being (*) the largest.

the interaction between nodes either as a balanced stance toward a semantic axis or very unbalanced leveraging the direction of each edge to encode how two nodes position themselves with respect to a certain aspect, for example "openness". Here nodes are the whole text production of a user during a given month of 2020. Partitioning the graph, we obtain blocks that agree almost perfectly and position themselves almost in the same way over the semantic axis, which may suggest an oversimplification of the method, as shown in Figure 5.6. There are also biases in the data, the topic even if rich in point of view has a limited lexicon, and so any automatic analysis system is weakened by much redundancy of vocabulary. In this context distinguishing various plots becomes very complicated. This reflect also, in the analysis of the interplay between layers, at this point is mostly a qualitative study, that needs refining and better tuning with reality.

Finally, we propose a strategy to frame an aspect inside a debate, ranking each user according to the ρ_a index. With the aspect of "openness" inside the lockdown topic, a mild pattern can be detected: culture, writing professions, university, and research rank at the bottom, while innovation, defense, marketing, and liberalism are at the top.

In addition to limitations in the data, there are also parts that can be improved in the structure of the method. For instance, starting with the choice of the initial group of users, that may be very arbitrary, more so since

it is the same for every month of 2020 at this stage of the work. It is still not clear how it impacts the convergence of the pipeline. More analysis must be structured in order to shape a more solid theoretical background over the convergence, and how the different parameters vary its outcome, starting for example by the selection of the threshold T that rules the finding of new users in the building of the engagement layer. We have seen that the debate can be very multilingual and so could be quite important to develop new mechanisms that keep this diversity into consideration.

The overall method ensembles two different drives, snowball sampling in the search of new users, and a bounding box control effect, whose effects can be found, for example, in the keywords selection during the Twitter API queries, or in the newspaper filtering throughout the collection of new posts. The success of the framework will be in finding the perfect trade-off between these two opposite fundamental thrusts.

Chapter 6

Ancient Civilization

6.1 Overview

In this Chapter, we combine two well-established frameworks: Linked Data to obtain a rich data structure, and Network Science to explore different research questions regarding the structure and the evolution of ancient societies. We propose a multi-disciplinary pipeline where starting from a semantically annotated prosopographic archive, a research question is translated into a query on the archive and the obtained dataset is the input to the network model. We applied this pipeline to different archives, a Hittite and a Kassite collection of cuneiform tablets, which represent official documents and that can be considered bureaucratic (and so technical) traces of a social and political system in place at the time.

Finally, network visualization is presented as a powerful tool to highlight both the data structure and the social network analysis results, adaptable to different research questions. The results obtained can be appraised in a more insightful way by domain experts.

This work was partially presented at the 1st Workshop on Artificial Intelligence for Cultural Heritage [141]. The building of the complete database is still in the making so the results presented are still partial, but nevertheless, they represent a useful step towards the final scope of the entire project: the study of networks operating at local and regional levels in the Mesopotamian-Anatolian region.

6.2 Research Questions

- We aim to build a multi-disciplinary pipeline where, starting from a semantically annotated prosopographic archive, a research question is translated into a query to the archive, and the obtained dataset is the input to a network framework.

- We will present how the pipeline can be adapted to different research

questions and tasks.

6.3 Ancient Near-Eastern Corpora

The Late Bronze Age (LBA) social structures were based on the court model, their internal networks, and the economic systems were controlled on the basis of selected epigraphic and archaeological sources. Cuneiform sources attest to the existence of hundreds of thousands of persons who lived millennia ago in ancient Mesopotamia. In most cases, we get only glimpses into their lives when they interact with an economic institution or administrative unit that decided to record this information on a clay tablet, which then fortuitously came down to us. In fact, for most of these people, we have so little information – often just a name – that it would be impossible to write proper biographies about them and fully describe the reality of their daily lives. Thus, they tend to disappear in the anonymous crowds that populate the indices of text editions, while they could be an extremely valuable resource for investigating the organization of ancient societies.

Late Bronze Age documents dealing with the administration and economy of the Near Eastern kingdoms and polities differ as regards their typology, contents, and aims.

Kassite [1] were a people who probably originated in the Zagros and who ruled Babylonia in the 16^{th}-12^{th} centuries BCE. Their corpus has been extracted from a group of 776 cuneiform tablets dating in the 14^{th}-13^{th} centuries BCE, which belong to the Rosen Collection and were formerly on loan at Cornell University (NY) [142]. They might have come from a town whose ancient name was Dūr-Enlilē, described as "an important economic center that was to a certain degree dependent on Nippur and played an important role in the administration" [143], with Nippur being the capital of the reign. The Collection texts are administrative records mainly dealing with the income, storage, and redistribution of agricultural products (mostly cereals, but also sesame, pulses, and cress) and by-products (beer and flour), animal husbandry, and textile production; smaller groups of texts include legal documents and letters.

The Hittites were an Anatolian people who played an important role in establishing first a kingdom in Kussara (before 1750 BC), then the Kanesh or Nesha kingdom (c. 1750–1650 BC), and next an empire centered on Hattusa in north-central Anatolia (around 1650 BC). The Hittite documentation includes very few administrative and economic records [144] but it is very informative as far as the governance of the state and the role played by the court and the officials are concerned. Little is known about administration in the Hittite Society. Administrative documents were presumably written on wooden tablets that are not preserved. Despite this, cuneiform tablets found

[1] https://www.iranicaonline.org/articles/kassites

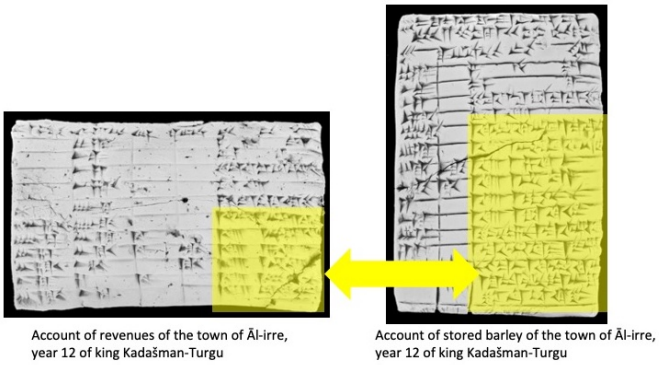


Figure 6.1: Excerpts of cuneiform texts referencing the same location (the town of Āl-irre)

in many Anatolian sites, offering interesting study material; we refer here to texts, such as the inventory tablets, the cadastral records, the lists of people and workers, the letters, the depositions recorded on the occasion of trials, the so-called “cult inventories" dealing with offerings and festivals, the donation tablets, the royal decrees, the texts dealing with foodstuffs, etc. The corpus of Hittite written documents mainly consists of the tablets and seals found at Hattusa and other Anatolian sites such as Maşat Höyük/Tapikka, Kuşaklı Höyük /Sarissa, Kayalıpinar/Samuha, and Ortaköy/Sapinuwa. The research takes also into consideration the documents from the Syrian polities subordinated to Hatti, namely, Karkemish, Ugarit, Alalah, and Emar. Hittite written documents and the sealings mention a huge number of personal names, titles, and professions.

6.4 Pipeline

To overcome the inherent limitations of fragmented evidence, the project is characterized by an interdisciplinary approach, combining traditional linguistic, archaeological, lineage, and historical research methods with methods and tools developed in the digital humanities, such as factoid-based approaches, to develop consistent models that provide a schematic description of the activities of the target population through direct links to sources. The representation of prosopography relies on the pioneering work carried out by [36] [36]. The Factoid-based Prosopographic Ontology[2] revolves around the notion of factoid, intended as a believed-to-be-true, reported event in some written source, a definition that fits very precisely the data inferred from the corpora investigated by the project. The Factoid model connects two basic entities: the Source, where the factoid is asserted (a Hittite or Kassite cuneiform text), and the Relation it describes (e.g., an administra-

[2]https://github.com/johnBradley501/FPO

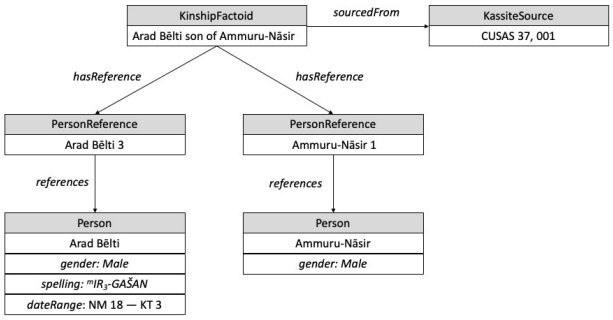


Figure 6.2: Representation of a kinship relationship in the Kassite dataset.

tive or kinship relation), further specialized in a Person reference and in a Location reference (respectively, the personages and places involved in the relation). Figure 6.2 illustrates an example factoid, namely the kinship relationship between two personages sourced from a cuneiform text from Kassite Babylonia.

1. **Record creation**: this step, carried out by a domain expert (namely, an *archaeologist* or a *historian*), starts with the ingestion into the Omeka S platform of the data extracted from the cuneiform texts encoded in the clay tablets, through the factoid paradigm. This is where the data are first interpreted, as experts are asked to identify the entities mentioned in the cuneiform script (often encoded with different spellings) and the relationships between them, aggregating possible homonyms. Omeka S platform is a relational database that allows collecting, publishing, and sharing data with Linked Open Data via built-in REST APIs [3] in JSON-LD format. This step does not require any familiarity with the Linked Data formats, since the ingestion is carried out via a set of interlinked web forms.

2. **Data extraction**. Omeka S allows formulating queries in a semantic format but, in practice this strategy falls short to translate the most complex research questions into queries on the archive. To bypass this limitation, the resulting knowledge graph is then stored in an Apache Fuseki [4] triple store. Through this step, the research questions formulated by the *domain experts* (historian and archaeologist) are translated by the *knowledge engineer* into SPARQL queries and executed on the knowledge graph.

3. **Network model**. Finally, a network analysis model is built with the data extracted from the SPARQL query.

[3] https://omeka.org/s/docs/developer/api/rest_api/
[4] https://jena.apache.org/documentation/fuseki2/

6.5 Tasks

In the following, we describe three different research questions that have been explored by employing the pipeline described above.

6.5.1 Investigating the co-occurrence between persons and locations sourced from the Kassite document collection.

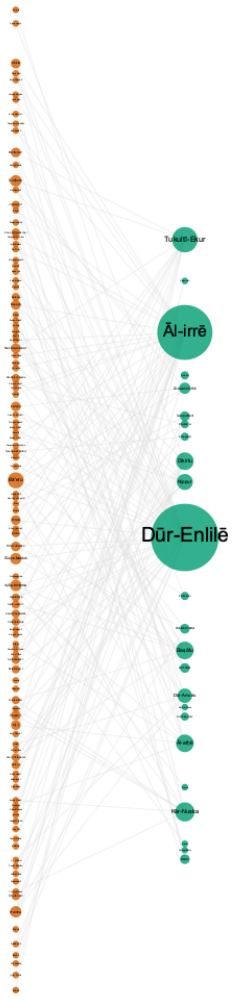
The original archaeological context of these documents is unknown, but it can be assumed that they originated from the same administrative center – if not from the same archive – because of the typological, prosopographic, geographical, and chronological features they share. To shed light on the geographical aspect of the archive it is possible to employ the pipeline to search for patterns that can reveal previously unknown aspects of the society of the time.

The raw data, shown in Figure 6.1, which represents examples of two cuneiform texts mentioning the same location in different contexts, has been ingested into records and stored in the knowledge graph. Through the query 6.5.1, we filter the information needed. For this task, we defined a bipartite network $G = (U, V, E)$, where U is the set of Kassite persons, V the set of locations, and E the set of edges (i, j) that exists between nodes $i \in U$ and $j \in V$ if the person i appear in some activity in location j. The network is shown in Figure 6.3. The further analysis involves both the possible projections of the bipartite network. A projection is a compressed version of the bipartite network that contains nodes of only either of the two sets, nodes are connected only when they have at least one common neighboring in the other set. Figures 6.3b and 6.3c show the projections obtained with simple weighting, that is where edges are weighted by the number of times a common association between two nodes of the same set with the same node of the other set is repeated. The projection networks can be leveraged to gain a deeper view of the twofold system.

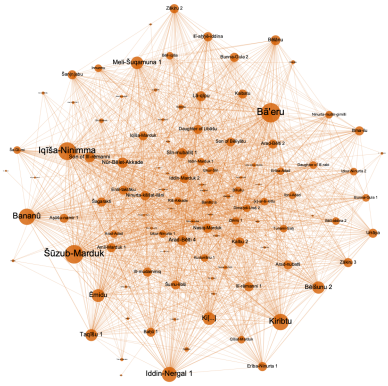
```
SELECT ?id  ?name ?location
WHERE {
   ?person rdf:type kfpo:KassitePerson .
   ?person  dcterms:title ?name.
   ?person  omeka:id ?id.
   OPTIONAL{?person kfpo:hasLocationName ?location}
}
```

6.5.2 Trade Network in the Hittite Empire

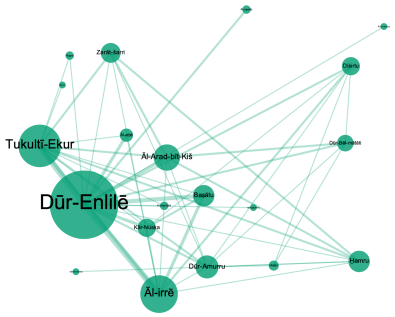
In this Section, we will frame how to investigate the geographical trade, in particular, how a specific administrative role distributes over the cities of



(a) Bipartite network.



(b) Projection over Kassite persons.



(c) Projection over Kassite Empire locations.

Figure 6.3: Kassite co-occurrence person-location bipartite network and projections. (b) represents the network projection over the persons ($|U| = 108$, $|E| = 1475$). (c) represents the network projection over the locations ($|U| = 20$, $|E| = 51$). Node dimensions scale over weighted degree.

the Hittite Empire. We filter the data through the query 6.5.2: we retain all the persons in the late Hittite period that appear as witnesses in some transactions. We define the network to be undirected, weighted with colored node, $G = (V, \varpi)$, where V is the set of Hittite persons and $\varpi : V \times V \to \mathbb{N}$ is the function defining for each pair of nodes $i, j \in V$ the weight of edge (i, j) that measures the times the nodes appear as witnesses in the same trade. A network is said to be *colored* if we associate colors, or labels, to each of its vertices and/or edges. We assign as labels the city to which the

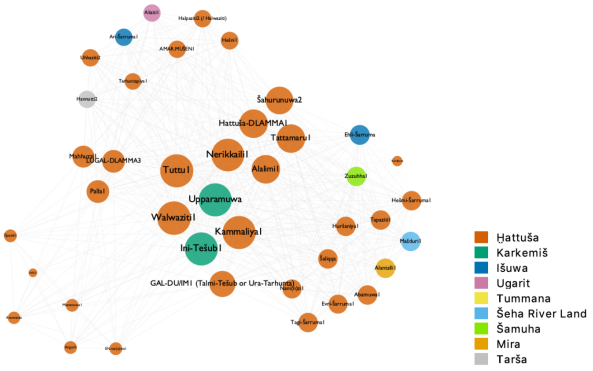


Figure 6.4: Late-period Hittite witness relationships. Nodes ($|U| = 40$) are Hittite persons; dimension scales over the weighted degree; color schema is location. Edges ($|E| = 449$) stand for two persons who appeared as witnesses in the same transaction.

person belongs. Discovering motifs in the resulting network would mean discovering interesting trade routes that were used in the Empire. A partial result is depicted in Figure 6.4.

```
SELECT ?name ?link ?to
WHERE {
   ?link dcterms:title ?name;
   hfpo:hasWitnessReference ?to.
   ?link hfpo:sourceLanguage "Late Hittite".
}
```

6.5.3 Aklu Document

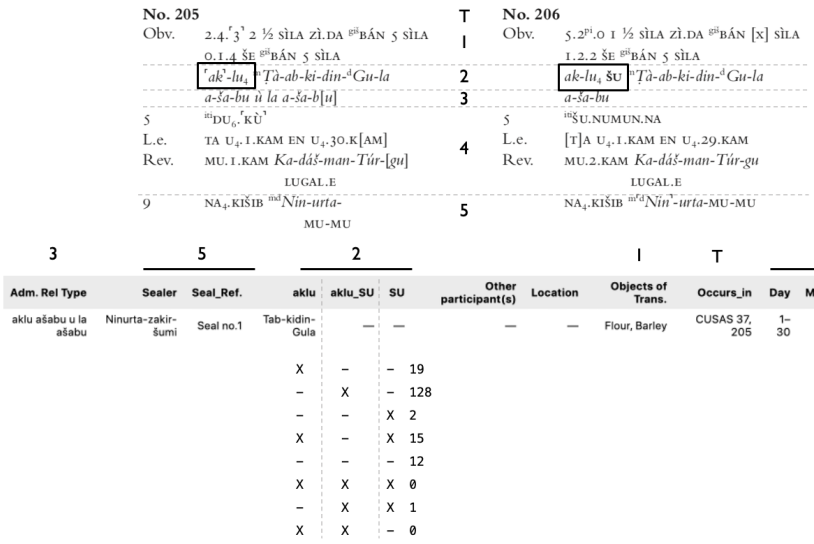
A particular group of documents, drafted by the Kassite administration, recording the issue of foodstuffs is represented by the "aklu-texts". Single aklu-expenditures are recorded on small, usually sealed tablets. The scheme described in Figure 6.5 applies to several documents in the overall corpus used for the following analysis, but nevertheless, it is difficult to define a type that would fit them all, especially because there is a significant degree of variation in the sequence of information conveyed by these documents. In this case, the final database includes multiple collections: 41 texts are from the Rosen Collection, already mentioned, and 110 aklu documents come from the Nippur [5] Collection [145], and finally 60 texts are documents from the Hilprecht Collection [146] in Jena. Merging documents from multiple sources is advocated by the nature of the data, we are dealing with administrative

[5] An important administrative Kassite site.

records that we can suppose to be coherent within a limited time span and the same ruling empire if only to ensure the governability of the kingdom.

The administration of the reign was directed from the palaces of various cities. The administrative centers or in many cases the temples, either in Nippur or in the other major cities, were in charge of the economic organization, which consisted in collecting numerous agricultural deliveries and imposts from broad parts of Babylonia. The gathered commodities were then used as allowance or loans, delivered to people possibly in return for labor. In this context, the term *aklu* is interpreted as the artisans' delivery of commodities to a facility, such as the storehouse of the palace. Then, the commodities are disbursed to the beneficiaries. Each document representing this type of transaction follows the schema 6.5, however, is still unclear how the people that appear in the documents relate to each other and with the administration. Names appear in different parts of the texts, in line 2 appear the persons involved in the trade, while in part 5 the Sealer. Each sealer may use a different seal. In line 2 the name of several characters may occur in three different forms: after the word "aklu", after the pair of words "aklu ŠU", or after the word "ŠU". The three forms could appear in different combinations in each text as shown at the bottom of Figure 6.5, but never at the same time. The bureaucratic meaning of these three forms is still unknown, as well as the direction of the trade, namely who trades with whom. In order to shed light on these open questions we propose a network framework that leverages multiedges and projection to represent different administrative circumstances. We define an administrative circumstance as the feature set of the trade, $A = [a_1, a_n]$, where A can be enriched with many different features, like the sealer, the seal, the date of a transaction, the object of a transaction, and so on. We encode this information in a bipartite network characterized by two sets of nodes P, A, the people that interact in the trade and the administrative circumstances. The edges are labeled according to the type of reference linked to the person, either "aklu", "aklu ŠU" or "ŠU". Figures 6.6, 6.7, and 6.8, depict different case studies. In 6.6 A is composed only by the Sealer name, we see that some of them appear also among the people involved in the trades, breaking in this way the bipartite structure of the network. 6.7 pairs sealer and the seal he uses to certify the document, the same sealer may use more types of seal. The last is built around the more rich type of A since A is characterized by the sealer, the seal, the type of transaction, the month (since commodities are food, they might follow some sort of seasonality), and objects of the transaction; nodes in the set of A are colored by the object of the transaction (or set of the objects of the transaction), gray nodes stand for the people making the exchange.

The interpretation of the final results may serve experts to find new patterns or connections among the different characters that lived during this period, or about the administrative guidelines needed to manage the kingdom.



3	5		2					I	T	4			
Adm. Rel Type	**Sealer**	**Seal_Ref.**	**aklu**	**aklu_SU**	**SU**	**Other participant(s)**	**Location**	**Objects of Trans.**	**Occurs_in**	**Day**	**Month**	**Year**	**King**
aklu ašabu u la ašabu	Ninurta-zakir-šumi	Seal no.1	Tab-kidin-Gula	—	—	—	—	Flour, Barley	CUSAS 37, 205	1–30	VII	1	KT
			X	–	–	19							
			–	X	–	128							
			–	–	X	2							
			X	–	X	15							
			–	–	–	12							
			X	X	X	0							
			–	X	X	1							
			X	X	–	0							

Figure 6.5: The principal scheme of two aklu documents, with all possible features, and the distribution over all the documents of the types of references of people making the trade, “aklu”, "aklu ŠU" or "ŠU".

6.6 Final Remarks

In this Chapter, we described a pipeline developed to support the study of ancient societies with visualizations from a semantic representation of prosopographic data. Starting from different research fields, we merge them into a pipeline, with the scope to study a specific historical domain. The methodology and the obtained visualizations, designed by a multi-disciplinary team involving knowledge representation experts, digital humanists, historians, and visualization experts, have been positively assessed by the domain experts, who have been able to confirm the research hypotheses behind the creation of the networks. However, the final validation must be postponed to the completion of the archives, which now account for a significant but still incomplete portion of the source data.

Of course, the picture we get of any ancient society through administrative records is just a glimpse of the actual lives of people during that time. Data is often incomplete, unreadable and of uncertain origin, so much so that any possible interpretation must be made with extreme caution and with the understanding that there is no universally verifiable ground truth. But still, glimpses of this socio-technical system can be enough to arise new research questions and new ideas for projects that with time and effort could lead to new discoveries.

In future work, we plan to fully automatize the pipeline, so that the extraction of data, the construction of the network, and its visualization can

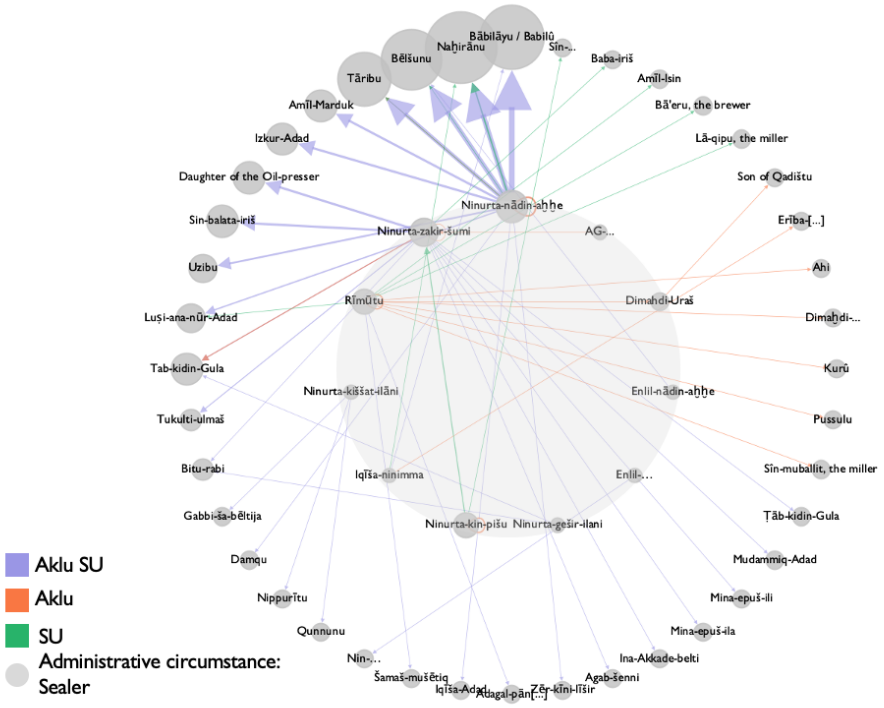


Figure 6.6: Network between Sealer and the people making the trade in aklu document. Nodes scale over in-strength. The edges are colored according to the type of reference linked to the person, either “aklu”, "aklu ŠU” or "ŠU”.

be executed in the same environment. We also plan to generalize as much as possible the approach behind the pipeline building and to further implement specific network analysis.

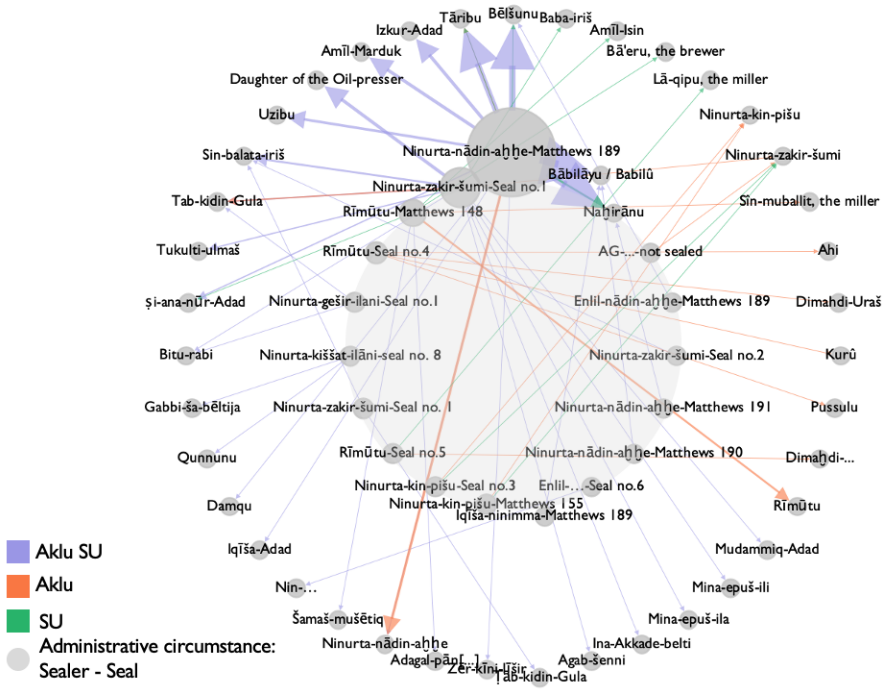


Figure 6.7: Network between Sealer+Seal and the people making the trade in aklu document. Nodes scale over in-strength. The edges are colored according to the type of reference linked to the person, either “aklu”, "aklu ŠU” or "ŠU”.

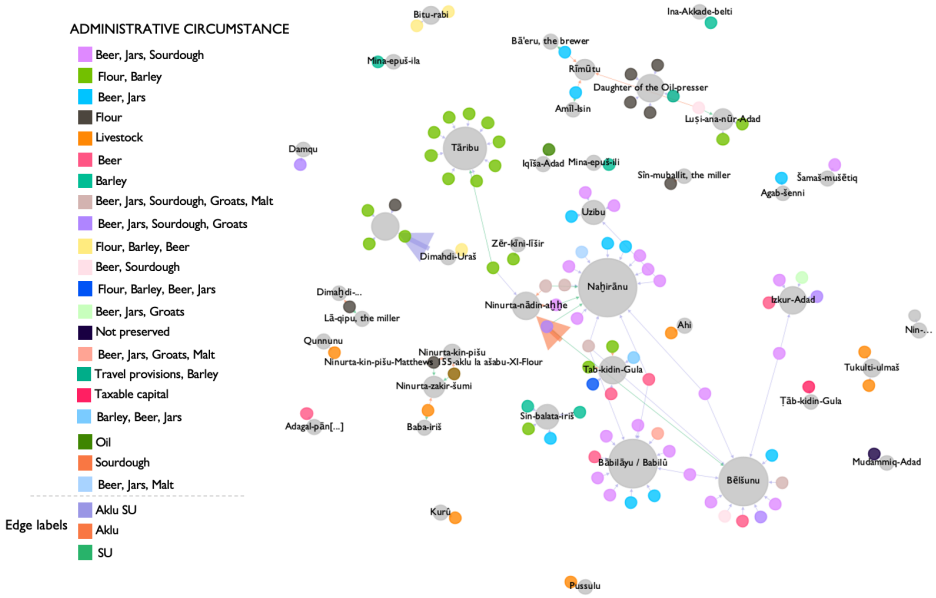


Figure 6.8: Network between the people making the trade in aklu document the administrative circumstance characterized by the sealer, the seal, the type, the month, and the object of the transaction. Nodes scale over in-strength and are colored by the object of the trade transaction. The edges are colored according to the type of reference linked to the person, either “aklu”, "aklu ŠU” or "ŠU”.

Chapter 7

Conclusions

All of the works reviewed in this thesis are empirical analyses that address very different contexts. But there is an underline motif that connects all the three systems presented, the scientific migration ecosystem, the ancient population archive analysis, and the online debate around Covid-19, which is the interplay between the human component and the technical regulation that originates the networks, or in a broader sense the conditions that partially shape the interaction. "Partially" means that still, the human component plays a fundamental role in the evolution of the systems, driving it. To seize the overall complexity we exploit data. However, data are always incomplete, with reasons that may be multiple, because we extract it from terracotta tablets being almost four thousand years old, or because they belong to a private company, that deals with content sharing, but at the same time regulates how everyone can access them. At the same time, the common underlined motif suggests common strategies for us to apply to extrapolate patterns and mechanisms. We have pursued two main general strategies to answer the different research questions:

1. Centrality ranking
2. Community detection

being both sometimes the ultimate goal of some analyses and sometimes an intermediate step in a pipeline or the building criteria of other lines of research, but nonetheless, they are the bases of investigations chosen to unearth the regulatory mechanisms of the systems studied. In particular, we exploit the topology of the respective networks, searching for outliers in the link distribution that could suggest a prominent role in the social fabric, and voids and fullness in the general link pattern, that we can use as guidelines to cut the network into different communities or block. Starting from these two well-established analysis frameworks we build and model more elaborated structures to encode the data, such as multilayer networks, coupled bipartite networks, colored networks, or temporal networks. The results we get are still partial, and improvable in many aspects, from the delineation of

a more formal theoretical background, such as the convergence analysis of the pipeline of Chapter 5, to a more refined use of the data, as best described in the final remarks of Chapter 4, where we delineate many possible analyses that take into account, for example, a researcher's first career transition after the educational period.

We can conclude by saying that efforts in designing, managing, auditing, and ultimately improving methods to study these systems may benefit greatly society, especially in those spaces where technical contexts interact with human dynamics.

Appendix A

A.1 ORCID

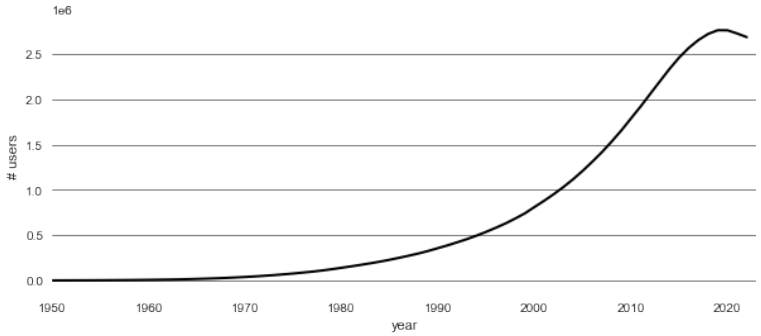


Figure A.1: Distribution of the number of ORCID members with at least one active affiliation per year, from 1950 to 2023.

A.2 Network Model

In Table A.1, we show some basic network statistics, grouped by year. For each year $y \in [2000, \ldots 2021]$ we show the number of *nodes*, i.e., countries that occur as a source or as a destination in that year at least once ($|V_y|$), the number of *links* established during that year between countries ($|E_y|$), and the related following measures: the *density* of the network ($d = \frac{2|E_y|}{|V_y|(|V_y|-1)}$); the *reciprocity*, i.e., the ratio of the number of edges pointing in both directions to the total number of edges in the graph ($r = \frac{|e=(i,j):(j,i))\in E_y|}{|E_y|}$); the size of the Strongly Connected Component (SCC); and the diameter of the network, i.e., the length of the longest path among the shortest ones.

year	nodes	links	density	reciprocity	SCC	diameter
2000	179	1560	0.048961	0.589744	133	6
2001	173	1612	0.054174	0.575682	138	5
2002	175	1709	0.056125	0.586308	140	6
2003	177	1755	0.056337	0.602849	140	6
2004	185	1981	0.058196	0.614841	150	6
2005	185	2101	0.061722	0.628272	149	5
2006	192	2327	0.063454	0.629996	159	5
2007	196	2536	0.066353	0.630915	161	5
2008	205	2742	0.065567	0.641138	168	5
2009	202	2976	0.073297	0.670027	175	5
2010	198	3098	0.079424	0.681085	173	5
2011	207	3340	0.078327	0.688024	181	5
2012	209	3615	0.083157	0.687690	184	4
2013	214	3874	0.084990	0.700568	191	5
2014	218	4004	0.084640	0.693806	194	5
2015	213	4044	0.089556	0.706231	193	4
2016	211	3895	0.087903	0.703979	183	5
2017	200	3457	0.086859	0.694822	171	4
2018	202	3121	0.076868	0.681833	168	5
2019	195	2908	0.076870	0.682256	166	6
2020	188	2369	0.067385	0.651752	149	5
2021	145	1587	0.076006	0.633900	113	6

Table A.1: Summary of some basic network statistics, grouped by year.

A.3 Drain Index

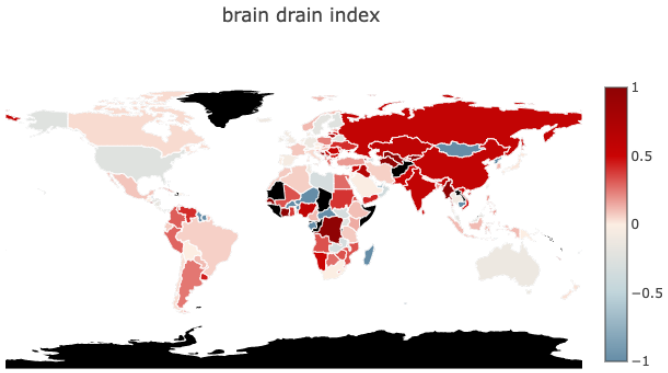


Figure A.2: Drain index β in 2000. Positive (negative) values of β are color coded with different shades of red (blue). Countries without data have been colored black.

A.4 HITS complete ranking

2000, 2010, and 2020 rankings of countries according to authority and hub scores are reported in this section for illustrative purposes.

1	United States	46	Czech Republic	91	Qatar	136	French Guiana
2	United Kingdom	47	Hungary	92	Gambia	137	French Southern T.
3	Germany	48	Poland	93	Ghana	138	Lithuania
4	France	49	Cuba	94	Bhutan	139	Andorra
5	Canada	50	Ukraine	95	Cambodia	140	Kazakhstan
6	Australia	51	Romania	96	Estonia	141	Latvia
7	Spain	52	Bangladesh	97	Honduras	142	South Sudan
8	Italy	53	Oman	98	Paraguay	143	Lesotho
9	Japan	54	Tunisia	99	Nepal	144	Syria
10	Brazil	55	Venezuela	100	North Korea	145	Timor-Leste
11	Netherlands	56	Cyprus	101	Dominican Rep.	146	Montenegro
12	South Korea	57	Slovakia	102	Macau	147	Suriname
13	Portugal	58	Kenya	103	Madagascar	148	Togo
14	Switzerland	59	Puerto Rico	104	Uganda	149	Libya
15	Sweden	60	Lebanon	105	Malta	150	Burkina Faso
16	Mexico	61	Bulgaria	106	Mozambique	151	Tonga
17	China	62	Ecuador	107	Mongolia	152	Moldova
18	Malaysia	63	Algeria	108	Fiji	153	Niger
19	Singapore	64	Croatia	109	Zimbabwe	154	Saint Martin
20	Indonesia	65	Luxembourg	110	Iraq	155	Guyana
21	Ireland	66	Vietnam	111	Botswana	156	Albania
22	Taiwan	67	Pakistan	112	Sri Lanka	157	Curacao
23	Colombia	68	Kuwait	113	Namibia	158	Swaziland
24	Hong Kong	69	Nigeria	114	Georgia	159	Eritrea
25	Austria	70	Ethiopia	115	Liechtenstein	160	Gabon
26	Iran	71	Belarus	116	Seychelles	161	Central African Rep.
27	Turkey	72	Philippines	117	Bermuda	162	French Polynesia
28	New Zealand	73	Malawi	118	Macedonia	163	Benin
29	Denmark	74	Palestinian Ter.	119	Yemen	164	Myanmar [Burma]
30	Belgium	75	Nicaragua	120	Panama	165	Turks and Caicos Is.
31	Greece	76	Jamaica	121	Zambia	166	Armenia
32	South Africa	77	Bahrain	122	Sudan	167	American Samoa
33	Thailand	78	Trinidad and Tobago	123	Azerbaijan	168	Bosnia and Herz.
34	India	79	Iceland	124	Rwanda	169	Congo [DRC]
35	Israel	80	Guatemala	125	Vatican City	170	Turkmenistan
36	Egypt	81	Morocco	126	Barbados	171	Cote dIvoire
37	Chile	82	Serbia	127	S Kitts and Nevis	172	El Salvador
38	Argentina	83	Costa Rica	128	Papua New Guinea	173	Senegal
39	Saudi Arabia	84	Tanzania	129	Bahamas	174	Maldives
40	Russia	85	Cameroon	130	Angola	175	Dominica
41	Finland	86	Cape Verde	131	Bolivia	176	Kyrgyzstan
42	Norway	87	Grenada	132	Belize	177	Monaco
43	Jordan	88	Guadeloupe	133	Mali	178	New Caledonia
44	United Arab Em.	89	Uruguay	134	Sint Maarten	179	Uzbekistan
45	Peru	90	Slovenia	135	Aruba		

Table A.2: Ranking of the countries by **authority** score in **2000**.

1	United States	46	Israel	91	Senegal	136	Papua New Guinea
2	United Kingdom	47	Bangladesh	92	Algeria	137	Togo
3	Germany	48	Russia	93	Zambia	138	Reunion
4	Australia	49	Vietnam	94	Estonia	139	Mali
5	France	50	Jordan	95	Malawi	140	Chad
6	Spain	51	Ecuador	96	Morocco	141	El Salvador
7	Canada	52	Peru	97	Cameroon	142	Nicaragua
8	China	53	Ghana	98	Zimbabwe	143	Slovakia
9	Japan	54	Kenya	99	Bulgaria	144	Guinea
10	Portugal	55	Czech Republic	100	Bhutan	145	Armenia
11	Italy	56	Ethiopia	101	Yemen	146	Cambodia
12	Switzerland	57	Philippines	102	Malta	147	Lesotho
13	Singapore	58	Cyprus	103	Trinidad and Tobago	148	Cote dIvoire
14	Sweden	59	Hungary	104	Mauritius	149	Moldova
15	South Korea	60	Nepal	105	Namibia	150	Dominica
16	Netherlands	61	Romania	106	Bolivia	151	Eritrea
17	Hong Kong	62	Iraq	107	Brunei	152	Bahamas
18	Brazil	63	Oman	108	Angola	153	Bermuda
19	India	64	Venezuela	109	Albania	154	Aruba
20	Denmark	65	Uganda	110	Croatia	155	Antigua and Barbuda
21	Colombia	66	Puerto Rico	111	Afghanistan	156	Cayman Islands
22	Malaysia	67	Syria	112	Latvia	157	Guam
23	Saudi Arabia	68	Luxembourg	113	Gambia	158	Northern Mariana Is.
24	Belgium	69	Sri Lanka	114	Jamaica	159	Liechtenstein
25	Mexico	70	Tanzania	115	Uzbekistan	160	Fiji
26	Austria	71	Cuba	116	Benin	161	Belarus
27	Taiwan	72	Iceland	117	Mozambique	162	Lithuania
28	Ireland	73	Sudan	118	Guatemala	163	Gabon
29	Turkey	74	Panama	119	Cape Verde	164	San Marino
30	Egypt	75	Libya	120	Burkina Faso	165	Guyana
31	New Zealand	76	Costa Rica	121	Bosnia and Herz.	166	Equatorial Guinea
32	Pakistan	77	Palestinian Ter.	122	Bahrain	167	Central African Rep.
33	Finland	78	Lebanon	123	Grenada	168	French Polynesia
34	Indonesia	79	Ukraine	124	Honduras	169	French Guiana
35	Chile	80	Serbia	125	Sierra Leone	170	Timor-Leste
36	South Africa	81	Kuwait	126	Haiti	171	Suriname
37	Greece	82	Kazakhstan	127	Madagascar	172	Congo [DRC]
38	Argentina	83	Botswana	128	Seychelles	173	Curacao
39	Norway	84	Tunisia	129	Kyrgyzstan	174	Niger
40	Poland	85	Macau	130	Macedonia [FYROM]	175	Azerbaijan
41	Nigeria	86	Paraguay	131	Sint Maarten	176	South Sudan
42	Thailand	87	Uruguay	132	Georgia	177	Myanmar [Burma]
43	Qatar	88	Dominican Rep.	133	Vatican City	178	Swaziland
44	United Arab Emirates	89	Rwanda	134	S Kitts and Nevis	179	Maldives
45	Iran	90	Slovenia	135	Mongolia	180	Samoa

Table A.3: Ranking of the countries by **authority** score in **2010**.

1	United States	46	Ecuador	91	Yemen	136	Guyana
2	China	47	Malaysia	92	Zimbabwe	137	Zambia
3	United Kingdom	48	Hungary	93	Cambodia	138	Congo [Republic]
4	Germany	49	Vietnam	94	Morocco	139	Myanmar [Burma]
5	Spain	50	Iran	95	Uganda	140	San Marino
6	Canada	51	Thailand	96	Falkland Is.	141	Laos
7	France	52	Philippines	97	Bahrain	142	Somalia
8	Australia	53	Luxembourg	98	Bhutan	143	Mozambique
9	India	54	Qatar	99	Botswana	144	Mali
10	Italy	55	Ghana	100	Brunei	145	Timor-Leste
11	Brazil	56	Kenya	101	Barbados	146	Guinea-Bissau
12	Switzerland	57	Argentina	102	Cote dIvoire	147	Lesotho
13	Netherlands	58	Nepal	103	Honduras	148	Togo
14	South Korea	59	Ukraine	104	Monaco	149	Azerbaijan
15	Japan	60	Slovenia	105	Malawi	150	Curacao
16	Sweden	61	Cyprus	106	Grenada	151	Georgia
17	Denmark	62	Iraq	107	Dominican Republic	152	Nicaragua
18	Portugal	63	Sri Lanka	108	Uzbekistan	153	Belarus
19	Belgium	64	Estonia	109	Saint Kitts and Nevis	154	Bosnia and Herz.
20	Austria	65	Costa Rica	110	Paraguay	155	Macedonia
21	Ireland	66	Macau	111	Sudan	156	Bolivia
22	Colombia	67	Ethiopia	112	Madagascar	157	Samoa
23	Turkey	68	Rwanda	113	Cuba	158	New Caledonia
24	Singapore	69	El Salvador	114	Sint Maarten	159	Fiji
25	Hong Kong	70	Croatia	115	Bermuda	160	Benin
26	Norway	71	Uruguay	116	French Guiana	161	Syria
27	Finland	72	Latvia	117	Isle of Man	162	Congo [DRC]
28	Israel	73	Jordan	118	Andorra	163	Burkina Faso
29	Taiwan	74	Romania	119	Mongolia	164	Gambia
30	Mexico	75	Iceland	120	Slovakia	165	Kyrgyzstan
31	Poland	76	Malta	121	Reunion	166	Belize
32	Egypt	77	Cameroon	122	Bulgaria	167	French Polynesia
33	New Zealand	78	Oman	123	Namibia	168	Kosovo
34	Saudi Arabia	79	Venezuela	124	Libya	169	Trinidad and Tobago
35	Greece	80	Lebanon	125	Tunisia	170	S Vincent and the Gren.
36	Chile	81	Lithuania	126	Cape Verde	171	Mauritania
37	Czech Republic	82	Angola	127	Senegal	172	Tonga
38	Indonesia	83	Algeria	128	Afghanistan	173	Sierra Leone
39	United Arab Emirates	84	Liberia	129	Palestinian Ter.	174	Gabon
40	Bangladesh	85	Montserrat	130	Papua New Guinea	175	Niger
41	Russia	86	Puerto Rico	131	Vanuatu	176	Maldives
42	South Africa	87	Tanzania	132	Albania	177	Armenia
43	Nigeria	88	Panama	133	Swaziland	178	Moldova
44	Peru	89	Serbia	134	Kuwait	179	Haiti
45	Pakistan	90	Kazakhstan	135	Guatemala	180	Liechtenstein

Table A.4: Ranking of the countries by **authority** score in **2020**.

1	China	46	Poland	91	Panama	136	Macedonia
2	United Kingdom	47	Hong Kong	92	Namibia	137	Yemen
3	Canada	48	Pakistan	93	Botswana	138	Bolivia
4	South Korea	49	Jordan	94	Honduras	139	Zambia
5	India	50	Bangladesh	95	Trinidad and Tobago	140	Macau
6	Germany	51	Norway	96	Mozambique	141	American Samoa
7	United States	52	Puerto Rico	97	Kyrgyzstan	142	French Polynesia
8	France	53	Lebanon	98	Turks and Caicos Islands	143	Barbados
9	Japan	54	Hungary	99	Azerbaijan	144	Turkmenistan
10	Italy	55	Finland	100	Belarus	145	Myanmar [Burma]
11	Brazil	56	Croatia	101	Tunisia	146	Rwanda
12	Spain	57	Czech Republic	102	Uganda	147	Maldives
13	Russia	58	Vietnam	103	Luxembourg	148	South Sudan
14	Mexico	59	Palestinian Ter.	104	Lesotho	149	Togo
15	Australia	60	Philippines	105	Algeria	150	New Caledonia
16	Turkey	61	Ecuador	106	Swaziland	151	Monaco
17	Colombia	62	Uruguay	107	Iraq	152	French Guiana
18	Switzerland	63	Bulgaria	108	Paraguay	153	Montenegro
19	Netherlands	64	Cyprus	109	United Arab Emirates	154	Cambodia
20	Portugal	65	Syria	110	Latvia	155	Guadeloupe
21	Egypt	66	Cuba	111	Moldova	156	Mongolia
22	Taiwan	67	Ethiopia	112	Senegal	157	Guyana
23	Sweden	68	Slovakia	113	Uzbekistan	158	Benin
24	Indonesia	69	Kenya	114	Morocco	159	Gabon
25	Greece	70	Ghana	115	Libya	160	Aruba
26	Austria	71	Sri Lanka	116	Sudan	161	Belize
27	South Africa	72	Lithuania	117	Kazakhstan	162	Grenada
28	Saudi Arabia	73	Guatemala	118	Angola	163	Sint Maarten
29	Israel	74	Bermuda	119	Kuwait	164	Andorra
30	Chile	75	Estonia	120	Georgia	165	Suriname
31	Peru	76	Dominica	121	Nicaragua	166	Tonga
32	Malaysia	77	Zimbabwe	122	Cameroon	167	Bahamas
33	Argentina	78	Oman	123	Dominican Rep.	168	Central African Rep.
34	New Zealand	79	Slovenia	124	Cape Verde	169	Madagascar
35	Denmark	80	Costa Rica	125	Nepal	170	French Southern Ter.
36	Ireland	81	Armenia	126	Cote dIvoire	171	Niger
37	Romania	82	Qatar	127	Eritrea	172	Liechtenstein
38	Iran	83	Malta	128	Papua New Guinea	173	Saint Martin
39	Venezuela	84	Tanzania	129	Bhutan	174	Burkina Faso
40	Singapore	85	Iceland	130	Albania	175	Seychelles
41	Ukraine	86	Jamaica	131	Fiji	176	North Korea
42	Serbia	87	Malawi	132	El Salvador	177	Timor-Leste
43	Nigeria	88	Gambia	133	Congo [DRC]	178	Saint Kitts and Nevis
44	Belgium	89	Mali	134	Vatican City	179	Curacao
45	Thailand	90	Bahrain	135	Bosnia and Herz.		

Table A.5: Ranking of the countries by **hub** score in **2000**.

1	China	46	Poland	91	Senegal	136	Nicaragua
2	India	47	Iraq	92	Malta	137	Uzbekistan
3	United Kingdom	48	Jordan	93	Lithuania	138	Sierra Leone
4	United States	49	Thailand	94	Cambodia	139	Macedonia
5	Germany	50	Finland	95	Kuwait	140	Macau
6	Canada	51	Norway	96	Kazakhstan	141	Niger
7	Italy	52	Ukraine	97	Tunisia	142	Namibia
8	Spain	53	Sri Lanka	98	Rwanda	143	Angola
9	France	54	Philippines	99	Oman	144	Madagascar
10	South Korea	55	Romania	100	Botswana	145	Cote dIvoire
11	Brazil	56	Hungary	101	Georgia	146	Mauritius
12	Colombia	57	Ghana	102	Libya	147	Brunei
13	Japan	58	Puerto Rico	103	Algeria	148	Togo
14	Portugal	59	Venezuela	104	Afghanistan	149	Barbados
15	Australia	60	Nepal	105	Malawi	150	Fiji
16	Netherlands	61	Lebanon	106	Yemen	151	Burundi
17	Turkey	62	Uganda	107	Bahrain	152	Congo [DRC]
18	Switzerland	63	Serbia	108	Burkina Faso	153	Timor-Leste
19	Iran	64	Ethiopia	109	Sint Maarten	154	Kyrgyzstan
20	Sweden	65	United Arab Emirates	110	Cayman Islands	155	Seychelles
21	Mexico	66	Palestinian Ter.	111	S Kitts and Nevis	156	Somalia
22	Russia	67	Kenya	112	Belarus	157	Belize
23	Greece	68	Czech Republic	113	El Salvador	158	Benin
24	Egypt	69	Costa Rica	114	Slovakia	159	Kiribati
25	Taiwan	70	Croatia	115	Honduras	160	Guinea
26	Indonesia	71	Qatar	116	Slovenia	161	Samoa
27	Ireland	72	Cuba	117	Mozambique	162	Moldova
28	Singapore	73	Bulgaria	118	Guatemala	163	Laos
29	Pakistan	74	Zambia	119	Albania	164	Cape Verde
30	Belgium	75	Panama	120	Swaziland	165	Vatican City
31	Nigeria	76	Tanzania	121	Latvia	166	Eritrea
32	Austria	77	Bhutan	122	Tajikistan	167	New Caledonia
33	Saudi Arabia	78	Cyprus	123	Morocco	168	Guadeloupe
34	Vietnam	79	Sudan	124	Armenia	169	French Guiana
35	Chile	80	Dominican Rep.	125	Bolivia	170	Andorra
36	Denmark	81	Syria	126	Zimbabwe	171	Turkmenistan
37	New Zealand	82	Uruguay	127	Mongolia	172	Guyana
38	Bangladesh	83	Estonia	128	Mali	173	Reunion
39	Israel	84	Paraguay	129	Montenegro	174	Lesotho
40	Hong Kong	85	Trinidad and Tobago	130	Djibouti	175	Gambia
41	Malaysia	86	Grenada	131	Gabon	176	Dominica
42	Argentina	87	Jamaica	132	Marshall Islands	177	Curacao
43	Ecuador	88	Haiti	133	Papua New Guinea	178	Bosnia and Herz.
44	South Africa	89	Iceland	134	Myanmar [Burma]	179	Maldives
45	Peru	90	Cameroon	135	Luxembourg	180	Liechtenstein

Table A.6: Ranking of the countries by **hub** score in **2010**.

1	United States	46	Ecuador	91	Rwanda	136	Myanmar [Burma]
2	United Kingdom	47	Greece	92	Sierra Leone	137	Congo [DRC]
3	Germany	48	Hungary	93	Falkland Is.	138	Liechtenstein
4	Spain	49	Argentina	94	Somalia	139	Bolivia
5	Australia	50	Vietnam	95	French Polynesia	140	Swaziland
6	Canada	51	Indonesia	96	Jordan	141	Benin
7	India	52	Qatar	97	Algeria	142	Nicaragua
8	Italy	53	Macau	98	Panama	143	Mongolia
9	France	54	Puerto Rico	99	Kuwait	144	Mozambique
10	Brazil	55	Philippines	100	Brunei	145	Greenland
11	Switzerland	56	Estonia	101	Bulgaria	146	Guinea-Bissau
12	China	57	Ukraine	102	Angola	147	Fiji
13	Netherlands	58	Thailand	103	Trinidad and Tobago	148	Togo
14	Japan	59	Ghana	104	Georgia	149	Monaco
15	Portugal	60	Luxembourg	105	Papua New Guinea	150	Guinea
16	Sweden	61	Kenya	106	Zambia	151	Albania
17	Singapore	62	Venezuela	107	Cuba	152	Curacao
18	Denmark	63	Cyprus	108	Malta	153	Niger
19	South Korea	64	Iceland	109	Bahrain	154	Macedonia
20	Belgium	65	Romania	110	Azerbaijan	155	Antarctica
21	Colombia	66	Ethiopia	111	Latvia	156	Tonga
22	Ireland	67	Croatia	112	Guyana	157	Jamaica
23	Austria	68	Costa Rica	113	Belize	158	Yemen
24	Hong Kong	69	Slovakia	114	Samoa	159	Mali
25	Iran	70	Sri Lanka	115	Gambia	160	Lesotho
26	Mexico	71	Lebanon	116	Grenada	161	Jersey
27	Finland	72	Serbia	117	Haiti	162	Uzbekistan
28	Turkey	73	Oman	118	Cayman Islands	163	Saint Lucia
29	Saudi Arabia	74	Slovenia	119	Cambodia	164	Bosnia and Herz.
30	New Zealand	75	Iraq	120	Armenia	165	Maldives
31	South Africa	76	Uruguay	121	Burkina Faso	166	Congo [Republic]
32	Norway	77	Nepal	122	Mauritania	167	Moldova
33	Israel	78	Paraguay	123	Gabon	168	S Kitts and Nevis
34	Russia	79	Morocco	124	Guatemala	169	New Caledonia
35	Poland	80	Lithuania	125	Cape Verde	170	Dominican Republic
36	Malaysia	81	Cameroon	126	Honduras	171	Vanuatu
37	Pakistan	82	Malawi	127	Belarus	172	Kosovo
38	Czech Republic	83	Tunisia	128	Andorra	173	Timor-Leste
39	Egypt	84	Zimbabwe	129	San Marino	174	Isle of Man
40	Chile	85	Uganda	130	Cote dIvoire	175	S Vincent and the Gren.
41	Taiwan	86	Tanzania	131	El Salvador	176	Syria
42	Bangladesh	87	Kazakhstan	132	Guadeloupe	177	Kyrgyzstan
43	Peru	88	Sudan	133	Reunion	178	Sint Maarten
44	Nigeria	89	Montserrat	134	Bhutan	179	Afghanistan
45	United Arab Emirates	90	Madagascar	135	Senegal	180	Barbados

Table A.7: Ranking of the countries by **hub** score in **2020**.

A.5 Gini Index

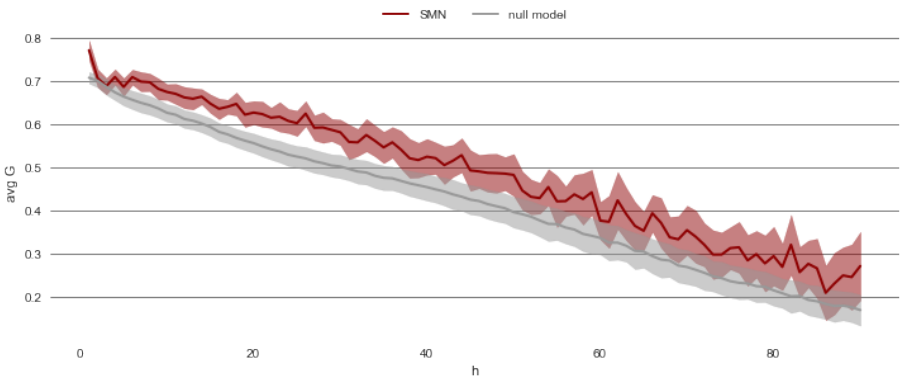


Figure A.3: Average Gini coefficient (and 95% confidence interval) as a function of the hub ranking of the scientific migration network and of the null model without self-loops. The population **W** is represented by the edge weights of outgoing edges and the average is computed over the time domain T.

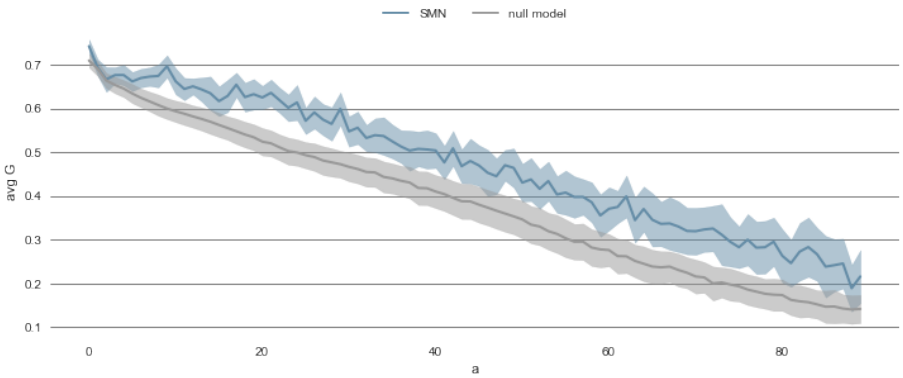


Figure A.4: Average Gini coefficient (and 95% confidence interval) as a function of the authority ranking of the scientific migration network and of the null model without self-loops. The population **W** is represented by the edge weights of outgoing edges and the average is computed over the time domain T.

A.6 Ranking by the number of returns

	1	2	3	4	5	6	7	8	9	10
1	70.0	114.0	178.0	215.0	251.0	224.0	240.0	247.0	203.0	170.0
2	66.0	113.0	160.0	213.0	242.0	220.0	218.0	205.0	176.0	140.0
3	52.0	94.0	146.0	177.0	188.0	205.0	194.0	159.0	124.0	112.0
4	50.0	81.0	135.0	156.0	144.0	158.0	172.0	144.0	122.0	108.0
5	50.0	81.0	125.0	142.0	139.0	125.0	133.0	118.0	105.0	84.0
6	45.0	79.0	111.0	138.0	133.0	124.0	129.0	115.0	97.0	81.0
7	36.0	78.0	100.0	132.0	130.0	122.0	110.0	99.0	88.0	73.0
8	34.0	71.0	86.0	126.0	121.0	118.0	95.0	94.0	87.0	64.0
9	29.0	63.0	76.0	82.0	104.0	90.0	90.0	79.0	83.0	62.0
10	27.0	39.0	76.0	79.0	93.0	83.0	85.0	78.0	77.0	58.0
11	26.0	38.0	64.0	77.0	79.0	80.0	76.0	73.0	74.0	57.0
12	19.0	36.0	53.0	56.0	77.0	71.0	70.0	72.0	62.0	57.0
13	19.0	32.0	44.0	55.0	55.0	64.0	63.0	65.0	58.0	47.0
14	18.0	29.0	44.0	55.0	53.0	64.0	52.0	55.0	54.0	45.0
15	16.0	24.0	43.0	47.0	47.0	58.0	49.0	49.0	37.0	38.0
16	11.0	23.0	42.0	46.0	47.0	55.0	43.0	36.0	33.0	32.0
17	11.0	22.0	41.0	45.0	40.0	55.0	38.0	34.0	32.0	28.0
18	10.0	22.0	40.0	39.0	39.0	45.0	37.0	30.0	30.0	28.0
19	9.0	21.0	28.0	37.0	38.0	45.0	35.0	30.0	29.0	26.0
20	9.0	20.0	28.0	33.0	35.0	37.0	33.0	30.0	27.0	21.0

Table A.8: Ranking by number of returns and time difference δ_t.

	1	2	3	4	5	6	7	8	9	10
1	0.0417	0.0667	0.0476	0.0476	0.1429	0.0435	0.0417	0.0417	0.0769	0.0233
2	0.0294	0.0488	0.0370	0.0412	0.0500	0.0345	0.0417	0.0357	0.0412	0.0172
3	0.0244	0.0156	0.0345	0.0339	0.0444	0.0294	0.0345	0.0238	0.0362	0.0169
4	0.0103	0.0156	0.0309	0.0337	0.0270	0.0286	0.0290	0.0234	0.0159	0.0166
5	0.0098	0.0140	0.0270	0.0263	0.0233	0.0274	0.0263	0.0213	0.0156	0.0161
6	0.0093	0.0122	0.0263	0.0244	0.0225	0.0250	0.0256	0.0204	0.0154	0.0141
7	0.0074	0.0119	0.0250	0.0229	0.0206	0.0238	0.0238	0.0196	0.0153	0.0140
8	0.0070	0.0118	0.0208	0.0217	0.0204	0.0200	0.0238	0.0154	0.0149	0.0132
9	0.0070	0.0113	0.0200	0.0206	0.0204	0.0192	0.0227	0.0148	0.0149	0.0130
10	0.0065	0.0111	0.0192	0.0204	0.0196	0.0182	0.0222	0.0146	0.0140	0.0118
11	0.0064	0.0103	0.0177	0.0204	0.0194	0.0169	0.0207	0.0137	0.0137	0.0113
12	0.0060	0.0103	0.0165	0.0185	0.0192	0.0169	0.0206	0.0132	0.0135	0.0112
13	0.0059	0.0103	0.0158	0.0177	0.0187	0.0167	0.0206	0.0125	0.0130	0.0110
14	0.0047	0.0100	0.0149	0.0176	0.0185	0.0163	0.0187	0.0118	0.0125	0.0106
15	0.0046	0.0098	0.0148	0.0175	0.0183	0.0159	0.0185	0.0116	0.0119	0.0105
16	0.0045	0.0093	0.0145	0.0175	0.0180	0.0158	0.0183	0.0113	0.0117	0.0105
17	0.0043	0.0093	0.0125	0.0172	0.0177	0.0154	0.0174	0.0111	0.0112	0.0097
18	0.0041	0.0092	0.0124	0.0157	0.0177	0.0151	0.0169	0.0109	0.0111	0.0097
19	0.0037	0.0092	0.0121	0.0143	0.0175	0.0145	0.0169	0.0108	0.0111	0.0087
20	0.0036	0.0088	0.0111	0.0143	0.0174	0.0143	0.0168	0.0106	0.0108	0.0087

Table A.9: Values of the Normalized Return Index 4.5 corresponding to Table 4.9.

Appendix B

B.1 Engagement Layer Network

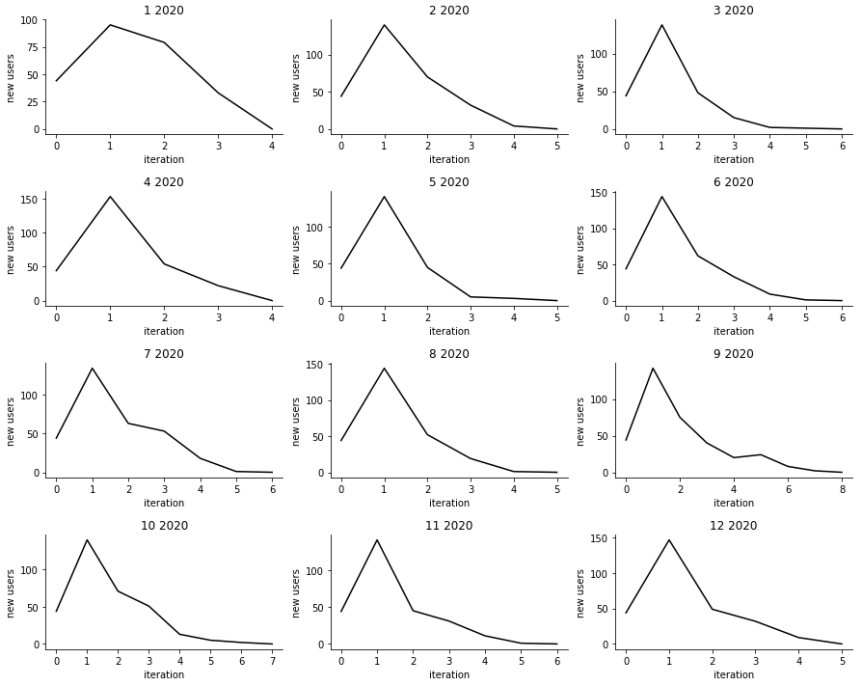


Figure B.1: Convergences in new user gathering of the Engagement Layer Network pipeline describe in Algorithm 11, given the experiment [1] from Table 5.2, for every month of 2020.

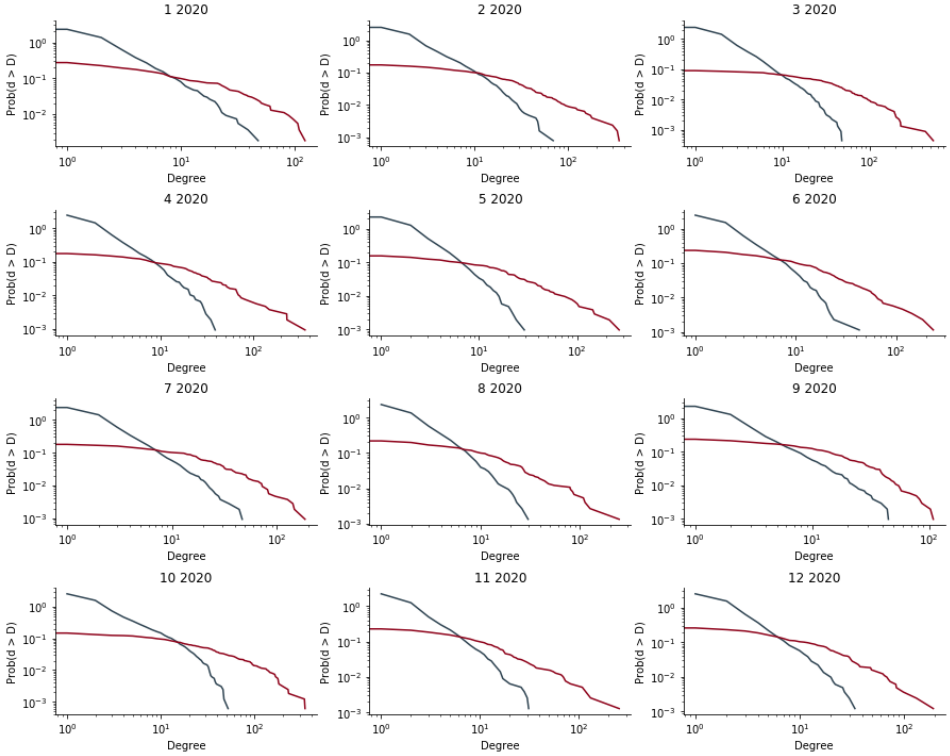


Figure B.2: Degree distribution of the Engagement Layer Network pipeline describe in Algorithm 11, given the experiment [1] from Table 5.2, for every month of 2020. Red line is for the out-degree, blue line for the in-degree.

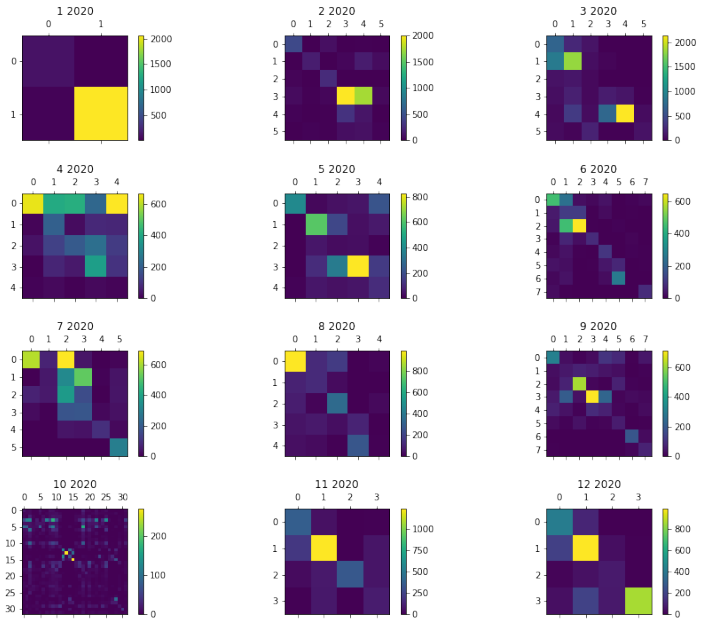


Figure B.3: Blocks of the Engagement Layer Networks given the experiment [1] 5.2, for every month of 2020.

Bibliography

[1] Andrea Baronchelli. The emergence of consensus: a primer. Royal Society open science, 5(2):172189, 2018.

[2] Alessandro Vespignani. Predicting the behavior of techno-social systems. Science, 325(5939):425–428, 2009.

[3] Giorgio Fagiolo and Marina Mastrorillo. International migration network: Topology and modeling. Physical Review E, 88(1):012812, 2013.

[4] Alexander Belyi, Iva Bojic, Stanislav Sobolevsky, Izabela Sitko, Bartosz Hawelka, Lada Rudikova, Alexander Kurbatski, and Carlo Ratti. Global multi-layer network of human mobility. International Journal of Geographical Information Science, 31(7):1381–1402, 2017.

[5] Laura Alessandretti, Ulf Aslak, and Sune Lehmann. The scales of human mobility. Nature, 587(7834):402–407, 2020.

[6] Chiara Franzoni, Giuseppe Scellato, and Paula Stephan. Foreign-born scientists: mobility patterns for 16 countries. Nature Biotechnology, 30(12):1250, 2012.

[7] Henk F Moed, Andrew Plume, et al. Studying scientific migration in scopus. Scientometrics, 94(3):929–942, 2013.

[8] Luca Verginer and Massimo Riccaboni. Brain-circulation network: The global mobility of the life scientists. Working Papers 10/2018, IMT School for Advanced Studies Lucca, 2018.

[9] James E Anderson. The gravity model. Annu. Rev. Econ., 3(1):133–160, 2011.

[10] Caleb Robinson and Bistra Dilkina. A machine learning approach to modeling human migration. In Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies, page 30. ACM, 2018.

[11] Massimiliano Luca, Gianni Barlacchi, Bruno Lepri, and Luca Pappalardo. A survey on deep learning for human mobility. ACM Computing Surveys (CSUR), 55(1):1–44, 2021.

[12] Riccardo Guidotti, Anna Monreale, Salvatore Rinzivillo, Dino Pedreschi, and Fosca Giannotti. Unveiling mobility complexity through complex network analysis. Social Network Analysis and Mining, 6(1):59, 2016.

[13] Luca Pappalardo, Filippo Simini, Salvatore Rinzivillo, Dino Pedreschi, Fosca Giannotti, and Albert-László Barabási. Returners and explorers dichotomy in human mobility. Nature communications, 6(1):1–8, 2015.

[14] David Pastor-Escuredo and Enrique Frias-Martinez. Flow descriptors of human mobility networks. arXiv preprint arXiv:2003.07279, 2020.

[15] Roy Cerqueti, Gian Paolo Clemente, and Rosanna Grassi. A network-based measure of the socio-economic roots of the migration flows. Social Indicators Research, pages 1–18, 2018.

[16] Aldo Geuna. Global mobility of research scientists: The economics of who goes where and why. Academic Press, Cambridge, Massachusetts, 2015.

[17] Luca Pappalardo, Dino Pedreschi, Zbigniew Smoreda, and Fosca Giannotti. Using big data to study the link between human mobility and socio-economic development. In 2015 IEEE International Conference on Big Data (Big Data), pages 871–878. IEEE, 2015.

[18] Giorgio Fagiolo and Gianluca Santoni. Human-mobility networks, country income, and labor productivity. Network Science, 3(3):377–407, 2015.

[19] AnnaLee Saxenian. From brain drain to brain circulation: Transnational communities and regional upgrading in india and china. Studies in comparative international development, 40(2):35–61, 2005.

[20] Ajay Agrawal, Devesh Kapur, John McHale, and Alexander Oettl. Brain drain or brain bank? the impact of skilled emigration on poor-country innovation. Journal of Urban Economics, 69(1):43–55, 2011.

[21] Alexander Subbotin and Samin Aref. Brain drain and brain gain in russia: Analyzing international migration of researchers by discipline using scopus bibliometric data 1996–2020. Scientometrics, 126(9):7875–7900, 2021.

[22] Eun Lee, Aaron Clauset, and Daniel B Larremore. The dynamics of faculty hiring networks. EPJ Data Science, 10(1):48, 2021.

[23] Khairunnisa Ibrahim, Samuel Khodursky, and Taha Yasseri. Gender imbalance and spatiotemporal patterns of contributions to citizen science projects: the case of zooniverse. Frontiers in Physics, 9:650720, 2021.

[24] Lu Liu, Nima Dehmamy, Jillian Chown, C Lee Giles, and Dashun Wang. Understanding the onset of hot streaks across artistic, cultural, and scientific careers. Nature communications, 12(1):1–10, 2021.

[25] Dany Bahar and Hillel Rapoport. Migration, knowledge diffusion and the comparative advantage of nations. The Economic Journal, 128(612):F273–F305, 2018.

[26] Ricardo Hausmann and Ljubica Nedelkoska. Welcome home in a crisis: Effects of return migration on the non-migrants' wages and employment. European Economic Review, 101:101–132, 2018.

[27] John Fitzgerald, Sanna Ojanperä, and Neave O'Clery. Is academia becoming more localised? the growth of regional knowledge networks within international research collaboration. Applied Network Science, 6(1):1–27, 2021.

[28] Chakresh Kumar Singh, Emma Barme, Robert Ward, Liubov Tupikina, and Marc Santolini. Quantifying the rise and fall of scientific fields. PloS one, 17(6):e0270131, 2022.

[29] Puyu Yang and Giovanni Colavizza. A map of science in wikipedia. In Companion Proceedings of the Web Conference 2022, pages 1289–1300, 2022.

[30] Allon Wagner, Yuval Levavi, Siram Kedar, Kathleen Abraham, Yoram Cohen, and Ran Zadok. Quantitative social network analysis (sna) and the study of cuneiform archives: A test-case based on the murašû archive. Akkadica, 134(2):117–134, 2013.

[31] Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. From Louvain to Leiden: guaranteeing well-connected communities. Scientific reports, 9(1):1–12, 2019.

[32] Tero Alstola, Shana Zaia, Aleksi Sahala, Heidi Jauhiainen, Saana Svärd, and Krister Lindén. Aššur and his friends: a statistical analysis of neo-assyrian texts. Journal of Cuneiform Studies, 71(1):159–180, 2019.

[33] Hilde De Weerdt, Brent Ho, Allon Wagner, Jiyan Qiao, and Mingkin Chu. Is there a faction in this list? Journal of Chinese History , 4(2):347–389, 2020.

[34] Eero Hyvönen, Petri Leskinen, Minna Tamper, Heikki Rantala, Esko Ikkala, Jouni Tuominen, and Kirsi Keravuori. Linked data: A paradigm shift for publishing and using biography collections on the semantic web. In Proceedings of the Third Conference on Biographical Data in a Digital World (BD 2019). CEUR-WS. org, 2022.

[35] Paola Ronzino. Harmonizing the crmba and crmarchaeo models. International Journal on Digital Libraries, 18(4):253–261, 2017.

[36] Michele Pasin and John Bradley. Factoid-based prosopography and computer ontologies: towards an integrated approach. Digital Scholarship in the Humanities, 30(1):86–97, 2015.

[37] Aming Li, Sean P Cornelius, Y-Y Liu, Long Wang, and A-L Barabási. The fundamental advantages of temporal networks. Science, 358(6366):1042–1046, 2017.

[38] Mark Newman, Albert-Laszlo Barabasi, and Duncan J Watts. The structure and dynamics of networks, volume 12. Princeton University Press, 2011.

[39] Tao Zhou, Jie Ren, Matúš Medo, and Yi-Cheng Zhang. Bipartite network projection and personal recommendation. Physical review E, 76(4):046115, 2007.

[40] Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P Gleeson, Yamir Moreno, and Mason A Porter. Multilayer networks. Journal of complex networks, 2(3):203–271, 2014.

[41] Tom Brughmans and Matthew A Peeples. Network Science in Archaeology. Cambridge University Press, 2022.

[42] Tom Brughmans, Anna Collar, and Fiona Coward. The connected past: challenges to network studies in archaeology and history. Oxford University Press, 2016.

[43] David Bamman, Adam Anderson, and Noah A Smith. Inferring social rank in an old assyrian trade network. arXiv preprint arXiv:1303.2873, 2013.

[44] Jasminko Novak, Isabel Micheel, Mark Melenhorst, Lars Wieneke, Marten Düring, Javier Garcia Morón, Chiara Pasini, Marco Tagliasacchi, and Piero Fraternali. Histograph – a visualization tool for collaborative analysis of networks from historical social multimedia collections. In 2014 18th International Conference on Information Visualisation, pages 241–250, 2014.

[45] Stefan Bornhofen and Marten Düring. Exploring dynamic multilayer graphs for digital humanities. Applied Network Science, 5(1):1–13, 2020.

[46] Ronald L Breiger and Philippa E Pattison. Cumulated social roles: The duality of persons and their algebras. Social networks, 8(3):215–256, 1986.

[47] Koji Mizoguchi. Nodes and edges: a network approach to hierarchisation and state formation in japan. Journal of Anthropological Archaeology, 28(1):14–26, 2009.

[48] Maximilian Schich, Chaoming Song, Yong-Yeol Ahn, Alexander Mirsky, Mauro Martino, Albert-László Barabási, and Dirk Helbing. A network framework of cultural history. science, 345(6196):558–562, 2014.

[49] Pedro Ribeiro and Fernando Silva. Discovering colored network motifs. In Complex networks V, pages 107–118. Springer, 2014.

[50] Jake Lever and Russ B Altman. Analyzing the vast coronavirus literature with coronacentral. Proceedings of the National Academy of Sciences, 118(23):e2100766118, 2021.

[51] Yian Yin, Jian Gao, Benjamin F Jones, and Dashun Wang. Coevolution of policy and science during the pandemic. Science, 371(6525):128–130, 2021.

[52] Massimo Stella, Michael S Vitevitch, and Federico Botta. Cognitive networks identify the content of english and italian popular posts about covid-19 vaccines: Anticipation, logistics, conspiracy and loss of trust. arXiv preprint arXiv:2103.15909, 2021.

[53] John Bollenbacher, Diogo Pacheco, Pik-Mai Hui, Yong-Yeol Ahn, Alessandro Flammini, and Filippo Menczer. On the challenges of predicting microscopic dynamics of online conversations. Applied Network Science, 6(1):1–21, 2021.

[54] Martino Trevisan, Luca Vassio, and Danilo Giordano. Debate on online social networks at the time of covid-19: An italian case study. Online Social Networks and Media, 23:100136, 2021.

[55] Kyriaki Kalimeri, Mariano G. Beiró, Alessandra Urbinati, Andrea Bonanomi, Alessandro Rosina, and Ciro Cattuto. Human values and attitudes towards vaccination in social media. In Companion Proceedings of The 2019 World Wide Web Conference, pages 248–254, 2019.

[56] Neil F Johnson, Nicolas Velásquez, Nicholas Johnson Restrepo, Rhys Leahy, Nicholas Gabriel, Sara El Oud, Minzhang Zheng, Pedro Manrique, Stefan Wuchty, and Yonatan Lupu. The online competition between pro-and anti-vaccination views. Nature, pages 1–4, 2020.

[57] Jacopo Lenti and Giancarlo Ruffo. Ensemble of opinion dynamics models to understand the role of the undecided about vaccines. Journal of Complex Networks, 10(3):cnac018, 2022.

[58] Hanjia Lyu, Junda Wang, Wei Wu, Viet Duong, Xiyang Zhang, Timothy D Dye, and Jiebo Luo. Social media study of public opinions on potential covid-19 vaccines: informing dissent, disparities, and dissemination. Intelligent medicine, 2(01):1–12, 2022.

[59] Alfonso Semeraro, Salvatore Vilella, Giancarlo Ruffo, and Massimo Stella. Emotional profiling and cognitive networks unravel how mainstream and alternative press framed astrazeneca, pfizer and covid-19 vaccination campaigns. Scientific reports, 12(1):1–12, 2022.

[60] Emilio Ferrara. What types of covid-19 conspiracies are populated by twitter bots? First Monday, 2020.

[61] Kai-Cheng Yang, Emilio Ferrara, and Filippo Menczer. Botometer 101: Social bot practicum for computational social scientists. arXiv preprint arXiv:2201.01608, 2022.

[62] Emily Chen, Kristina Lerman, and Emilio Ferrara. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. JMIR Public Health and Surveillance, 6(2):e19273, 2020.

[63] Anatoliy Gruzd and Philip Mai. Going viral: How a single tweet spawned a covid-19 conspiracy theory on twitter. Big Data & Society, 7(2):2053951720938405, 2020.

[64] Wasim Ahmed, Josep Vidal-Alaball, Joseph Downing, and Francesc López Seguí. Covid-19 and the 5g conspiracy theory: social network analysis of twitter data. Journal of Medical Internet Research, 22(5):e19458, 2020.

[65] Julie Jiang, Emily Chen, Shen Yan, Kristina Lerman, and Emilio Ferrara. Political polarization drives online conversations about covid-19 in the united states. Human Behavior and Emerging Technologies, 2(3):200–211, 2020.

[66] Hao Sha, Mohammad Al Hasan, George Mohler, and P Jeffrey Brantingham. Dynamic topic modeling of the covid-19 twitter narrative among us governors and cabinet executives. arXiv preprint arXiv:2004.11692, 2020.

[67] Jon Green, Jared Edgerton, Daniel Naftel, Kelsey Shoub, and Skyler J Cranmer. Elusive consensus: Polarization in elite communication on the covid-19 pandemic. Science Advances, 6(28):eabc2717, 2020.

[68] Kristina Gligorić, Manoel Horta Ribeiro, Martin Müller, Olesia Altunina, Maxime Peyrard, Marcel Salathé, Giovanni Colavizza, and Robert

West. Experts and authorities receive disproportionate attention on twitter during the covid-19 crisis. arXiv preprint arXiv:2008.08364, 2020.

[69] Ryan J Gallagher, Larissa Doroshenko, Sarah Shugars, David Lazer, and Brooke Foucault Welles. Sustained online amplification of covid-19 elites in the united states. arXiv preprint arXiv:2009.07255, 2020.

[70] Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. The covid-19 social media infodemic. arXiv preprint arXiv:2003.05004, 2020.

[71] Pierluigi Sacco Manlio De Domenico. Covid19 Infodemics Observatory. https://covid19obs.fbk.eu/#/, 2020. [Online; accessed 27-August-2020].

[72] Riccardo Gallotti, Francesco Valle, Nicola Castaldo, Pierluigi Sacco, and Manlio De Domenico. Assessing the risks of" infodemics" in response to covid-19 epidemics. arXiv preprint arXiv:2004.03997, 2020.

[73] Martin Müller, Marcel Salathé, and Per E Kummervold. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. arXiv preprint arXiv:2005.07503, 2020.

[74] Catherine Ordun, Sanjay Purushotham, and Edward Raff. Exploratory analysis of covid-19 tweets using topic modeling, umap, and digraphs. arXiv preprint arXiv:2005.03082, 2020.

[75] Han Woo Park, Sejung Park, and Miyoung Chong. Conversations and medical news frames on twitter: Infodemiological study on covid-19 in south korea. Journal of Medical Internet Research, 22(5):e18897, 2020.

[76] Philipp Wicke and Marianna M Bolognesi. Framing covid-19: How we conceptualize and discuss the pandemic on twitter. arXiv preprint arXiv:2004.06986, 2020.

[77] Long Chen, Hanjia Lyu, Tongyu Yang, Yu Wang, and Jiebo Luo. In the eyes of the beholder: Sentiment and topic analyses on social media use of neutral and controversial terms for covid-19. arXiv preprint arXiv:2004.10225, 2020.

[78] Leonard Schild, Chen Ling, Jeremy Blackburn, Gianluca Stringhini, Yang Zhang, and Savvas Zannettou. " go eat a bat, chang!": An early look on the emergence of sinophobic behavior on web communities in the face of covid-19. arXiv preprint arXiv:2004.04046, 2020.

[79] Jay J Van Bavel, Katherine Baicker, Paulo S Boggio, Valerio Capraro, Aleksandra Cichocka, Mina Cikara, Molly J Crockett, Alia J Crum, Karen M Douglas, James N Druckman, et al. Using social and behavioural science to support covid-19 pandemic response. Nature Human Behaviour, pages 1–12, 2020.

[80] Albert-László Barabási et al. Network science, chapter 1. Cambridge university press, 2016.

[81] Michele Coscia. The atlas for the aspiring network scientist. arXiv preprint arXiv:2101.00863, 2021.

[82] Gary William Flake, Steve Lawrence, C Lee Giles, and Frans M Coetzee. Self-organization and identification of web communities. Computer, (3):66–71, 2002.

[83] Mark EJ Newman. Modularity and community structure in networks. Proceedings of the national academy of sciences, 103(23):8577–8582, 2006.

[84] David Easley, Jon Kleinberg, et al. Networks, crowds, and markets, volume 8, chapter 3. Cambridge university press Cambridge, 2010.

[85] Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. Virality prediction and community structure in social networks. Scientific reports, 3:2522, 2013.

[86] Elizabeth E Bruch and MEJ Newman. Aspirational pursuit of mates in online dating markets. Science Advances, 4(8):eaap9815, 2018.

[87] Johan Bollen, Bruno Gonçalves, Ingrid van de Leemput, and Guangchen Ruan. The happiness paradox: your friends are happier than you. EPJ Data Science, 6(1):4, 2017.

[88] Walter Quattrociocchi, Antonio Scala, and Cass R Sunstein. Echo chambers on facebook. Available at SSRN 2795110, 2016.

[89] Alberto Aleta and Yamir Moreno. Multilayer networks in a nutshell. Annual Review of Condensed Matter Physics, 2018.

[90] Martina Contisciani, Eleanor A Power, and Caterina De Bacco. Community detection with node attributes in multilayer networks. Scientific reports, 10(1):1–16, 2020.

[91] Michele Coscia and Michael Szell. Multiplex graph association rules for link prediction. arXiv preprint arXiv:2008.08351, page 27, 2020.

[92] Manlio De Domenico, Albert Solé-Ribalta, Elisa Omodei, Sergio Gómez, and Alex Arenas. Centrality in interconnected multilayer networks. arXiv preprint arXiv:1311.2906, 2013.

[93] Michele Coscia. Generalized euclidean measure to estimate distances on multilayer networks. ACM Transactions on Knowledge Discovery from Data (TKDD), 16(6):1–22, 2022.

[94] Petter Holme and Jari Saramäki. A map of approaches to temporal networks. In Temporal Network Theory, pages 1–24. Springer, 2019.

[95] Petter Holme and Jari Saramäki. Temporal networks. Physics reports, 519(3):97–125, 2012.

[96] Mikko Kivelä, Jordan Cambe, Jari Saramäki, and Márton Karsai. Mapping temporal-network percolation to weighted, static event graphs. Scientific reports, 8(1):1–9, 2018.

[97] Paolo Boldi and Sebastiano Vigna. Axioms for centrality. Internet Mathematics, 10(3-4):222–262, 2014.

[98] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. Computer networks and ISDN systems, 30(1-7):107–117, 1998.

[99] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[100] Jon M Kleinberg. Hubs, authorities, and communities. ACM computing surveys (CSUR), 31(4es):5, 1999.

[101] Linton C Freeman, Douglas Roeder, and Robert R Mulholland. Centrality in social networks: Ii. experimental results. Social networks, 2(2):119–141, 1979.

[102] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. Reviews of modern physics, 74(1):47, 2002.

[103] Tiago P Peixoto. Bayesian stochastic blockmodeling. Advances in network clustering and blockmodeling, pages 289–332, 2019.

[104] KR1442 Chowdhary. Natural language processing. Fundamentals of artificial intelligence, pages 603–649, 2020.

[105] Wahab Khan, Ali Daud, Jamal A Nasir, and Tehmina Amjad. A survey on the state-of-the-art machine learning models in the context of nlp. Kuwait journal of Science, 43(4), 2016.

[106] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[107] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9, 2019.

[108] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020.

[109] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big?. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pages 610–623, 2021.

[110] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. Transactions of the association for computational linguistics, 5:135–146, 2017.

[111] Basile Valerio, Lai Mirko, and Sanguinetti Manuela. Long-term social media data collection at the university of turin. In Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), pages 1–6. CEUR-WS, 2018.

[112] OECD. A profile of immigrant populations in the 21st century: data from OECD countries. OECD Paris, Paris, France, 2008.

[113] Jeni Klugman. Human development report 2009. overcoming barriers: Human mobility and development, 2009.

[114] Salvatore Rinzivillo, Simone Mainardi, Fabio Pezzoni, Michele Coscia, Dino Pedreschi, and Fosca Giannotti. Discovering the geographical borders of human mobility. KI-Künstliche Intelligenz, 26(3):253–260, 2012.

[115] Fabio Schiantarelli. Global economic prospects 2006: economic implications of remittances and migration. The World Bank, 2005.

[116] Emanuele Pugliese, Giulio Cimini, Aurelio Patelli, Andrea Zaccaria, Luciano Pietronero, and Andrea Gabrielli. Unfolding the innovation system for the development of countries: co-evolution of science, technology and production. Scientific reports, 9:16440, 2019.

[117] Pierre Deville, Dashun Wang, Roberta Sinatra, Chaoming Song, Vincent D Blondel, and Albert-László Barabási. Career on the move: Geography, stratification, and scientific impact. Scientific reports, 4:4770, 2014.

[118] Alessandra Urbinati, Edoardo Galimberti, and Giancarlo Ruffo. Measuring scientific brain drain with hubs and authorities: A dual perspective. Online Social Networks and Media, 26:100176, 2021.

[119] April Tash. Global standards now exist for a healthy ecosystem of research and innovation. UNESCO Science Report: The race against time for smarter development, 2021:24, 2021.

[120] J Bohannon and K Doran. Data from: Introducing orcid, 2017.

[121] John Bohannon and Kirk Doran. Introducing orcid. Science, 356(6339):691–692, 2017.

[122] Mark EJ Newman. The structure and function of complex networks. SIAM review, 45(2):167–256, 2003.

[123] M Ángeles Serrano and Marián Boguñá. Weighted configuration model. In AIP conference proceedings, volume 776, pages 101–107. American Institute of Physics, 2005.

[124] M. Ángeles Serrano, Marián Boguñá, and Alessandro Vespignani. Extracting the multiscale backbone of complex weighted networks. Proceedings of the National Academy of Sciences, 106(16):6483–6488, 2009.

[125] Gerardo Iñiguez, Carlos Pineda, Carlos Gershenson, and Albert-László Barabási. Universal dynamics of ranking. arXiv preprint arXiv:2104.13439, 2021.

[126] Alon Altman and Moshe Tennenholtz. Ranking systems: The pagerank axioms. In Proceedings of the 6th ACM Conference on Electronic Commerce, EC '05, page 1–8, New York, NY, USA, 2005. Association for Computing Machinery.

[127] Corrado Gini. Variabilità e mutabilità. Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi, 1912.

[128] David Easley and Jon Kleinberg. Networks, crowds, and markets: Reasoning about a highly connected world. Cambridge university press, 2010.

[129] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment, 2008(10):P10008, 2008.

[130] Andrea Tacchella, Matthieu Cristelli, Guido Caldarelli, Andrea Gabrielli, and Luciano Pietronero. A new metrics for countries' fitness and products' complexity. Scientific reports, 2(1):1–7, 2012.

[131] Ana Fernández-Zubieta, Aldo Geuna, and Cornelia Lawson. Productivity pay-offs from academic mobility: should i stay or should i go? Industrial and Corporate Change, 25(1):91–114, 2016.

[132] Richard Van Noorden. Global mobility: Science on the move. Nature News, 490(7420):326, 2012.

[133] Farshad Kooti, Winter A Mason, Krishna P Gummadi, and Meeyoung Cha. Predicting emerging social conventions in online social networks. In Proceedings of the 21st ACM international conference on Information and knowledge management, pages 445–454, 2012.

[134] Danah Boyd, Scott Golder, and Gilad Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In 2010 43rd Hawaii International Conference on System Sciences, pages 1–10. IEEE, 2010.

[135] Censis. 17° rapporto sulla comunicazione. 2021.

[136] Simon Kemp. Digital 2020: Italy. by Hootstat and We Are Social, Datareportal.com, 2019.

[137] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. Modern information retrieval, volume 463. ACM press New York, 1999.

[138] Haewoon Kwak, Jisun An, Elise Jing, and Yong-Yeol Ahn. Frameaxis: characterizing microframe bias and intensity with word embedding. PeerJ Computer Science, 7:e644, 2021.

[139] Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. Semaxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment. arXiv preprint arXiv:1806.05521, 2018.

[140] Elise Jing and Yong-Yeol Ahn. Characterizing partisan political narrative frameworks about covid-19 on twitter. EPJ data science, 10(1):53, 2021.

[141] Alessandra Urbinati, Enrico Burdisso, Thomas Edward Capozzi Lupi Arthur, Claudio Mattutino, Salvatore Vilella, Alfonso Semeraro, Giancarlo Francesco Ruffo, Carlo Corti, Stefano De Martino, Elena Devecchi, et al. Bridging representation and visualization in prosopographic

research: A case study. In CEUR WORKSHOP PROCEEDINGS, pages 80–92. Sun SITE Central Europe, Technical University of Aachen, 2022.

[142] Elena Devecchi. Middle Babylonian Texts in the Cornell Collections, Part 2. Penn State University Press, 2020.

[143] Wilfred Hugo van Soldt. Middle babylonian texts in the cornell university collections/part 1 the later kings. Middle Babylonian texts in the Cornell University collections, 30, 2015.

[144] Theo Van den Hout. Administration and writing in hittite society. In Balza, ME; Giorgieri, M.; Mora, C.(a cura di), Archivi, depositi, magazzini presso gli Ittiti. Nuovi materiali e nuove ricerche= Proceedings of the Workshop held at Pavia, pages 41–58, 2009.

[145] Leonhard Sassmannshausen. Beiträge zur Verwaltung und Gesellschaft Babyloniens in der Kassitenzeit, volume 2100. von Zabern, 2001.

[146] Herbert Petschow. Mittelbabylonische rechts-und wirtschaftsurkunden der hilprecht-sammlung jena. In Mittelbabylonische Rechts-und Wirtschaftsurkunden der Hilprecht-Sammlung Jena. De Gruyter, 2022.