

**CORPORATE PROFILING:
TEXT ANALYTICS FOR EMPLOYABILITY SKILLS MINING**

A DOCTORAL THESIS BY

FRANCESCO SMALDONE



UNIVERSITY OF SALERNO



DEPARTMENT
OF BUSINESS SCIENCE
MANAGEMENT
& INNOVATION SYSTEMS

PHD in BIG DATA MANAGEMENT
XXXIII Cycle

DOCTORAL THESIS

CORPORATE PROFILING:
TEXT ANALYTICS FOR EMPLOYABILITY SKILLS MINING

BY

FRANCESCO SMALDONE

SUPERVISOR

PROF. MARCO PELLICANO

PHD COORDINATOR

PROF. VALERIO ANTONELLI

Firmato digitalmente ex dlgs 82/2005

ACADEMIC YEAR 2019/2020

CORPORATE PROFILING: TEXT ANALYTICS FOR EMPLOYABILITY SKILLS MINING

Francesco Smaldone

Abstract: In the current scenario, the increasing attention to big data realities is causing firms to develop new tools for corporate management, touching several realities from the marketing compartment to the HR management. In the context of today's rapid technological development and its application in a growing array of fields, the role of big data is simultaneously evolving. The present doctoral thesis provides insights into the current expectations of employers seeking to hire individuals. Profiling was conducted by harvesting relevant data from job ads published in a US employment website, which currently attracts the US market's highest recruitment traffic. This research aims to identify the skills, experience, and qualifications sought by employers in several industries and for several professionals, also indicating to candidates the tangible parameters that would increase their employability in such a role.

Keywords: Big Data Analytics, Corporate Profiling, Employability, Labor Market, Market Analytics and Metrics, Skills, Text Analytics, Text Mining.

immersus emergo

Index

PREFACE	6
PREMISE	7
1. INTRODUCTION	9
1.1 STATEMENT OF THE PROBLEM	9
1.2 RESEARCH QUESTION AND OBJECTIVES	13
1.5 SOFTWARE, TOOLS, AND ALGORITHMS FOR METHODOLOGY	19
2. THEORETICAL BACKGROUND	23
2.1 EMPLOYABILITY SKILLS FOR EMERGING PROFESSIONALS: A LITERATURE REVIEW	23
2.2 TRANSVERSAL CAPABILITIES TO REPAIR THE SKILLS MISMATCH	26
2.3 HARD & SOFT SKILLS ENHANCING CANDIDATES' EMPLOYABILITY	45
2.4 LITERATURE GAP.....	48
3. METHODS, EXTRACTION TECHNIQUES, AND PROFILING TECHNIQUES	49
3.1 BIG DATA ANALYTICS TO A NEW APPROACH TO THE APPLIED RESEARCH REGARDING TRANSVERSAL COMPETENCIES.....	49
3.2 TEXT ANALYTICS TO ANALYZE AND INTERPRETATE BIG DATA	79
3.3 GRAPH THEORIES TO VISUALIZE AND CONSTRUE EMPLOYABILITY SKILLS	99
3.4 DATA EXTRACTION TECHNIQUES AND ARCHITECTURE: DATA COLLECTION AND SAMPLING	113
3.5 SKILLS PROFILING TECHNIQUES AND ARCHITECTURE FOR DATA TREATMENT.....	119
4. RESULTS	133
4.1 DATA COLLECTION AND SAMPLING	133
4.2 RESULTS FROM THE PROFILING PROCESS	138
4.2.1 <i>MARKETING</i>	138
4.2.2 <i>ACCOUNTING & FINANCE</i>	154
4.2.3 <i>DATA SCIENCE</i>	169
4.2.4 <i>BIOINFORMATICS</i>	184
4.2.5 <i>SOFTWARE ENGINEERING & CLOUD COMPUTING (SECC)</i>	199
4.2.6 <i>TOURISM MANAGEMENT</i>	214
4.2.7 <i>PSYCHOLOGY</i>	229
4.2.8 <i>LAW</i>	244
5. DISCUSSION	259
6. RESEARCH IMPLICATION	304
6.1 DATA SCIENCE IMPLICATIONS	304
6.2 MANAGERIAL IMPLICATIONS	305
6.3 MARKETING IMPLICATIONS.....	311
7. CONCLUSIONS	313
7.1 LIMITS & EARLY REMARKS	315
REFERENCES	317

Preface

“Power, time, gravity, love.

The forces that really kick ass are all invisible.

And my dreams are the single unpredictable factor in my zoned days and nights.

Nobody allots them or censors them.

Dreams are all I have ever truly owned.”

David Mitchell

This doctoral thesis path was perhaps one of the most adrenaline filled of my life—not only for the project that I have carried out in these three years, but also for all the knowledge that the Ph.D. program has brought me. I think there is a very specific reason why this project managed to be completed, despite everything. I certainly mark one of the reasons as the day I arrived at this University and took the entrance interview. I vividly remember it; it was the most important day of my life. At least for me, it was a dream. In that dream, only one person always believed in me, and she was not with me. How dare I miss. I had always believed in studying, believed that there was something more in those books. And it was certainly not a question of status. Studying gives you a very great gift: it makes you *free*. Education makes you rich at the very moment it allows you to think critically. And that is where freedom lies: Galileo's famous freedom, that of thought, the one you cannot close. And that is why today I feel free, because, at the end of this academic gym, I still remember how badly I wanted it. I am finally sure I really wanted something, despite every challenge, and I will never forget.

Premise

This research project was born during the development of a thesis project started in 2016 at the International University of Languages and Media (IULM) of Milan, at the beginning of the research path of the Author. The project initially focussed on the Data Science scientific area (ING-INF/05), which was the domain chosen by the Author as the specialization for his master's thesis. After starting the Ph.D. program, the Author reassessed the project to hybridize it. In fact, under the guidance of Prof. Marco Pellicano, the project started to also concern the sector of Business Administration and Management (SECS-P/08).

Basing on this consideration, this doctoral thesis aims to respond to both mentioned scientific domains. The doctoral thesis refers to data science concerning methodological aspects and the development of a decision support system and to business administration and management concerning and associated knowledge advancements. The thesis takes mainly into account the employability theory from Fugate (Fugate et al., 2004; Fugate, 2006, Fugate & Kinicki, 2008; Fugate et al., 2021), and the Author heavily considers this construct, enhancing the actuality of this theory.

In advance, the Author must thank several scientific figures who helped in the development of this project.

The Author thanks Prof. Marco Pellicano for the imparted knowledge regarding economics, management, and philosophy and for the constant support during these years. The Author also thanks Prof. Vittoria Marino for the lessons regarding the marketing field, for helping keep my interest for this discipline alive, and for having supported me during these years. The Author thanks Prof. Adelaide Ippolito for conducting joint research in the field of employability and for the constant support.

The Author invites the Reader to consider this work as a supportive instrument to Corporate Profiling, one concerning both corporate and market analysis. The developed software integrates text analytics to conduct an employability skills mining on job ads to give back to managers and marketers a brand-new decision support system ready made for the market.

The Author already tested the scientific validity of the proposed methodology through several international publications (Smaldone, 2019; Smaldone, 2020; Smaldone et al., 2021a; Smaldone et al., 2021b; Smaldone et al., 2021c).

1. Introduction

1.1 Statement of the problem

Nowadays, the continuous process of digitalization and globalization are plunging firms into a reality requiring the analysis and selection of massive data produced by individuals. Raw data produced by consumers through Web 2.0 technologies are increasingly large and digital. Therefore, data production forces several companies to hire professionals who can read and interpret this data or create new hybrid professional figures to oversee modern technologies, revolutionizing the job market towards a digital renaissance (Brown et al., 2011).

Chiefly, this overabundance of data impacts the job market and modern industries (Yin & Kaynak, 2015). It continues to extend far beyond business administration, touching other essential fields such as law (Porat & Strahilevitz, 2014; Perrons & Jensen, 2015; Custers & Uršič, 2016; Devins et al., 2017), agriculture (Sonka, 2014; O'Connor & Kelly 2017), public health (Jee & Kim, 2013; Alonso et al., 2017; Vayena et al., 2018; Galetsi et al., 2020), public governance (Kim et al., 2014; Desouza & Jacob, 2017; Fredriksson et al., 2017; Klievink et al., 2017), banking and finance (Sarlin & Marghescu, 2011; Sarlin & Peltonen, 2013; Sarlin, 2016; Trelewicz, 2017; Riikkinen et al., 2018; Pejić Bach et al., 2019), fashion (Silva et al., 2019), oil and gas (Perrons & Jensen, 2015, Sumbal et al.,

2017; Patel et al., 2020), smart cities and energy (Mohammed et al., 2019), retail (Aktas & Meng, 2017), and tourism (Xiang & Fesenmaier, 2017; Del Vecchio et al., 2018; Talón-Ballesteros et al., 2018, Li et al., 2018; Alaei et al., 2019).

The emergence of these new technologies undoubtedly modified consumers' habits and firms' standards. Consumers constantly adopt emerging digital purchasing processes, and firms provide new tools to facilitate customers in their journey.

Social Media (SM) are digital tools able to spread information worldwide and are often defined as characteristically free, simple to use, and as facilitating simultaneous communication. “Web 2.0 explains the set of all those online applications, such as SM, which allows a high level of interaction between the website and the user such as blogs, forums, chats, wikis, and media sharing platforms” (Smaldone et al., 2020; p. 19). Furthermore, SM phenomena has habituated people to hyper-connection, bringing individuals to new forms of consciousness and desires. Concerning this, SM touched several realities, such as the job market. The primary example of this phenomenon is represented by LinkedIn, an SM platform allowing the exchange of professional information and its users to search for job offerings and open positions on the market. This example from the job market is a clear expression of the bright side of social media (Kietzmann et al., 2011; Baccarella et al., 2018).

Considering the impervious implications of SM, a dark side of these online platforms emerged in the present reality (Baccarella et al., 2018; Smaldone et al., 2020), as “there are several aspects involving the dark side of SM in general and affecting indistinctly all people around the globe who belong to developed society, and who are exposed daily to these online networking tools because of emerging technologies and hyper-economic development (Durham & Kellner, 2012; Hutton and Fosdick, 2011; Schroeder, 2016; Servaes, 2007) that institute real online territories (Christensen et al., 2011).” (Smaldone et al., 2020; p. 21). The dark side of SM emerges in relation to online hate speech, fake news, misinformation, trolls, and other relevant distortions of the daily life, affecting the quality of data and the analysis’ reliability (Smaldone et al., 2020). Together with its bright side, the possible danger of social media should be considered when talking regarding digital data analytics.

With the evolution of Job Search Websites (JSW), platforms allowing the exchange of information between those who offer and those looking for work, the use and analysis of big data is a valuable tool to test to understand what characteristics each must possess. According to entrepreneurs' desiderata, individuals aspiring to a specific open position should have the hard and soft skills or the technical and transversal skills necessary to complete the requisite work. In recent years, the world of work has considerably changed:

companies require and hire more flexible personnel, adaptable to production needs and who relate to colleagues to pursue the set objectives.

Thus, the reality has emerged within which young people are looking for a functional interface every day. In this increasingly complex and global scenario, it is unrealistic to think of traditional skills and competencies; instead, it has become essential to focus on necessary, soft and transversal skills inherent to everyone but, at the same time, learnable and refined during university programs and training courses and extracurricular activities such as workshops and internships.

Necessary skills, such as knowing a second language, using IT tools, and personal computers, are increasingly requested by companies and form part of the transversal skills. There are "personal" characteristics that tend to facilitate work: taking care of one's appearance and having good dexterity and copious availability.

The job market is increasingly influenced by subtle logic, which requires more and more skills and abilities that lead the individual to a timely and effective response to daily changes. The issue of technological and digital revolution impacts the workplace, the organization, working methods, and conditions therein.

The analysis of local and global trends relating to the economy and the company allows us to identify future work and skills changes but only in a broad way not related to individuals' real competencies.

1.2 Research question and objectives

With a research focus on people, industries, and technologies, corporate profiling finds a perfect collocation in this work. Corporate profiling is defined in this doctoral thesis as that process delivering in-depth insights of the organizational structure and its technology, people, and processes through the extraction of data and implementation of data analytics.

It will be crucial to map the current skills demanded in an increasingly complex job market to provide a blueprint of the currently demanded skills. Through this process, it will be possible to identify the relationships occurring in the dataset.

Profiling activities also identifies any common or causal factors between corporations, businesses, and information technology, and more significantly, it elucidates hidden or not-so-obvious factors that would otherwise be overlooked.

These considerations have led to the investigation of the following five research questions:

RQ1: What are the most effective employability skills demanded from firms offering work in the digital labor market?

RQ2: What are the crucial thematic areas conducive to the employability skills that candidates should possess to be hired by a firm in the digital labor market?

RQ3: What are the most correlated skills in open positions offered by the firms in the digital labor market?

RQ4: What are the central relationships intercurrent in the skillset of a candidate who should be hired by the firms offering work in the digital labor market?

RQ5: What is the possible scenario of a job interview based on the retrieved skillset of candidates that firms ideally seek?

Therefore, this research aims to understand which skills are most sought after by employers and companies in eight sectors, especially when individuals and groups can find these skills in JSWs.

1.3 Organization of the thesis and research design

The research design section explains the structure and instruments used to conduct this research. Several qualitative and quantitative research methods were employed for this study. The qualitative approach was selected to find research gaps in the literature and to understand previous theoretical frameworks in the field to provide insights and considerations to discuss results in more detail. Quantitative research methods were selected to collect, organize, clean, and analyze data from a descriptive, diagnostic, cognitive, prescriptive, and strategic perspective.

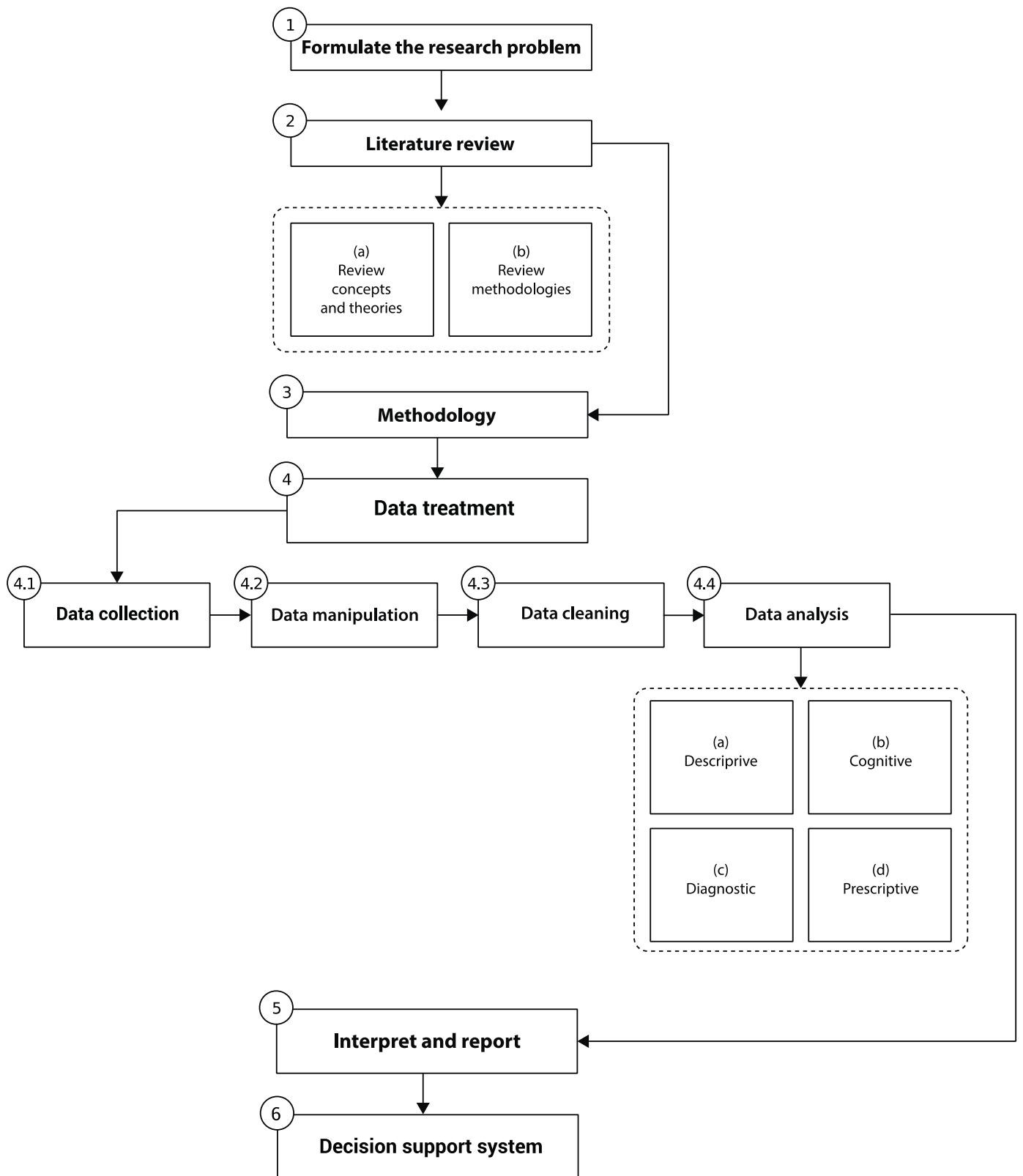
The present study is an exploratory and comparative research on eight sectors, providing results useful for both practitioners and researchers, conducted via textual analytics and comparative methods.

To mine the most required skills from the entrepreneurs, the SM platform Indeed has been interrogated via web scraping to extract the most relevant features from job ads. In **Section 2**, a systematic literature review is presented to find research gaps and theoretical frameworks. After that, job descriptions have been analyzed through several described textual analytics. Methodology and applications are explained in **Section 3**. Results are exposed in **Section 4** and discussed in **Section 5**. Conclusions and implications are presented

in **Section 6**. Early remarks, limitations, and further required research are highlighted in **Section 7**.

Because employability skills are a vast and multidisciplinary subject and the research is focused first on methodological perspectives, applications, comparisons, and advancements regarding the topic, and only secondarily on exploratory advancements of the knowledge regarding the topic that are gained from crossings through the literature review and data insights, the present work excluded other branches of research and theories—for instance, those which deal with related topics, theoretical frameworks, and theories (e.g., organizational theories, behavioral theories, value management). A blend of research methods from big data analytics was used: web scraping, text cleaning, text mining, stylometry, topic modeling, multiple correlation analysis, social network analysis, modularity detection, co-membership, clustering, partial correlations networks, and Monte Carlo Markov chains.

Fig. 1: Research design.



Source: Elaboration adapted from Kothari (2004).

1.4 Systematic literature review

This subparagraph aims to describe the Systematic Literature Review process brought to the literature overview reported in **Section 2**. The review process of academic articles, conference proceedings, and books is crucial to understanding the concepts, theories, and previous methodological advancements regarding employability skills so as to give an overview of this fluid concept in developing a conceptual framework. This review process follows the methodology proposed by vom Brocke et al. (2009).

The systematic literature review has been conducted as follows. Selected articles from the Google Scholar database were chosen (i). Thus, only relevant contributions were selected according to the aim of the research (ii). The search spanned multiple disciplines, including business sciences, statistics, data science, and software engineering research fields (iii). The most relevant contributions were selected to provide an exhaustive overview of the existing research on employability skills. Articles have been analyzed first based on the abstracts (iv), followed by Boolean search strings in titles, keywords, and abstracts (v). Unfitting contributions, duplicated loadings, off-topic research, and non-scientific publications were removed from the dataset at the sixth stage (vi). Then, the remaining articles were examined to judge whether they could contribute to the work (vii). Contributions were analyzed in depth, and full-reading analysis was conducted to reduce the literature to a proper amount for the detailed analysis of the contributions and to reveal research gaps and proper frameworks (viii).

1.5 Software, tools, and algorithms for methodology

To develop the methodology proposed in **Section 3**, two software programs were built. Python software was employed for the web scraping program with 3.9.2 release (<https://www.python.org/downloads/release/python-392/>), and an automated scraper was coded and developed via beautifulsoup and requests packages. The packages find applications in many aspects of the scraper, as requests employing the get function allow parsing HTML pages and the relative content. The output of the get function is transferred to the beautifulsoup environment. Thus, beautifulsoup provides the function find, detecting the relevant tags responding to the pre-settled parameters, as the name of the HTML tag and the relative tag's attributes (e.g., id, class, name), to mine the desired tag or HTML page from which to parse data.

Other crucial packages were re and pandas, where re stands for regular expressions. re has been employed to process text and to remove occasional special characters. Notwithstanding, pandas was utilized to build the final database in .csv format.

R software was utilized to develop a customized program script for analytics. R found various applications in this research work. After scraping data, the database was stored in external memory and then loaded for the analysis. Job descriptions were isolated and then organized, cleaned, and analyzed through R software via textual analytics. Many R packages were employed at this stage, such as stringr, a crucial part in many data cleaning and preparation tasks. The stringr package offers a cohesive set of functions studied to work

with string objects efficiently. tm was employed to conduct text mining through several functions: tm_map, gsub, inspect, Corpus, content_transformer, stripWhitespaces, bracketX, replace_number, replace_abbreviation, replace_contraction, replace_symbol. SnowballC is a powerful stemmer for text mining. qdap permits the implementation of regular expressions to clean data properly. tokenizers package allows terminological combination in document-term matrices, and stylo can be a valuable package for those who will perform stylometry on small and massive amounts of text. hlust, cluster, and mclust are packages designed to perform cluster analyses, providing a set of methods, functions, and specific libraries. topicmodels was employed to perform LDA and topic modeling. corrplot provides graphical visualizations of correlations in the model. igraph and ggplot2 were employed as graphic libraries to compute figures and outputs of the thesis. RColorBrewer is a graphic library to build palettes and to customize the output of the analyses. lsa performed greedy, spectral, and optimal modularity detection, co-membership, and assortativity measures. qqgraph was utilized to build the partial correlations network. The R version adopted for this doctoral thesis is the 4.0.5 release of this open-source software (<https://cran.r-project.org/bin/windows/base/>).

1.6 Acknowledgements on data treatment, processing, and profiling activities

The data treatment and profiling activities process was conducted according to the General Data Protection Regulation (GDPR), the current European legislation in force since 2018/25/05, and for any further modifications or integrations.

Data processing was carried out in compliance with the requirements relating to the PhD Program.

The data controller is the Author of this doctoral thesis, the undersigned Francesco Smaldone, and is the only person authorized to carry out this task, conducted in the public interest of the scientific research community and in accordance with the Art. 45 of the GDPR. Data will not be made public and/or shared with anyone in respect of the interests of the author of this doctoral thesis and his right to privacy in accordance with Arts. 8 and 46 of the GDPR, and because data content regards personal data that are not made blinded, according to the Author's desire, and are not owned by the interested part (the Reader, GDPR, Art. 63). The programs and scripts built using R and Python software are the result of the ingenuity of the Author of this doctoral thesis, and specific applications of software packages and functions requiring a process of original coding by means of the author are part of his intellectual property, as his derived moral and patrimonial rights, according to the Italian law 633 1941/22/04. These will also not be made public and/or shared with anyone, as they are useful to the private interests of the Author, who retains

the copyright as intellectual property as they are useful for the development of commercial strategies and not relevant to any public interest or for vital or social purposes. The previous concept also considers the expectations nurtured by the interested party (the Reader) based on his relationship with the data controller and owner (the Author), which is aimed exclusively at the didactic use of the research results in accordance with Arts. 45 and 46 of the GDPR. The data controller reserves the right not to keep data for the sole purpose of being able to respond to potential requests (GDPR, Art. 64) and reserves the right to keep them, delete them, transfer them to other processing or storage systems, and make them inaccessible to anyone (GDPR, Arts. 67, 68) to strengthen the right to be forgotten of potential interested parties and data creators (GDPR, Arts. 65, 66), ensuring that the processing has been carried out according to relevant legal principles (GDPR, Art. 69). According to Art. 72, profiling is subject to the provisions of GDPR, governing the processing of personal data, such as legal basis for the processing or the principles of data protection. The protection of the rights and freedoms of the data subjects is always guaranteed during the processing of data, as well as the general responsibility of the data controller, which is considered lawful, (GDPR, Art. 79) as scraping activities are legal and benign to, in this case, the economic interests of companies. The companies holding the job advertisements are never mentioned, thus respecting and protecting their corporate assets and their image.

2. Theoretical Background


The research for this thesis adopted a Systematic Literature process described in the introduction to explore the relationship between relevant skills sought by the labor market and employability. It provides the reader with a spectrum of empirical analyses conducted with the use of several methods. Starting with an analysis of the literature on employability, human capital, and the change supported by SM platforms, the review continues in a specifically designed way, considering the analyses developed by Fugate, Kinicki, and Ashforth (2004) and Fugate (2006), and then considers the types of skills required for employability, paying close attention to the theoretical framework of Pool and Sewel (2007). Other studies from the literature review have been selected to understand the main gaps in the literature and previous analyses.

2.1 Employability skills for emerging professionals:

a literature review

In current times, firms worldwide are paying great attention to human capital, which is crucial for economic systems (Quintini, 2014; Shields & Kameshwara, 2020). It is possible to divide human capital into three dimensions:


i) individual, defined as "the knowledge, skills, competencies and other attributes embodied in individuals or groups of individuals acquired during

their life and used to produce goods, services or ideas in market circumstances" (Westphalen, 1999, p. 4);

ii) institutional, representing the core competencies of laborers employed in an organization (Hamel & Prahalad, 1990); and

iii) national, where workers' competencies are measured via academic qualifications (Healy & Côté, 2001).

Consequently, a balance between the workforce's hard and soft skills is crucial in the upkeep of the labor market in a narrow sense (Zhou, 2020).

The ouds and softs of the labor markets have shown an overwhelming discrepancy between skills supply and skills demand in the last decade. Sparreboom & Tarvid (2017) reported a considerable mismatch between skills demand and availability, mirroring a growing employability discrepancy (Denrell & Le Mens, 2020; Acosta, 2018; Mahmud & Härtel, 2014). Such a discrepancy is due to the spread of the Internet, SM, and Web 2.0 technologies and the associated digitalization and intensified by globalization (Baccarella et al., 2018; Stathopoulou et al., 2019; Smaldone et al., 2019; Baccarella et al., 2020).

Even if these phenomena are democratizing education, there are numerous social aspects to consider as well. Netizens' behavior patterns are a clear signal of lifestyle changes. Individuals tend to assume often elevated

¹ Westphalen, S. Å. (1999). Reporting on human capital Objectives and trends. Amsterdam.

expectations about several aspects of the daily life due to a distortion generated by SM (Stringfield & Stone, 2017; Smaldone et al., 2020). As a consequence, applicants and employers constantly need to navigate the Web to evaluate different scenarios.

Some significant questions emerge from this dynamic: how can people become seen as qualified professionals, and what parameters yield an employable profile candidate?

On one side, the relationship between ability levels and employment chances is unquestionable (Mason et al., 2009; Korczynski, 2005; Rosenberg et al., 2012). Desirable skills are correlated to the needs of the labor market (Chaibate et al., 2020; Ijaola et al., 2020).

Another side, as with the macro-level analysis of human capital, concerns how candidates' level of academic qualification has long been considered a guarantee for skills evaluation (Danford et al., 2009; Edwards et al., 2009; Goetsch & Davis, 2014). Candidates' employability was subsequently weighted via these academic criteria (Acosta, 2018).

For businesspeople, these two frameworks prop up the rife lack of consensus regarding the perception and definition of relevant attributes that workers must present, in relation to the traditional measurement of skills given by academic qualifications (Winkler, König & Kleinmann, 2013). Thus, it is crucial to understand the proper skillsets required in the job market.

2.2 Transversal capabilities to repair the skills mismatch

The emerging mismatch between entrepreneurs' desiderata and skills availability stresses the concept of employability as a construct and an instrument to reduce this gap.

The effect of employability on the skills required in the labor market will be investigated here.

Fugate et al. (2004) used the term employability to denote several concepts enabling the employee to adapt to changes in the workforce environment, benefitting the laborer in gaining career opportunities. This process of adaptation facilitates the mobility-related process from within one organization and between multiple organizations. The employability construct highlights candidates, moving the responsibility for career building and development from the employer to the employee. The employment process, under this theory, does not regard the parts mutually, only the candidate and his relative personal and technical capabilities, as much as his ability to adapt.

The dimensions of employability are structured using an initiative-taking development of workers' situation and flexibility, allowing the ideal candidate to correspond to a wage earner's environment desiderata (Chan, 2000; Savickas, 2005). Underlining this is Hall's protean career concept (1986), which describes flexibility and adaptiveness required to each laborer. These two attitudes are connected to a considerable elasticity and multitasking ability regarding job tasks (Hall and Mirvis, 1995), and they

elucidate features enabling workers to build a boundaryless work path, transitioning between different positions and multiple organizations (Arthur, 1994; Mirvis and Hall, 1994; Hall, 2002).

Employability is a unique attribute between individuals, and it is strongly related to individuals' attitudes² (Chan, 2000). Fugate et al. (2004) described employability's dimensions assuming their efficacy concerning considered contexts. Employability disassembles simultaneously through the subsequent constructs:

a) career identity, regarding the employee's self-representation build around experiences and aspirations;

b) individual adaptability, or the ability of the employee to redesign in front of changes in the working environment;

c) human and social capital, constituting the social and collective side of employability.

Howbeit, Fugate, & Kinicki (2008) considered the concept of dispositional employability as “a constellation of individual differences that predispose employees to (pro) actively adapt to their work and career environments” (p. 20).³ Dispositional employability enables laborers to evolve in response to changes or better predict sudden modifications in the work environment and

² i.e., a worker that is more initiative-taking than others.

³ Fugate, M. (2006), Employability, in Greenhaus, J., and Callanan, G. (Eds.), *Encyclopedia of career development* (Vol. 1, pp. 267–271). Sage, Thousand Oaks, CA.

the resulting uncertainty in the workplace, representing a valuable tip for job applicants hoping to get hired (Fugate and Kinicki, 2008).

Skills are a crucial aspect of the labor market, as they can increase the whole country's growth, as stated before (Quintini, 2014). Skills constitute a collection of fluid and constantly changing abilities, mutating concerning a country's economic, social, and technological evolutions.

Socio-economical phenomena are increasingly touching labor environments.

Elucidating, an example of how the labor market has been deeply affected by a socio-economical phenomenon is the financial crisis of 2008 following the bankruptcy of Lehman Brothers. The collapse of the market resulted in modifications to the required professional profiles. Similarly, insistent changes in the job market structure have been facilitated by the spread of SM, both to the benefit and detriment of job seekers (Baccarella et al., 2018; Smaldone et al., 2020).

It is possible to divide skills into hard and soft (Heckman & Kautz, 2012; Mishra, 2014), and both are crucial to the organization and to workers' effectiveness. Hard and soft skills are interrelated constructs, and the lack of one of these could harm a worker's performance, productivity, and job balance (Mishra, 2014). Currently, soft skills are most demanded by employers, as they are connected to a higher probability that the selected candidate will quickly join the work team, achieving high-level performances (Bacon & Blyton, 2003; Heckman & Kautz, 2012; Groeneveld et al., 2020).

Table 1: Detailed articles analysis.

Author(s)	Year	Main purpose of the paper	Analyzed Skills	Investigated Aspects in the Labor Market	Suggestions for Employability and the Labor Market
Acemoglu & Restrepo	2020	The paper discusses how the recent technological change has been biased towards automation, with insufficient focus on creating new tasks where laborers can be productively employed.	Digital skills, Robotic skills.	Labor demand, declining labor share in national income, rising inequality, and lowering productivity growth.	The study suggests that developing AI in the direction of further automation might mean missing out on the promise of the 'right kind of AI' with better economic and social outcomes.
Adeyinka-Ojo	2018	Identifying employability skills deficits in rural hospitality and tourism (RHT) and developing a framework for employability skills in RHT destinations.	Employability skills, relational skills, industry-specific skills.	The experience economy, human resource management, employability skills, competencies in hospitality and tourism, meetings, incentives, conventions, and exhibition.	Findings indicate fourteen employability skills deficits in RHT and identified the skills valued most by employers in the hospitality and tourism sector.
Adeyinka-Ojo et al.	2020	The paper addresses the strategic industry challenge relating to new education frameworks in the industry of hospitality.	Employability skills, industry-specific skills.	Digital revolution, hospitality 4.0, educational schemes.	The paper identifies critical digital literacy and employability skills that students and candidates need to develop regarding the hospitality and tourism industry.
Alrifai & Raju	2019	The paper aims to find out the required skillset for enhancing the employability of graduates and employees.	Employability skills, hard skills, soft skills, technical skills.	The key employability skills need to be identified and categorized according to a specific industry, like interior design.	The study suggests several implications regarding employability, focussing upon six sets of skills: communication skills, problem-solving skills, teamwork skills, design skills, project management skills, computer skills.

Arnedillo-Sánchez et al.	2017	The paper presents rESSuME: Employability Skills Social Media Survey, which was developed to understand how employers screen candidates' social media profiles.	Employability skills, hard skills, soft skills, technical skills.	Debating the correlation between these skills and employment, employability skills are perceived as more critical than job-specific skills.	Findings remarked that employers investigate candidates online, and their findings affect hiring decisions with rejection, rather than hiring, being the more likely outcome. Studies on graduates highlight a gap between employers' and candidates' perspectives on employability skills, the emerging mismatch between skills expected by the employers and those displayed by candidates.
Benson et al.	2014	The paper concludes an empirical study of UK business graduates and their use of social networking based on the employability construct.	Employability skills, hard skills, social skills, soft skills.	Social capital, online social networking sites, career, and skills management.	The paper presents for discussion an employability skillset for contemporary business professionals and calls for higher education to address the skill gap.
Bongomin et al.	2019	The study explores disruptive technologies of industry 4.0 and quantifies the impact on the skillsets in terms of the number of their appearances in published literature.	Employability skills, 4.0 skills.	Disruptive technologies, literature review, industry 4.0.	The study identified the need to investigate the capability and the readiness of developing countries in adapting industry 4.0 in terms of the changes in the education systems and industrial manufacturing settings.
Bruun & Duka	2018	The paper proposes an instrument to mitigate future technological unemployment by introducing a Basic Income scheme, accompanied by reforms in school curricula and retraining programs.	Digitalization, digital skills, mechanical skills, changes in skills.	Basic Income Studies, educational schemes, retraining programs, technological unemployment.	The authors suggest a special tax on industries using robotic labor, including a practical roadmap that would see a government take this proposal from the conceptual phase and implement it nationwide in one decade.

Cernusca	2020	Studying students and employers' perception of hard and soft skills needed in view of the accounting graduates' access to the labor force market	Employability skills.	Penetrating the labor market in the field of accounting.	Employers would be increasingly interested in hiring young graduates in accounting who hold hard and soft skills and subsequently invest in training to develop the hard skills that they need daily in the chosen job.
Chen et al.	2017	The research aims to analyze the current information management functions in insurance companies in China and the skills needed by information managers, along with the way to improve information management function in the future.	Employability skills.	Insurance companies, information management, employment.	The study highlights the importance of insurance companies implementing information management and seeking practical approaches to improve information management functions.

Clayton & Harris	2019	The editorial includes papers from Switzerland, the USA, Indonesia, Germany, and Australia, all focusing directly or indirectly on this skills theme. However, they exemplify a variety of perspectives that highlight the diversity of views on this topic.	Digital skills, employability skills, hard skills, soft skills, technical skills.	Perspectives on skills diversities between countries and sectors.	The whole editorial raises essential questions about the skills needed to navigate the present and future world of work. Despite the endorsement of employability skills by industry and government, their embedding in Training Packages has concluded that many of the commonly held beliefs surrounding these skills include the transfer of skills across work contexts and the ability to compartmentalize skills and the adoption of an instrumentalist approach to education. Significant concerns appear not to have been addressed by industry or policymakers and implementation of employability skills.
Crisp & Powell	2016	The paper presents a critical analysis of the current policy focus on promoting employability among young people in the UK.	Employability skills.	Conditional welfare as colonization undermines the value of employability as an academic tool for understanding why young people face difficulties in entering the labor market.	The paper suggests that the notion of youth transitions offers more potential for understanding youth unemployment. Policymakers could provide a fruitful avenue for future research, requiring a longer-term, spatially informed perspective as well as a greater emphasis on the changing power relations that mediate young people's experiences of broader social and economic transformations. The paper concludes that promoting employment among urban young people requires a marked shift to address the historically and geographically inadequate knowledge and assumptions on which policies are based.

Cukier et al.	2015	This study provides a systematic review of the extant literature on soft skills in Canada.	Employability skills, soft skills.	Science, Technology, Engineering, and Math (STEM), the Social Sciences, and Humanities, skills gap, expectations and perceptions of employers, formal and informal learning, experiential learning.	The lack of consistency in definitions and fragmentation of stakeholders involved in soft skills development compounds the problem, and more coordination is needed to develop shared expectations and bridge the gap between supply and demand.
Degryse	2016	The paper gives an overview of the new possibilities opened up by the fourth industrial revolution and tackles some specific questions about its effects on the labor market.	Digitalization, digital skills, 4.0 skills.	Employees' status, working conditions, training, digital economy, trade unions.	The paper highlights the main initiatives proposed at the European trade union level in the context of the labor market.
Di Gregorio et al.	2019	Understanding how digital transformation has disrupted the marketing career path by analyzing the most in-demand marketing skills and identifying opportunities for future marketing professionals.	Job-specific skills; twenty-nine skills in five employability areas.	The impact of digitalization on marketing professionals, the skillset of marketing professionals, capabilities.	The study proposes a framework defining the skillset required of marketing professionals.
Finch et al.	2016	The authors propose to adopt the Dynamic Capabilities framework to analyze the competitive advantage of graduates.	Personal resources, meta-skills, job-specific skills, dynamic capabilities.	How new graduates can enhance their competitive advantage when entering the employment market.	The study proposes the adoption of an approach based on Dynamic Capabilities to enhance candidates' competitiveness and employability in the labor market.

Galloway	2017	This research canvasses the state of play in graduates' employability and for legal professions specifically.	Employability skills, attorney skills.	Digital revolution, legal professions, graduates' employability.	The research suggests that new professions anticipate a future of work that is disrupted by digital technologies.
Hollister et al.	2017	The paper aims to report North Florida employers' perceptions of information technology (IT) program graduates' workplace readiness.	Digital skills, employability skills, educational schemes.	Employers' perception, IT, workplace readiness.	Findings suggest the implementation of technological tools to enhance stakeholder benefits.
Imene & Imhanzenobe	2020	The paper discusses how IT has affected the accountancy profession, arguing on the traditional duty of accountants in preparing financial statements, and the several tasks that are carried out throughout that function.	Digital skills, accounting skills, technical skills.	How IT, AI, and big data impacted the accounting sector and the relative skills required.	The paper argues that, in light of the continuous advancements in IT, future accountants and accounting processes are likely to be cloud-based, communicate with and through Artificial Intelligence machines invest in big data and cyber-security, and explore the potentials of Virtual Reality and Augmented reality in meeting users' information needs.
Jones	2013	The present article identifies the alignment of transferable skills developed through international experience.	Employability skills, intercultural skills, soft skills, transferable skills.	The value of domestic intercultural contexts for similar-skills learning.	The paper offers a comprehensive reading of intercultural and transferable skills to address worldwide universities, policymakers, and academics, offering key pointers for policy and practice.
Komljenovic	2018	Understanding the effects of SM on higher education as a sector.	Digital skills.	The impact of SM on the employability of university graduates.	The paper suggests introducing the term 'qualification altimetric' and that SM is building a global marketplace for skills.

Lucianelli & Citro	2018	Understanding professional accountants' views on quality in accounting education reporting results from an empirical study.	Employability skills.	The future expectation of professional competencies in the accounting sector.	Demanding cooperation with the university world to broaden the programs of accounting education with new technical competencies for undergraduate and postgraduate degrees.
Makki et al.	2015	The paper discusses several employability skillsets from existing studies and proposes a conceptual framework that integrates engineering graduates' work readiness skills, career self-efficacy, and career exploration.	Employability skills, engineering skills, industry-specific skills, readiness skills.	Conceptualization of the interrelationship between work readiness skills, career self-efficacy, and career exploration.	The study gives recommendations for further research, underlining how career exploration and readiness skills enhance possibilities for engineering graduates.
Metilda & Neena	2017	The paper analyzes the significance of Learning with Digital Technology to enhance the employability potential of business graduates as digital competence is expected for better employment prospects.	Digital skills, employability skills, hard skills, soft skills, technical skills.	Digital technologies, digital competencies, ICT.	The study analyzed the variation of the process skills of the graduates and the impact of digital technology on institutional implementations. It also identified the instance of high variation in the process skills of those graduates who are not given exposure to Digital Technology facilities.
Minocha et al.	2017	The research discusses various virtual reality technologies and, through examples, argues how virtual reality technology is transforming work styles and workplaces.	Employability skills, 4.0 skills, work styles.	The impact of virtual reality technologies on skills, work styles, and workplaces.	The paper suggests that better awareness of virtual reality technology and its integration in curriculum design will enhance employability skills for current and future workplaces.

Misra & Khurana	2017	The purpose of this paper is to find out the required skill set for enhancing the employability of graduates and employees, majorly focusing on the IT sector.	Digital skills, employability skills, IT skills.	Graduates' employability and the necessity to have the right set of employability skills in the IT sector.	The paper explores the theoretical concepts and models of employability to ascertain gaps between the knowledge and skills imparted by academia and knowledge and skills considered as necessary by employers while hiring.
Muhamad & Seng	2019	Exploring students' critical thinking and problem-solving skills and the awareness of Industry 4.0 to understand their perspective of the future workplace.	Employability skills, critical thinking, problem-solving skills.	The level of critical thinking and problem-solving skills and awareness about the current industrial challenges.	The paper suggests that students with a higher capacity to think critically will be appreciated as a workplace asset and have brighter occupational prospects.
Needham & Papier	2018	The article draws on Bernstein's (1999) theorization of practical and disciplinary learning to show how a curriculum impacts pedagogies, assessment, and quality assurance structures to understand employability skills in the insurance industry.	Employability skills.	Insurance companies, information management, employment.	After examining why college candidates who had succeeded in the first-level occupational qualification with its significant workplace component struggled to complete subsequent university levels, the article concludes that divergent curricula and pedagogies will need serious attention if aspirations for more seamless articulation and more straightforward progression are to become reality.

Nikolaichuk et al.	2019	The purpose of the article is to analyze new challenges, underlying transformations, and recent trends in the Russian labor market towards more training and employability in the era of the digital economy.	Digital skills, employability skills.	Determinants of the Russian job market development, identifying the levers to combat unemployment.	The paper hints at the necessity to make university training more adaptable to the labor market and changing economic conditions; modern higher educational institutions can provide a set of stimulating resources that could help realize students' full potential, give them the edge over other employees, and help them succeed in achieving their professional goals.
Nisha & Rajasekaran	2018	The study highlights the significance of employability skills by reviewing various papers pertinent to these skills.	Employability skills.	Employability skills as perceived by employers, like communication skills, teamwork skills, problem-solving skills, emotional intelligence skills, self-assessment skills, leadership skills, computational skills, interpersonal skills, entrepreneurial skills, and analytical skills.	The paper aims to illuminate the role of employability skills in shaping students' careers and emphasizes how possessing employability skills can help young graduates reach greater heights in their careers. The study consolidates suggestions that students can follow to acquire the employability skills that are essential at workplaces.
Ojanperä et al.	2018	The paper offers a review of the recent literature on the future of work and employability skills.	Digital skills, employability skills, technical skills.	Changes in the nature and creation of jobs, impacts of technology on different tasks. challenges for young people to boost employability, governance models to manage the transitions related to the future of work.	The findings highlights suggestions regarding the susceptibility of tasks and assignments to computerization, industrial diversification, data-driven policy tools, and the development of online labor markets.

Osmani et al.	2019	The paper aims to contribute to the debate regarding employability by mapping and contrasting the rankings of graduate attributes, employing a systematic review of the literature and focused scanning of the job market.	Employability skills.	Employability mismatch, job market, graduates' attributes.	The paper explores the significant variations between employers' wants and the attributes possessed by new graduates, highlighting solutions for the employability skills mismatch.
Pani & Das	2015	Exploring the required skillset for employability in the Tourism industry of the next-gen youth in India by considering the present education system.	Employability skills.	Digitalization, developing economy, tourism skills.	The paper suggests investing in quality training and education to generate efficient professionals in the hospitality sector.
Pejich-Bach & Krstić	2019	This paper aims to develop a profile of Industry 4.0 job advertisements, using text mining on publicly available job advertisements to collect relevant information about the required knowledge and skills in rapid-changing industries.	4.0 skills, industry-specific skills, job-specific skills.	Corporate profiling, text mining, job-search websites.	The paper presents two professional profiles for 4.0 workers, the highlighted skills included in supply change management, customer satisfaction, and enterprise software.

Pieterse & van Eekelen	2016	The paper describes some technical and employability skills essential for students to succeed in a career in software development.	Employability skills, hard skills, soft skills, technical skills.	Understanding the students' problems when in developing these skills.	The research proposes techniques for observing skills gaps and collecting knowledge about these gaps to intervene and suggest remedial action. The authors discuss how to create opportunities for students to enhance their skills.
Rampersard	2020	The purpose of this study is to investigate the key factors driving innovation among work-integrated learning students and the impact on their skills.	Employability skills, digital skills, soft skills.	Skills impacting innovation, changes in required competencies.	The study found that critical thinking, problem-solving, communication, and teamwork have significant impacts on innovation development.
Rosenberg et al.	2012	The authors aim to contribute to the debate by mapping and contrasting the rankings of graduate attributes among academic and practitioner communities, focusing on the United Kingdom.	Employability skills, hard skills, soft skills, technical skills.	National employability strategies, employability mismatch.	The authors suggest reducing significant variations between what employers want and the attributes possessed by new graduates.

Selvam	2017	The study discusses promoting factors obtained from the participating students, explaining promoting factors and their corresponding promoted skills with their magnitude.	Employability skills, hard skills, soft skills, technical skills.	Promoting factor, employability, skills magnitude, participation, responsibility.	The paper highlights that promoting factors are taking up responsibility, self-interest, participating in sports, the college education system, reading books, being confident, joint family system, participation in stage programs, group experiences, family economic conditions, involvement and commitment, industrial interactions, participating in co-curricular activities, peer influences and supports, writing habits, parents' guidance, participation in educational tours, facilities at home, leading activities, facilities available at college, creative tendency, special training, and observation of events.
---------------	------	--	---	---	--

Sharma	2018	The study attempts to understand graduates' awareness about soft skills, their importance, and the impact that soft skills have on building their professional career.	Employability skills, soft skills.	Career building, digitalization, industry-readiness.	A major thrust is put on finding out ways and means to develop soft skills for increasing the employability quotient of graduates.
Shmatko et al.	2020	The research investigates how the career prospects of engineers and researchers have changed, updating and developing the individuals' "portfolio of competencies."	Aggregate competencies, employability skills, experience, soft organizational skills.	Individuals' experience, competencies, and portfolio.	The study advocates those skills received during the study period at the university or dissertation research can no longer be considered sufficient for a career and that firms should consider the aggregate portfolio of competencies between different professionals.

Suarta et al.	2017	The paper discusses the importance of graduates' employability skills in entering the workforce according to employers' perceptions through a literature review. In the 21st century workplace, the occupation-specific skills are no longer sufficient for graduates to meet the needs of labor markets. Workers are nowadays expected to have an additional set of skills and attributes, called employability skills.	Employability skills, hard skills, soft skills, technical skills.	Employability skills as an issue at the national, regional, and international labor market.	The literature review found several employability skills attributes required by graduates in entering the workforce. Communication skills, problem-solving and decision-making skills, and teamwork skills are the attributes of employability skills with the highest importance level. Graduates are also expected to have several personal attributes: self-awareness, self-confidence, independence, emotional intelligence, flexibility and adaptability, stress tolerance, creativity, initiative, willingness to learn, reflectiveness, lifelong learning, and professional behavior.
Tan & Laswad	2018	Examining the employability skills of accountants cited in job advertisements in Australia and New Zealand indicates the skills that employers most value.	Accounting skills, sought-after skills, ability to collaborate with colleagues.	The changing roles of accountants.	Teamwork with a positive attitude and good communication skills seems to be the most valued behavioral skill perceived by employers.
Vanhercke et al.	2014	This paper aims to define employability within the psychological literature with a focus on perceived employability.	Employability skills.	Psychological approaches for the definition of employability, employability perception, skills perception.	The literature review compares the perceived employability approach to other approaches in the psychological field, concluding with integrating the three approaches into a process model to demonstrate their interrelationships and derive a specific view on employability skills.

Varshney	2020	The paper examines the techniques and methods used by select companies to encourage and develop their employees to become attuned with the digital transformation processes implemented.	Digital skills, digital awareness, digital abilities, employability skills.	The most sophisticated techniques employed by most new companies in their workforce to upgrade their digital awareness and capabilities.	The author developed a model to sustain, evolve, and update employees in the digital era.
William	2015	The rationale of this phenomenological study is to investigate the perceptions of students and employers related to the soft skills needed to be successful in future employment.	Employability skills, soft skills.	Employability, Mezirow's transformational, Daloz's mentorship theories, successful skills.	Communication was the most important and the most lacking soft skill. The recommendations informed the creation of a mandatory three-day professional development training program to help students enhance their soft skills before entering their future careers. This study directly affects positive social change by enhancing the quality of soft skills for future employees who enter the local workforce.
Wilton	2011	The paper focuses on how UK policymakers offer dominant rationales for the continued expansion of higher education: to service the high-skill labor requirements of a knowledge economy and increase educational and employment opportunities for underrepresented groups.	Employability skills, knowledge skills.	High-skill labor, policymakers, employment.	Data suggest that traditional labor market disadvantages still appear to impede achievement, regardless of the extent to which graduates develop employability skills during their undergraduate studies.

Table 1 presents a detailed analysis of the literature to understand the main differences and similarities between skills' roles in the employability process.

Soft skills play a pivotal role in individuals' employability, while hard skills have gradually seen a lessening in their importance (Heckman & Kautz, 2012; Mishra, 2014).

Several models for the analysis of skills have been developed in the literature. One of the most intriguing is the Key to Employability Model, or CarrerEDGE, developed by Pool and Sewell (2007), in which dimensions of the model assume the fanciful form of a 'key' to open the door of the job market, tracing the path of continuous learning.

The model considers five dimensions that are considerably correlated: (a) career development learning; (b) work and life experience; (c) degree subject knowledge, understanding, and skills; (d) soft skills; and (e) emotional intelligence.

A deep connection between the dimensions of the model is given by self-efficacy, self-confidence, and self-esteem. These are closely interrelated, because learning experiences allow reflections and evaluations regarding one's own condition (Pool and Sewell, 2007).

2.3 Hard & soft skills enhancing candidates' employability

Hard skills recur in labor marketers' argumentations. Hard skills are defined as "specific capabilities to perform a particular job" (Cimatti, 2016, p. 98). Technical skills depend on training and earlier work experience (Forde & MacKenzie, 2004). These kinds of skills can be learned quickly: languages, interface software, coding, etc. Collectively, hard skills refer to the ability to use combined knowledge to undertake a specific job or profession in efficiency and effectiveness.

Technical skills are crucial for specific job profiles, considered as prerequisites when selecting human resources. Hard skills are measurable, quantifiable, and can be learned through study and practice. Therefore, specific professional skills differentiate workers in different areas (Wolf and Archer, 2013; Harris and King, 2015; Stringfield and Stone, 2017). In sum, hard skills are strongly related to employability. The literature analysis supports this supposition, highlighting a wide range of papers identifying the corresponding hard skills required for employability for technical works (Woya, 2019; Oviawe et al., 2017; Entika, 2017, Marsithi and Alias, 2013; Gokuladas, 2011; Hinchliffe & Jolly, 2011; Wolf and Archer, 2013; Harris and King, 2015; Stringfield and Stone, 2017; Nurlaela et al., 2017). Technical aspects from the laborers also appear in research analyzing hard skills in IT professionals and legal workers (Siddoo et al., 2017; Alamelu et al., 2017; Murdoch, 2015).

The spread of digital economies strongly modified the job market, requiring highly skilled workers. This aspect is found in several papers highlighting that hard skills required to study politics are statistical and quantitative elements of Excel. On the other hand, Atkins (2013) notes that, for English teenagers without academic degrees, practical skills are necessary to have a chance to join the world of work.

Additionally, the literature review highlighted interesting studies regarding soft skills. Soft skills are defined as “non-technical and not reliant on abstract reasoning, involving interpersonal and intrapersonal abilities to facilitate mastered performance in particular contexts” (Hurrell et al., 2013, p. 161), suggesting the strong customized connotation of this construct. More specifically, Trilling and Fade (2009) define soft skills as “21st-century skills.” The authors divide them into learning and innovation skills, digital literacy, and life and career skills. Matteson et al. (2016) discussed an undefinable number of soft skills in candidates’ profiles, examining and finding differing interpretations due to context.

A precious aspect characterizing soft skills is that they are not teachable, treasured by the strong social and psychological characterization that personal skills produce.

Soft skills play a vital role in work requiring human interaction, where direct contact is integral and often determines the service's quality. Personal skills, however, are also fundamental in jobs not requiring customer interaction (e.g., manufacturing), considering that corporate activities increasingly

require teamwork and worker interaction (Heckman & Kautz, 2012). Personal skills broaden the versatility of hard skills as a sort of liquid construct.

In example, Heckman and Kautz (2012) researched the influential role performed by soft skills for employability. The authors considered soft skills such as communication, teamwork, problem-solving, and ethics in the sports industry.

Although definitions vary in the literature, depending on different social and psychological characteristics of workers and on the working contexts, the literature remarks on many aspects representing foundation to soft skills, such as communication, leadership, problem-solving, initiative, self-regulation, expertise, know-how, and ambition (Hirsch, 2017; Hogan et al. 2013; De Fruyt et al. 2015; Alvarez & Alvarez, 2018), even if employability is not always increased by skills owned by workers, especially those skills developed with discretionary work (Petrovski et al., 2017).

Factors characterizing soft skills dimensions are unraveled in various research developed in specific industries and contextual scenarios.

Exemplifying this, Ngoma and Dithan Ntale (2016) and Dejaeghere et al. (2016) investigated the general importance of social capital in the underdeveloped countries of Uganda and Tanzania, while Ramirez-Pérez et al. (2015) highlighted the role of communication, teamwork, and effective organization for the practical employability of Mexican youths.

In the hospitality industry, Mahfud et al. (2017) revealed the critical role assumed by soft social (ability to cooperate, focus on doing work, communication skills) and psychological (honesty, responsibility, creativity) skills, highlighting the pivotal role of soft skills for employability in sectors involving direct contact with customers and therefore require workers to engage in personal interaction regularly.

2.4 Literature gap

The gap in the literature to which this doctoral thesis contributes is mainly represented by the necessity of proposing a new methodology to profile skills in order to enhance employability, emerged from the comparison of several studies, after a systematic literature review, revealing a lack of a precise and proper measurement of candidates competitiveness based on digital data. From a managerial perspective, this doctoral thesis is also aimed at confuting the reflections from Crisp & Powell (2016) that employability is not a proper tool for skills' evaluation, defending the position of Fugate et al. (2021), demonstrating once again the validity of this construct. From a methodological perspective, this doctoral thesis develops several methodologies concerning text analytics in a complete way, distancing this study from other works in the extant literature employing profiling techniques on job ads (Suarta et al., 2018; Kamaru Zaman et al., 2019).

3. Methods, Extraction Techniques, and Profiling Techniques

The methodology employed for this analysis can be divided into three central moments: the extraction methodologies, textual analysis, and graph analysis. Combining these three methods helps obtain a correct estimate and representation of the data processed with R and Python software. The methodology starts with describing the current big data analytics in **Section 3.1**, explains the role of text analytics in **Section 3.2**, shows the pivotal role of graph theories in **Section 3.3**, exploits the extraction techniques in **Section 3.4**, and profiles techniques and the relative software architecture in **Section 3.5**.

3.1 Big data analytics: a new approach to the applied research regarding transversal competencies

The history of big data dates to 2001 when Doug Loney, vice president and Service Director at META GROUP, attempted to explain in a corporate report the phenomenon of the rapid growth of available data, identifying three characteristics for big data: volume, velocity, and variety (McAfee et al., 2012; Kapil et al., 2016; Younas, 2019). Over time, following new studies, two other peculiarities were then added to these three: veracity and value (Hitzler & Janowicz, 2013; Saha & Srivastava, 2014; Younas, 2019). The 5V have

been integrated with further dimensions identifiable in the variability and visualization (Owais & Hussein, 2016).

Since 2011, big data has attracted many academics and has often been the subject of essential projects, being the main focus of various international papers, conferences, and books (Porat & Strahilevitz, 2014; Perrons & Jensen, 2015; Custers & Uršič, 2016; Devins et al., 2017; Sonka, 2014; O'Connor & Kelly 2017; Jee & Kim, 2013; Alonso et al., 2017; Vayena et al., 2018; Galetsi et al., 2020; Kim et al., 2014; Desouza & Jacob, 2017; Fredriksson et al., 2017; Klievink et al., 2017; Sarlin & Marghescu, 2011; Sarlin & Peltonen, 2013; Sarlin, 2016; Trelewicz, 2017; Riikkinen et al., 2018; Pejić Bach et al., 2019; Perrons & Jensen, 2015, Sumbal et al., 2017; Patel et al., 2020; Mohammed et al., 2019; Aktas & Meng, 2017; Xiang & Fesenmaier, 2017; Del Vecchio et al., 2018; Talón-Ballesteros et al., 2018, Li et al., 2018; Alaei et al., 2019).

Researchers have demonstrated that a large amount of data is essential to know more about the external environment and the surrounding society and gaining a better idea of all those factors beyond customers, potential customers, and suppliers that impact companies (Davenport & Diché, 2013). Regarding managers, “only 23 percent said their organizations had a strategy for big data” (Davenport, 2014, p. 6).⁴ In this scenario, the pivotal role of

⁴ Davenport, T. (2014). Big data at work: dispelling the myths, uncovering the opportunities. Harvard Business Review Press.

data analysis emerges to support firms in current times, specifically by developing tailor-made software (Davenport & Harris, 2017).

There is no precise and rigorous definition of big data. There are those who consider it a marketing term and an advancement of technological trends that permits a new understanding of the world and lends to decision making. Mckinsey Global Institute, in 2011, stated that it is possible to call big data a characteristic data set that has such a large volume that it exceeds the ability of relational database systems to capture, store, manage, and analyze data (Brown et al., 2011). More precisely, it argued that this size must surpass a certain number of terabytes (Brown et al., 2013). But the quantity and complexity making a set of data qualifiable as big data remains a debated topic. Some take the petabyte (1,000 terabytes) as the threshold, and others operate in the field of exabytes (1,000 petabytes).

However, considering only the size of the database may expose a strong bias misleading companies, mainly because, although they do not have such vast archives in terms of dimensionality, content characteristics, such as variability, enable the employment of technologies and techniques inherent to big data management and profiling (e.g., unstructured data) that can lead to new and relevant findings. In this case, analysts adopt a ‘little data approach’ to process data accurately and swiftly.

The relevance of big data finds a brilliant metaphor in Rick Smolan’s (1986) photographic series *Day in the Life*, where he defined it as “the dashboard of humanity, an intelligent tool to be able to fight problems such as crime,

poverty, and environmental pollution.” McAfee et al. (2012) compare the importance of big data to the importance of the invention of the microscope, stating, “You can’t manage what you don’t measure” (p. 63).⁵ Geoffrey Moore, a marketing theorist from Silicon Valley, wrote that, “without big data, you are blind and deaf in the middle of a highway” (2014, p. 316).⁶ Contrary to popular belief, all organizations can benefit from big data analytics, deriving knowledge from their data production.

Apart from profiling based on corporate data, big data offers the possibility of measuring behaviors and sentiments of SM users.

Big data is heterogeneous. It does not originate from a single source but from different matrices. Most of big data is generated by mobile devices, and a substantial percentage stems from sales and pricing data. Other kinds of big data derives from fidelity cards and promotional events and other from computer log data and from popular SM (e.g., Twitter, Facebook, and Instagram).

It is customary to classify data generation into three categories:

- i. Human-generated: deriving from social networking platforms (Mansour, 2016), blogging and microblogging (Broniatowski et al., 2014), social news (Oboler et al., 2012), social bookmarking (Yanbe, 2007), multimedia sharing (Pouyanfar et al., 2018; Wang et al., 2018), wikis (Percell, 2016), review sites

⁵ McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data: the management revolution. *Harvard business review*, 90(10), 60-68.

⁶ Moore, G., (2014). “Without “big data”, you are blind and deaf and in the middle of a freeway”, pp. 316-334, in: *The Business Book*, 2014. Dorling Kindersley Ltd. ISBN: 978-1-4654-1585-1

(Özköse et al., 2015), and e-commerce portals (Akter & Wamba, 2016; Zheng et al., 2020). Therefore, it is possible to consider human-generated data when every human experience can be recorded in photographs, audio, and video, and subsequently digitized and archived.

- ii. Machine-generated: produced by sources such as GPS sensors (Zhou et al., 2016; Kan et al., 2018; Liu et al., 2020), IoT (Cai et al., 2016; Mourtzis et al., 2016; Marjani et al., 2017), RFID (Zhong et al., 2014; Zhong et al., 2015; Zhong et al., 2016), monitoring stations for meteorological events (Zheng et al., 2015; Guo, 2016; Huang & Jiao, 2017), scientific instruments, and biomedical devices. Data collected in this way is usually well structured and includes transactions data and reference tables and is generally considered highly reliable, even considering occurrences of faulty sensors and missing data. Its well-structured nature makes it predisposed to computer processing.
- iii. Business-generated: derived from human or machine generation internally to a company and able to record all the data-driven activities of corporate business processes. Historical data is statically stored in relational databases, representing payments, orders, production, inventory, sales, and financial data (Bumblauskas et al., 2017; Ebner et al., 2017; Majeed et al., 2019; Trabucchi & Buganza, 2019). The company collects and processes data relating to the demographic and psychographic aspects of customers, purchasing behavior, and website visits (Bradlow et al., 2017; Liu et al., 2019; Holmund et al., 2020). This data includes: the level of customer satisfaction and any assistance problems encountered (Mariani et al., 2018;

Park et al., 2019; Park, 2019; Zhao et al., 2019); sales, cost, and cash flow data: production and data related to production volumes (Spiess et al., 2014); and shipments, stocks, and the activities of competitors (Kotler et al., 2019). Moreover, there is an intercurrent distinction between unstructured (a), structured (b), and semi-structured data (c).

(a). The greater the adherence to reality, the greater the possibility of transforming it into actual knowledge. Mirroring the reality enhances the achievement of efficiency, process optimization, and the development of new products and services. Not by chance, the core function of big data is providing the best possible representation of reality through data analysis. Unstructured data internal structure does not follow a predefined scheme. It can be of textual or non-textual nature and is generated by humans or machines (e.g., images, email, social media data, audio, video, sensor-generated data). Compared to structured data, it is more challenging to analyze. However, specific analytics based on machine learning are utilized for structuring the data, allowing the extraction of valuable information.

(b) Structured data, which usually reside in relational databases, has a structure that makes them much easier to investigate through queries and algorithms.

(c) Semi-structured data exists between the two types and is characterized by internal marks and tags that identify separate data elements, allowing groupings and hierarchical formations.

According to the fifth rule of programming by Rob Pike (1989), “data dominates. If the right data structures are chosen the right data structures and

organized well, the algorithms will almost always be self-evident. Data structures, not algorithms, are central to programming,” reflecting the crucial role of the data manipulation process and the choice of the proper data structure.

There are different types of data structures, such as Data frame,⁷ Matrix,⁸ List,⁹ Array,¹⁰ Table,¹¹ Time series,¹² and Distance matrix.¹³ It is appropriate to explain the difference between a dataset, database, data stream, data flow, and data source in this context.

The dataset is a collection of data, or a set of data structured in a relational form, organized according to specific characteristics or attributes. The dataset

⁷ Data frame: a list of vectors (variables) that must all have the same length (number of cases) but can be of different types, such as nominal, cardinal, or numeric. A data frame can be constructed by entering data directly into R or importing data from other applications; the elements of the data frame (row-vectors, column-vectors, and cells) are indexed.

⁸ Matrix: information organization tool for statistical analysis, it is a table showing all the information collected for all the cases studied. These cases are organized by arranging the cases in rows and the variables in columns; to perform the necessary algebraic operations required by statistical analysis on the matrix, it must not have empty cells.

⁹ List: abstract and dynamic data structure denoting a collection of homogeneous elements or data containers. The size is not known a priori and can vary over time. A list can contain one or more fields containing information and must contain a pointer through which it is linked to the next element. The primary operations are insertion, removal, search, access to random by index, sequential access, and element count.

¹⁰ Array: a vector that indicates a complex, static, and homogeneous data structure. Arrays are inspired by the mathematical notion of vectors or matrices, allowing one to define new data types starting from pre-existing types through the aggregation of different objects, all of which are the same type. One can imagine an array as a sort of container whose boxes are called cells of the array itself, and each of the cells is identified by an index value.

¹¹ Table: the primary function of tabulating the data present in a dataset is to produce frequency or contingency tables.

¹² Time Series: a set of random variables ordered concerning time and expressing the dynamics of a specific phenomenon over time. Historical series are studied to interpret a phenomenon, identify components of trends, cyclicity, seasonality, and accidentality and predict future trends. In the historical series, it is assumed that n observations are coming from as many dependent random variables; the time series can be deterministic if the values of the variable can be precisely determined since the previous values or of the stochastic type if the values of the variables can be determined based on the previous values only partially.

¹³ Distance matrix: tries to model the data as distances between points in a geometric space, identifying an x -dimensional space representing the n points in a smaller dimension's coordinates.

corresponds to the content of a single database table or a single statistical data matrix in which each row of the table corresponds to a distinct member of the dataset and each column to a particular variable. The number of members and variables constitute the dimension of the dataset. The dataset is independent of the data source to which it relates.

Currently, the main problems that need to be addressed regarding the collection of data on the internet are of both a local nature—such as the accessibility of data by applications and the lack of availability for a specific format (e.g., XML, JSON, CSV)—and of a global nature—such as the heterogeneity of the set of datasets from different sources.

A certain standard has been implemented to facilitate and approve the construction of datasets: The Statistical Data and Metadata Exchange (SDMX). SDMX represents an international cooperation initiative to develop and use the most efficient processes for the exchange and sharing of statistical data and metadata between the most important international organizations. The initiative, launched in 2001, aims to establish a series of standards recognized and observed by all operators to facilitate access to statistical data. However, this standard can only be associated with using a data exchange language such as XML.

The term dataset can also be used more generically to indicate data in a set of closely related tables relating to a particular experiment or event.

A database is an archive of structured and homogeneous data in terms of content and format, a harbinger of multiple functions and performances. It

contains both information and relevant relationships. The contents can be transferred, modified, copied, or reorganized without modifying the data files. As far as the connection between the numerous data is concerned, this can be relational, hierarchical, or reticular.

In the 1960s, the first databases were born. These are those of a hierarchical, tree type; in the 70s, the reticular and relational ones developed. Finally, in the 90s, object and semantic databases emerged. The use of databases spread throughout the 70s and 80s.

Unlike files and old computer archives, the database allows one to manage data entry, update, and search operations in a standardized form using unique Database Management System (DBMS) software. The management operations are generally carried out using a data management language based on a query language, and the most used is the SQL language.

Data flows represent a set of activities, each of which modifies some of the information. The hardware development of dataflow architectures was one of the significant research lines of the 70s and early 80s. The first to deal with static dataflows was Jack Dennis from the Massachusetts Institute of Technology (MIT), while the Manchester Dataflow Machine and MIT Tagged Token architecture were the first projects to deal with dynamic dataflow architectures.

Subsequently, there is the data flow diagram, a type of diagram defined in 1978 by Tom DeMarco that defines the data flows within the information system. The diagram is used in computer systems and descriptions of the data

flow. Through data flow diagrams, one can define how information flows and is processed within the system, which is essential for understanding where the data is stored, from what source it came, at which source it arrived, and which system components processed it. This diagram, combined with the Entity-Relationship diagram, is used to design data and functions within an information system. A data flow diagram consists of processes, agents, data flows, and data repositories, where the data flow represents the movement of information between agents, processes, and data repositories. In this doctoral thesis, data flows are found in the data extraction process.


Data streams contain raw data collected by users' browsers from websites, as they are algorithm providers. The most significant statistical and financial database is Thomson Reuters Data stream (TRD), which covers numerous asset classes, estimates, economic and financial fundamentals, indices, and other economic data. TRD contains over 25,000,000 reports, allowing the analysis of economic phenomena not only of a numerical type but also broken down into individual components.

Data streams collect data from the most important national and international statistical institutes. Its main characteristic is the breadth of the range of traceable information: from macroeconomic data to data on derivatives, from government bonds to bonds of private companies, from the price of primary goods on the international market to the consumer price of a single good in a domestic market, and much more. In particular, the following data series are available in a data stream form: equities, bonds, stock and bond market

indices, interest and exchange rates, macroeconomic data, futures and options, warrants, and commodities.

Data sources represent the source of the obtained data. They are used in the context of databases and database management systems. The data source is defined in the application in such a way as to find the location of the data. The data source for a computer program can be a file, datasheet, spreadsheet, XML file, data encoded within the program, or XM file. It is possible to collect all the technical information necessary to access data: driver name, network address, and network software through the data source. There are two types of data sources: machine data sources and file data sources. Furthermore, data sources may differ according to the application or field of operation. The data source is represented by a string or by the complete structure.

Methodologically, the data management and analysis process (big data value chain) is quite complex. It is divided into a series of phases, each of which is preparatory to the other.

There are six basic  steps to the data management and analysis process:

1. Acquisition: the raw data is collected from available, reliable, and well-constructed sources and then transformed to make it as homogeneous as possible. The acquisition of digital data can take place through different means, such as the APIs of those who provide data from social networks, sharing applications, ETL tools, and web-scraping tools. This phase of the

value chain requires significant financial efforts for the necessary infrastructure investments in some circumstances.

2. Preparation and extraction: once collected, the data enters the data preparation phase, sometimes also called *pre-processing*. During this phase, the raw data acquired is cleaned and organized for the next step. During the preparation, the data is rigorously checked. It is necessary to select the best quality items and eliminate the redundant, incomplete, and wrong ones.
3. Storage and Integration: clean data is uploaded to the target system and then decoded into a language readable to the processing system.

From a semantic point of view, data acquired from the previous phase can be enriched, trying to give it meaning. Storage takes place in powerful hardware or servers, data warehouses, data lakes, and cloud storage. Data warehouses are the traditional systems on which business applications record their data and are the sources from which big data applications draw information. Data lakes are repositories or containers of information capable of containing vast volumes of data in their native format, and it is necessary to perform the requisite processing, obtaining the information for business applications. The requested information is extracted from data. Data lakes are born as a valuable paradigm to exploit the potential generated by the characteristics of big data to overcome the strong rigidities present in traditional approaches to data management (data silos, data warehouses, data marts) typically centered on the creation of banks structured data.

Data lakes facilitate and speed up data sharing because they are built on raw data (structured, semi-structured and unstructured) in its original format and

allow analysis and value extraction. Clouds are one of the most used ways to store more corporate data, sometimes in object-storage mode.

4. Processing and analysis: data loaded into the system is processed for interpretation. The processing phase uses semantic learning algorithms and varies according to the processed data source and its purpose. The process takes place through cloud technology. The cloud is based on the convenience of electronic data processing methods, increasing their speed and effectiveness. Not all data is analyzed, both for practical and timing reasons. Therefore, the necessary quantity is identified concerning the objectives of the analysis.
5. Interpretation: data is transformed into information and knowledge that the analyst can use; it is translated and readable. At this point, companies can use the data themselves and move on to the next phase of analysis.
6. Reporting and decision-making: the process ends with choosing how to use the information in the decision-making process. It is deduced that what matters is not the quantity of data but how it is analyzed and used. The importance of the analysis is strictly connected to use. It is therefore essential to choose and use cutting-edge, fast, and precise techniques and technologies. Obtaining value from big data upstream requires careful planning that must outline and support both *data at rest* and *data in motion*.

Big data methodologies represent the critical element in a data-driven economy. There are many areas, both public and private, in which the use of big data analysis techniques has made it possible to create new services,

improve existing ones, innovate production and distribution processes, and make the offer of all products and services more responsive to the needs of consumers and citizens. However, several market failures have been recorded due to the existence of diversified barriers.

There are several barriers for researchers and practitioners in the use of big data methodologies:

I. Skills-related barriers: there is a substantial lack of professionals with the necessary knowledge to exploit the intrinsic characteristics of big data. The ability of analysts to collect, manage, evaluate, and use data in various contexts of application is called *data literacy*; simply put, data literacy is the ability to derive meaningful information from data. Data literacy encompasses a series of skills, including discriminating data based on specific use, correctly interpreting graphs and tables, knowing the practical techniques and technologies of data analysis, recognizing when data is tampered with, and effective communication of information. There are three pillars of Data Literacy: knowledge of data, data-based decision-making, and dissemination of the use of data within the company.

Sophisticated methodologies for big data analyses radically change the decision-making process, incorporating multiple factors—i.e., in highly competitive sectors, the presence of data scientists able to understand and extract the value of data implies having high barriers to entry for other companies, because it is not possible to fill this gap simply by changing the

requirements of the graduates. Therefore, it is necessary to retrain a significant amount of talent on the spot.

II. Technical and technological barriers: the universe of big data is constantly evolving. Consequently, technological platforms employed for big data management are subjected to dynamic obsolescence and continuous research regarding methodologies. The relative newness in the relationship between big data, research, and business models is also part of the technical and technological barriers.

III. Privacy barriers: laws, codes of conduct, and regulations are all barriers that hinder the process of using big data, containing a series of norms and principles to be observed to protect the people involved.

Big data methodologies contain a duality: they are a source of advantages and a source of disadvantages under the methodologies of collection and analysis and the purposes of their use. If effectively and efficiently extrapolated and used, big data methodologies can offer considerable opportunities for the benefit of both users and recipients; however, if misinterpreted and misused, they create essential risks.

The main disadvantages include a) *false discoveries*: businesses can use data aggregations to gain an unfair competitive advantage and harm consumers; b) *privacy violations*: as the value of big data grows, the importance of protecting privacy grows—businesses, companies, and individuals will have to come to a compromise between privacy and usefulness; and c) *infringement of intellectual property and creation of highly concentrated*

markets: data can be copied perfectly and efficiently, meaning the same data can be used by several people simultaneously, and such elements distinguish data from other physical resources, but those who hold a vital information asset could try to defend it at any cost, and, furthermore, the data available to some companies can be made available to other economic actors who can gain value in a different context than the company that made them available;

d) *liability and security issues*: sensitive data should be secure—big data explosion paved the way for the Cambridge Analytica case, a case study in insecure data: through some flaws in the Facebook system and violating the terms of use of the data that had been collected for research purposes, the profiles of millions of American citizens were used by the British consultancy for the political marketing campaign linked to the election of Donald Trump, and the construction of politically oriented messages and the continuous exposure of the masses to such messages have had a manipulative and polarizing effect on opinions. After the scandal, Cambridge Analytica filed for bankruptcy.

Among the advantages are: a) *economic benefits*: organizations capable of using big data are estimated to increase their operating margin by more than 60%—this is the case with Tesco in the UK; b) *ability to discover hidden behaviors and consumer needs*: companies are increasingly using CRM applications that allow them to manage the relationship with the customer under different aspects to develop a profitable and lasting relationship—specifically, companies can analyze consumer behavior from a multichannel

marketing perspective to improve customer satisfaction, collateral sales, and offer higher quality after-sales services, thus increasing loyalty;

c) *enhancement of forecasts*: when registering on social networks and apps of Facebook-partner companies, users provide authorizations for the processing of personal data and thus become part of a database—the information one provides, the pages one visits, and the posts one likes, together with the geolocation functions of smartphones, allows for the construction of a psychological profile of users in a straightforward and precise way; d) *greater support in decision-making*: in the agricultural sector, the weather data and agronomic information of the individual vineyards of interest collected through special sensors placed on drones, if properly used and analyzed, can feed a decision support system in the wine sector; e) *creation of new products and services*: sentiment analysis, in particular, studies what is posted on social media and reveals the attitudes and propensities of users, thus discovering new needs and new requests to be satisfied through the creation of new products and services, reducing a company's risk of failure by garnering knowledge of the consumer and foresight necessary to anticipate consumer needs; f) *enabling of new business models*; g) *trade development*; h) *improvement of the national economy*: data can be used to make the city smarter, ecological and liveable—i.e., to focus on renewable energy through platforms powered by citizens that provide data and reports on the electricity grid, traffic, pollution, water consumption, and public lighting to plan green areas, traffic management, and making the energy supply more efficient by developing innovative solutions (also

applicable to the private sector); i) *improvement of the efficiency and quality of products and services*; l) *improvement of the quality-price ratio of goods and services*; m) *improvement of the productivity of the public and private sectors*: the development of the internet has accentuated the company-community paradigm of reference: the company advertises a product or service by simply clicking on the Facebook share button or an individual can suggest a friend's LinkedIn profile for a particular job posting—all this creates economic value at no cost; n) *greater transparency*: making big data more easily accessible to stakeholders promptly creates extraordinary value—for example, by integrating data from R&D, design, and production units, companies and governments can identify suspicious activities by recognizing patterns that may indicate fraudulent behavior, thus preventing its occurrence, or identifying the culprit; o) *more specific segmentation of the consumer target*: through big data, it is possible to identify the key factors that move people to purchase a particular good or use a specific service.

Table 2: Main phases of a big data Analysis.

PHASE	DATA TYPE	TYPE OF ANALYSIS	ANALYTIC METHODS
1.	Hindsight¹⁴	<u>Descriptive</u>	<u>Descriptive Statistics</u>
2.	Insights¹⁵	<u>Diagnostics</u>	<u>Correlation Analysis</u>
	Insights	<u>Cognitive</u>	<u>Cognitive Analytics</u>
3.	Foresight¹⁶	<u>Prescriptive</u>	<u>Strategic Forecasting</u>

¹⁴ Personal/collective data referring to the past, obtained from the extraction of meaning.

¹⁵ The set of raw data and data analyzed through descriptive analysis.

¹⁶ Prospective prescriptive data is useful for predictive analysis.

3.1.2 Big data processing

Generally, big data processing occurs in three central moments: (1) data extraction, (2) storage and warehousing, and (3) data analysis and the possible applications of meaning extractions obtained from data processing.

There are two fundamental characteristics for the preliminary data analysis:

[1]. Transparency: data must be real and correspond correctly to the analyzed phenomena.

[2]. Cleanness: tendentially, various data sets must be clean for the necessary metadata to be correctly analyzed, avoiding bias and erroneous findings

3.1.3 Big data analytics

Table 2 shows the main phases of the big data analysis (BDA). BDA usually involves three phases distinguished by the type of processed data, possible types of analysis, and different methodologies of analysis.

In the first phase, there is data defined as hindsight. This data is of a raw nature, personal or collective, and obtained from the extraction of meaning in metadata from big data series. Descriptive analyzes can be developed on this data to outline the data series and proceed with further analyses.

In the second phase, data is defined as insights, which concerns sets of data and preliminary analyses. Therefore, this data allows for diagnostic analyses with an analysis of correlations or cognitive analyses.

The third phase deals with predictive analyses with prospective prescriptive data, permitting strategic predictive analyses based on applications on real data, also called foresight.

3.1.4 Web scraping

Web scraping is a technique of extracting data of a digital nature from a website using code or software. Big data is generally extracted in small series to analyze phenomena and correlations and provide predictive analysis. One of the primary purposes of web scraping is related to data indexing on the internet, involving a phase of data collection and a later stage of transforming raw data into metadata. Information describing a whole database can be analyzed locally by creating datasets through the most common analysis software. Some of the scraping uses can be online price comparison, data mining, scientific research, and web data integration.

Tables are usually presented in HTML format, not easily manageable for data processing but easily transferable to a spreadsheet. However, the information is spread over hundreds of pages, and, without web scraping, it would be an enormous operation of “copy and paste.”

The main feature of web scraping is to extract data from pages with a defined pattern through a crawling process highlighting the desired hyperlinks. After the URLs’ extraction, there will be a normalized table with the number of objects extracted per row, and the variables will correspond to the fields of interest selected before the crawling phase per column.

3.1.5 Metadata

The rapid change in the ways of accessing information, mainly caused by the development of new technologies and the transition to the digital age—which began with the advent of the World Wide Web (WWW) and began to see concrete evolutions from the early 00s—has undoubtedly led to a structural change in the management of resources related to the information obtainable through data. Metadata, or synthesis data of digital information, is a primary tool within data content management and represents a considerable connection to the value chain of knowledge in economics. The main question that arises when it comes to metadata could be how to manage it. To answer these questions, it is useful to understand how metadata is born. Metadata first became a classification during the Dublin Core Metadata Initiative (DCMI) in 1995. Since then, fifteen terms have been maintained, indicating 15 groups of standard metadata, which are shown in **Table 3**:

Table 3: Dublin Core metadata.

<u>CONTENT</u>	<u>INTELLECTUAL PROPERTY</u>	<u>INSTANCES</u>
1. Title	8. Creator	12. Date
2. Subject	9. Publisher	13. Format
3. Description	10. Contributor	14. Identifier (ID)
4. Type	11. Rights	15. Language
5. Source		
6. Relation		
7. Coverage		

Source: adapted from Weibel et al., (1998).

3.1.5.1 Content

- 1) Title: the name given to the resource, such as *security*, will be a term by which the resource will be known formally.
- 2) Subject: parent topic of the resource—i.e., a subject can be expressed by words or keywords which explain the topic of the resource. Usually, these terms are chosen from the values of a purpose-built thesaurus.
- 3) Description: free descriptive text that can include an analytical summary, an index, or a graphical representation of the content.
- 4) Type: nature or genre of the content of the resource.
- 5) Source: the source from which the resource in question came, usually defined as the source resource.
- 6) Relation: reference to a related resource.
- 7) Coverage: extent or purpose of the resource content. Typically, in coverage, we include the spatial location (the name or polar coordinates), the time span (the specification of a period, a date, or a series of dates), and a jurisdiction (for example, the name of an administrative entity, such as a federal department or division).

3.1.5.2 Intellectual property

8) Creator: legal entity that has the primary responsibility to produce the content of the resource. This can refer to a person, an organization, or a service body directly responsible for the resource's intellectual content.

9) Publisher: legal entity responsible for the publication or legal deposit (patent, invention patent) of the resource.

10) Contributor: legal entity responsible for producing one or more contributions to the resource's content.

11) Rights: usually, a Rights element contains an indication of the management of rights on the resource or a reference to the protection service that provides this information. This field includes Intellectual Property Rights (IPR), copyright, and other intellectual property rights. If the Rights element is unavailable, no assumptions can be made about the resource's rights. It, therefore, create multiple variables for the processing of sensitive data.

3.1.5.3 Creating instances

12) Date: date related to an event of the life cycle of the resource. Usually, the date is indicated with the specific date when the digital asset was created.

13) Format: generally, in this field, the type of support and the resource format are indicated, outlining the physical or digital manifestation of the resource content.

14) Identifier (ID): consisting of the unique reference to the resource within a given context. A sequence of alphanumeric characters usually identifies resources according to a formally defined identification system. An example of such identification systems include the Uniform Resource Identifier (URI).

15) Language: the code of the resource's intellectual contents expressed in the relative writing coding. Among the types of language, we do not find coding languages since we do not intend to classify the type of digital data in C language, for example, by making a distinction between the acquisition of languages, such as short int or signed int. It is essential to understand and acquire the type of coding with which the data is treated, whereby coding we mean that process of attribution of meaning based on international predefined synthesis codes that assign an alphanumeric code to each alphabetic symbol (e.g., UTF-8, Unicode, ANSI). This knowledge is crucial and will be discussed as part of the methodology, as wrong language knowledge can lead to a loss of data both at the level of structure and at the level of content; in fact, Unicode type, with single code, although available for all types of processors, UTF-8 encoding can lead to the loss of the whole document if switching from Macintosh to Windows. The loss of the whole document, as some symbols are not decoded by the UTF-8 lists, leads to a corrupt document or, in the most common cases, to a deprecated document.

3.1.6 Brief History of Machine Learning

In the science methodology, well-known authors such as Blaise Pascal and Von Leibniz have always believed in the possibility that a machine, a programmed artificer created by men, could learn much more than a human being due to the limits of the human mind concerning the speed of processing information.

The following bullet points represent the historical excursus of the characterizing moments in the history of artificial learning.

[1]. In 1950, Alan Turing invented and conceived the “Turing Test” to determine whether artificial intelligence can have real intelligence. To pass the test, a computer must cheat a human being, making the subject in question believe that he is, in turn, a human being. Alan Turing is universally recognized as a founding thinker regarding computer science, artificial intelligence, machine learning, and computer science.

[2]. In 1952, Arthur Samuel wrote the first programming code for machine learning. The program was essentially based on the selection criteria that are implemented in the game of checkers, allowing the IBM company to grow exactly as the number of new players grows, deriving from these the winning strategies created by studying the moves that guaranteed the game's victory in fewer steps and by incorporating the optimal combinations within the program.

[3]. In 1957, Frank Rosenblatt designed the first neural network for computers: the Perceptron. This machine simulated the processes taking place within the human brain.

[4]. In 1967, the nearest neighbor algorithm was developed, an algorithm that allows computers to start using a reconnaissance pattern, although an initially very basic one. This algorithm can be used, for example, to draw a map and give directions to a person traveling, ensuring that, starting from a city chosen at random among those sampled, one can be sure to pass through each of them.

[5]. In 1979, some Stanford University students invented the “Stanford Cart,” an intelligence that independently recognized the obstacles placed on a path to be traveled.

[6]. 1981 is the year of Explanation Based Learning (EBL), in which a computer analyzes the process of generating data from learning dynamics and creates a general principle that can be followed, discarding irrelevant data.

[7]. In 1985, Terry Sejnowski invented NeTTalk, an intelligence that made it possible to learn to pronounce words through infants’ learning method.

[8]. The 1990s were the years of statistics, and indeed the years in which the focus on machine learning changed the approaches based on

learning models to models based on data trends. Statisticians and scholars began to create countless software to analyze large amounts of data.

[9]. In 2006, Geoffrey Hinton coined the term Deep Learning to explain how new computers can distinguish objects, text, images, and videos.

[10]. In 2010, Microsoft's Kinect was revealed to detect twenty human functions every thirty seconds, allowing users to interact with computers through movements and gestures made by users and detected by the machine.

3.1.7 Smart Data

Smart data's existence directly influences big data's future. Smart data is defined as those data extracted and obtained in massive series through big data and reworked to extract meaning from it, highlighting the synergistic capacity and versatility of this data. Undoubtedly, the extraction techniques will evolve, new algorithms will be formulated, and methods will be found to put together unstructured data of a very different nature.

The entry into the data of the study of semantics has now deeply marked the concept of transformation of big data into smart data; this is today a common situation that can be analyzed through various fields of analysis that are always in rapid evolution, among which must be mentioned cognitive computing, semantic graph database, and data lakes.

The addition of further data certainly would not be the ideal solution. While semantics' introduction can be defined as that step that provides value to the

data, trying to consider and treat the data in an intelligent way, including appropriate tools, with smart data, it is possible to create a model to map real data to a predefined model. Therefore, meaning is created from the model, and the transition to smart data is fundamental for the development of big data considering the vast numbers and complexity of such data that push analyses to ever tighter times. The extraction of meaning is a useful tool for understanding big data and processing such data to achieve the purpose of analysis.

3.2 Text analytics to analyze and interpret big data

3.2.1 Encoding

In the statistical analysis of textual data, the main problem is the identification of the meaning that characterizes the corpus, if any is present. Therefore, to recognize the possible meanings that a text can assume, it is essential to know all the parts and fragments that make it up.

First, the basic unit of any writing is the word, which can be learned from any dictionary of any language and is defined as the oral or written expression of a concept or information or the representation of an idea through a conventional reference. As highlighted in morphology, which is the part of linguistics that studies the grammatical structure of words, these can take on two forms: either that of a free morpheme, the smallest element of the word itself that can no longer be divided, or that formed by a sequence of linked morphemes. To define the word in more detail, it is also necessary to distinguish two elements that characterize it: the sign and the meaning.

The linguistic sign derives from the correlation between a signifier, distinguished from the linguistic form adopted of phonic (spoken) and graphic (written) type, and meaning, in turn distinct from the point of view of form and substance. Specifically, the signifier is the outer plane of language or the visual dimension that allows the inner plane, the meaning, to manifest itself perceptually and therefore to be read, seen, heard, or even touched.

The meaning of a word, on the other hand, derives from the terms that surround it (syntagmatic axis) but also from the identification and selection of other terms that could replace it in the same sentence without changing the sentence and, therefore, the meaning (paradigmatic axis).

Considering that we will deal with the statistical analysis of textual data, it is fundamental to know that the meaning of a text is formed by the system of meanings obtained from the co-occurrences in the entire corpus of textual data. All this derives from the fact that in textual statistics, unlike other analyses generically carried out on texts, the study of the graphic forms of words, i.e., of the sign, is conducted regardless of the meaning of the text units, but, above all, it is distinct and independent from that of language—that is, from sense.

Continuing to define the units of this analysis, the set of all the different graphic forms—of the words as they are written in the text—and of the headwords—the reduction of the words of the text to the corresponding word present in a dictionary of the language—which appears in a corpus is called the text vocabulary, and every single term in turn constitutes the breadth of the vocabulary.

The union of two words connected (subject and predicate) constitutes a minimal sentence with full meaning without necessarily being inserted in a verbal or situational context. Several propositions, which differ in coordinates and subordinates, constitute the periods; the set of several periods

give shape to the utterances; and, finally, the union of the latter makes up the text.

In the specific case of textual analysis, the text is the set of speech fragments, whose elementary units (words) are defined as occurrences and can be generated both from writings and from the transcription of oral speeches or the translation of specific codes. There are various types of texts that differ in content, some of which can be mentioned. Descriptive texts may contain either subjective descriptions of the writer or objective descriptions, technical or scientific descriptions, for example.

- [1]. Narrative texts, such as stories, reports, or topical texts.
- [2]. Argumentative texts, such as scientific arguments or comments.
- [3]. Informative texts, which can be verbal, textual interpretations, or explanatory essays.
- [4]. Conative texts, which contain regulations and instructions.
- [5]. Hypertexts, which are made up of parts of text linked together so that the reader can create his path.
- [6]. Documents, such as invoices, contracts, or certificates.

Regardless of genre, the text should be understood as the essential elements of one of the corpus's many possible partitions.

Analysis is meant to collect context units or fragments of the corpus—that is, any set of coherent, pertinent, and comparable writings under some point of interest or property. The corpus, like any other database definable as a complex of large amounts of homogeneous information, can be browsed in

different ways according to one's objectives, and each different reading generates, from the point of view of textual statistics, a set of profiles or lexical units that become the basis for the analysis.

This definition of the corpus applies to all the various existing textual sources regardless of whether they can be entire texts, documents, sections, or simple sentences. Texts are divided into literal and technical-scientific texts, political and institutional discourses, periodical press years, technical and sectoral documents, collections, or collections of short texts such as bibliographies, political posters, advertisements and headlines, field surveys such as non-directive interviews and life stories, transcripts of non-text messages and audiovisual products, and direct transcriptions of spoken language employing speech recognition.

3.2.2 Linguistics

The text's automatic analysis derives from the study of the different models taken into consideration under two different aspects, which in turn determine two kinds of analysis. These studies are based on the qualitative analysis and which, due to the objectivity of the measurements, is based on the quantitative one and are distinguished by the properties and characteristics of each.

[1]. Lexical analysis analyzes the text units, i.e., the words or lexemes making up the vocabulary, whereby lexeme we mean the basic unit of the

lexicon that can be a root (for example, the root written—in written, writer, written, writer), an autonomous word (for example father, knife, ladle) or a sequence of well-defined words that express a specific concept and cannot be replaced (for example: after lunch, for or not, beyond).

[2]. On the other hand, the textual analysis examines the context units: the written documents or fragments of the discourse making up the corpus as an entire set of occurrences.

The same and/or very similar operations can be applied to both.

3.2.3 *Lexical analysis*

Lexical analysis is the paradigmatic study—that is, of language—of a corpus. This is a *vertical* type of analysis, i.e., a verification in which the representation of the text derives exclusively from the extraction and examination of the individual words without considering the speech's development.

Lexical variety is an essential feature for a text:

- a. A corpus can be lexically varied, containing many different words with few repetitions, or conversely
- b. Use a reduced vocabulary, repeating the same terms numerous times.

Therefore, the size of the vocabulary provides the first bit of information on the variety of the text.

Generally, the following relation is used for the definition of the vocabulary:

$$V_i = V_1 + V_2 + V_3 + \dots + V_{fmax}$$

V_i indicates the number of different words that are present i times in the vocabulary. Consequently, V_1 represents the set of terms that appear only once, V_2 those that appear twice, V_3 those that occur three times, up to indicating with $V(fmax)$ the last set where $fmax$ expresses the value of the occurrences of the word with the most significant number of occurrences in the vocabulary. So, for example, comparing two texts A and B of the same length in words ($[N = 10,000]$), and knowing that A contains 2,000 different words ($[V(A) = 2,000]$), while B has 3,000 ($[V(B) = 3,000]$), text B is lexically more varied than A.

The comparison between texts of different lengths is complex; comparing a text C (10,000 words long) with D (50,000 words in length) and knowing that both have a vocabulary of width $V = 5,000$, it is not possible to say that they have the same variety, because in C a word is repeated twice on average twice ($10,000/5,000 = 2$), while the frequency in D is 10 ($50,000/5,000 = 10$). So, it is noted that, in D, there is more significant repetition and thus less variety.

The relationship, in the example just cited, between the number of occurrences (N), i.e., the length of the text, and the number of different words (V), i.e., the width of the vocabulary, also expresses the concept of the average frequency of words of a text, which is one of the first indices of lexical variety.

Therefore, the study of the vocabulary of a corpus consists of producing statistics and descriptions of some constants of the language, in terms of the percentage incidence of some classes of words able to differentiate the original texts, to identify their level and type.

The lexical analysis is carried out, at a later stage, on the morphology, or on the study of the grammatical forms of a language, traditionally grouped into classes, which are defined parts of speech: nouns, verbs, adverbs, and adjectives. Generally, this approach favors the theoretical and practical considerations that automatic text processing poses. The classes are divided into:

[1]. nouns, adjectives, and derived adverbs.

[2]. verbs.

[3]. other (e.g., articles, pronouns, prepositions, conjunctions, adverbs).

Available inventories characterize the first two classes as flexible words that change rapidly over time. The first is the most changeable, because standard terms that belong to it are born and die. Moreover, words that have limited flexibility are collected (singular or plural number for nouns, masculine or feminine gender and singular number or plural for adjectives, and a single form for derived adverbs). On the other hand, the second class is made up of innumerable forms of verbs, each of which can present many inflections and variations and is therefore moderately open, as new verbs are less frequent than nouns. Unlike the previous ones, the third class comprises closed inventories, consisting of a set of generally invariable elements: articles,

prepositions, pronouns, and primitive adverbs are the essential elements of grammar that never change. Subsequently, lexical analysis is based on the morphological productivity of a word—that is, on the ability to generate a variety of forms starting from its lexeme or root, belonging to certain morphological groups, such as verbal derivatives or enclitics connected to personal pronouns.

The level of lexical analysis, therefore, is that of words and how they appear in the text; at this point, we arrive at the lexical analysis, which consists of a complex of successive and parallel procedures, such as:

- [1]. Segmentation.
- [2]. Numbering.
- [3]. Indexing.
- [4]. Computing frequencies.
- [5]. The different arrangements of the vocabulary thus obtained.

Segmentation is the tool used to divide all the elementary signs that successively constitute the corpus of the text into two categories:

- i. The set of word-forming signs (i.e., the letters and other symbols that appear within a term).
- ii. The set of word-separator signs (i.e., spaces and punctuation marks).

The segmentation has the purpose of scanning the corpus, grapheme by grapheme, in search of the first word-forming sign. Once one finds the first word sign, he/she then takes this one and takes the others up to the next separator, which indicates the end of the word itself. Thus, the graphic forms,

or the single words cut out of the text, are gradually recorded each on a line, also indicating the position in which they were found in the text, which is called the address. The result of all this leads to the generation of a succession of registrations, which will then be organized in a table of the type “form, address”:

Table 4: Example of a table of a kind "form, address."

Form	Address
1	<i>textual</i>
2	<i>statistics</i>
3	<i>identifies</i>
4	<i>new</i>
5	<i>disciplinary</i>
6	<i>sectors</i>

3.2.4 Textual analysis

The first studies towards textual statistics date back to the period of the development of automatic text analysis. It is, in fact, between the 1950s and 1960s that Guiraud (1954), Herdan (1958; 1964), Yule (2014),¹⁷ and Zipf, (1935; 1949) pioneered a quantitative analysis approach in the linguistic field. To define the various stages of the development of textual statistical

¹⁷ Please note: the reference referred to Yule (2014) is referred to the current edition, reprinted from the original manuscript from 1944.

analysis, it is necessary to mention other authors who have approached the subject, including:

Benzecri (1963) carried out the first experiments of the 'analyse des données' based on the linguistic data learning. Chomsky (1965) moved away from Benzecri's ideas but approached what was done by Harris (1977) in his work, which was based on the statistical approach of natural language treatment. Muller (1973) & Lafon (1984) developed the classical indices and measurements of linguistic statistics; from the 1930s–50s, these indices have been exploited in the studies of language properties, such as lexemes, morphemes, and n-grams, both in linguistics and lexical statistics in which the analysis of language consists in learning the lemmas. Zampolli & De Mauro (1992) are the two Italian authors who, thanks to their particular interest in calculating the frequencies of headwords, give rise to the development of quantitative linguistics and statistical-linguistic resources. Following the 70s and 80s, thanks to the advent but above all to the evolution of information technology, the statistical analyses expressed on natural or textual data were subject to significant changes and developments that allowed the affirmation of automatic analysis texts and textual statistics. In recent years, the growing availability of computerized linguistic resources, i.e., the texts that can be consulted and analyzed online, has further contributed to developing this new family of techniques. The new theories and applications in text mining no longer focus only on the methods and tools of statistics but, given the strong multidisciplinary that characterizes them,

are strongly conditioned by the theoretical reference apparatus. Retracing the various phases, the objective of quantitative studies on the language has moved from a linguistic logic, developed up to the 1960s, to a lexical type, which established itself in the 1970s, to become then type analysis textual or lexical-textual in the 80s and 90s. The techniques and units of analysis covered by these studies have also progressively changed thanks to the development of software tools and supply chains for data processing.

The above excursus moved from the study of the classics of literature to that of non-texts or artificial texts. Non-texts are data expressed in natural language from different sources, such as field surveys (open questions or interviews) or analysis of short texts (posters, bibliographies, messages), grouped in a collection of documents constituting a corpus of textual data.

Finally, at the end of the 1980s, Lebart & Salem (1988) defined the boundaries of textual statistics based on the analysis for graphic forms and repeated segments and no longer for headwords. At the same time, they begin to design and create software for analyzing textual data.

In the study of linguistics, it is necessary to know that the text acquires a specific and different meaning from time to time as a series of factors influence it; the text has the characteristic of being contextualized both in the communicative situation and in the cultural scenario in which it is produced and consumed. All this, of course, is also reflected in the individual words that make up the document itself, giving rise to different types of relationships through which they are connected. Therefore, before delving

into the phases and arguments of the textual analysis process, it is appropriate to distinguish between the extra-linguistic context and the linguistic context.

The extra-linguistic environment, as the word itself expresses, is external to the text and refers to three central aspects: to the thematic field, or the conceptual field that forms the background of the text and that is necessary to know in order to understand and interpret the document itself correctly (for example, with the expressions political text, sports text, geometric text, chemical text, etc., the adjectives indicate the area of reference to the conceptual universe); the communicative situation, i.e., the physical situation in which the document is produced and consumed; and environmental culture, or the doctrine almost always completely implicit, shared, and generically familiar to the interlocutors and made up of a set of knowledge, traditions, world views, and artistic references.

The linguistic context (or internal context, or verbal context, or only context) of any textual unit, ranging from the single syllable to the sentence or the entire period of the document, is, on the other hand, internal, or intrinsic to the text. It is considered a larger part of the document that interacts with the textual segment itself by physical proximity or logical link.

At this stage, it is important to define and illustrate the content and tools of textual analysis. This process aims to provide a syntagmatic representation of the text and therefore examines all operations directly addressed to the corpus. In particular, it:

1. Analyzes the concepts of the document, identifies the essential ones, and answers the possible and complex questions posed on the text.
2. Highlights the fundamental elements found.
3. Distinguishes and classifies the text fragments based on the different categories of belonging to then create new textual variables that will increase the structured database's productivity.

The initial and primary phase of this kind of analysis is that of the study of concordances, which aims to reconstruct the meaning of the words within the context in which they are found and, consequently, to be useful for interpreting the text itself and for a semiautomatic analysis of contexts. Concordances are, conceptually, simple tools developed to distinguish graphic forms from two different points of view that still interact with each other based on the environmental relationships these units have with the language they belong to and on the verbal context in which they currently appear.

The first type of relationship is called paradigmatic, and the second is syntagmatic. This process is distinguished by the aim of analyzing the agreement from a semantic and syntactic point of view; it displays, in fact, first the relationships that a word has with other nearby textual units and, secondly, the meanings that the same word can assume when it is present several times in the corpus, thus creating a list of the different portions of text that occur before or after the chosen term.

At this point, it is appropriate to introduce the concept of co-frequency, which is linked to the study of frequencies and can be distinguished as absolute co-frequency $cofreq(X, Y)$, which indicates the number of co-occurrences in each text, and co-frequency relative $cofreq(X, Y)$, which represents the ratio between the absolute co-frequency and the length (N) of the text.

Finally, different approaches have been advanced to establish which are the most effective co-occurrences. Among these, Church & Hanks (1990) presented the measurement through Shannon's mutual information given by the formula:

$$I_i \stackrel{\text{def}}{=} -\log_b P_i = \log_b \frac{1}{P_i}$$

where I_i represents the self-information of the message x while b is the logarithmic base.

Two types of results can be obtained from this report. If $I(X, Y) = 1$, mutual information is null. This result expresses the independence between the two forms, so their co-frequency will tend to be equal to the product of the individual frequencies. Conversely, if $I(X, Y) > 1$, mutual information exists and is positive. This result highlights the positive association of the two forms, whereby the recurrence of one increases the probability of the other's recurrence.

The process is commonly known as Text Mining (TM) or Text Data Mining (TDM), and it is a specific application of text analysis that has established

itself and evolved over the last few years, representing one of the automatic forms of textual analysis.

3.2.5 Text Mining

The term TM refers to the analysis of the texts that can be consulted online, an analysis made possible by the joint use of IT and linguistic tools, without forgetting the statistical solutions. Textual or lexical-textual analysis developed between the 80s and 90s and led to an important change in the software sector; the progressive developments of which have been projected towards the management of knowledge and business intelligence.

To analyze texts online, software automatically analyzes a text using multidimensional methods independent of its size; it is the techniques of multidimensional analysis, such as correspondence analysis, cluster analysis, discriminant analysis, and multidimensional scaling, that solve the problem of ambiguity inherent in language.

The solution adopted to reduce the level of ambiguity of individual words consists in isolating polysemantic words and adverbs, prepositions, and adjectives; to group the most recurrent expressions, a frequency lexicon was developed, taking as a starting point a corpus of standard Italian texts.

The process of automatic text analysis, after the initial parsing, involves the carrying out of a series of phases which, integrated with each other, give rise to the so-called analysis chain.

The main steps to follow are 4:

1) Preparation of the text: an essential phase which consists in cleaning and normalizing the text and in associating (annotating the text) meta-information to the words.

2) Lexical analysis: a vertical type of analysis in which the representation of the text takes place by extracting the words; in this case, we are talking about a bag of words. During the second step, statistics are built on verbs, nouns, and adjectives, highlighting the most frequent and the particularly significant ones, and studying empty words such as punctuation or incipit of a sentence.

3) Information extraction: focuses on that percentage of the most significant terms in the distribution and selects the specific vocabulary. The latter corresponds to 12-15% of the most relevant vocabulary for conducting textual analysis.

4) Textual analysis: concerns all the activities carried out directly on the body of the text capable of providing a syntagmatic representation of the text. The distinction between textual analysis and lexical analysis is important; the first is applied to the documents or fragments of the speech, better known as context units constituting the corpus as a total set of occurrences, while the second instead is applied to the text units, or to the words that make up the vocabulary.

In this thesis, the interest concerning TM was to transform the set of unstructured job ads' texts into a set of structured text data. TM is a specific

application of text analysis that allows one to process unstructured data and, by creating data encoded in structured fields, allows you to extract information that creates value in the sense of business and competitive intelligence.

TM represents the most elaborate extension of Data Mining. It has been developing since the mid-90s and is especially essential for companies and institutions to cope with the excess of information stemming from the availability of IT resources (such as electronic dictionaries). TM simultaneously connects Information Retrieval and Information Extraction operations and is structured in different phases.

A) Phase of pre-processing of the texts (in which IT prevails): consists of retrieving the text sources from the web or intranet (e.g., news or press articles, website content, messages, chats, forums, or other document bases), in their formatting (e.g., transformation into XML), and in the establishment of the document warehouse.

B) Lexical processing phase (in which linguistics prevails): consists of recognizing words (with the use of dictionaries and knowledge bases, semantic networks, sensigraphs, or other methods), identifying already known keywords or concepts (with the use of rules and ontologies), and making lemmatizations (recognition of the main parts of speech, mostly nouns, adjectives, and verbs). This is not a necessary step for all applications because, sometimes, linguistic processing of the text is not carried out.

C) Real TM phase (in which Statistics and Data Mining techniques play a crucial role): consists of one or more of the following steps:

- i. Automatic categorization of documents for subsequent retrieval of information.
- ii. Search for entities (terms) in multilingual texts, regardless of the terms' language of origin. This presupposes the availability and alignment of specific linguistic resources in the various languages investigated.
- iii. Queries in natural language, interpreted by NLP processes also based on artificial intelligence algorithms.

Regarding the lexical processing phase, one of the operations that characterize it is lemmatization. This represents an algorithm process that can automatically determine the lemma of each single term given the word, where the lemma is the visual representation of all the inflected forms that a class of words can take. Finally, the concept of lemmatization is also linked to that of stemming—that is, the phase of reducing the inflected form of a given term in its root, which takes the name of the theme. As far as the application fields of TM are concerned, they are various, and some bear mentioning, such as customer relationship management, the classification and addressing of e-mails, the field of customer opinion surveys, the analysis of reports, questions and the level of customer satisfaction, human resource management, the survey on the opinions of employees, and the curriculum vitae that the company receives.

In conclusion, TM is used in all areas of investigation, consisting of both small and large amounts of data whose content one wants to know and understand and involves extracting only the relevant and interesting information ranging from publishing and media to telecommunications, such as websites and call centers, from those of banks, financial markets, public institutions, and public administration to those of the internet, such as automatic translators and language resources online, and to pharmaceutical and healthcare data.

3.3 Graph theories to visualize and construe employability skills

Among the various models proposed by graph theories, in this doctoral thesis the Author considers the random graph models for social networks proposed by Robins et al. in 2007. In this model, the occurring links between the nodes of a network are considered as random variables, and the dependency assumption regarding relations between these random variables determines the shape of exponential random graphs for a social network model. Examples of different dependency assumptions and their associated models are Bernoulli, independent dyads, and random Markov chart models.

Incorporating the attributes of the actors in the social selection of models is also revisited for an exponential model.

Lately, interest has grown in exponential random graph models for social networks, commonly called p^* model class. According to Robins et al. (2007), these probability models for social networks on a given set of actors allow for generalization, beyond the hypothesis of dyadic independence, of the class of the p^* model. Consequently, they allow models to be created from a more realistic construct of social behavior's structural foundations.

There have been more theoretical and technical developments in the literature since the presentation of the p^* models, also called Anderson models (Barnes, 1976). Several known techniques measure the properties of networks, nodes, or subsets of nodes, such as density and centrality measurements. These techniques offer valuable contributions in describing and understanding the characteristics of the network (Robins et al., 2007).

Furthermore, adding a small amount of randomness to an otherwise routine process makes it possible to modify the properties of the possible outcomes of that process, as demonstrated by Watts (1999).

Perhaps more importantly, a specific stochastic model allows one to understand the uncertainty associated with observed outcomes. According to Robins et al. (2007), it is possible to observe the distribution of possible outcomes for a given model or estimate the parameters of the hypothesized model for the observed data from which the data could have been generated to obtain quantitative estimates of the uncertainty associated with the estimate.

It is difficult to investigate such questions without a hybrid model in all but the simplest cases given by the overall results from the combinations of many small structures that are not immediately obvious, even from a qualitative point of view (Robins et al., 2007).

It is possible to overcome those questions crossing this micro-macro gap with locally specified models for social networks, often through simulation. This process is vital to develop models that can be estimated from the data and therefore are plausible and empirically formulated models. Many models in the networking literature are essential tools for simulation, hypothesis generation, and thought experiments. However, the main goal is to estimate the model parameters from the data and then evaluate how they adequately represent the data. Exponential random graphs are especially useful due to the distinctive value of their data-driven approach (Robins et al., 2007).

According to Robins et al. (2007), in placing an exponential random graph model for a social network, five necessary steps must be followed.

Step 1: Each network link is viewed as a random variable.

This step implies a stochastic frame with a fixed number of nodes. Assuming that a bond is a random variable does not imply that people form ad hoc bonds: some relationships may be more likely. With possible network links established to be random variables, some important notions need to be addressed.

For every i and j , which are unique numbers of a set N with n actors, having a random variable Y_{ij} where $Y_{ij} = 1$ if there is a network link between actor i and actor j , and $Y_{ij} = 0$ if there is no it is bond. Specifying y_{ij} as the observed value of the variable Y_{ij} and letting Y be the matrix of all variables with y the matrix of observed bonds, of course, y could also be constructed as a graph on the node-set N , with the set limit specified by (i, j) for which $y_{ij} = 1$.

Step 2: A hypothesis of dependence for the contingency between network variables.

This hypothesis incorporates local social processes that are assumed to be generated by network ties, i.e., bonds can be independent of each other, as people create social connections independently of other social bonds. This assumption is not very realistic. In the previous example of the school class with reciprocal processes in place, if student A likes student B, then it will be

likely that B likes A, thus having a dyadic dependence. The links also depend on attributes such as homophily in a class. These processes can be represented as a configuration of small graphs: a mutual bond or a bond between two girls.

Step 3: Addition hypothesis is applied to a particular form of the model.

It can be proven that well-defined dependence assumptions involve a particular class of models.

Each parameter corresponds to a network configuration, a small subset of possible network links (and/or actor attributes). These configurations are the structural features of interest. The model represents a distribution of random graphs that arise from localized patterns represented by bond configurations—i.e., a single bond is a configuration, as could be a mutual bond (in a direct graph), a triad, and a two-star. Parameters relating to each of these configurations in the observed graph can be included in the model.

Step 4: Simplification of the parameters through homogeneity or other constraints.

To define an exact model, one must reduce the number of parameters. This is often done by imposing homogeneity constraints.

Step 5: Estimate and interpret model parameters.

Estimation and interpretation are often the focus of research applications. This step is difficult if the dependence relationship between nodes is labile, as happens in models applied to reality. Having the estimates of the parameter

provides several advantages over having a statistical model for the network built from specific dependency assumptions. The parameter is estimated from the data observed on the network; in looking at the range of results of the network predicted by the model, an inference on the model parameters is feasible. It is possible to deduce if each parameter of the model is significantly different from zero and, thus, if the corresponding configuration is present in the observed graph to a greater or lesser extent than expected, given other parameter values (Robins et al., 2007).

Any random graph has the following formula:¹⁸

$$\Pr(\mathbf{Y} = \mathbf{y}) = \frac{1}{k} \exp \left\{ \sum_A^n \eta_A g_A(\mathbf{y}) \right\}$$

where:

η_A is the corresponding parameter of configuration A , and it is not equal to zero only if each pair of variables A is conditionally dependent; $g_A(\mathbf{y}) = \prod_{ij \in A} y_{ij}$ is the statistical network corresponding to configuration A ; $g_A(\mathbf{y}) = 1$ if the configuration is observed in the network \mathbf{y} but is otherwise equal to 0; k is the normalized quantity that ensures that 1 is the correct probability distribution.

All exponential graph models have this equation, which describes a general probability distribution of graphs with n nodes. The equation gives the

¹⁸ Robins, G., Pattison, P., Kalish, Y., & Lusher, D. (2007). An introduction to exponential random graph (p^*) models for social networks. *Social networks*, 29(2), 173-191.

probability of observing a graph y in this distribution, and this probability is dependent on the statistic $g_A(y)$ in the network y and on the various parameters not equal to 0, η_A for each configuration of A within the model.

Different dependency assumptions have the purpose of choosing different types of configurations relevant to the model, and the only configurations relevant to the model are those in which any possible link of the configuration is mutually contingent with one another. It is worth noting that if a set of possible bonds represents a configuration in the model, then this implies that any subset of possible bonds is also a configuration. Thus, the dependency model is crucial in understanding/constraining possible configurations in the model (Robins et al., 2007).

Configuration A refers to a subset of tie variables and corresponds to a small network subset. Applying the theory of dyadic dependence to a direct network, it will follow those reciprocity parameters in the model. In this case, one pattern in the model is a set of variables $\{Y_{12}, Y_{21}\}$, another $\{Y_{13}, Y_{31}\}$, and so on, with each dyad providing its configuration. If both bonds are present in the observed graph for each of these configurations, we could observe a mutual bond, so the configuration represents a type of network sub-structure observed in the graph y (Robins et al., 2007).

The statistical graph $g_A(y)$, on the other hand, tells if the configuration A is observable in the network y . For a reciprocity configuration A , that statistic tells us whether there are reciprocal links between the relevant pair of nodes or not. Nonetheless, the strength and direction of any bond's value will affect

the frequency with which the corresponding configuration is observed. If the parameter is large and positive, it is expected to observe the corresponding configuration in distribution graphs more frequently than a parameter equal to zero. Moreover, if a reciprocity parameter were large and positive, one would expect to see some reciprocal links in the observed network. Similarly, when a parameter is large and negative, a less frequent configuration than a zero parameter is expected (Robins et al., 2007).

Since there is an exponential term, these distributions refer to exponential graphic models. Random Markov charts are a particular class of exponential graphic models directly related to the Monte Carlo Markov chain (MCMC). The network analytic community also refers to p * exponential random models since they are a generalization of the dyadic models, of which $p1$ is an example.

The following sub-sections describe models and algorithms for exponential graphs that have been implemented in representing and analyzing employability skills.

3.3.1 Watts-Strogatz model and algorithm

The Watts-Strogatz model has as its conceptual basis Small World, a theory introduced by sociologist Stanley Milgram mainly useful for analyzing complex networks and small-world networks. The model considers two of the main characteristics of a complex network:

[1]. The clustering coefficient (CC), which measures the normality and the locality of the network. The higher the coefficient, the more the relationships will connect nodes that are close to each other. On the contrary, tending towards zero indicates relationships between nodes that are more distant from each other.

[2]. The distance between the vertices of the network.

The higher the CC, the more the average distance between two nodes is likely high, even if the inverse phenomenon usually occurs in the standard networks observed—that is, a high value of the CC and a low distance between nodes.

There are three formal preconceptions of the model:

[1]. A regular network with a reticular configuration has a strong aggregation, but it is not a small-world type network.

[2]. A random graph is a small-world network but does not have any aggregation.

[3]. The network must appear in a chaotic nature to keep the degree of separation between each node low.

Watts and Strogatz propose a hybrid model at the limit between a grid-like model and a random graph (Watts & Strogatz, 1998; Newman et al., 2002):

[1]. It starts from a ring with n arches.

[2]. Each vertex connects with the k closest to each ring.

[3]. Each vertex is 'rewound' with probability equal to p , one of the vertices is kept fixed, and a new target is chosen as another vertex, among all vertices, at random.

Watts-Strogatz algorithm:

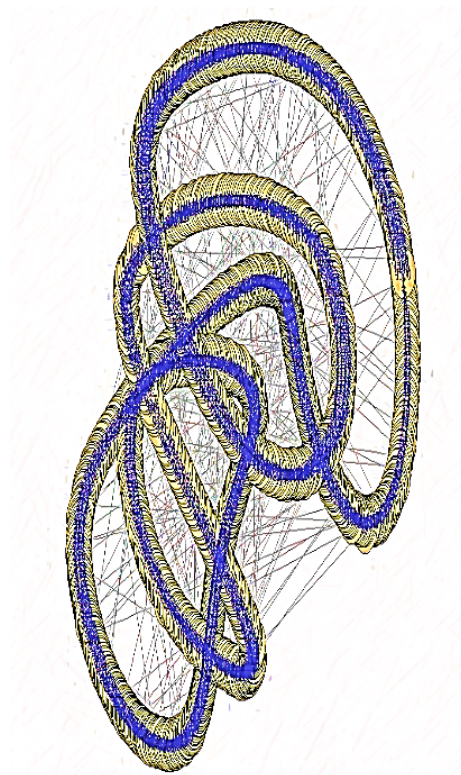
[1]. Let $V = \{V_1, V_2, \dots, V_n\}$ be the vertices of the graph.

[2]. Let K be an even value.

[3]. Let $n \gg k \gg \ln(n) \gg 1$.

- [4]. The V vertices are ordered as a ring.
- [5]. It connects each vertex to its first $k/2$ neighbors on the left on the ring clockwise and its neighboring $k/2$ s on the right counter-clockwise. This is how a G graph is defined.
- [6]. With probability p , we replace an arc $\langle u, v \rangle$ with an arc $\langle u, w \rangle$, where $w \neq u$ is randomly chosen from the vertices of the repeatedly reiterated graph G , such that $\langle u, w \rangle \in E(G)$.

Figure 2: Graphic representation of the model.



Algorithm 1: Watts-Strogatz algorithm application in R language.

```
set.seed(1)
avg = function(nei, p) {
  result = replicate(1000, {
    wsg = watts.strogatz.game(1, 100, nei, p)
    c(average.path.length(wsg),
      transitivity(wsg))
  })
  apply(result, 1, quantile, probs = c(0.5, 0.05, 0.95))
}
```

Above is a graphical representation of the Watts-Strogatz model on a random sample of $n=1500$, $k=10$, and $p=0.01$ computed via R software.

The model is at the base of the MCMC algorithm for simulation described in **Section 3.5** and of the three algorithms implemented for the graphical representation and analysis of employability skills, described in the following sub-sections, that are usually employed to represent small-word models (Prettejohn et al., 2011).

The below algorithms (DrL, Kamada-Kawai, and Fruchterman-Reingold) belong to force-directed graph-drawing algorithms, a particular class of algorithms for simplification and aesthetically pleasing graphs-drawing.

The main scope in implementing these algorithms is in positioning the graphs' nodes in a two-dimensional space so that all the ties are almost homogeneously sized, reducing crossing edges as much as possible. The algorithms employ

edges and nodes' position in assigning forces among them to simulate the motion of the edges and nodes, which are typically spring-like attractive forces based on Hooke's law (Grandjean, 2015).

There are many advantages in employing these algorithms, such as flexibility, good-quality results, intuitiveness, simplicity, interactivity, and strong theoretical foundations. In fact, there is strong research on the latter. Force-directed algorithms often appear in the literature and in daily life because of their intuitive nature and because they are easy to read.

In addition, there are two disadvantages to mention when employing force-directed graph-drawing algorithms: the high running time and the poor local minima. The second issue particularly affected this analysis, as in many cases the local minimum was found worse than the global minimum, providing low-quality drawing. This is because, in many algorithms, the output can be strongly influenced by the randomly generated initial layout. The problem of poor local minima becomes more consistent as the number of vertices of the graph increases. The literature suggests that a combined application of different algorithms is a helpful solution to this problem (Colberg et al., 2003).

As suggested in the literature, to methodologically overcome the problem, in this doctoral thesis the Kamada–Kawai (Kamada & Kawai, 1989) and DrL (Martin et al., 2007) algorithms were first employed to quickly generate a reasonable initial layout, and then the Fruchterman–Reingold algorithm (Fruchterman & Reingold, 1991) was employed to present the final outputs from the Social Network Analysis.

Algorithm 2: Usage and example of the Kamada-Kawai algorithm in R software.

```
layout_with_kk(  
  graph,  
  coords = NULL,  
  dim = 2,  
  maxiter = 50 * vcount(graph),  
  epsilon = 0,  
  kkconst = vcount(graph),  
  weights = NULL,  
  minx = NULL,  
  maxx = NULL,  
  miny = NULL,  
  maxy = NULL,  
  minz = NULL,  
  maxz = NULL,  
  niter,  
  sigma,  
  initemp,  
  coolexp,  
  start  
)  
with_kk()  
{  
  g = make_ring(10)  
  E(g)$weight <- rep(1:2, length.out = ecount(g))  
  plot(g, layout = layout_with_kk, edge.label = E(g)$weight)  
}
```

Algorithm 3: Usage and example of the Fruchterman–Reingold algorithm in R software.

```
layout_with_fr(  
  graph,  
  coords = NULL,  
  dim = 2,  
  niter = 500,  
  start.temp = sqrt(vcount(graph)),  
  grid = c("auto," "grid," "nogrid"),  
  weights = NULL,  
  minx = NULL,  
  maxx = NULL,  
  miny = NULL,  
  maxy = NULL,  
  minz = NULL,  
  maxz = NULL,  
  coolexp,  
  maxdelta,  
  area,  
  repulserad,  
  maxiter  
)  
with_fr(...)  
co=layout_with_fr(g, minx=minC, maxx=maxC,  
                  miny=minC, maxy=maxC)  
co[1,]  
plot(g, layout=co, vertex.size=30, edge.arrow.size=0.2,  
      vertex.label=c("ego", rep("", vcount(g)-1)), rescale=FALSE,  
      xlim=range(co[,1]), ylim=range(co[,2]), vertex.label.dist=0,  
      vertex.label.color="red")  
axis(1)  
axis(2)  
}
```

Algorithm 4: Usage and example of the DrL algorithm in R software.

```
layout_with_drl(  
  graph,  
  use.seed = FALSE,  
  seed = matrix(runif(vcount(graph) * 2), ncol = 2),  
  options = drl_defaults$default,  
  weights = E(graph)$weight,  
  fixed = NULL,  
  dim = 2  
)  
with_drl()  
{  
  g <- as.undirected(sample_pa(100, m=1))  
  l <- layout_with_drl(g, options=list(simmer.attraction=0))  
  plot(g, layout=l, vertex.size=3, vertex.label=NA)  
}
```

3.4 Data extraction techniques and architecture: Data collection and sampling

As reported in the introduction, Python software was employed to build the web scraper. An automated scraper was coded and developed via Beautiful Soup and requests packages. The packages find applications in many aspects of the scraper, as requests employ get function that allows parsing of the HTML page and the relative content. The output of get function is transferred to the beautifulsoup environment. Thus, beautifulsoup provides the function find, detecting the fitting tags responding to the pre-settled parameters in the name of the HTML tag and the relative tag's attributes (e.g., id, class, name) to mine the desired tag or HTML page to parse data.

Other crucial packages were RE and pandas, where re stands for regular expressions. re has been employed to process text and to remove occasional special characters. Pandas was utilized to build the final database in .csv format.

The scraping process in Python started loading libraries that were necessary to the development and running of the program (pandas, beautifulsoup, re, and requests). After that, a personal class for the scraping task was created and imported to keep the code clean. The name of the personal class was 'Scraper.' The exemplificative code below explains how the specific extrapolation process starts with the loading of the personal class Scraper and ends with writing each retrieved job. Data writing happens inside the personal class, providing this function thanks to pre-set codes.

Algorithm 1: Scraper's script main.

```
from scrapers.Scraper import Scraper
import pandas as pd
sects = pd.read_excel('./data/Careers.xlsx')
location = 'United States'
for sector in sects:
    *   jobs = sector['jobs']
        for job in jobs:
            Scraper.scrape_job_data(job, location)
```

The very core logic of the scraping program built-in Python language is contained in the implemented method `scrape_job_data` of the class `Scraper`, retrieving data from the job parameter filtered by the settled location (US). The method researches the open position named 'job,' geocoded with geographical criteria indicated by 'location' parameter.'

Algorithm 2: URL's formatting, page request, and relative parsing instance.

```
query_job = job.replace(' ', '+')
query_location = location.replace(' ', '+')
BASE_URL= 'https://www.jobportal.com/jobs?q=' + query_job + '&l=' +
query_location
page = requests.get(BASE_URL)
soup = BeautifulSoup(page.text, 'html.parser')
```

This block of codes is the initial part of the developed method `scrape_job_data` of the class `Scraper` described before. Particularly, in this phase it constructed the URL based on the input parameters represented by 'job' and 'location.' After that, through the implementation of the request library, it made a request to the constructed URL (`BASE_URL`) to obtain the HTML code of the website. Then, this HTML code is transferred to a `beautifulsoup` instance, enabling the HTML parsing, obtaining as output an object (`soup`) through which it is possible to conduct the effective detection of desired data and information. Software learns that for each symbol contained in the URLs and specified as a parameter.

Algorithm 3: URL's formatting, page request, and relative parsing instance.

```
job_type_menu = soup.find(name='ul', attrs={'id': 'job-type-menu'})
rows = job_type_menu.find_all(name='li', attrs={'class': 'menu-option'})
jobs_type = []
for row in rows:
    job_type = row.find(name='span', attrs={'class': 'label'})
    jobs_type.append(job_type.text)
ul_exp = soup.find(name='ul', attrs={'id': 'experience-level-menu'})
rows = ul_exp.find_all(name='li', attrs={'class': 'menu-option'})
exp_levels = []
for row in rows:
    exp_level = row.find(name='span', attrs={'class': 'label'})
    exp_levels.append((exp_level.text))
```

The above code concern finding the different types (e.g., full-time, part-time, stage) of the given job and the different level of experience (e.g., no experience, medium, high) that the platform provides for the given job. The above code sheds light on how the BeautifulSoup object, called soup, is used to interact and parse the HTML given as parameter in the constructor. In detail, the interaction is carried out by a find function that takes as its input parameter a tag name (name) and the attributes (attrs) to identify the right tag we are looking for, the ul tag representing a list of options. The result of the find is another BeautifulSoup object called the find_all function, which finds all the list options containing the desired data, in this case the aforementioned job types. At the end, the text found in each list option tag (li) is added to a list named job_types. The same approach is used for the extraction of the experience level.

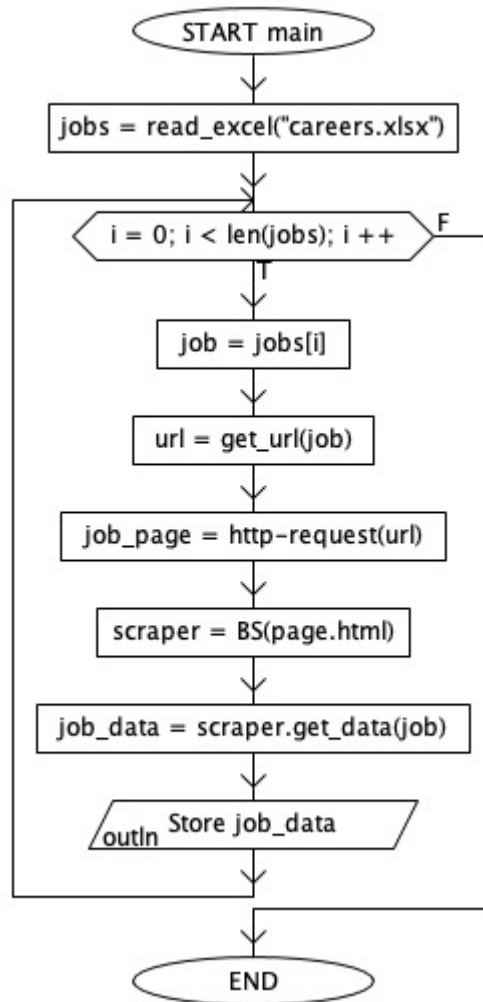
The following code block describes how, in general, data scraping is performed. Once it has obtained the list of all types of jobs and all experience levels for the searched job, for each job type and for each level of experience it constructs the detailed URL to get the web page of all jobs filtered by type and experience level. After that, through the requests library, the web page HTML is obtained and passed to the BeautifulSoup constructor to parse it. The BeautifulSoup object (soup) extracts from the web page all the job advertisements, then for each job found uses the data manager object to store the scraped data on Firebase.

The functions `get_all_jobs()` and `extract_job_data()` are not detailed here for copyright reasons reported in **Section 1.4**.

Algorithm 4: URL's formatting, page request, and relative parsing instance.

```
data_manager = DataManager()
for job_type in jobs_type:
    url = BASE_URL + job_type
    for exp_level in exp_levels:
        final_url = url + exp_level
        web_page = requests.get(final_url)
        soup = BeautifulSoup(web_page)
        jobs = get_all_jobs_ads(soup)
        for job in jobs:
            data_manager.insert(extract_job_data(job))
```

Fig 3: Macroscopic system structure of the web scraper.



3.5 Skills profiling techniques and architecture for data treatment

As reported in the introduction, R software was employed to profile data to analyze exhaustively the job ads' derived data.

The text was treated via several packages as tm and qdap.

First, data was uploaded and encoded. The following block of codes describes the loading and encoding process via gsub function and qdap package.

Algorithm 5: Data loading and encoding.

```
ads = read.csv(job_data.csv, sep=";", header=T)
desc = as.character(ads$Description)
desc = gsub("[]", ads, ignore.case = T)
desc = gsub("'", ads, ignore.case = T)
```

After that, data was first cleaned using regular expressions. The following algorithm shows how data was preliminarily cleaned, structured, and encoded under UTF-8 parameters, building an *ad hoc* function named cleanString.

Algorithm 6: Data treatment with regular expressions and cleanStrings function.¹⁹

```
cleanStrings = function(s) {  
  s = gsub("", "", s, perl=T)  
  s = iconv(s, "", "?")  
  s = gsub("[^[]", " ", s)  
  s = tolower(s)  
  return(s)  
}  
desc = cleanStrings(desc)
```

The above code assigns to the function cleanStrings a set of operations and instructions to manipulate data. This passage is useful in standardizing text, ensuring the remotion of unproper terms, strange characters, and blank spaces.

Subsequently, data was listed and unlisted to facilitate its processing. The following code describes the listing process.

Algorithm 7: List/unlist.

```
a = list()  
for (i in seq_along(desc)) {  
  a[i] = gettext(desc[[i]][[1]])  
}  
desc = unlist(a)
```

¹⁹ The complete usage of cleanStrings function has been omitted and shown in an exemplificative form.

After unlisting data, the text was refined via tm package to build a clean corpus. The following algorithm demonstrates the text refining and the corpus construction.

Algorithm 8: Text refining and corpus building.

```
desc = stripWhitespace(desc)
desc = bracketX(desc)
desc = replace_number(desc)
desc = replace_abbreviation(desc)
desc = replace_contraction(desc)
desc = replace_symbol(desc)
desc = apply(as.matrix(desc),1, function(x){paste(" ", x, " ")})
corpus = Corpus(VectorSource(desc))
corpus
inspect(corpus)
```

Once the corpus was built, several data operations were implemented to give a proper sense to the data, and a cleaned corpus was built. URLs were eliminated, punctuation deleted, text standardized, and English *stopwords* removed.

The following algorithm shows the text refining with the tm package, employing tm_map function, gsub function, and its content transformer.

Algorithm 9: Corpus cleaning and structuring process.²⁰

```
corpus = tm_map(corpus, content_transformer(gsub), pattern=" %amp",
replacement = " ")
inspect(corpus)
removeURL = function(x) gsub()
removeURL1 = function(x) gsub()
corpus = tm_map(corpus, content_transformer(removeURL))
corpus = tm_map(corpus, content_transformer(removeURL1))
corpus = tm_map(corpus, removePunctuation)
corpus = tm_map(corpus, tolower)
inspect(corpus)
corpus = tm_map(corpus, removeWords, stopwords("en"))
inspect(corpus)
class(corpus)
```

After the phase of corpus refining and the removal of specific residuals derived from text cleaning, the text was ready to be analyzed. To begin, stylometry was performed on the text. To conduct the stylometry, the *stylo* package was employed (Eder et al., 2016). First, unigram and bigram datasets were constructed, ordered, and then plotted to visualize results from the stylometry.

After stylometry, the analysis moves to the construction of a tokenized document-term matrix (DTM). To tokenize the DTM, a specific function was coded to combine terms when the corpus is transformed to a DTM. The function

²⁰ The complete usage of `removeURL` and `removeURL1` functions has been omitted and shown in an exemplificative form.

(NPL_tokenizer) collapses words into a terminological combination and their respective occurrences. The DTM will be employed for the topic modeling and for further textual network analysis.

The following algorithm shows the stylometry process.

Algorithm 10: Stylometry application and visualization.

```
unig = data.frame(table(make.ngrams(corpus,
  ngram.size = 1)))
dig = data.frame(table(make.ngrams(corpus,
  ngram.size = 2)))
sort_unig = unig[order(unig),]
sort_dig = dig[order(dig),]
top24dig = sort_dig[1:24,]
ggplot (top24dig) +
  geom_bar() +
  geom_text() +
  xlab() +
  ylab() +
  theme (axis.text.x = element_text())
```

Algorithm 11: Document-Term Matrix construction.

```
dtm = DocumentTermMatrix(corpus, control_list_ngram)
```

Once the DTM was built, topic modeling was conducted on the DTM after a proper sparsity reduction. Each corpus of ads required different sparsity treatments to guarantee the same number of terms (twenty-four) in the matrix.

Algorithm 12: Topic modeling with LDA application.

```
res_lda = LDA(dtm, k, control=list())
summary(res_lda)
topics = tidy(res_lda, matrix = "beta")
topics
```

Algorithm 13: DTM conversion to adjacency matrix and weighted graph object construction.

```
dtm2 = dtm %*% t(dtm)
g1 = graph.adjacency(dtm2, "undirected", weighted=T, diag=F)
g1$edge.width = E(g1)$weight
g2 = simplify(g1, remove.multiple = T, remove.loops = T,
edge.attr.comb=igraph_opt())
```

After topic modeling was conducted and correlation analysis was performed to detect the strongest linear relations in the model, graph theories were applied to the extracted skills (Robins et al., 2007). To transform the DTM into a graph object, the matrix (in the form of an affiliation matrix) was converted into an adjacency matrix, as shown in **Algorithm 13**.

The squared matrix was then transformed into a weighted, undirected, exponential random graph via the *igraph* package (Csardi, 2013).

After the skill sets were plotted, the modularity of the graphs was computed following three different methods: greedy modularity (Brandes et al., 2007; Schuetz et al., 2008; Ovelgönne et al., 2010; Chen et al., 2014), spectral modularity (Newman, 2006; Van Mieghem et al., 2010; Nadakuditi & Newman,

2012; Newman, 2013; Sarkar et al., 2014), and optimal modularity (Arenas et al., 2007; Görke et al., 2013; Nematzadeh et al., 2014).

Algorithm 14: Modularity detection and comparison.

```
B = modularity_matrix(g2, membership(c1))
round(B[1,],2)
membership(c1)
C = modularity_matrix(g2, membership(c2))
round(C[1,],2)
membership(c2)
D = modularity_matrix(g2, membership(c3))
round(D[1,],2)
membership(c3)
EG = round(modularity(c1) / modularity(c3),3)
ES = round(modularity(c2) / modularity(c3),3)
EB = round(modularity(c3) / modularity(c2),3)
```

In the literature, the modularity is described “as the fraction of the edges that fall within the given groups minus the expected fraction if edges were distributed at random”²¹ (Dekker et al., 2019, p. 6; Newman, 2006). To compare the performances of modularity detections, three indicators were constructed as follows:

²¹ Dekker, M., Panja, D., Dijkstra, H., & Dekker, S. (2019). Predicting transitions across macroscopic states for railway systems. PLoS One, 14(6), 1-26, e0217710.

$$\xi_G = \frac{\textit{Greedy Modularity}}{\textit{Optimal Modularity}}$$

representing the performance indicator for greedy modularity methods;

$$\xi_S = \frac{\textit{Spectral Modularity}}{\textit{Optimal Modularity}}$$

representing the performance indicator for spectral modularity methods; and

$$\xi_O = \frac{\textit{Optimal Modularity}}{\textit{Spectral Modularity}}$$

representing the performance indicator for optimal modularity methods.

In the construction of ξ_O , spectral modularity appears at the denominator because, based on empirical observation, in the model the greedy modularity tends to one. This tendency could generate a distortion of the indicator's value. The best performing indicator will be employed in selecting the dendrogram for skills' clustering.

Algorithm 15: Centrality measures computation.

```
centRes = centrality(g2)
centRes$OutDegree
centRes$Closeness
centRes$Betweenness
centRes$Strength
clusteringPlot(g2)
clusteringTable(g2)
```

Algorithm 15 exploits the computation of centrality measures of the model, considering the Out Degree, the Closeness, the Betweenness, the node's strength, and the normalized degree.

Algorithm 16: Monte Carlo Markov Chain simulation employing text generation methods.

```
fit_markov = markovchainFit(text_term, method="map")
for (i in 1:24) {
  set.seed(i)
  markovchainSequence(n,
    markovchain,
    t0, include.t0 = T) %>%
  paste() %>%
  str_replace_all(pattern, replacement) %>%
  str_to_sentence() %>%
  print()
}
```

After computing measures, skills have been clustered comparing three methods: Zhang (Zhang et al., 1996; Zhang et al., 1997; Sheikholeslami et al., 1998; Zhang et al., 2006), Onnela (Onnela et al., 2002; Onnela et al., 2004; Heimo et al., 2007), and Barrat (Barrat et al., 2008).

After computing measures and plots, a simulation of an ideal job interview based on the skills mining has been developed through the implementation of Monte Carlo Markov methods, generating a Monte Carlo Markov Chain (MCMC).

Algorithm 16 considers t_0 as the most frequent term, including it in the text generation and computing probabilities in the speech starting from t_0 . To generate the MCMC, maximum a posteriori probability (MAP) estimation models have been employed. The MAP method is part of Bayesian statistics. The MAP estimate is usually employed in obtaining a point estimate of an unobserved quantity from empirical data observation and is closely related to the method of maximum likelihood estimation (MLE). The difference is that MAP employs an augmented optimization having a prior distribution. It quantifies all the information available through the prior knowledge of a correlated event over the estimate. The MAP estimation could be grasped as a regularization of the MLE. Thus, MAP is considered the most fitting method for the purpose of this doctoral thesis if compared to other methods implementable for an MCMC (e.g., MLE, bootstrap, or La Place method).

To describe this method, one should start by assuming that the analysis is aimed at estimating an unobserved population parameter θ with observations x . f should be the sampling distribution of x , since $f(x | \theta)$ is the probability of x when the underlying population parameter is θ . Then the function

$$\theta \rightarrow f(x | \theta)$$

is the likelihood function, and the estimate

$$\hat{\theta}_{MLE}(x) = \arg \max_{\theta} f(x | \theta)$$

is the maximum likelihood estimate of θ .

If a prior distribution g over θ exists, θ is considered a random variable, as in Bayesian statistics. It is possible to compute the posterior distribution of θ employing Bayes's theorem:

$$p(\theta | x) = \frac{f(x | \theta)g(\theta)}{\int_{\theta} f(x | \vartheta) g(\vartheta)d\vartheta}$$

where g is the density function of θ , and θ is the parametric domain of g .

The MAP method then estimates θ as the mode of its posterior distribution:²²

$$\begin{aligned} \hat{\theta}_{MAP}(x) &= \arg \max_{\theta} f(\theta | x) \\ &= \arg \max_{\theta} \frac{f(x | \theta)g(\theta)}{\int_{\theta} f(x | \vartheta)g(\vartheta)d\vartheta} \\ &= \arg \max_{\theta} f(x | \theta)g(\theta) \end{aligned}$$

The denominator of the posterior distribution²³ is always positive, is not dependent on θ , and has no influence during the optimization. Therefore, the MAP estimate of θ coincides with the MLE when the prior g is uniform (e.g., g is a constant function).

The loss function of MAP optimization is of the form:

$$L(\theta | a) = \begin{cases} 0, & \text{if } |a - \theta| < c \\ 1, & \text{otherwise} \end{cases}$$

²² θ is treated as a random variable.

²³ Cf. Marginal likelihood.

When c goes to zero, the Bayes estimator approaches the MAP estimator if the distribution of θ is quasi-concave.

Empirically, a MAP estimator is not a Bayes estimator unless θ is discrete.

After the theoretical explanation of the research methodology, the Author is going to show the macroscopic system structure of the profiling software in **Figure 4**, and the overall software architecture in **Figure 5**. Results are discussed in **Section 4**.

Figure 4: Macroscopic system structure of the profiling software.

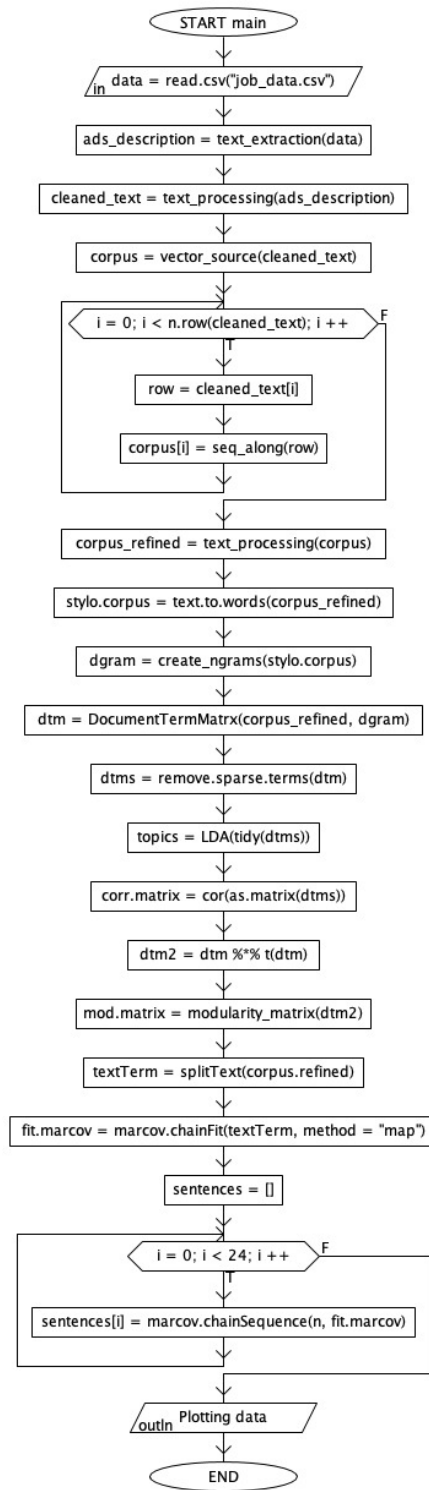
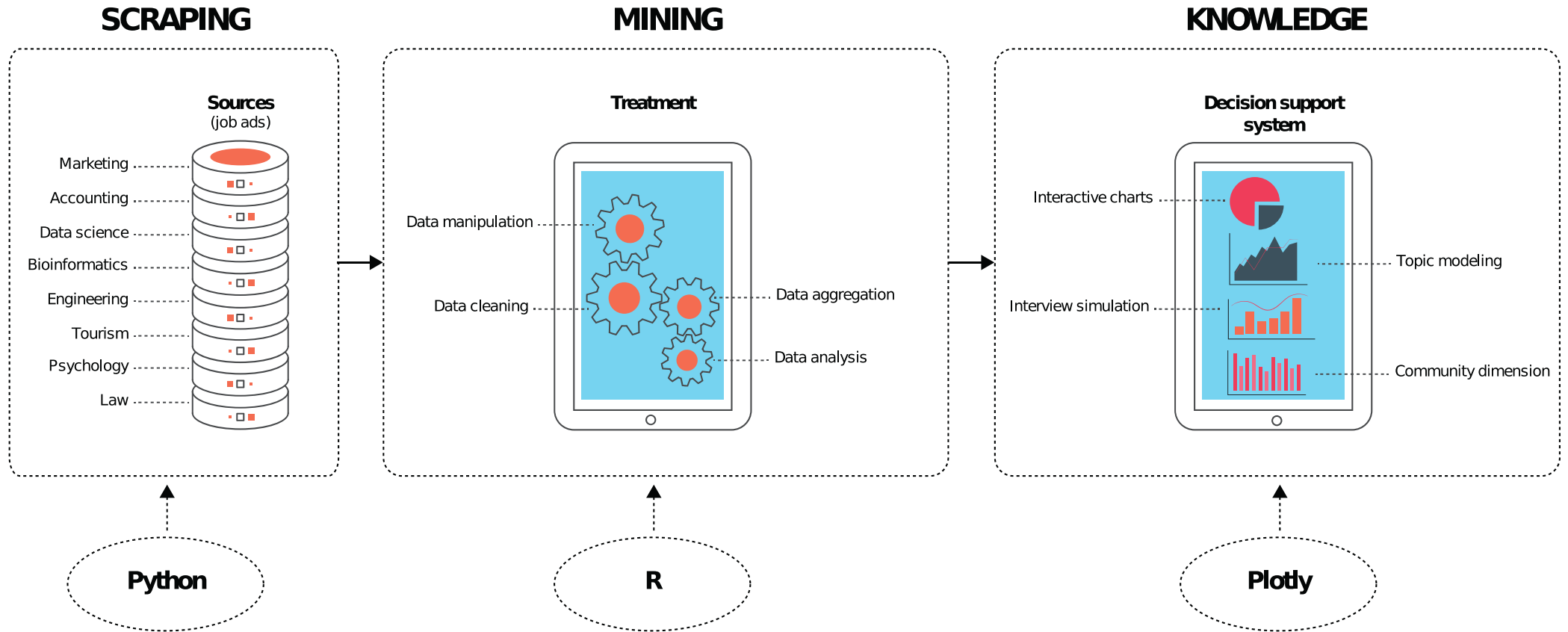


Fig 5: Overall software architecture.



4. Results

The results section reports the main figures from the analysis. The section is structured as follows: the first paragraph is dedicated to data collection and sampling exposure; then, for each sector comments on the reported statistics; and then the output of the analysis is sequentially presented. The first output of each sector regards the bigrams from the corpus, the second shows topic modeling, and in the third we have the corrplot, followed by SNA applications and techniques (network representation with modularity detection methods, dendrogram, clustering lines plot, and centrality measures). Lastly, MCMC is presented to simulate the job interview.

4.1 Data collection and sampling

Data mining has been conducted on the Indeed portal to extract and process data (Blázquez Soriano et al., 2012). Web scraping, defined in the literature as “a technique of digital data extraction from a website through a software parsing in real-time instances wide-spreading hyperlinks reality” (Munzert et al., 2014, p. 11), was carried out on a population of 144k job ads, extracting almost 71k ads from the Indeed portal in the American (US) market, recording 14.6kk terms in 70,449 documents, excluding invalid cases. The sampling was conducted between 1 Nov. 2019 and 31 May 2020. Furthermore, the data collection has been conducted on the American job ads listed on the Indeed portal for several reasons, as, for example, the higher

availability of job ads in the US market in relation to other markets presenting fewer open positions in the market. The choice of the American market (US) is then motivated by the fact that, concerning the keyword “data scientist” for job vacancy searches on the Indeed research portal, the American market offers around 60k more job listings than the English, French and Italian markets (144k observations in the American market against 30k observations in the English market, 25k observations in the French market, and just over 8k for the Italian market).

According to the rankings (<https://www.roberthalf.com/blog/job-market/10-best-job-search-websites>), adequacy is ensured because Indeed is classified as the third-best job search website in the US, as well as the most accessible website for web scraping to the exerted software (Liu et al., 2010).

After choosing the population, the existence of job titles as a pre-defined keyword for the opened job vacancies was verified to avoid a self-selection bias. Then, the sampling method to employ was defined. Based on the literature (Madow, 1949), systematic geographical sampling was employed, choosing the first sample according to random sampling and then keeping a fixed interval of four, thus ensuring reliability.

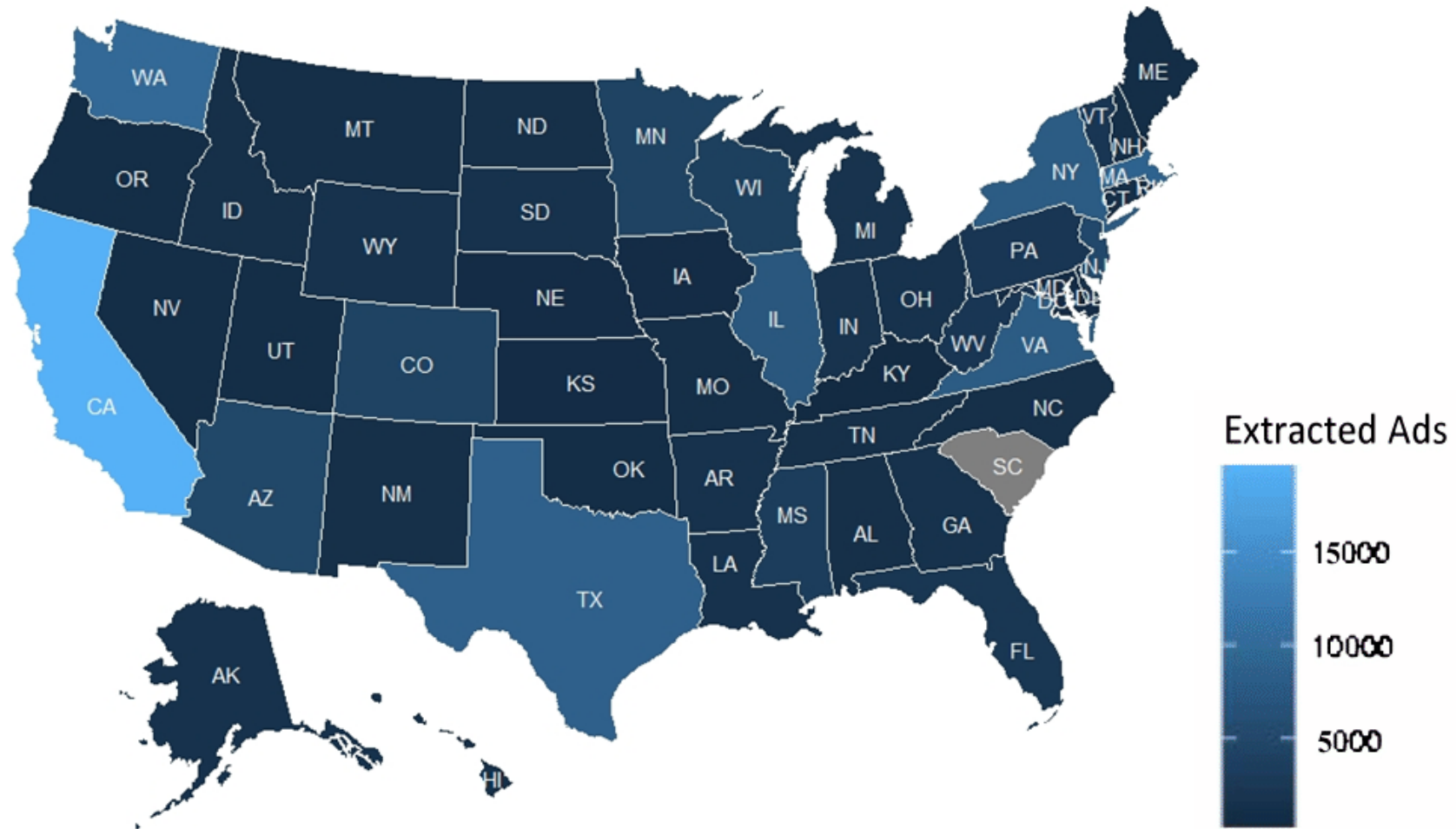
The observations were extracted in five dimensions: firm, location, open position, job description, and recommendations. According to the research aim, the contents were analyzed to profile skills for the exploratory study of the job description.

To extract information from Indeed, a two-step process was conducted: 1) detecting specific page URLs and 2) extraction of information and target data completing the pre-designed software task.

Table 5: Figures from the sampling.

Industry	Job 1	Job 2	Job 3	Job	Job 5
Marketing, Sales, & Development	Marketing Manager 2,645	Brand Manager 1,425	Presales Consultant 1,436	Market Analyst 1,489	Digital Marketing Manager 1,789
Accounting & Finance	Actuary 2,352	External Auditor 2,245	Forensic Accountant 1,756	Private Banker 1,896	Stockbroker 1,652
Leisure, Travel, & Tourism	Destination Manager 2,423	Tourism Business Consultant 1,806	Digital Event Manager 1,502	Reservations Manager 1,918	Arts & Hospitality Manager 1,523
Data Science & Artificial Intelligence	Data Scientist 1,856	Cybersecurity Specialist 1,326	Blockchain Specialist 1,465	big data Manager 2,228	Growth Hacker 1,899
Bioinformatics & Genetics	Bioinformatician 1,736	Computational Biologist 1,326	Genomics Programmer 2,051	Pharmacovigilance Specialist 1,142	Genome Architect 1,806
Engineering & Cloud Computing	Mechatronic 1,704	Software Engineer 1,365	Development operations engineer 1,789	Full Stack developer 1,564	Additive Manufacturing Engineer 1,896
Law	Attorney 2,352	Corporate Attorney 1,956	Corporate Lawyer 1,555	Securities Lawyer 1,829	Tax Law Attorney 1,736
Psychology	Psychologist 1,545	Psychoanalyst 1,899	Work Psychologist 1,485	Mental Health Counsellor 1,689	Corporate Counsellor 1,484

Fig. 6: Concentration map of the geographical sampling.



Source: own elaboration.

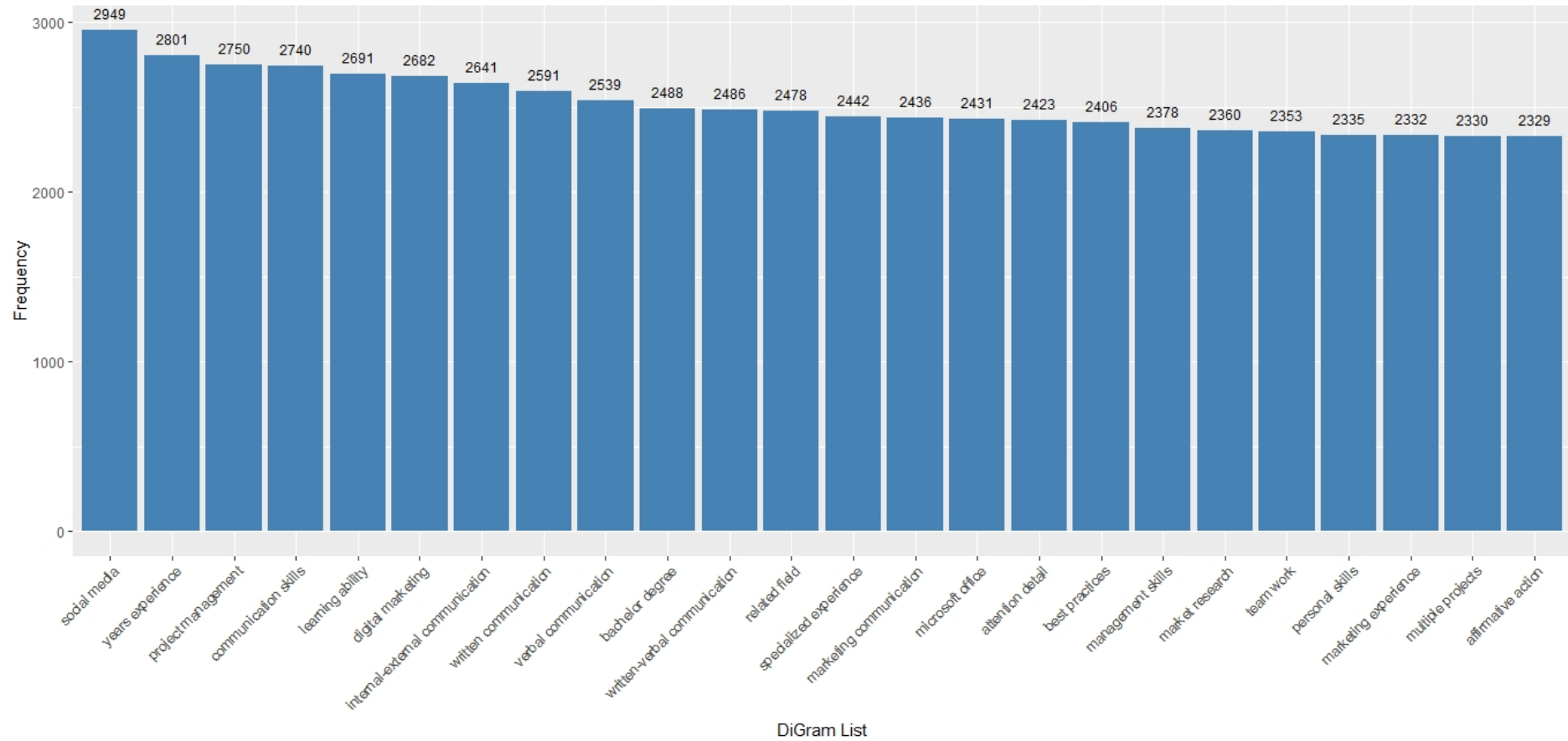
4.2 Results from the profiling process

4.2.1 Marketing

Results from the marketing industry have been obtained analyzing the subset corpus from the extracted ads regarding the sector. A tokenized Document-Term Matrix (DTM) has been built, and sparsity was removed till 78%. **Fig. 7** shows the bigrams from the corpus. The most frequent terminological combinations were social media (2949), years experience (2801), project management (2750), communication skills (2740), and learning ability (2691). Topic modeling is presented in **Fig. 8** with four thematic areas. **Fig. 9** highlights the main correlations through the skills set. **Fig. 10** detects greedy modularity in the skillset, dividing it in three groups, and the relative memberships are shown in **Fig. 11**. Application of spectral modularity is presented in **Fig. 12**, and the relative memberships are reported in **Fig. 13**. The employment of optimal modularity detection is shown in **Fig. 14** and their memberships highlighted in **Fig. 15**. Modularity indicators were compared to define the most proper method to give sense to the analysis. Having $\xi_G > \xi_O > \xi_S$, the dendrogram in **Fig. 16** was built with greedy modularity, and partial correlations were used for the weighted network in **Fig. 17**. Thus, a clustering plot with Zhang, Onnela, and Barrat methods is reported in **Fig. 18**. Centrality measures are exposed in **Fig. 19**. The most between skills in the set were project management (8.3%), communication skills (8%), learning ability (2.3%), best practices (1.6%), and written

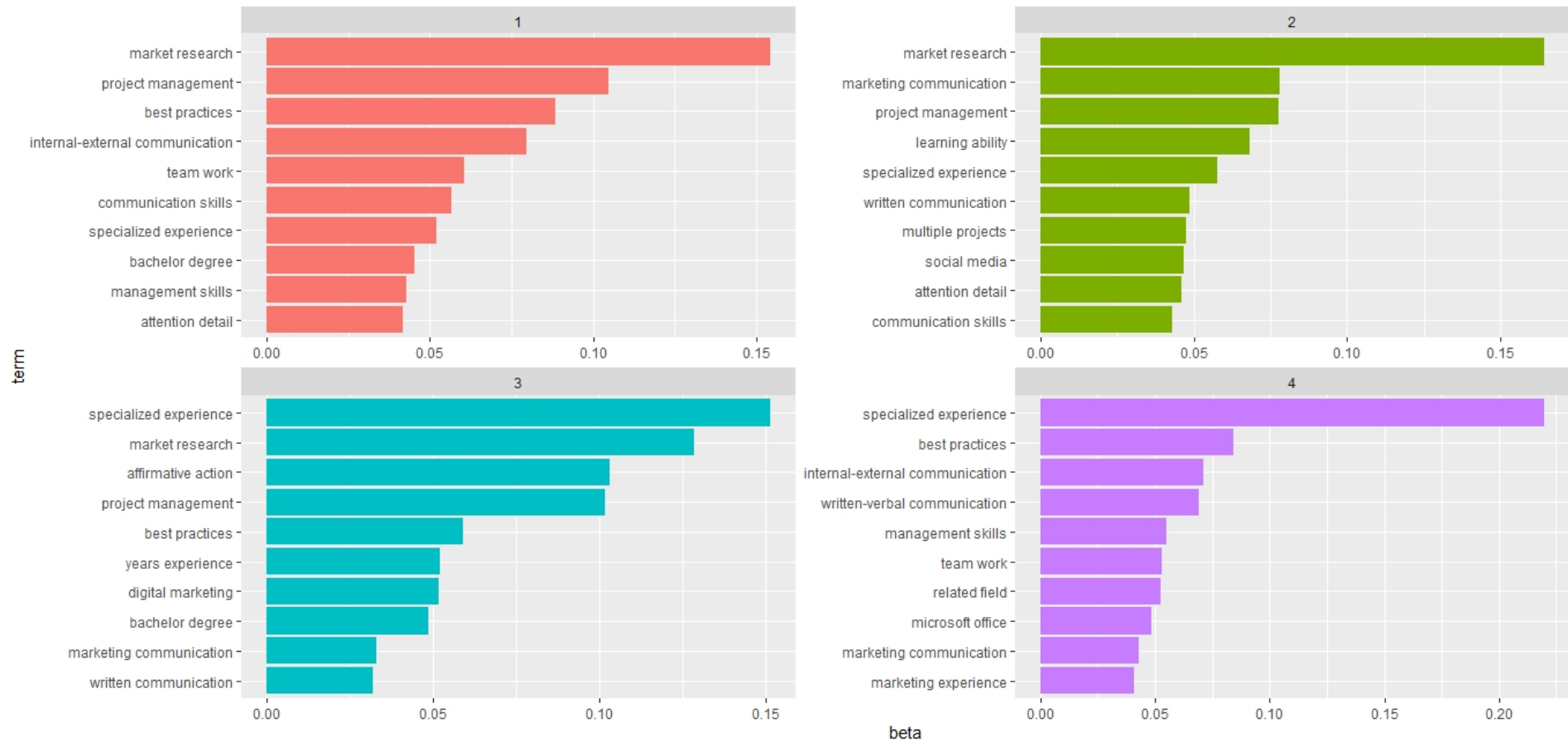
communication (0.3%). The closest skills were learning ability (18%), communication skills (16.5%), project management (16.2%), internal-external communication (14.7%), and written communication (1.5%). MCMC with MAP method is shown in **Fig. 20** to forecast and simulate a possible job interview for the Marketing & Sales industry.

Fig 7: Bigrams of the Marketing skillset.



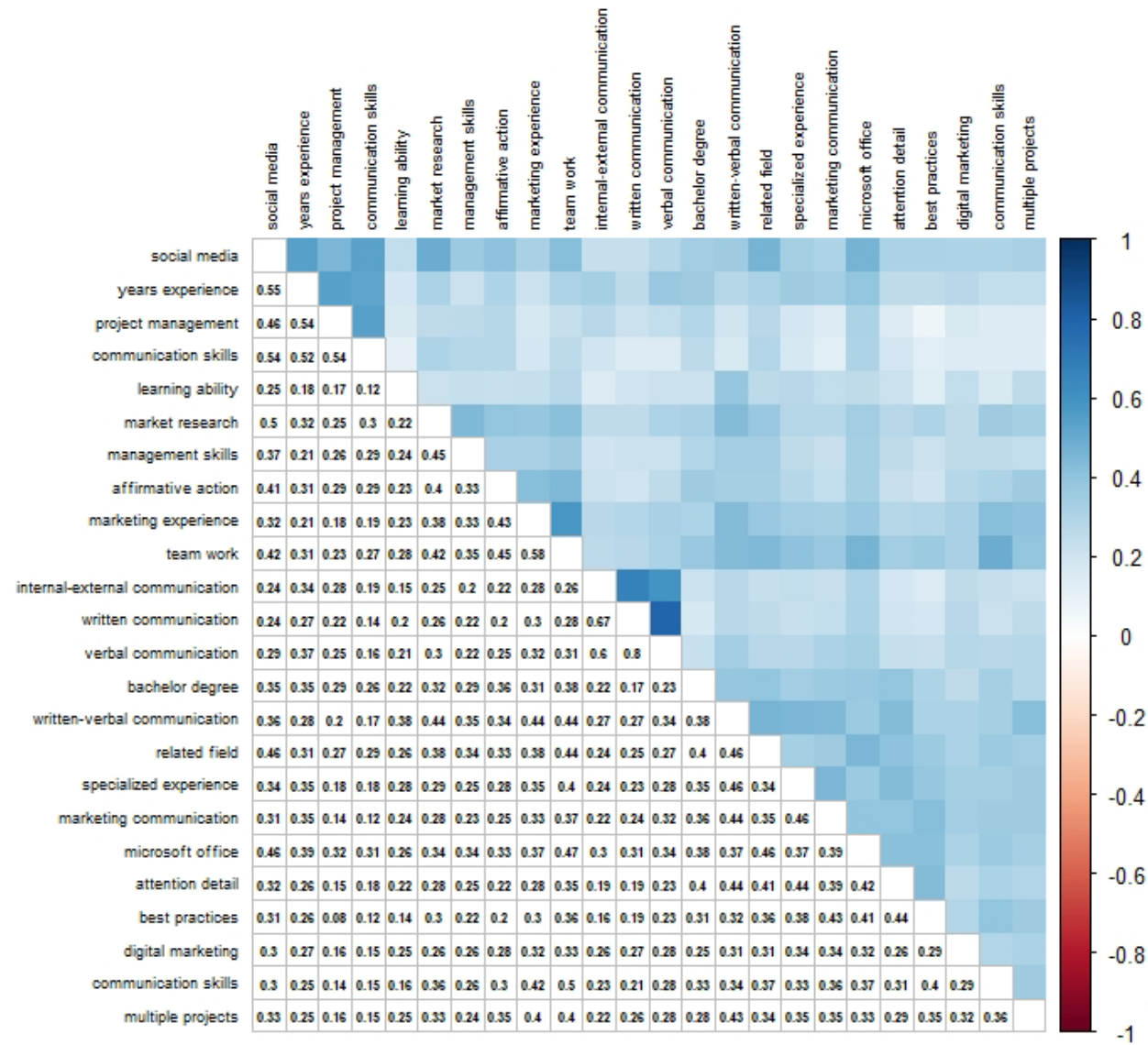
Source: own elaboration.

Fig. 8: Topic modeling of the Marketing skillset.



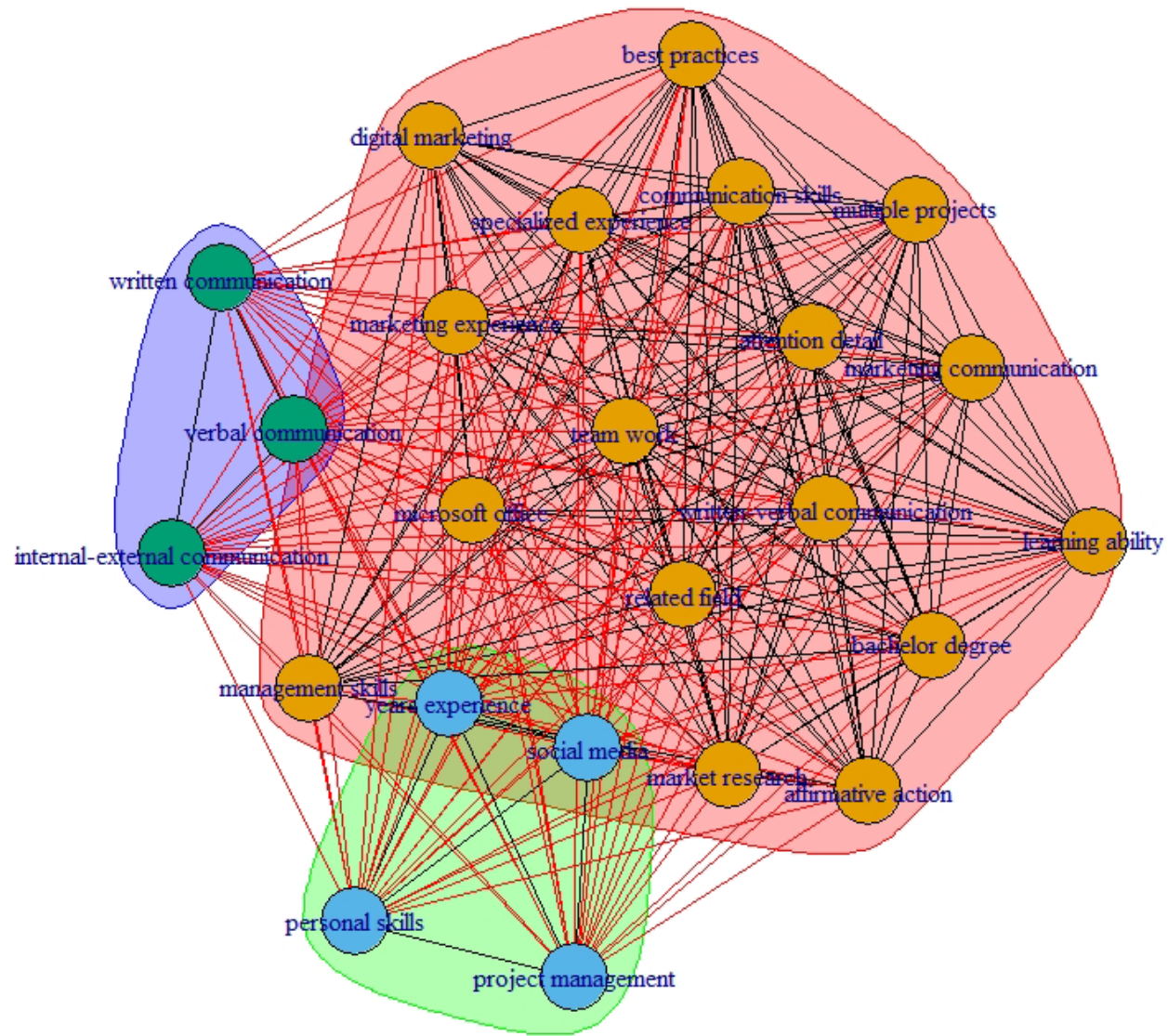
Source: own elaboration.

Fig 9: Corrplot of the Marketing skillset.



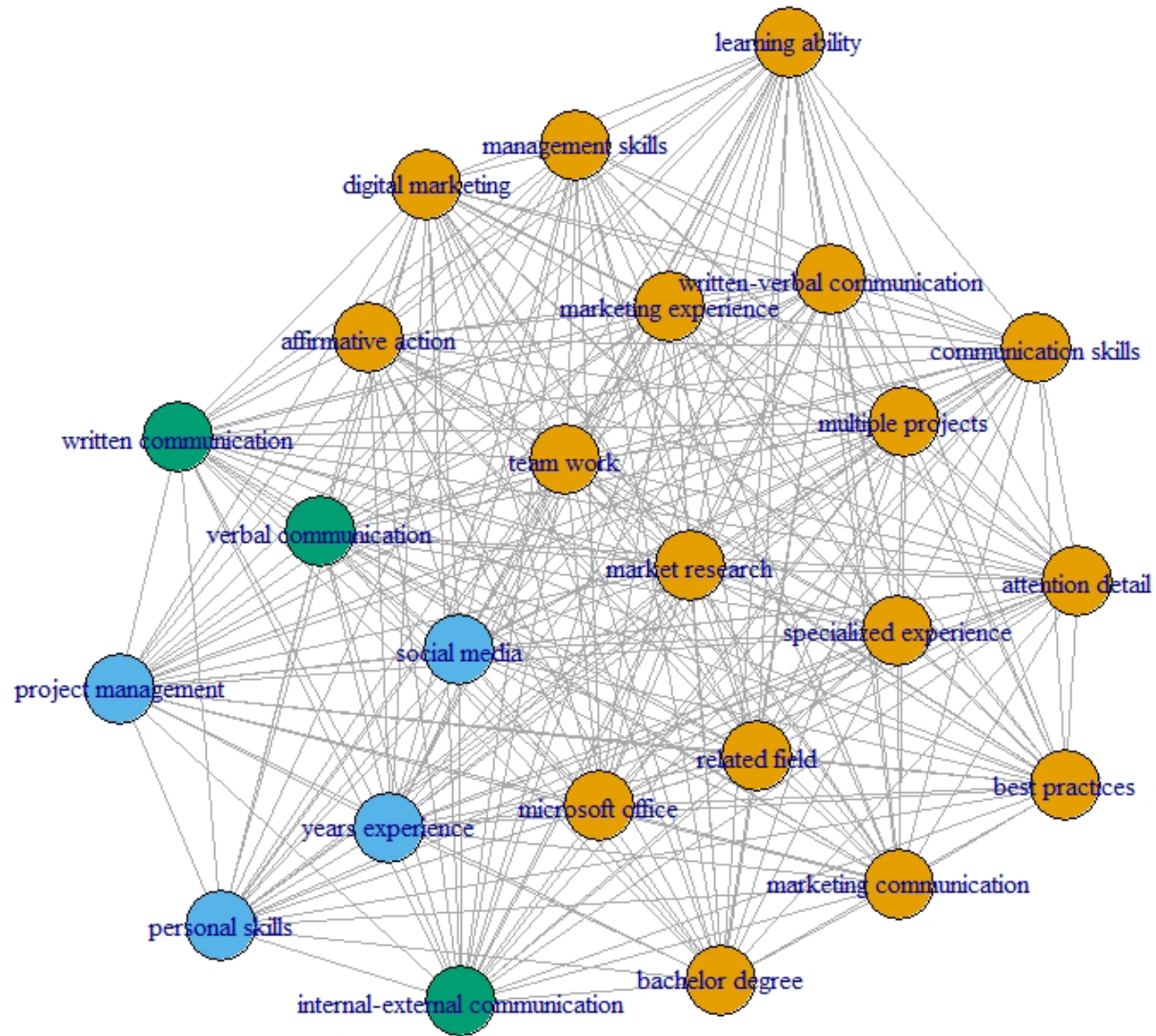
Source: own elaboration.

Fig. 10: Skills network with greedy modularity community detection.



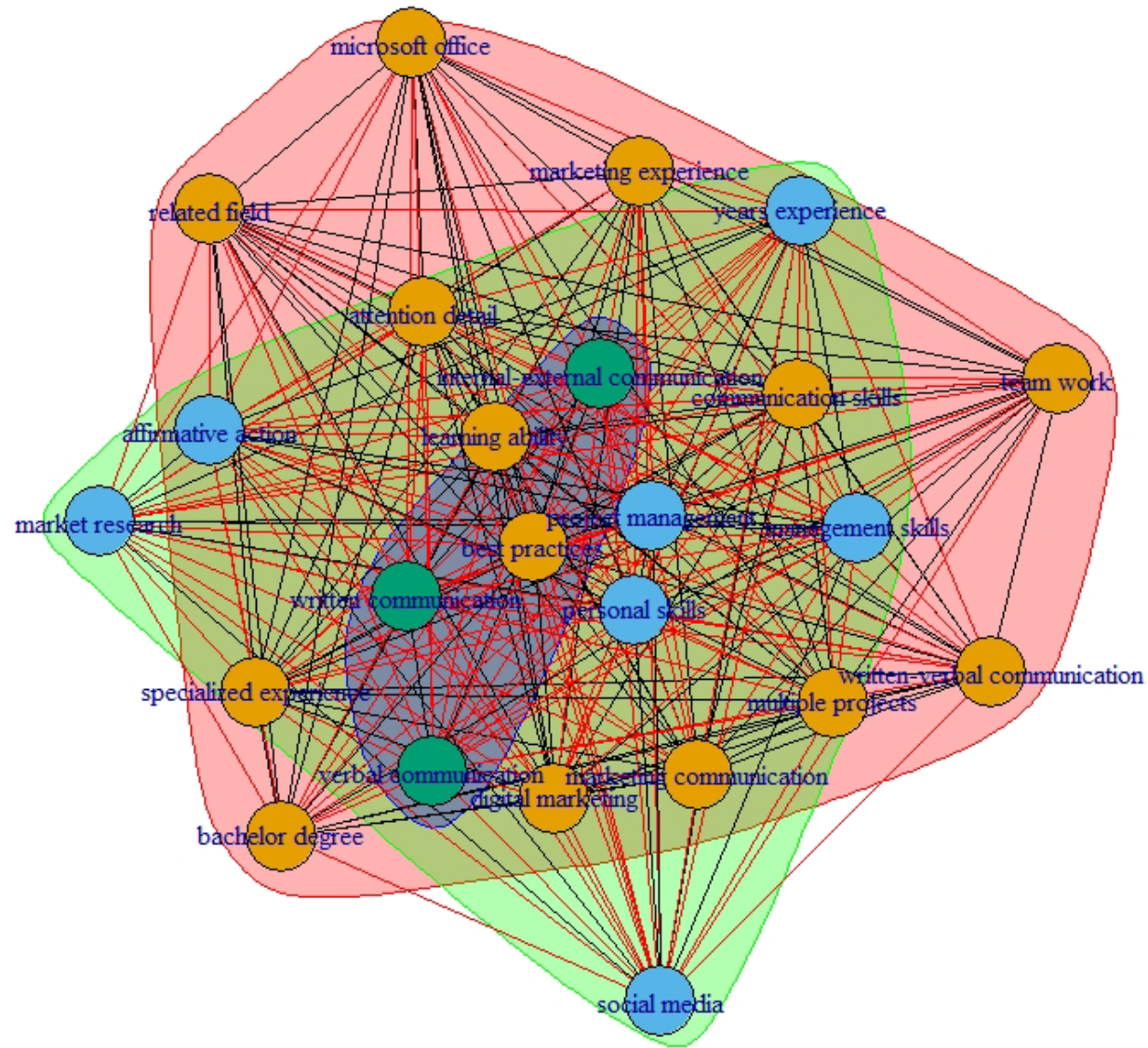
Source: own elaboration.

Fig. 11: Skills network community membership according to greedy modularity.



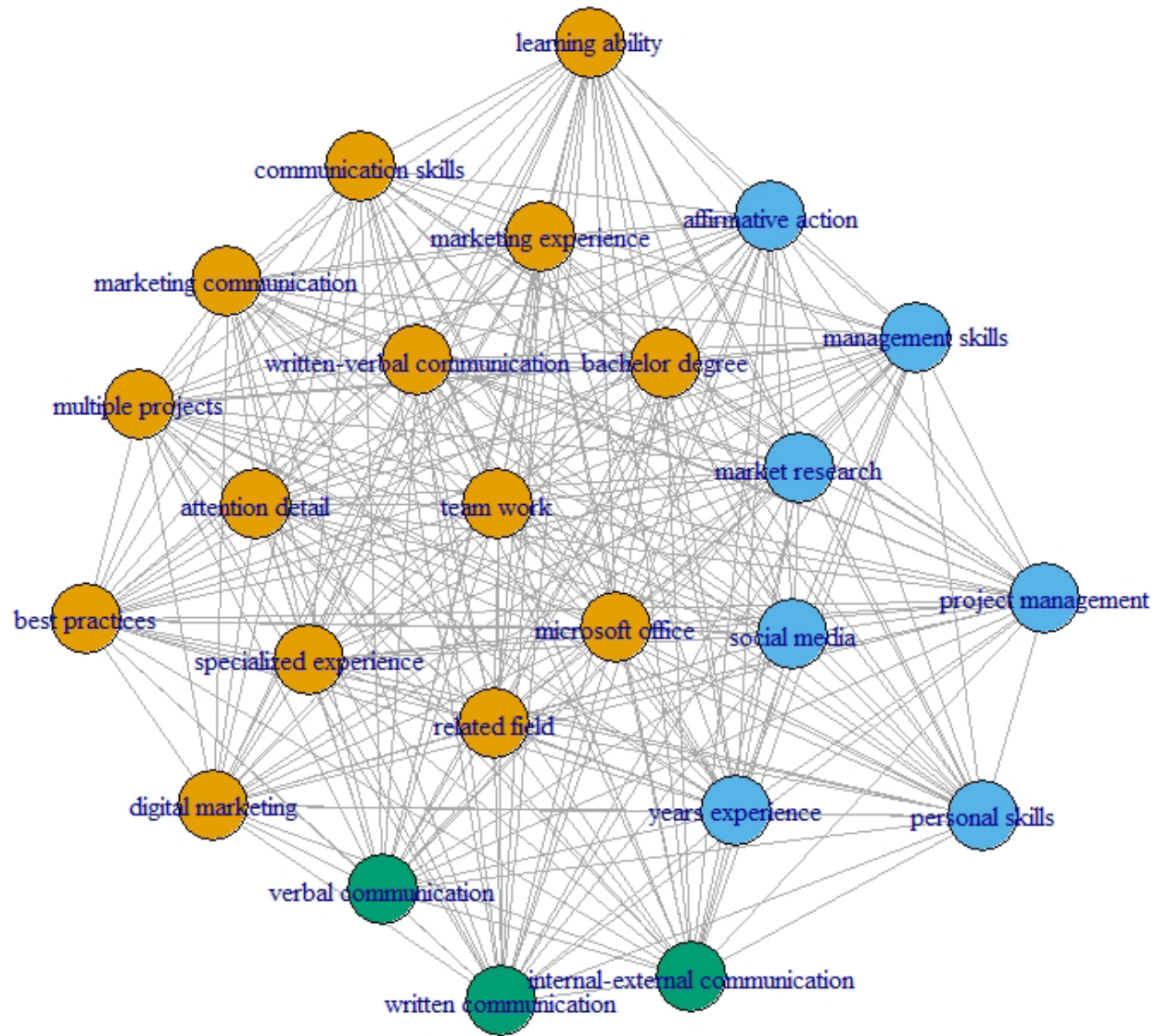
Source: own elaboration.

Fig. 12: Skills network with spectral modularity community detection.



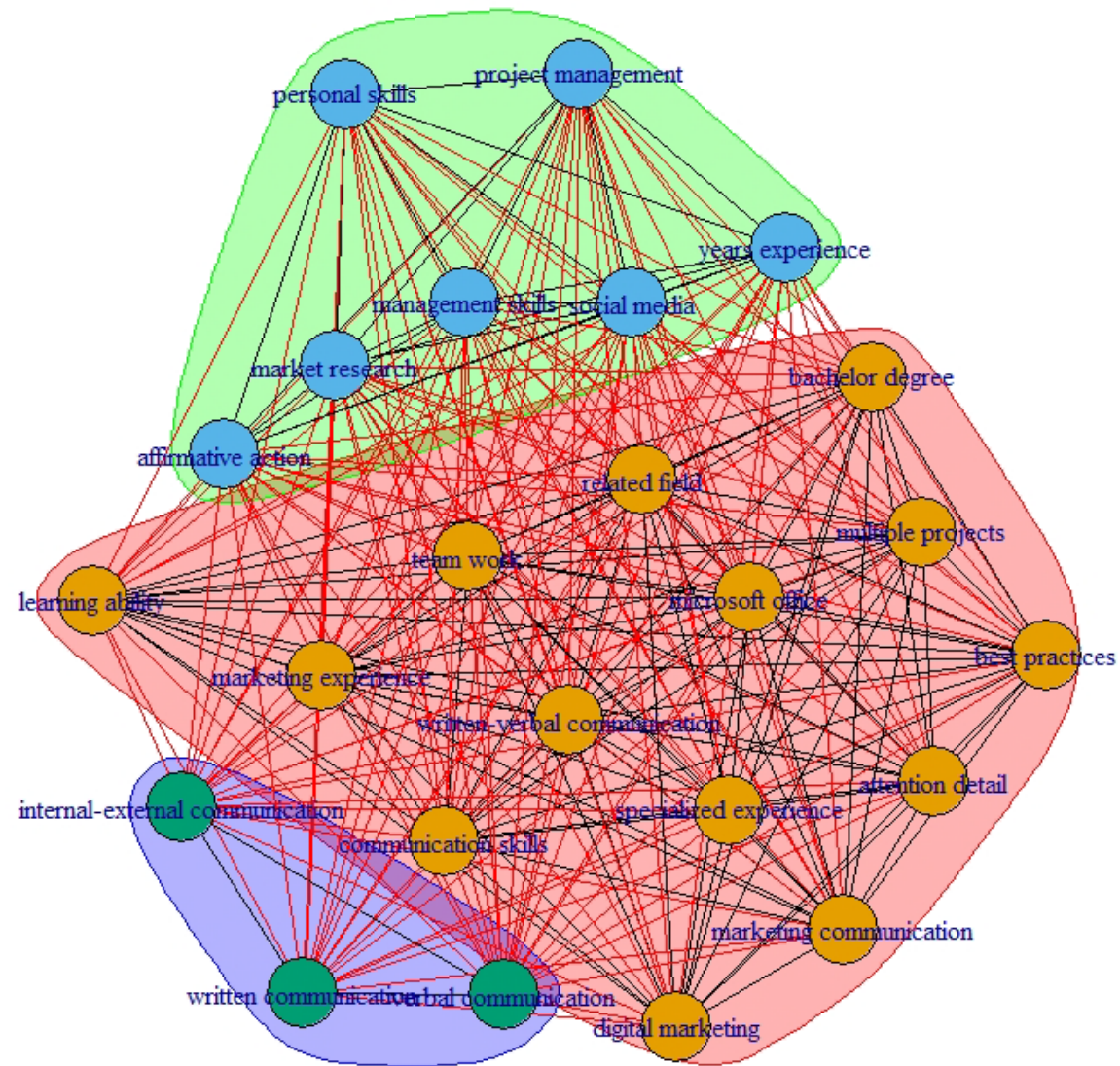
Source: own elaboration.

Fig. 13: Skills network community membership according to spectral modularity.



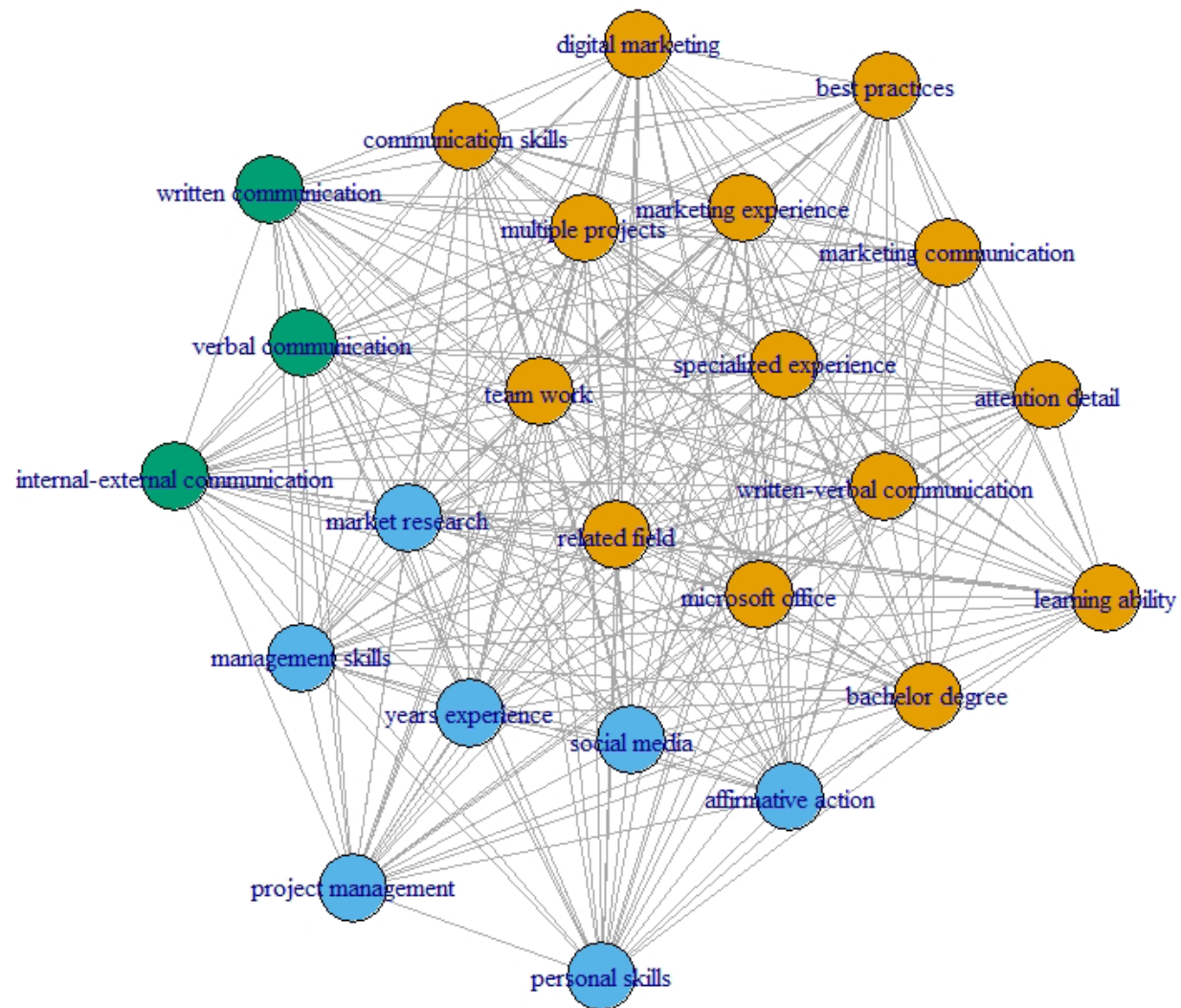
Source: own elaboration.

Fig. 14: Skills network with optimal community detection.



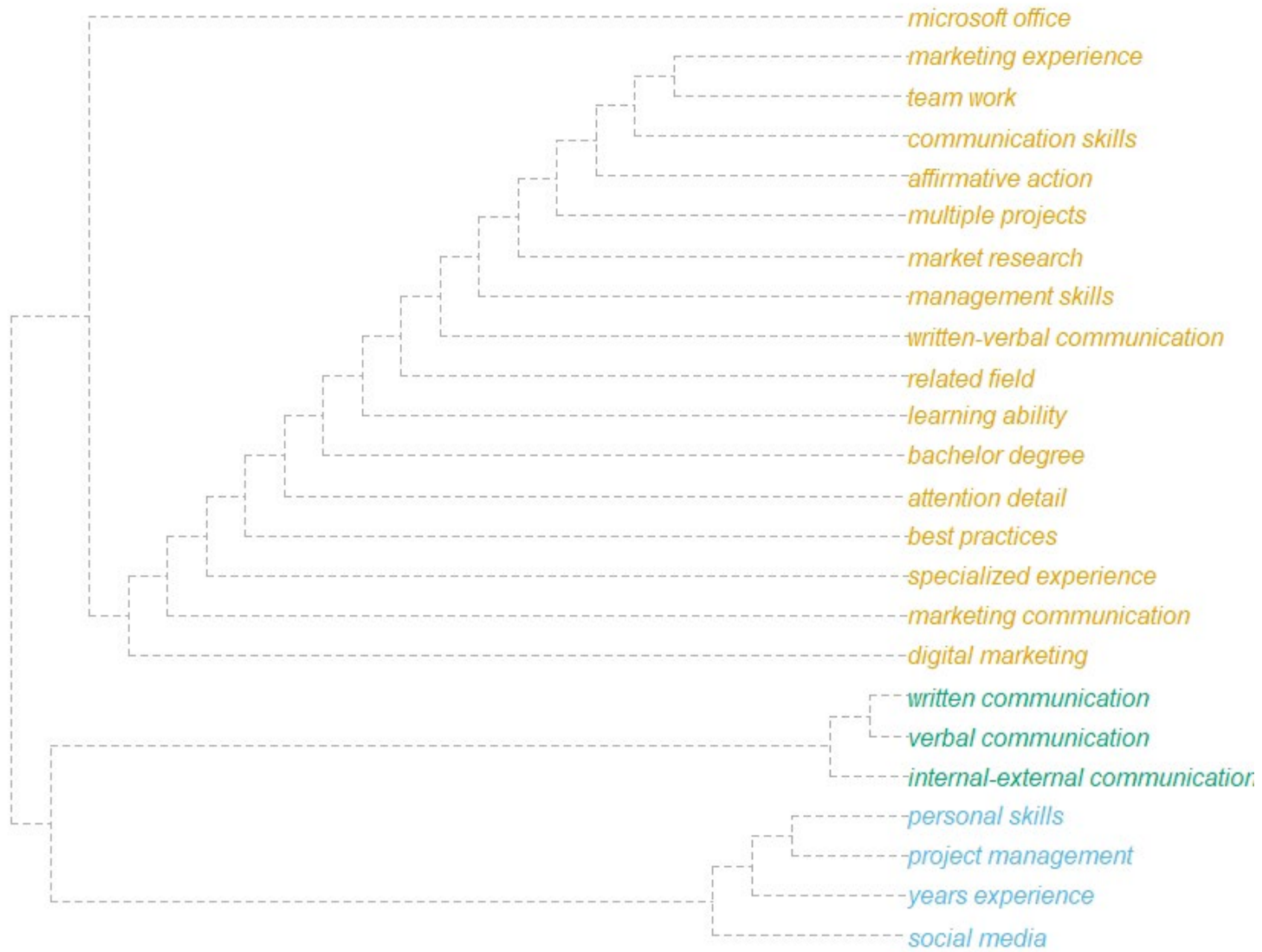
Source: own elaboration.

Fig. 15: Skills network community membership according to optimal modularity.



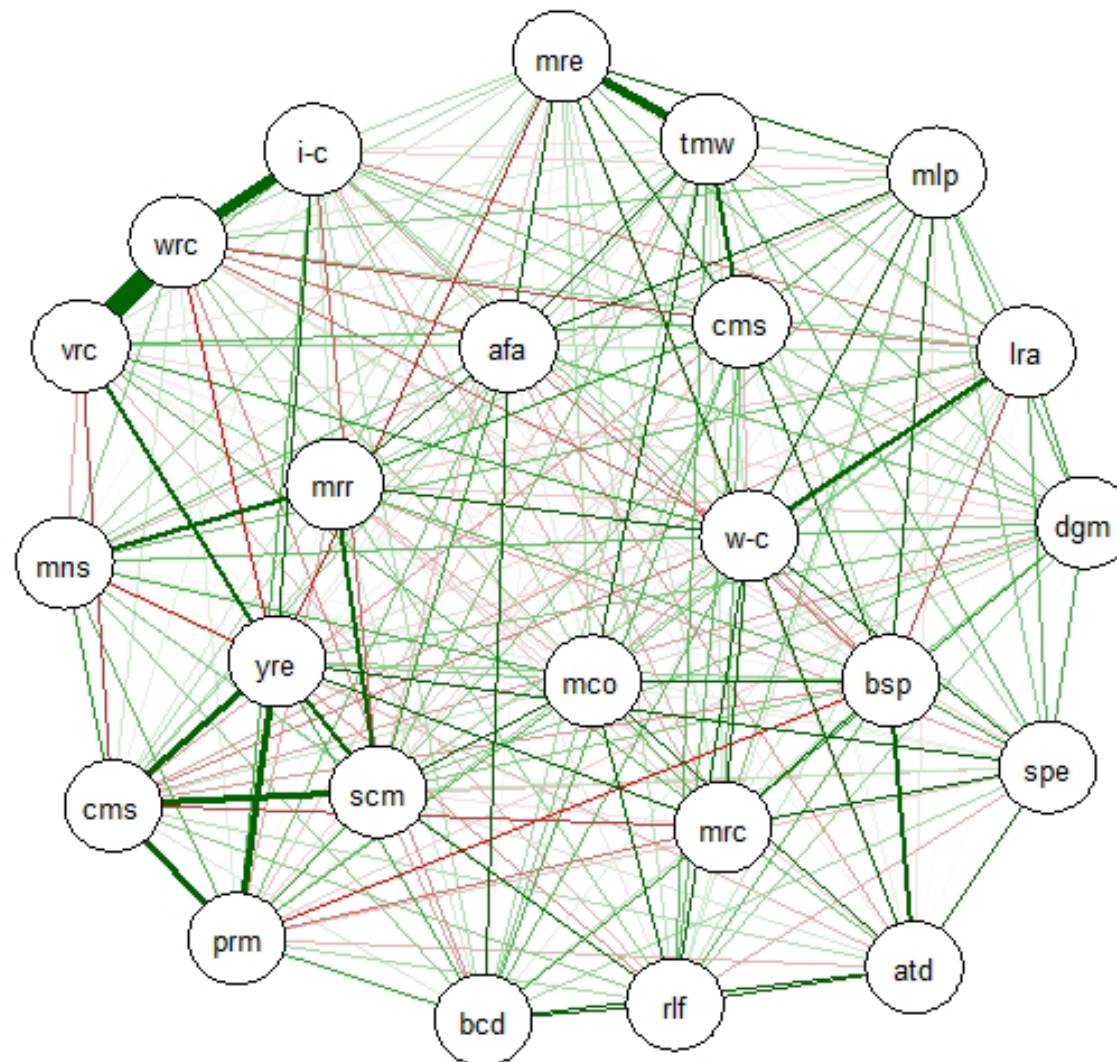
Source: own elaboration.

Fig. 16: Dendrogram with greedy modularity.



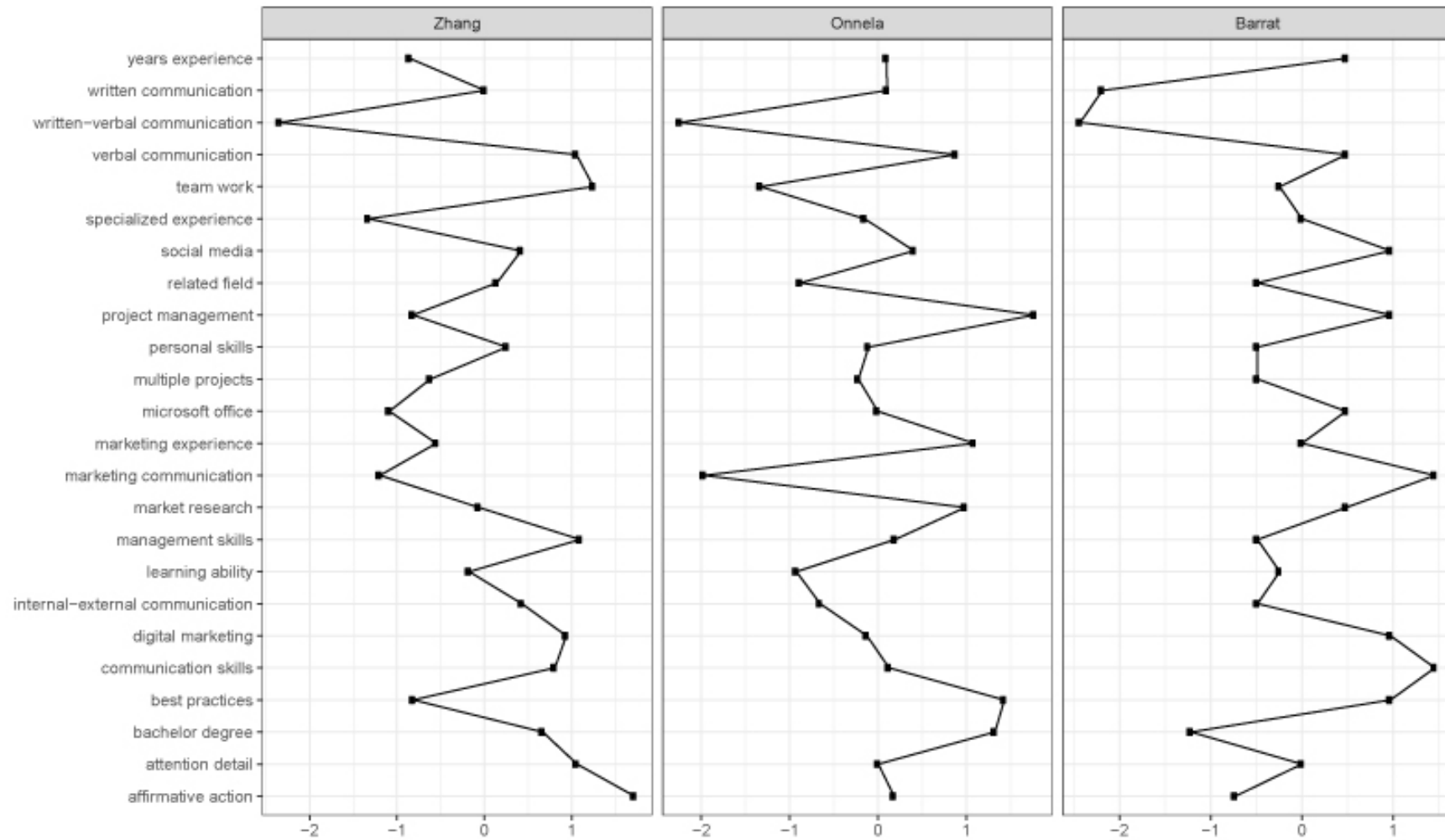
Source: own elaboration.

Fig. 17: Weighted skills network via partial correlations clustering.



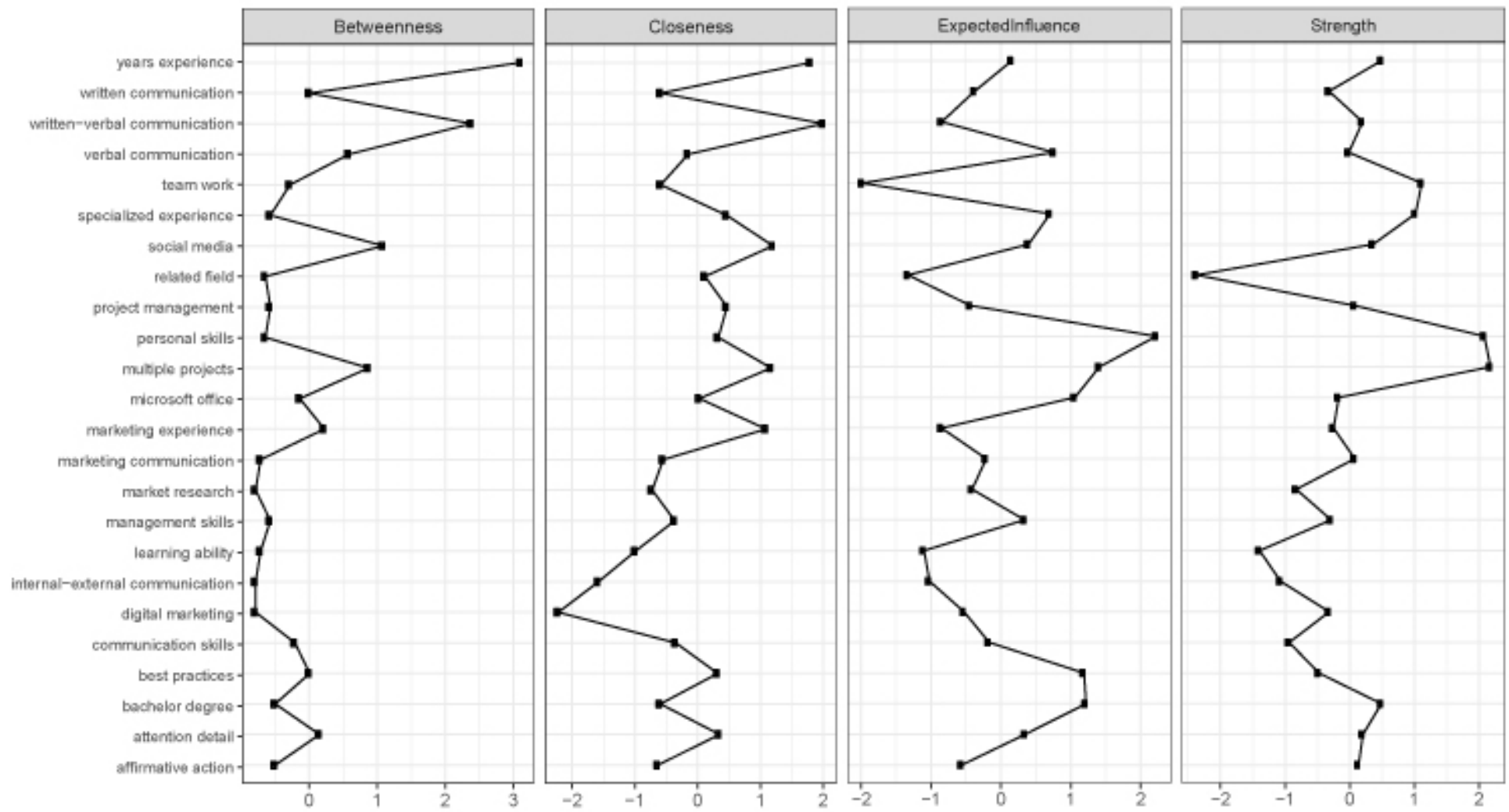
Source: own elaboration.

Fig. 18: Clustering plot with compared methods.



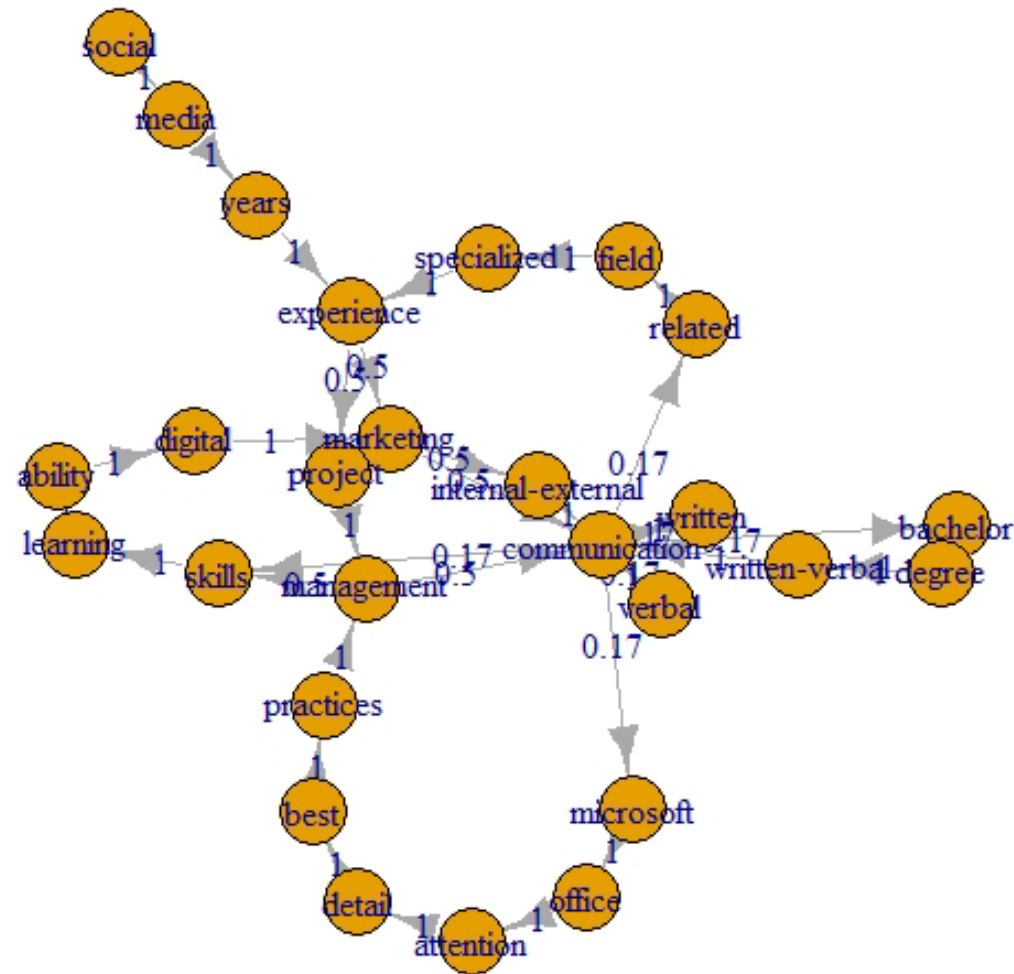
Source: own elaboration.

Fig. 19: Centrality measures plot.



Source: own elaboration.

Fig. 20: Monte Carlo Markov Chain with MAP method.

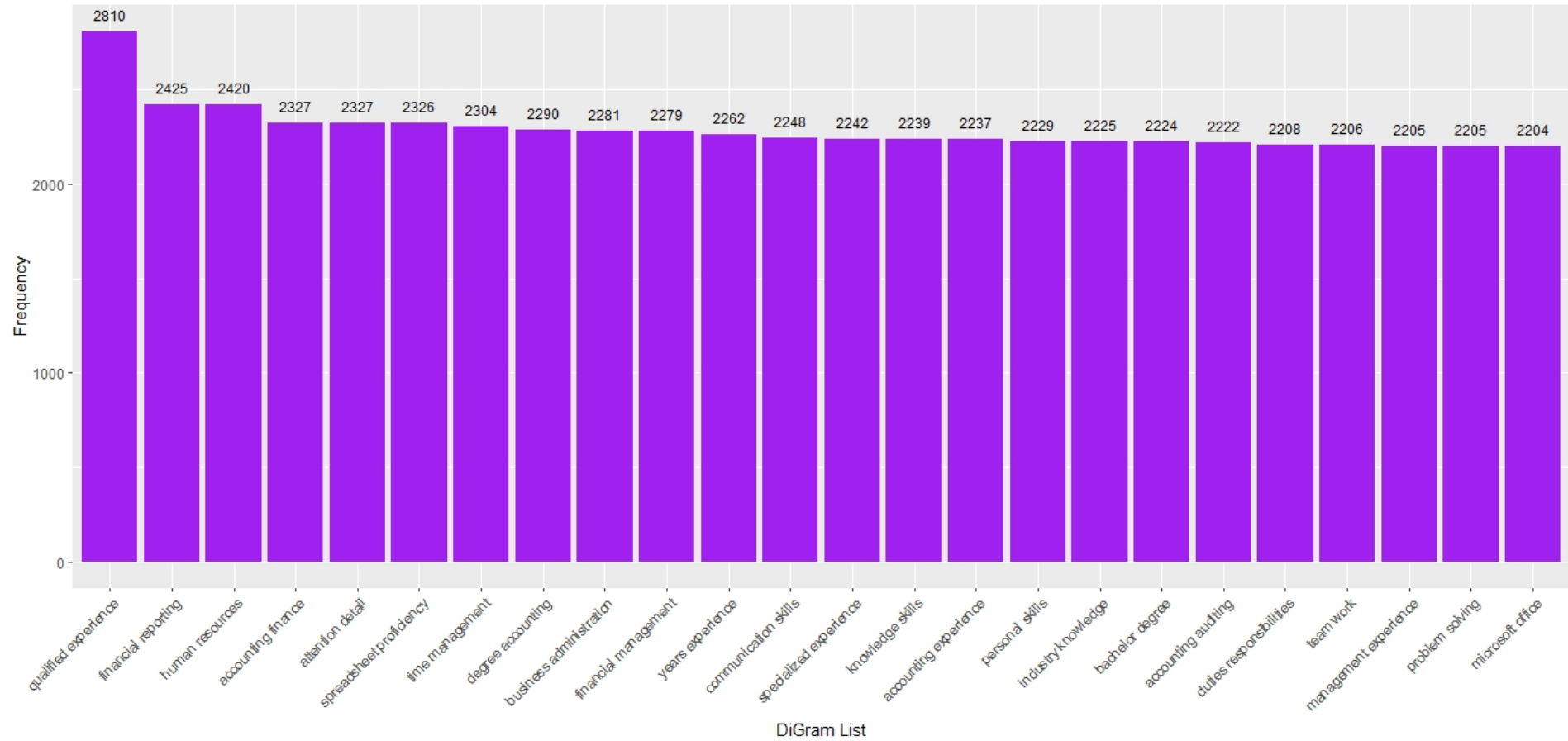


Source: own elaboration.

4.2.2 Accounting & finance

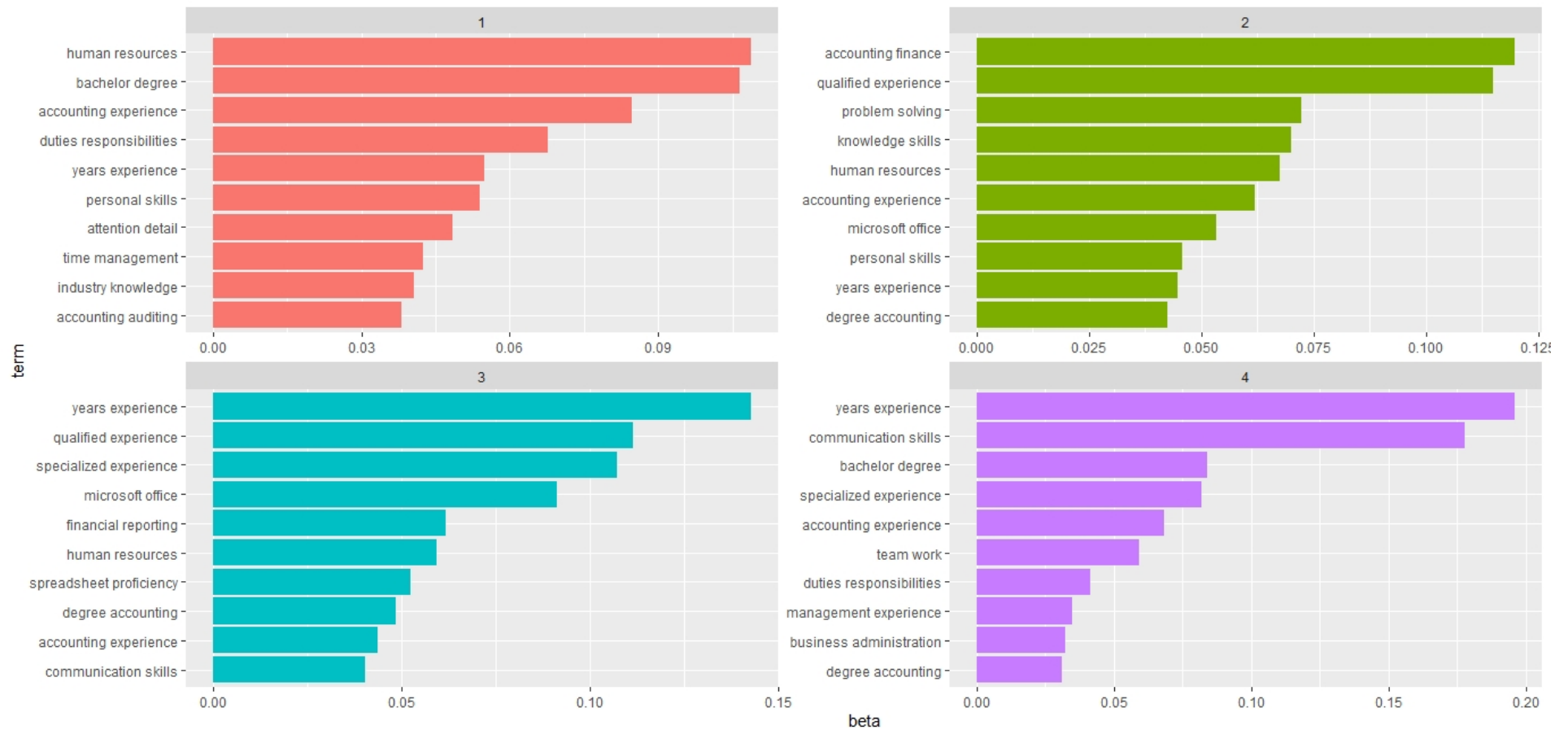
Results from the Accounting industry have been obtained analyzing the subset corpus from the extracted ads regarding the sector. A tokenized Document-Term Matrix (DTM) has been built, and sparsity was removed till 74%. **Fig. 21** shows the bigrams from the corpus. The most frequent terminological combinations were qualified experience (2810), financial reporting (2425), human resources (2420), accounting finance (2327), and attention detail (2327). Topic modeling is presented in **Fig. 22** with four thematic areas. **Fig. 23** highlights the main correlations through the skills set. **Fig. 24** detects greedy modularity in the skillset, dividing it in three groups, and the relative memberships are shown in **Fig. 25**. Application of spectral modularity is presented in **Fig. 26**, and the relative memberships are reported in **Fig. 27**. The employment of optimal modularity detection is shown in **Fig. 28** and their memberships highlighted in **Fig. 29**. At this point, modularity indicators were compared to define the most proper method to give sense to the analysis. Having $\xi_G > \xi_O > \xi_S$, the dendrogram in **Fig. 30** was built with greedy modularity, and partial correlations will be used for the weighted network in **Fig. 31**. Thus, a clustering plot with Zhang, Onnela, and Barrat methods is reported in **Fig. 32**. Centrality measures are exposed in **Fig. 33**. The most between skills in the set were accounting finance (8.9%), financial management (7.1%), financial reporting (5.3%), duties responsibilities (3.1%), and teamwork (1.1%). The closest skills were accounting finance (22%), teamwork (20%), financial management (16.2%), time management (18.5%), and duties responsibilities (17.5%). MCMC with MAP method is shown in **Fig. 34** to forecast and simulate a possible job interview for the Accounting industry.

Fig 21: Bigrams of the Accounting skillset.



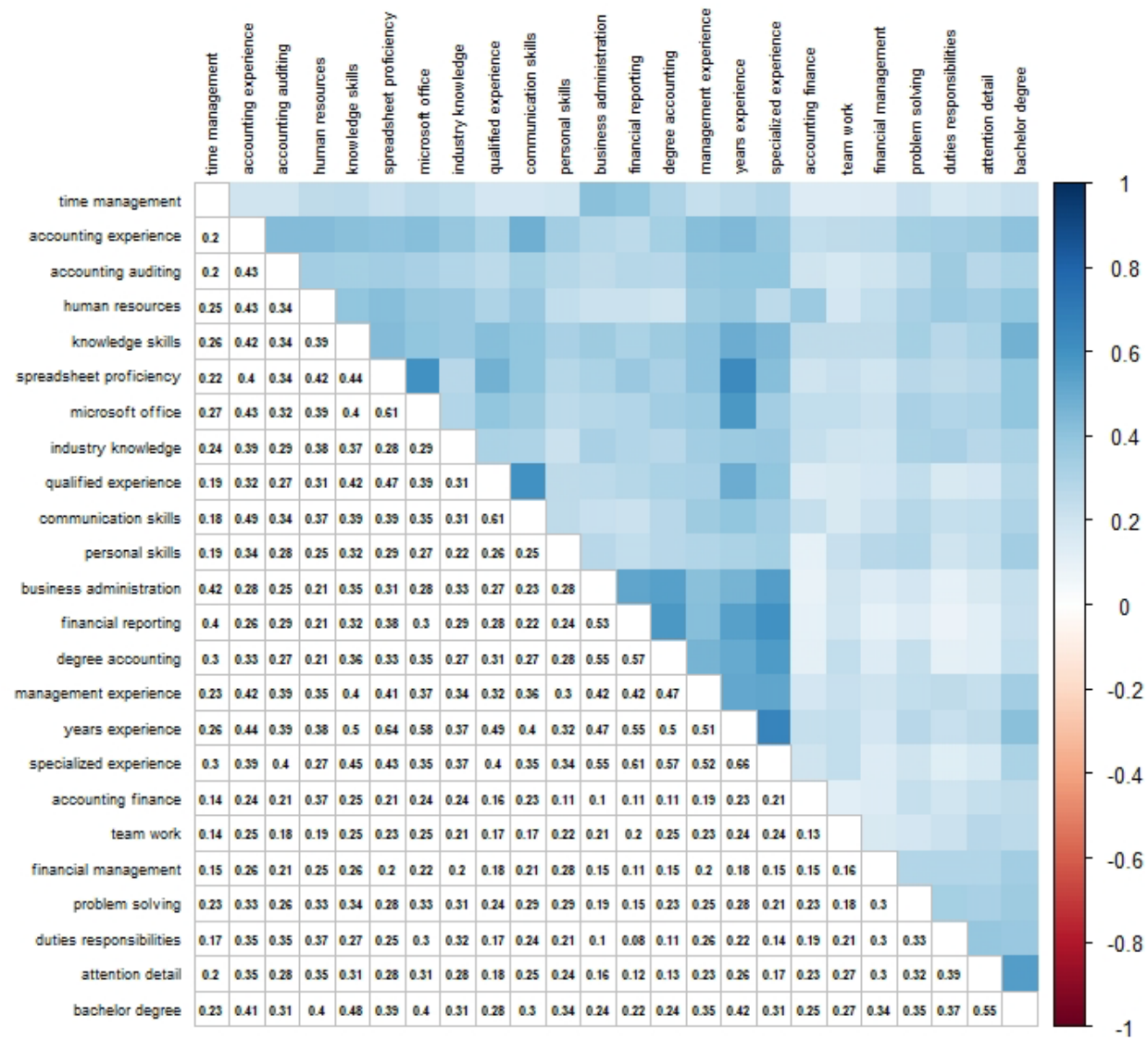
Source: own elaboration.

Fig. 22: Topic modeling of the Accounting skillset.



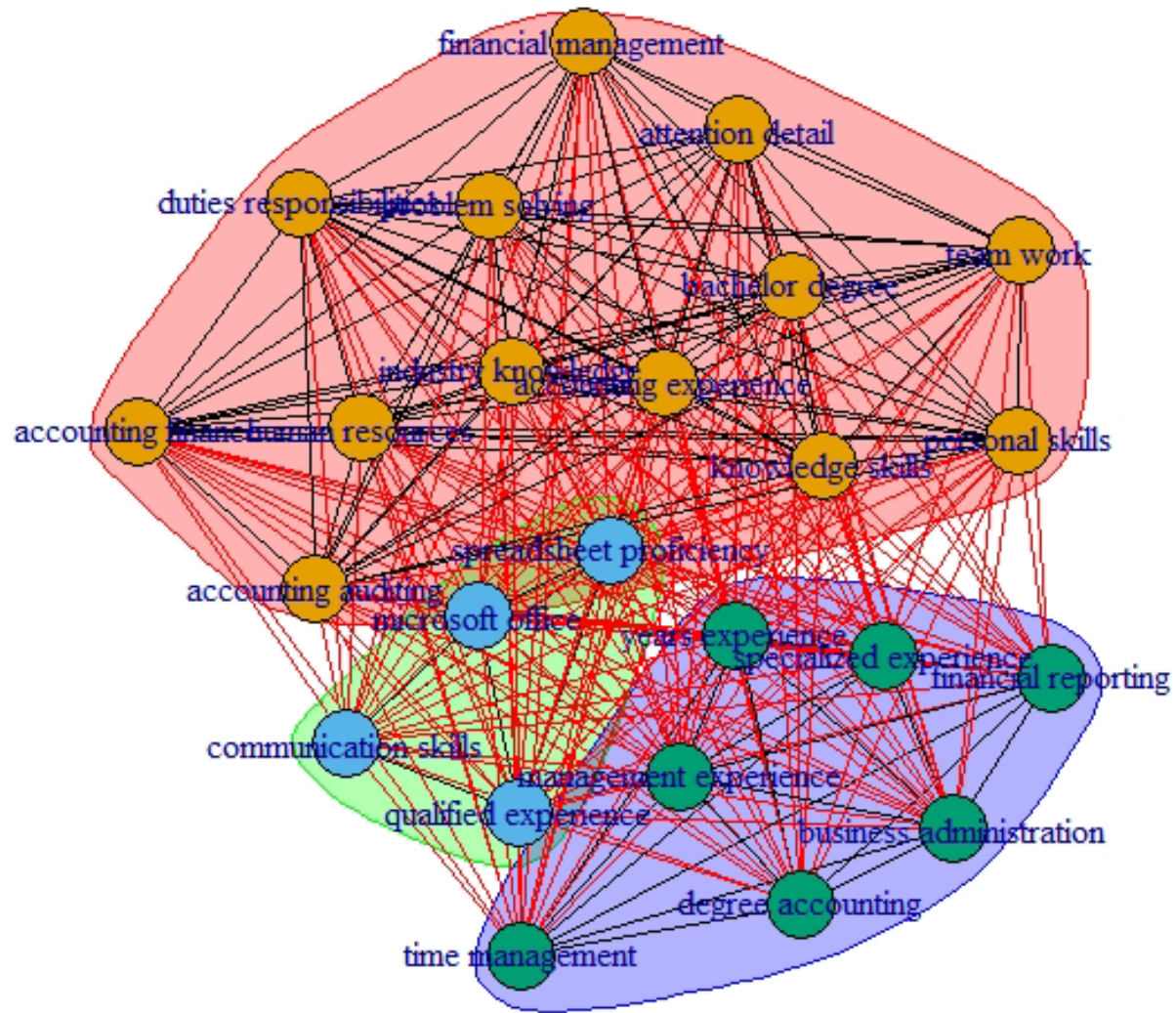
Source: own elaboration.

Fig. 23: Corrplot of the Accounting skillset.



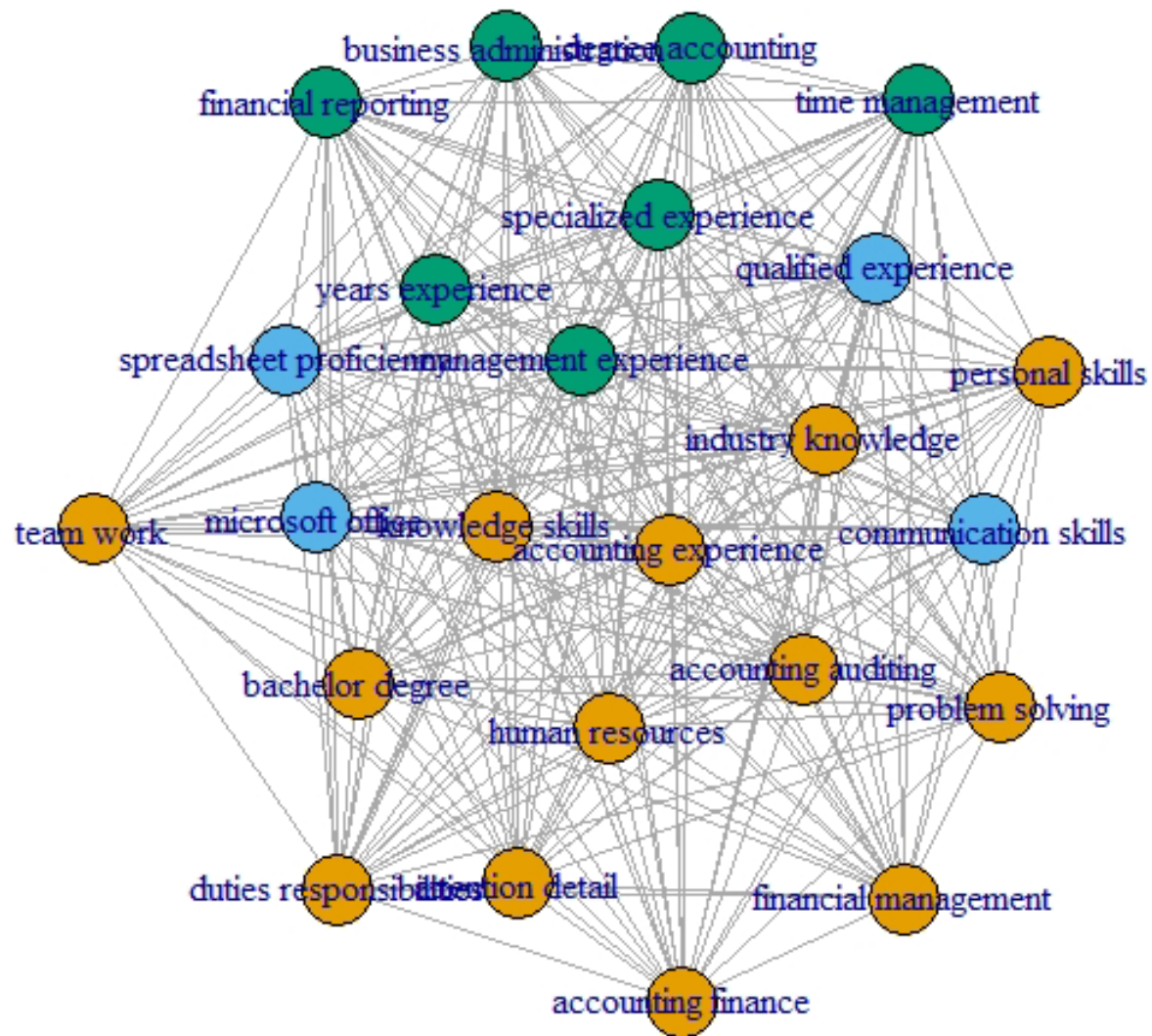
Source: own elaboration.

Fig. 24: Skills network with greedy modularity community detection.



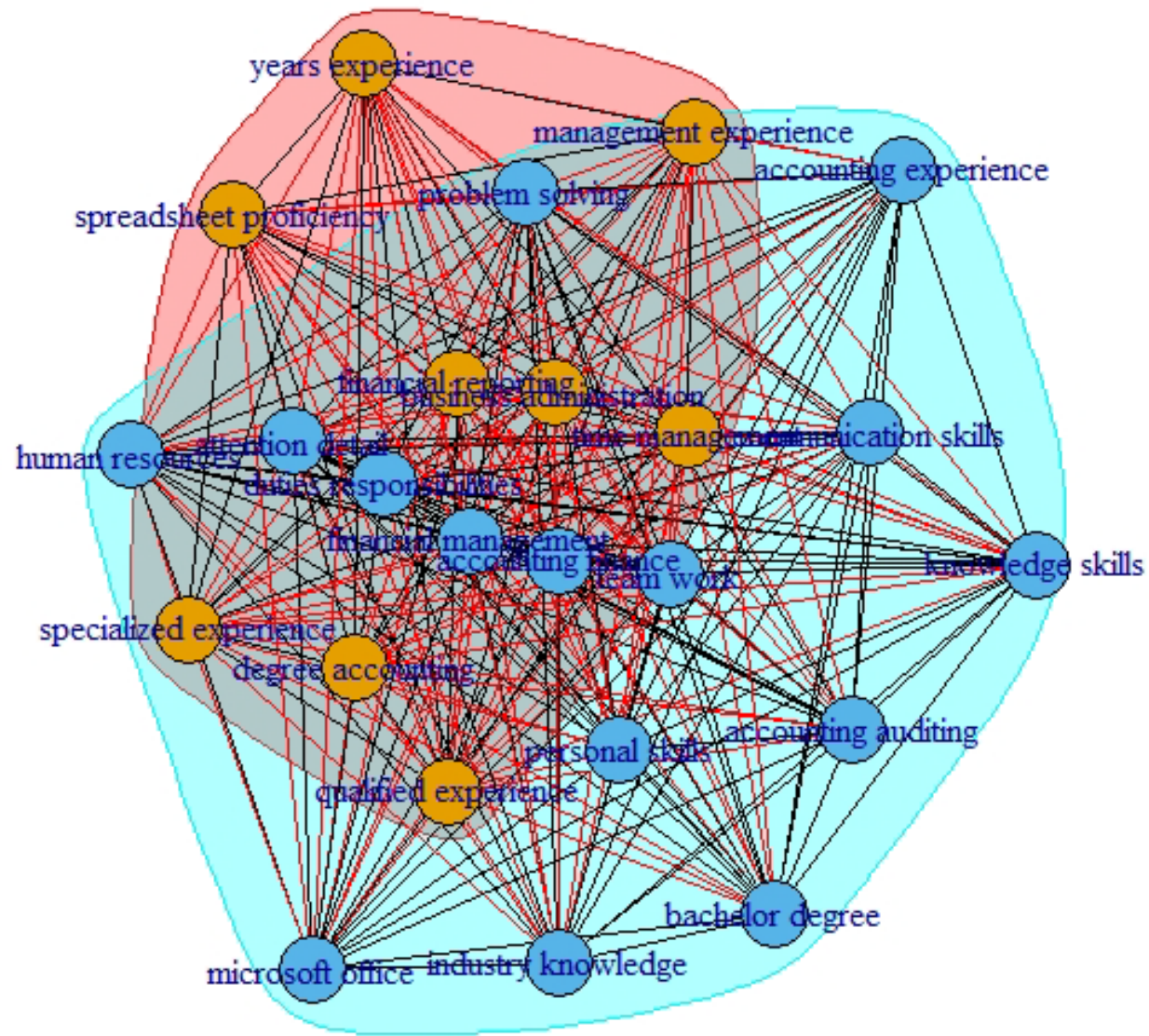
Source: own elaboration.

Fig 25: Skills network community membership according to greedy modularity.



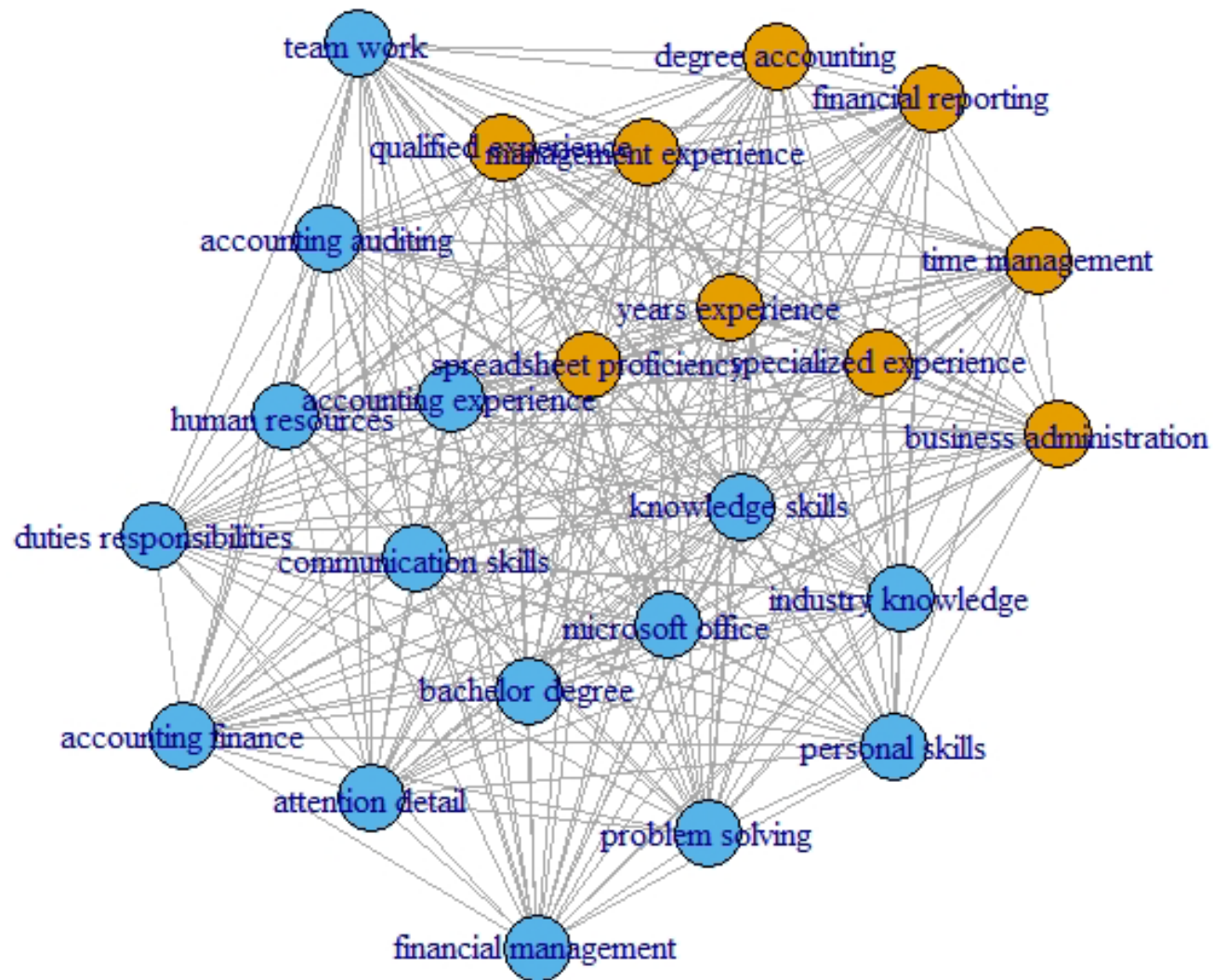
Source: own elaboration.

Fig. 26: Skills network with spectral modularity community detection.



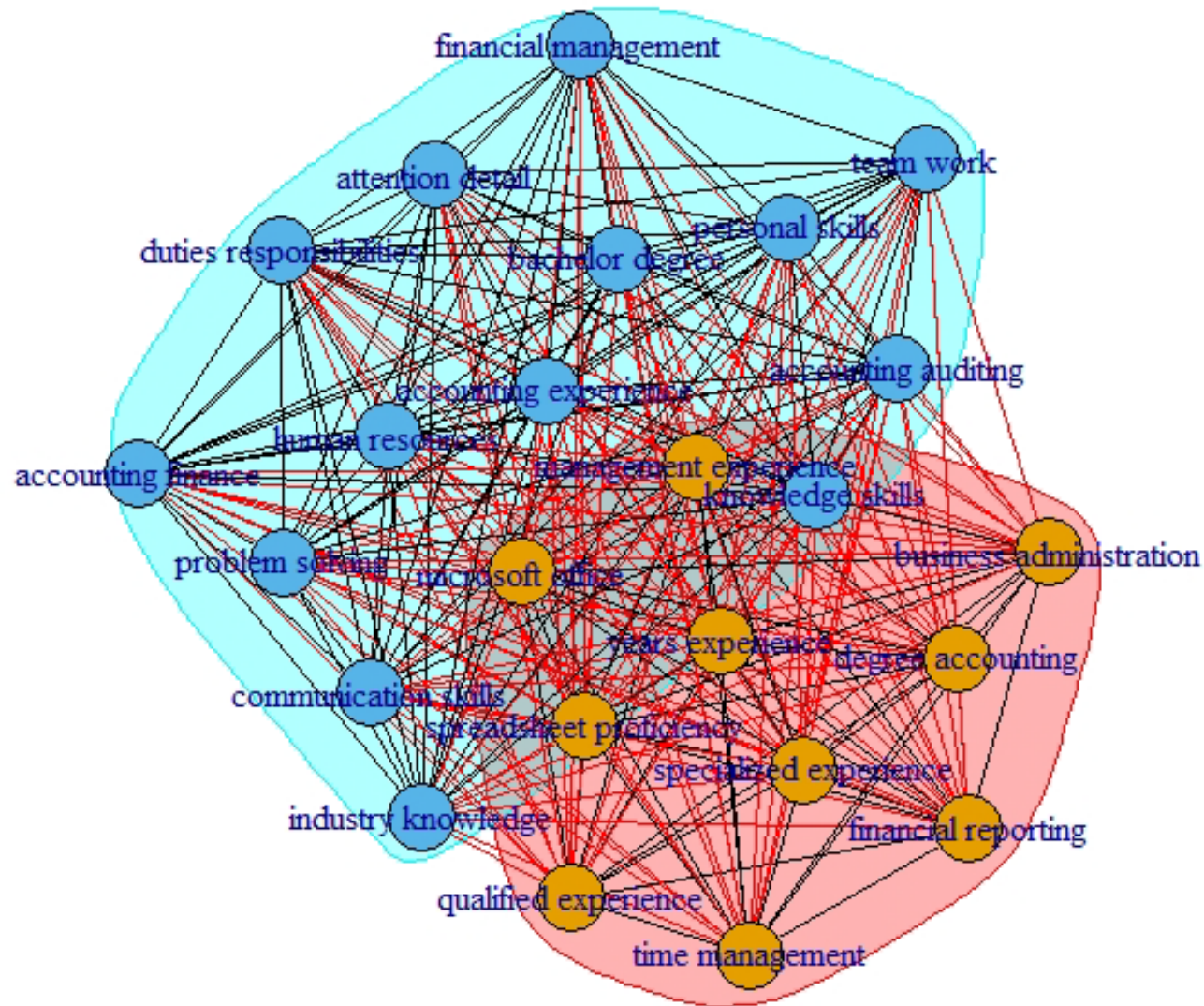
Source: own elaboration.

Fig. 27: Skills network community membership according to spectral modularity.



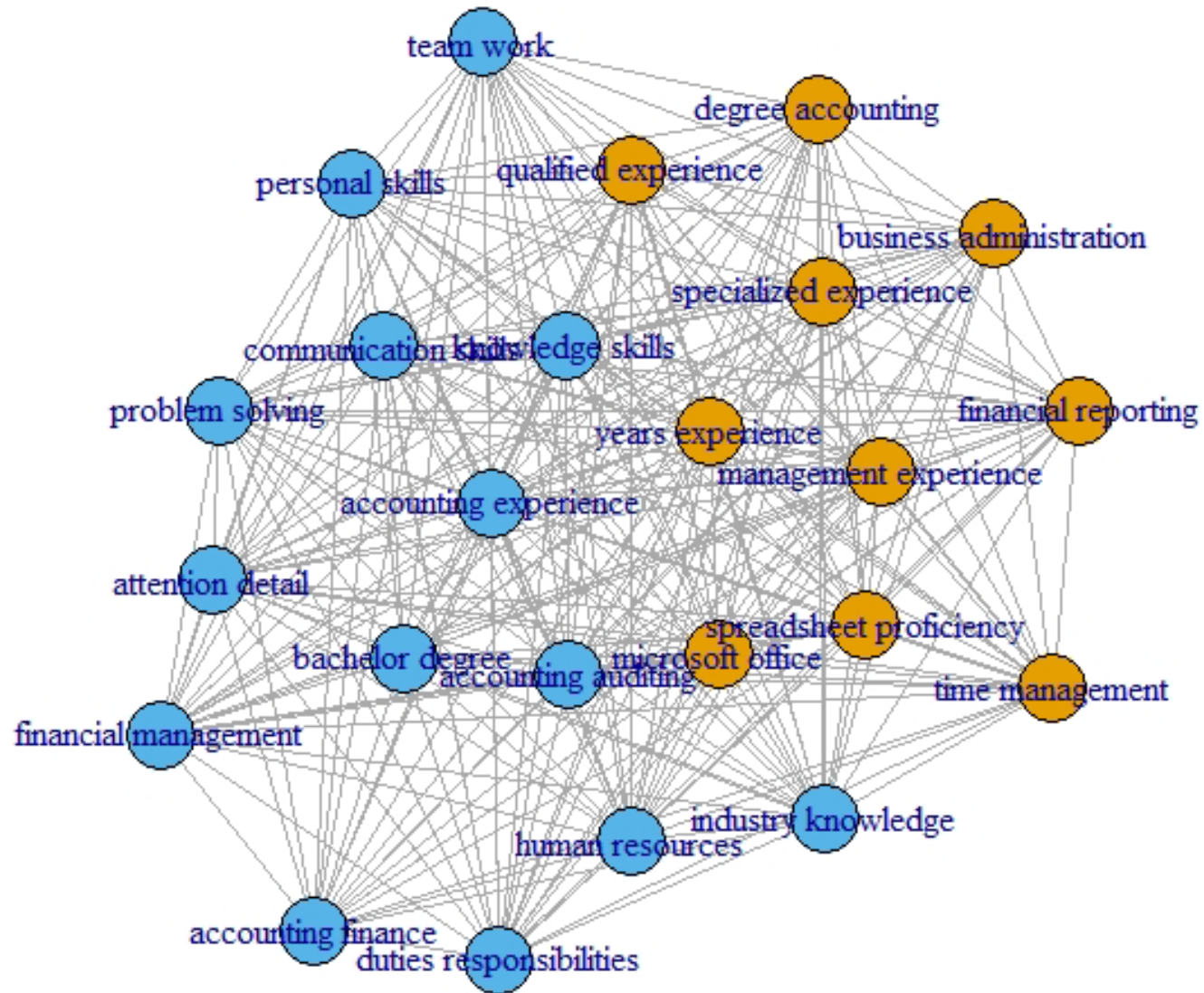
Source: own elaboration.

Fig. 28: Skills network with optimal community detection.



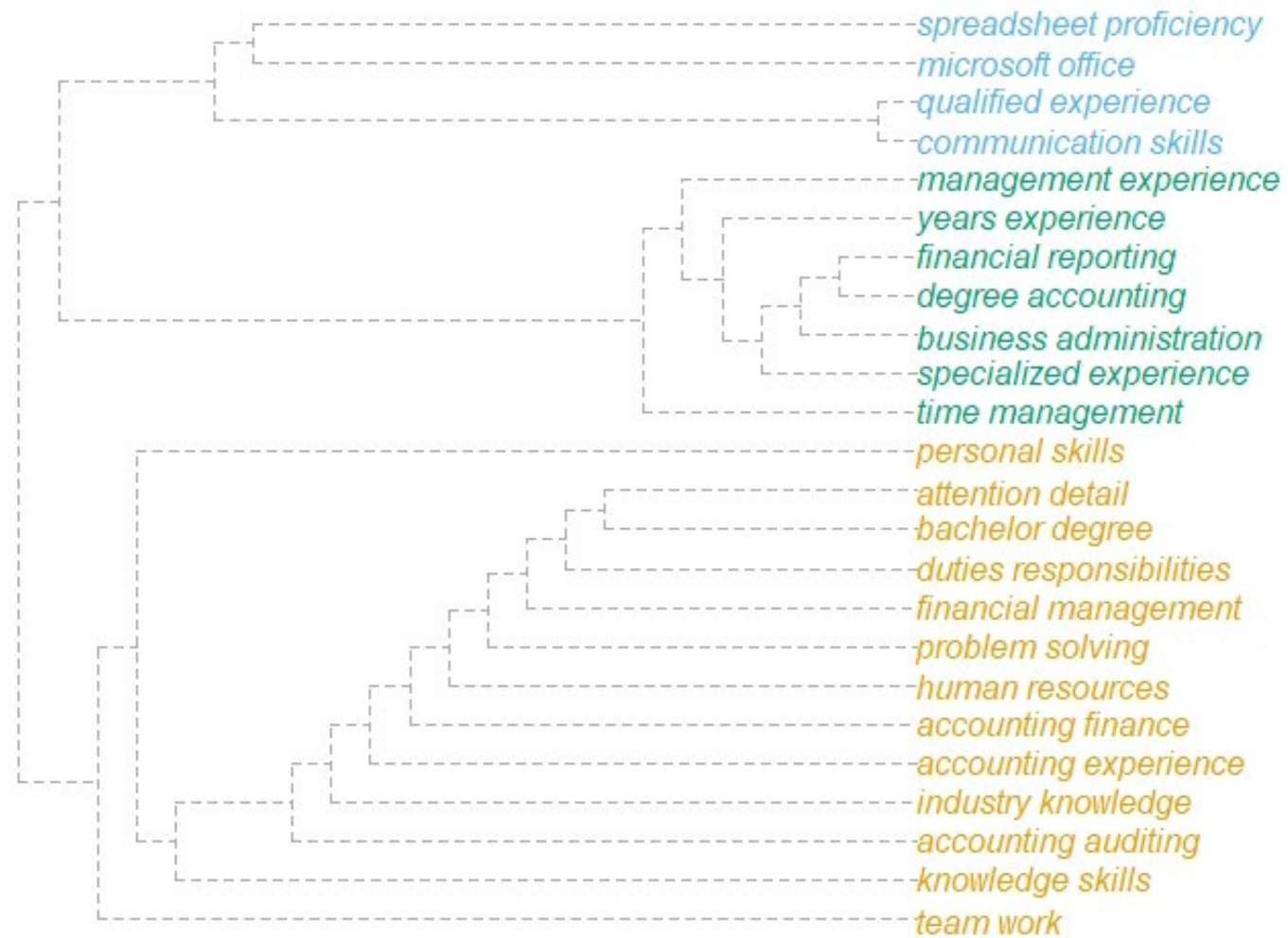
Source: own elaboration.

Fig. 29: Skills network community membership according to optimal modularity.



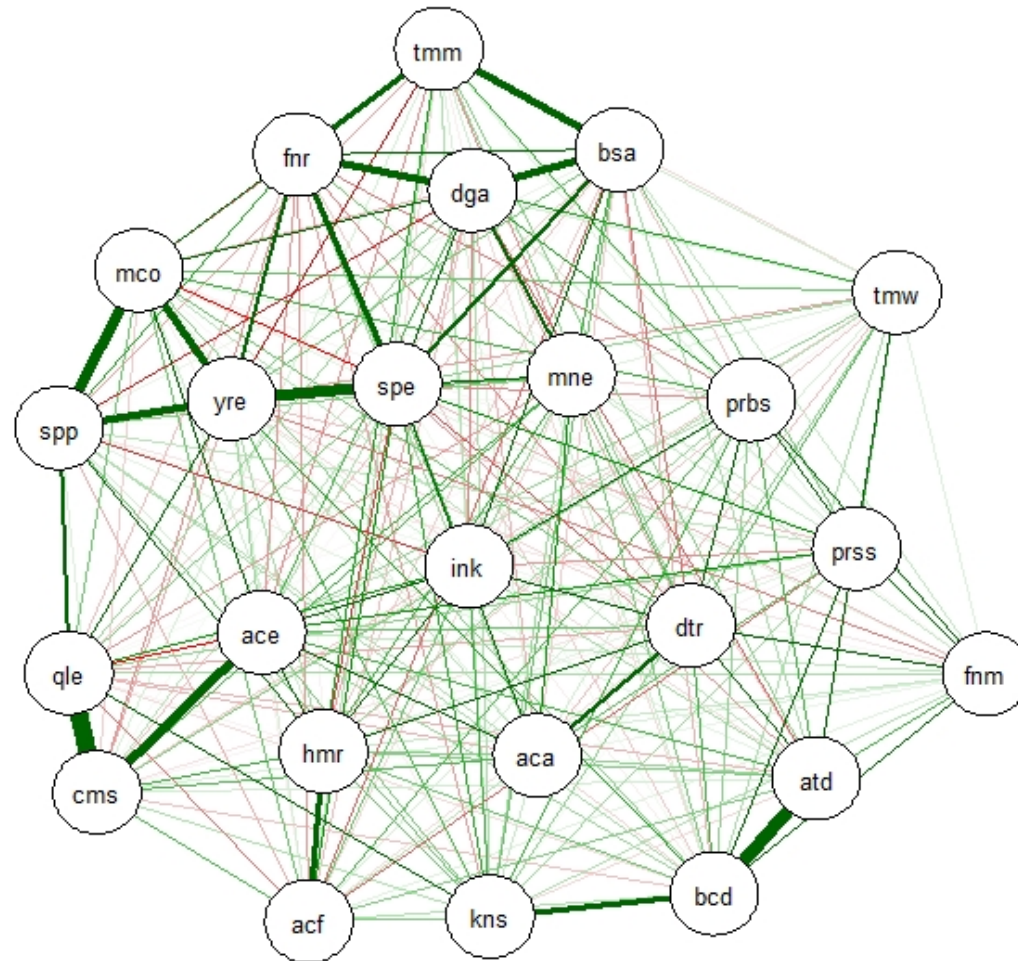
Source: own elaboration.

Fig. 30: Dendrogram with greedy modularity.



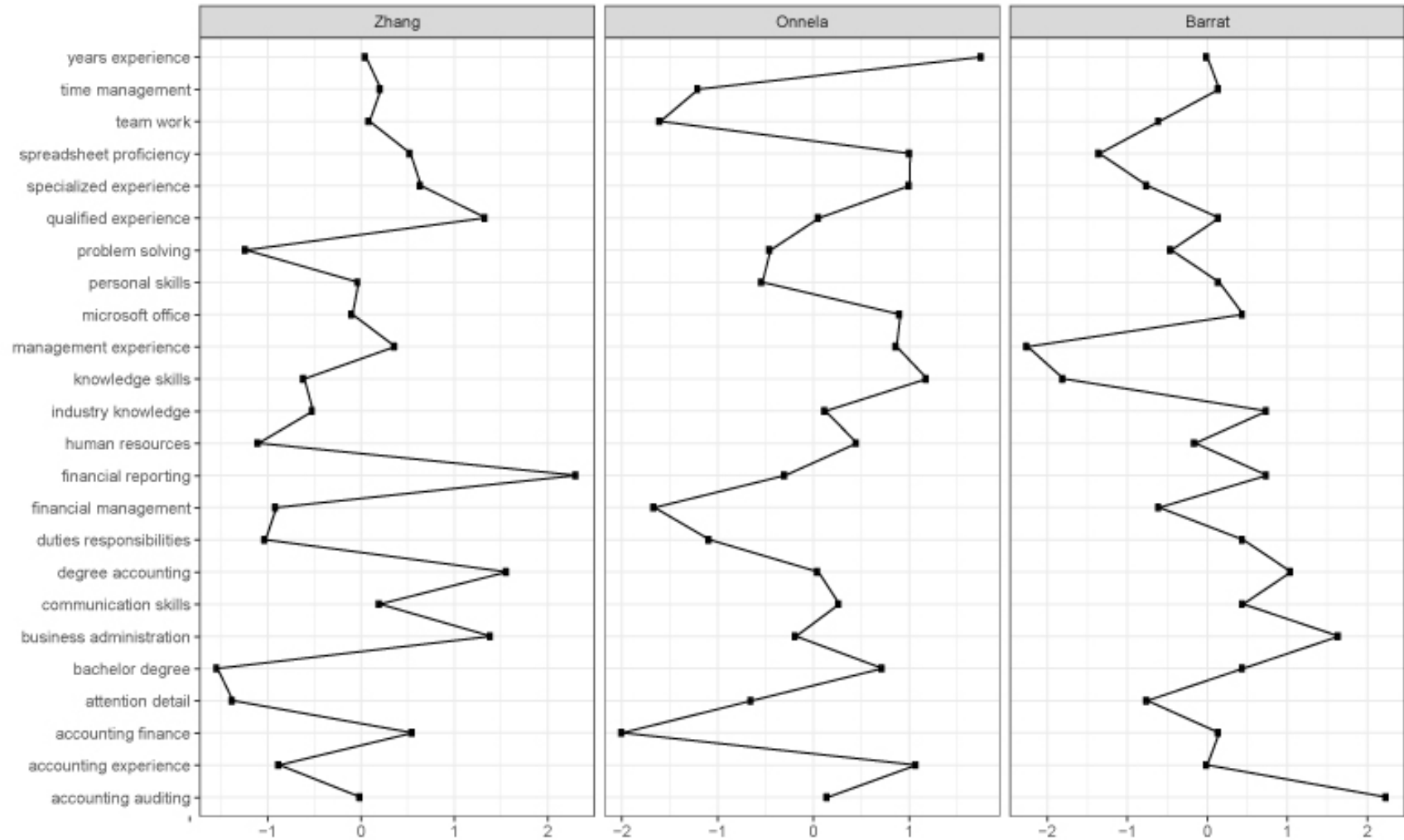
Source: own elaboration.

Fig. 31: Weighted skills network via partial correlations clustering.



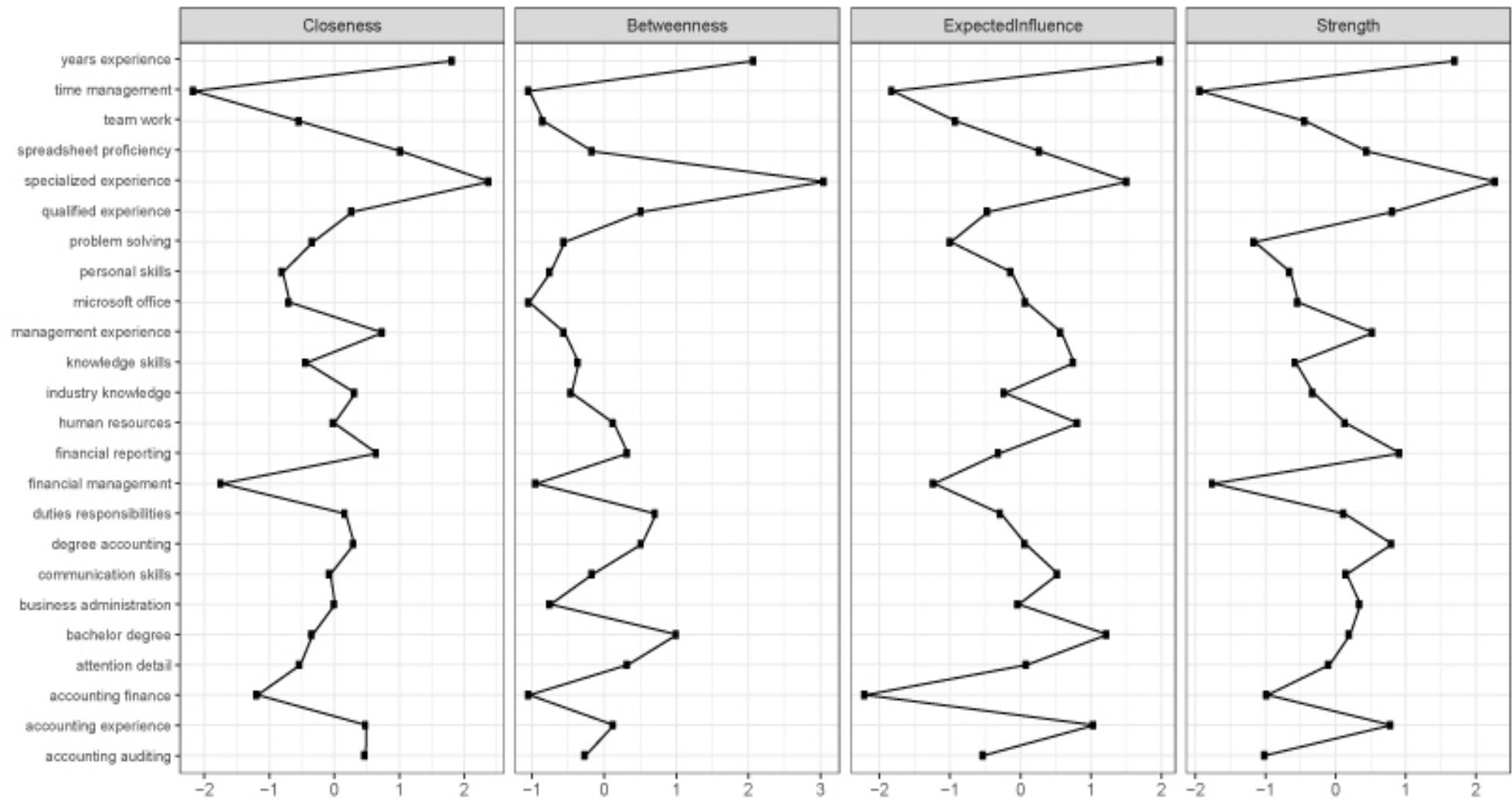
Source: own elaboration.

Fig. 32: Clustering plot with compared methods.



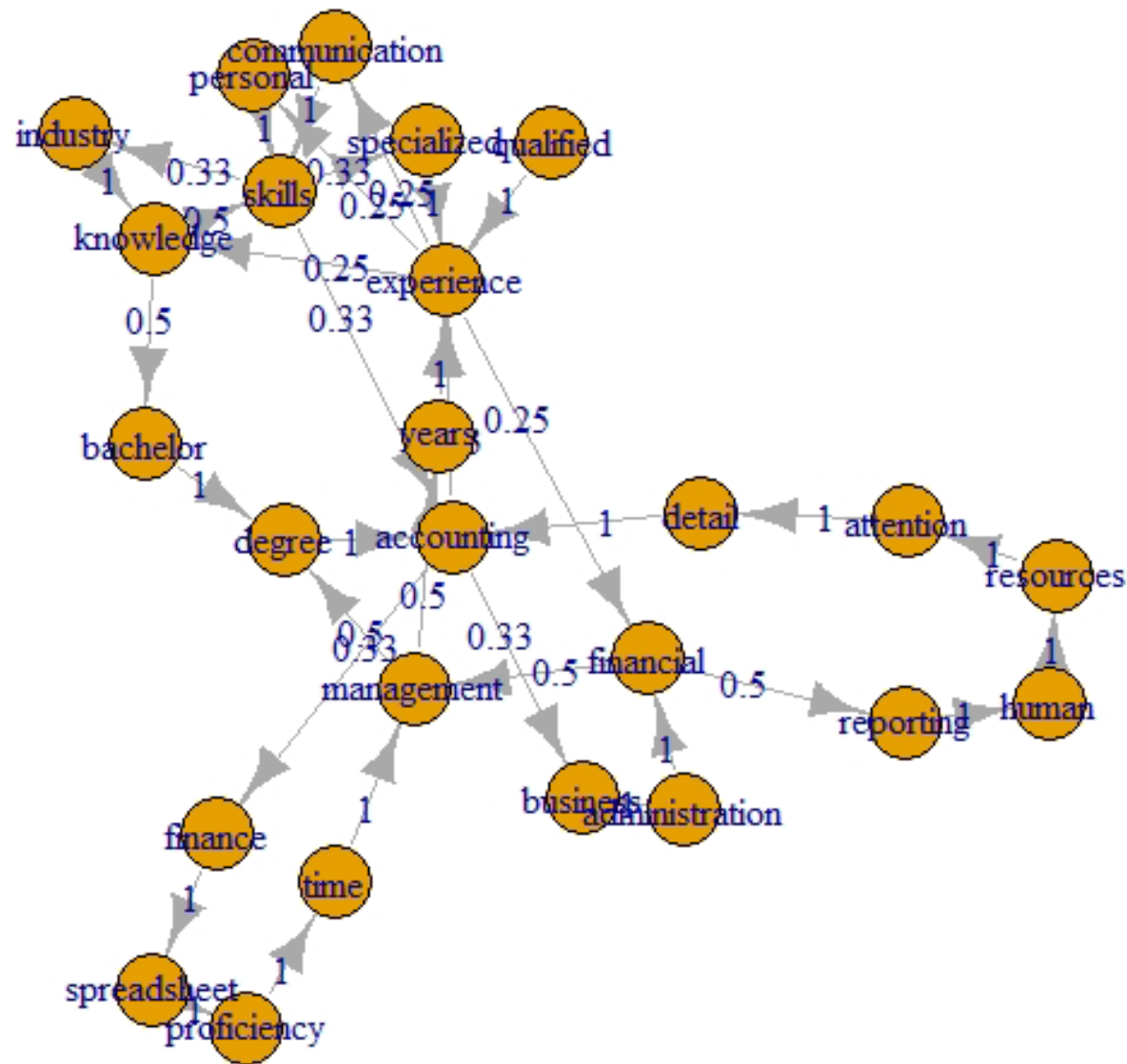
Source: own elaboration.

Fig. 33: Centrality measures plot.



Source: own elaboration.

Fig. 34: Monte Carlo Markov Chain with MAP methods.

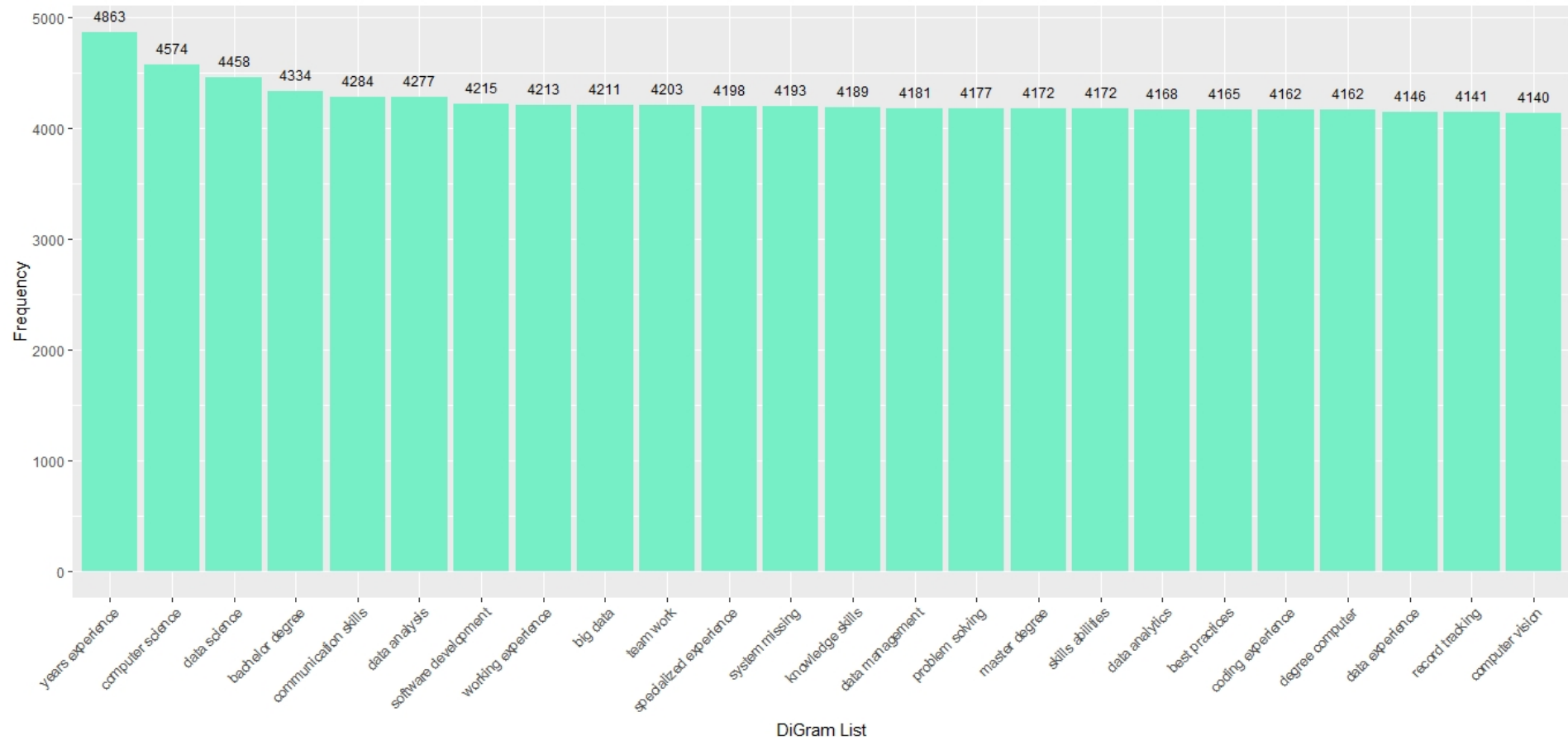


Source: own elaboration.

4.2.3 Data Science

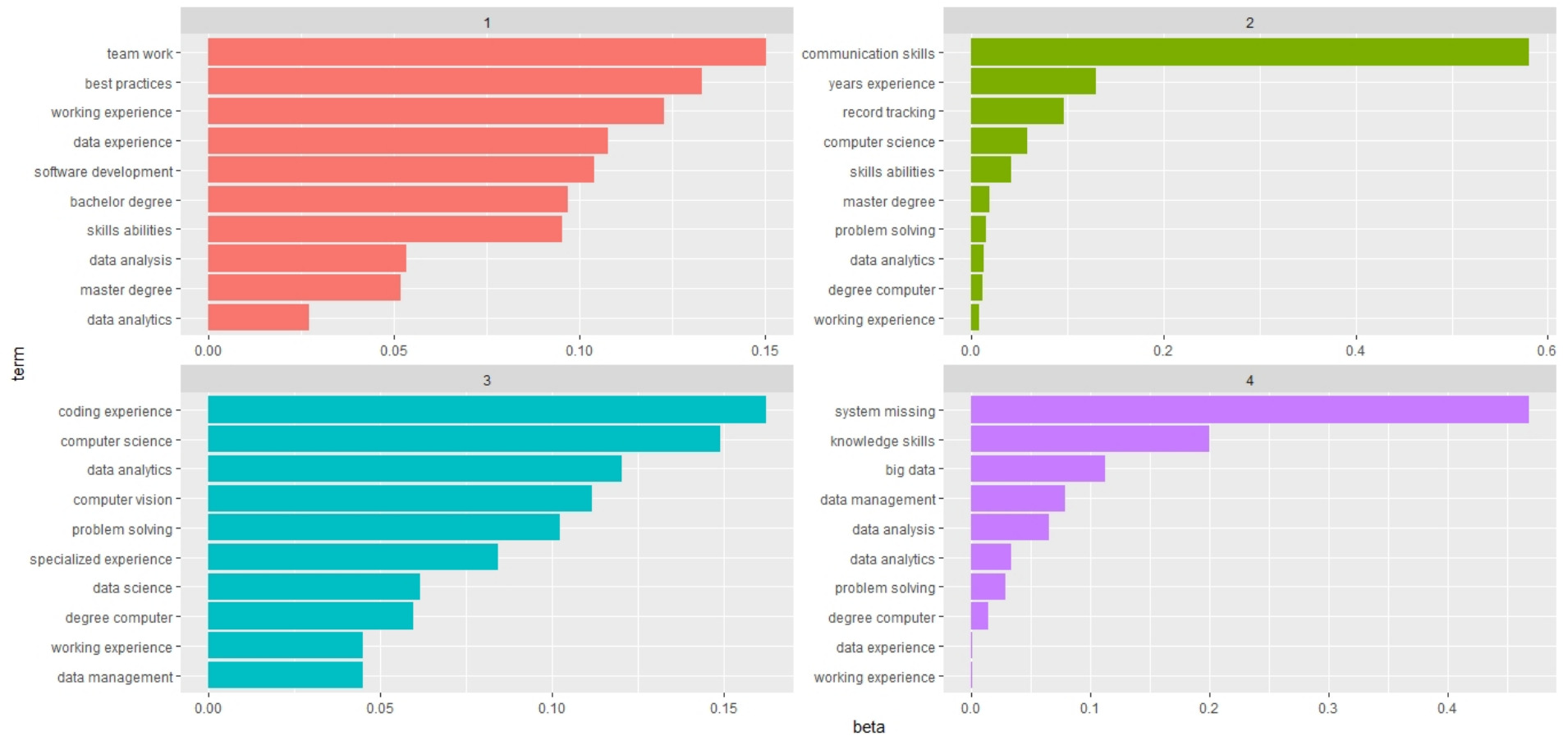
Results from the Data Science industry have been obtained analyzing the subset corpus from the extracted ads regarding the sector. A tokenized Document-Term Matrix (DTM) has been built, and sparsity was removed till 61%. **Fig. 35** shows the bigrams from the corpus. The most frequent terminological combinations were years experience (4863), computer science (4575), data science (2750), bachelor's degree (4334), and communication skills (4284). Topic modeling is presented in **Fig. 36** with four thematic areas. **Fig. 37** highlights the main correlations through the skills set. **Fig. 38** detects greedy modularity in the skillset, dividing it in three groups, and the relative memberships are shown in **Fig. 39**. Application of spectral modularity is presented in **Fig. 40** and the relative memberships reported in **Fig. 41**. The employment of optimal modularity detection is shown in **Fig. 42** and their memberships highlighted in **Fig. 43**. Modularity indicators were then compared to define the most proper method to give sense to the analysis. Having $\xi_G > \xi_O > \xi_S$, the dendrogram in **Fig. 44** is built with greedy modularity, and partial correlations will be used for the weighted network in **Fig. 45**. Thus, a clustering plot with Zhang, Onnela, and Barrat methods is reported in **Fig. 46**. Centrality measures are exposed in **Fig. 47**. The most between skills in the set were big data (67.5%), specialized experience (51.38%), teamwork (7.3%), knowledge skills (7.1%), and data science (4.3%). The closest skills were big data (48.3%), specialized experience (48.8%), data science (29.9%), computer vision (28.36%), and coding experience (27.2%). MCMC with MAP method is shown in **Fig. 48** to forecast and simulate a possible job interview for the Data Science industry.

Fig. 35: Bigrams of the Data Science skillset.



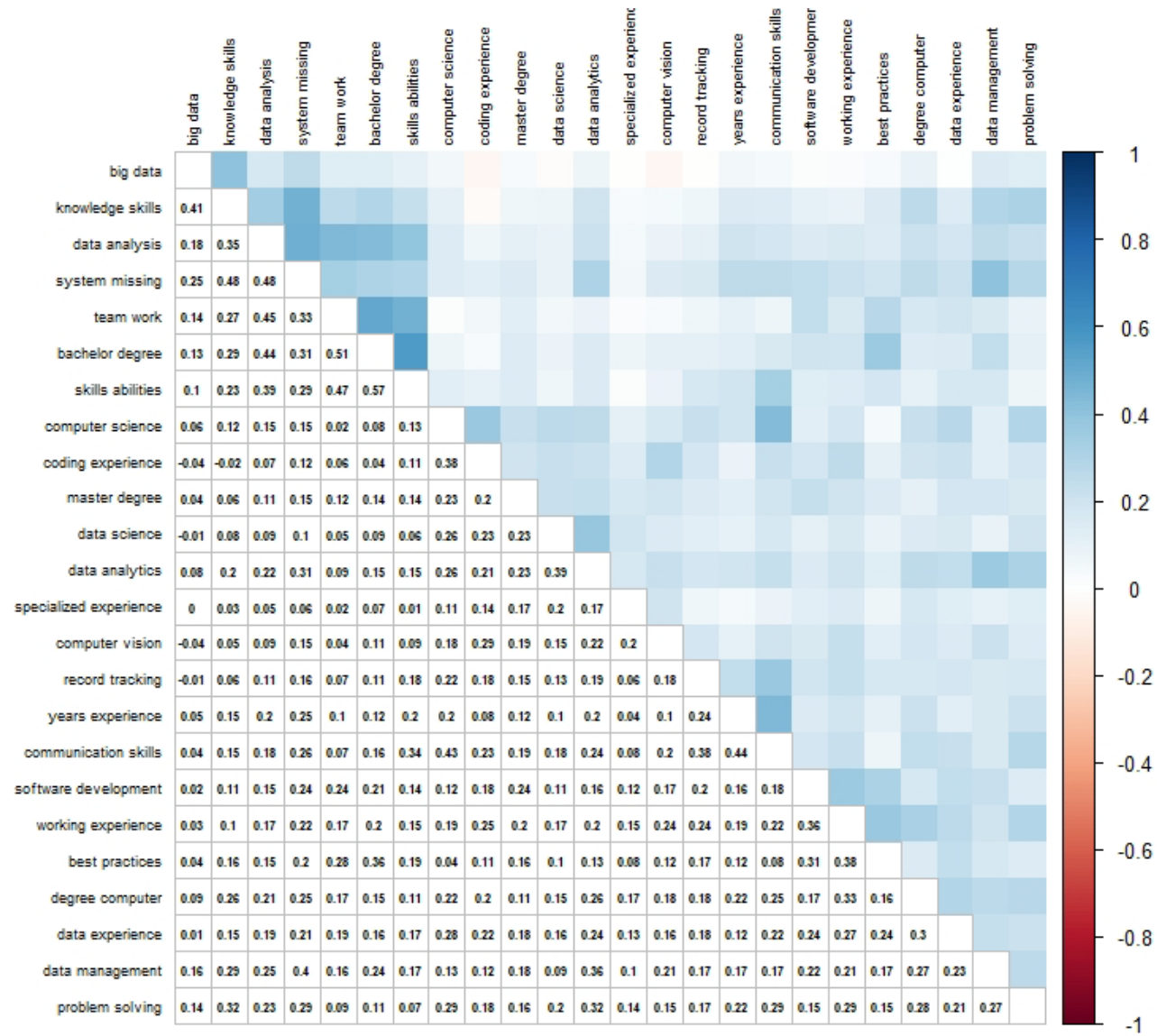
Source: own elaboration

Fig. 36: Topic modeling of the Data Science skillset.



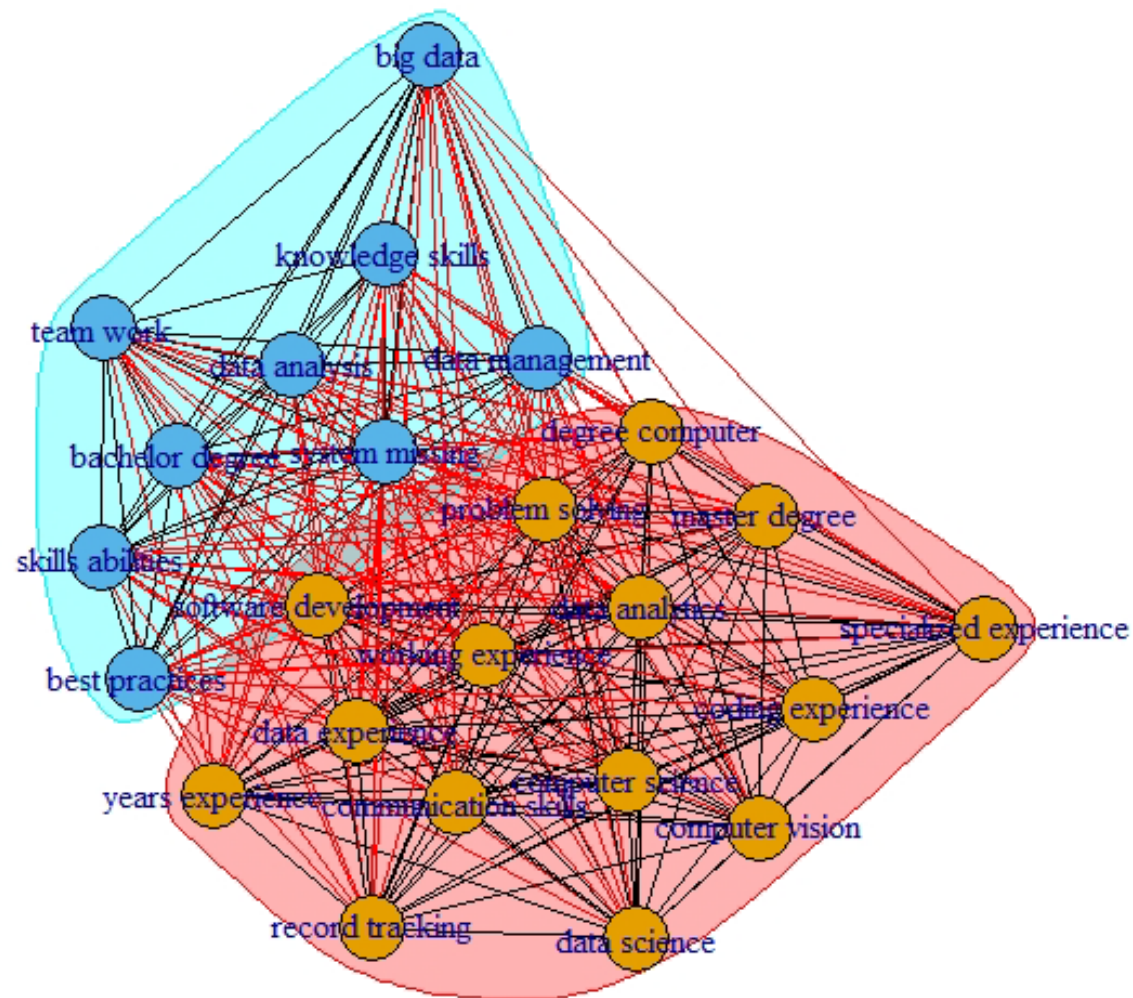
Source: own elaboration.

Fig. 37: Corrplot of the Data Science skillset.



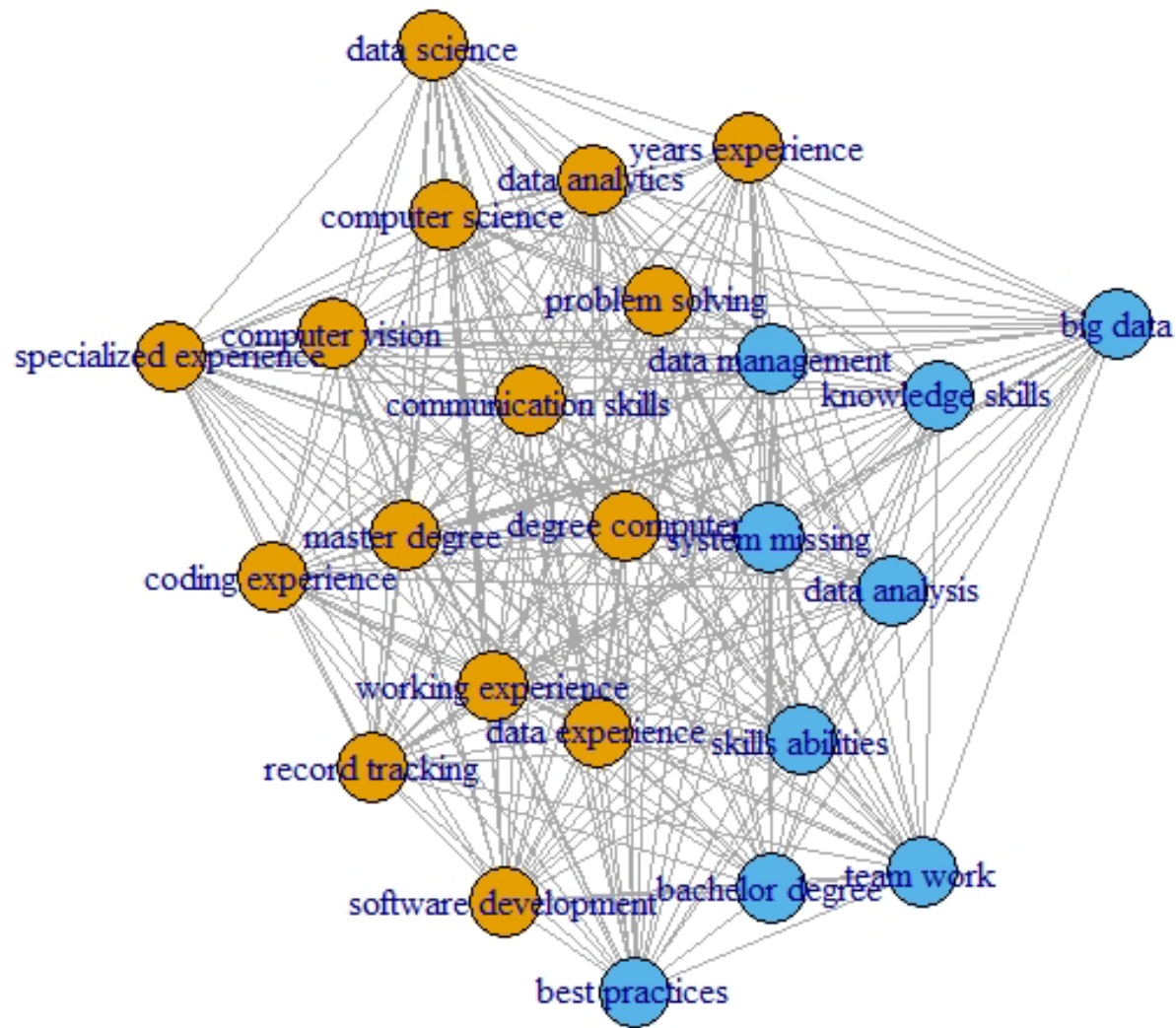
Source: own elaboration.

Fig. 38: Skills network with greedy modularity community detection.



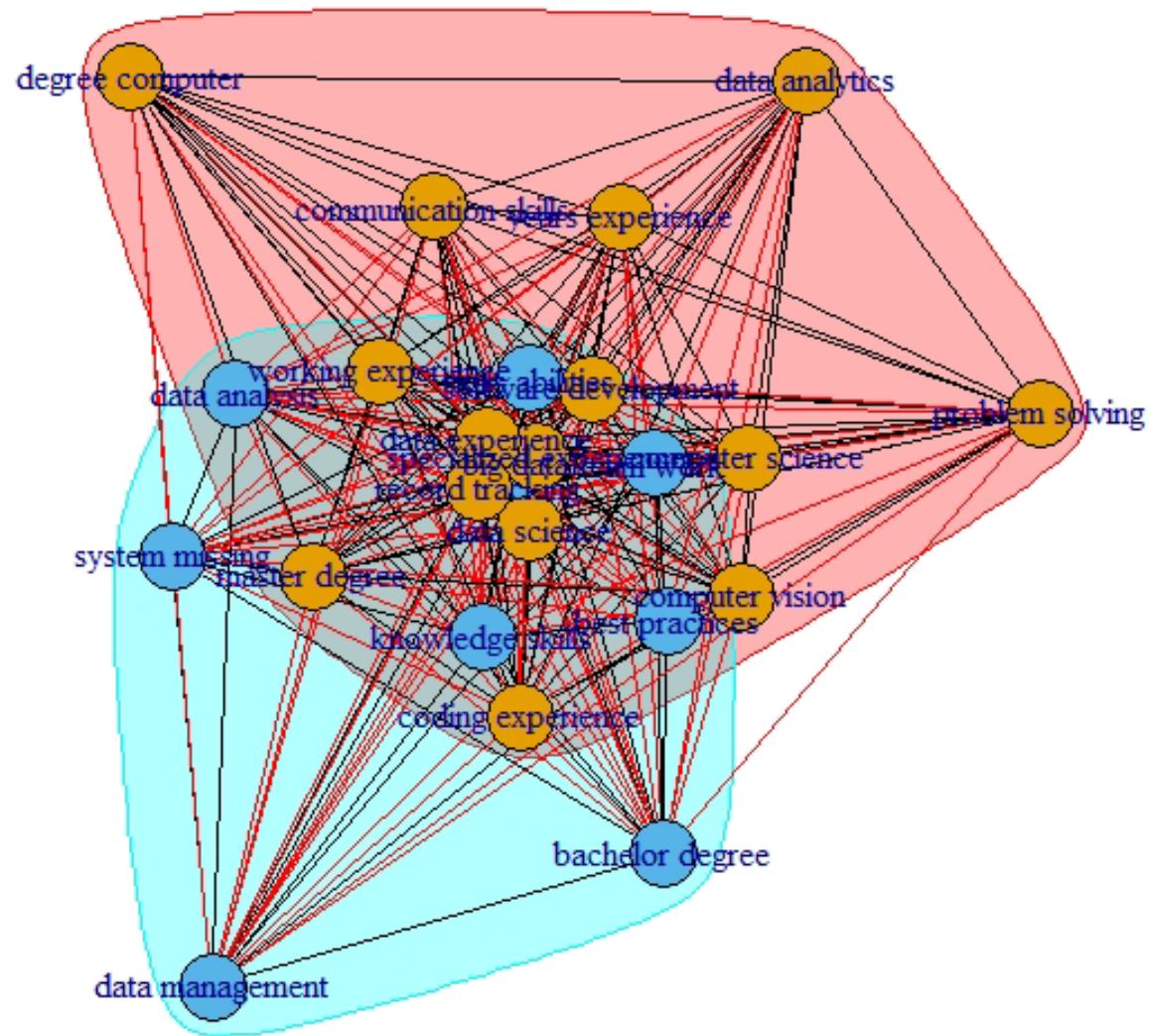
Source: own elaboration.

Fig. 39: Skills network community membership according to greedy modularity.



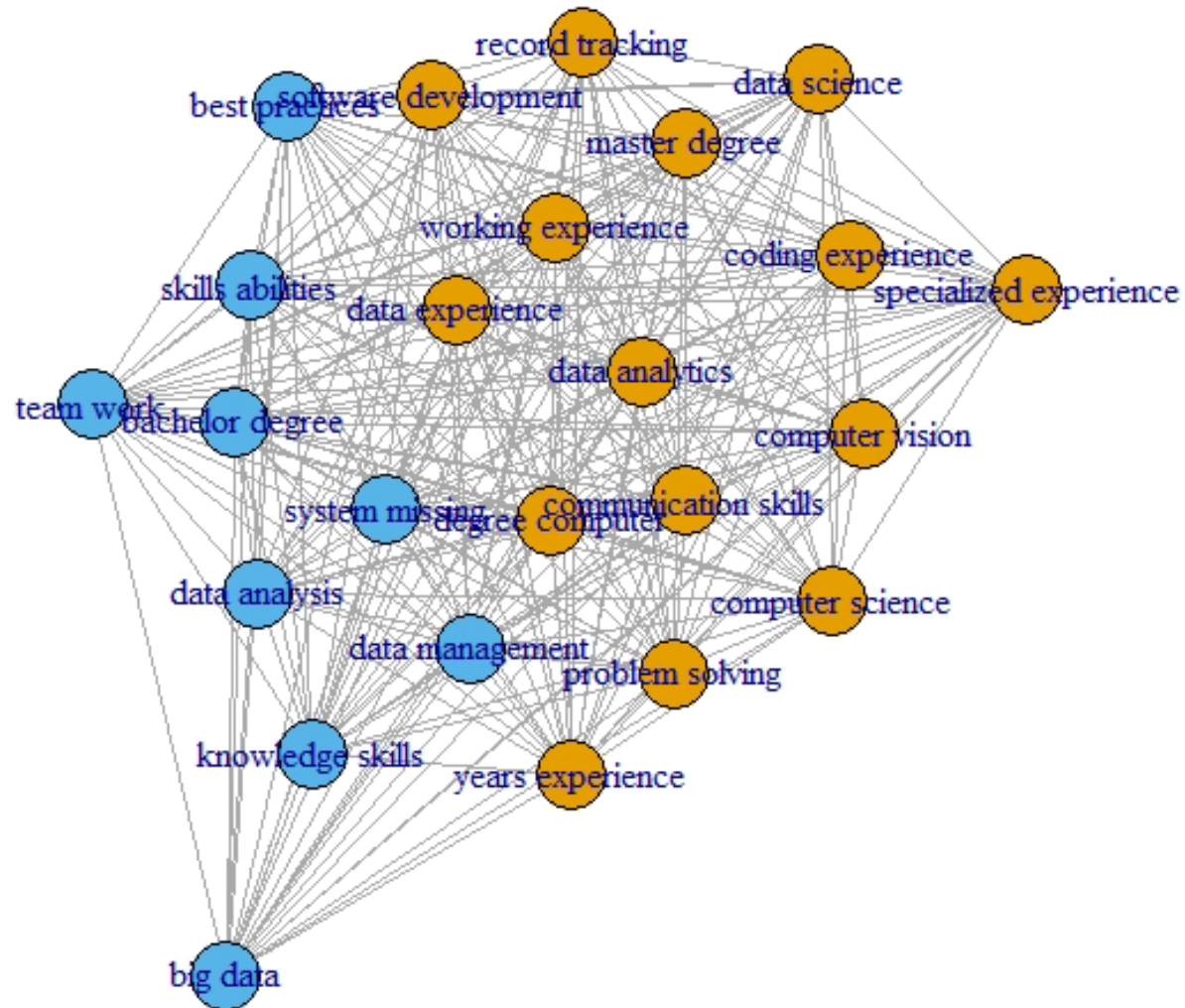
Source: own elaboration.

Fig. 40: Skills network with spectral modularity community detection.



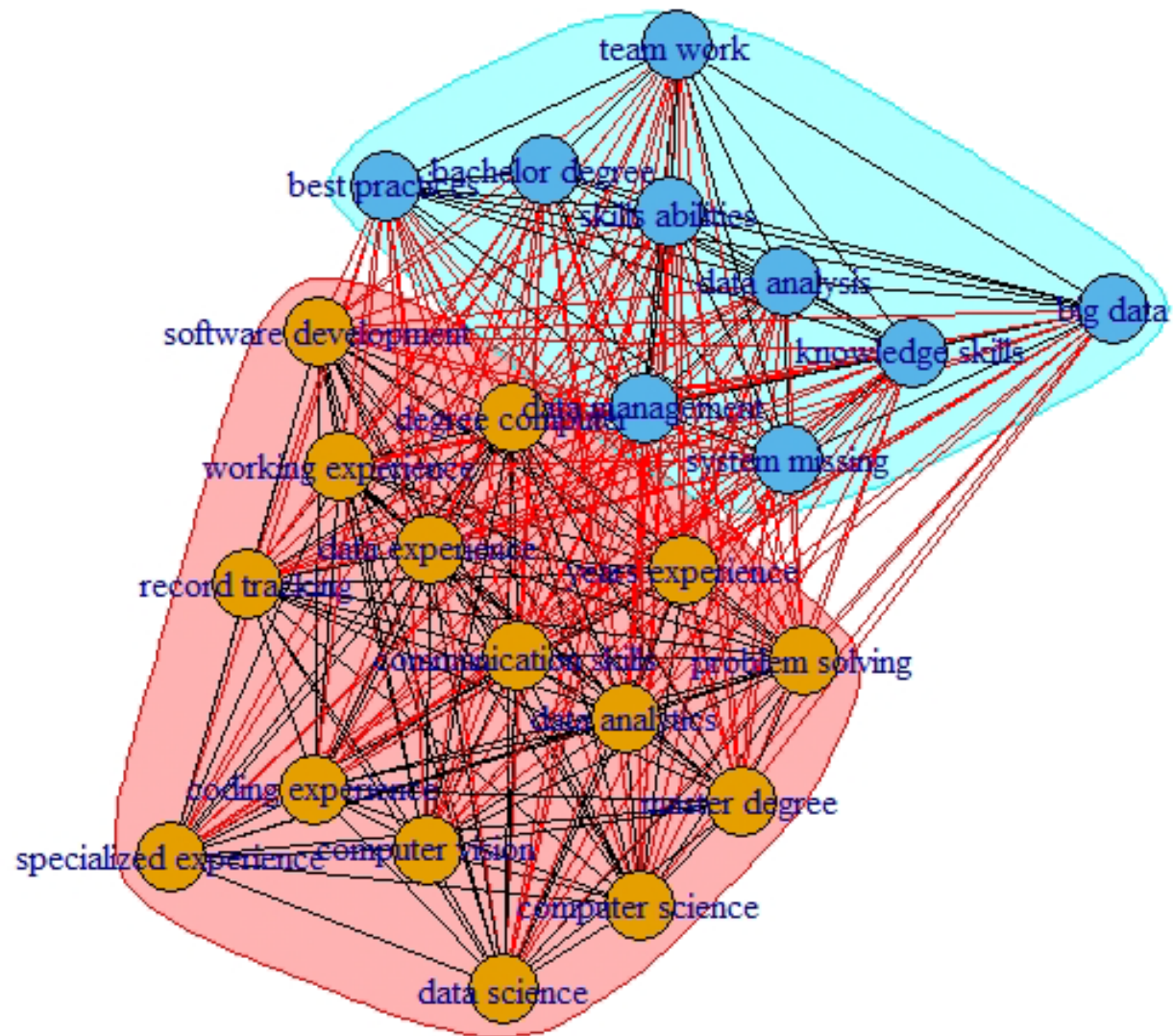
Source: own elaboration.

Fig. 41: Skills network community membership according to spectral modularity.



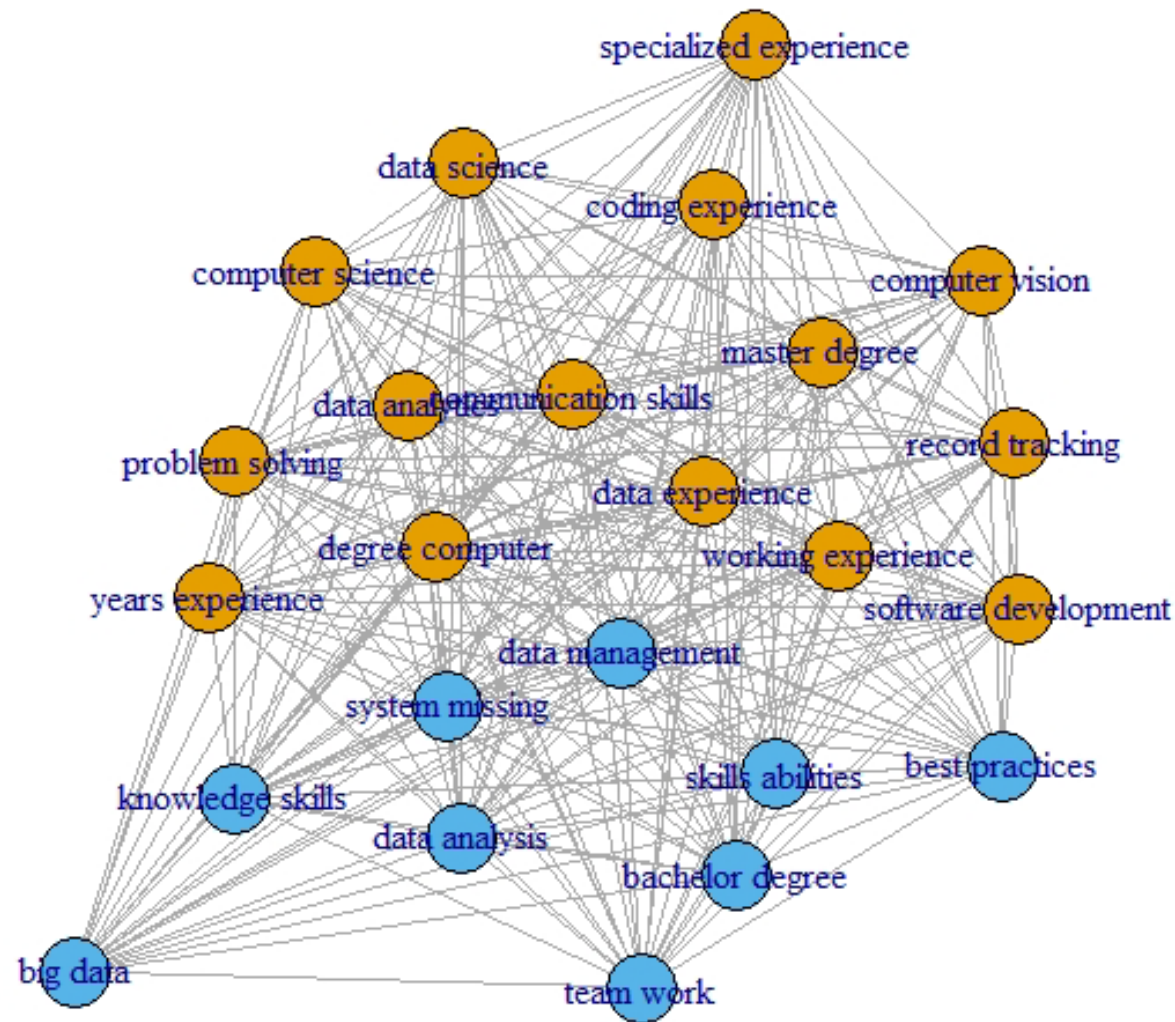
Source: own elaboration.

Fig. 42: Skills network with optimal community detection.



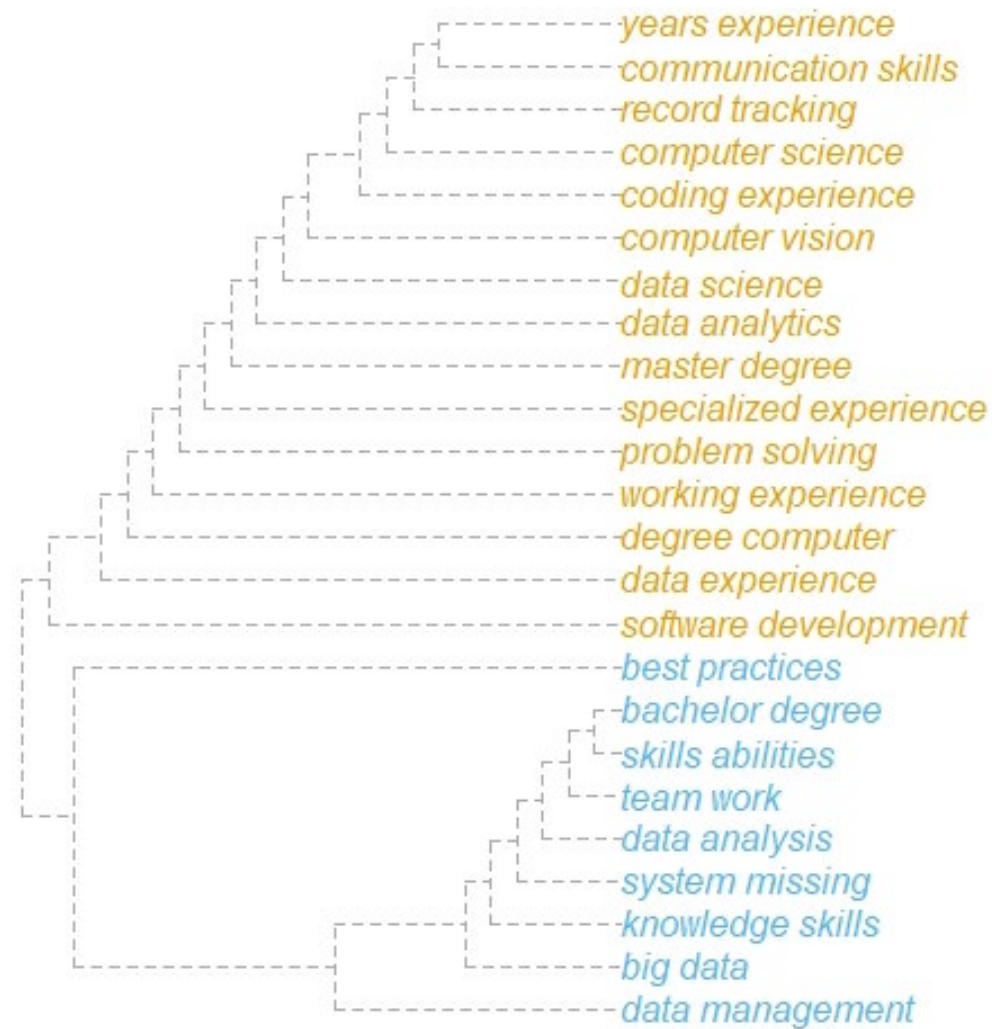
Source: own elaboration.

Fig. 43: Skills network community membership according to optimal modularity.



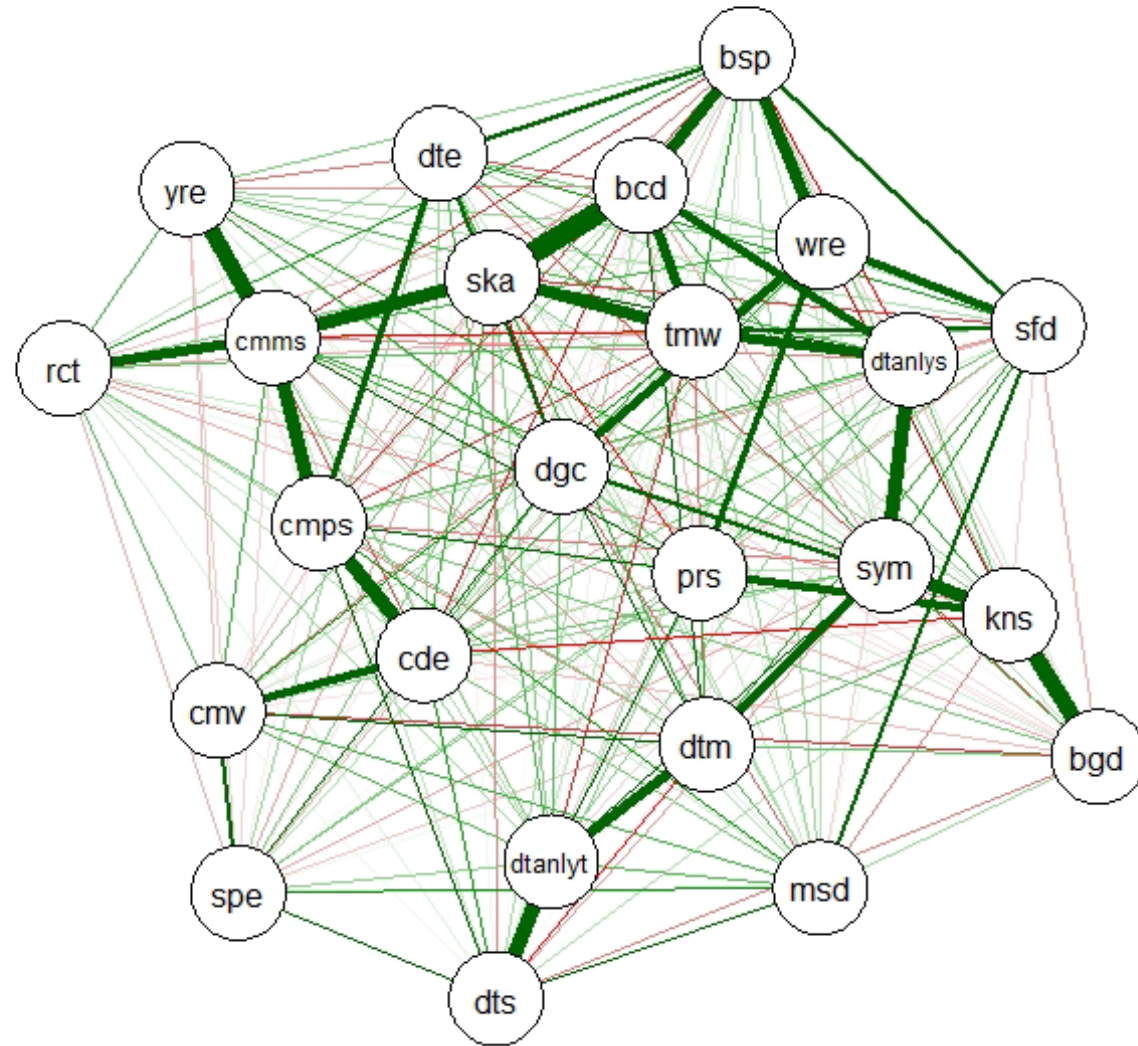
Source: own elaboration.

Fig. 44: Dendrogram with greedy modularity.



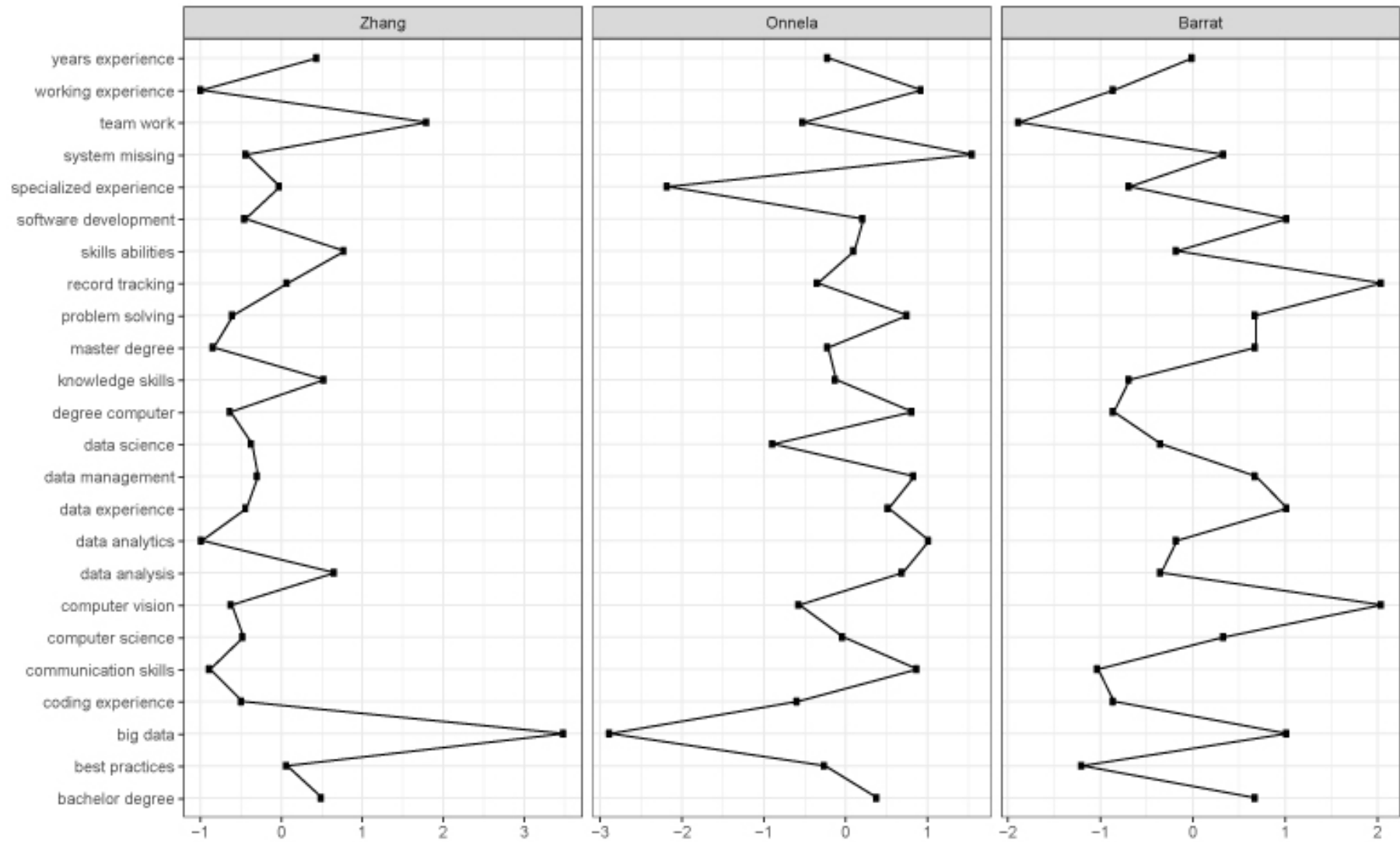
Source: own elaboration

Fig. 45: Weighted skills network via partial correlations clustering.



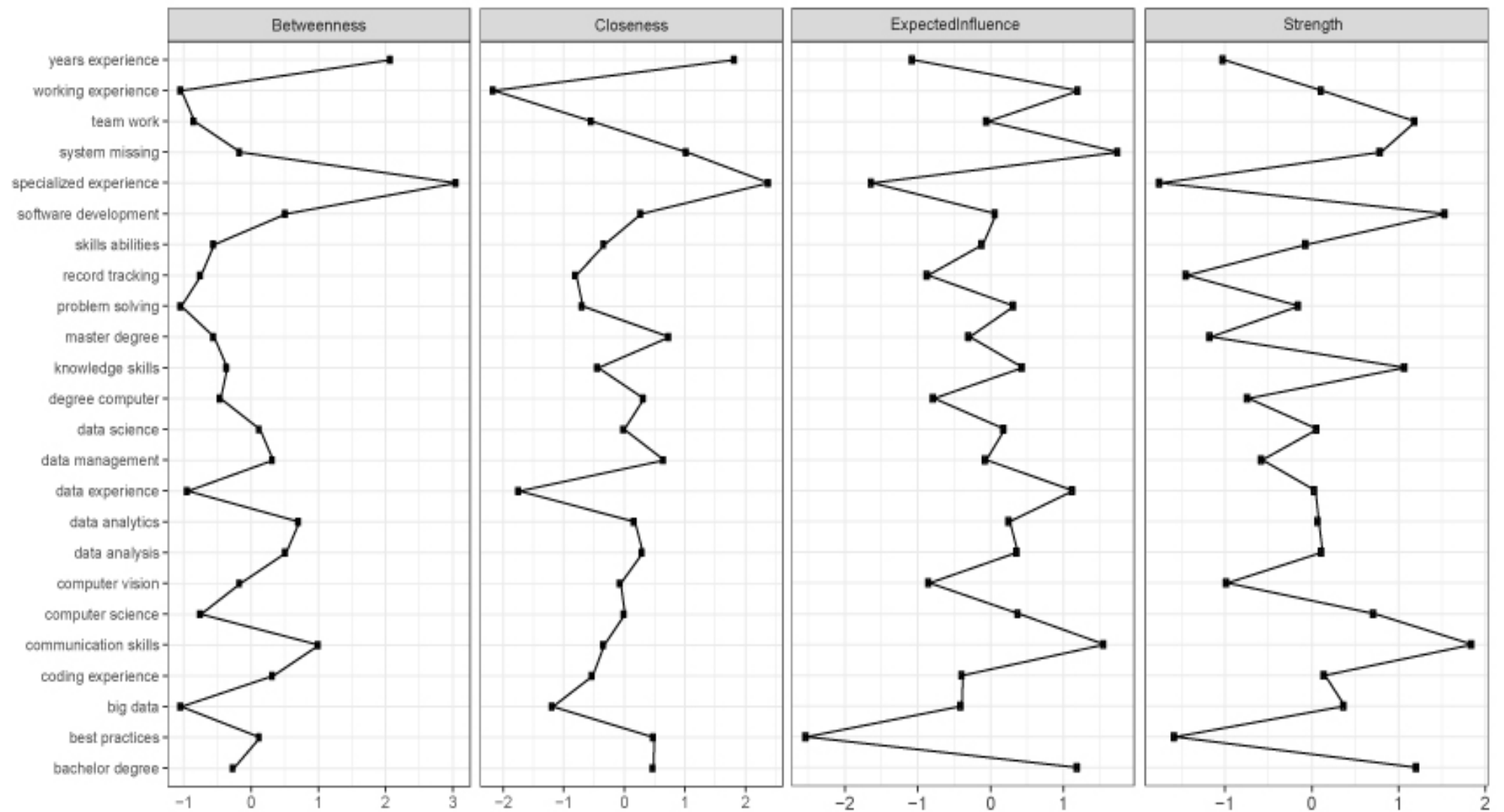
Source: own elaboration.

Fig. 46: Clustering plot with compared methods.



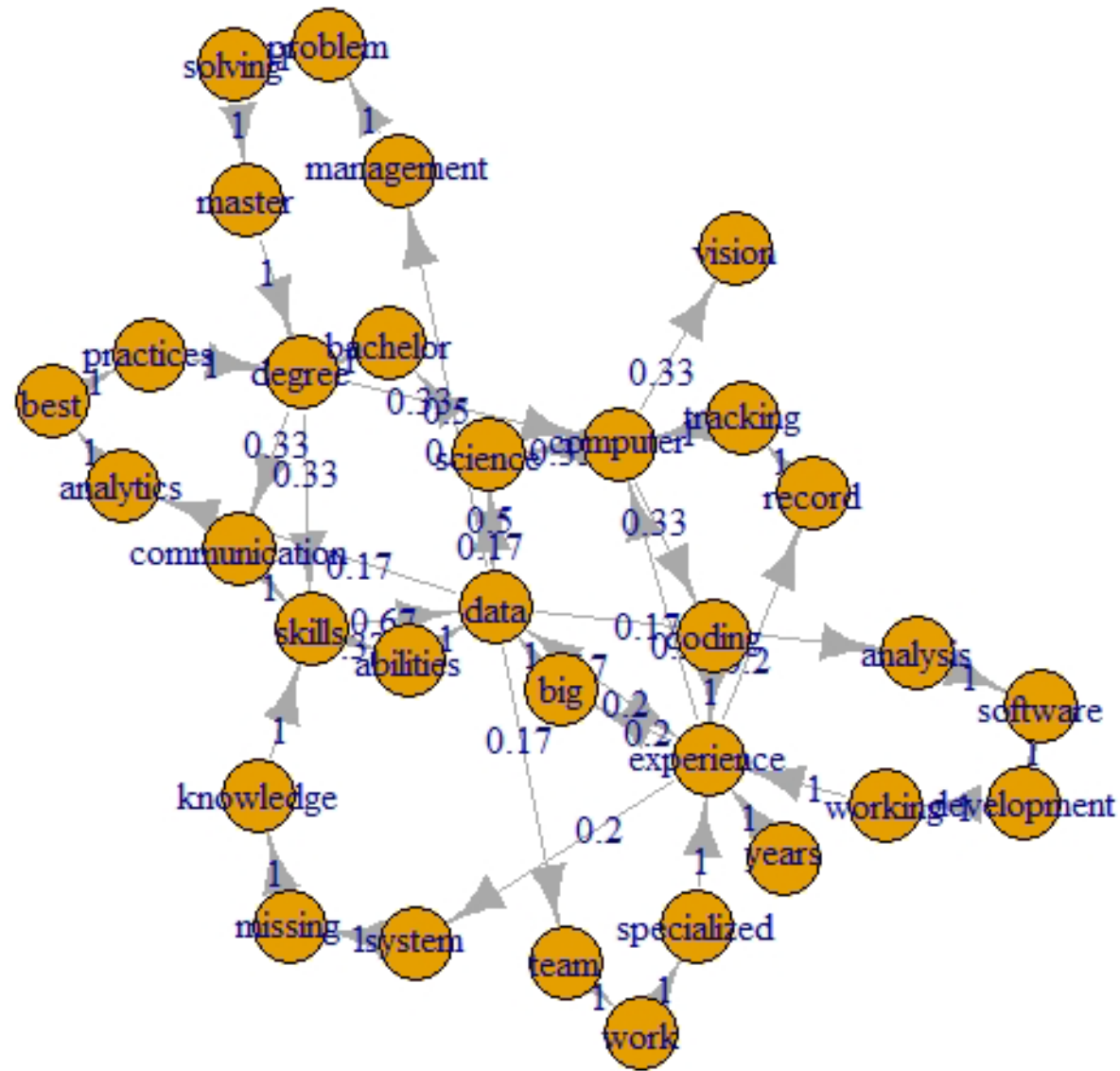
Source: own elaboration.

Fig. 47: Centrality measures plot.



Source: own elaboration.

Fig. 48: Monte Carlo Markov Chain with MAP method.

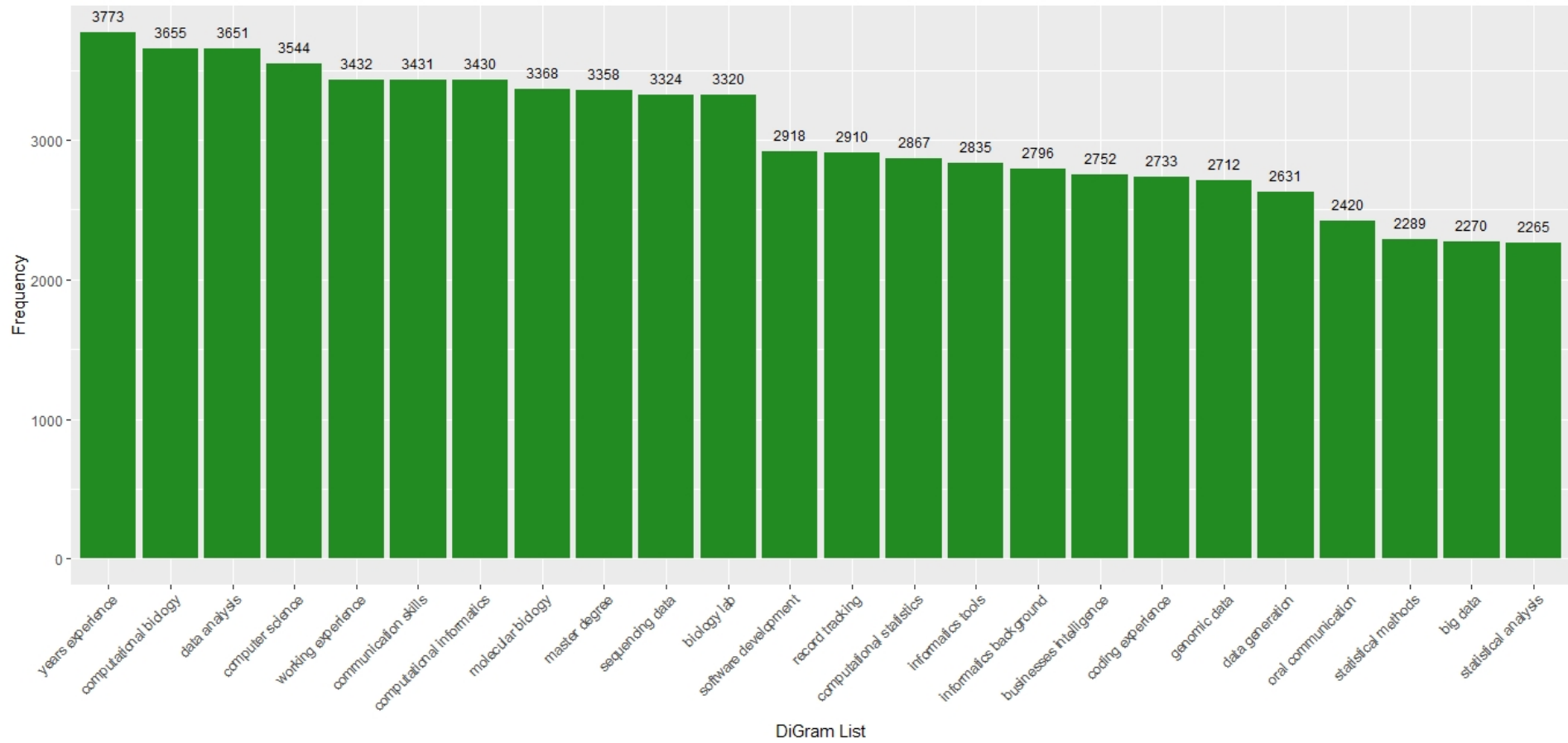


Source: own elaboration.

4.2.4 Bioinformatics

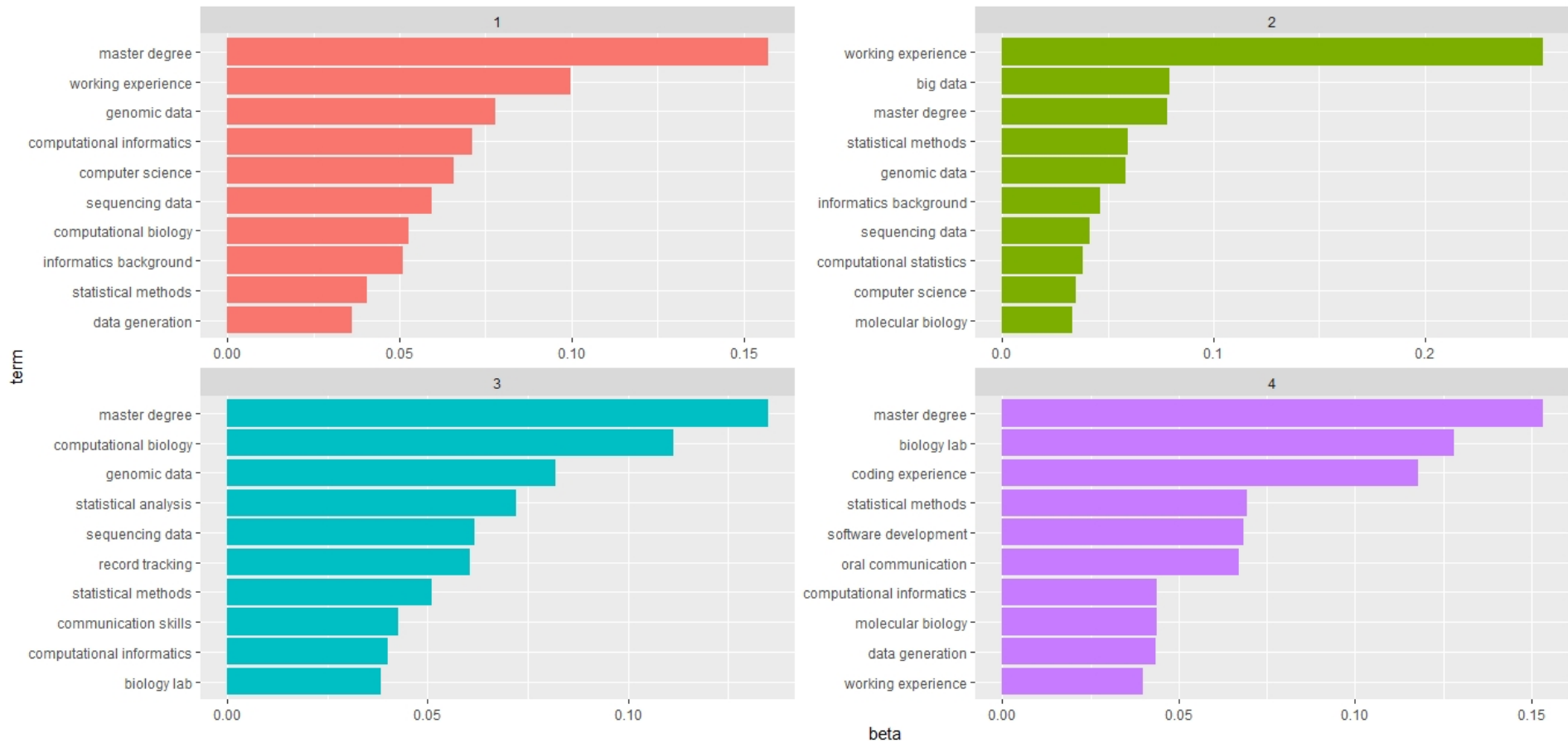
Results from the Bioinformatics industry have been obtained analyzing the subset corpus from the extracted ads regarding the sector. A tokenised Document-Term Matrix (DTM) has been built, and sparsity was removed till 58%. **Fig. 49** shows the bigrams from the corpus. The most frequent terminological combinations were years experience (3773), computational biology (3655), data analysis (3651), computer science (3544), and working experience (3432). Topic modeling is presented in **Fig. 50** with four thematic areas. **Fig. 51** highlights the main correlations through the skills set. **Fig. 52** detects greedy modularity in the skillset, dividing it in three groups, and the relative memberships are shown in **Fig. 53**. Application of spectral modularity is presented in **Fig. 54** and the relative memberships reported in **Fig. 55**. The employment of optimal modularity detection is shown in **Fig. 56** and their memberships highlighted in **Fig. 57**. Modularity indicators were compared to define the most proper method to give sense to the analysis. Having $\xi_G > \xi_O > \xi_S$, the dendrogram in **Fig. 58** is built with greedy modularity, and partial correlations will be used for the weighted network in **Fig. 59**. Thus, a clustering plot with Zhang, Onnela, and Barrat methods is reported in **Fig. 60**. Centrality measures are exposed in **Fig. 61**. The most between skills in the set were informatic tools (39.5%), years experience (32%), record tracking (31.6%), coding experience (21.3%), and molecular biology (17.7%). The closest skills were years experience (53.1%), informatic tools (52.5%), molecular biology (37.2%), coding experience (32.7%), and data analysis (30.3%). MCMC with MAP method is shown in **Fig. 62** to forecast and simulate a possible job interview for the Bioinformatics industry.

Fig. 49: Bigrams of the Bioinformatics skillset.



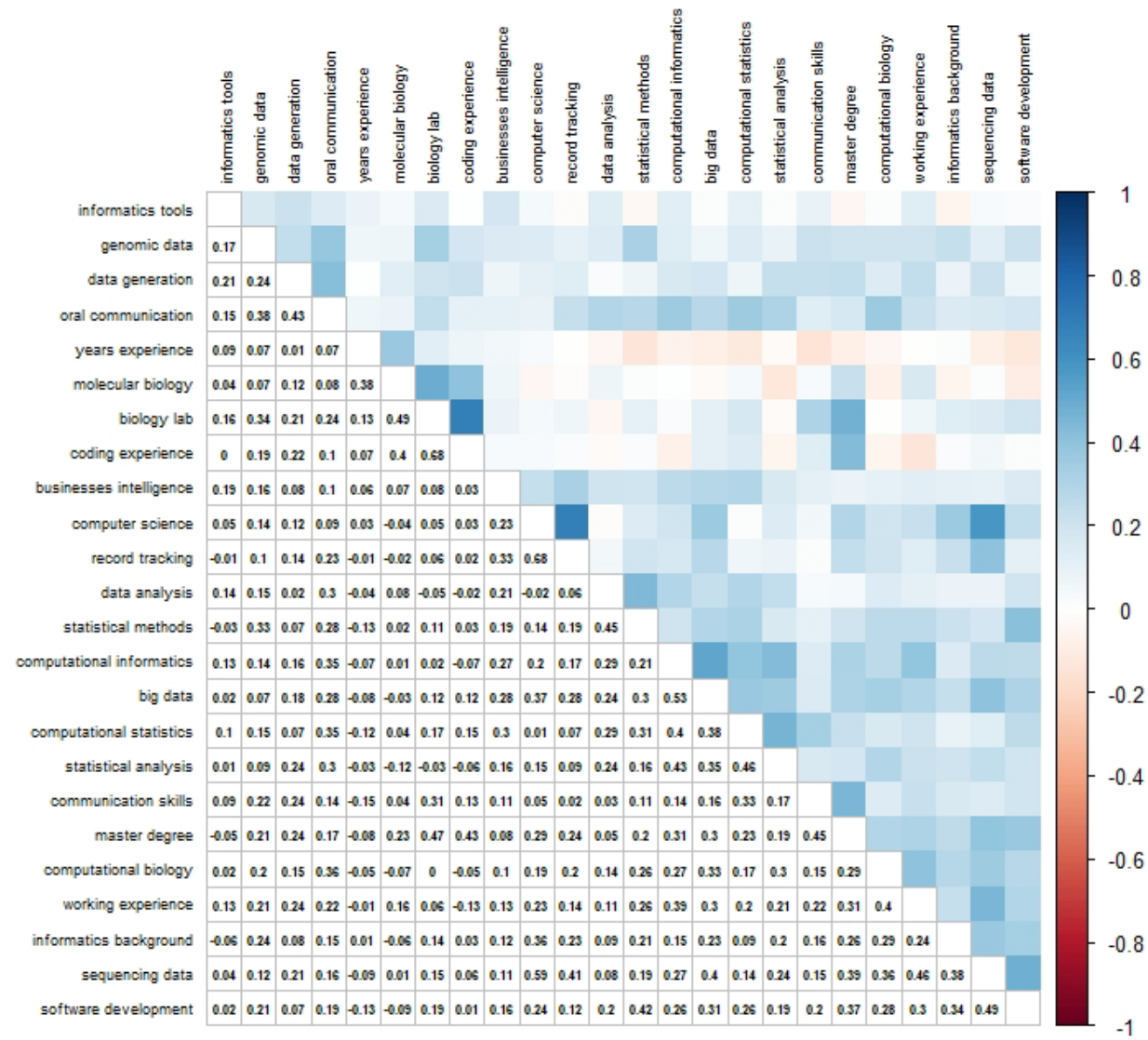
Source: own elaboration

Fig. 50: Topic modeling of the Bioinformatics skillset.



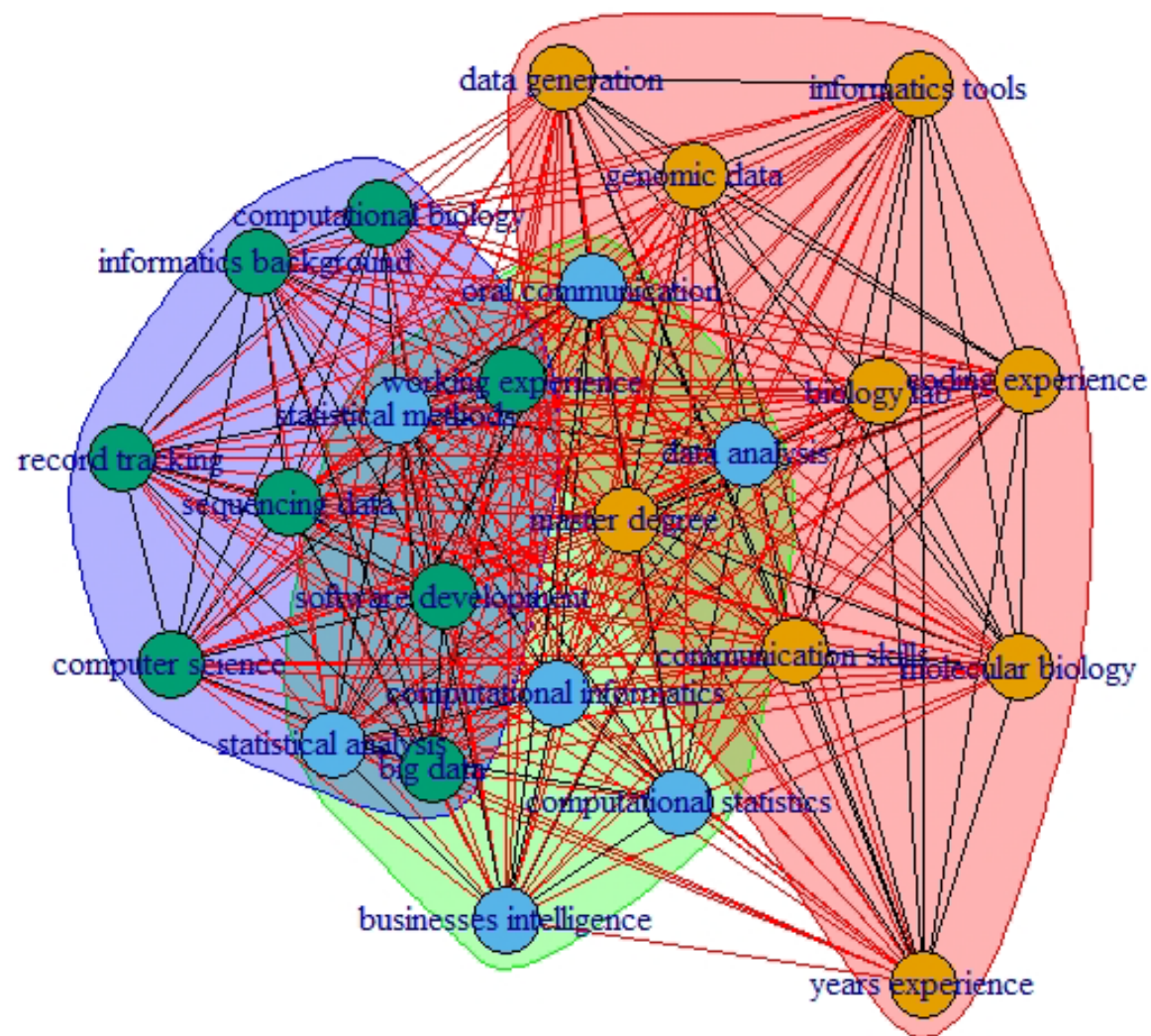
Source: own elaboration.

Fig. 51: Corrplot of the Bioinformatics skillset.



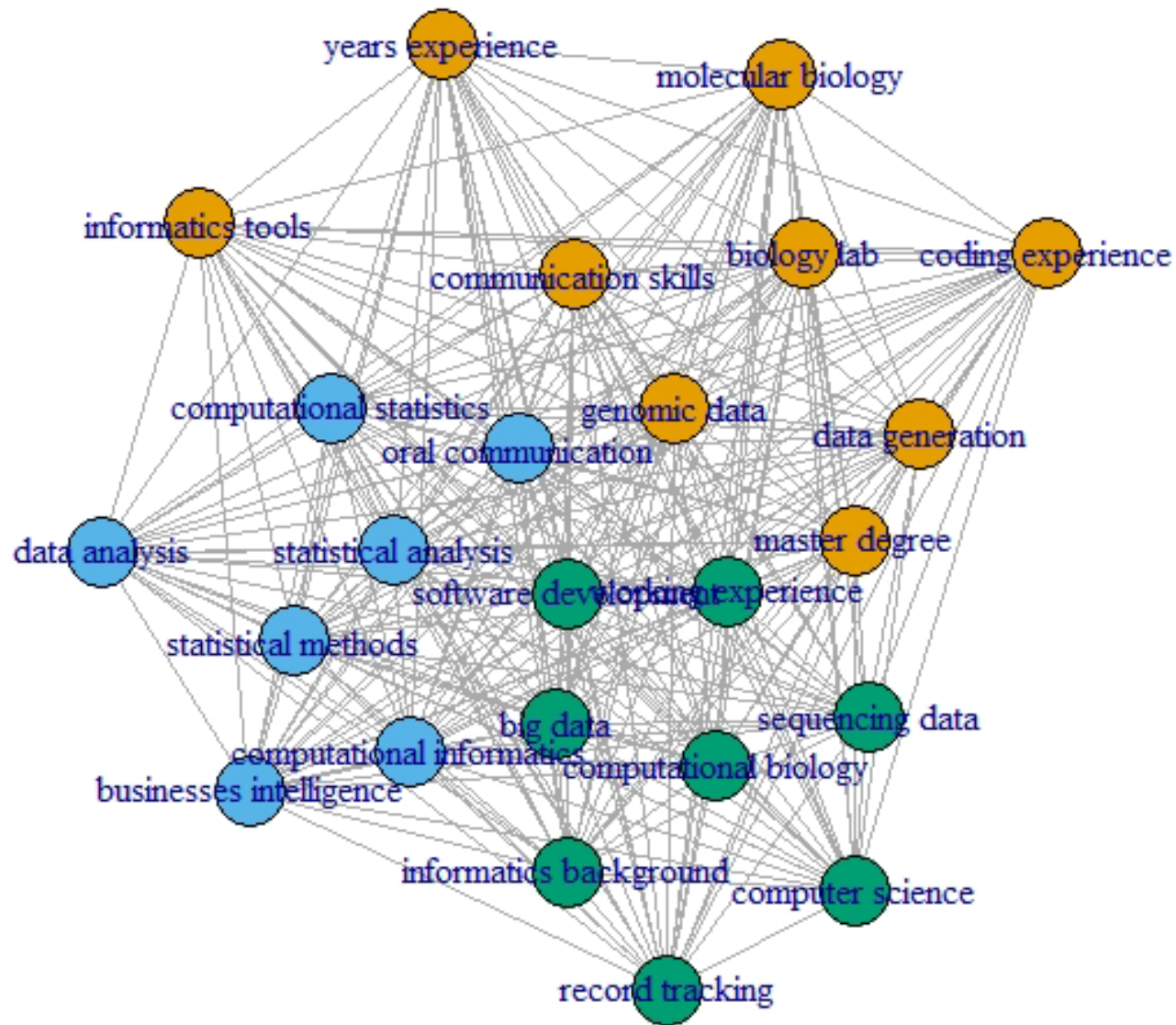
Source: own elaboration.

Fig. 52: Skills network with greedy modularity community detection.



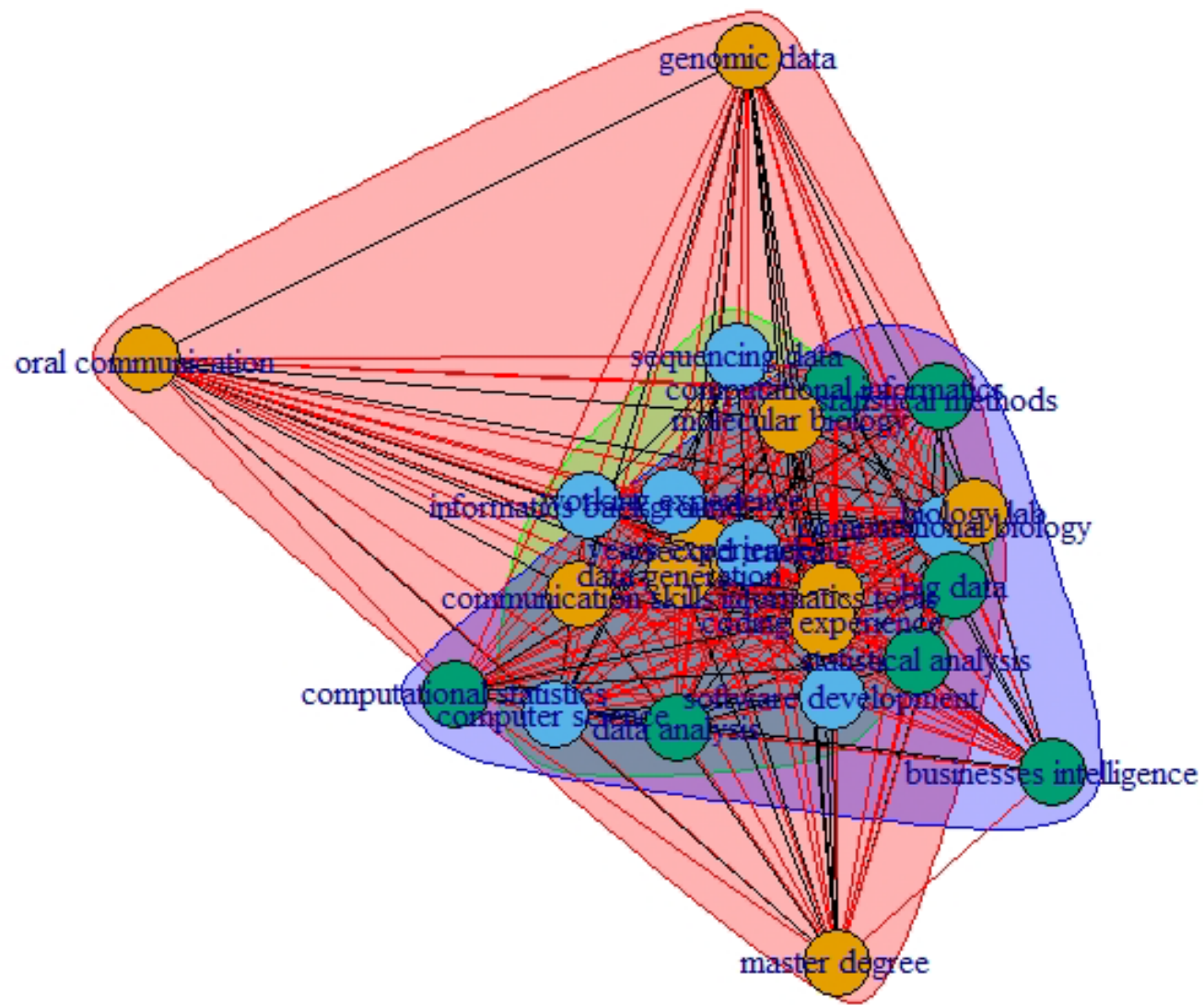
Source: own elaboration.

Fig. 53: Skills network community membership according to greedy modularity.



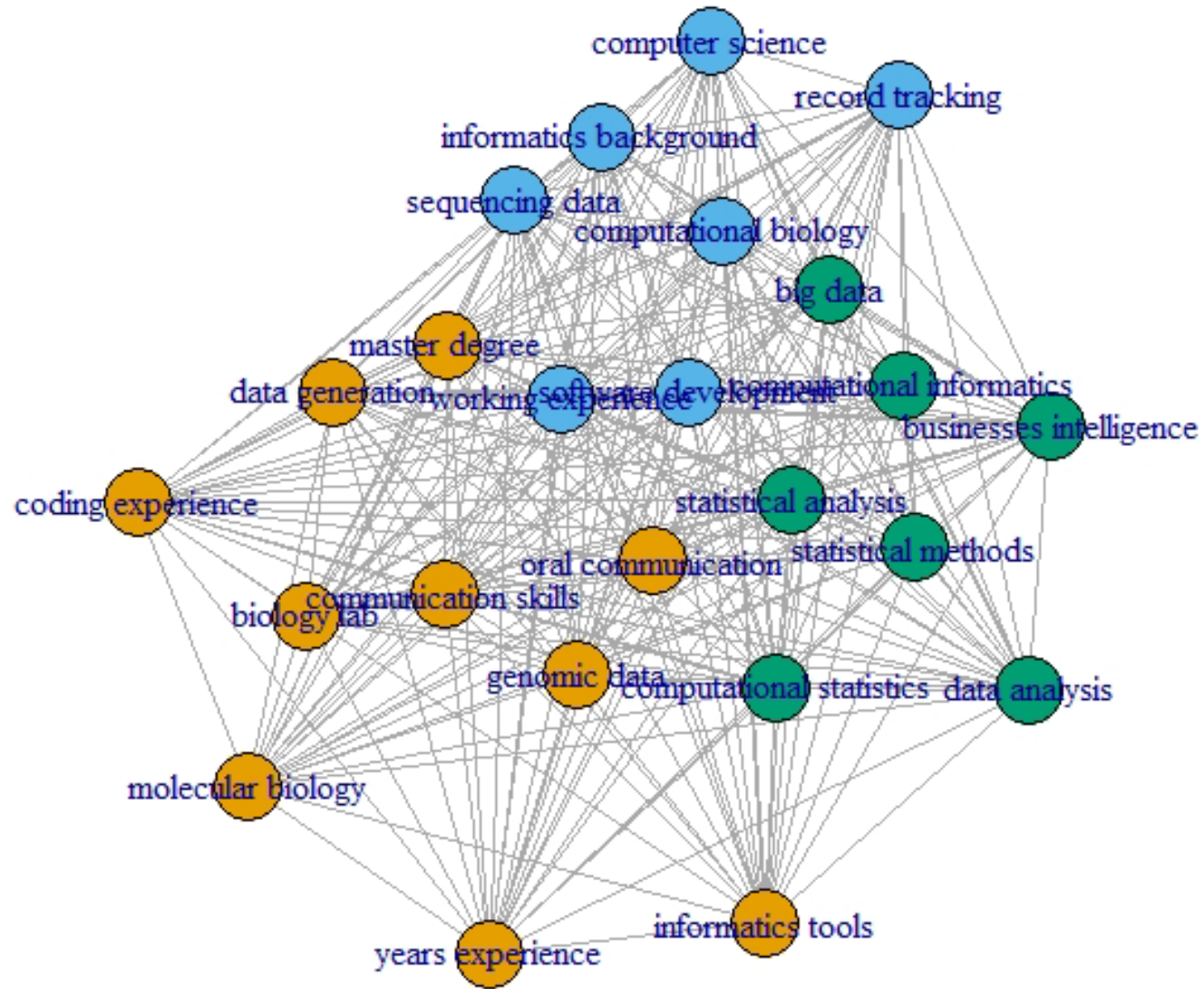
Source: own elaboration.

Fig. 54: Skills network with spectral modularity community detection.



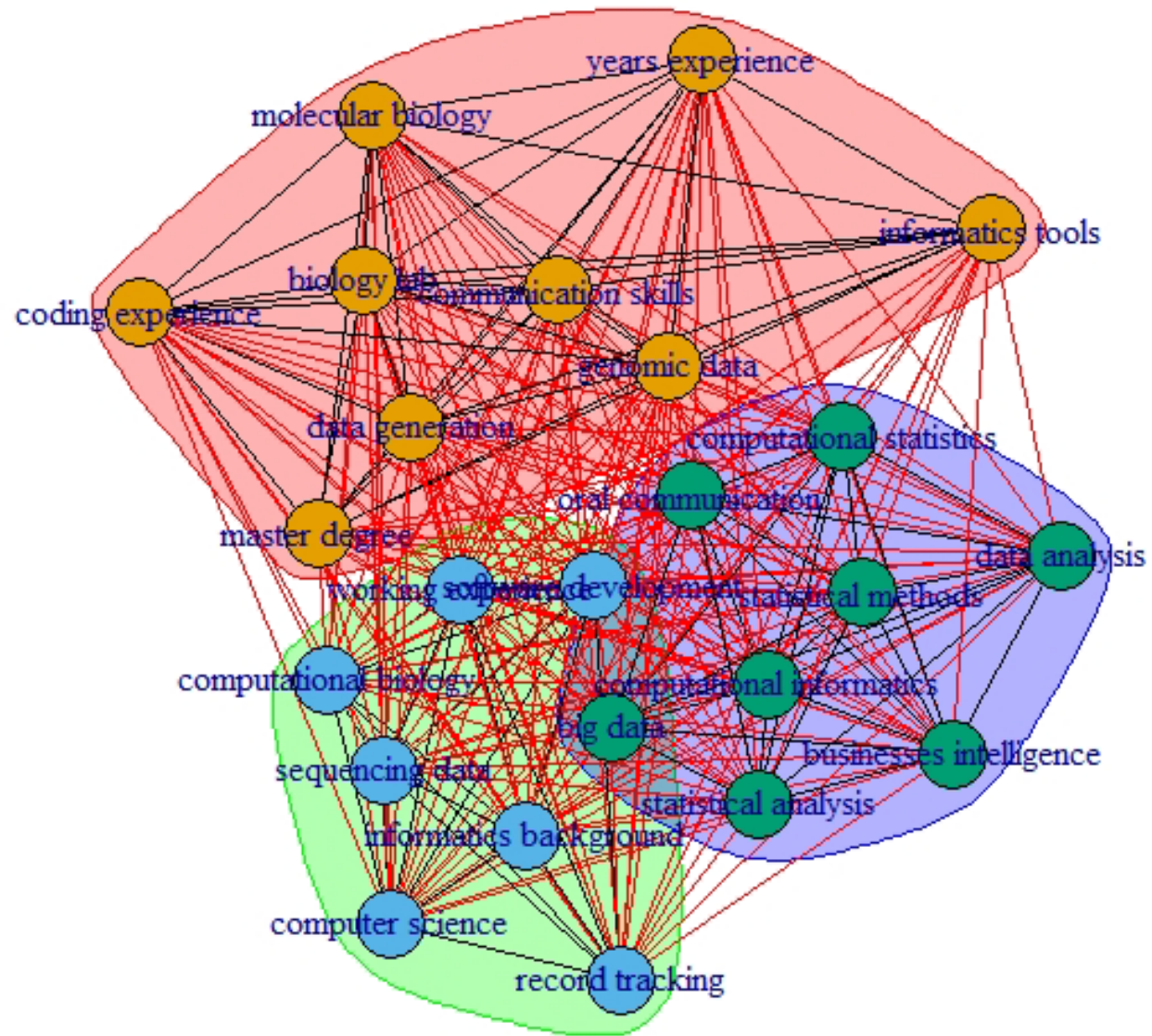
Source: own elaboration.

Fig. 55: Skills network community membership according to spectral modularity.



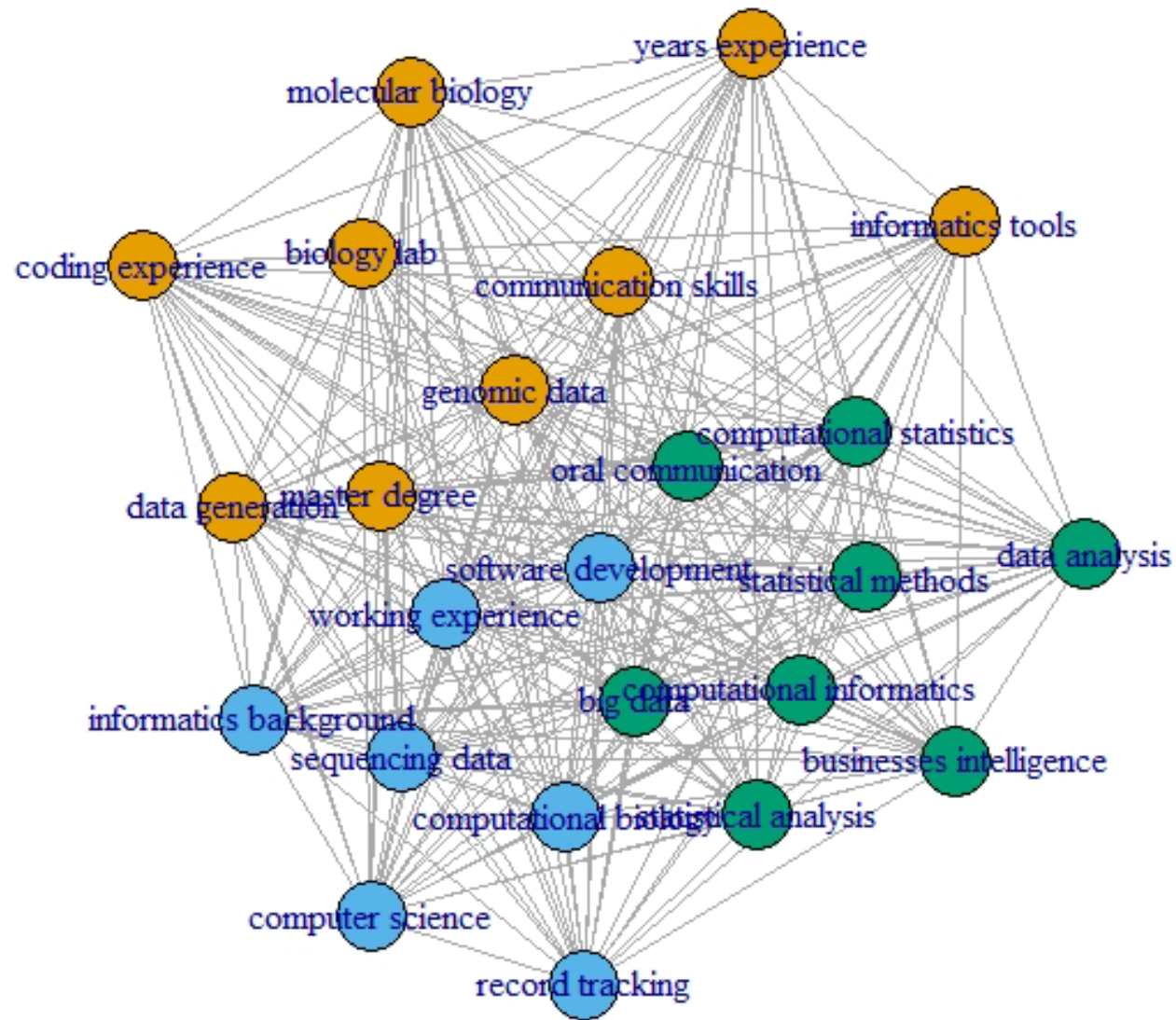
Source: own elaboration.

Fig. 56: Skills network with optimal community detection.



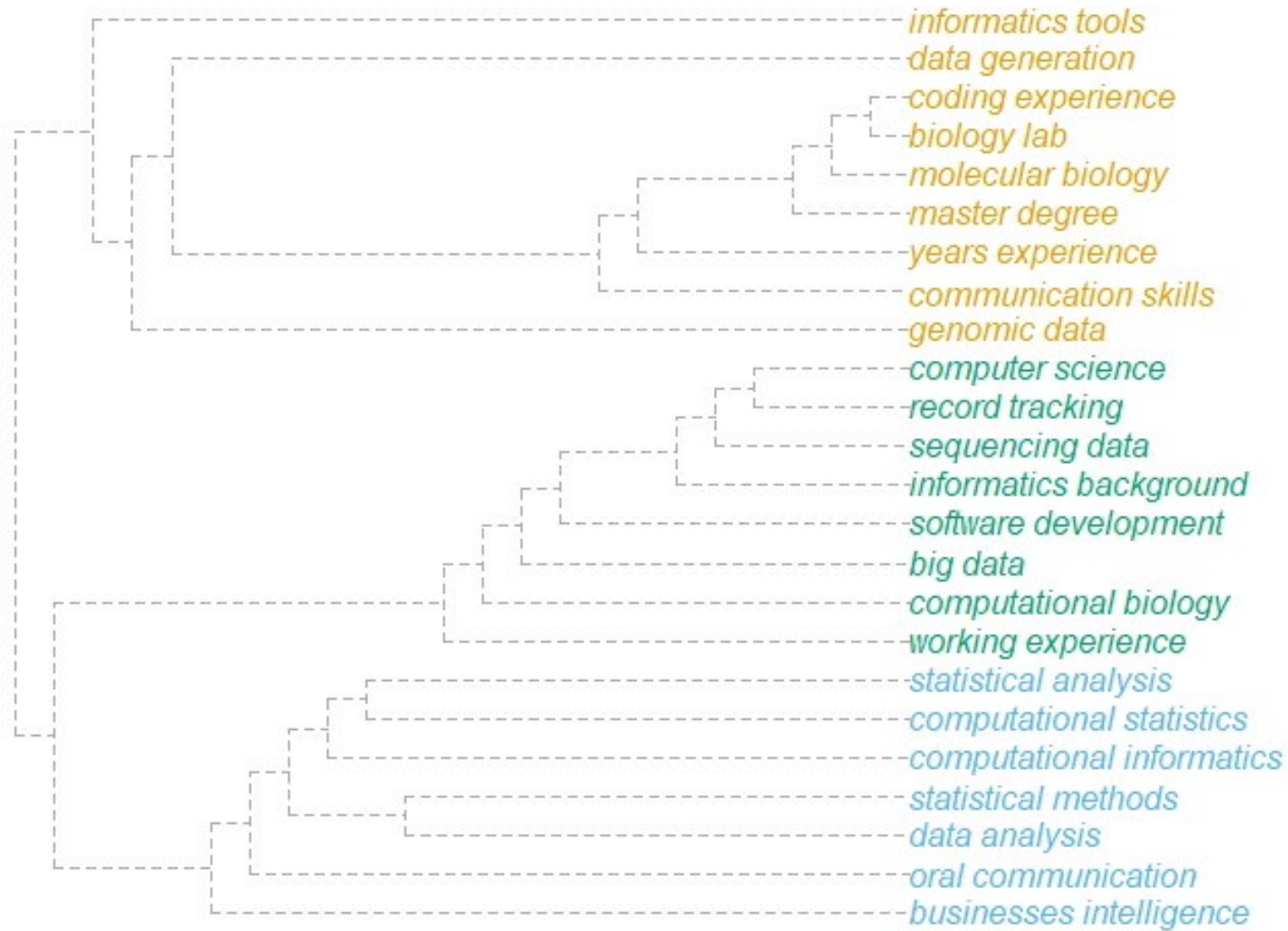
Source: own elaboration.

Fig. 57: Skills network community membership according to optimal modularity.



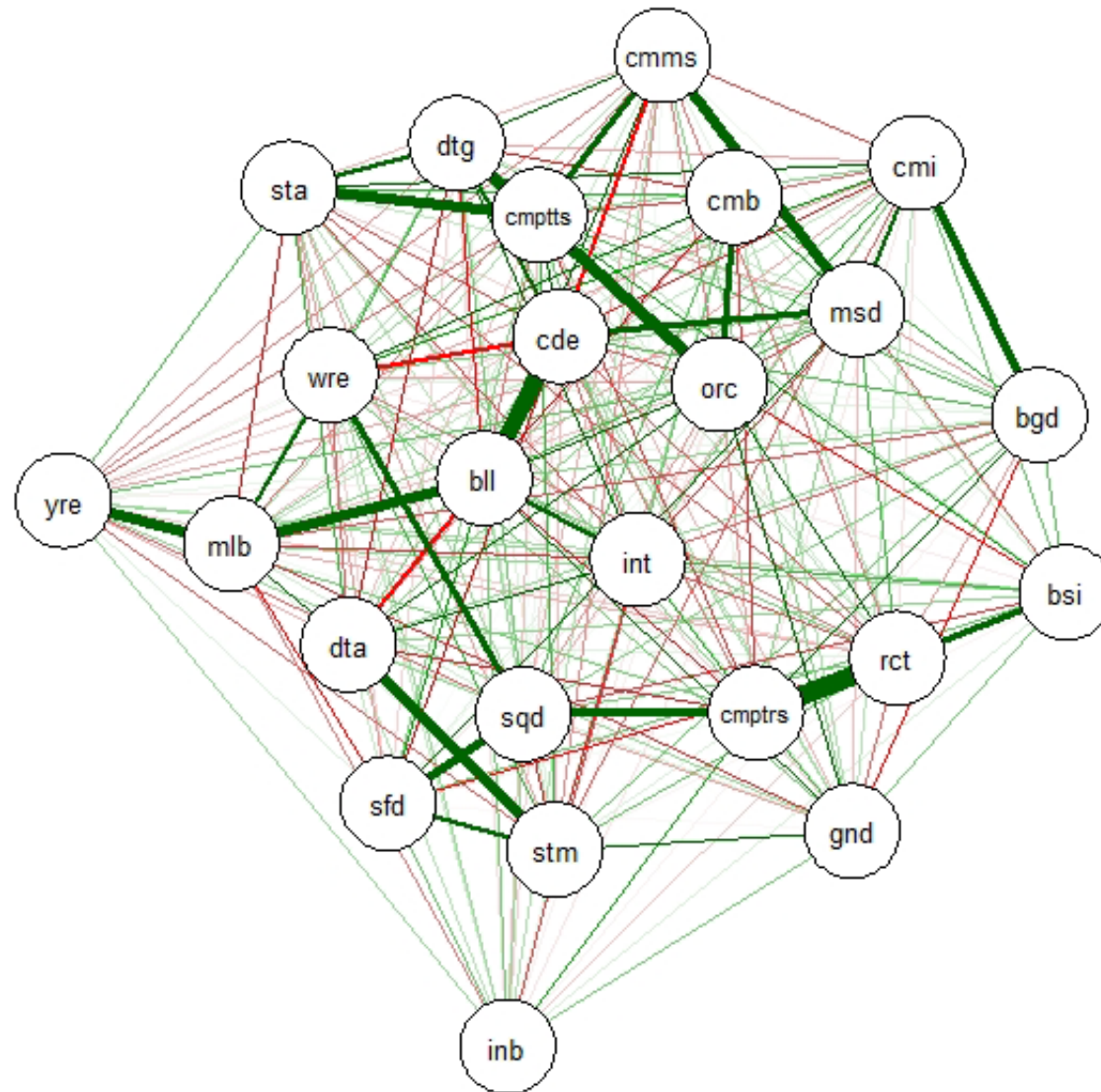
Source: own elaboration.

Fig. 58: Dendrogram with greedy modularity.



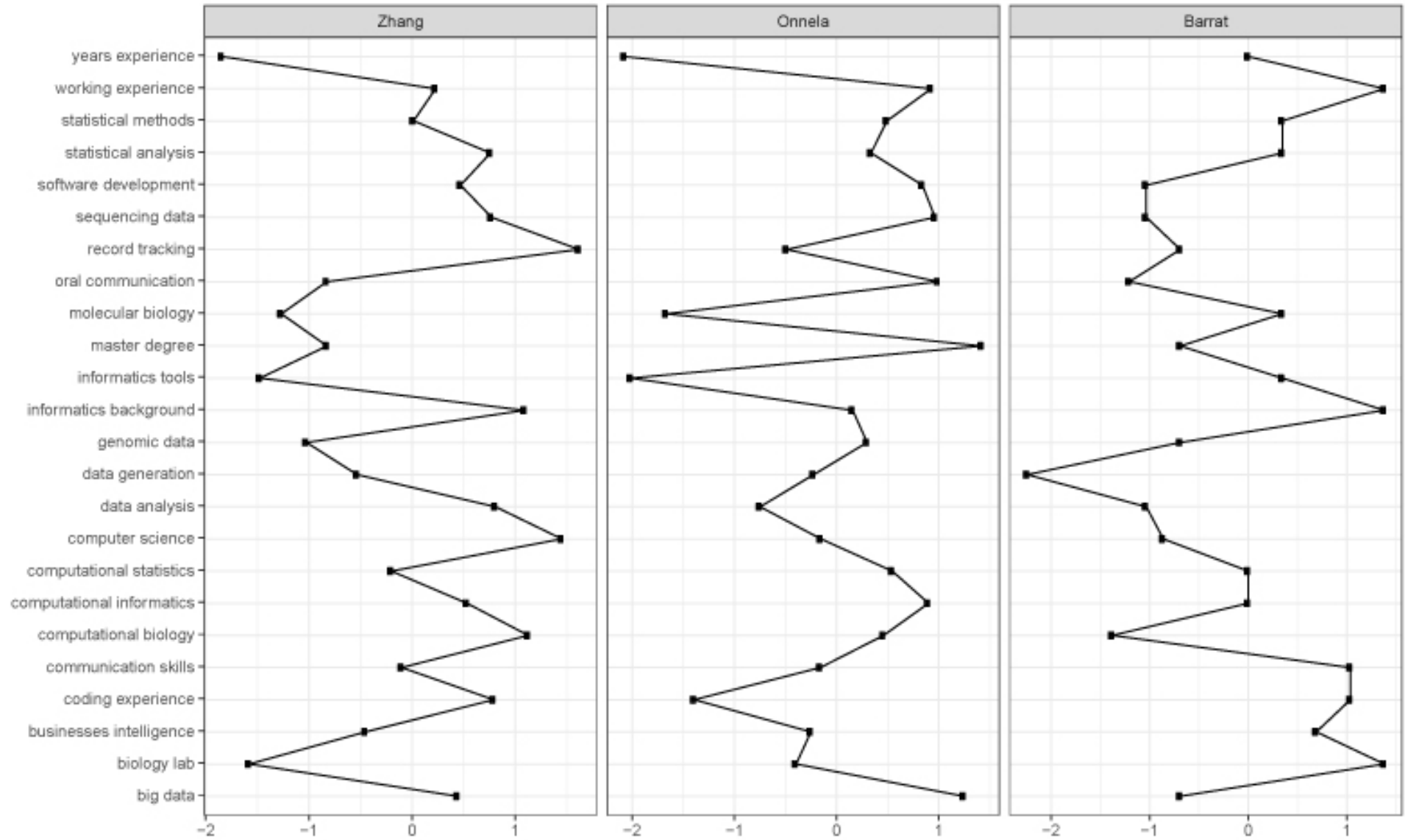
Source: own elaboration

Fig. 59: Weighted skills network via partial correlations clustering.



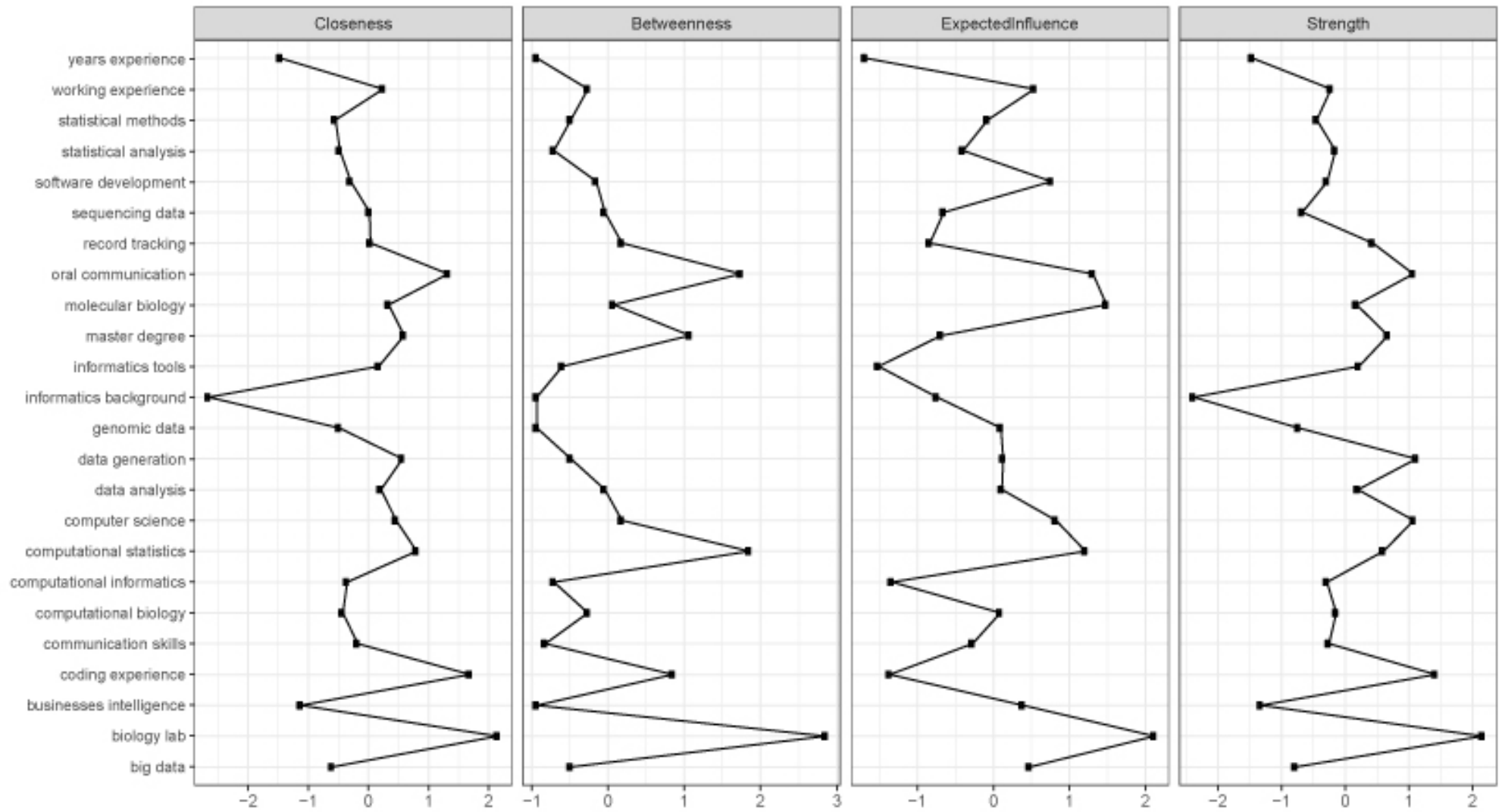
Source: own elaboration.

Fig. 60: Clustering plot with compared methods.



Source: own elaboration.

Fig. 61: Centrality measures plot.

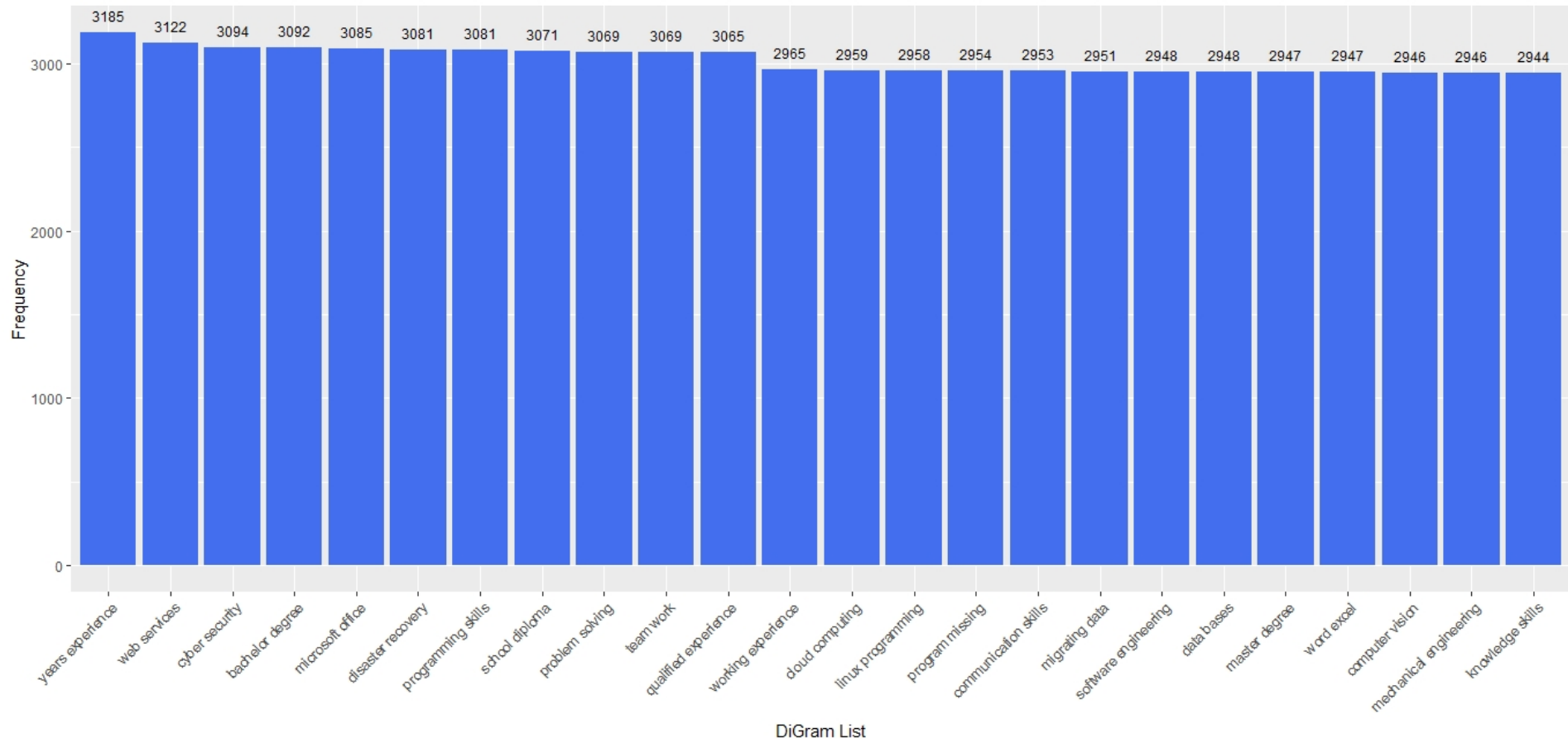


Source: own elaboration.

4.2.5 Software engineering & cloud computing (SECC)

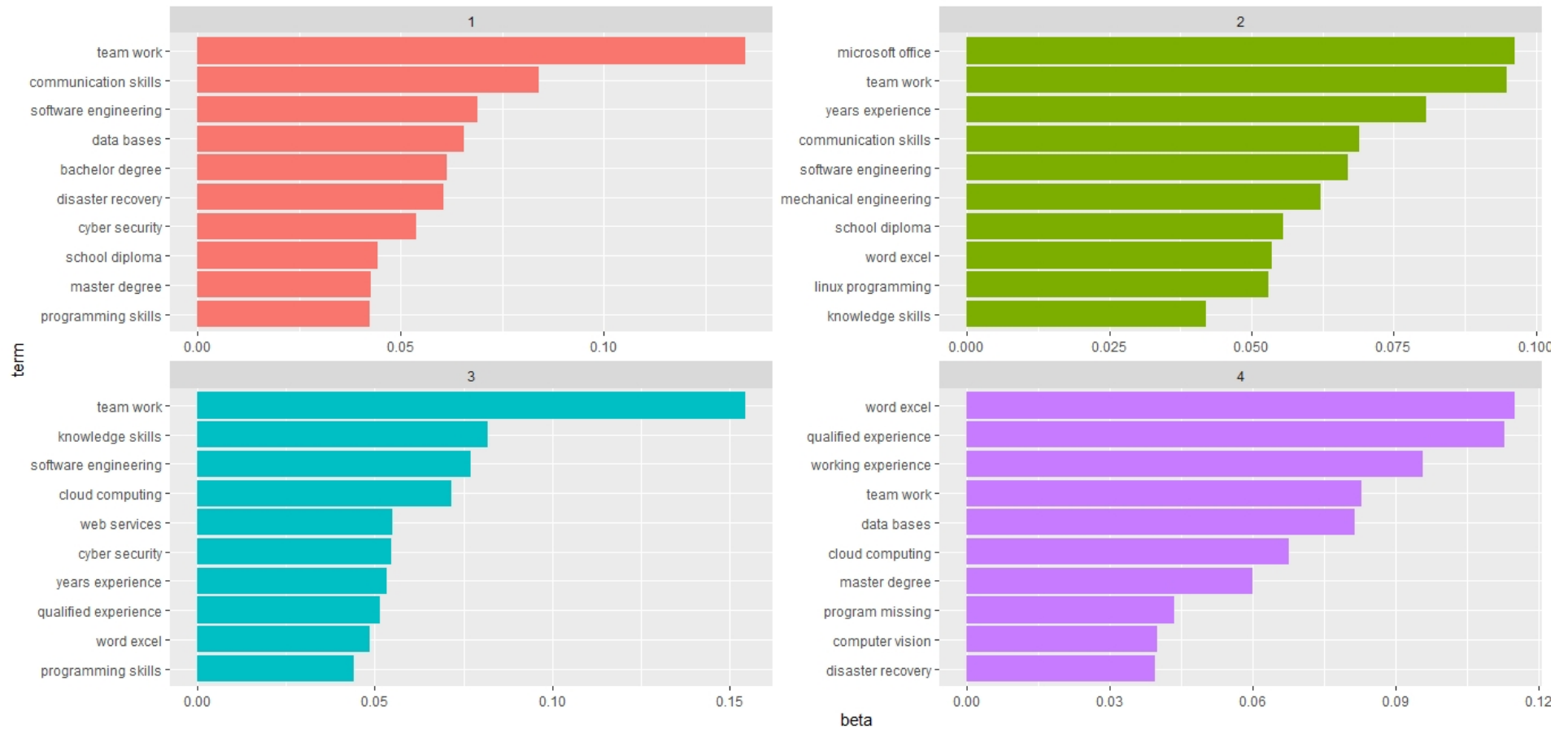
Results from the SECC industry have been obtained analyzing the subset corpus from the extracted ads regarding the sector. A tokenized Document-Term Matrix (DTM) has been built, and sparsity was removed till 61%. **Fig. 63** shows the bigrams from the corpus. The most frequent terminological combinations were years experience (3185), web services (3122), cyber security (3098), bachelor's degree (3092), and Microsoft Office (3085). Topic modeling is presented in **Fig. 64** with four thematic areas. **Fig. 65** highlights the main correlations through the skills set. **Fig. 66** detects greedy modularity in the skillset, dividing it in three groups, and the relative memberships are shown in **Fig. 67**. Application of spectral modularity is presented in **Fig. 68** and the relative memberships reported in **Fig. 69**. The employment of optimal modularity detection is shown in **Fig. 70** and their memberships highlighted in **Fig. 71**. Modularity indicators were compared to define the most proper method to give sense to the analysis. Having $\xi_G > \xi_O > \xi_S$, the dendrogram in **Fig. 72** is built with greedy modularity, and partial correlations will be used for the weighted network in **Fig. 73**. Thus, a clustering plot with Zhang, Onnela, and Barrat methods is reported in **Fig. 74**. Centrality measures are exposed in **Fig. 75**. The most between skills in the set were program missing (71.1%), years experience (22.9%), school diploma (20.5%), computer vision (19.3%), and Linux programming (13.04%). The closest skills were program missing (18%), Microsoft Office (16.5%), computer vision (16.2%), web services (35.3%), and written communication (34.3%). MCMC with MAP method is shown in **Fig. 76** to forecast and simulate a possible job interview for the SECC industry.

Fig. 63: Bigrams of the SECC skillset.



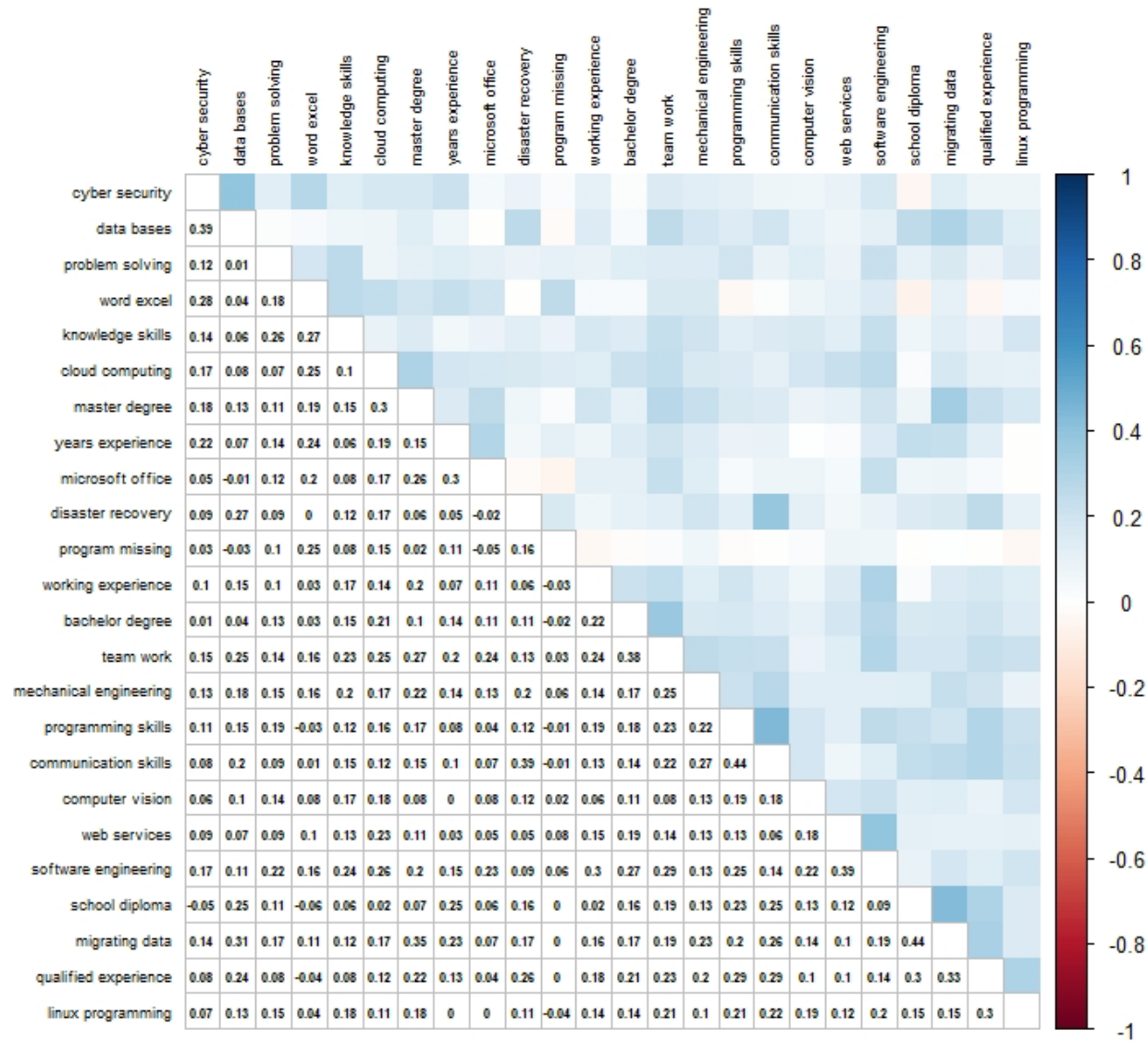
Source: own elaboration

Fig. 64: Topic modeling of the SECC skillset.



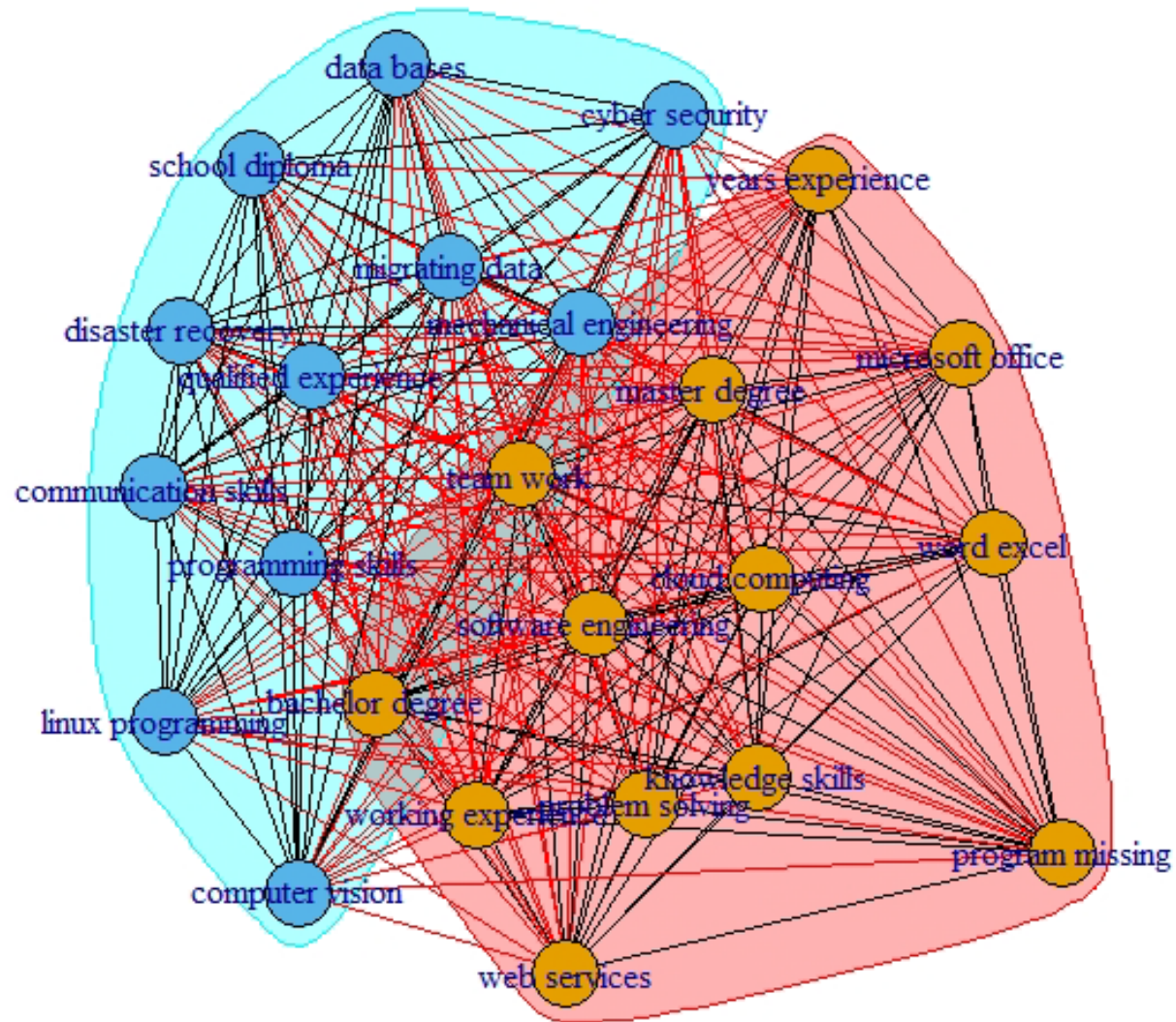
Source: own elaboration.

Fig. 65: Corrplot of the SECC skillset.



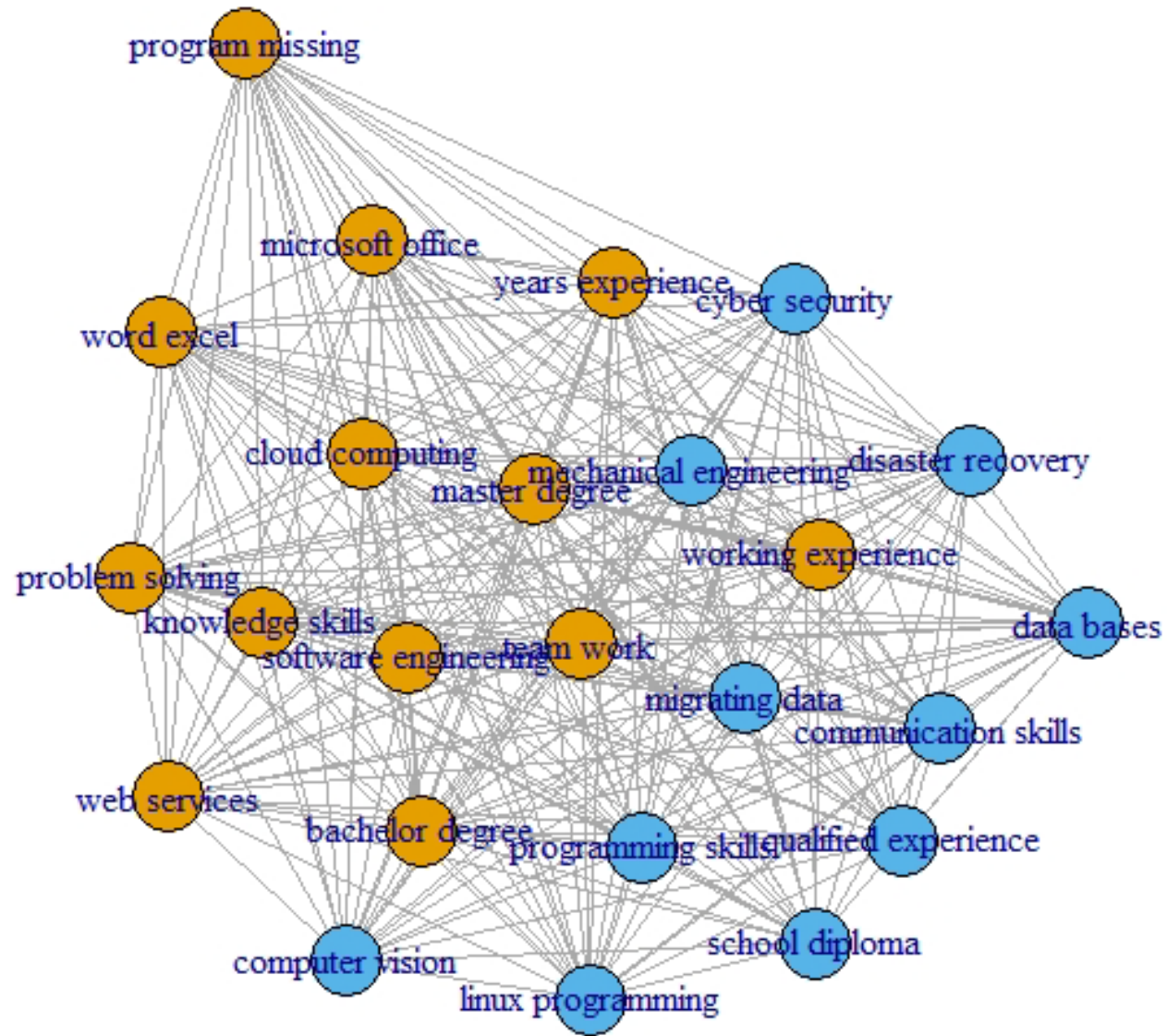
Source: own elaboration.

Fig. 66: Skills network with greedy modularity community detection.



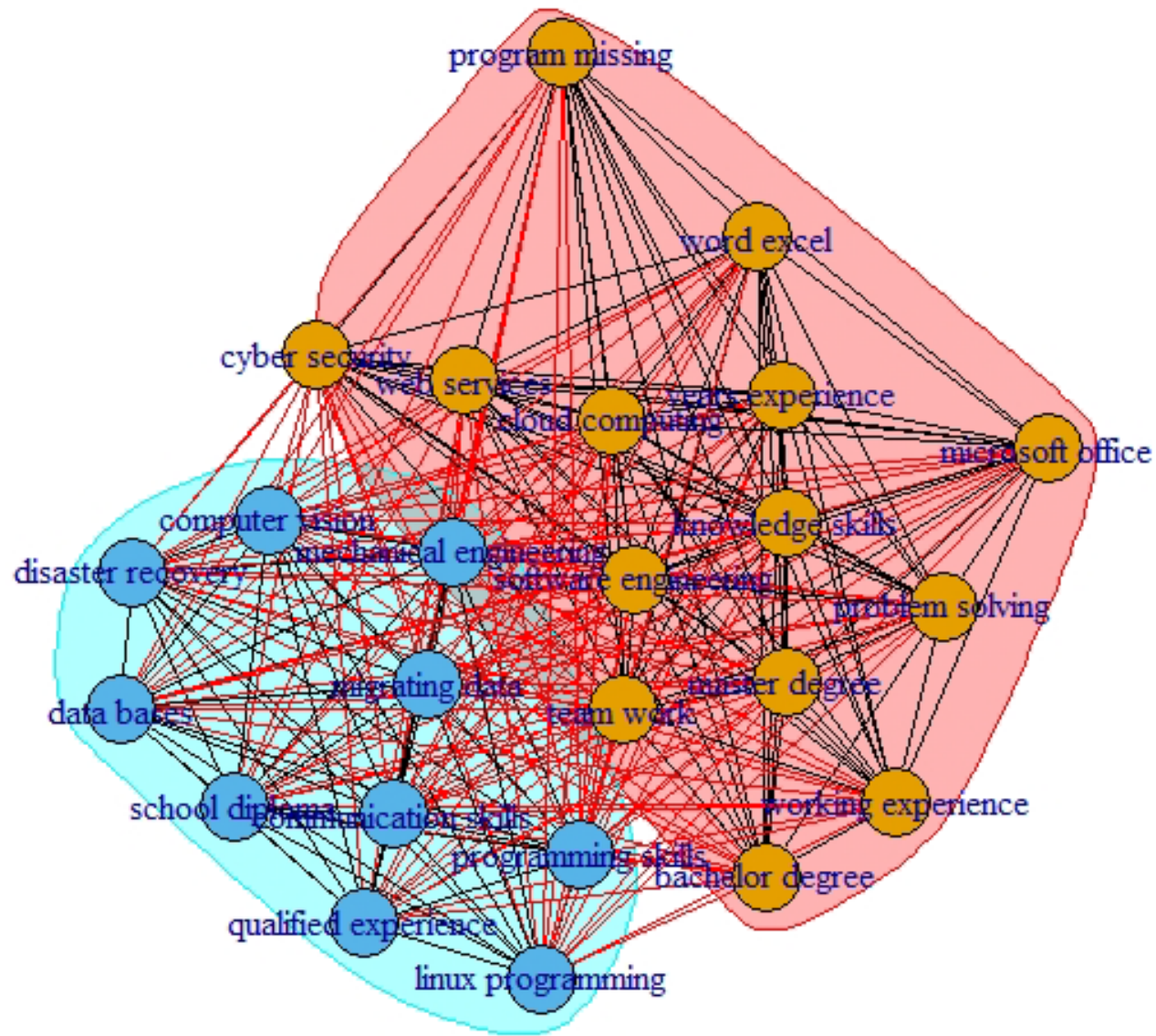
Source: own elaboration.

Fig. 67: Skills network community membership according to greedy modularity.



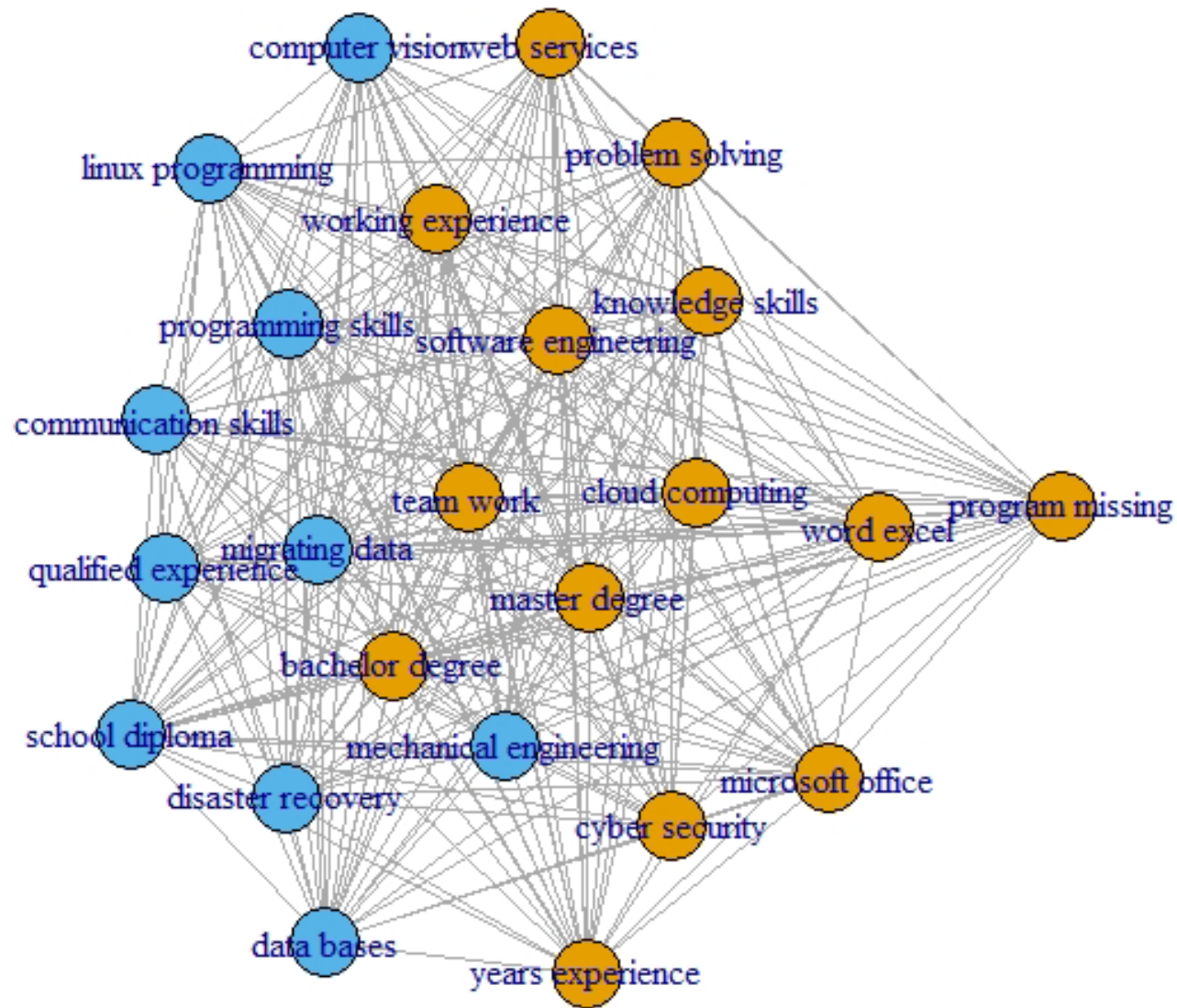
Source: own elaboration.

Fig. 68: Skills network with spectral modularity community detection.



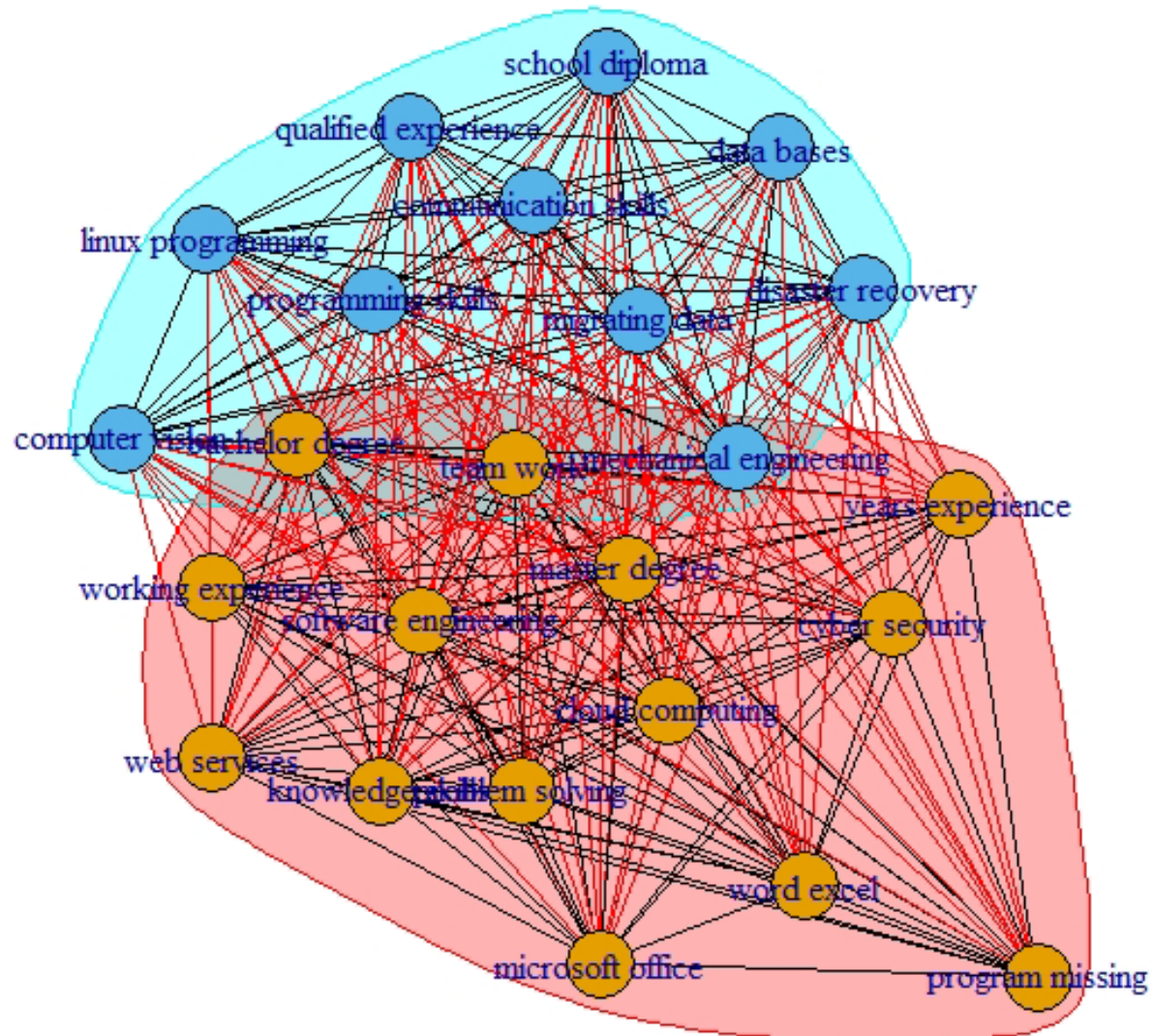
Source: own elaboration.

Fig. 69: Skills network community membership according to spectral modularity.



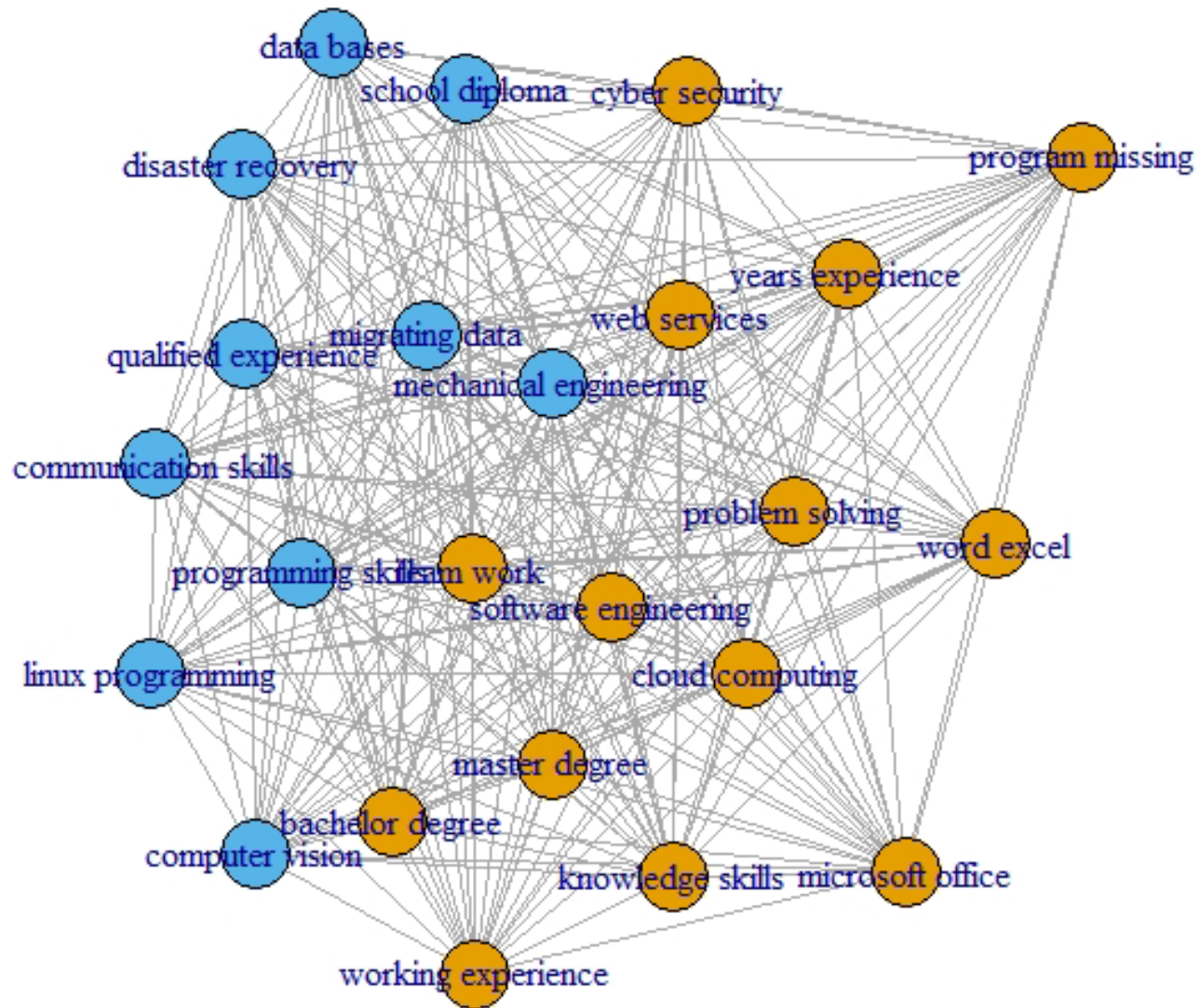
Source: own elaboration.

Fig. 70: Skills network with optimal community detection.



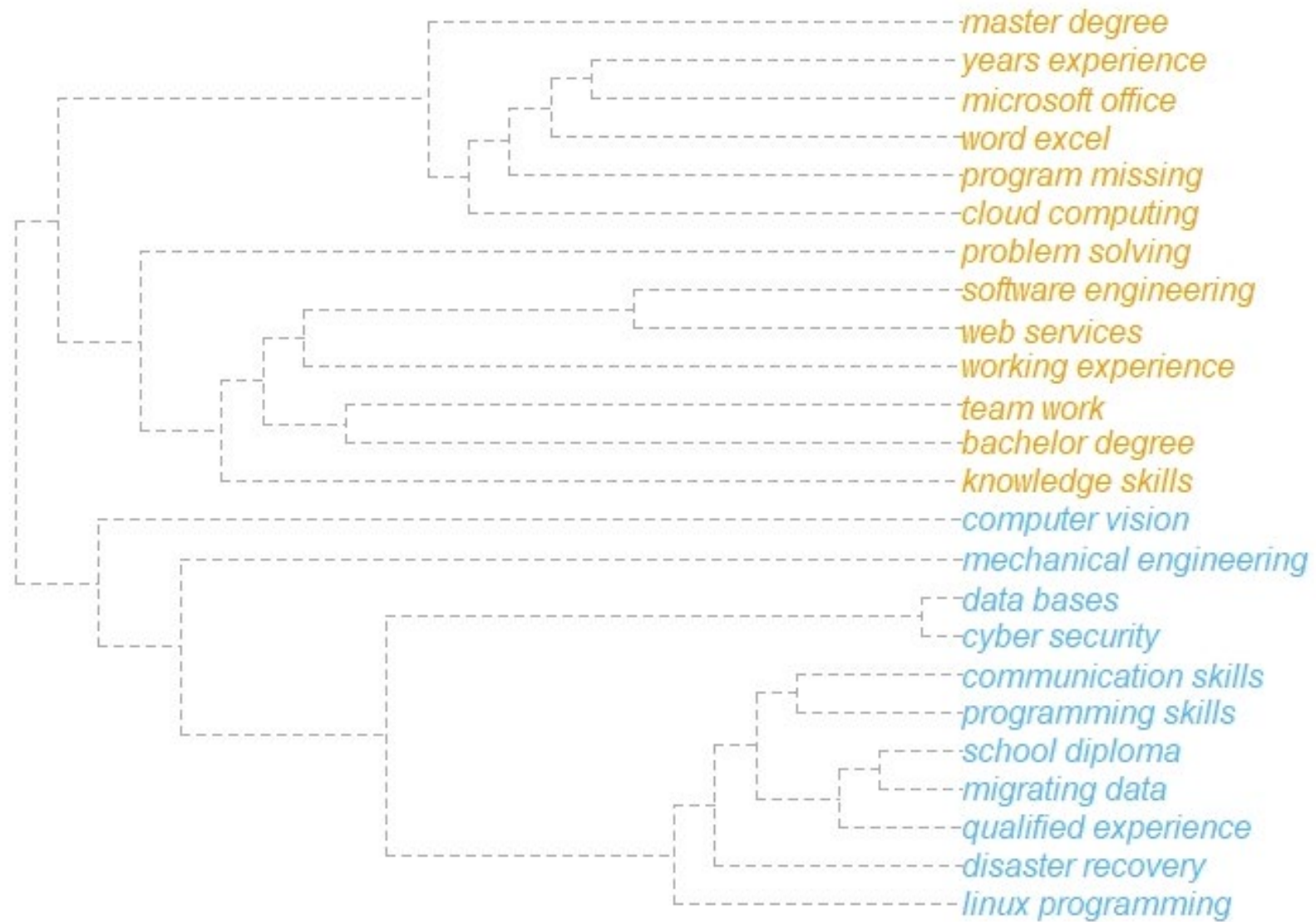
Source: own elaboration.

Fig. 71: Skills network community membership according to optimal modularity.



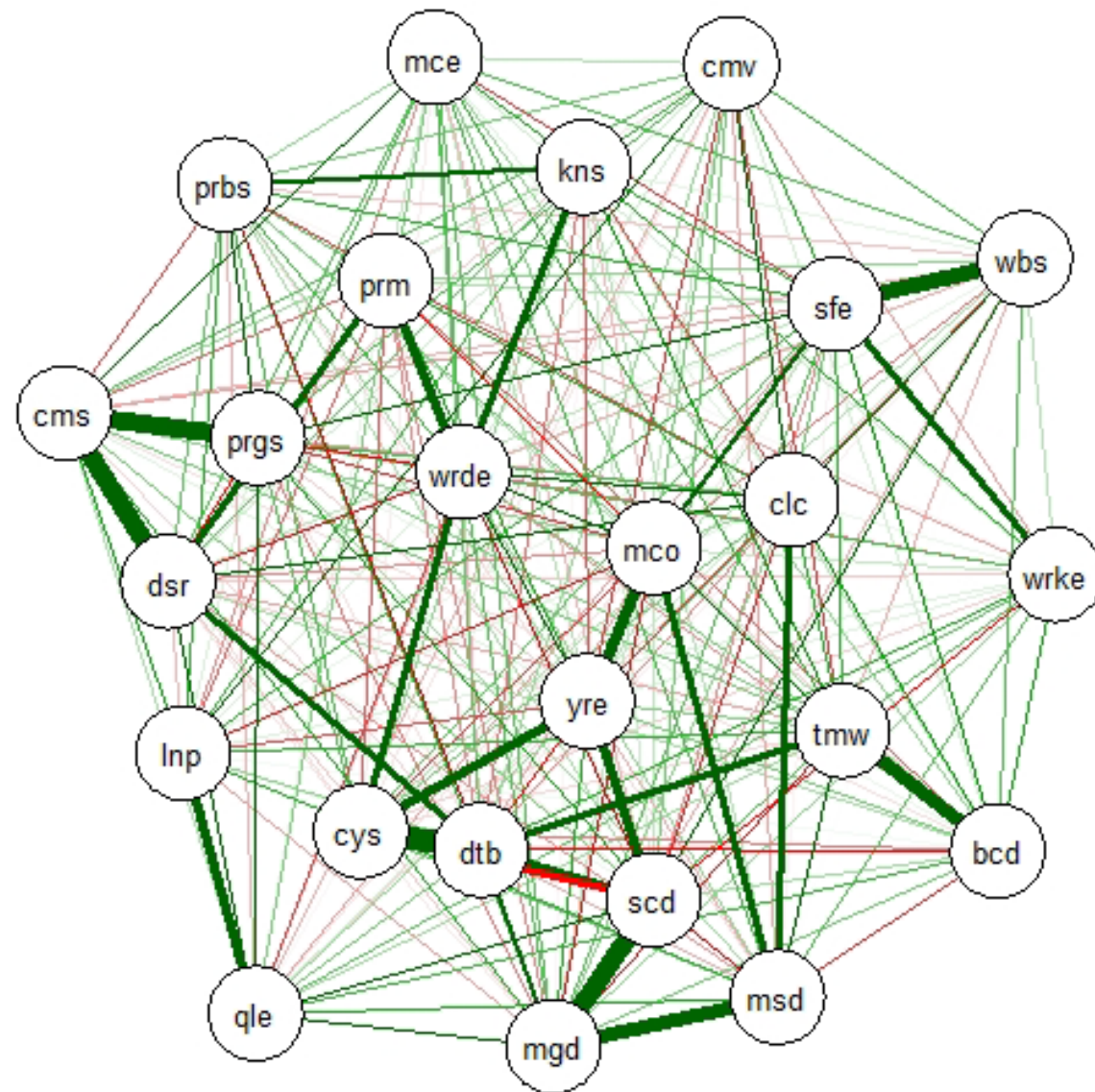
Source: own elaboration.

Fig. 72: Dendrogram with greedy modularity.



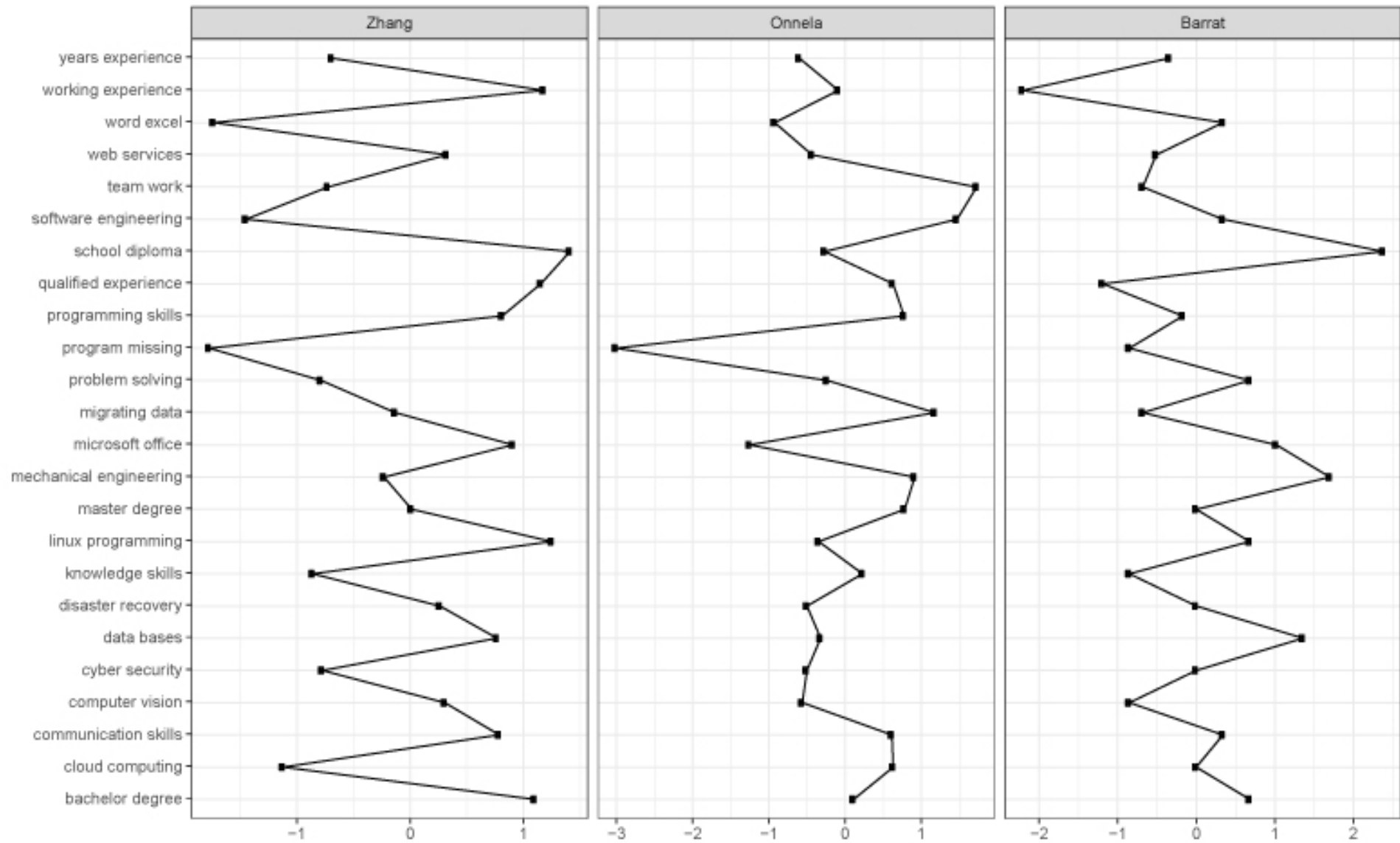
Source: own elaboration

Fig. 73: Weighted skills network via partial correlations clustering.



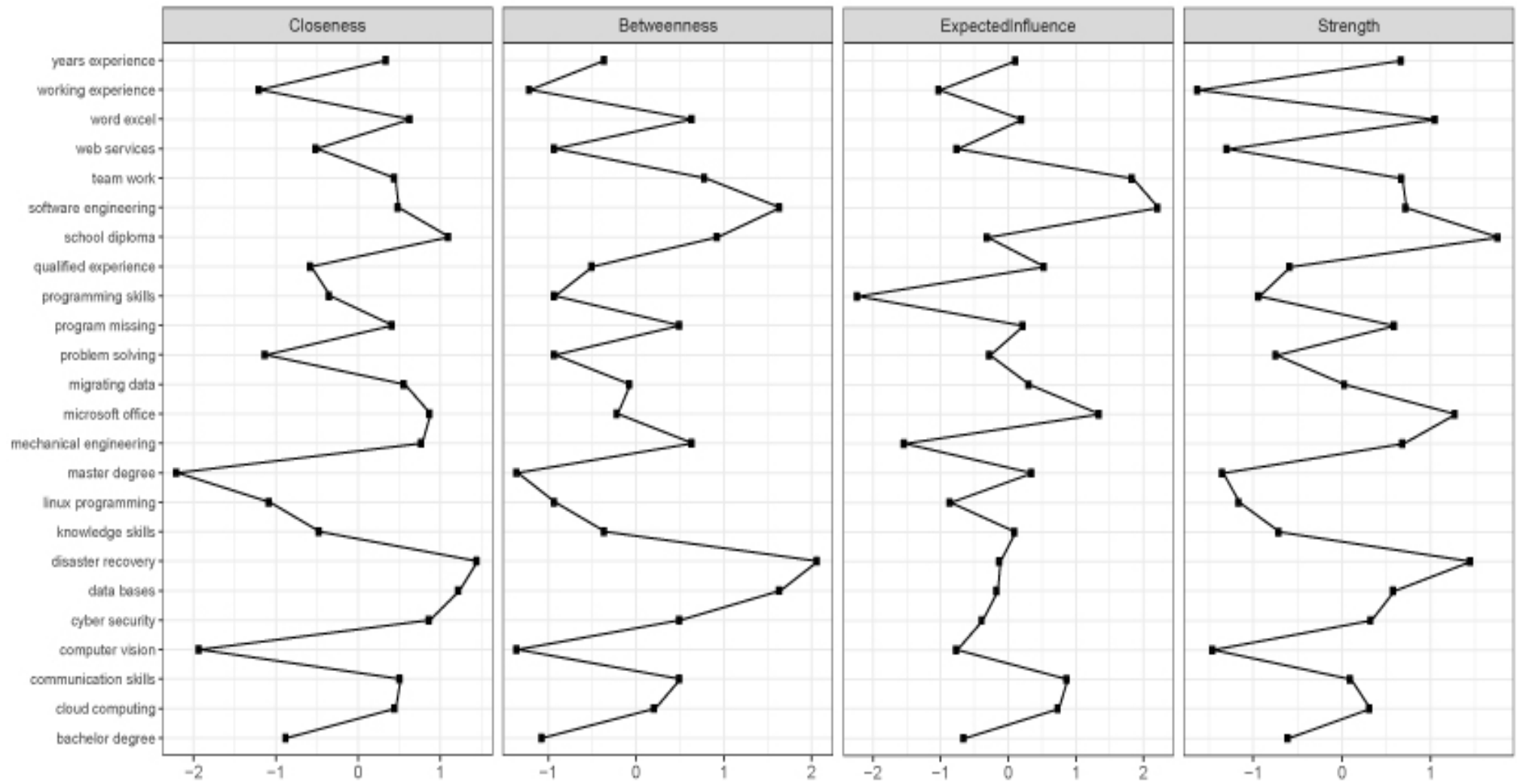
Source: own elaboration.

Fig. 74: Clustering plot with compared methods.



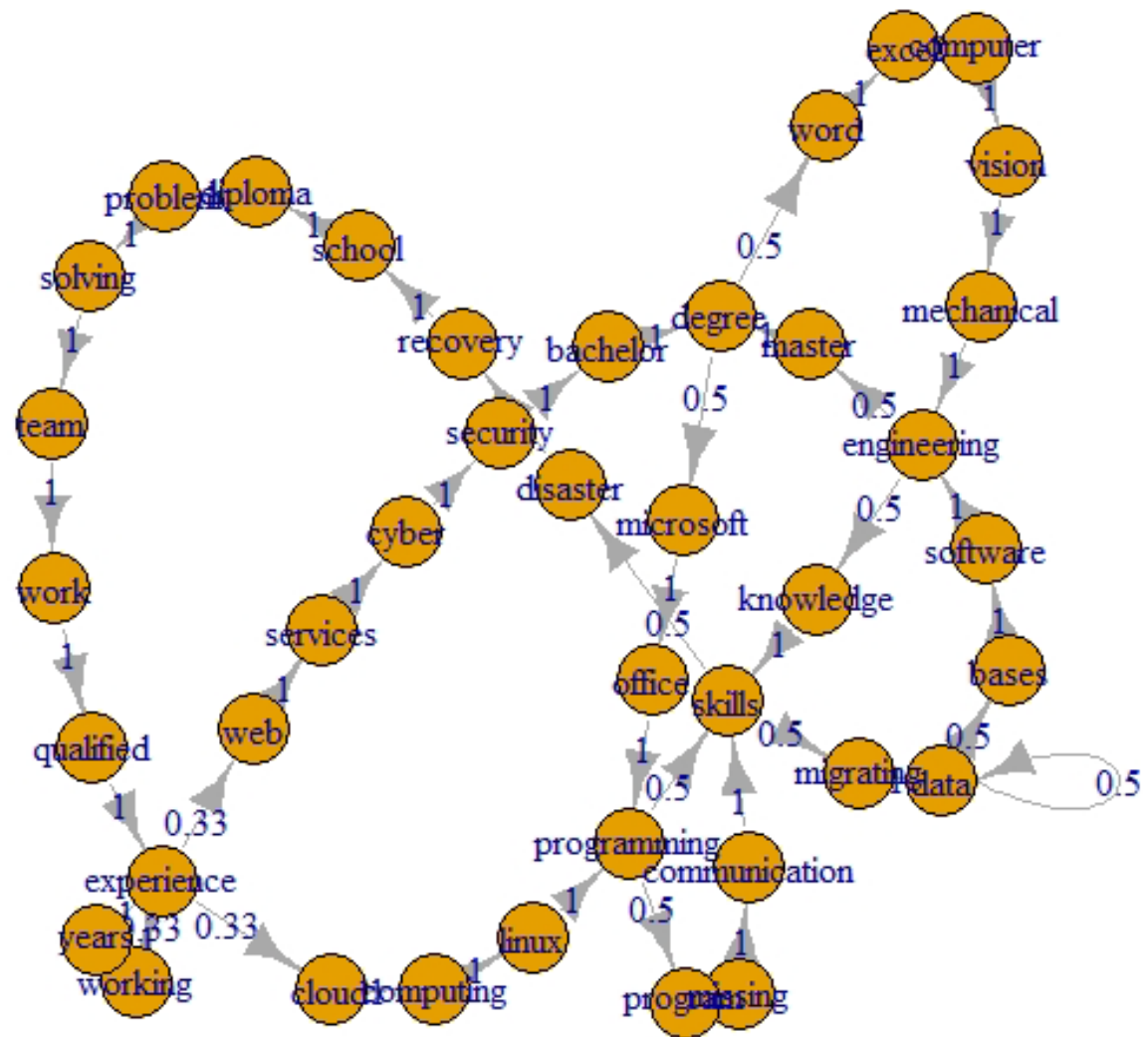
Source: own elaboration.

Fig. 75: Centrality measures plot.



Source: own elaboration.

Fig. 76: Monte Carlo Markov Chain with MAP method.

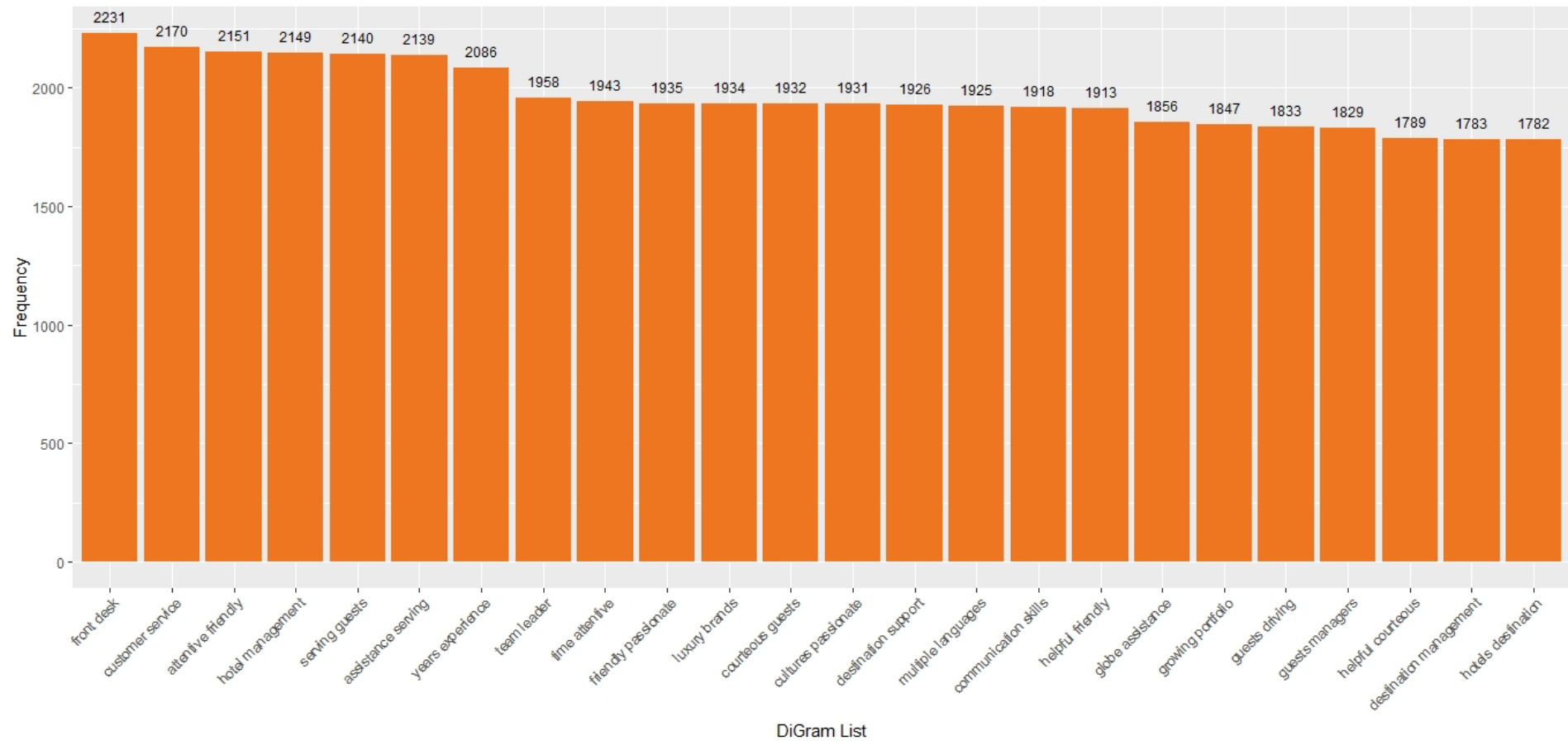


Source: own elaboration.

4.2.6 Tourism management

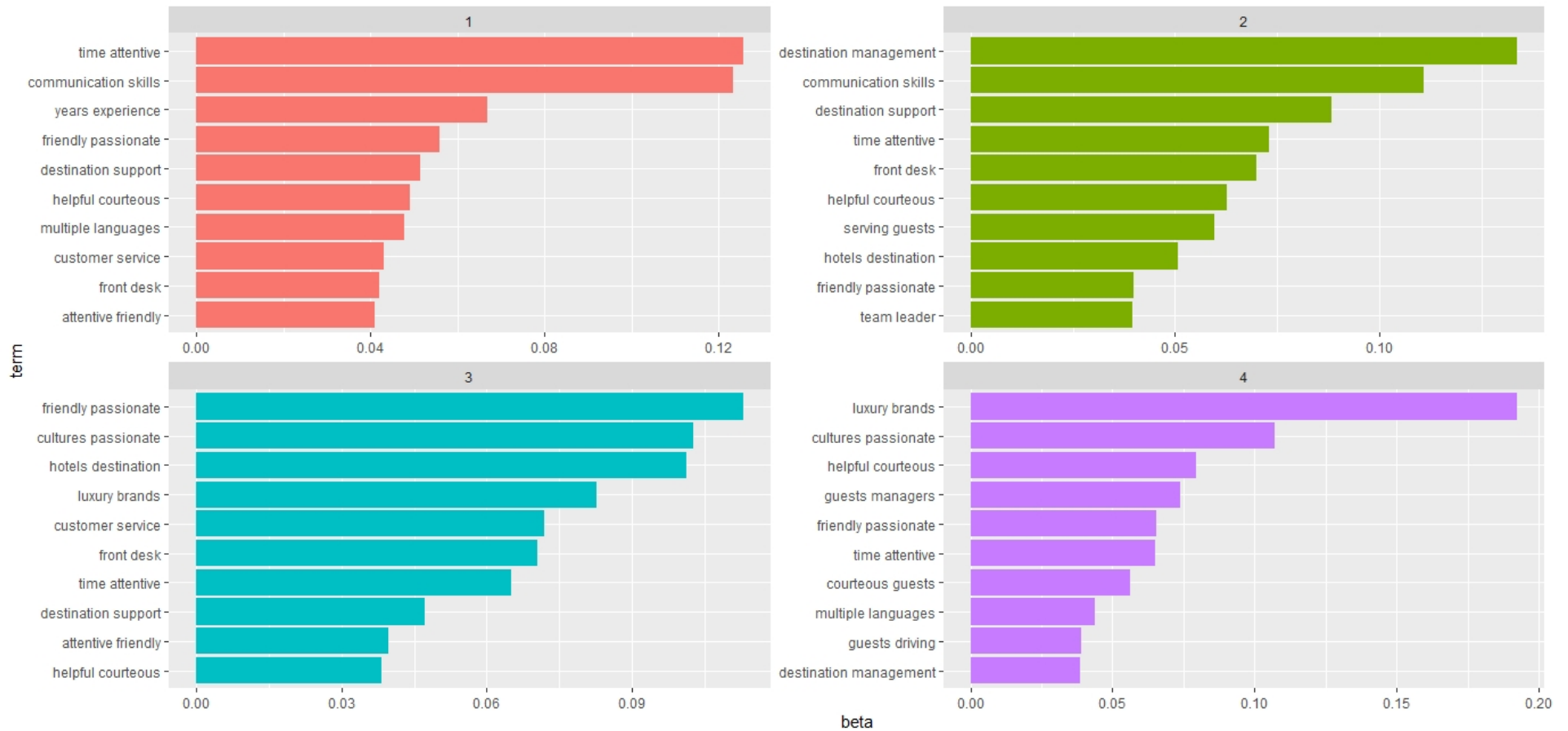
Results from the Tourism Management industry have been obtained analyzing the subset corpus from the extracted ads regarding the sector. A tokenized Document-Term Matrix (DTM) has been built, and sparsity was removed till 72%. **Fig. 77** shows the bigrams from the corpus. The most frequent terminological combinations were front desk (2231), customer service (2170), attentive friendly (2151), hotel management (2149), and serving guests (2140). Topic modeling is presented in **Fig. 78** with four thematic areas. **Fig. 79** highlights the main correlations through the skills set. **Fig. 80** detects greedy modularity in the skillset, dividing it in three groups, and the relative memberships are shown in **Fig. 81**. Application of spectral modularity is presented in **Fig. 82**, and the relative memberships are reported in **Fig. 83**. The employment of optimal modularity detection is shown in **Fig. 84** and their memberships highlighted in **Fig. 85**. Modularity indicators were compared to define the most proper method to give sense to the analysis. Having $\xi_G > \xi_O > \xi_S$, the dendrogram in **Fig. 86** is built with greedy modularity, and partial correlations will be used for the weighted network in **Fig. 87**. Thus, a clustering plot with Zhang, Onnela, and Barrat methods is reported in **Fig. 88**. Centrality measures are exposed in **Fig. 89**. The most between skills in the set were serving guests (68.3%), time attentive (57.1%), years experience (38.6%), customer service (22%), and multiple language (19.9%). The closest skills were years experience (81%), customer service (73.7%), serving guests (63.4%), attentive friendly (56.2%), and time attentive (48%). MCMC with MAP method is shown in **Fig. 90** to forecast and simulate a possible job interview for the Tourism Management industry.

Fig. 77: Bigrams of the Tourism Management skillset.



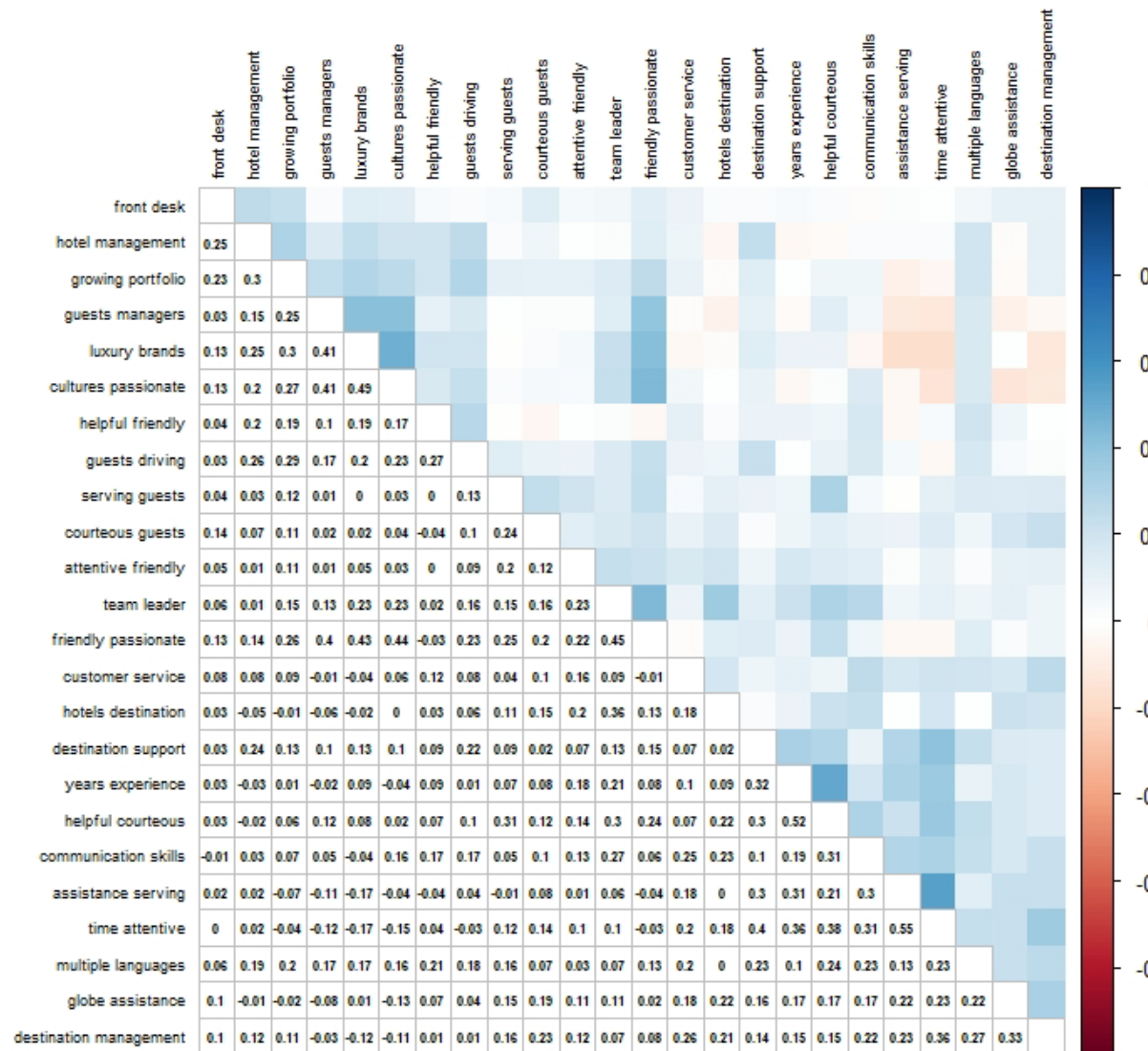
Source: own elaboration

Fig. 78: Topic modeling of the Tourism Management skillset.



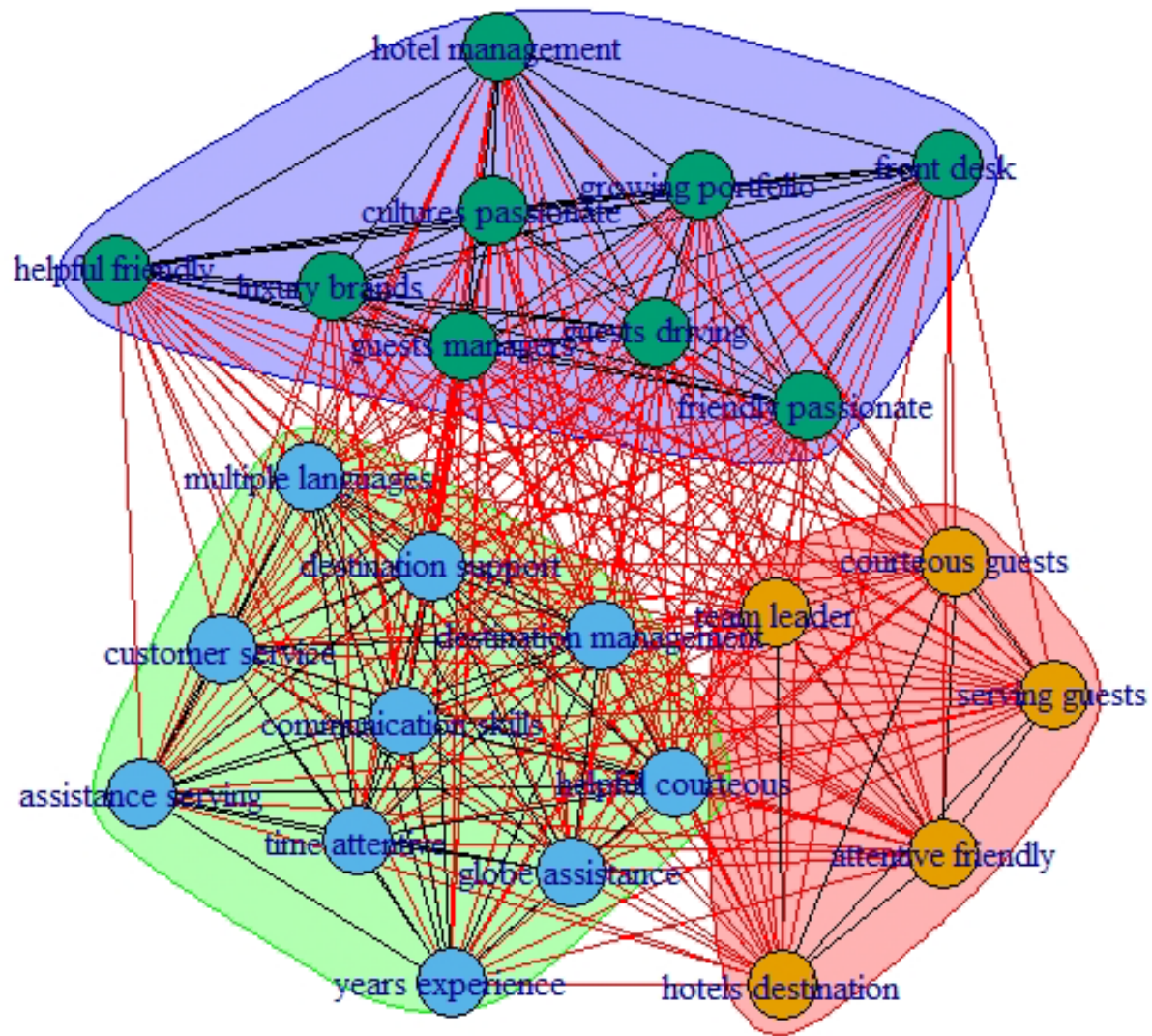
Source: own elaboration.

Fig. 79: Corrplot of the Tourism Management skillset.



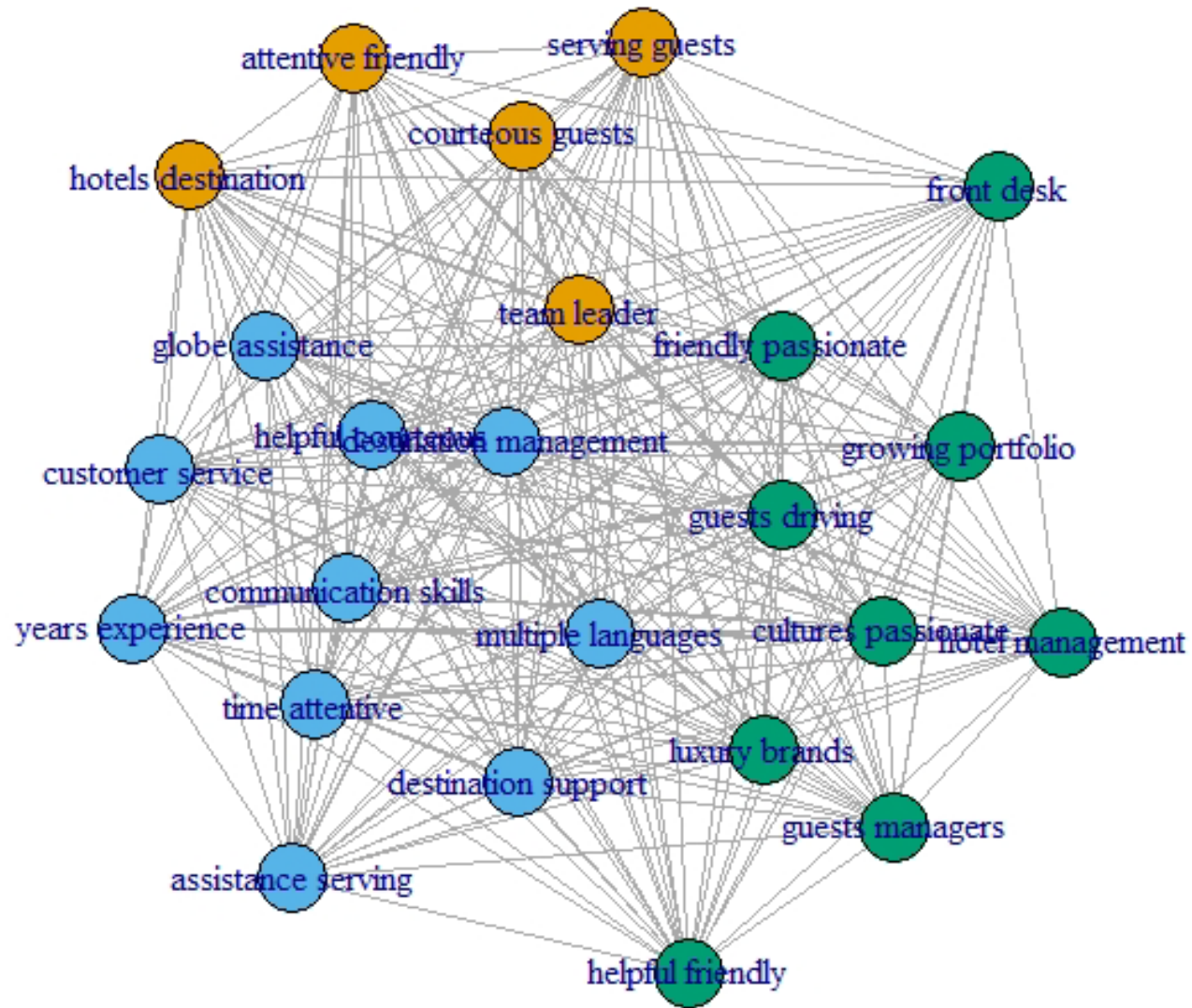
Source: own elaboration.

Fig. 80: Skills network with greedy modularity community detection.



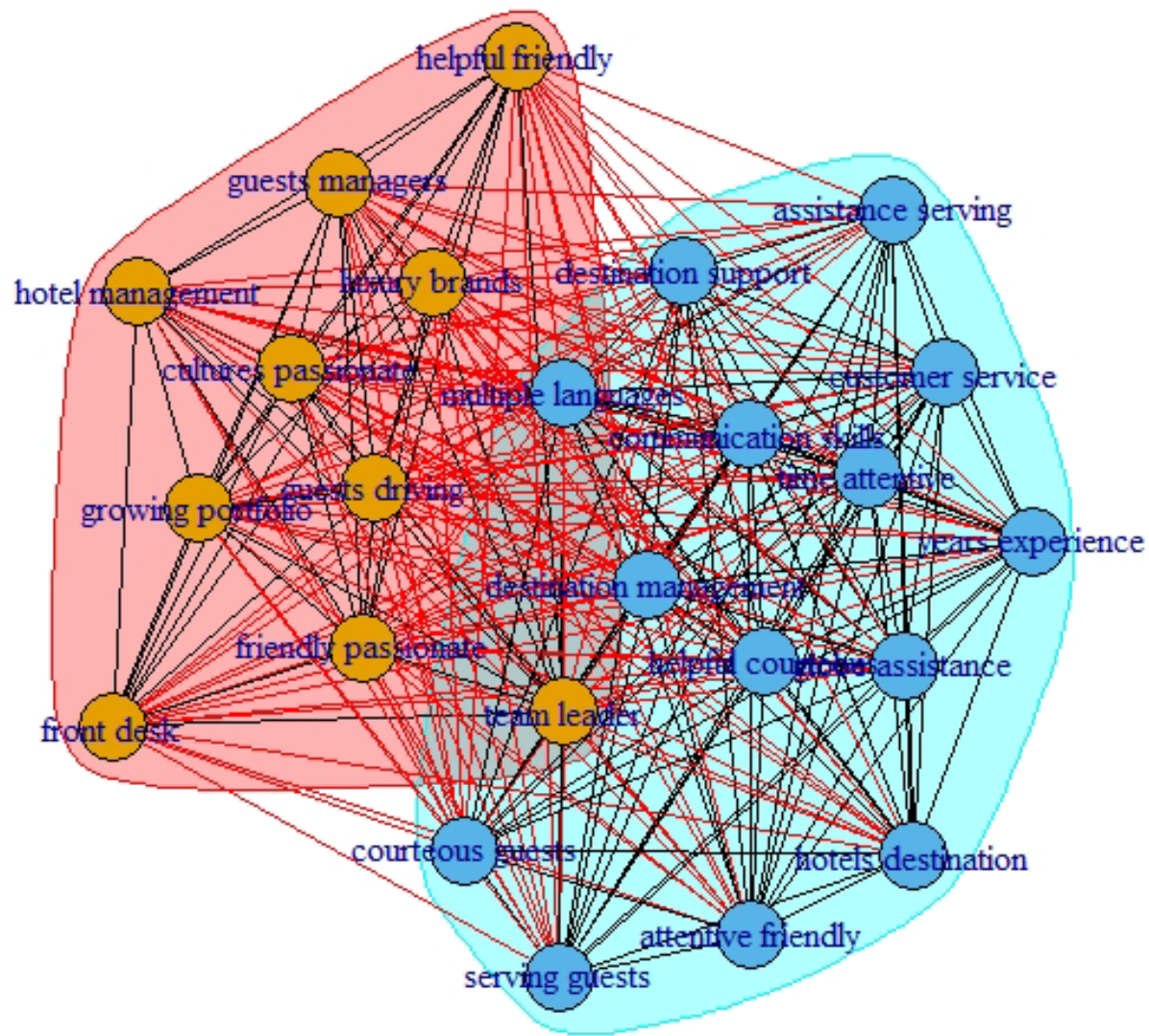
Source: own elaboration.

Fig. 81: Skills network community membership according to greedy modularity.



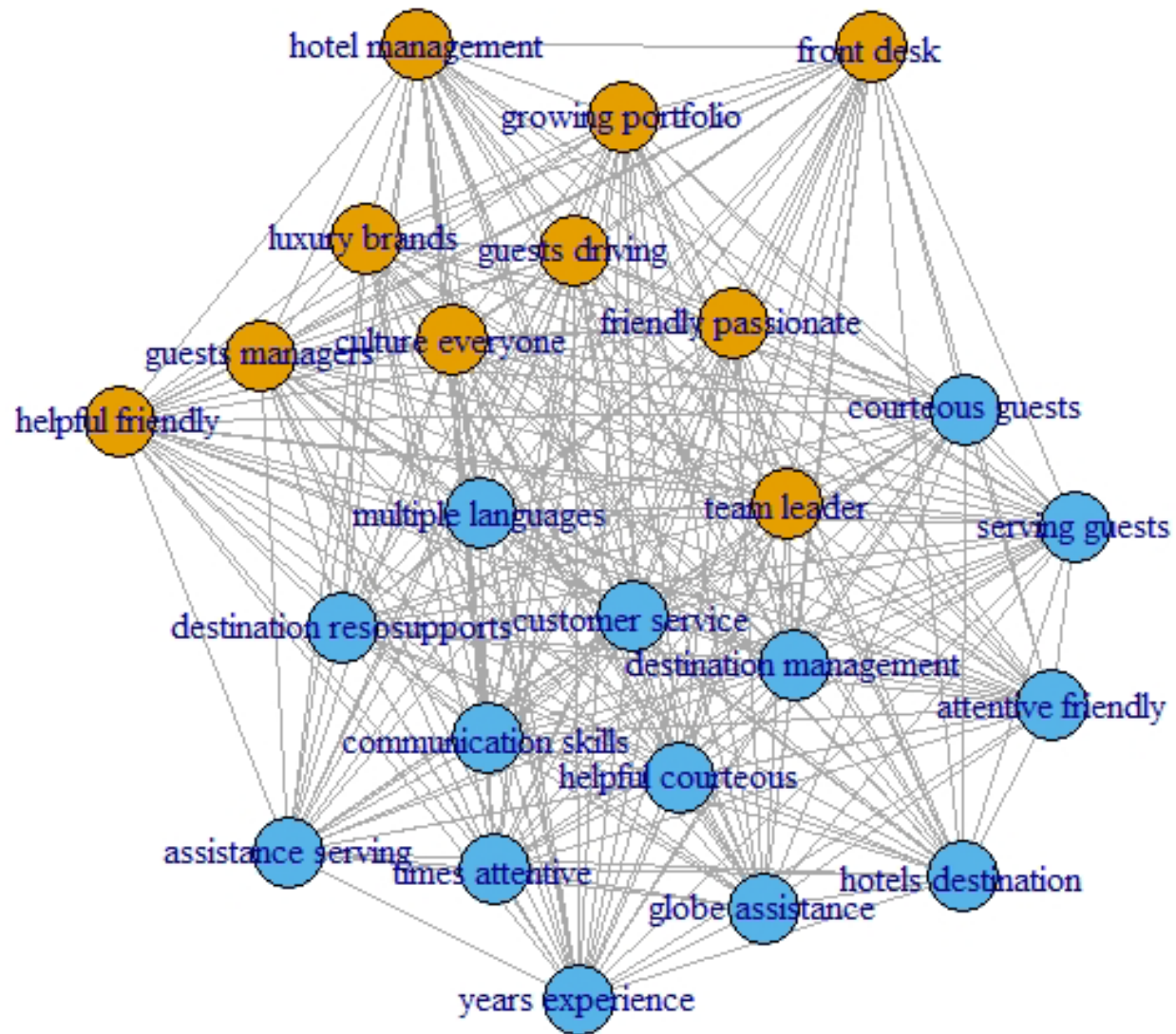
Source: own elaboration.

Fig. 82: Skills network with spectral modularity community detection.



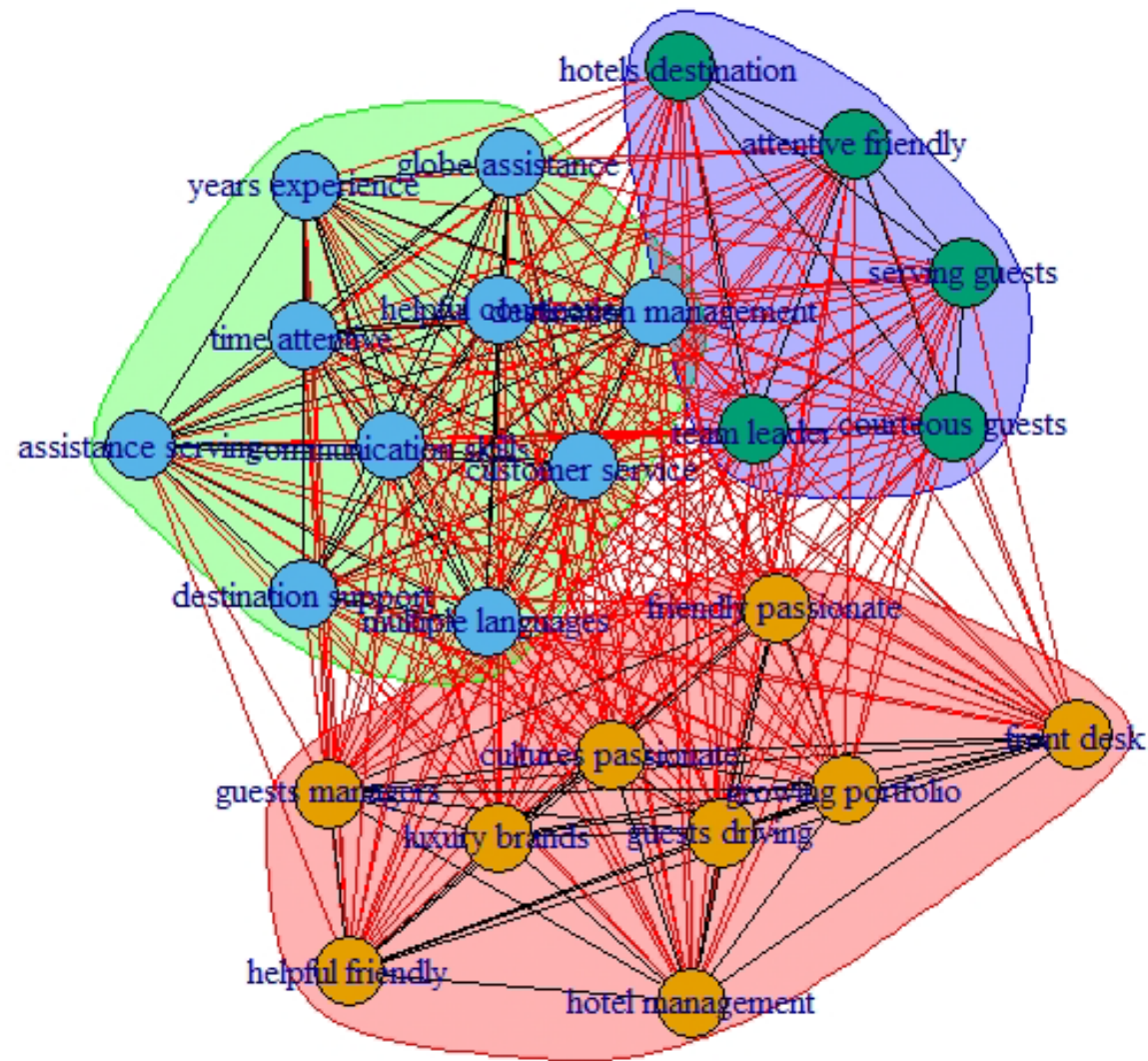
Source: own elaboration.

Fig. 83: Skills network community membership according to spectral modularity.



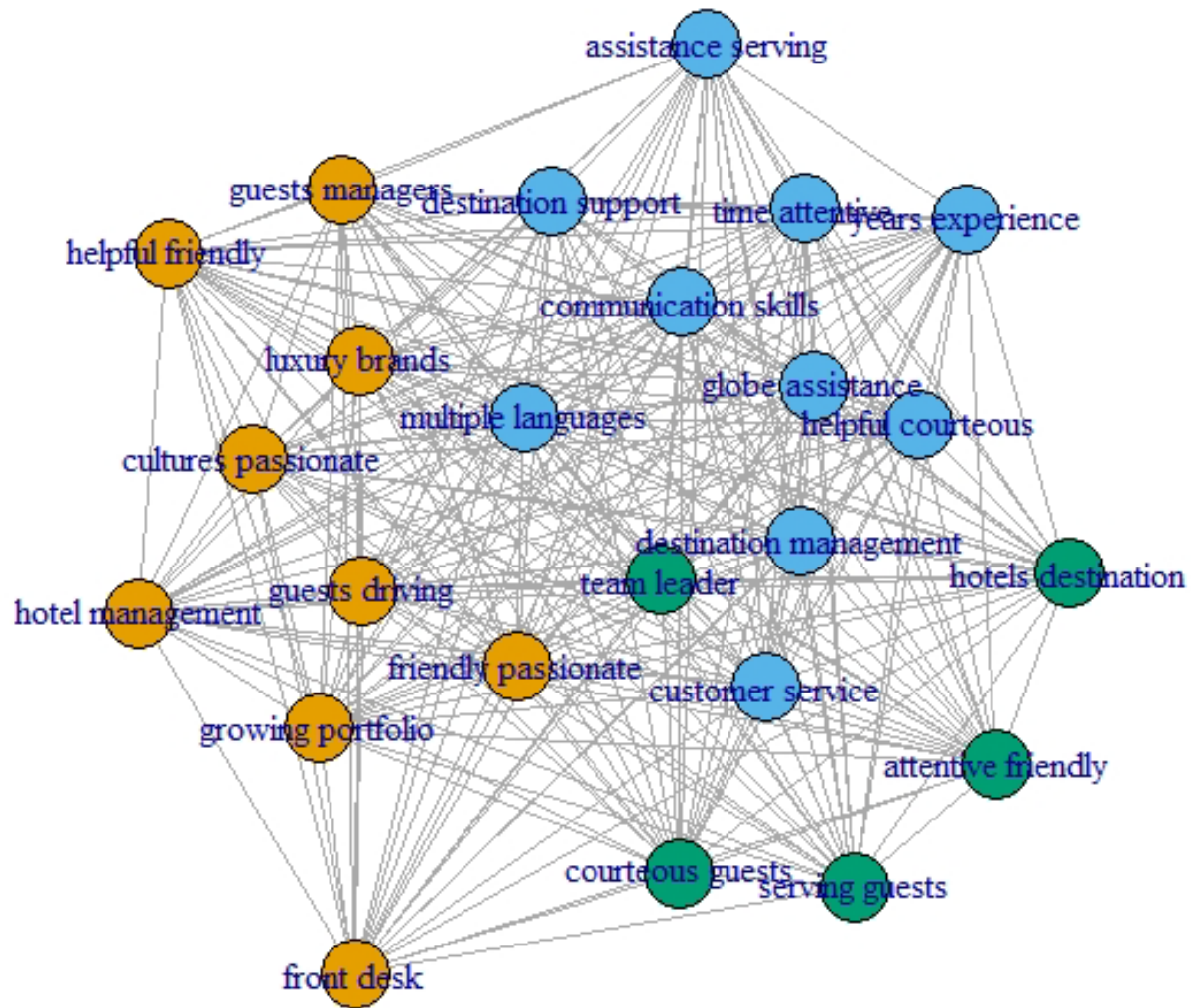
Source: own elaboration.

Fig. 84: Skills network with optimal community detection.



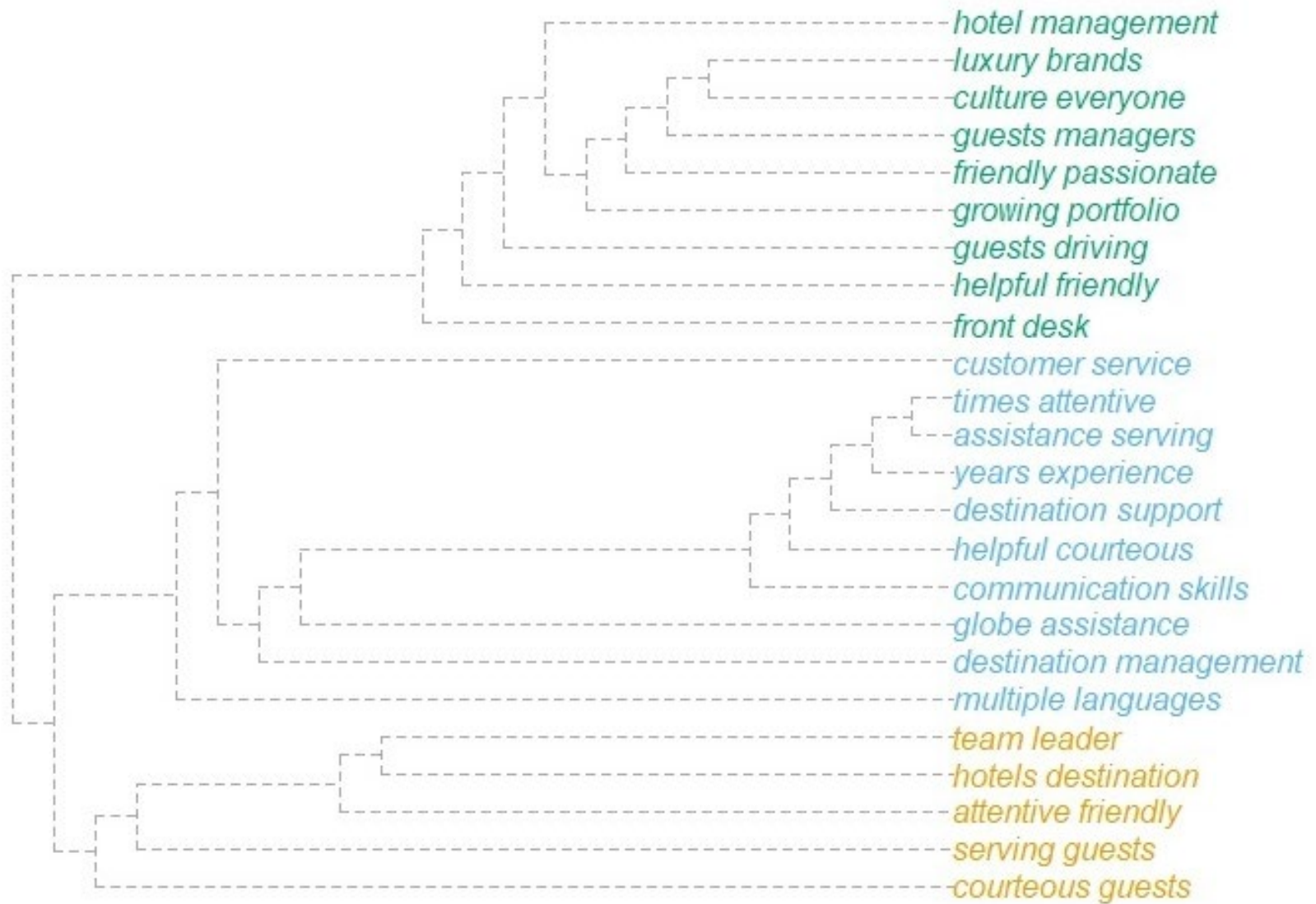
Source: own elaboration.

Fig. 85: Skills network community membership according to optimal modularity.



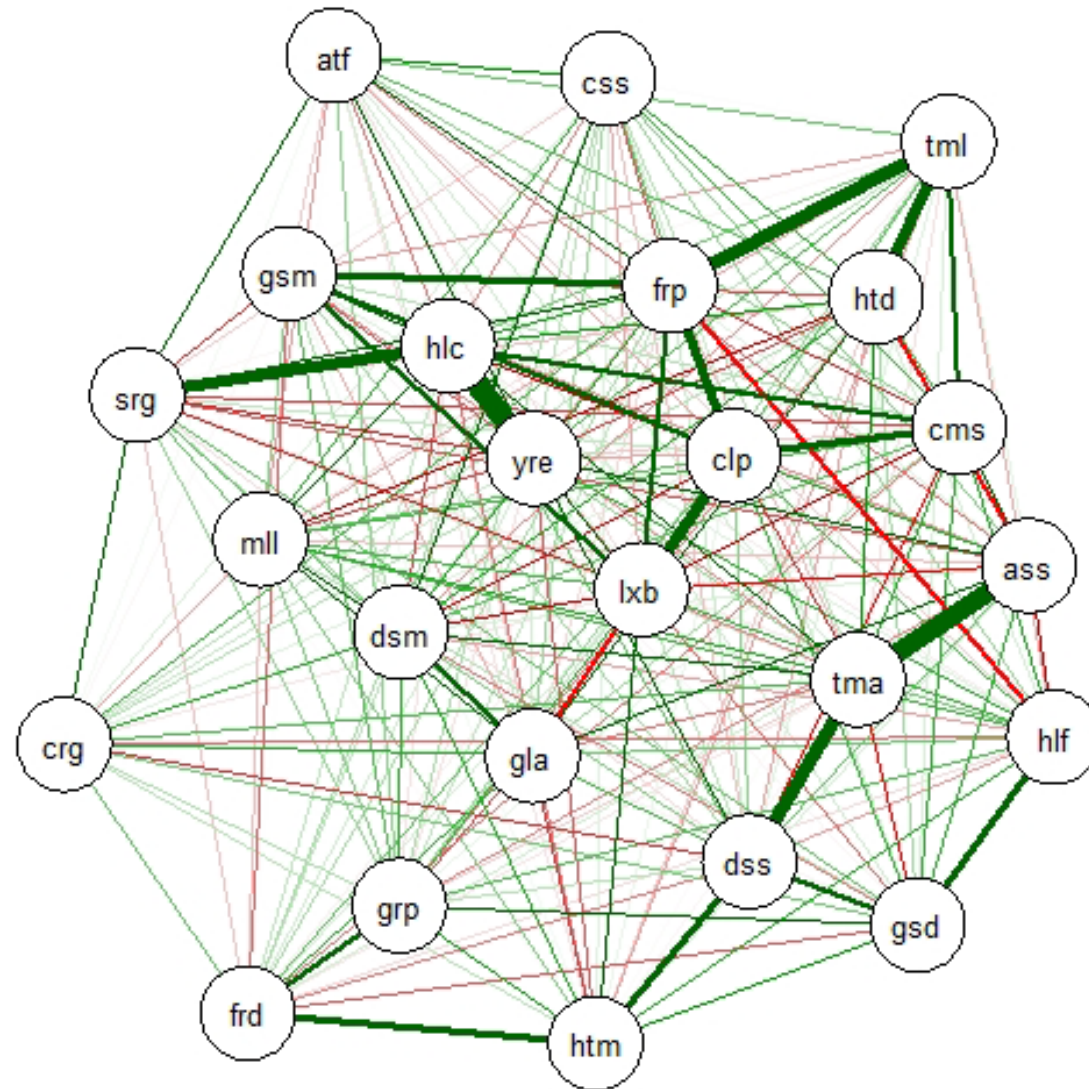
Source: own elaboration.

Fig. 86: Dendrogram with greedy modularity.



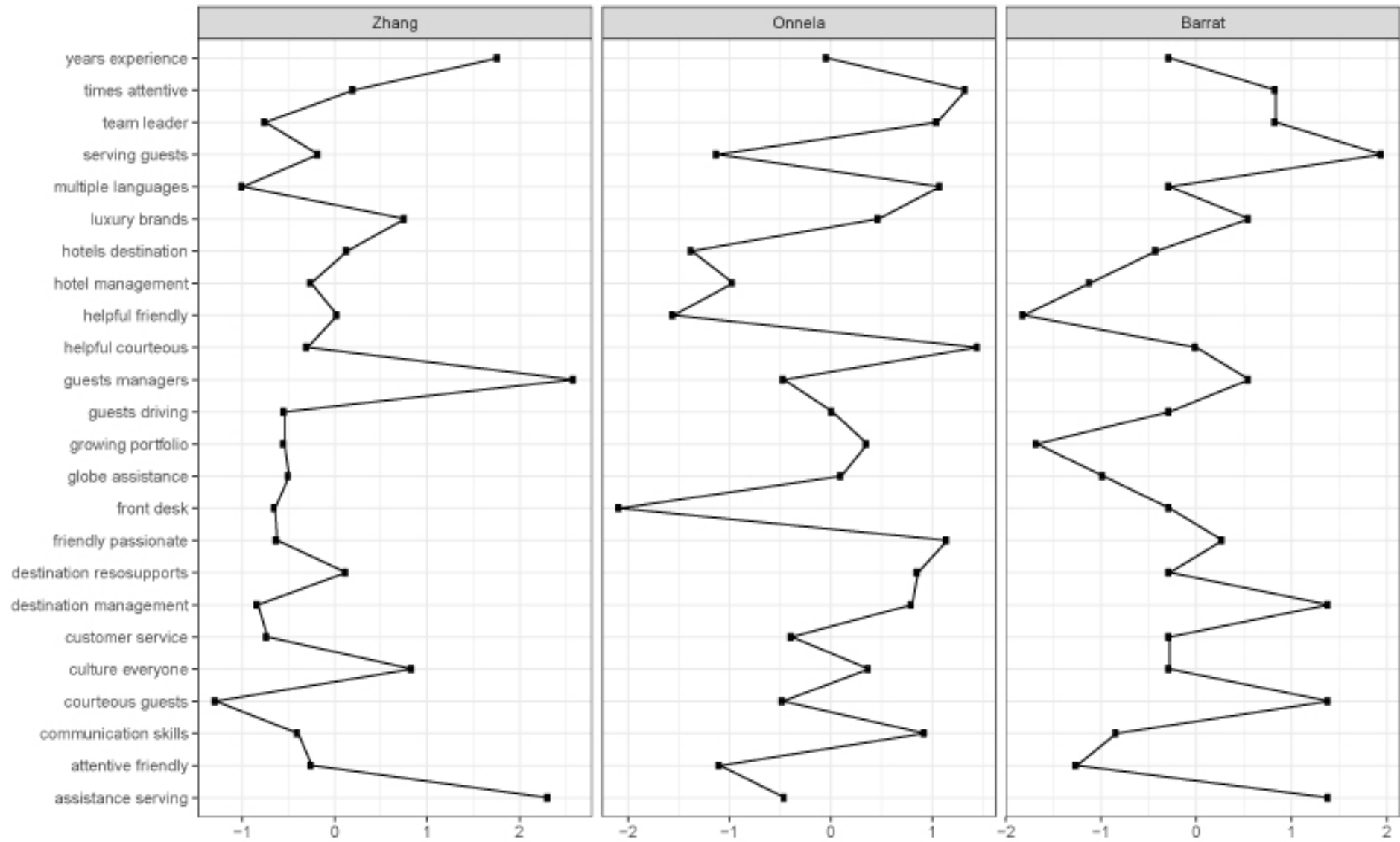
Source: own elaboration

Fig. 87: Weighted skills network via partial correlations clustering.



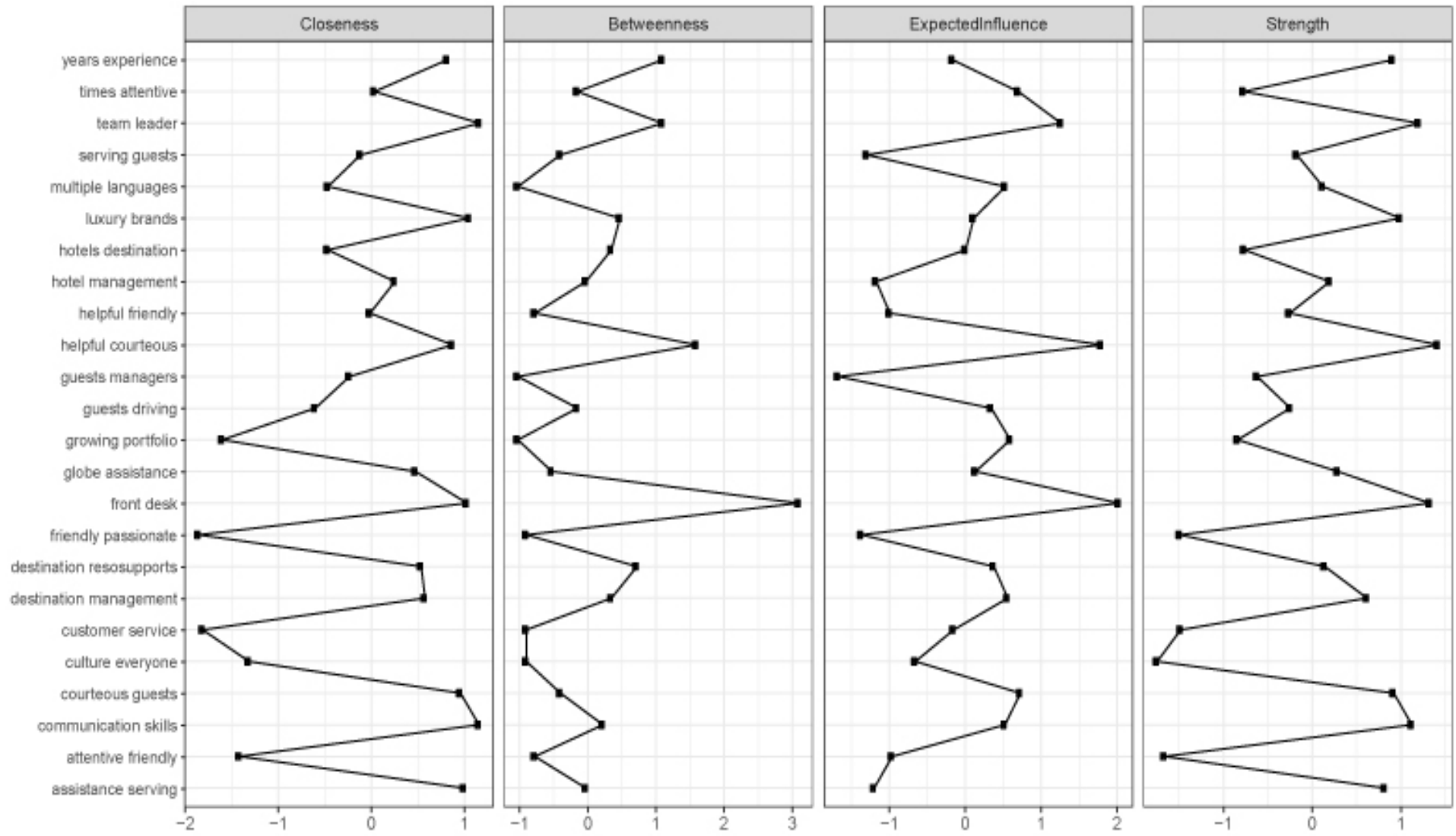
Source: own elaboration.

Fig. 88: Clustering plot with compared methods.



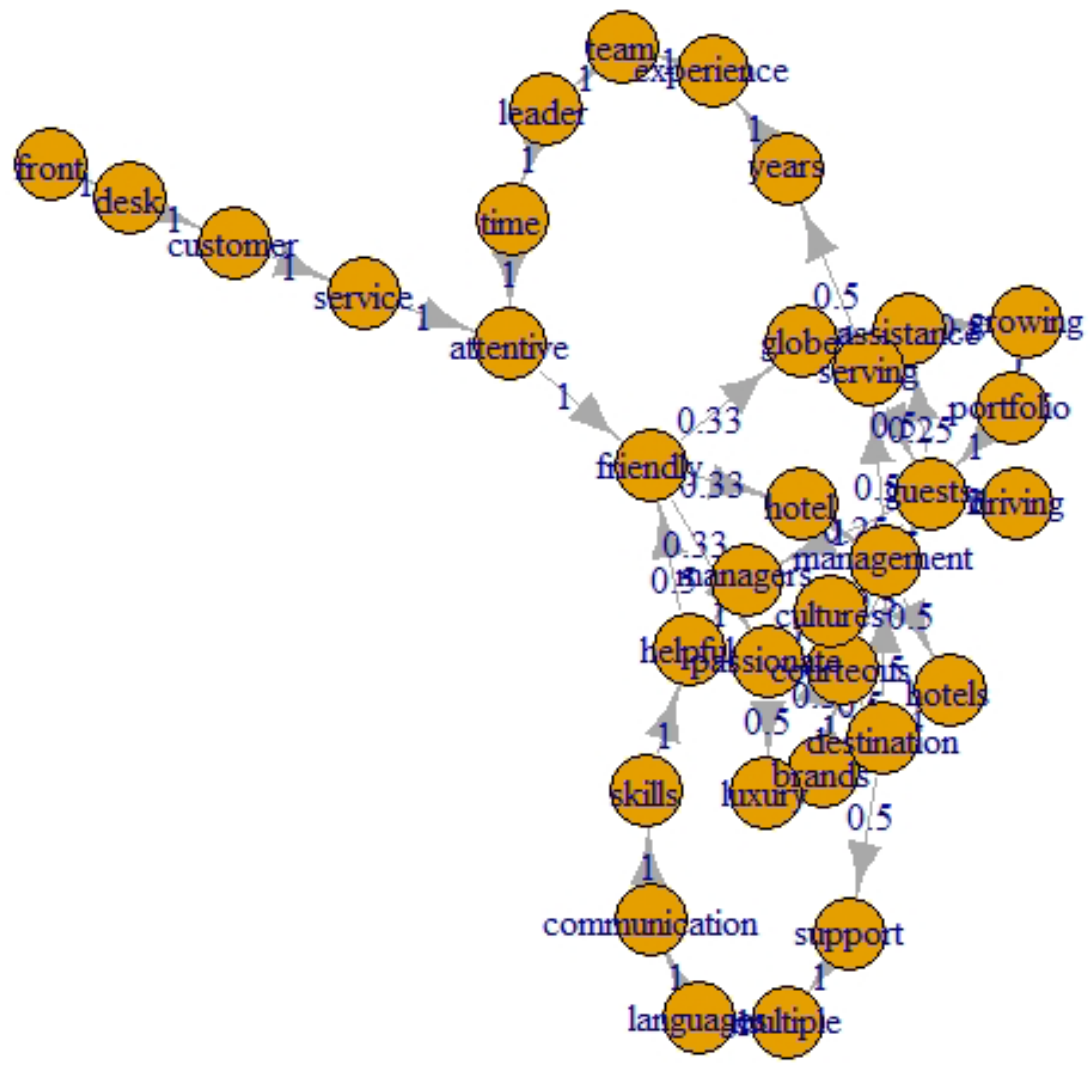
Source: own elaboration.

Fig. 89: centrality measures plot.



Source: own elaboration.

Fig. 90: Monte Carlo Markov Chain with MAP method.

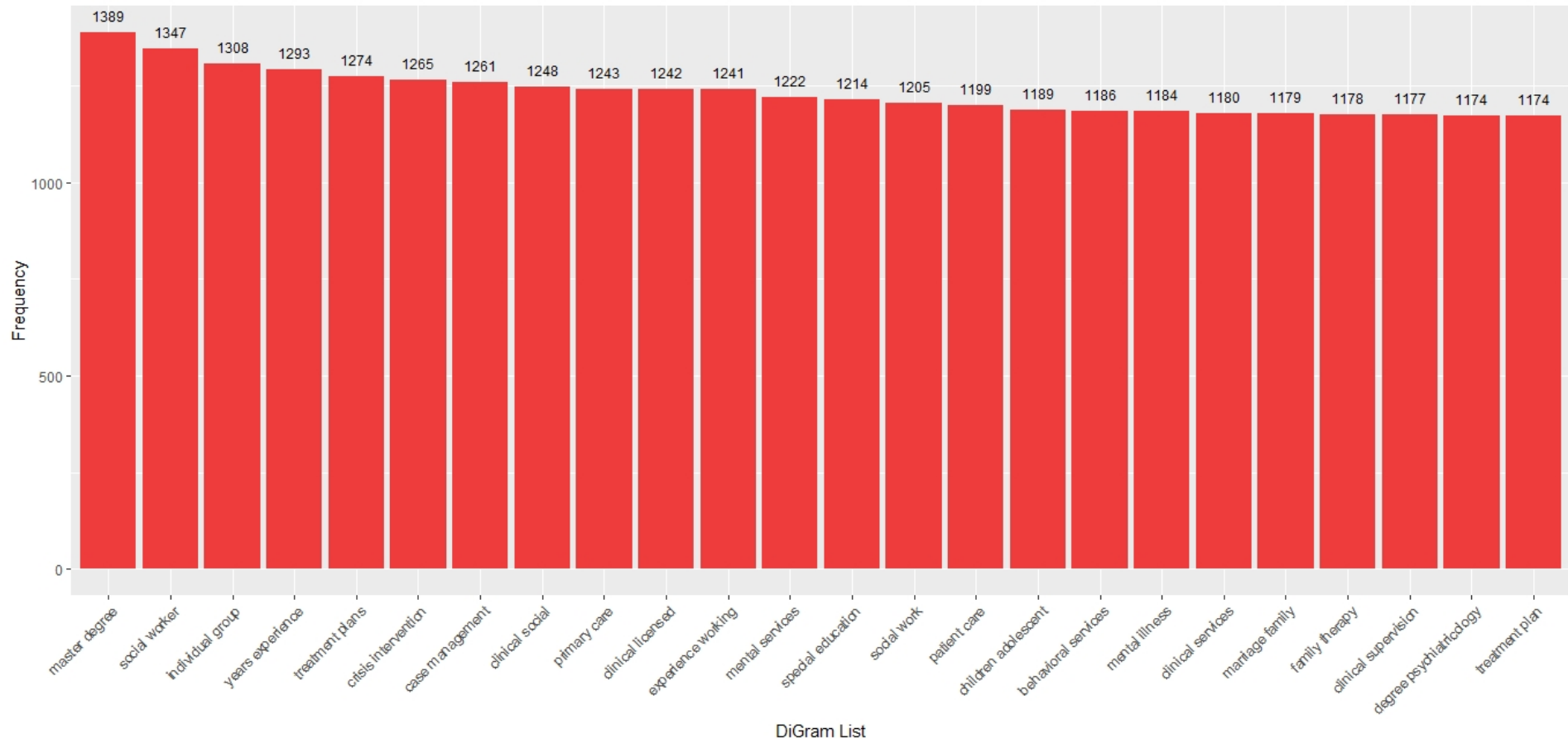


Source: own elaboration.

4.2.7 Psychology

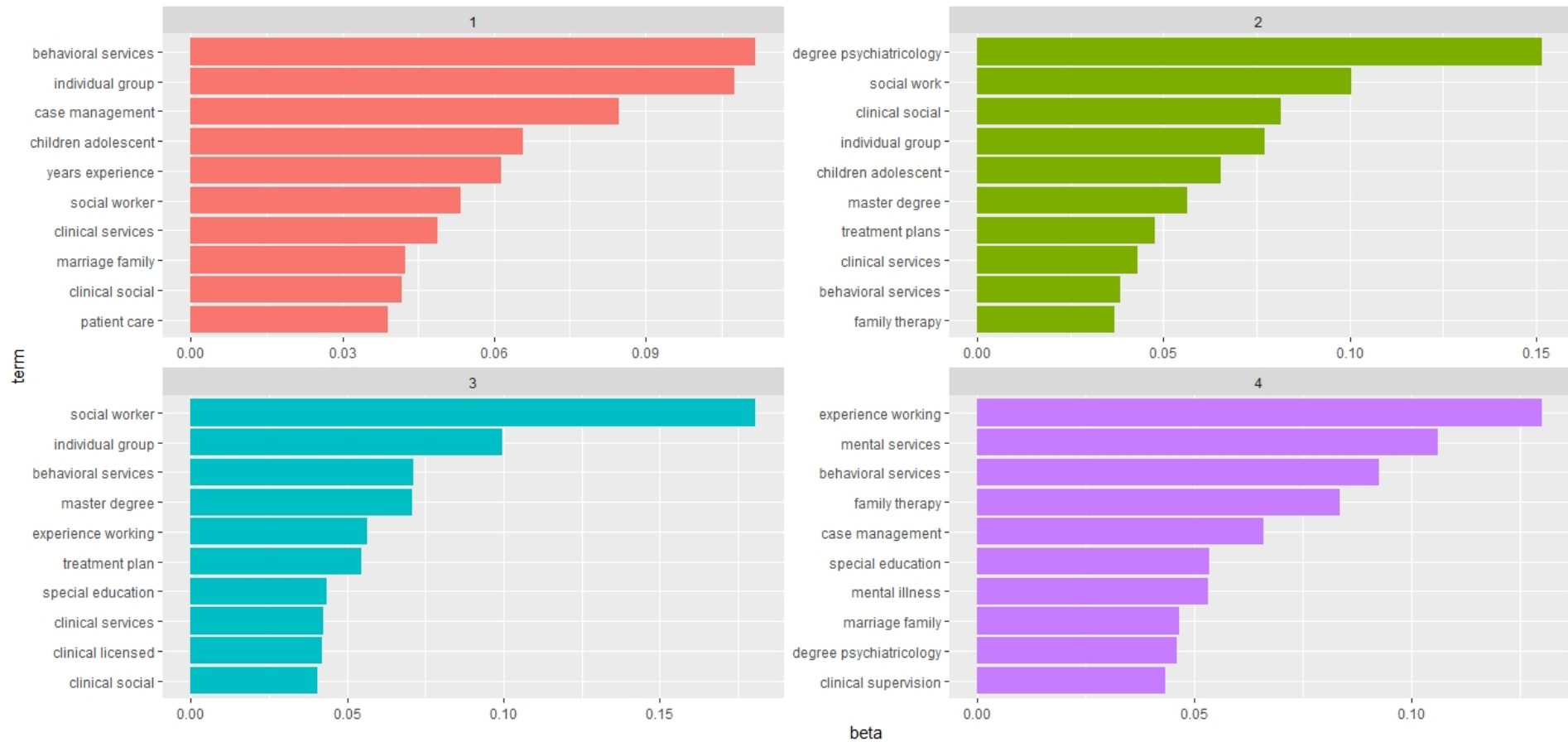
Results from the Psychology industry have been obtained analyzing the subset corpus from the extracted ads regarding the sector. A tokenized Document-Term Matrix (DTM) has been built, and sparsity was removed till 73%. **Fig. 91** shows the bigrams from the corpus. The most frequent terminological combinations were master's degree (1389), social work (1347), individual group (1308), years experience (1293), and treatment plan (1274). Topic modeling is presented in **Fig. 92** with four thematic areas. **Fig. 93** highlights the main correlations through the skills set. **Fig. 94** detects greedy modularity in the skillset, dividing it in three groups, and the relative memberships are shown in **Fig. 95**. Application of spectral modularity is presented in **Fig. 96** and the relative memberships reported in **Fig. 97**. The employment of optimal modularity detection is shown in **Fig. 98** and their memberships highlighted in **Fig. 99**. Modularity indicators were compared to define the most proper method to give sense to the analysis. Having $\xi_G > \xi_O > \xi_S$, the dendrogram in **Fig. 100** is built with greedy modularity, and partial correlations will be used for the weighted network in **Fig. 101**. Thus, a clustering plot with Zhang, Onnela, and Barrat methods is reported in **Fig. 102**. Centrality measures are exposed in **Fig. 103**. The most between skills in the set were project children adolescent (53.3%), marriage family (34.3%), mental services (18.9%), clinical licensed (12.6%), and social worker (11%). The closest skills were mental services (48.8%), children adolescent (45.5%), crisis intervention (34%), clinical licensed (33.6%), and treatment plan (33.1%). MCMC with MAP method is shown in **Fig. 104** to forecast and simulate a possible job interview for the Psychology industry.

Fig. 91: Bigrams of the Psychology skillset.



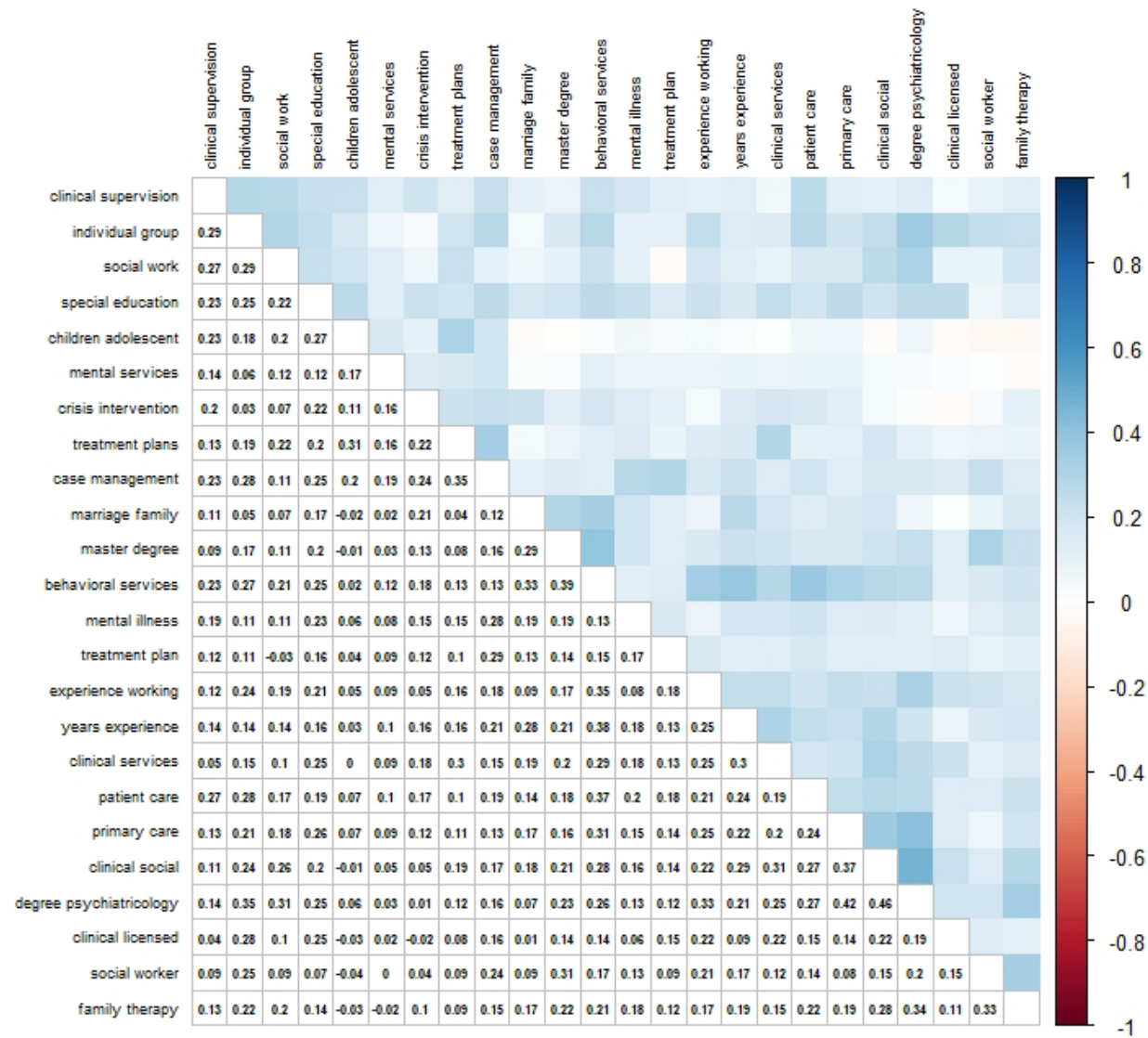
Source: own elaboration

Fig. 92: Topic modeling of the Psychology skill set.



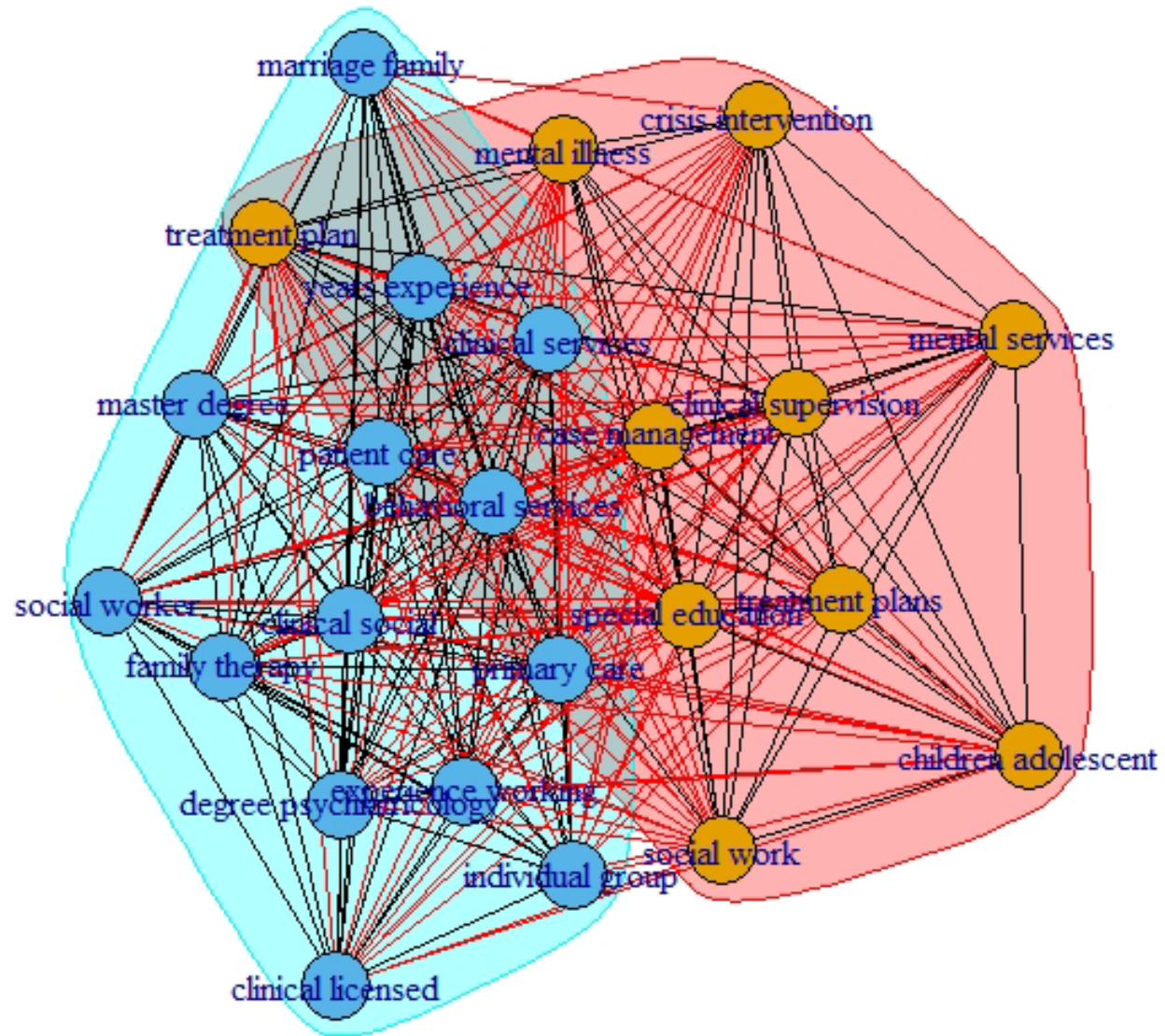
Source: own elaboration.

Fig. 93: Corrplot of the Psychology skill set.



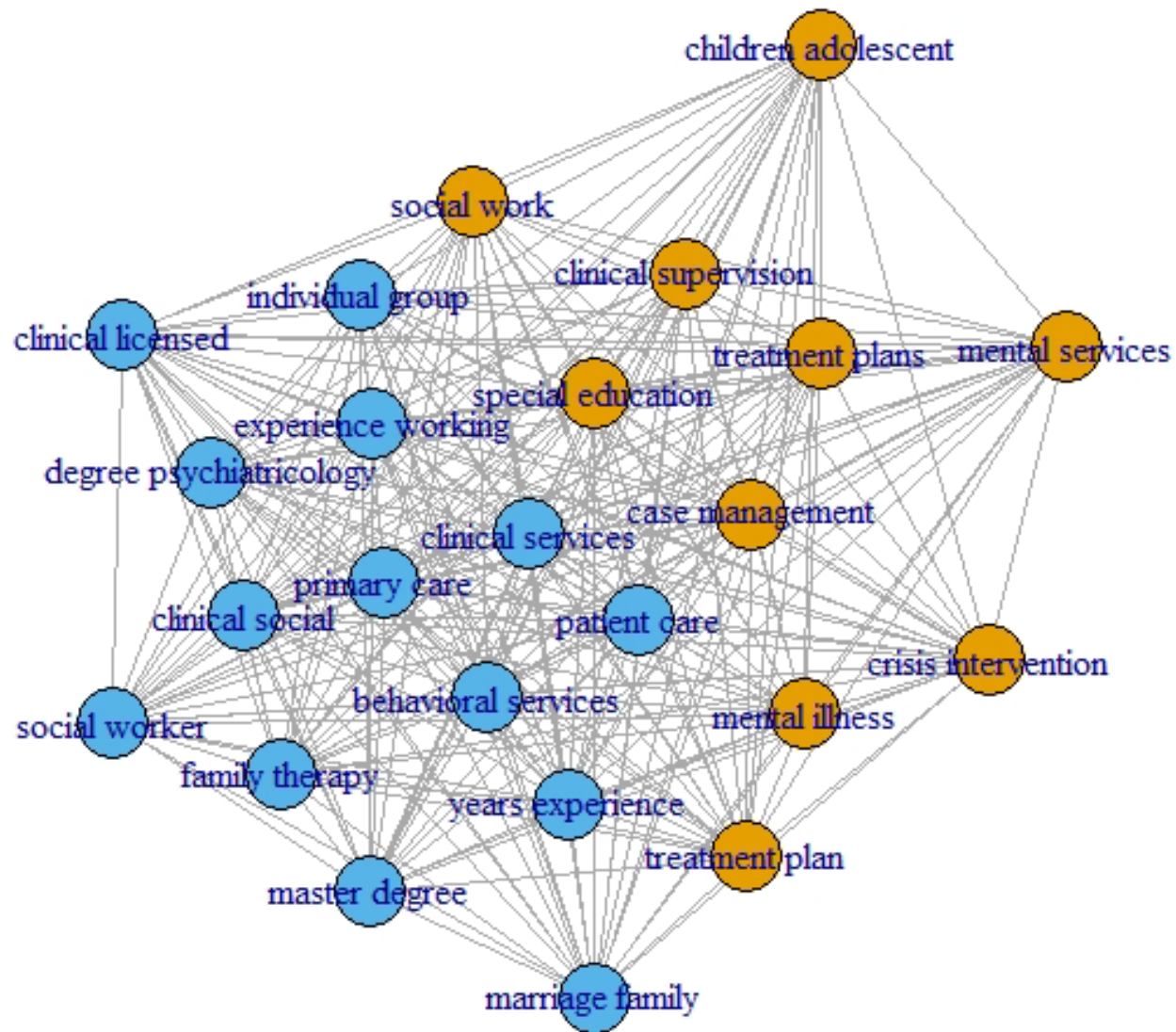
Source: own elaboration.

Fig. 94: Skills network with greedy modularity community detection.



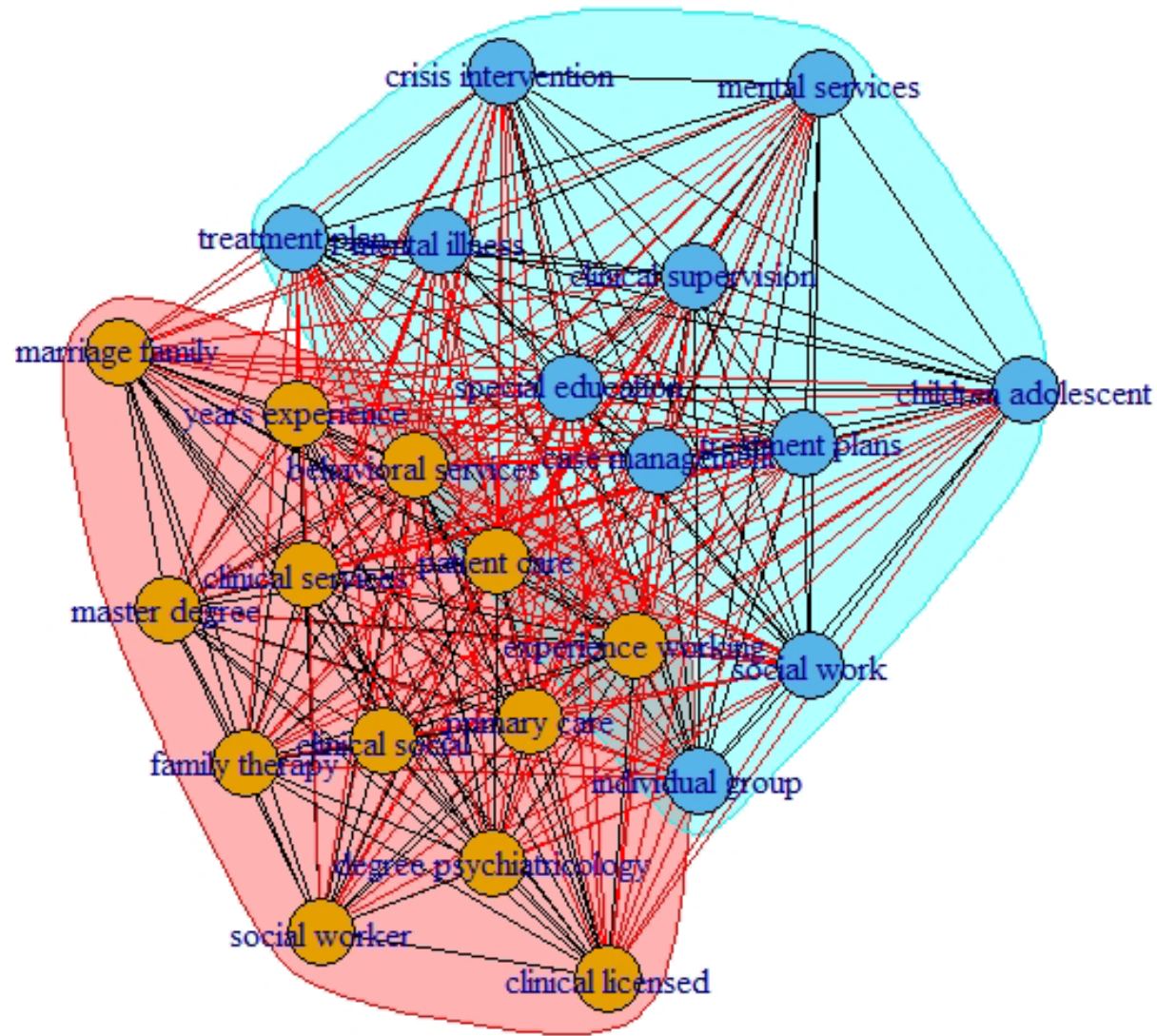
Source: own elaboration.

Fig. 95: Skills network community membership according to greedy modularity.



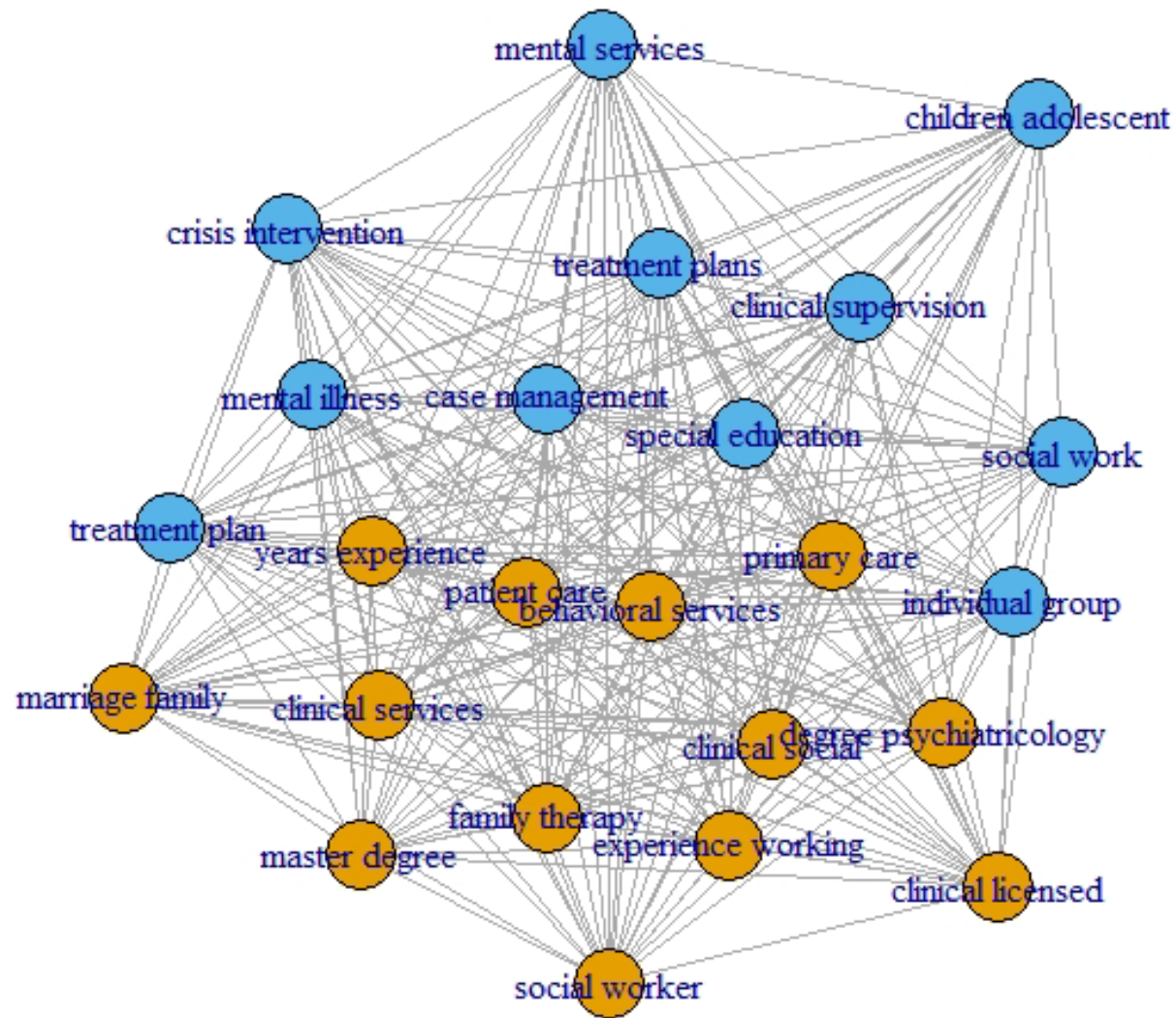
Source: own elaboration.

Fig. 96: Skills network with spectral modularity community detection.



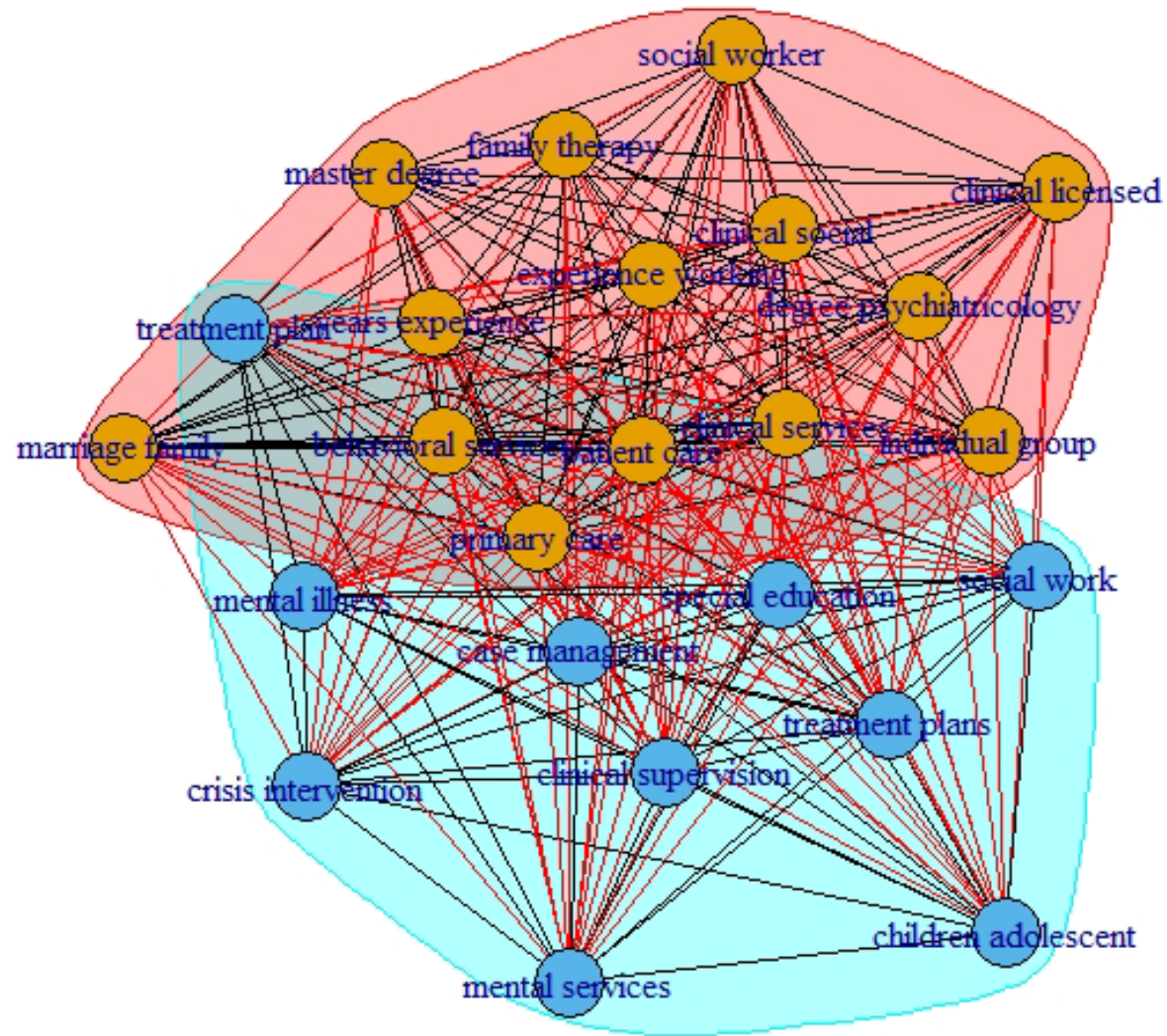
Source: own elaboration.

Fig. 97: Skills network community membership according to spectral modularity.



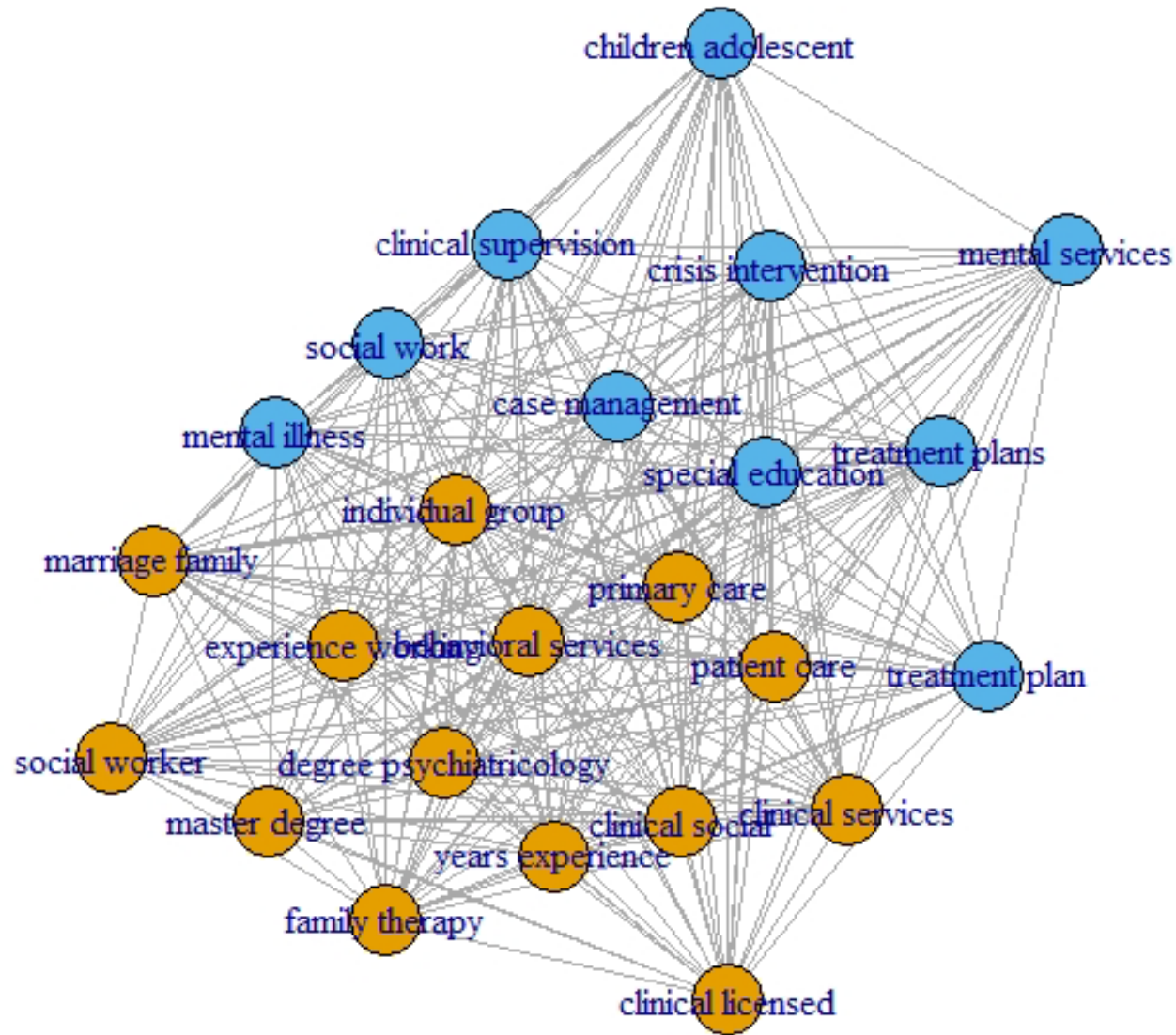
Source: own elaboration.

Fig. 98: Skills network with optimal community detection.



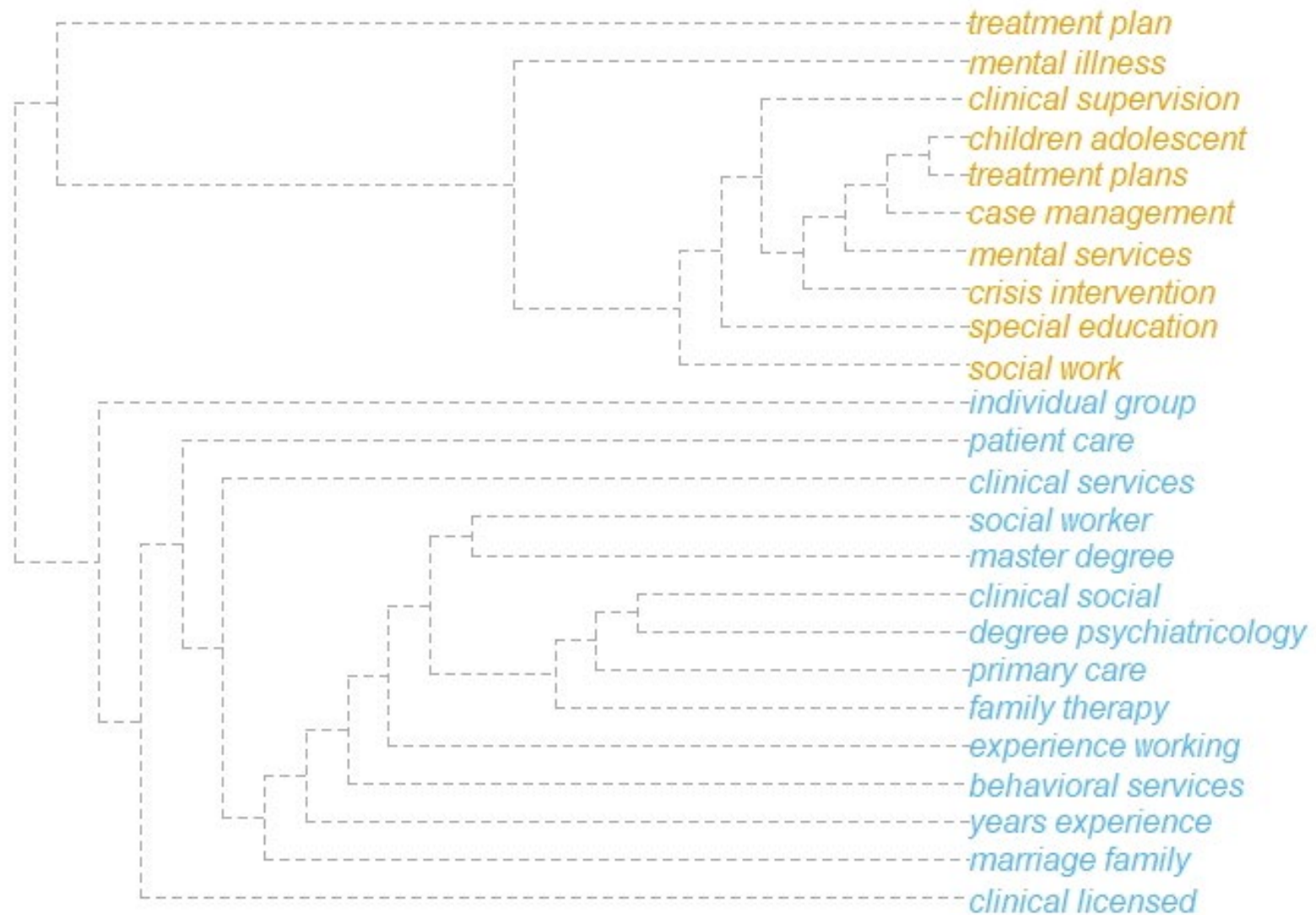
Source: own elaboration.

Fig. 99: Skills network community membership according to optimal modularity.



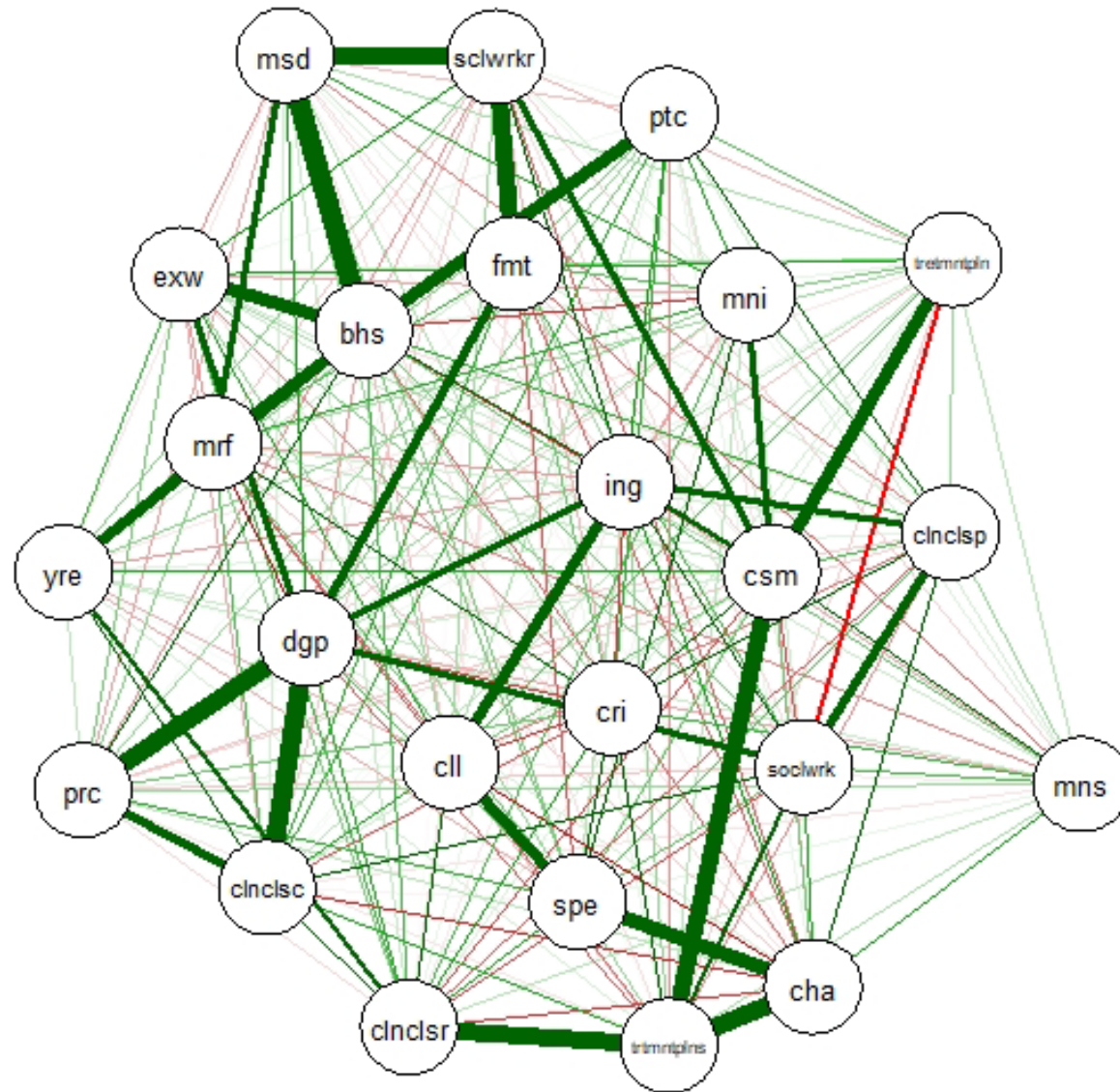
Source: own elaboration.

Fig. 100: Dendrogram with greedy modularity.



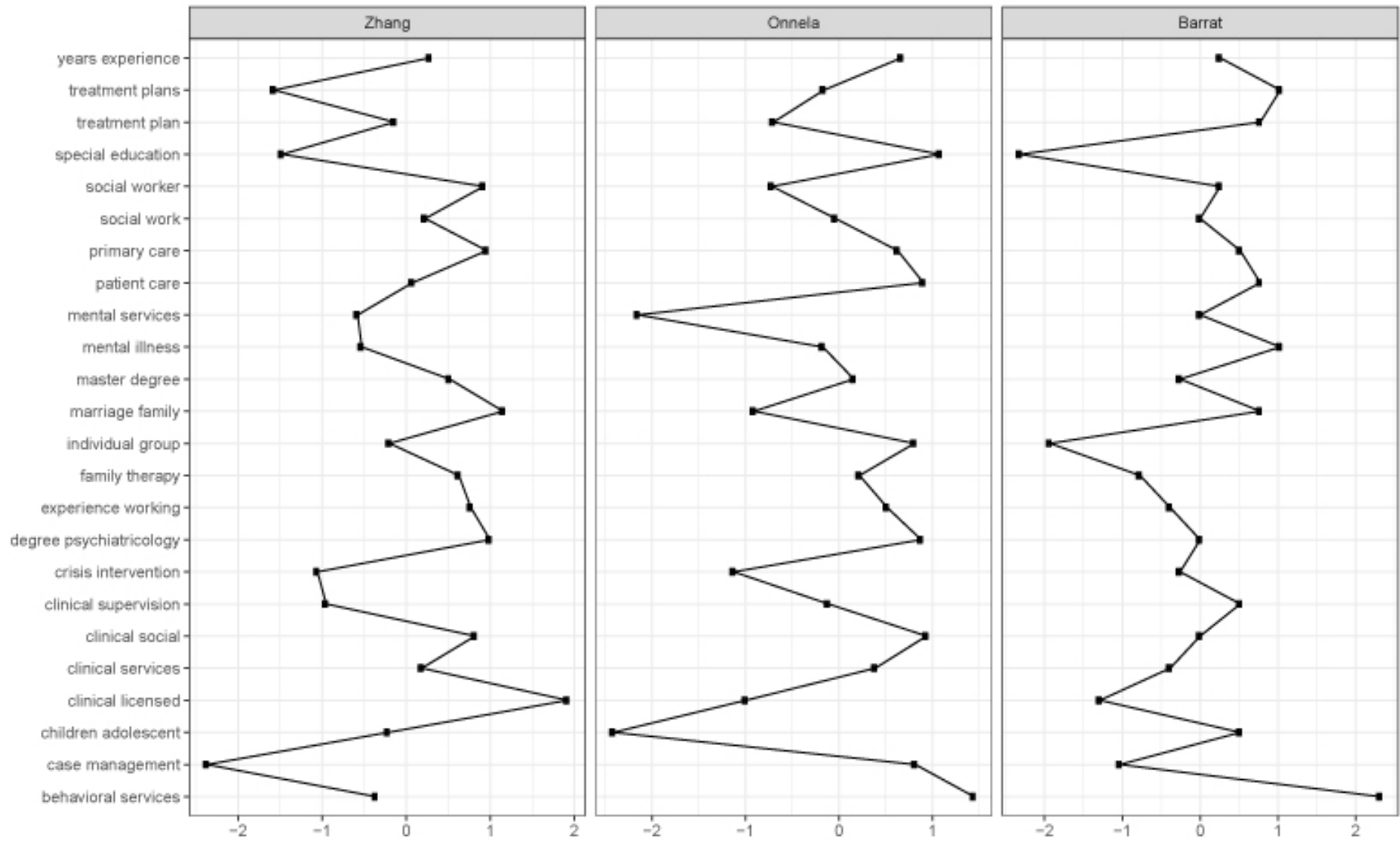
Source: own elaboration

Fig. 101: Weighted skills network via partial correlations clustering.



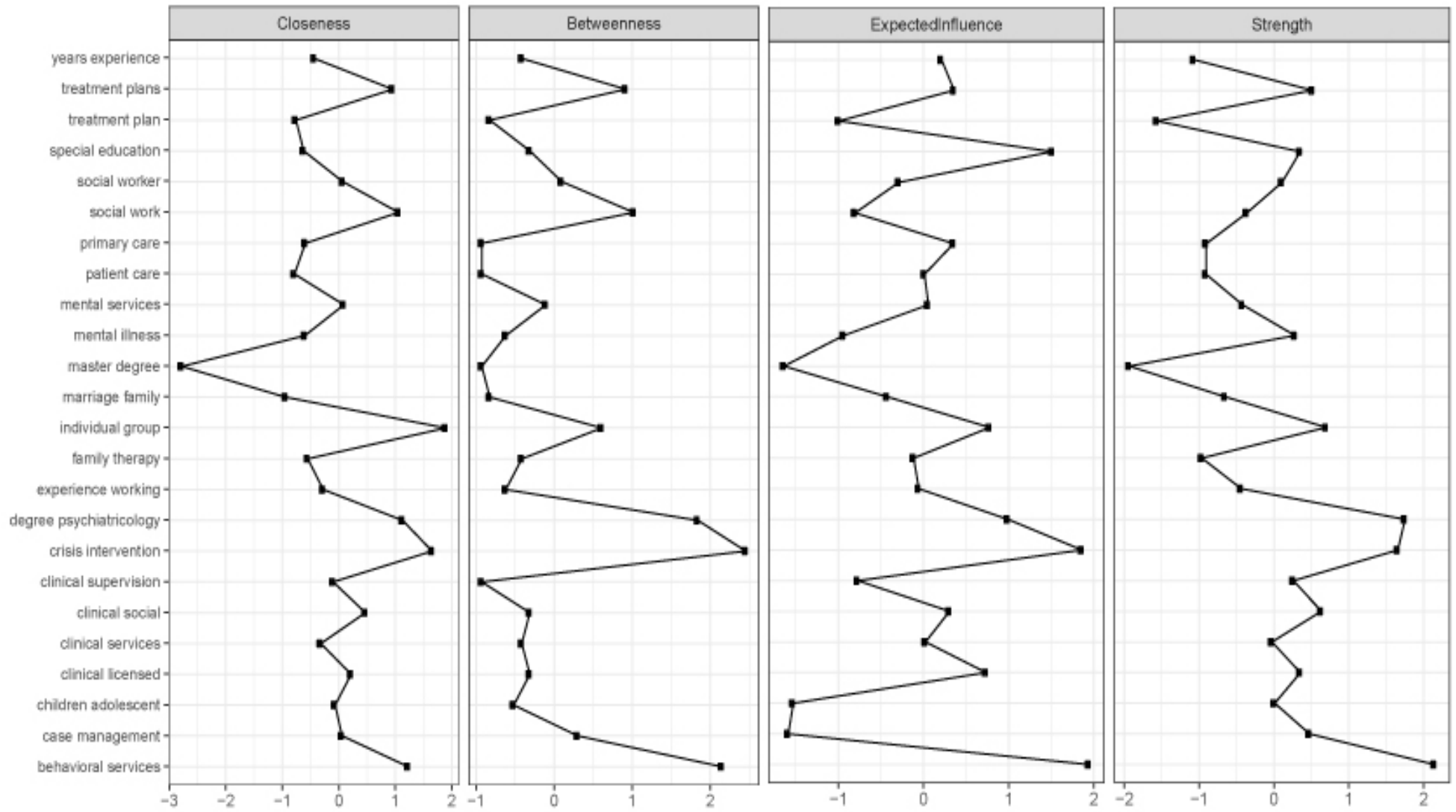
Source: own elaboration.

Fig. 102: Clustering plot with compared methods.



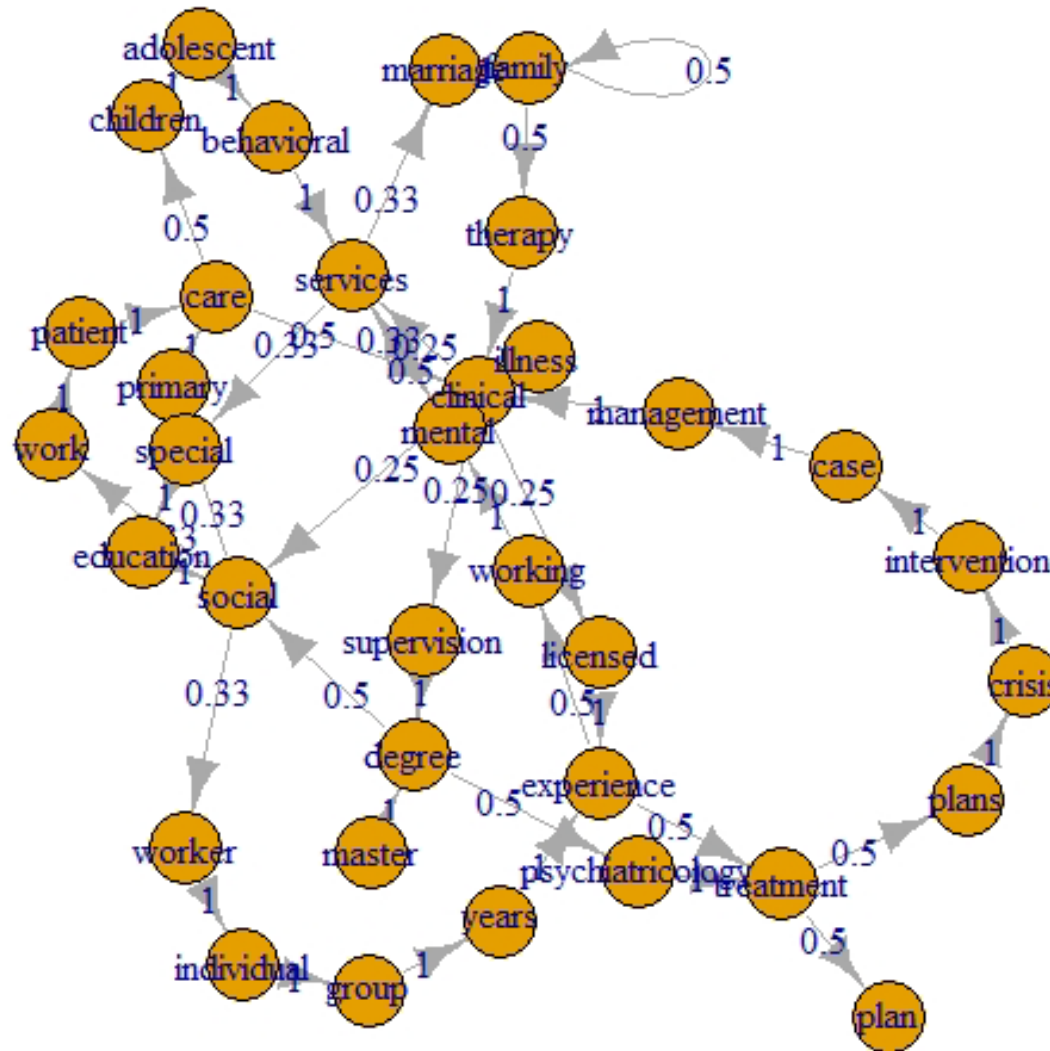
Source: own elaboration.

Fig. 103: Centrality measures plot.



Source: own elaboration.

Fig. 104: Monte Carlo Markov Chain with MAP method.

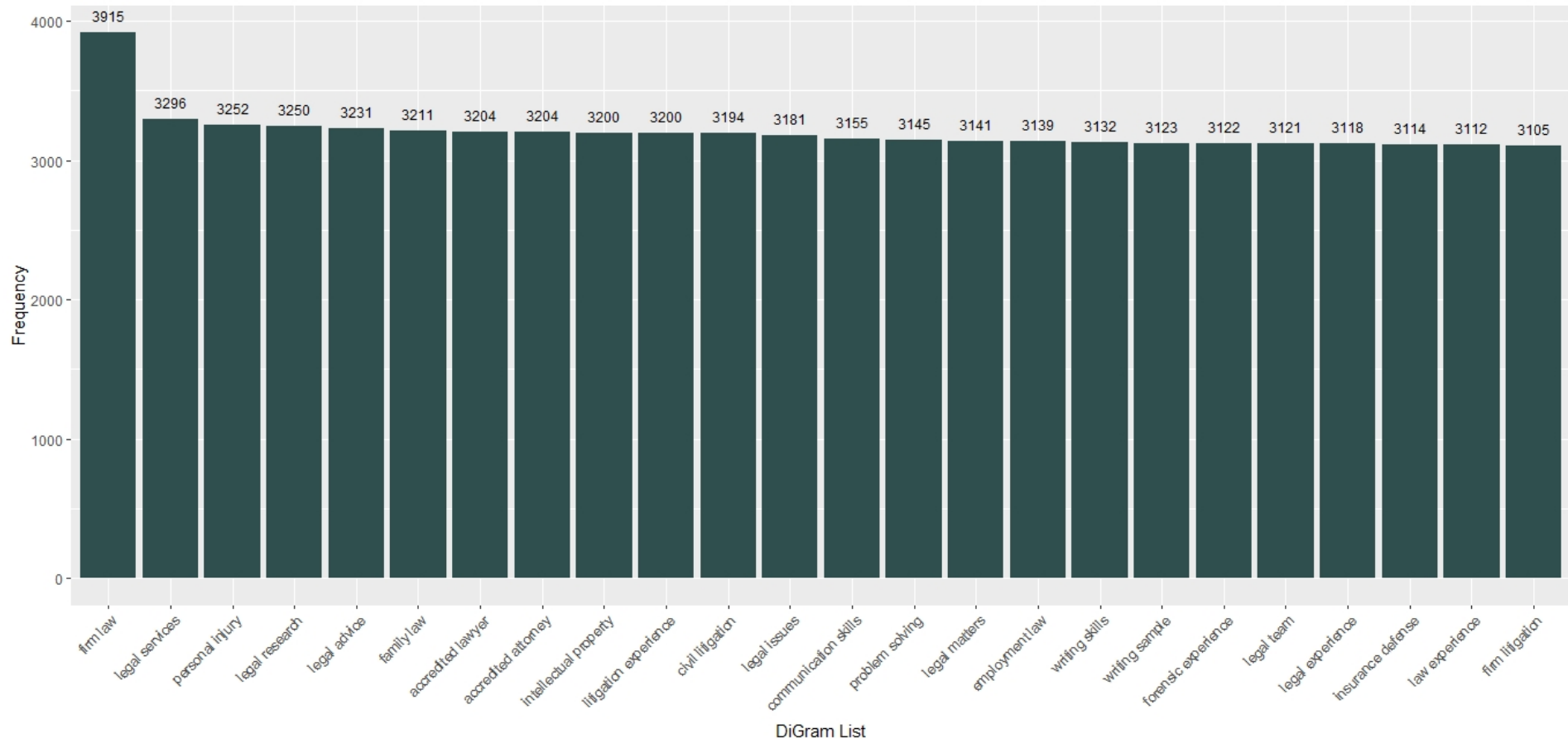


Source: own elaboration.

4.2.8 Law

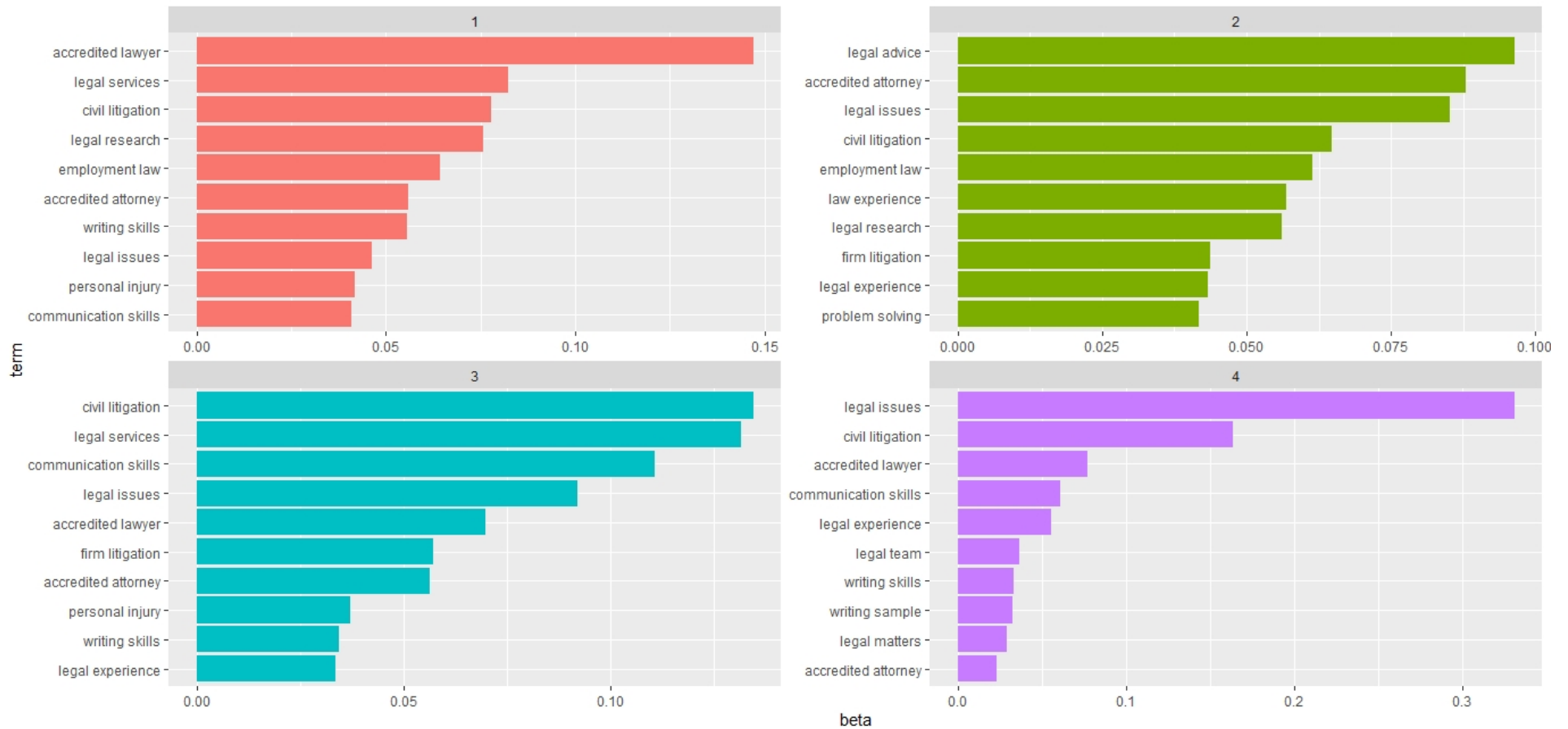
Results from the Law industry have been obtained analyzing the subset corpus from the extracted ads regarding the sector. A tokenized Document-Term Matrix (DTM) has been built, and sparsity was removed till 68%. **Fig. 105** shows the bigrams from the corpus. The most frequent terminological combinations were firm law (3915), legal services (3296), personal injury (3252), legal research (3250), and legal advice (3231). Topic modeling is presented in **Fig. 106** with four thematic areas. **Fig. 107** highlights the main correlations through the skills set. **Fig. 108** detects greedy modularity in the skillset, dividing it in three groups, and the relative memberships are shown in **Fig. 109**. Application of spectral modularity is presented in **Fig. 110**, and the relative memberships are reported in **Fig. 111**. The employment of optimal modularity detection is shown in **Fig. 112** and their memberships highlighted in **Fig. 113**. Modularity indicators were compared to define the most proper method to give sense to the analysis. Having $\xi_G > \xi_O > \xi_S$, the dendrogram in **Fig. 114** is built with greedy modularity and partial correlations will be used for the weighted network in **Fig. 115**. Thus, a clustering plot with Zhang, Onnela, and Barrat methods is reported in **Fig. 116**. Centrality measures are exposed in **Fig. 117**. The most between skills in the set were family law (32.4%), personal injury (26.8%), problem-solving (16.2%), firm law (16%), and legal team (15.4%). The closest skills were firm law (67.7%), communication skills (57.2%), personal injury (56.6%), family law (54.7%), and employment law (51.7%). MCMC with MAP method is shown in **Fig. 118** to forecast and simulate a possible job interview for the Law industry.

Fig. 105: Bigrams of the Law skillset.



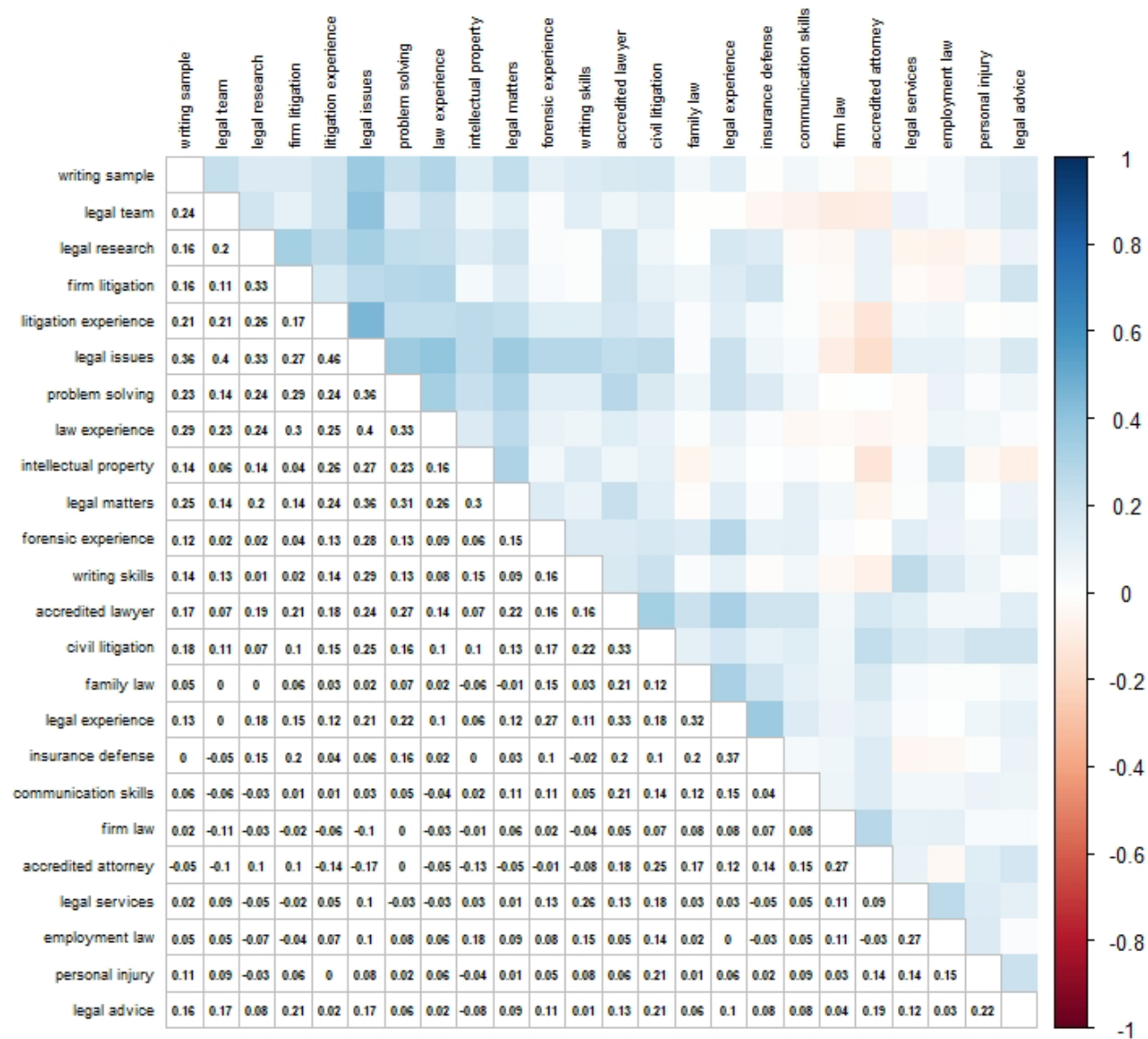
Source: own elaboration

Fig. 106: Topic modeling of the Law skillset.



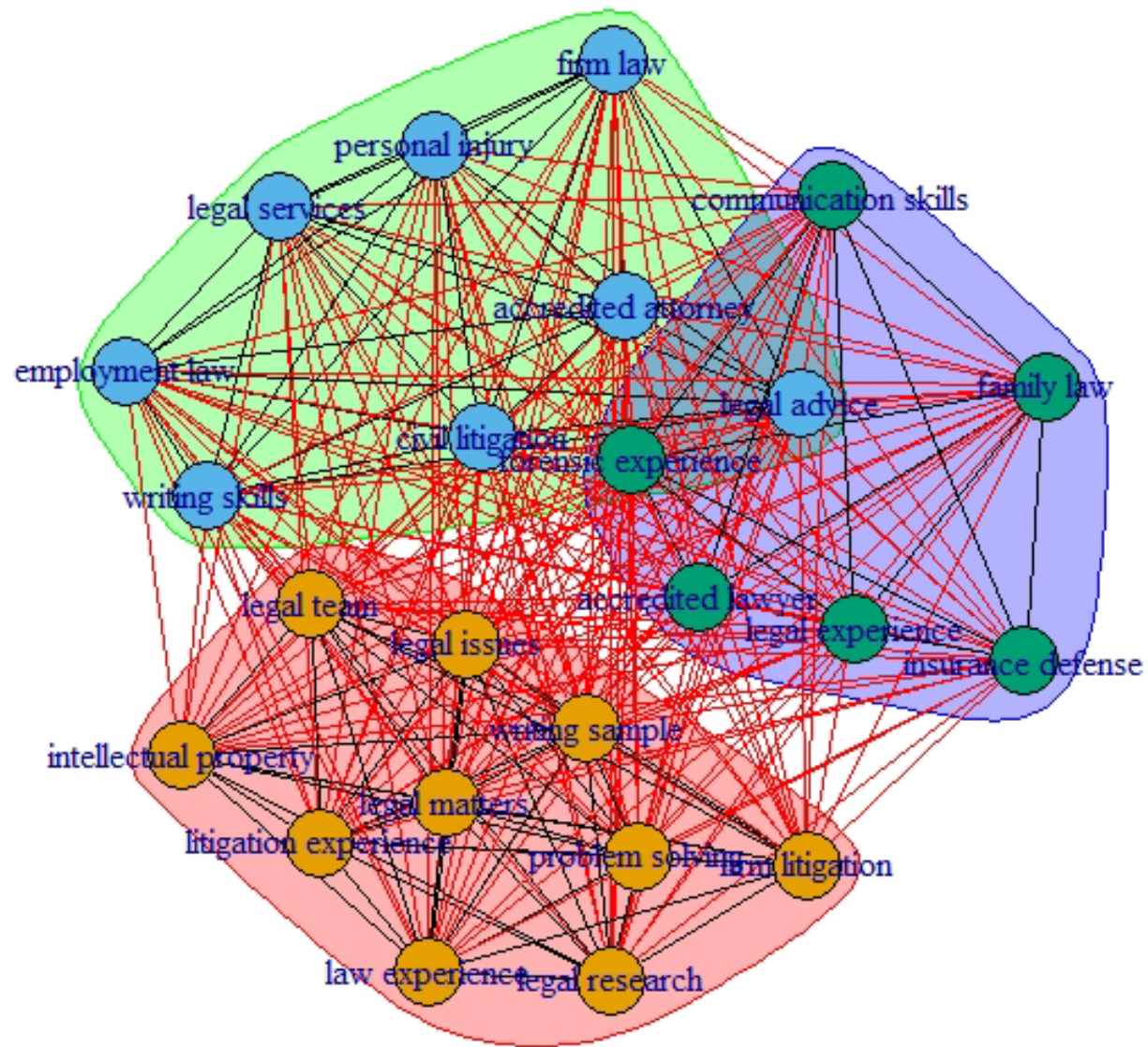
Source: own elaboration.

Fig. 107: Corrplot of the Law skillset.



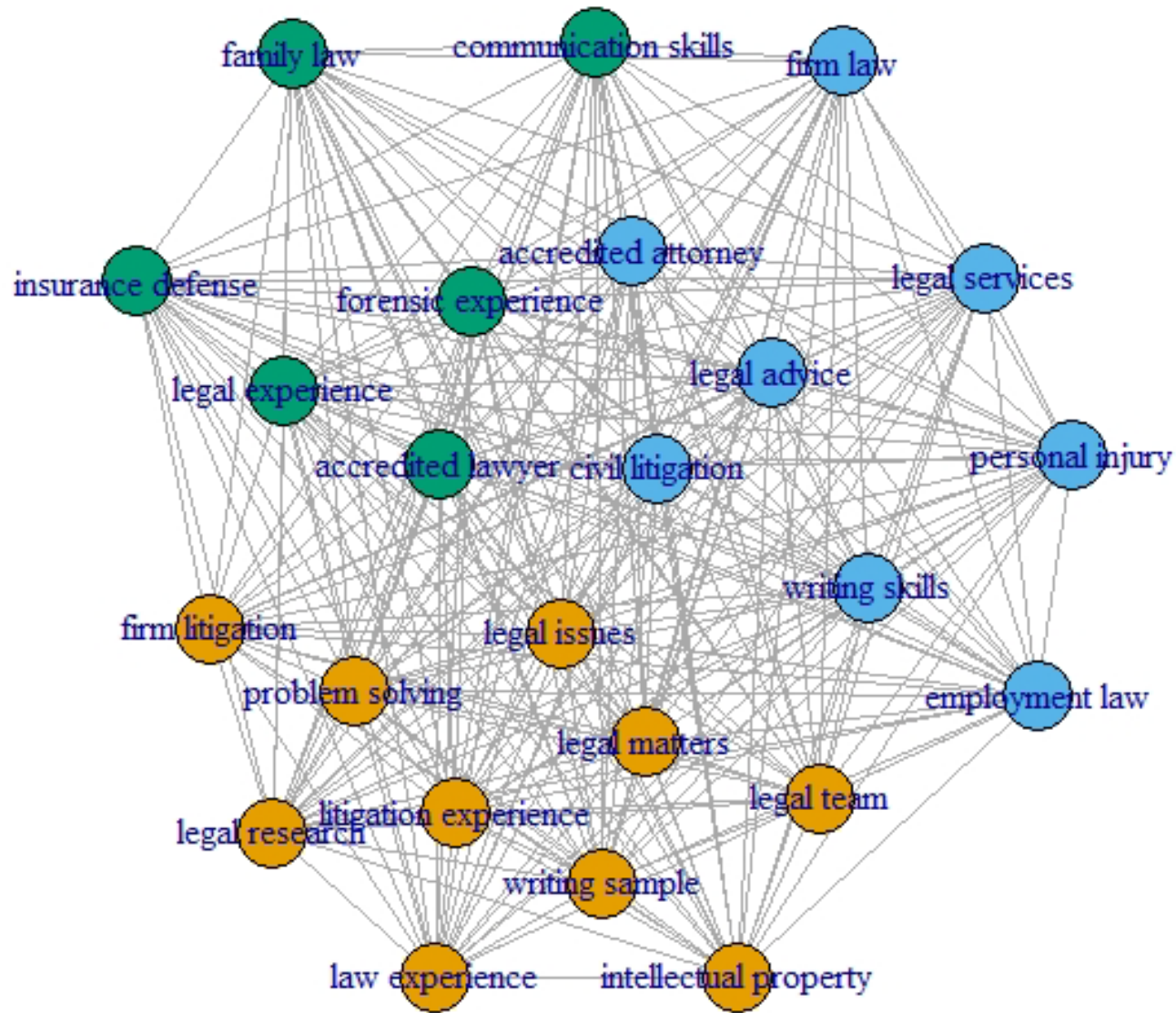
Source: own elaboration.

Fig. 108: Skills network with greedy modularity community detection.



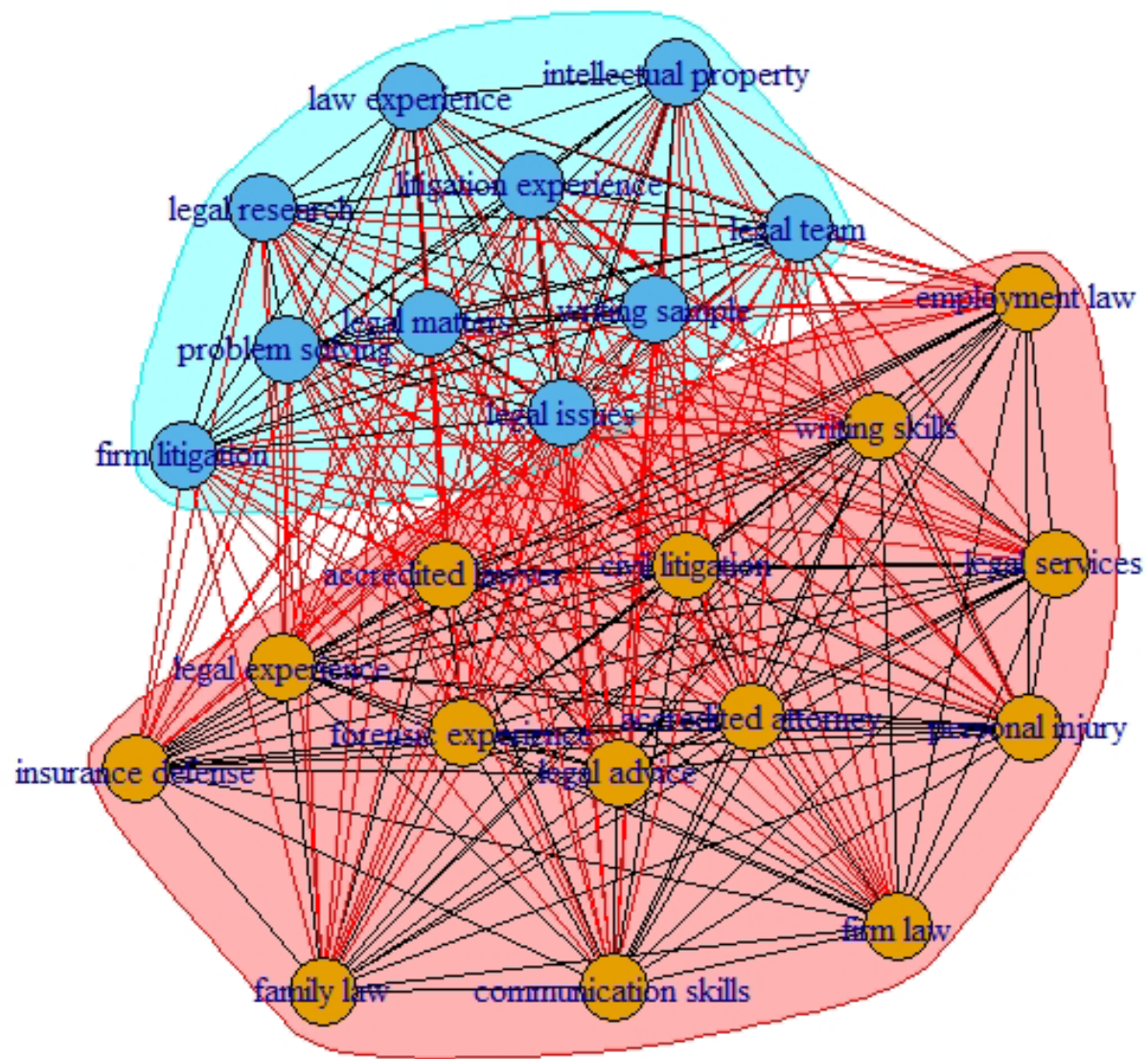
Source: own elaboration.

Fig. 109: Skills network community membership according to greedy modularity.



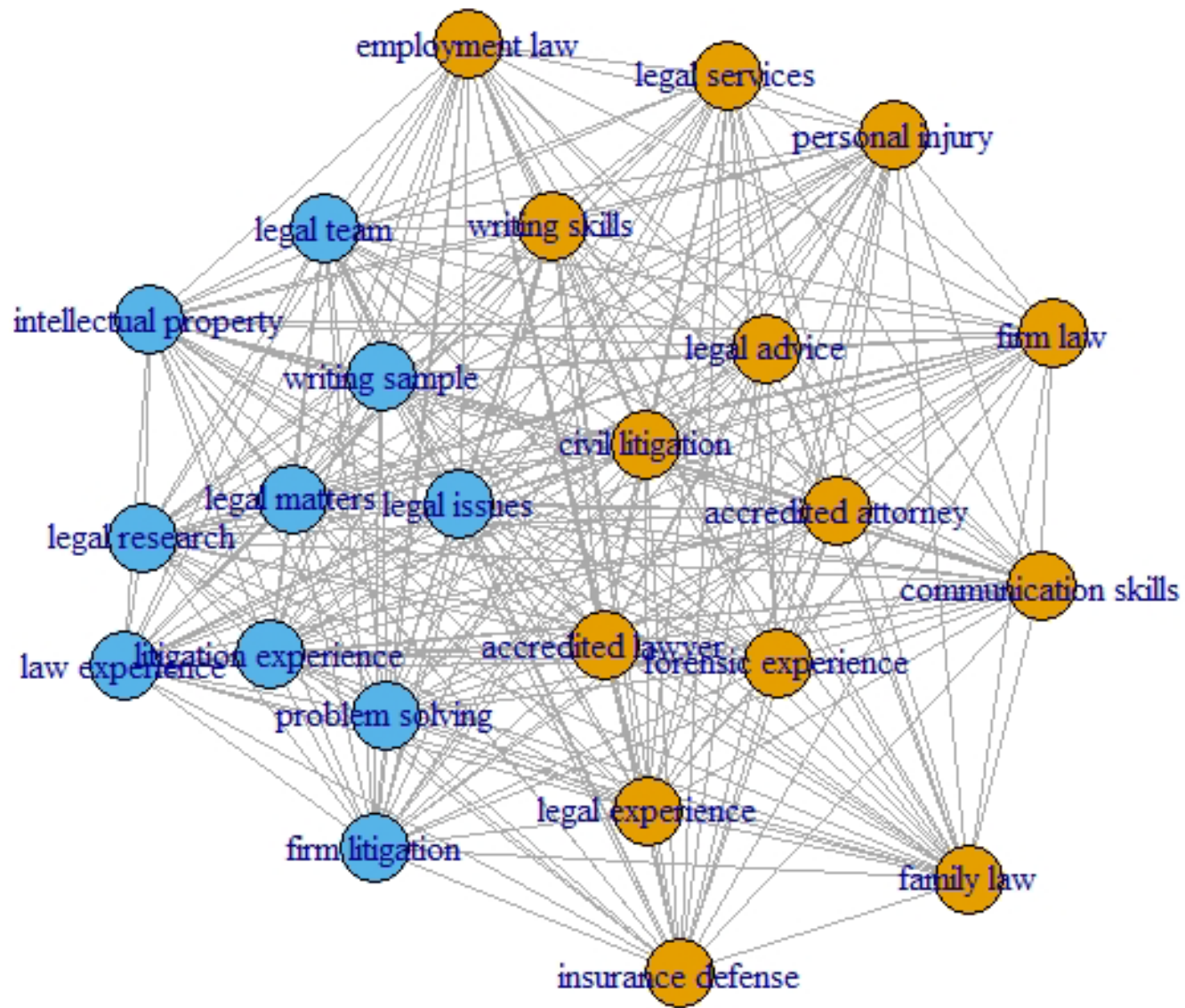
Source: own elaboration.

Fig. 110: Skills network with spectral modularity community detection.



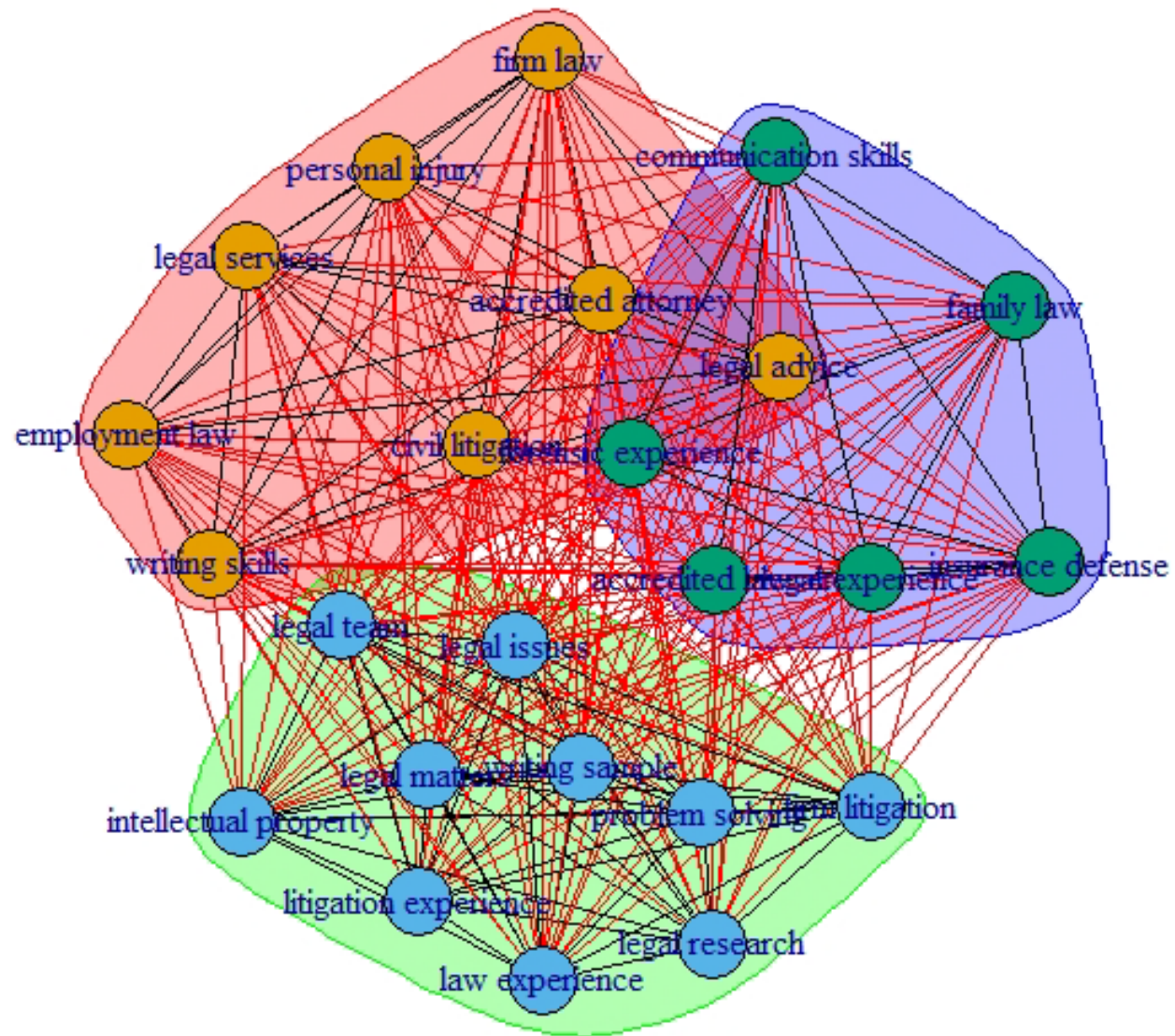
Source: own elaboration.

Fig. 111: Skills network community membership according to spectral modularity.



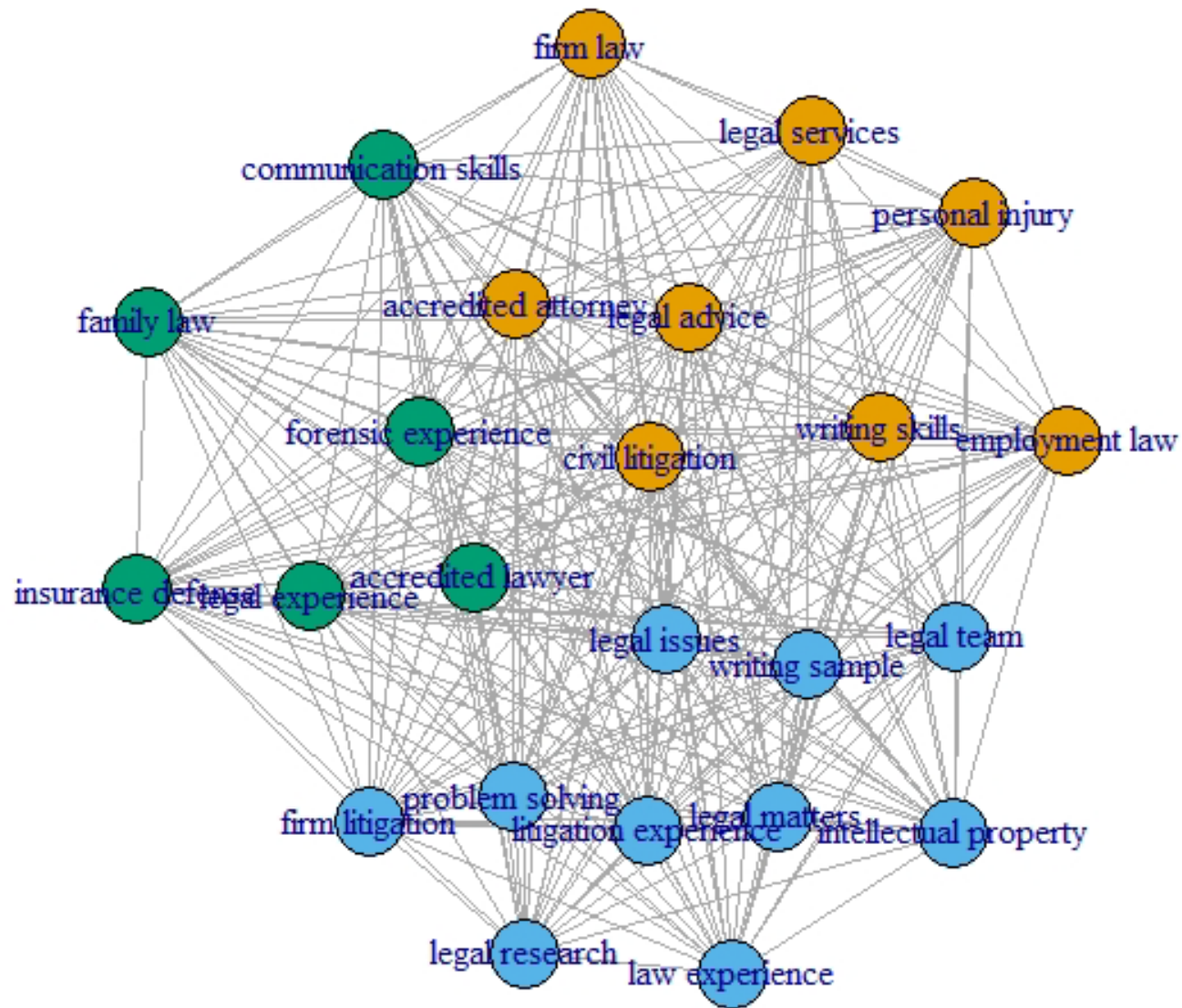
Source: own elaboration.

Fig. 112: Skills network with optimal community detection.



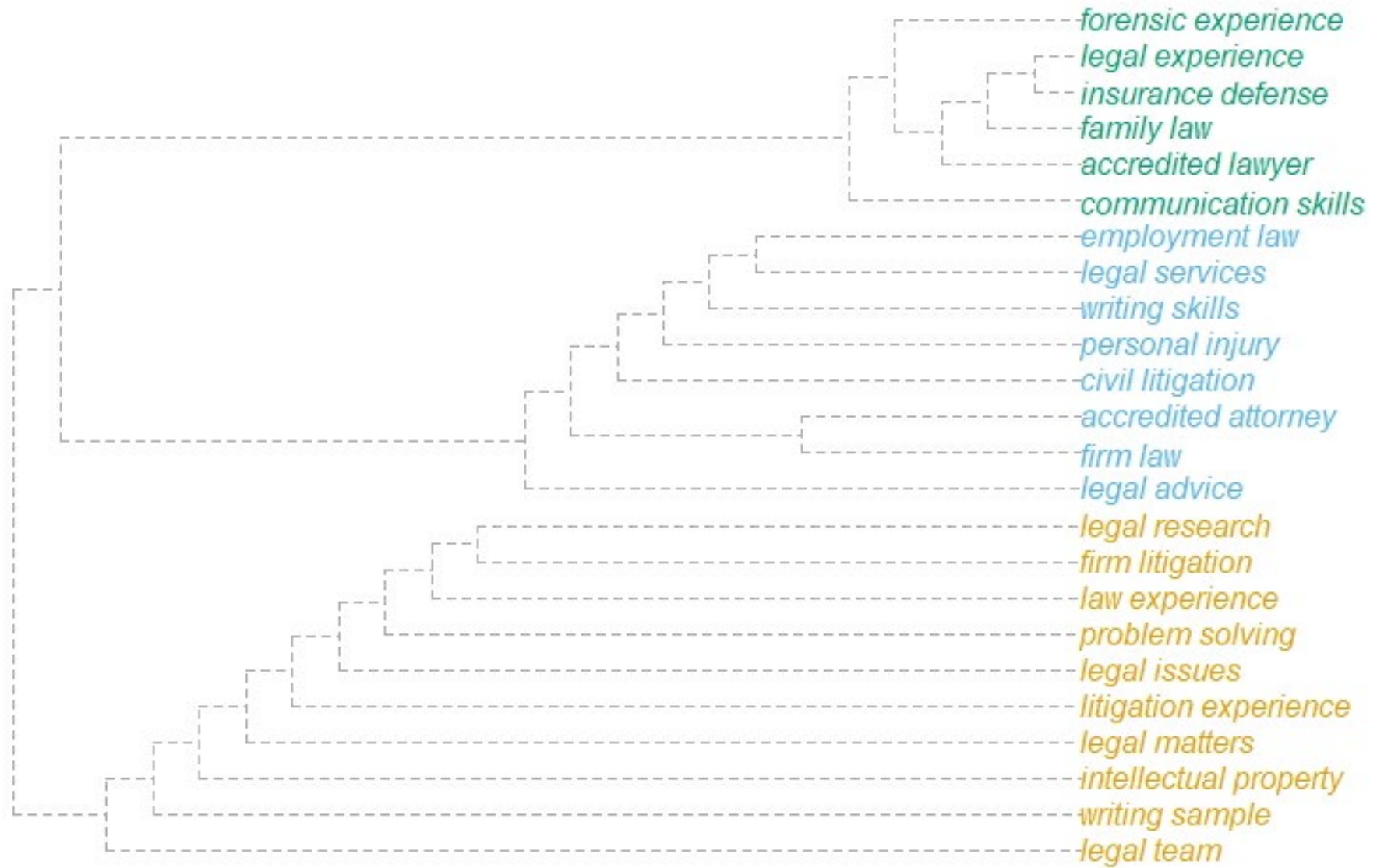
Source: own elaboration.

Fig. 113: Skills network community membership according to optimal modularity.



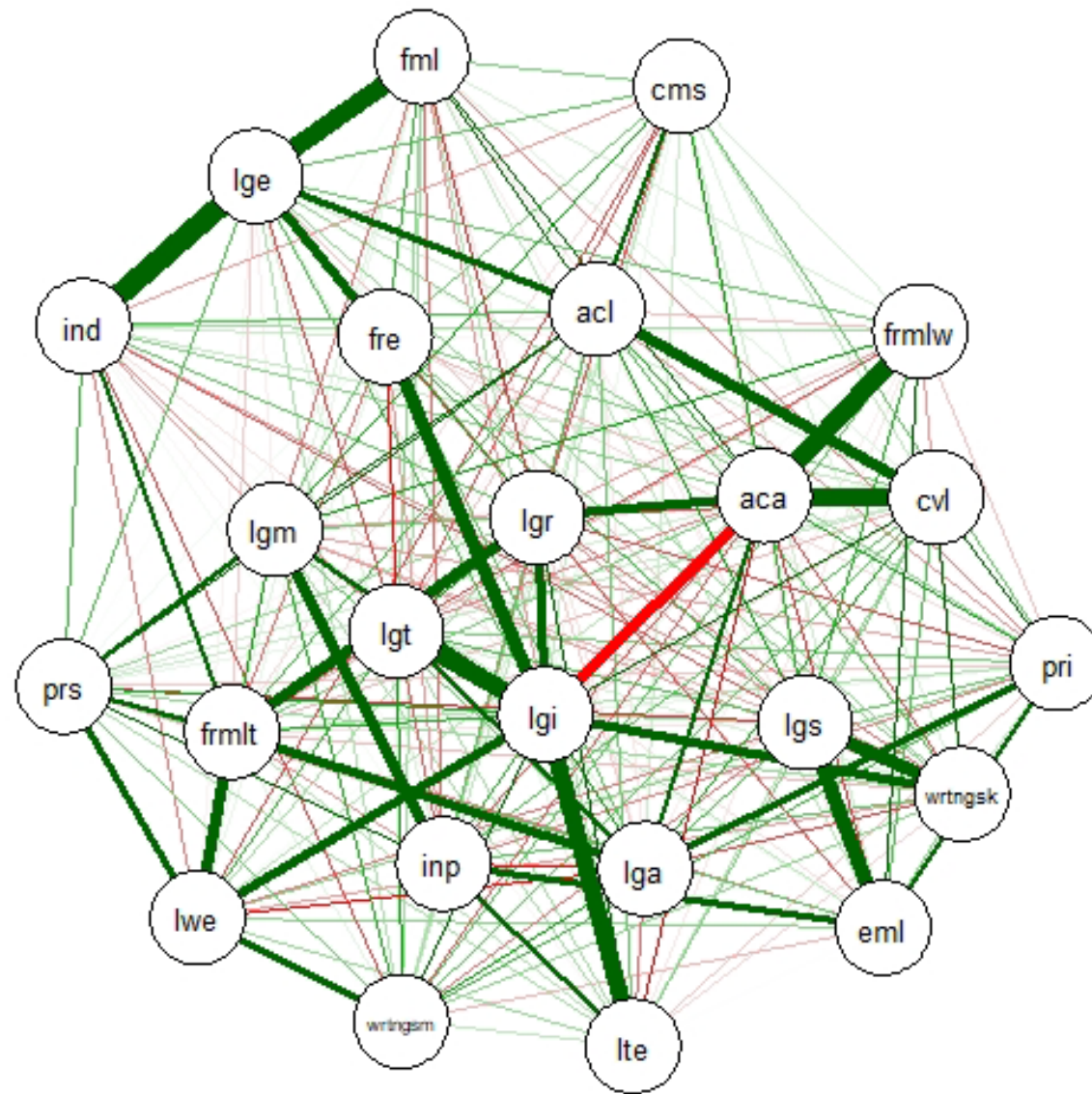
Source: own elaboration.

Fig. 114: Dendrogram with greedy modularity.



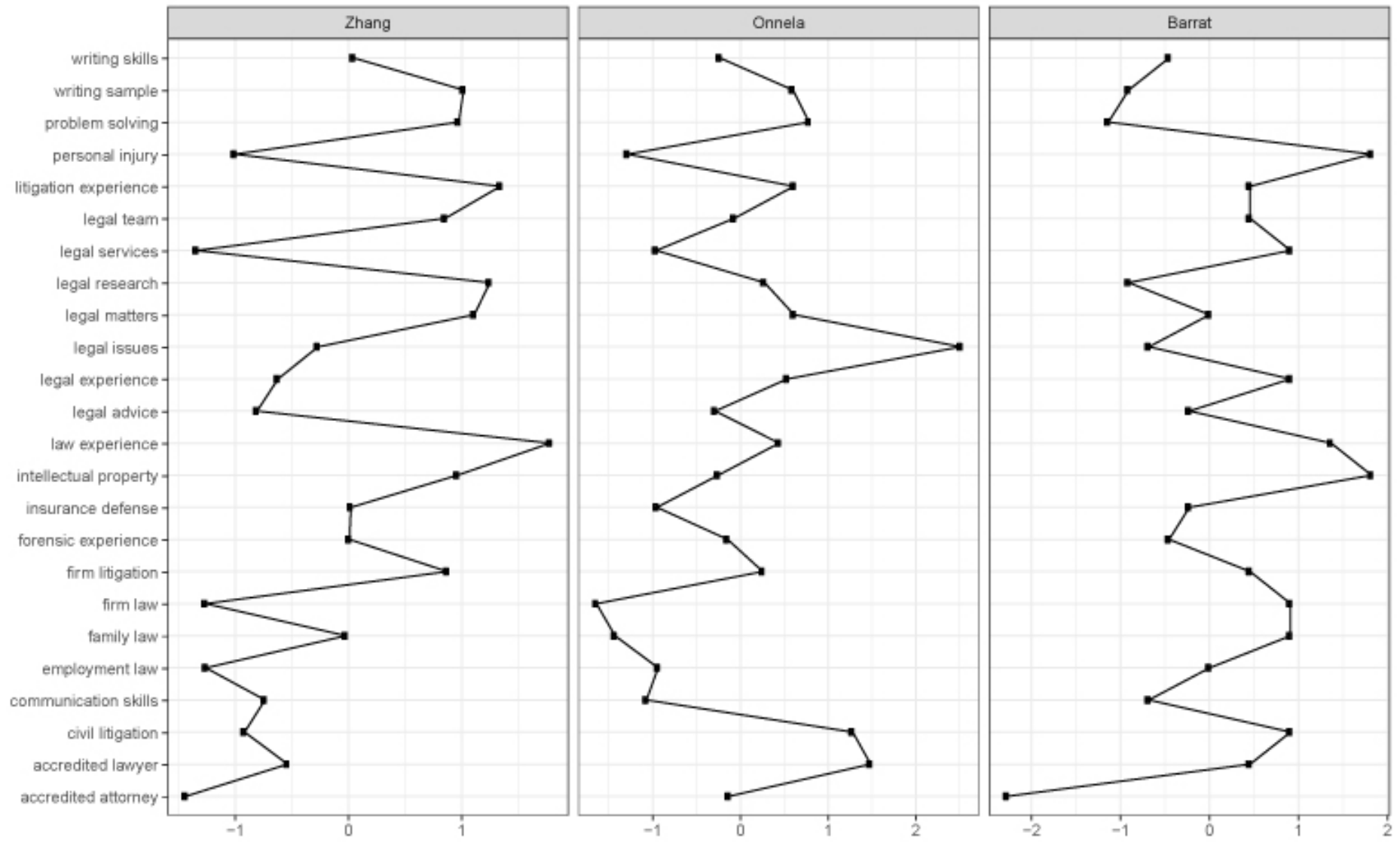
Source: own elaboration

Fig. 115: Weighted skills network via partial correlations clustering.



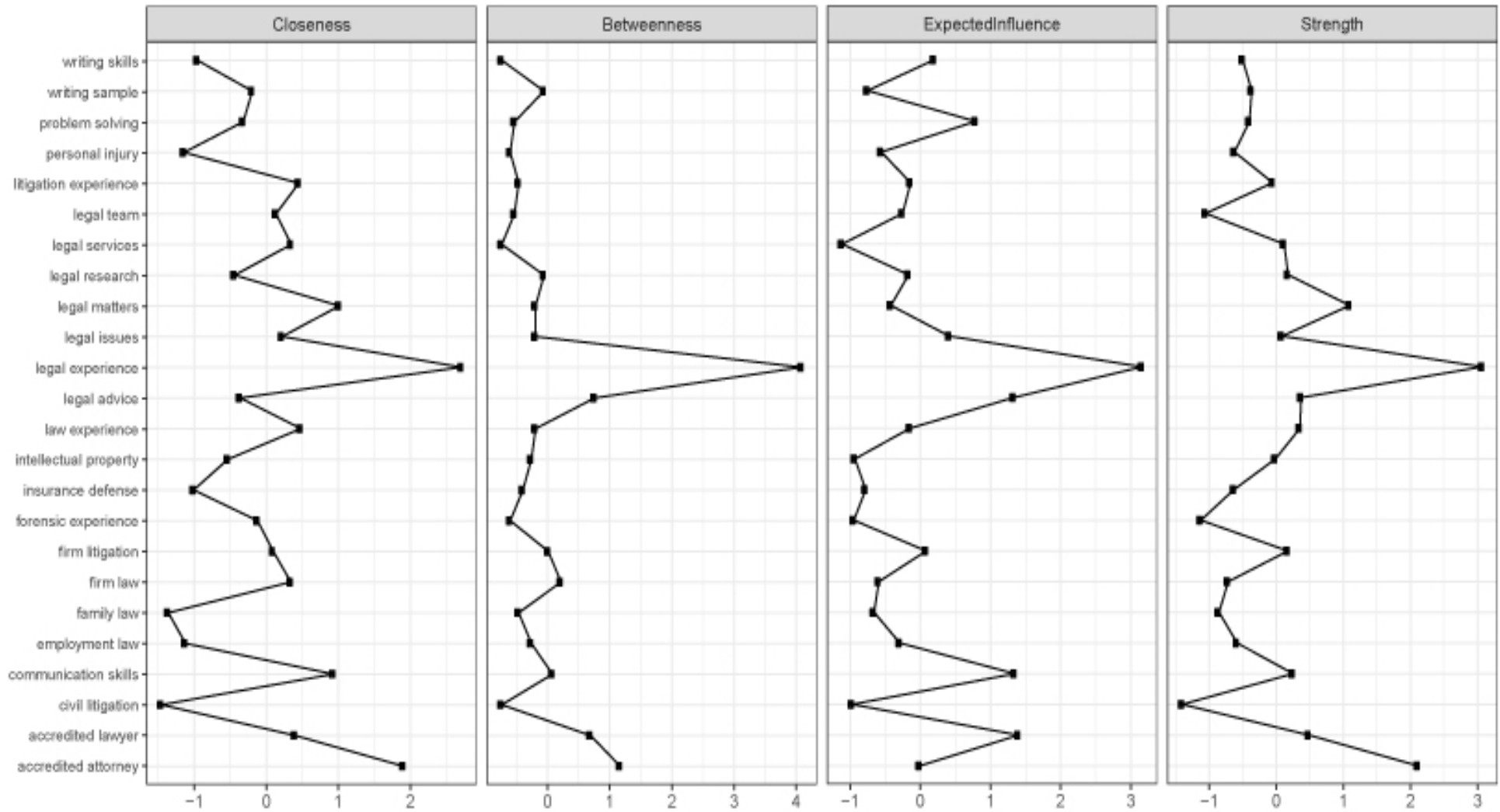
Source: own elaboration.

Fig. 116: Clustering plot with compared methods.



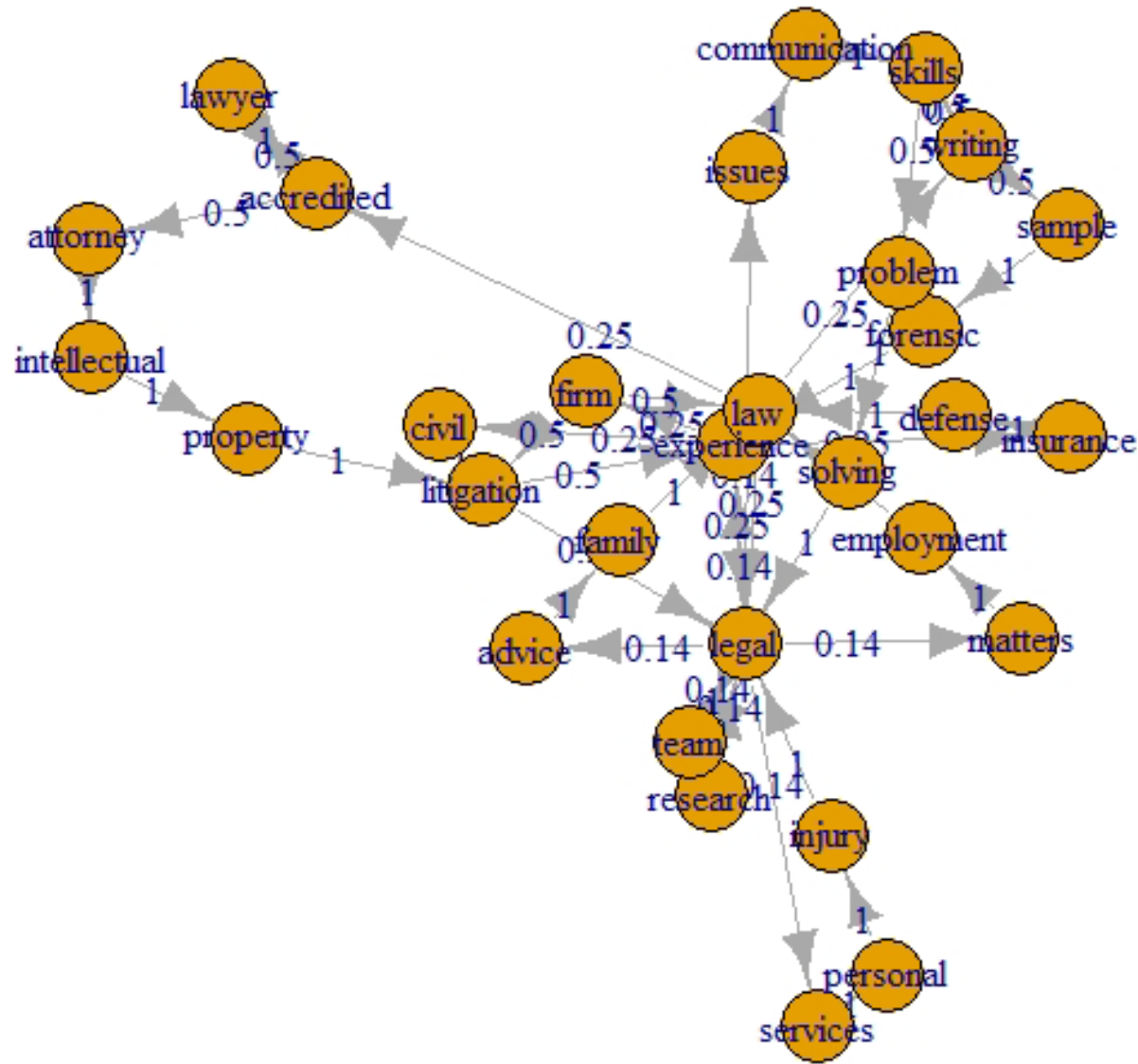
Source: own elaboration.

Fig. 117: Centrality measures plot.



Source: own elaboration.

Fig. 118: Monte Carlo Markov Chain with MAP method.



Source: own elaboration.

5. Discussion

As reported in the introduction, this doctoral thesis aimed to investigate job ads to extract those considered requirements for candidates seeking open job vacancies in eight sectors to build a DSS.

Thus, to analyze the extracted skills, the analysis results will be discussed here to respond to the research questions. To ensure readability, the discussion will be divided by sector.

5.1 Marketing

First, the results from the marketing sector allow an understanding of which are the most requested skills for the open job vacancies for marketing manager, brand manager, presales consultant, market analyst, and digital marketing manager. Almost 9k ads were extracted with a low standard deviation, ensuring reliability. The US state with the highest concentration of job ads was California.

Responding to RQ1, results from TM and stylometry show that the most requested skills by firms operating in the marketing sector or a marketing division were social media, years of experience, project management, communication skills, and learning. Again, results are aligned with the

literature review (Rosenberg et al., 2012; Metilda & Neena, 2017; Nisha & Rajasekaran, 2018; Ojanaperä et al., 2018; Clayton & Harris, 2019; Di Gregorio et al., 2019; Osmani et al., 2019). In particular, the knowledge extraction process highlighted a hybrid skillset melting hard and soft skills (Wolf and Archer, 2013; Alamelu et al., 2017, Entika, 2017; Hirsch, 2017; Mahfud et al., 2017; Woya, 2019).

According to the literature, the results are coherent with the concept of redefining skills in the current market composition (Hall, 2002), enhancing attention to soft skills like learning ability, teamwork, and communication skills (Savickas, 2005; Pool and Sewell, 2007; Mishra, 2014; De Fruyt et al., 2015; Hirsch, 2017).

Hard skills, such as project management, written communication, specialized experience, market research, and Microsoft Office, have emerged from the textual analysis, highlighting that, at present, there is still great attention placed upon technical skills, according to the literature review (Mishra, 2014; Harris and King, 2015; Cimatti, 2016; Alamelu et al., 2017; Oviawe et al., 2017; Woya, 2019).

There is a melting between hard, soft, and managerial skills in managerial skills, digital marketing, related field, multiple projects, and marketing experience. According to the literature review (Wolf and Archer, 2013; Mishra, 2014; Nurlaela et al., 2017; Hirsch, 2017; Petrovski et al., 2017;

Chaibate et al., 2020; Ijaola et al., 2020), these results are aligned with the new market trends in terms of employability (Fugate et al., 2004; Fugate 2006; Fugate and Kinicki, 2008).

Furthermore, responding to RQ2, the results highlighted four thematic areas regarding the skillset that a marketer should possess. (1) The first thematic area includes market research, project management, best practices, teamwork, communication skills, attention to detail, and specialized experience. The first topic is comprised of both hard and soft skills (Nurlaela et al., 2017; Hirsch, 2017; Petrovski et al., 2017; Chaibate et al., 2020; Ijaola et al., 2020). (2) The second topic presents skills such as marketing communication, learning ability, written communication, multiple projects, social media, attention to detail, and communication skills. Even if there is the presence of hard skills, such as market research, project management, and specialized experience, soft skills are predominant in this thematic area (Savickas, 2005; Pool and Sewell, 2007; Mishra, 2014; De Fruyt et al., 2015; Hirsch, 2017). (3) The third group of skills highlights specialized experience, market research, affirmative action, project management, best practices, years experience, digital marketing, bachelor's degree, and marketing and written communication. This topic represents the most requested skills for a candidate in the marketing field (Mishra, 2014; Harris and King, 2015; Cimatti, 2016; Alamelu et al., 2017; Oviawe et al., 2017; Woya, 2019).

(4) The fourth topic included internal-external communication, written-verbal communication, management skills, marketing communication, and marketing experience. The presence of these kinds of skills highlighted the pivotal role of industry-specific and job-specific skills that are fundamental for candidates who are going to be employed within that field—in this case, marketing (Savickas, 2005; Pool and Sewell, 2007; Harris and King, 2015; Alamelu et al., 2017; Nurlaela et al., 2017; Oviawe et al., 2017; Woya, 2019).

Thus, responding to RQ3, results from the correlation analysis highlighted several medium or strong correlations through the skillset. The most considerable correlations between years experience and social media, years experience, and project management ($r=.55$) are between communication skills and social media ($r=.54$). The strongest correlations in the model regard communication skills. Written communication and verbal communication present a strong correlation ($r=.80$), and written communication and internal-external communication do as well ($r=.67$). These results are consistent with the literature review (Makki et al., 2015; Nurlaela et al., 2017; Oviawe et al., 2017; Woya, 2019; Di Gregorio et al., 2019).

Responding to RQ4, several methods were compared in configuring skills as a social network. Greedy, spectral, and optimal modularity were detected and compared, building-specific indicators to evaluate their performance. The best-fitting indicator was $\xi_G=.97$, representing the performance of greedy modularity detection, which was found almost with a perfect model fit.

Indicators to evaluate spectral modularity and optimal modularity performance were relatively lower than the greedy one ($\xi_s=.87$, $\xi_o=.93$). For this reason, the ultimate dendrogram to group skills basing on their modularity was plotted following this clustering, and the social network has been replotted based on partial correlations in the model. According to greedy modularity detection and the dendrogram, the network could be divided into three groups based on skills' co-membership: the first regards skills concerning the communication area (written communication, verbal communication, internal-external communication); the second presents a melting between personal skills, project management, social media, and years experience; and the third includes the remaining skills that represent the essential skills for a candidate looking for a job in the marketing field. The weighted skills network based on partial correlations highlighted a strong correlation between the communication skills area and several negative correlations between hard skills (e.g., project management and best practices, written communication and years experience, verbal communication, and Microsoft Office). Results are coherent and align with the literature review (Savickas, 2005; Pool and Sewell, 2007; Harris and King, 2015; Makki et al., 2015; Alamelu et al., 2017; Nurlaela et al., 2017; Oviawe et al., 2017; Woya, 2019; Di Gregorio et al., 2019).

Comparing centrality measures in the skills network, the most between skills in the set were project management (8.3%), communication skills (8%),

learning ability (2.3%), best practices (1.6%), and written communication (0.3%). The closest skills were learning ability (18%), communication skills (16.5%), project management (16.2%), internal-external communication (14.7%), and written communication (1.5%). Based on their centrality measures, the best-fitting clustering method for the skills network was the Barrat one (Barrat et al., 2008).

Responding to RQ5, a proper simulation of a job interview in the marketing field is given by the MCMC text generation simulation reported in the results section. Having t_0 =social (having social media as the most frequent bigram and keeping the text split to perform MCMC), the job interview could ideally start by discussing social media competencies, move to the years of experience of a candidate, and then continue by talking about marketing-specific skills. At this point, the interview could transition to questions regarding digital and learning ability or move to communication-related skills and the educational attainments of the candidate. Investigating deeply regarding communication-related skills, an HR manager could ask questions regarding the candidate's ability with Microsoft Office, best practices, and attention to detail. Finally, the interview would ideally end by discussing the candidate's work experience, specialized experience, and related fields. With this format, the simulation is in line with the literature review (Savickas, 2005; Pool and Sewell, 2007; Harris and King, 2015; Makki et al., 2015;

Alamelu et al., 2017; Nurlaela et al., 2017; Oviawe et al., 2017; Woya, 2019; Di Gregorio et al., 2019).

5.2 Accounting & finance

The accounting industry's results enable the exploration of the most requested skills for the open job vacancies of actuary, external auditor, forensic accountant, private banker, and stockbroker. Almost 10k ads were extracted with a low standard deviation, ensuring reliability. The US state with the highest concentration of job ads was New York.

Responding to RQ1, results from TM and stylometry show that the most requested skills by firms operating in the accounting sector, or having an accounting division, were qualified experience, financial reporting, human resources, accounting and finance, and attention to detail. These results align with the literature review (Rosenberg et al., 2012; Lucianelli & Citro, 2018; Needham & Papier, 2018; Nisha & Rajasekaran, 2018; Ojanaperä et al., 2018; Clayton & Harris, 2019; Osmani et al., 2019; Imene, 2020; Cernusca, 2020).

Also, in the accounting industry, both soft and hard skills emerged regarding candidates' employability (Wolf and Archer, 2013; Alamelu et al., 2017, Entika, 2017; Hirsch, 2017; Mahfud et al., 2017; Woya, 2019).

As discussed in the systematic literature review, the results are an expression of the skills' redefinition regarding accountancy professionals (Hall, 2002; Lucianelli & Citro; Imene, 2020; Cernusca, 2020) when considering soft skills like attention to details, communication skills, and personal skills (Savickas, 2005; Pool and Sewell, 2007; Mishra, 2014; De Fruyt et al., 2015; Hirsch, 2017).

Hard skills are predominant in the accounting set, including time management, financial reporting, business administration, specialized experience, and Microsoft Office. Thus, according to the literature review, there is still great emphasis placed upon technical skills (Mishra, 2014; Harris and King, 2015; Cimatti, 2016; Alamelu et al., 2017; Oviawe et al., 2017; Woya, 2019).

There is also a melting between hard, soft, and managerial skills in this skillset, according to the literature review (Wolf and Archer, 2013; Mishra, 2014; Nurlaela et al., 2017; Hirsch, 2017; Petrovski et al., 2017; Chaibate et al., 2020; Ijaola et al., 2020).

Moreover, responding to RQ2, results highlighted four thematic areas regarding the skillset that an accountant should possess. (1) The first thematic area includes human resources, bachelor's degree, accounting experience, duties responsibilities, years experience, time management, industry knowledge, accounting, and auditing. This thematic area presents a strong

presence of hard skills that are predominant with respect to soft skills, notwithstanding those present in the area (Hirsch, 2017; Petrovski et al., 2017; Lucianelli & Citro, 2018; Cernusca, 2020; Chaibate et al., 2020; Ijaola et al., 2020; Imene, 2020). (2) The second topic presents, on the one hand, skills such as accounting and finance, problem-solving, knowledge skills, personal skills, and, on the other hand, skills such as qualified experience, accounting experience, and years of experience. According to the literature, it is possible to state that, in this topic, both soft skills and candidates' experience assume a pivotal role (Hirsch, 2017; Petrovski et al., 2017; Lucianelli & Citro, 2018; Cernusca, 2020; Chaibate et al., 2020; Ijaola et al., 2020; Imene, 2020). (3) The third topic highlights skills including specialized experience, qualified experience, and accounting experience. Other relevant skills appear in human resources, Microsoft Office, spreadsheet proficiency, financial reporting, and accounting degree. In this topic, there is a mix of both hard skills and candidates' experience. These findings are aligned to the literature review (Hirsch, 2017; Petrovski et al., 2017; Lucianelli & Citro, 2018; Cernusca, 2020; Chaibate et al., 2020; Ijaola et al., 2020; Imene, 2020). This topic represents the most requested skills for a candidate in the accounting and finance industry for an open position at the primary employment stage (Mishra, 2014; Harris and King, 2015; Cimatti, 2016; Alamelu et al., 2017; Chen, 2017; Oviawe et al., 2017; Woya, 2019).

(4) The fourth thematic area clusters skills such as communication skills, teamwork, management experience, business administration, and management experience. The presence of this skills typing highlighted the pivotal role of industry-specific and job-specific skills fundamental for candidates who are going to be employed in the accounting field (Hirsch, 2017; Petrovski et al., 2017; Lucianelli & Citro, 2018; Cernusca, 2020; Chaibate et al., 2020; Ijaola et al., 2020; Imene, 2020).

Nevertheless, responding to RQ3, results from the correlation analysis highlighted several medium or strong correlations within the skillset. The most significant correlations were found between specialized experience and years experience ($r=.66$), years experience and spreadsheet proficiency ($r=.64$), Microsoft Office and spreadsheet proficiency ($r=.61$), years experience and Microsoft Office ($r=.58$), a degree in accounting and financial reporting ($r=.57$), a degree in accounting and business administration ($r=.55$), and between financial reporting and business administration ($r=.53$). In sum, the most correlated skills pertain to the experience area and to hard skills; only one medium correlation regards soft skills (attention to detail ~ bachelor's degree, where $r=.55$). Results from the correlations analysis align with the literature review (Mishra, 2014; Harris and King, 2015; Cimatti, 2016; Alamelu et al., 2017; Chen, 2017; Oviawe et al., 2017; Lucianelli & Citro; Woya, 2019; Cernusca, 2020; Imene, 2020).

Responding to RQ4, various graphical methods were tested in configuring skills as a social network. Greedy, spectral, and optimal modularity were detected and compared, building-specific indicators to evaluate their performance. The best-fitting indicator was $\xi_G=.98$, representing the performance of greedy modularity detection, which was found almost with a perfect model fit. Indicators to evaluate spectral modularity and optimal modularity performance were relatively lower than the greedy one, even if good performing ($\xi_S=.91$, $\xi_O=.94$). Based on this observation, the ultimate dendrogram to group skills based on their modularity was plotted following this clustering, and the social network has been replotted based on partial correlations in the model. According to greedy modularity detection and the dendrogram, the network could be divided into three groups based on skills' co-membership: the first involves skills regarding the experience, the use of digital instruments, and communication skills, representing a combination of soft and hard skills and experience; the second group presents skills such as management experience, years experience, financial reporting, a degree in accounting, business administration, specialized experience, and time management—this group pertains to industry and job-specific skills and their related field of experience; the third group clusters the remaining skills representing the essential skills for a candidate looking for a job in the accounting and finance field. The weighted skills network based on partial correlations highlighted a strong correlation between communication skills

and job-specific experience (accounting experience, qualified experience) and between spreadsheet proficiency, years experience, specialized experience, and Microsoft Office. Other positive correlations are found between bachelor's degree, knowledge skills, and attention to detail. Negative correlations are found between specialized experience and Microsoft Office, time management, and years of experience. Results are in line and coherent with the literature review (Savickas, 2005; Pool and Sewell, 2007; Harris and King, 2015; Makki et al., 2015; Alamelu et al., 2017; Chen, 2017; Nurlaela et al., 2017; Oviawe et al., 2017; Lucianelli & Citro, 2018; Woya, 2019; Imene, 2020; Cernusca, 2020).

Regarding centrality measures in the skills network, the most between skills in the set were accounting finance (8.9%), financial management (7.1%), financial reporting (5.3%), duties responsibilities (3.1%), and teamwork (1.1%). The closest skills were accounting finance (22%), teamwork (20%), financial management (16.2%), time management (18.5%), and duties responsibilities (17.5%). Based on their centrality measures, the most coherent clustering method for the skills network was the Barrat one (Barrat et al., 2008).

Responding to RQ5, the MCMC text generation simulation, reported in the results section, provided a proper simulation of a job interview in accounting. Having t_0 =qualified (because the qualified experience was the most frequent bigram, and because of MCMC's computational necessity), the HR manager

could ideally start with questions concerning the qualified experience of the candidate, and then move to financial reporting, human resources, and attention to detail. The interview could then turn to the candidate's knowledge of accounting and finance, spreadsheet proficiency, and time management. The candidate's experience is now again demanded, with questions about the candidate's years of experience and any specialized experience, before shifting to communication and personal skills. The final questions could cover the candidate's industry knowledge starting from the experience matured during his bachelor's degree.

5.3 Data Science

Results from the data science field highlighted the most requested skills for the open job vacancies of data scientist, cybersecurity specialist, blockchain specialist, big data manager, and growth hacker. Almost 9.5k ads were extracted with a low standard deviation, ensuring reliability. The US state with the highest concentration of job ads was California.

Responding to RQ1, results from TM and stylometry show that the most requested skills by firms operating in the data science sector, or having a data science division, were years experience, computer science, bachelor's degree, and communication skills. Such results align with the literature review (Degryse, 2016; Pieterse & Eekelen, 2016; Hollister et al., 2017;

Minocha & Tudor, 2017; Misra & Khurana, 2017; Nisha & Rajasekaran, 2018; Ojanaperä et al., 2018; Osmani et al., 2019; Pejich-Bach & Kristić, 2019).

Also, in the data science industry, both soft and hard skills emerged regarding candidates' employability (Wolf and Archer, 2013; Alamelu et al., 2017, Entika, 2017; Hirsch, 2017; Mahfud et al., 2017; Woya, 2019).

As demonstrated in the systematic literature review, results are an expression of the skills that a candidate should possess to be employed for an open data science job (Degryse, 2016; Pieterse & Eekelen, 2016; Hollister et al., 2017; Minocha & Tudor, 2017; Misra & Khurana, 2017; Pejich-Bach & Kristić, 2019), and these include soft skills like teamwork, communication skills, and problem-solving (Savickas, 2005; Pool and Sewell, 2007; Mishra, 2014; De Fruyt et al., 2015; Hirsch, 2017).

Hard skills are also predominant in the data science set, including computer science, data science, data analysis, software development, big data, system missing, data management, data analytics, record tracking, and computer vision. Thus, technical skills play a pivotal role in the data science industry, according to the literature review (Degryse, 2016; Pieterse & Eekelen, 2016; Hollister et al., 2017; Minocha & Tudor, 2017; Misra & Khurana, 2017; Nisha & Rajasekaran, 2018; Ojanaperä et al., 2018; Osmani et al., 2019; Pejich-Bach & Kristić, 2019).

Moving deeper and answering RQ2, results highlighted four topics concerning the skillset that a data scientist should possess. (1) The first thematic area includes teamwork, best practices, working experience, data experience, software development, data analysis, and data analytics. This topic presents skills concerning the areas of soft skills, industry and job-specific skills, and experience (Degryse, 2016; Pieterse & Eekelen, 2016; Hollister et al., 2017; Minocha & Tudor, 2017; Misra & Khurana, 2017; Nisha & Rajasekaran, 2018; Ojanaperä et al., 2018; Osmani et al., 2019; Pejich-Bach & Kristić, 2019). (2) The second topic presents a joint of soft skills, hard skills, educational and working experience (e.g., communication skills, problem-solving, years experience, record tracking, computer science, computer science, master's degree), which aligns the topic with the literature review (Degryse, 2016; Pieterse & Eekelen, 2016; Hollister et al., 2017; Minocha & Tudor, 2017; Misra & Khurana, 2017; Pejich-Bach & Kristić, 2019; Chaibate et al., 2020; Ijaola et al., 2020). (3) The third topic group includes coding experience, computer science, data analysis, computer vision, specialized experience, working experience, and data management. In this topic, both hard skills and candidates' experience are mentioned, maintaining an industry-specific emphasis with the present job-specific skills, according to the literature review (Degryse, 2016; Pieterse & Eekelen, 2016; Hollister et al., 2017; Minocha & Tudor, 2017; Misra & Khurana, 2017; Nisha & Rajasekaran, 2018; Ojanaperä et al., 2018; Osmani et al.,

2019; Pejich-Bach & Kristć, 2019). (4) The fourth thematic area includes skills such as system missing, knowledge skills, big data, data management, data analysis, data analytics, problem-solving, a degree in computer science, data experience, and working experience. This topic represents the most requested skills for a candidate for a job vacancy in the data science field (Mishra, 2014; Harris and King, 2015; Cimatti, 2016; Hollister et al., 2017; Minocha & Tudor, 2017; Misra & Khurana, 2017; Nisha & Rajasekaran, 2018; Ojanaperä et al., 2018; Osmani et al., 2019; Pejich-Bach & Kristć, 2019; Woya, 2019).

Responding to RQ3, results from the correlation analysis showed several medium correlations between skills. The most significant correlations were found between skills abilities and bachelor's degree ($r=.57$) and between bachelor's degree and teamwork ($r=.51$). Other correlations are under the threshold of $r=.5$. Thus, results from the correlations analysis are aligned with the literature review (Hollister et al., 2017; Minocha & Tudor, 2017; Misra & Khurana, 2017; Nisha & Rajasekaran, 2018; Ojanaperä et al., 2018; Osmani et al., 2019; Pejich-Bach & Kristć, 2019).

Responding to RQ4, several approaches have been implemented in visualizing skills as a social network. Greedy, spectral, and optimal modularity were tested and evaluated, building specific indicators to compare their performance. The best indicator for community detection was $\xi_G=.97$, representing the performance of greedy modularity detection, which was

found almost with a perfect model fit. On the other hand, indicators to evaluate spectral modularity and optimal modularity performance were relatively lower than the greedy one, even if good performing ($\xi_s=.88$, $\xi_o=.96$). Having these findings, the final dendrogram to group skills based on their modularity was plotted following this clustering, and the social network has been replotted based on partial correlations in the model. According to greedy modularity detection and the dendrogram, the network could be divided into two groups based on skills' co-membership. The first cluster includes skills such as best practices, bachelor's degree, skills abilities, teamwork, data analysis, system missing, knowledge skills, big data, and data management. This group pertains to soft skills, industry and job-specific skills, education, and working experience. The other group clusters the remaining skills representing the essential skills for a candidate looking for a job in the data science field. The weighted skills network based on partial correlations reports strong correlations between bachelor's degree, skills, abilities, communication skills, years experience and between system missing, data analysis, teamwork, and knowledge skills. Results are in line and coherent with the literature review (Hollister et al., 2017; Minocha & Tudor, 2017; Misra & Khurana, 2017; Nisha & Rajasekaran, 2018; Ojanaperä et al., 2018; Osmani et al., 2019; Pejich-Bach & Kristó, 2019).

Discussing centrality measures in the skills network, the most between skills in the set were big data (67.5%), specialized experience (51.38%), teamwork

(7.3%), knowledge skills (7.1%), and data science (4.3%). The closest skills were big data (48.3%), specialized experience (48.8%), data science (29.9%), computer vision (28.36%), and coding experience (27.2%). The most coherent clustering method for the skills network, based on their centrality measures, was the Barrat one (Barrat et al., 2008).

Responding to RQ5, the MCMC text generation simulation, reported in the results section, provided a proper simulation of a job interview in the data science field. Having t_0 =years (because years experience was the most frequent bigram, and because of MCMC's computational necessity), candidates for an open position in the data science industry could be questioned concerning their years experience, with a focus on their coding experience, and then move on to record tracking, computer vision, and computer science. The interview could continue questioning candidates' bachelor's degree, communication skills, best practices, and data science. Other questions could concern candidates' problem-solving abilities, their master's degree, and their ability to work in a team.

5.4 Bioinformatics

Results from the Bioinformatics field highlight the crucial skills for candidates seeking a job as a bioinformatician, computational biologist, genomics programmer, pharmacovigilance specialist, and genome architect. Almost 10k were extracted with a low standard deviation, ensuring reliability. The US state with the highest concentration of job ads for the engineering sector was California, followed by Washington and Illinois.

Responding to RQ1, results from TM and stylometry report that the most requested skills by corporations operating in the bioinformatic field were years of experience, computational biology, data analysis, computer science, and working experience. These results align with the literature review (Jones, 2013; Degryse, 2016; Pieterse & Eekelen, 2016; Chen, 2017; Hollister et al., 2017; Lee & Janna, 2017; Minocha & Tudor, 2017; Misra & Khurana, 2017; Nisha & Rajasekaran, 2018; Selvam, 2018; Bongomin et al., 2019; Muhamad, 2019; Osmani et al., 2019; Pejich-Bach & Krist c, 2019).

Both soft and hard skills emerged regarding candidates' employability (Wolf and Archer, 2013; Degryse, 2016; Pieterse & Eekelen, 2016; Alamelu et al., 2017, Entika, 2017; Hirsch, 2017; Mahfud et al., 2017; Osmani et al., 2019; Pejich-Bach & Krist c, 2019; Woya, 2019)

Hard skills are predominant also in the abovementioned skills set, through skills such as computational biology, computational informatics, molecular

biology, sequencing data, biology lab, software development, record tracking, computational statistics, informatics tools, business intelligence, data generation, statistical methods, big data, and statistical analysis. Thus, technical skills play a pivotal role in the data science industry, according to the literature review (Jones, 2013; Degryse, 2016; Pieterse & Eekelen, 2016; Chen, 2017; Hollister et al., 2017; Lee & Janna, 2017; Minocha & Tudor, 2017; Misra & Khurana, 2017; Nisha & Rajasekaran, 2018; Selvam, 2018; Bongomin et al., 2019; Muhamad, 2019; Osmani et al., 2019; Pejich-Bach & Kristić, 2019).

Also, there is a combination of hard, soft, and job-specific skills in the bioinformatics skillset, according to the literature review (Wolf and Archer, 2013; Mishra, 2014; Nurlaela et al., 2017; Hirsch, 2017; Petrovski et al., 2017; Chaibate et al., 2020; Ijaola et al., 2020).

Continuing to RQ2, results highlighted four topics concerning the skillset that a bioinformatician should possess. (1) The first topic of the bioinformatics skillset includes skills such as working experience, genomic data, computational informatics, computer science, sequencing data, computational biology, informatics background, statistical methods, data generation. In addition, this topic presents skills concerning the areas of hard skills, industry and job-specific skills, and candidates' experience (Degryse, 2016; Pieterse & Eekelen, 2016; Hollister et al., 2017; Minocha & Tudor, 2017; Misra & Khurana, 2017; Nisha & Rajasekaran, 2018; Ojanaperä et al.,

2018; Osmani et al., 2019; Pejich-Bach & Kristć, 2019). (2) The second topic presents a joint of hard skills and job-specific skills (e.g., big data, statistical methods, genomic data, informatics background, sequencing data, computational statistics, molecular biology). The topic is in line with the literature review (Degryse, 2016; Pieterse & Eekelen, 2016; Hollister et al., 2017; Minocha & Tudor, 2017; Misra & Khurana, 2017; Pejich-Bach & Kristć, 2019; Chaibate et al., 2020; Ijaola et al., 2020). (3) The third thematic area clusters skills such as master's degree, computational biology, genomic data, statistical analysis, sequencing data, record tracking, communication skills, and biology lab. This topic highlights job-specific skills, technical skills, and communication skills, all of which are in line with findings from the literature review (Degryse, 2016; Pieterse & Eekelen, 2016; Hollister et al., 2017; Minocha & Tudor, 2017; Misra & Khurana, 2017; Nisha & Rajasekaran, 2018; Ojanaperä et al., 2018; Osmani et al., 2019; Pejich-Bach & Kristć, 2019). (4) The fourth thematic area includes biology lab, coding experience, statistical methods, software development, computational informatics, data generation, oral communication, and working experience, meaning it comprises communication skills, hard skills, and job-specific skills. This thematic area clearly shows the most requested skills for a candidate for a job vacancy in the bioinformatics field (Mishra, 2014; Cimatti, 2016; Degryse, 2016; Hollister et al., 2017; Minocha & Tudor, 2017;

Misra & Khurana, 2017; Nisha & Rajasekaran, 2018; Ojanaperä et al., 2018; Osmani et al., 2019; Pejich-Bach & Kristć, 2019).

Responding to RQ3, results from the correlation analysis showed several medium correlations between skills, both positive and negative. The most significant correlations were found between coding experience and biology lab ($r=.68$), between record tracking and computer science ($r=.68$), between sequencing data and computer science ($r=.59$), and between big data and computational informatics ($r=.53$). Other correlations are under the threshold of $r=.5$. Thus, results from the correlations analysis are aligned with the literature review (Jones, 2013; Degryse, 2016; Pieterse & Eekelen, 2016; Chen, 2017; Hollister et al., 2017; Lee & Janna, 2017; Minocha & Tudor, 2017; Misra & Khurana, 2017; Nisha & Rajasekaran, 2018; Selvam, 2018; Bongomin et al., 2019; Muhamad, 2019; Osmani et al., 2019; Pejich-Bach & Kristć, 2019).

Responding to RQ4, several methods have been employed in representing skills as a social network graph. Greedy, spectral, and optimal modularity were tested and evaluated, building specific indicators to compare their performance. The best indicator for community detection was $\xi_G=.99$, representing the performance of greedy modularity detection, which was found almost with a perfect model fit. Indicators to evaluate spectral modularity and optimal modularity performance were relatively lower than the greedy one, even if good performing ($\xi_S=.91$, $\xi_O=.98$). With these

findings, the final dendrogram to group skills based on their modularity was plotted following this clustering, and the social network has been replotted based on partial correlations in the model. According to greedy modularity detection and the dendrogram, the network could be divided into three groups based on skills' co-membership. The first group includes skills such as informatics tools, data generation, coding experience, biology lab, molecular biology, master's degree, years experience, communication skills, and genomic data. This group pertains to soft skills, experience, and job-specific skills. The second cluster includes computer science, record tracking, sequencing data, informatics background, software development, big data, computational biology, and working experience. The last group clusters the remaining skills representing the basic skills for a candidate looking for a job in the bioinformatics field. The weighted skills network based on partial correlations reports strong correlations between statistical methods and computer science, and coding experience, biology lab, and molecular biology. Results are in line and coherent with the literature review (Jones, 2013; Degryse, 2016; Pieterse & Eekelen, 2016; Chen, 2017; Hollister et al., 2017; Lee & Janna, 2017; Minocha & Tudor, 2017; Misra & Khurana, 2017; Nisha & Rajasekaran, 2018; Selvam, 2018; Bongomin et al., 2019; Muhamad, 2019; Osmani et al., 2019; Pejich-Bach & Kristé, 2019).

Discussing centrality measures in the skills network, the most between skills in the set were informatic tools (39.5%), years experience (32%), record

tracking (31.6%), coding experience (21.3%), and molecular biology (17.7%). The closest skills were years experience (53.1%), informatics tools (52.5%), molecular biology (37.2%), coding experience (32.7%), and data analysis (30.3%). The most coherent clustering method for the skills network, based on their centrality measures, was the Zhang one (Zhang, 1996; Zhang,1997; Zhang, 2006).

Responding to RQ5, the MCMC text generation simulation, reported in the results section, provided a proper simulation of a job interview in bioinformatics. Having t_0 =years (because years experience was the most frequent bigram, and because of MCMC's computational necessity), the director of the job interview could question candidates regarding their years experience and continue with inquiries about their skills in computational biology, their master's degree, their ability in sequencing data, and their knowledge of statistical methods and big data. Following, interviewers might query candidates about the process of data generation could be demanded of candidates, together with their skills in data analysis, computer science, statistical analysis, and their working experience. Finally, the interviewer could question candidates regarding their knowledge of a biology lab, software development, record tracking, computational informatics, and informatics tools. Throughout the interview, candidates could also be questioned regarding their background in informatics, business intelligence, and coding experience.

5.5 Software Engineering and Cloud Computing (SECC)

Results from the SECC sector highlighted the most requested skills for open positions as a mechatronic, software engineer, development operations engineer, full stack developer, and additive manufacturing engineer. Almost 9k ads were extracted with a low standard deviation, ensuring reliability. The US state with the highest concentration of job ads for the engineering sector was California, immediately followed by Michigan.

Responding to RQ1, results from TM and stylometry report that the most requested skills by corporations operating in the SECC industry were years experience, web services, cybersecurity, bachelor's degree, and Microsoft Office. These results align with the literature review (Jones, 2013; Degryse, 2016; Pieterse & Eekelen, 2016; Chen, 2017; Hollister et al., 2017; Lee & Janna, 2017; Minocha & Tudor, 2017; Misra & Khurana, 2017; Nisha & Rajasekaran, 2018; Selvam, 2018; Bongomin et al., 2019; Muhamad, 2019; Osmani et al., 2019; Pejich-Bach & Kristć, 2019).

Concerning the SECC industry, hard skills and candidates' experience emerged as the most requested prerequisites (Wolf and Archer, 2013; Alamelu et al., 2017, Entika, 2017; Hirsch, 2017; Mahfud et al., 2017; Woya, 2019).

Results are an expression of the skills that a candidate should possess to be employed in the SECC industry (Degryse, 2016; Pieterse & Eekelen, 2016;

Hollister et al., 2017; Minocha & Tudor, 2017; Misra & Khurana, 2017; Selvam, 2018; Pejich-Bach & Kristć, 2019; Bongomin et al., 2019; Muhamad, 2019; Osmani et al., 2019; Pejich-Bach & Kristć, 2019).

Hard skills are predominant in the abovementioned skills set, including cybersecurity, disaster recovery, programming skills, cloud computing, Linux programming, program missing, migrating data, software engineering, databases, and mechanical engineering. Thus, technical skills play a pivotal role in the SECC industry, according to the literature review (Jones, 2013; Degryse, 2016; Pieterse & Eekelen, 2016; Chen, 2017; Hollister et al., 2017; Lee & Janna, 2017; Minocha & Tudor, 2017; Misra & Khurana, 2017; Nisha & Rajasekaran, 2018; Selvam, 2018; Bongomin et al., 2019; Muhamad, 2019; Osmani et al., 2019; Pejich-Bach & Kristć, 2019).

Answering RQ2, results highlighted four topics concerning the skillset that a SECC candidate should possess. (1) The first topic of the SECC skillset includes teamwork, communication skills, software engineering, databases, disaster recovery, cybersecurity, school diploma, bachelor's and master's degree, and programming skills. This topic primarily presents skills concerning the areas of industry and job-specific skills and candidates' education (Degryse, 2016; Pieterse & Eekelen, 2016; Hollister et al., 2017; Minocha & Tudor, 2017; Misra & Khurana, 2017; Nisha & Rajasekaran, 2018; Ojanaperä et al., 2018; Osmani et al., 2019; Pejich-Bach & Kristć, 2019). (2) The second topic presents a melting of soft and hard skills (e.g.,

communication skills, knowledge skills, teamwork, Microsoft Office, software engineering, mechanical engineering, Linux programming, Word, and Excel). The topic is in line with the literature review (Degryse, 2016; Pieterse & Eekelen, 2016; Hollister et al., 2017; Minocha & Tudor, 2017; Misra & Khurana, 2017; Pejich-Bach & Krist c, 2019; Chaibate et al., 2020; Ijaola et al., 2020). (3) The third thematic area clusters skills such as teamwork, knowledge skills, cloud computing, web services, cybersecurity, years experience, and qualified experience. This topic highlights job-specific skills, technical skills, soft skills and experience, which aligns with the literature review (Degryse, 2016; Pieterse & Eekelen, 2016; Hollister et al., 2017; Minocha & Tudor, 2017; Misra & Khurana, 2017; Nisha & Rajasekaran, 2018; Ojanaper  et al., 2018; Osmani et al., 2019; Pejich-Bach & Krist c, 2019). (4) The fourth thematic area includes qualified experience, working experience, teamwork, master's degree, Word, and Excel. Including databases, cloud computing, program missing, computer vision, and disaster recovery, this thematic area clearly shows the most demanded skills to be employed in the SECC industry (Mishra, 2014; Cimatti, 2016; Degryse, 2016; Chen, 2017; Hollister et al., 2017; Lee & Janna, 2017; Minocha & Tudor, 2017; Misra & Khurana, 2017; Nisha & Rajasekaran, 2018; Ojanaper  et al., 2018; Osmani et al., 2019; Pejich-Bach & Krist c, 2019).

Furthermore, responding to RQ3, results from the correlation analysis showed that every correlation for this sector is under the threshold of $r=0.5$. A

justification for this kind of results in the model can be found in two reasons: (a) the SECC industry includes jobs requiring a high tax of specialization, and so the skillset presents many small groups of positively correlated skills, and, (b) due to this consideration, in-depth investigation with text mining revealed that the content of job ads presented an overall high sparsity rate²⁴ (sparsity=.99), affecting correlations' values also after the implementation of an algorithm for the sparsity reduction.

Responding to RQ4, several methods have been employed in representing skills as a social network graph. Greedy, spectral, and optimal modularity were tested and evaluated, building specific indicators to compare their performance. The best indicator for community detection was $\xi_G=.98$, representing greedy modularity detection performance, which was found almost with a perfect model fit. Indicators to evaluate spectral modularity and optimal modularity performance were relatively lower than the greedy one, even if good performing ($\xi_S=.93$, $\xi_O=.96$). With these findings, the final dendrogram to group skills based on their modularity was plotted following this clustering, and the social network has been replotted based on partial correlations in the model. According to greedy modularity detection and the dendrogram, the network could be divided into two groups based on skills' co-membership. The first group includes skills such as computer vision, mechanical engineering, databases, cybersecurity, communication skills,

²⁴ Sparsity, in text mining, indicates several cells with zero entries in large matrices.

programming skills, school diploma, migrating data, qualified experience, disaster recovery, Linux programming. This group contains hard and communication skills. The second cluster groups the other skills representing the basic skills for a candidate looking for a job in the SECC industry. The weighted skills network based on partial correlations reports strong correlations between communication skills, programming skills, and disaster recovery and between Linux programming and qualified experience. Other significant correlations are found between cybersecurity, years of experience, and databases and between school diplomas and migrating data. A significant negative correlation is found between databases and a school diploma. Results are in line and coherent with the literature review (Jones, 2013; Degryse, 2016; Pieterse & Eekelen, 2016; Chen, 2017; Hollister et al., 2017; Lee & Janna, 2017; Minocha & Tudor, 2017; Misra & Khurana, 2017; Nisha & Rajasekaran, 2018; Selvam, 2018; Bongomin et al., 2019; Muhamad, 2019; Osmani et al., 2019; Pejich-Bach & Krist c, 2019).

Discussing centrality measures in the skills network, the most between skills in the set were program missing (71.1%), years experience (22.9%), school diploma (20.5%), computer vision (19.3%), and Linux programming (13.04%). The closest skills were program missing (18%), Microsoft Office (16.5%), computer vision (16.2%), web services (35.3%), and written communication (34.3%). The most coherent clustering method for the skills

network, based on their centrality measures, was the Zhang one (Zhang, 1996; Zhang,1997; Zhang, 2006).

Responding to RQ5, the MCMC text generation simulation, reported in the results section, provided a proper simulation of a job interview in bioinformatics. Having t_0 =years (because years experience was the most frequent bigram, and because of MCMC's computational necessity), candidates for an open position in the SECC industry would likely be questioned concerning their years experience and their skills in web services and cybersecurity, their bachelor's degree, their knowledge of Word and Excel, computer vision, and mechanical engineering. After that, disaster recovery abilities and the possession of a school diploma could be queried of candidates, together with their problem-solving skills, their ability to work in a team, and if they have qualified experience. The interview could move to questioning candidates' knowledge of cloud computing, Linux programming, and programming skills, migrating data, having skills regarding databases, and software engineering. Candidates could even be questioned regarding their master's degree, knowledge of Microsoft Office and programming, program missing, and communication skills.

5.6 Tourism Management

The tourism management sector highlighted the most requested skills for open positions as a destination manager, tourism business consultant, digital event manager, reservations manager, and arts & hospitality manager. Almost 11k ads were extracted with a low standard deviation, ensuring reliability. The US states with the highest concentration of job ads for the tourism sector were Florida, followed by California and Nevada.

Responding to RQ1, results from TM and stylometry report that the most requested skills by corporations operating in the tourism industry were front desk, customer service, attentive friendly, hotel management, and serving guests. Such results coincide with the literature review (Jones, 2013; Finch et al., 2016; Lee & Janna, 2017; Metilda & Neena, 2017; Adeyinka-Ojo, 2018; Bruun & Duka, 2018; Nisha & Rajasekaran, 2018; Ojanaperä et al., 2018; Clayton & Harris, 2019; Osmani et al., 2019).

Concerning the tourism industry, hard skills and job-specific skills emerged as the most requested prerequisites (Wolf and Archer, 2013; Alamelu et al., 2017, Entika, 2017; Hirsch, 2017; Mahfud et al., 2017; Adeyinka-Ojo, 2018; Woya, 2019).

Hard skills are predominant in the abovementioned skills set, even if strongly melted with soft skills. Due to this consideration, it is the case to talk regarding industry-specific skills, having the presence of skills such as front

desk, customer service, hotel management, luxury brands, growing portfolio, and destination management. Thus, industry-specific skills play a pivotal role in the tourism industry, according to the literature review (Jones, 2013; Finch et al., 2016; Lee & Janna, 2017; Metilda & Neena, 2017; Adeyinka-Ojo, 2018; Bruun & Duka, 2018; Nisha & Rajasekaran, 2018; Ojanaperä et al., 2018; Clayton & Harris, 2019; Osmani et al., 2019).

Moreover, answering RQ2, the results highlighted four topics concerning the skillset that a tourism manager should possess. (1) The first topic of the tourism skillset includes skills such as time attention, communication skills, years of experience, friendly passion, destination support, helpful courteous, multiple languages, customer service, front desk, and attentive and friendly. This topic presents skills concerning the areas of industry and job-specific skills and soft skills (Jones, 2013; Finch et al., 2016; Lee & Janna, 2017; Metilda & Neena, 2017; Adeyinka-Ojo, 2018; Bruun & Duka, 2018; Nisha & Rajasekaran, 2018; Ojanaperä et al., 2018; Clayton & Harris, 2019; Osmani et al., 2019). (2) The second topic presents a melting of soft and hard skills (e.g., destination management, communication skills, destination support, time attentive, front desk, serving guests, hotels destination, team leader). The topic is in line with the literature review (Jones, 2013; Finch et al., 2016; Lee & Janna, 2017; Metilda & Neena, 2017; Adeyinka-Ojo, 2018; Bruun & Duka, 2018; Nisha & Rajasekaran, 2018; Ojanaperä et al., 2018; Clayton & Harris, 2019; Osmani et al., 2019). (3) The third thematic area

clusters skills such as cultural passion, hotels destination, luxury brands, customer service, front desk, time attentive, and destination support. This topic highlights job-specific and technical skills, making it in line with the literature review (Jones, 2013; Finch et al., 2016; Lee & Janna, 2017; Metilda & Neena, 2017; Adeyinka-Ojo, 2018; Bruun & Duka, 2018; Nisha & Rajasekaran, 2018; Ojanaperä et al., 2018; Clayton & Harris, 2019; Osmani et al., 2019). (4) The fourth thematic area includes skills such as luxury brands, cultural passion, helpful courteous, guest manager, time attentive, multiple languages, guests driving, and destination management. This topic highlights the most demanded skills to be employed in the tourism industry (Jones, 2013; Finch et al., 2016; Lee & Janna, 2017; Metilda & Neena, 2017; Adeyinka-Ojo, 2018; Bruun & Duka, 2018; Nisha & Rajasekaran, 2018; Ojanaperä et al., 2018; Clayton & Harris, 2019; Osmani et al., 2019).

Furthermore, responding to RQ3, results from the correlation analysis showed several correlations between skills, both positive and negative. The most significant correlations were found between time attentive and assistance serving ($r=.55$), helpful, courteous, and years of experience ($r=.52$), and between cultures passionate and luxury brands ($r=.5$). Other correlations are under the threshold of $r=.5$. Results from the correlations analysis are aligned with the literature review (Jones, 2013; Finch et al., 2016; Lee & Janna, 2017; Metilda & Neena, 2017; Adeyinka-Ojo, 2018;

Bruun & Duka, 2018; Nisha & Rajasekaran, 2018; Ojanaperä et al., 2018; Clayton & Harris, 2019; Osmani et al., 2019).

Responding to RQ4, several methods have been implemented in visualizing skills as a social network graph. Greedy, spectral, and optimal modularity were tested and evaluated, building specific indicators to compare their performance. The best indicator for community detection was $\xi_G=.99$, representing the performance of greedy modularity detection, which was found almost with a perfect model fit. Indicators to evaluate spectral modularity and optimal modularity performance were relatively lower than the greedy one, even if good performing ($\xi_S=.94$, $\xi_O=.97$). The final dendrogram to group skills based on their modularity was plotted following this clustering, and the social network has been replotted based on partial correlations in the model. According to greedy modularity detection and the dendrogram, the network could be clustered into three groups based on skills' co-membership. The first group includes skills such as team leader, hotels destination, attentive friendly, serving guests, and courteous to guests. This group pertains to soft skills and job-specific skills. The second cluster group includes hotel management, luxury brands, everyone's culture, guest manager, friendly passionate, growing portfolio, guests driving, helpful, friendly, and front desk. The group pertains to communication skills, soft skills, and job-specific skills. The last group includes skills representing the common skills for a candidate looking for a job in the tourism industry. The

weighted skills network based on partial correlations reports strong correlations between serving guests, helpful and courteous, and years of experience. Other significant correlations are found between assistance serving, time attentive, destination supports, and hotel management. Results are in line and coherent with the literature review (Jones, 2013; Finch et al., 2016; Lee & Janna, 2017; Metilda & Neena, 2017; Adeyinka-Ojo, 2018; Bruun & Duka, 2018; Nisha & Rajasekaran, 2018; Ojanaperä et al., 2018; Clayton & Harris, 2019; Osmani et al., 2019).

Discussing centrality measures in the skills network, the most between skills in the set were serving guests (68.3%), time attentive (57.1%), years experience (38.6%), customer service (22%), and multiple languages (19.9%). The closest skills were years experience (81%), customer service (73.7%), serving guests (63.4%), attentive friendly (56.2%), and time attentive (48%). The most coherent clustering method for the skills network, based on their centrality measures, was the Barrat one (Barrat et al., 2008).

Responding to RQ5, the MCMC text generation simulation, reported in the results section, provided a proper simulation of a job interview in bioinformatics. Having t_0 =front (because the front desk was the most frequent bigram, and because of MCMC's computational necessity), an HR manager from the tourism industry could question candidates regarding their front-desk skills, customer service, and if they are attentive and friendly in their hotel management. These questions could follow with ones about their

skills in destination support, their knowledge of multiple languages, and their communication skills. The interview would ideally move on to the availability of global assistance, the candidates' years of experience, their attitude in being a team leader, and time attention.

5.7 Psychology

Results from the psychology field highlighted the most requested skills for open positions as a psychologist, psychoanalyst, work psychologist, mental health counsellor, and corporate counsellor. Almost 8.5k ads were extracted with a low standard deviation, ensuring reliability. The US states with the highest concentration of job ads for the psychology sector were California and New York.

Responding to RQ1, results from TM and stylometry report that the most requested skills by corporate psychologists were master's degree, social work, individual group, years experience, and treatment plan, and these results are aligned with the literature review (Jones, 2013; Vanhercke et al.; 2014; Finch et al., 2016; Lee & Janna, 2017; Metilda & Neena, 2017; Bruun & Duka, 2018; Nisha & Rajasekaran, 2018; Ojanaperä et al, 2018; Selvam, 2018; Clayton & Harris, 2019; Osmani et al., 2019).

Concerning the psychology field, soft skills and job-specific skills emerged as the most requested prerequisites (Wolf and Archer, 2013; Alamelu et al.,

2017, Entika, 2017; Hirsch, 2017; Mahfud et al., 2017; Adeyinka-Ojo, 2018; Woya, 2019).

Hard skills and job-specific skills are not predominant in the psychology skills set, demanding skills such as treatment plans, crisis intervention, case management, primary care, clinical social, mental services, patient care, children adolescent, behavioral services, and clinical services. Thus, technical skills play a pivotal role in the case of psychology professionals. Findings are aligned to the literature review (Jones, 2013; Vanhercke et al.; 2014; Finch et al., 2016; Lee & Janna, 2017; Metilda & Neena, 2017; Bruun & Duka, 2018; Nisha & Rajasekaran, 2018; Ojanaperä et al., 2018; Selvam, 2018; Clayton & Harris, 2019; Osmani et al., 2019).

Answering RQ2, the results highlighted four topics concerning the skillset that a someone in the psychology field should possess. (1) The first topic of the psychology skillset includes skills such as behavioral services, individual group, case management, children adolescent, years experience, social worker, clinical services, marriage and family, clinical and social, and patient care. This topic presents job-specific skills and soft skills (Jones, 2013; Vanhercke et al.; 2014; Finch et al., 2016; Lee & Janna, 2017; Metilda & Neena, 2017; Bruun & Duka, 2018; Nisha & Rajasekaran, 2018; Ojanaperä et al., 2018; Selvam, 2018; Clayton & Harris, 2019; Osmani et al., 2019). (2) The second topic presents job-specific skills and educational prerequisites (e.g., psychology, social work, clinical, social, individual and group, master's

degree, treatment plans, clinical services, behavioral services). The topic is in line with the literature review (Jones, 2013; Vanhercke et al.; 2014; Finch et al., 2016; Lee & Janna, 2017; Metilda & Neena, 2017; Bruun & Duka, 2018; Nisha & Rajasekaran, 2018; Ojanaperä et al., 2018; Selvam, 2018; Clayton & Harris, 2019; Osmani et al., 2019). (3) The third topic includes social workers, behavioral services, master's degree, working experience, special education, and clinical license. This topic highlights the pivotal role of experience and educational attainments, which is in line with the literature review (Jones, 2013; Vanhercke et al.; 2014; Finch et al., 2016; Lee & Janna, 2017; Metilda & Neena, 2017; Bruun & Duka, 2018; Nisha & Rajasekaran, 2018; Ojanaperä et al., 2018; Selvam, 2018; Clayton & Harris, 2019; Osmani et al., 2019). (4) The fourth thematic area includes mental services, behavioral services, family therapy, case management, mental illness, marriage, and family clinical supervision. This topic highlights the most demanded skills to be employed in the psychology field (Jones, 2013; Vanhercke et al.; 2014; Finch et al., 2016; Lee & Janna, 2017; Metilda & Neena, 2017; Bruun & Duka, 2018; Nisha & Rajasekaran, 2018; Ojanaperä et al., 2018; Selvam, 2018; Clayton & Harris, 2019; Osmani et al., 2019).

Furthermore, responding to RQ3, results from the correlation analysis showed sparse low correlations, indicating a sort of independency between skills in the psychology skillset.

Responding to RQ4, several methods have been developed to represent skills as a social network graph. Greedy, spectral, and optimal modularity were tested and evaluated, building specific indicators to compare their performance. The best indicator for community detection was $\xi_G=.98$, representing the performance of greedy modularity detection, which was found almost with a perfect model fit. Indicators to evaluate spectral modularity and optimal modularity performance were relatively lower than the greedy one, even if good performing ($\xi_S=.95$, $\xi_O=.96$). The final dendrogram to group skills based on their modularity was plotted following this clustering, and the social network has been replotted based on partial correlations in the model. According to greedy modularity detection and the dendrogram, the network could be clustered into groups based on skills' co-membership. The first group includes skills such as treatment plan, mental illness, clinical supervision, children adolescent, treatment plans, case management, mental services, crisis intervention, special education, and social work. The other cluster includes skills representing the essential skills for a candidate looking for a job in the psychology industry. The weighted skills network based on partial correlations reports strong correlations between master's degree, behavioral services, working experience, social worker, psychology, clinical, social, and clinical licensed. Results are in line and coherent with the literature review (Jones, 2013; Finch et al., 2016; Lee & Janna, 2017; Metilda & Neena, 2017; Adeyinka-Ojo, 2018; Bruun &

Duka, 2018; Nisha & Rajasekaran, 2018; Ojanaperä et al., 2018; Clayton & Harris, 2019; Osmani et al., 2019).

Discussing centrality measures in the skills network, the most between skills were children adolescent (53.3%), marriage family (34.3%), mental services (18.9%), clinical licensed (12.6%), and social worker (11%). The closest skills were mental services (48.8%), children adolescent (45.5%), crisis intervention (34%), clinical licensed (33.6%), and treatment plan (33.1%). The most coherent clustering method for the skills network, based on their centrality measures, was the Onnela one (Onnela et al., 2002; Onnela et al., 2004).

Responding to RQ5, the MCMC text generation simulation, reported in the results section, provided a proper simulation of a job interview in bioinformatics. Having t_0 =master (because master's degree was the most frequent bigram, and because of MCMC's computational necessity), candidates in the psychology field could be questioned concerning their master's degree, experience in social work, patient care, clinical licensure, experience with treatment plans, the establishment of crisis plans for intervention, and case management. The candidates could also be interrogated regarding clinical services for marriage and family therapy, clinical supervision, and a degree regarding social education.

5.8 Law

Results from the attorney sector highlighted the most requested skills for open positions as an attorney, corporate attorney, corporate lawyer, securities lawyer, and tax law attorney. Almost 9k ads were extracted with a low standard deviation, ensuring reliability. The US states with the highest concentration of job ads for the law sector were New York and California.

Responding to RQ1, TM and stylometry report that the most requested skills by corporations operating in the law industry were firm law, legal services, personal injury, legal research, and legal advice. Results are aligned with the literature review (Jones, 2013; Finch et al., 2016; Lee & Janna, 2017; Galloway, 2017; Metilda & Neena, 2017; Bruun & Duka, 2018; Nisha & Rajasekaran, 2018; Ojanaperä et al, 2018; Clayton & Harris, 2019; Osmani et al., 2019).

Concerning the attorney industry, hard skills and job-specific skills emerged as the most requested prerequisites (Wolf and Archer, 2013; Alamelu et al., 2017, Entika, 2017; Hirsch, 2017; Mahfud et al., 2017; Adeyinka-Ojo, 2018; Woya, 2019).

Hard skills are predominant also in the law skills set, having the presence of skills such as legal services, personal injury, legal research, legal advice, family law, intellectual property, employment law. These findings are aligned to the literature review (Jones, 2013; Finch et al., 2016; Lee & Janna,

2017; Galloway, 2017; Metilda & Neena, 2017; Bruun & Duka, 2018; Nisha & Rajasekaran, 2018; Ojanaperä et al., 2018; Clayton & Harris, 2019; Osmani et al., 2019).

Answering RQ2, results highlighted four topics concerning the skillset that a candidate as an attorney should possess. (1) The first topic of the law skillset includes skills such as an accredited lawyer, legal services, civil litigation, legal research, employment law, accredited attorney, writing skills, legal issues, personal injury, and communication skills. This topic presents skills concerning the areas of industry and job-specific skills and some soft skills (Jones, 2013; Finch et al., 2016; Lee & Janna, 2017; Galloway, 2017; Metilda & Neena, 2017; Bruun & Duka, 2018; Nisha & Rajasekaran, 2018; Ojanaperä et al., 2018; Clayton & Harris, 2019; Osmani et al., 2019). (2) The second topic presents soft skills, hard skills, and experience (e.g., legal advice, accredited attorney, legal issues, civil litigation, employment law, legal research, firm litigation, problem-solving). The topic is in line with the literature review (Jones, 2013; Finch et al., 2016; Lee & Janna, 2017; Galloway, 2017; Metilda & Neena, 2017; Bruun & Duka, 2018; Nisha & Rajasekaran, 2018; Ojanaperä et al., 2018; Clayton & Harris, 2019; Osmani et al., 2019). (3) The third thematic area clusters skills such as civil litigation, legal services, legal issues, firm litigation, personal injury, writing skills, and legal experience. This topic highlights job-specific skills, technical skills, and experience, which is in line with the literature review (Jones, 2013; Finch et

al., 2016; Lee & Janna, 2017; Galloway, 2017; Metilda & Neena, 2017; Bruun & Duka, 2018; Nisha & Rajasekaran, 2018; Ojanaperä et al., 2018; Clayton & Harris, 2019; Osmani et al., 2019). (4) The fourth topic includes skills such as civil litigation, accredited lawyer, legal experience, legal team, and legal matters. This topic highlights the most demanded technical skills to be employed in the law industry (Jones, 2013; Finch et al., 2016; Lee & Janna, 2017; Galloway, 2017; Metilda & Neena, 2017; Bruun & Duka, 2018; Nisha & Rajasekaran, 2018; Ojanaperä et al., 2018; Clayton & Harris, 2019; Osmani et al., 2019).

Responding to RQ3, results from the correlation analysis showed that every correlation for this sector is under the threshold of $r=.5$. A justification for this kind of results in the model can be due to two main reasons: (a) the law industry includes different types of attorneys having different specializations, and so the skillset presents many small groups of positively and negatively correlated skills, and, (b) because of this reason, deep investigation with text mining revealed that the content of job ads presented an overall high sparsity rate (sparsity=.98), affecting correlations' values also after the implementation of an algorithm for sparsity reduction.

Responding to RQ4, several methods have been tested in visualizing skills as a social network graph. Greedy, spectral, and optimal modularity were tested and evaluated, building specific indicators to compare their performance. The best indicator for community detection was $\xi_G=.98$, representing the

performance of greedy modularity detection, which was found almost with a perfect model fit. Indicators to evaluate spectral modularity and optimal modularity performance were relatively lower than the greedy one ($\xi_S=.88$, $\xi_O=.96$). The final dendrogram to group skills based on their modularity was plotted following this clustering, and the social network has been replotted based on partial correlations in the model. According to greedy modularity detection and the dendrogram, the network could be clustered into three groups based on skills' co-membership. The first group includes skills such as forensic experience, legal experience, insurance defense, family law, and accredited lawyer. This group pertains to job-specific skills, experience, and prerequisites. The second cluster group includes employment law, legal services, writing skills, personal injury, civil litigation, accredited attorney, firm law, and legal advice. The group pertains to job-specific skills and communication skills. The last group includes skills representing the essential skills for a candidate looking for a job in the law industry. The weighted skills network based on partial correlations reports strong correlations between forensic experience, insurance defense, and firm litigations. Other significant correlations are found between legal issues, legal matters, and the legal team. A strong negative correlation is found between legal issues and accredited attorneys. Results are in line and coherent with the literature (Jones, 2013; Finch et al., 2016; Lee & Janna, 2017; Galloway, 2017; Metilda & Neena, 2017; Bruun & Duka, 2018; Nisha &

Rajasekaran, 2018; Ojanaperä et al., 2018; Clayton & Harris, 2019; Osmani et al., 2019).

Discussing centrality measures in the skills network, the most between skills in the set were family law (32.4%), personal injury (26.8%), problem-solving (16.2%), firm law (16%), and legal team (15.4%). The closest skills were firm law (67.7%), communication skills (57.2%), personal injury (56.6%), family law (54.7%), and employment law (51.7%). The most coherent clustering method for the skills network, based on their centrality measures, was the Zhang one (Zhang, 1996; Zhang, 1997; Zhang, 2006).

Responding to RQ5, the MCMC text generation simulation, reported in the results section, provided a proper simulation of a job interview in bioinformatics. Having t_0 =firm (because firm law was the most frequent bigram, and because of MCMC's computational necessity), candidates for an open position as an attorney could be questioned regarding their skills in firm law, law issues, communication skills, sample writing, forensic experience, problem-solving, and their attitude toward working in a legal team. The interview would ideally move on to their knowledge of legal services, personal injury law on personal injury, legal advice, family law, and civil litigation and firm experience, potentially culminating with accreditation as a lawyer and knowledge of intellectual property matters.

6. Research Implication

This doctoral thesis gave rise to several research implications, both regarding the methodology and knowledge advancement. As exposed in the premise section, the thesis aimed to respond to a multidisciplinary area.

For this reason, data science, management, and marketing implications are reported in the following sections.

6.1 Data science implications

Empirical evidence from the data analysis and the coding process raised several implications for both scholars and practitioners.

Starting from the scraping process, Python software has been considered the most proper tool for the extraction of job advertisements because of its devotion to automation rather than to data analysis.

Regarding text mining, several considerations emerged regarding the nature of the data. In fact, data from job ads presented very sparse DTMs, requiring an accurate process of sparsity reduction and text cleaning prior to data analysis. The presence of a high sparsity affected semantics and correlations and required the implementation of several algorithm to extract knowledge from row data. The Author's advice regarding sparsity is to treat data in a

sequential way, defining steps and thresholds for the sparsity reduction, and then testing sparsity following the *ad hoc* parameters.

The visualization of the skills' network implies the comparison and the implementation of several methods to obtain an optimal graph. Several algorithms were tested and employed, and the Fruchterman-Reingold algorithm resulted in the most proper visualization.

Modularity has also been tested and compared to evaluate the skills' network, building specific indicators for the comparison. The greedy modularity method proved the most proper for clusters detection.

Concerning MCMC, as explained in the methodological section, the MAP method was considered the most effective in simulating an ideal job interview.

6.2 Managerial implications

First, focussing on public management, countries must adopt education policies to avoid creating a gap between the skills possessed by young people and those required by the world of work (Delavande et al., 2020). This consideration implies a continuous monitoring of the skills requested and a consequent commitment in the field of education, which can only be achieved if the education and labor sectors work together (Brown et al., 2003; Aleksynska and Tritah, 2013; Kupets, 2015).

Secondly, the research and labour sectors must interact with each other to define practical skills mapping applicable to employability, as it is a constantly evolving construct (Hogan et al., 2013; De Fruyt et al., 2015; Hirsch, 2017; Petrovski et al., 2017, Fugate et al., 2021). This kind of skill mapping is a new method in the literature, as the definition of skills that are valuable to employability has never been analyzed concerning a specific professional figure or sector, preferring the classification of communicative, employment, hard, and soft skills (Harris and King, 2015; Alamelu et al., 2017; Woya, 2019). In conclusion, mapping the primary skills required by a professional figure allows companies to support the decision when selecting and recruiting human resources, independently if referred to multinationals, large companies, or SMEs.

Together with statistics, data analysis, and software engineering, it is possible to realize an existing decision support system for HR management, which allows companies to streamline the decisional, managerial, and organizational processes. This means they could possess an accurate dossier of candidates at their disposal, which can be classified more efficiently by using digital tools up to the semi-automatic selection of resources by defining thresholds in the analyzed values. Managers from the examined sectors could benefit from this research by employing a DSS to manage and channel the current skills demanded by corporate necessities. Prototypes of the Decision

Support System are supplied online as interactive plots, which are available by scanning the following QR Codes.

Prototype 1: Interactive bar chart of the data science industry.



Prototype 2: Simple count chart of the data science industry.



Prototype 3: Waterfall chart of the data science industry.



Prototype 4: Heatmap of the data science industry.



Prototype 5: Interactive bar chart of the accounting industry.



Prototype 6: Simple count chart of the accounting industry.



Prototype 7: Waterfall chart of the accounting industry.



Prototype 8: Heatmap of the accounting industry.



Prototype 9: Interactive bar chart of the engineering industry.



Prototype 10: Simple count plot of the engineering industry.



Prototype 11: Waterfall chart of the engineering industry.



Prototype 12: Skills heatmap of the engineering industry.



The proposal of visualizing the DSS after scanning the QR code in this doctoral thesis could enhance time management in handling job interviews, because QR codes could be printed as a sticker and be placed over candidates' CVs or directly implemented as an HTML object for tablets.

6.3 Marketing implications

Concerning the marketing side, this research could be useful to marketers and researchers in evaluating the scouting process for all those companies employed in the market comparison concerning human capital (e.g., Adecco, Manpower, GiGroup). By analyzing lexical and linguistic structures in the job ads, this research could be useful to marketers employed in the creation and development of the corporate offer regarding recruitment and placement. In fact, an employer from the marketing division could benefit from this research through implementing a proper and competitive structure for job advertisement, making them able to attract more and more applicants.

Moreover, this research could be useful to all digital marketing managers employed in the construction of the *engagement* between firms and universities (Marino & Lo Presti, 2018a) and the potential human resources and corporate management (Marino & Lo Presti, 2018b; Marino & Lo Presti, 2018c). Knowing which are the most requested skills by the labor market and from competitors could facilitate the development of communication strategies helpful in retaining a significant number of competitive resources. In sum, marketers could differentiate and diversify strategies for the growth of corporate awareness concerning the creation of brand affection and the psychological need in considering one firm more challenging than another based on its recruitment reputation. Elucidating, using this data firms could make their company seem easier to apply to, or to find employment in than other firms.

7. Conclusions

The emergence of big data and the knowledge economy gave rise to several changes in the current markets. Regarding the world of work, the job market is increasingly evolving concerning the way in which candidates seek employment. Usually, candidates search for employment on online platforms, and they commonly have their first touchpoint with the firm in a digital way. Based on this consideration, this doctoral thesis was aimed at reviewing and developing a methodology to map and profile the most required skills in eight sectors of the American labor market.

Concerning data science, the main scope and implication of this thesis revolved around extracting job ads from an open portal to derive and analyze the primary skills candidates should possess and building a DSS able to support the corporate management in the choice of potential human resources and in improving employers' knowledge and abilities in hiring.

Concerning business administration and management, the knowledge advancement in the marketing and management field stems from exploring the current skills demanded in the American market and raising comparisons with current studies from the extant literature.

Outputs from this thesis could be employed in the development of commercial strategies and digital apps to help facilitate corporate management.

7.1 Limits & early remarks

This doctoral thesis employed several methodologies to build a DSS considering the managerial literature and found several limitations. One limitation concerned the sampling. Conducting this research in a developed country was optimal, but its utility for growing countries remains unknown.

The second limitation of the study regarded the implementation of scraping techniques with R software, obliging the author to use solely Python software.

The third limitation concerned critiques advanced in a study regarding the use of employability as a tool for skills' evaluation (Crisp & Powell, 2016), but, to respond to this critique, as Fugate et al. (2021) intimated, results from this research highlights the validity of the employability construct.

In the future, this research will extend the methodological application to more sectors and countries to develop a proper mapping of employability skills in the western domain.

References

- Acemoglu, D. & Restrepo, P. (2020). 'Robots and jobs: Evidence from US labor markets.' *Journal of Political Economy*, 128(6), 2188-2244.
- Acosta, P. M. (2018). 'The role of cognitive and socio-emotional skills in labor markets.' *IZA World of Labor*.
- Adeyinka-Ojo, S. (2018). 'A strategic framework for analysing employability skills deficits in rural hospitality and tourism destinations.' *Tourism Management Perspectives*, 27, 47-54.
- Adeyinka-Ojo, S *et al.* (2020). 'Hospitality and tourism education in an emerging digital economy.' *Worldwide Hospitality and Tourism Themes*.
- Aktas, E. & Meng, Y. (2017). 'An exploration of big data practices in retail sector.' *Logistics*, 1(2), 12.
- Akter, S. & Wamba, S. F. (2016). 'Big data analytics in E-commerce: a systematic review and agenda for future research.' *Electronic Markets*, 26(2), 173-194.
- Alaei, A. R., Becken, S., & Stantic, B. (2019). 'Sentiment analysis in tourism: capitalizing on big data.' *Journal of Travel Research*, 58(2), 175-191.
- Alamelu, R., Lakshminarayanan, K.V. & Badrinath, V. (2017). 'Persistence of employability skills among IT software professionals - An analysis.' *International Journal of Applied Business and Economic Research*, 15(13), 325-333.
- Aleksynska, M. & Tritah, A. (2013). 'Occupation–education mismatch of immigrant workers in Europe: Context and policies.' *Economics of Education Review*, 36(1), 229-244.

Alonso, S. G., *et al.* (2017). 'A systematic review of techniques and sources of big data in the healthcare sector.' *Journal of Medical Systems*, 41(11), 1-9.

Alrifai, A. A. & Raju, V. (2019). 'The employability skills of higher education graduates: A review of literature.' *Argument*, 6(3), 83-88.

Alvarez, S. M. & Alvarez, J. F. (2018). 'Leadership development as a driver of equity and inclusion.' *Work and Occupations*, 45(4), 501-528.

Arenas, A., *et al.* (2007). 'Size reduction of complex networks preserving modularity.' *New Journal of Physics*, 9(6), 176.

Arnedillo-Sánchez, I., de Aldama, C., & Tseloudi, C. (2017). 'Mapping employability attributes onto Facebook: rESSuME: Employability skills social media survEy.' *European Conference on Technology Enhanced Learning*. Springer, Cham.

Arthur, M. B. (1994). 'The boundaryless career: A new perspective for organisational inquiry.' *Journal of Organizational Behavior*, 15(4), 295-306.

Atkins, L. (2013). 'From marginal learning to marginal employment? The real impact of "learning" employability skills.' *Power and Education*, 5(1), 28-37.

Baccarella, C. V., *et al.* (2018). 'Social media? It's serious! Understanding the dark side of social media.' *European Management Journal*, 36(4), 431-438.

Baccarella, C. V., *et al.* (2020). Averting the rise of the dark side of social media: The role of sensitization and regulation. *European Management Journal*, 38(1), 3-6.

Bacon, N. & Blyton, P. (2003). 'The impact of teamwork on skills: Employee perceptions of who gains and who loses.' *Human Resource Management Journal*, 13(2), 13-29.

Barnes, S. E. (1976). 'New method for the Anderson model.' *Journal of Physics F: Metal Physics*, 6(7), 1375.

Barrat, A., Barthelemy, M., & Vespignani, A. (2008). *Dynamical processes on complex networks*. Cambridge University Press.

Benson, V., Morgan, S., & Filippaios, F. (2014). 'Social career management: Social media and employability skills gap.' *Computers in Human Behavior*, 30, 519-525.

Bongomin, O., et al. (2020). 'Exponential disruptive technologies and the required skills of industry 4.0.' *Journal of Engineering*, 2020.

Bradlow, E. T., et al. (2017). 'The role of big data and predictive analytics in retailing.' *Journal of Retailing*, 93(1), 79-95.

Brandes, U., et al. (2007). 'On modularity clustering.' *IEEE Transactions on Knowledge and Data Engineering*, 20(2), 172-188.

Broniatowski, D. A., Paul, M. J., & Dredze, M. (2014). 'Twitter: big data opportunities.' *Inform*, 49, 255.

Brown, B., Chui, M., & Manyika, J. (2011). 'Are you ready for the era of "big data."' *McKinsey Quarterly*, 4(1), 24-35.

Brown, B., Sikes, J., & Willmott, P. (2013). 'Bullish on digital: McKinsey global survey results.' *McKinsey Quarterly*, 12, 1-8.

Büchi, M., et al. (2019). Chilling effects of profiling activities: Mapping the issues. Available at SSRN 3379275.

Bumblauskas, D., et al. (2017). 'Smart maintenance decision support systems (SMDSS) based on corporate big data analytics.' *Expert Systems with Applications*, 90, 303-317.

Cai, H., *et al.* (2016). 'IoT-based big data storage systems in cloud computing: perspectives and challenges.' *IEEE Internet of Things Journal*, 4(1), 75-87.

Cernușca, L. (2020). 'Soft and Hard Skills in Accounting Field-Empiric Results and Implication for the Accountancy Profession.' *Studia Universitatis 'Vasile Goldis' Arad–Economics Series*, 30(1), 33-56.

Chaibate, H., *et al.* (2020). 'A comparative study of the engineering soft skills required by Moroccan job market.' *International Journal of Higher Education*, 9(1), 142-152.

Chan, D. (2000). Understanding adaptation to changes in the work environment: Integrating individual difference and learning perspectives, in Ferris, G.R. (Ed.), *Research in personnel and human resources management*, 18, 1–42, Stamford, CT: JAI Press.

Chen, M., Kuzmin, K., & Szymanski, B. K. (2014). 'Community detection via maximization of modularity and its variants.' *IEEE Transactions on Computational Social Systems*, 1(1), 46-65.

Chen, Y., *et al.* (2017). 'Family caregiver contribution to self-care of heart failure: an application of the information-motivation-behavioral skills model.' *Journal of Cardiovascular Nursing*, 32(6), 576-583.

Church, K., & Hanks, P. (1990). 'Word association norms, mutual information, and lexicography.' *Computational Linguistics*, 16(1), 22-29.

Cimatti, B. (2016). 'Definition, development, assessment of soft skills and their role for the quality of organisations and enterprises.' *International Journal for Quality Research*, 10(1), 97-130.

Collberg, C., *et al.* (2003, June). 'A system for graph-based visualization of the evolution of software.' *Proceedings of the 2003 ACM symposium on Software Visualization* (pp. 77-ff).

Crisp, R., & Powell, R. (2017). 'Young people and UK labour market policy: A critique of "employability" as a tool for understanding youth unemployment.' *Urban Studies*, 54(8), 1784-1807.

Cukier, W., Hodson, J., & Omar, A. (2015). *Soft Skills are Hard. A Review of the Literature*. Ryerson University, 50.

Custers, B., & Uršič, H. (2016). 'Big data and data reuse: a taxonomy of data reuse for balancing big data benefits and personal data protection.' *International Data Privacy Law*, 6(1), 4-15.

Danford, A., *et al* (2009). "'Everybody's talking at me": The dynamics of information disclosure and consultation in high-skill workplaces in the UK.' *Human Resource Management Journal*, 19(4), 337-354.

Davenport, T. (2014). *Big data at work: dispelling the myths, uncovering the opportunities*. Harvard Business Review Press.

Davenport, T. H. & Dyché, J. (2013). 'Big data in big companies.' *International Institute for Analytics*, 3, 1-31.

De Fruyt, F., Wille, B. & John, O. P. (2015). 'Employability in the 21st century: Complex (interactive) problem solving and other essential skills.' *Industrial and Organisational Psychology*, 8(2), 276-281.

Degryse, C. (2016). 'Digitalisation of the economy and its impact on labour markets.' ETUI research paper – working paper.

Dejaeghere, J., Wiger, N. P., & Willemsen, L. W. (2016). 'Broadening educational outcomes: social relations, skills development, and employability for youth.' *Comparative Education Review*, 60(3), 457-479.

Dekker, M., Panja, D., Dijkstra, H., & Dekker, S. (2019). 'Predicting transitions across macroscopic states for railway systems.' *PLoS One*, 14(6), 1-26, e0217710.

Delavande, A., Del Bono, E., & Holford, A. (2020). 'Academic and non-academic investments at university: The role of expectations, preferences and constraints.' *Journal of Econometrics*.

DeMarco, T. (1979). *Structure analysis and system specification*. In *Pioneers and Their Contributions to Software Engineering* (pp. 255-288). Springer, Berlin, Heidelberg.

Denrell, J. & Le Mens, G. (2020). 'Revisiting the competency trap.' *Industrial and Corporate Change*, 29(1), 183-205.

Desouza, K. C. & Jacob, B. (2017). 'Big data in the public sector: Lessons for practitioners and scholars.' *Administration & Society*, 49(7), 1043-1064.

Devins, C., et al. (2017). *The law and big data*. Cornell JL & Public Policy, 27, 357.

Di Gregorio, A., et al. (2019). 'Employability skills for future marketing professionals.' *European Management Journal*, 37(3), 251-258.

Duka, A. & Bruun, E. P. (2018). Artificial Intelligence, Jobs and the Future of Work: Racing with the Machines. *Basic Income Studies*, 13(2), 1-15.

Ebner, K., Bühnen, T., & Urbach, N. (2014, January). 'Think big with big data: Identifying suitable big data strategies in corporate environments.' *2014 47th Hawaii International Conference on System Sciences* (pp. 3748-3757). IEEE.

Eder, M., Rybicki, J. and Kestemont, M. (2016). 'Stylometry with R: a package for computational text analysis.' *R Journal*, 8(1), 107-121.

Edwards, P., Sengupta, S., & Tsai, C. J. (2009). 'Managing low-skill workers: a study of small UK food manufacturing firms.' *Human Resource Management Journal*, 19(1), 40-58.

Entika, C. L., *et al.* (2017). 'Preliminary study on the prominent entrepreneurial skills set in the context of civil engineering practice.' *Journal of Technical Education and Training*, 9(2), 94-104.

Finch, D. J., *et al.* (2016). A dynamic capabilities view of employability. *Education+ Training*, 58(1), 61-81.

Forde, C. & MacKenzie, R. (2004). 'Cementing skills: training and labour use in UK construction.' *Human resource management journal*, 14(3), 74-88.

Fredriksson, C., *et al.* (2017). 'Big data in the public sector: A systematic literature review.' *Scandinavian Journal of Public Administration*, 21(3), 39-62.

Fruchterman, T. M. & Reingold, E. M. (1991). 'Graph drawing by force-directed placement.' *Software: Practice and experience*, 21(11), 1129-1164.

Fugate, M. (2006), 'Employability,' in Greenhaus, J., and Callanan, G. (Eds.), *Encyclopedia of career development* (Vol. 1, pp. 267–271). Sage, Thousand Oaks, CA.

Fugate, M. & Kinicki, A. J. (2008). 'A dispositional approach to employability: Development of a measure and test of implications for employee reactions to organizational change.' *Journal of Occupational and Organizational Psychology*, 81(3), 503-527.

Fugate, M., Kinicki, A. J., & Ashforth, B. E. (2004). 'Employability: A psycho-social construct, its dimensions, and applications.' *Journal of Vocational behavior*, 65(1), 14-38.

Fugate, M., *et al.* (2021). 'Is what's past prologue? A review and agenda for contemporary employability research.' *Academy of Management Annals*, 15(1), 266-298.

Galetsis, P., Katsaliaki, K., & Kumar, S. (2020). 'Big data analytics in health sector: Theoretical framework, techniques and prospects.' *International Journal of Information Management*, 50, 206-216.

Galloway, K. (2017). 'A rationale and framework for digital literacies in legal education.' *Legal Educ. Rev.*, 27, 117.

Goetsch, D. L. & Davis, S. B. (2014). *Quality management for organizational excellence*. Upper Saddle River, NJ: Pearson.

Gokuladas, V. K. (2011). 'Predictors of employability of engineering graduates in campus recruitment drives of Indian software services companies.' *International Journal of Selection and Assessment*, 19(3), 313-319.

Görke, R., *et al.* (2013). 'Dynamic graph clustering combining modularity and smoothness.' *Journal of Experimental Algorithmics (JEA)*, 18, 1-1.

Grandjean, M. (2015). 'Introduction à la visualisation de données: l'analyse de réseau en histoire.' *Geschichte und Informatik*, (18/19), 109-128.

Greg, W. W. (1944). 'The statistical study of literary vocabulary.' *The Modern Language Review*, 39(1), 291-293.

Groeneveld, W., Becker, B. A., & Vennekens, J. (2020, June). 'Soft skills: What do computing program syllabi reveal about non-technical expectations of undergraduate students?' *Proceedings of the 2020 ACM Conference on Innovation and Technology in Computer Science Education* (pp. 287-293).

Guiraud, P. (1954). 'Stylistiques.' *Neophilologus*, 38(1), 1-11.

Guo, X. (2016, September). 'Application of meteorological big data.' *2016 16th international Symposium on Communications and information technologies (ISCIT)* (pp. 273-279). IEEE.

Hall, D. T. (2002). *Careers in and out of organisations*. Sage, Thousand Oaks, CA.

Hall, D.T. (1986). 'Dilemmas in linking succession planning to individual executive learning.' *Human Resource Management*, 25(2), 235-265.

Hall, D.T. & Mirvis, P.H. (1995). 'The new career contract: Developing the whole person at midlife and beyond.' *Journal of Vocational Behavior*, 47(3), 269-289.

Hamel, G. & Prahalad, C. K. (1990). 'The core competence of the corporation.' *Harvard business review*, 68(3), 79-91.

Harris, C. R. & King, S. B. (2015). 'Rural Mississippi community college students' perceptions of employability skills.' *Community College Journal of Research and Practice*, 39(4), 383-386.

Harris, L. R. (1977). 'User oriented data base query with the ROBOT natural language query system.' *International Journal of Man-Machine Studies*, 9(6), 697-713.

Harris, R. & Clayton, B. (2018). 'The importance of skills—but which skills?.' *Education and training*, 16(2), 99-102.

Healy, T. & Côté, S. (2001). *The Well-Being of Nations: The Role of Human and Social Capital. Education and Skills*. Organisation for Economic Cooperation and Development, Paris, France.

Heckman, J. J., & Kautz, T. (2012). 'Hard evidence on soft skills.' *Labour economics*, 19(4), 451-464.

Heimo, T., *et al.* (2007). 'Spectral and network methods in the analysis of correlation matrices of stock returns.' *Physica A: Statistical Mechanics and its Applications*, 383(1), 147-151.

Herdan, G. (1958). 'The relation between the dictionary distribution and the occurrence distribution of word length and its importance for the study of quantitative linguistics.' *Biometrika*, 45(1-2), 222-228.

Herdan, G. (1964). 'Quantitative Linguistics or Generative Grammar?.' *Linguistics*, 2(4), 56-65.

Hinchliffe, G. W. & Jolly, A. (2011). 'Graduate identity and employability.' *British Educational Research Journal*, 37(4), 563-584.

Hirsch, B. J. (2017). 'Wanted: Soft skills for today's jobs.' *Phi Delta Kappan*, 98(5), 12-17.

Hitzler, P. & Janowicz, K. (2013). 'Linked data, big data, and the 4th paradigm.' *Semantic Web*, 4(3), 233-235.

Hogan, R., Chamorro-Premuzic, T., & Kaiser, R. B. (2013). 'Employability and career success: Bridging the gap between theory and reality.' *Industrial and Organisational Psychology*, 6(1), 3-16.

Hollister, J. M., *et al.* (2017). 'Employers' perspectives on new information technology technicians' employability in North Florida.' *Education+ Training*, 59(9), 929-945.

Holmlund, M., *et al.* (2020). 'Customer experience management in the age of big data analytics: A strategic framework.' *Journal of Business Research*, 116, 356-365.

Huang, T. & Jiao, F. (2017, August). 'Data transfer and extension for mining big meteorological data.' *International Conference on Intelligent Computing* (pp. 57-66). Springer, Cham.

Hurrell, S. A., Scholarios, D., & Thompson, P. (2013). 'More than a "humpty dumpty" term: Strengthening the conceptualisation of soft skills.' *Economic and Industrial Democracy*, 34(1), 161-182.

Ijaola, I. A., Omolayo, O. H., & Zakariyyh, K. I. (2020). 'Project manager's skills acquisition: A comparative study of indigenous and multinational

construction firms.’ *Journal of Engineering, Project, and Production Management*, 10(1), 71-79.

Imene F., & Imhanzenobe, J. (2020). ‘Information technology and the accountant today: What has really changed?’ *Journal of Accounting and Taxation*, 12(1), 48-60.

Jee, K. & Kim, G. H. (2013). ‘Potentiality of big data in the medical sector: focus on how to reshape the healthcare system.’ *Healthcare informatics research*, 19(2), 79.

Jones, E. (2013). ‘Internationalization and employability: The role of intercultural experiences in the development of transferable skills.’ *Public Money & Management*, 33(2), 95-104.

Kamada, T. & Kawai, S. (1989). ‘An algorithm for drawing general undirected graphs.’ *Information processing letters*, 31(1), 7-15.

Kamaru Zaman, *et al.* (2019). ‘Staff employment platform (StEP) using job profiling analytics.’ *Communications in Computer and Information Science*, 937, 387-401.

Kan, Z., *et al.* (2018). ‘Estimating vehicle fuel consumption and emissions using GPS big data.’ *International journal of environmental research and public health*, 15(4), 566.

Kapil, G., Agrawal, A., & Khan, R. A. (2016, October). ‘A study of big data characteristics.’ *2016 International Conference on Communication and Electronics Systems (ICCES)* (pp. 1-4). IEEE.

Kim, G. H., Trimi, S., & Chung, J. H. (2014). ‘Big-data applications in the government sector.’ *Communications of the ACM*, 57(3), 78-85.

Klievink, B., *et al.* (2017). ‘Big data in the public sector: Uncertainties and readiness.’ *Information systems frontiers*, 19(2), 267-283.

Korczynski, M. (2005). 'Skills in service work: An overview.' *Human Resource Management Journal*, 15(2), 3-14.

Kotler, P., Kartajaya, H., & Setiawan, I. (2019). 'Marketing 3.0: From products to customers to the human spirit.' In *Marketing Wisdom* (pp. 139-156). Springer, Singapore.

Kupets, O. (2015), 'Skill mismatch and overeducation in transition economies.' *IZA World of Labor*, 224.

Li, J., *et al.* (2018). 'Big data in tourism research: A literature review.' *Tourism Management*, 68, 301-323.

Liu, H., *et al.* (2019). 'Personality or value: A comparative study of psychographic segmentation based on an online review enhanced recommender system.' *Applied Sciences*, 9(10), 1992.

Liu, H., *et al.* (2020). 'When Gaussian process meets big data: A review of scalable GPs.' *IEEE transactions on neural networks and learning systems*, 31(11), 4405-4423.

Lucianelli, G. & Citro, F. (2018). 'Accounting education for professional accountants: Evidence from Italy.' *International Journal of Business and Management*, 13(8), 1-15.

Mahfud, T., Kusuma, B. J., & Mulyani, Y. (2017). 'Soft skill competency map for the apprenticeship programme in the Indonesian Balikpapan hospitality industry.' *Journal of Technical Education and Training*, 9(2), 16-34.

Mahmud, S., Alam, Q., & Härtel, C. (2014). 'Mismatches in skills and attributes of immigrants and problems with workplace integration: a study of IT and engineering professionals in Australia.' *Human Resource Management Journal*, 24(3), 339-354.

Majeed, A., Lv, J., & Peng, T. (2019). 'A framework for big data driven process analysis and optimization for additive manufacturing.' *Rapid Prototyping Journal*.

Makki, B. I., *et al.* (2015). 'The relationship between work readiness skills, career self-efficacy and career exploration among engineering graduates: A proposed framework.' *Research Journal of Applied Sciences, Engineering and Technology*, 10(9), 1007-1011.

Mansour, R. F. (2016). 'Understanding how big data leads to social networking vulnerability.' *Computers in Human Behavior*, 57, 348-351.

Mariani, M., Di Fatta, G., & Di Felice, M. (2018). 'Understanding customer satisfaction with services by leveraging big data: The role of services attributes and consumers' cultural background.' *IEEE Access*, 7, 8195-8208.

Marino, V. & Presti, L. L. (2018a). 'Approaches to university public engagement in the online environment: Insights from Anglo-Saxon higher education.' *International Journal of Educational Management*, 32(5), 734-748.

Marino, V. & Presti, L. L. (2018b). 'Engagement, satisfaction and customer behavior-based CRM performance: An empirical study of mobile instant messaging.' *Journal of Service Theory and Practice*, 28(5), 682-707.

Marino, V. & Presti, L. L. (2018c). 'From citizens to partners: the role of social media content in fostering citizen engagement.' *Transforming Government: People, Process and Policy*, 12(1), 39-60.

Marjani, M., *et al.* (2017). 'Big IoT data analytics: architecture, opportunities, and open research challenges.' *IEEE access*, 5, 5247-5261.

Marsithi, A. & Alias, M. (2013). 'Successful intelligence via Problem-Based Learning: Promoting employability skills of engineering graduates.' *Proceedings of the Research in Engineering Education Symposium*, Kuala Lumpur, Malaysia.

Martin, S., Brown, W. M., & Wylie, B. N. (2007). Dr. 1: Distributed recursive (graph) layout (No. dR1; 002182MLTPL00). Sandia National Laboratories.

Mason, G., Williams, G., & Cranmer, S. (2009). 'Employability skills initiatives in higher education: what effects do they have on graduate labour market outcomes?' *Education Economics*, 17(1), 1-30.

Matteson, M. L., Anderson, L., & Boyden, C. (2016). 'Soft skills: A phrase in search of meaning.' *Portal: Libraries and the Academy*, 16(1), 71-88.

McAfee, A., *et al.* (2012). 'Big data: the management revolution.' *Harvard business review*, 90(10), 60-68.

Metilda, R. M. & PC, N. (2017). 'Impact of digital technology on learning to enhance the employability skills of business management graduates.' *The Online Journal of Distance Education and e-Learning*, 5(2), 35.

Minocha, S., Tudor, A. D., & Tilling, S. (2017). 'Affordances of mobile virtual reality and their role in learning and teaching.' *Proceedings of the 31st British Computer Society Human Computer Interaction Conference* (pp. 1-10).

Mishra, K. (2014). 'Employability skills that recruiters demand.' *IUP Journal of Soft Skills*, 8(3), 50-55.

Misra, R. K. & Khurana, K. (2018). 'Analysis of employability skill gap in information technology professionals.' *International Journal of Human Capital and Information Technology Professionals (IJHCITP)*, 9(3), 53-69.

Mohammed, T. A., *et al.* (2019, December). 'Big data challenges and achievements: applications on smart cities and energy sector.' *Proceedings of the Second International Conference on Data Science, E-Learning and Information Systems* (pp. 1-5).

Moses, L. B. & Chan, J. (2014). 'Using big data for legal and law enforcement decisions: Testing the new tools.' *UNSWLJ*, 37, 643.

Mourtzis, D., Vlachou, E., & Milas, N. J. P. C. (2016). 'Industrial big data as a result of IoT adoption in manufacturing.' *Procedia cirp*, 55, 290-295.

Muhamad, M. & Seng, G. H. (2019). 'Teachers' perspective of 21st century learning skills in Malaysian ESL classrooms.' *International Journal of Advanced and Applied Sciences*, 6(10), 32-37.

Munné, R. (2016). 'Big data in the public sector.' *New Horizons for a Data -Driven Economy* (pp. 195-208). Springer, Cham.

Murdoch, J. (2015). 'Using self-and peer assessment at honours level: bridging the gap between law school and the workplace.' *The Law Teacher*, 49(1), 73-91.

Nadakuditi, R. R. & Newman, M. E. (2012). 'Graph spectra and the detectability of community structure in networks.' *Physical review letters*, 108(18), 188701.

Needham, S. & Papier, J. (2018). 'Professional qualifications for the insurance industry: Dilemmas for articulation and progression.' *Journal of Vocational, Adult and Continuing Education and Training*, 1(1), 52-70.

Nematzadeh, A., *et al.* (2014). 'Optimal network modularity for information diffusion.' *Physical review letters*, 113(8), 088701.

Newman, M. E. (2006). 'Modularity and community structure in networks.' *Proceedings of the national academy of sciences*, 103(23), 8577-8582.

Newman, M. E. (2013). 'Spectral methods for community detection and graph partitioning.' *Physical Review E*, 88(4), 042822.

Newman, M. E., Watts, D. J., & Strogatz, S. H. (2002). 'Random graph models of social networks.' *Proceedings of the national academy of sciences*, 99(suppl 1), 2566-2572.

Ngoma, M. & Dithan Ntale, P. (2016). 'Psychological capital, career identity and graduate employability in Uganda: the mediating role of social capital.' *International Journal of Training and Development*, 20(2), 124-139.

Nikolaichuk, O. A., Lizunova, N. M., & Obukhova, L. Y. (2019, December). 'Labour Market in the Era of Digital Economy.' *Institute of Scientific Communications Conference* (pp. 540-552). Springer, Cham.

Nisha, S. M. & Rajasekaran, V. (2018). 'Employability skills: A review.' *IUP Journal of Soft Skills*, 12(1), 29-37.

Nurlaela, L., Romadhoni, I. F., & Widodo, W. (2017). 'Application-based instructional tools for enhancing students' problem-solving skills in home economics.' *Journal of Technical Education and Training*, 9(3), 46-56.

O'Connor, C. & Kelly, S. (2017). 'Facilitating knowledge management through filtered big data: SME competitiveness in an agri-food sector.' *Journal of Knowledge Management*, 21(1), 156-179.

Oboler, A., Welsh, K., & Cruz, L. (2012). *The danger of big data: Social media as computational social science*. First Monday.

Ojanperä, S., O'Clery, N., & Graham, M. (2018). *Data science, artificial intelligence and the futures of work*. The Alan Turing Institute.

Onnela, J. P., *et al.* (2002). 'Dynamic asset trees and portfolio analysis.' *The European Physical Journal B-Condensed Matter and Complex Systems*, 30(3), 285-288.

Onnela, J. P., Kaski, K., & Kertész, J. (2004). Clustering and information in correlation based financial networks. *The European Physical Journal B*, 38(2), 353-362.

Osmani, M., *et al.* (2019). 'Graduates employability skills: A review of literature against market demand.' *Journal of Education for Business*, 94(7), 423-432.

Ovelgönne, M., Geyer-Schulz, A., & Stein, M. (2010). 'Randomized greedy modularity optimization for group detection in huge social networks.' *Proc. SNA-KDD*, 10, 117-130.

Oviawe, J. I., Uwameiye, R., * Uddin, P. S. (2017). 'Best practices in technical education programme for students' capacity building and sustainable development in the 21st century.' *Journal of Technical Education and Training*, 9(3), 57-68.

Owais, S. S. & Hussein, N. S. (2016). 'Extract five categories CPIVW from the 9V's characteristics of the big data.' *International Journal of Advanced Computer Science and Applications*, 7(3), 254-258.

Özköse, H., Arı, E. S., & Gencer, C. (2015). 'Yesterday, today and tomorrow of big data.' *Procedia-Social and Behavioral Sciences*, 195, 1042-1050.

Pani, A., Das, B., & Sharma, M. (2015). 'Changing dynamics of hospitality & tourism education and its impact on employability.' *Parikalpana: KIIT Journal of Management*, 11(1), 1-12.

Park, E. (2019). 'The role of satisfaction on customer reuse to airline services: An application of big data approaches.' *Journal of Retailing and Consumer Services*, 47, 370-374.

Park, E., *et al.* (2019). 'Determinants of customer satisfaction with airline services: An analysis of customer feedback big data.' *Journal of Retailing and Consumer Services*, 51, 186-190.

Patel, H., *et al.* (2020). 'Transforming petroleum downstream sector through big data: a holistic review.' *Journal of Petroleum Exploration and Production Technology*, 10(6), 2601-2611.

Pejić Bach, *et al.* (2019). 'Text mining for big data analysis in financial sector: A literature review.' *Sustainability*, 11(5), 1277.

Pejic-Bach, *et al.* (2020). 'Text mining of industry 4.0 job advertisements.' *International journal of information management*, 50, 416-431.

Percell, J. C. (2016). 'Data collaborative: A practical exploration of big data in course wikis.' *Quarterly Review of Distance Education*, 17(4), 63.

Perrons, R. K. & Jensen, J. W. (2015). 'Data as an asset: What the oil and gas sector can learn from other industries about "big data."' *Energy Policy*, 81, 117-121.

Petrovski, E., Dencker-Larsen, S., & Holm, A. (2017). 'The effect of volunteer work on employability: a study with Danish survey and administrative register data.' *European Sociological Review*, 33(3), 349-367.

Pieterse, V. & Van Eekelen, M. (2016, July). 'Which are harder? Soft skills or hard skills?' *Annual Conference of the Southern African Computer Lecturers' Association* (pp. 160-167). Springer, Cham.

Pool, L. D. & Sewell, P. (2007). 'The key to employability: developing a practical model of graduate employability.' *Education + Training*, 49(3), 277-289.

Porat, A., & Strahilevitz, L. J. (2014). 'Personalizing default rules and disclosure with big data.' *Michigan Law Review*, 112(8), 1417-1478.

Pouyanfar, S., *et al.* (2018). 'Multimedia big data analytics: A survey.' *ACM computing surveys (CSUR)*, 51(1), 1-34.

Prettejohn, B. J., Berryman, M. J., & McDonnell, M. D. (2011). 'Methods for generating complex networks with selected structural properties for simulations: A review and tutorial for neuroscientists.' *Frontiers in computational neuroscience*, 5, 11.

Quintini, G. (2014). Skills at work: How skills and their use matter in the labour market, oecd social, employment and migration working papers. OECD Publishing, Paris.

Ramírez-Pérez, H. X. *et al.* (2015). 'Effects of training method and age on employability skills of Mexican youth entrepreneurs.' *Journal of Entrepreneurship Education*, 18(3), 125-139.

Rampersad, G. (2020). 'Robot will take your job: Innovation for an era of artificial intelligence.' *Journal of Business Research*, 116, 68-74.

Riikkinen, M., *et al.* (2018). 'Using artificial intelligence to create value in insurance.' *International Journal of Bank Marketing*.

Robins, G., *et al.* (2007). 'An introduction to exponential random graph (p*) models for social networks.' *Social networks*, 29(2), 173-191.

Rosenberg, S., Heimler, R., & Morote, E. S. (2012). 'Basic employability skills: A triangular design approach.' *Education+ Training*, 54(1), 7-20.

Saha, B. & Srivastava, D. (2014, March). 'Data quality: The other face of big data.' *2014 IEEE 30th international conference on data engineering* (pp. 1294-1297). IEEE.

Sarkar, S., *et al.* (2014). 'Spectral characterization of hierarchical modularity in product architectures.' *Journal of Mechanical Design*, 136(1).

Sarlin, P. (2016). 'Macroprudential oversight, risk communication and visualization.' *Journal of Financial Stability*, 27, 160-179.

Sarlin, P. & Marghescu, D. (2011). 'Visual predictions of currency crises using self-organizing maps.' *Intelligent Systems in Accounting, Finance and Management*, 18(1), 15-38.

Sarlin, P. & Peltonen, T. A. (2013). 'Mapping the state of financial stability.' *Journal of International Financial Markets, Institutions and Money*, 26, 46-76.

Savickas, M.L. (2005). 'The theory and practice of career construction,' in Lent, R.W. and Brown, S.D. (Eds), *Career Development and Counseling: Putting Theory and Research to Work* (pp. 42-70). John Wiley and Sons, Hoboken, NJ.

Schuetz, P. & Caflisch, A. (2008). 'Efficient modularity optimization by multistep greedy algorithm and vertex mover refinement.' *Physical Review E*, 77(4), 046112.

Selvam, T. (2017). 'Promoting Factors of Employability Skills.' *International Journal for Research in Engineering Application & Management*, 4(3).

Sharma, V. (2018). 'Soft skills: An employability enabler.' *IUP Journal of Soft Skills*, 12(2), 25-32.

Sheikholeslami, G., Chatterjee, S., & Zhang, A. (1998, August). 'Wavecluster: A multi-resolution clustering approach for very large spatial databases.' *VLDB* (Vol. 98, pp. 428-439).

Shields, R. & Kameshwara, K. K. (2020). 'Social justice perspectives on education, skills, and economic inequalities.' *Handbook on promoting social justice in education*, Springer International Publishing, Cham, Switzerland, 1-15.

Shmatko, N., Gokhberg, L., & Meissner, D. (2020). 'Skill-sets for prospective careers of highly qualified labor.' *Handbook of labor, human resources and population economics*, 1-14.

Siddoo, V., *et al.* (2017). 'Exploring the competency gap of IT students in Thailand: The employers view of an effective workforce.' *Journal of Technical Education and Training*, 9(3), 1-15.

Silva, E. S., Hassani, H., & Madsen, D. Ø. (2019). 'big data in fashion: Transforming the retail sector.' *Journal of Business Strategy*, 41(4), 21-27.

Smaldone, F. (2019). 'Employability skills at the age of big data: exploring data scientists' requirements in the digital labor market.' *Ebiss2019 Conference*. <https://cs.ulb.ac.be/conferences/ebiss2019/posters.html>.

Smaldone, F. (2020). 'Profiling Patients' Care Satisfaction in Outpatient Settings via Comparative Twitter Analysis.' *Integrative Journal of Global Health*, 9th International Conference on Hospital Management and Healthcare. Barcelona, Spain, 2020.

Smaldone, F., D'Arco, M., & Marino, V. (2021, June). 'Fight against corona: Exploring consumer-brand relationship via Twitter textual analysis.' In *Digital Marketing & eCommerce Conference* (pp. 104-111). Springer, Cham.

Smaldone, F., D'Arco, M., & Marino, V. (2021, June). 'I am free to be in a grocery store: Profiling consumers' spending during covid-19 pandemic via big data market basket analysis,' in *National Brand and Private Label Marketing Conference* (pp. 47-54). Springer, Cham.

Smaldone, F., *et al.* (2021). 'Virulence of two infectious diseases: Covid-19 and inequality,' In *The International Research & Innovation Forum* (pp. 503-512). Springer, Cham.

Smaldone, F., Ippolito, A., & Ruberto, M. (2020). 'The shadows know me: Exploring the dark side of social media in the healthcare field.' *European Management Journal*, 38(1), 19-32.

Sonka, S. (2014). 'Big data and the ag sector: More than lots of numbers.' *International Food and Agribusiness Management Review*, 17(1030-2016-82967), 1-20.

Sparreboom, T. & Tarvid, A. (2017). Skills mismatch of natives and immigrants in Europe. International Labour Office, Conditions of Work and Equality Department, ILO. Switzerland: Geneva.

Spiess, J., *et al* (2014). 'Using big data to improve customer experience and business performance.' *Bell Labs Technical Journal*, 18(4), 3-17.

Stathopoulou, A., Siamagka, N. T., & Christodoulides, G. (2019). 'A multi-stakeholder view of social media as a supporting tool in higher education: An educator–student perspective.' *European Management Journal*, 37(4), 421-431.

Stringfield, S. and Stone III, J. R. (2017). 'The labor market imperative for CTE: Changes and challenges for the 21st century.' *Peabody Journal of Education*, 92(3), 166-179.

Suarta, I. M., *et al.* (2017, September). 'Employability skills required by the 21st century workplace: A literature review of labor market demand.' *International Conference on Technology and Vocational Teachers (ICTVT 2017)*. Atlantis Press.

Suarta, I. M., *et al.* (2018). 'Employability skills for entry level workers: a content analysis of job advertisements in Indonesia.' *Journal of Technical Education and Training*, 10(2).

Sumbal, M. S., Tsui, E., & See-to, E. W. (2017). 'Interrelationship between big data and knowledge management: an exploratory study in the oil and gas sector.' *Journal of Knowledge Management*.

Talón-Ballester, P., *et al.* (2018). 'Using big data from customer relationship management information systems to determine the client profile in the hotel sector.' *Tourism Management*, 68, 187-197.

Tan, L. M. & Laswad, F. (2018). 'Professional skills required of accountants: what do job advertisements tell us?' *Accounting Education*, 27(4), 403-432.

Trabucchi, D. & Buganza, T. (2019). 'Data-driven innovation: switching the perspective on big data.' *European Journal of Innovation Management*, 22(1), 23-40.

Trelewicz, J. Q. (2017). 'Big data and big money: The role of data in the financial sector.' *IT Professional*, 19(3), 8-10.

Trilling, B. & Fadel, C. (2009). *21st century skills: Learning for life in our times*. John Wiley & Sons, San Francisco, CA.

Van Mieghem, *et al.* (2010). 'Spectral graph analysis of modularity and assortativity.' *Physical Review E*, 82(5), 056113.

Vanhercke, D., *et al.* (2014). 'Defining perceived employability: a psychological approach.' *Personnel Review*.

Varshney, D. (2020). 'Digital transformation and creation of an agile workforce: exploring company initiatives and employee attitudes.' In Turkmenoglu, M.A. and Cicek, B. (Eds.), *Contemporary Global Issues in Human Resource Management* (pp. 89-105). Emerald Publishing Limited.

Vayena, E., *et al.* (2018). 'Policy implications of big data in the health sector.' *Bulletin of the World Health Organization*, 96(1), 66.

Wang, Z., Mao, S., Yang, L., & Tang, P. (2018). 'A survey of multimedia big data.' *China Communications*, 15(1), 155-176.

Watts, D. J., & Strogatz, S. H. (1998). 'Collective dynamics of "small-world" networks.' *Nature*, 393(6684), 440-442.

Weibel, S., *et al.* (1998). 'Dublin core metadata for resource discovery.' *Internet Engineering Task Force RFC*, 2413(222), 132.

Westphalen, S. Å. (1999). *Reporting on human capital Objectives and trends*. Amsterdam.

Williams, A. M. (2015). *Soft skills perceived by students and employers as relevant employability skills*. Doctoral dissertation. Walden University.

Wilton, N. (2011). 'Do employability skills really matter in the UK graduate labour market? The case of business and management graduates.' *Work, Employment and Society*, 25(1), 85-100.

Winkler, S., König, C. J., & Kleinmann, M. (2012). 'New insights into an old debate: Investigating the temporal sequence of commitment and performance at the business unit level.' *Journal of Occupational and Organizational Psychology*, 85(3), 503-522.

Wolf, K. & Archer, C. (2013). 'Into the unknown: A critical reflection on a truly global learning experience.' *Issues in Educational Research*, 23(3), 299-314.

Woya, A. A. (2019). 'Employability among Statistics Graduates: Graduates Attributes, Competence, and Quality of Education.' *Education Research International 2019*.

Xiang, Z. & Fesenmaier, D. R. (2017). Big data analytics, tourism design and smart tourism. In *Analytics in smart tourism design* (pp. 299-307). Springer, Cham.

Yanbe, Y., *et al.* (2007, June). 'Can social bookmarking enhance search in the web?' *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries* (pp. 107-116).

Yin, S. & Kaynak, O. (2015). 'Big data for modern industry: challenges and trends [point of view].' *Proceedings of the IEEE*, 103(2), 143-146.

Younas, M. (2019). *Research challenges of big data*. Springer.

Yule, C. U. (2014). *The statistical study of literary vocabulary*. Cambridge University Press.

Zhang, T., Ramakrishnan, R., & Livny, M. (1996). 'BIRCH: an efficient data clustering method for very large databases.' *ACM SIGMOD Record*, 25(2), 103-114.

Zhang, T., Ramakrishnan, R., & Livny, M. (1997). 'BIRCH: A new data clustering algorithm and its applications.' *Data Mining and Knowledge Discovery*, 1(2), 141-182.

Zhang, Z., Huang, K., & Tan, T. (2006, August). 'Comparison of similarity measures for trajectory clustering in outdoor surveillance scenes.' *18th International Conference on Pattern Recognition (ICPR'06)* (Vol. 3, pp. 1135-1138). IEEE.

Zhao, Y., Xu, X., & Wang, M. (2019). 'Predicting overall customer satisfaction: Big data evidence from hotel online textual reviews.' *International Journal of Hospitality Management*, 76, 111-121.

Zheng, K., Zhang, Z., & Song, B. (2020). 'E-commerce logistics distribution mode in big-data context: A case analysis of JD. COM.' *Industrial Marketing Management*, 86, 154-162.

Zheng, Y., *et al.* (2015, August). 'Forecasting fine-grained air quality based on big data.' *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 2267-2276).

Zhong, R. Y., Huang, G. Q., & Dai, Q. (2014, May). 'A big data cleansing approach for n-dimensional RFID-Cuboids.' *Proceedings of the 2014 IEEE 18th*

International Conference on Computer Supported Cooperative Work in Design (CSCWD) (pp. 289-294). IEEE.

Zhong, R. Y., *et al.* (2015). 'A big data approach for logistics trajectory discovery from RFID-enabled production data.' *International Journal of Production Economics*, 165, 260-272.

Zhong, R. Y., *et al.* (2016). 'Visualization of RFID-enabled shopfloor logistics big data in cloud manufacturing.' *The International Journal of Advanced Manufacturing Technology*, 84(1-4), 5-16.

Zhou, Y. (2020). 'Preparing students for the global workforce: Chinese and non-Chinese working professionals on key employability skills.' *Global Business Languages*, 20, 66-83.

Zhou, Z., *et al.* (2016). 'A method for real-time trajectory monitoring to improve taxi service using GPS big data.' *Information & Management*, 53(8), 964-977.

Zipf, G. K. (1946). 'Outline of a theory of linguistic change.' *Modern Language Notes*, 61(8), 565-569.

*I cannot be around people who think my growth is competition.
If we can't be happy for one another, we have nothing in common.*

Ioiην