IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Digital Object Identifier

Extracting Knowledge from on-line Sources for Software Engineering Labor Market: A Mapping Study

MARIA PAPOUTSOGLOU,[1] APOSTOLOS AMPATZOGLOU,[2] NIKOLAOS MITTAS[3] AND LEFTERIS ANGELIS.[1]

[1]Department of Informatics, Aristotle University of Thessaloniki, 54124, Thessaloniki, Greece (e-mail: mpapouts, lef@csd.auth.gr)
[2]Department of Applied Informatics, University of Macedonia, Thessaloniki, Greece (email: apostolos.ampatzoglou@gmail.com)
[3]Department of Chemistry, International Hellenic University, Kavala, Greece (email: nmittas@teiemt.gr)

Corresponding author: Maria Papoutsoglou (e-mail: mpapouts@csd.auth.gr).

ABSTRACT Software engineering is a continuously evolving sector and the demands of the related labor market result in a wide variety of job openings, ranging from developers to customer service positions. Thus, there is a need to continuously monitor labor market trends using data and analytics. Both employers and employees can benefit by capturing emerging trends which can facilitate continuous learning and training in new technologies, support of better matching between a job offer and the ideal candidate and expertise detection. To fulfill these needs, the results of labor market analytics need to reach the stakeholders timely and accurately. However, often delays occur, which stem from time-consuming approaches based on collecting data from traditional sources, such as questionnaires or interviews. Recently, researchers started leveraging content from digital sources, which are easily accessed and contain a wealth of information. This paper presents the results of a Systematic Mapping Study on digital sources that can be used to address the data analytics needs of the labor market. It provides a multifaceted categorization of the issues involved in the analysis of digital sources of the software engineering labor market. It aims to identify digital labor market sources for data retrieval which are appropriate for employers and employees analytics. Additionally, it aims to connect different skill types, needs and goals of labor market with the utilization of digital sources and data analysis methods. In total 86 primary studies were selected and each one was evaluated and classified aiming to identify the: (a) digital sources that are used for labor market analytics; (b) type of skills they examine; (c) methods which are used to utilize the raw digital content; (d) goals for which every primary study is conducted; (e) beneficiaries (stakeholder) of the results; and (f) time trends for all the above.

INDEX TERMS human factor, labor market analytics, digital sources, skills, software engineering

I. INTRODUCTION

DIGITAL labor market information sources that create and curate knowledge in real-time can increase the accuracy of official labor market statistics [1] and provide a new field for data collection and information science methods. Before entering the digital age, such statistics were usually survey-based, on an annual basis and were made publicly available periodically. However, recent technological developments that include knowledge sharing that is achieved through online communities can reduce the "time-to-market" for official statistics, as open digital data are available in real-time [2]. The real-time information that is offered can result into faster recognition of new trends compared to official labor market statistics. The combination of quick trend recognition and the rapid development of skills required by the IT business community, can lead to a new form of labor market, the *digital labor market*.[1] Based upon such sources, CEDEFOP[2] (the EU agency for the development of vocational education and training) investigates how online sources can be used to create a tool, which would be able to capture current skill trends and to extract almost real-time statistics on skills demand. As we can understand, utilizing digital content for labor market analysis provides a new

[1] http://www.cedefop.europa.eu/en/events-and-projects/projects/digitalisation-and-future-work.
[2] https://www.cedefop.europa.eu/en.

emerging need for automation process.

In the digital labor market, two major types of skills can be identified: hard and soft-skills. On the one hand, **hard-skills** are those that a person can acquire through formal learning processes, such as how to use a programming language [3]. On the other hand, **soft-skills** [3], such as teamwork or communication, cannot be easily acquired through formal learning. Yet, these skills can play a decisive role in determining the qualitative matching of a worker and an open job vacancy, or the amount of time the worker stays in the job [4]. According to the results of a 2017 LinkedIn survey, [3] the soft-skills constitute the third most crucial labor market aspect while in the next year's survey pointed out the lack of soft-skills, with special reference to *communication* or *interpersonal skills*.[4] The ICT sector faces the most significant soft-skills gap, in comparison with other occupations, such as account executives, sales development or project management.

Soft-skills gap becomes even more evident in the software engineering domain, which is one of the most dynamic domains of ICT sector, since software development and maintenance includes various and complex human-centric aspects. Spinellis [5] pointed out that soft-skills are closely connected to software engineering, in the sense that successful software development relies heavily on interpersonal skills, whereas the importance of the human factor in software development has led to the creation of dedicated events, such as the HuFo workshop.[5] Since software engineering is a dynamically evolving sector, there is a need to extract statistical results in real-time, using available digital information sources aiming to identify the latest trends of the software engineering labor market.

At the same time, there is also a need to map digital sources for online labor market analytics, which arises from the population growth and the impact of new Web channels for information flow. Population between 18-67 is considered as the workforce, but within this range there is a technological gap, which affects labor market information. More specifically, younger people use digital media in their everyday life to communicate with other professionals and seek for jobs. Furthermore, employers attract the new workforce using digital media so as to disseminate their requirements. This is accomplished through the post of online job advertisements, or by searching for future candidates' profiles, following their digital traces. Concluding, digital media are essentially sources of information containing a wealth of data that can be analyzed. Under this perspective, their analysis can be beneficial for providing better labor market statistics with reduced time of publication and for identifying the most informative and popular sources.

To obtain a comprehensive understanding on the use of digital labor market analysis in the software engineering sector, we consider six main questions, as presented in Figure 1, based on the 5W + 1H questions model for problem solving [6].

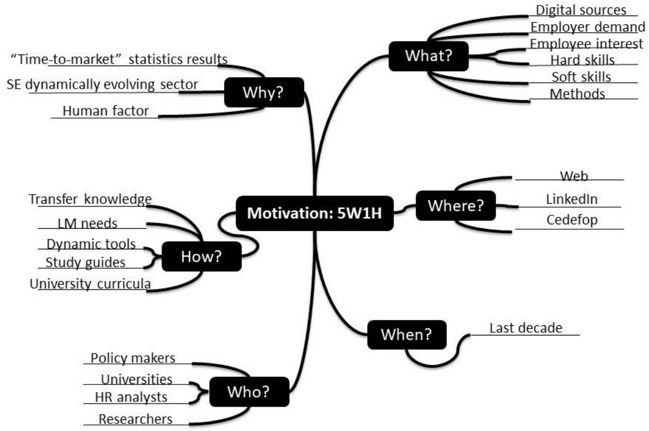

FIGURE 1. Use of Digital Labor Market Information Sources for Software Development

Q$_1$: *Where digital labor market data can be extracted from?* (**Sources**)
The Web contains a plethora of digital sources available for different aspects of labor market analytics. LinkedIn places particular emphasis on the field since 2003. One of the most important needs for efficient labor market analysis is to identify the sources that can be used. Authors in [7] claim that the web is the most dynamic source of information for assessing software engineers, and according to empirical evidence, there are different sources appropriate for extracting trends related to employer demands, employee interest and hard- and soft-skills identification (e.g. Stack Overflow, LinkedIn).
Q$_2$: *What digital labor market data can be identified?* (**Skills**)
Using the digital information leveraged from professional social networks, web portals and social coding platforms, we can identify skills that are required from the employer's perspective and skills that are acquired, from employees' point of view.
Q$_3$: *How can labor market data be analyzed?* (**Methods**)
This question is related to the appropriate methods and techniques that can be adopted to analyze the extracted raw data. The data related to labor market issues are characterized by high volume and velocity of updated content which is explainable due to the continuously evolving character of the labor market [2]. That is why their analysis requires different and usually advanced methods for content analysis [8]. Through the current study and by identifying the appropriate methods adopted to extract labor market knowledge (e.g., machine learning algorithms, statistical measures and methods, etc.) we can recognize potentials of other methods that have not yet been implemented in the field and could be

[3] LinkedIn is the leading professional social network platform, available for use from both sides (employee and employer needs); every year it releases official statistics for labor trends and needs, which rely on real-time community data such as member profiles and job postings.

[4] https://www.cnbc.com/video/2018/04/19/linkedin-ceo-on-the-soft-skills-gap.html.

[5] https://hufo2017.serandp.com/.

challenging to apply.

Q$_4$: *Who is interested in the analysis of labor market data?* (**Stakeholders**)

There is a large pipeline of parties interested in the information provided by labor market analytics. Students need to understand the current trends of required skills in order to focus on qualifications that would guarantee an entry-level position after graduation. At the same time, universities can provide courses tailored to the needs of the labor market to support their students' future successful careers and educators can update their "arsenal" with trendy programming languages and technologies. Companies want to ensure their innovation, competitiveness, sustainability and profitability by recruiting personnel having skills required in the majority of the labor market, in alignment with the most recent technological developments, by training their current staff and by hiring experts specialized in labor market data analysis. Finally, the research community (researchers and policy makers) is highly interested since new types of data create new problems and subsequently emerge new solutions (e.g., statistical models, machine learning algorithms, software engineering metrics for human factor productivity, etc.).

Q$_5$: *Why is the analysis of labor market data important?* (**Goal**)

The analysis of labor market data is important since it facilitates the identification and systematic study of notions related to the human factor and its role in professional environment, such as the expertise, the competencies, the social skills, etc. These notions are related and interact with other factors (educational or professional) and they have impact on the competitiveness, efficiency and productivity of teams or individuals. Different stakeholders (e.g. policy makers, researchers, young graduates etc.) perceive differently the importance of this information and use it differently.

Q$_6$: *When are online labor market analytics changing over time?* (**Time**)

The labor market sector is a dynamic and continuously-evolving sector and it has the potential to align with other universal trends such as the technological evolution and digitalization of many aspects of daily life. Through this study, we can explore the trend of labor market analytics from digital sources within time and locate peaks which signify rapid growth and significant changes.

To answer the aforementioned questions, we have performed a systematic mapping study (SMS), following the guidelines by Petersen et al. [9]. The rest of the paper is organized as follows: Section 2 describes related work and background information. Section 3 provides details on the employed methodology, whereas Section 4 presents the results of the study and discusses the results by providing tentative interpretations and implications for researchers and practitioners. Section 5 provides a discussion with implications using case studies of real-world needs. Section 6 reports the threats to validity and Section 7 concludes the paper.

II. BACKGROUND INFORMATION

In this section, we present background information and related work for this study. Since this paper is a secondary study, formally other literature reviews or mapping studies can be considered as direct related work. In Section 2.1 we present secondary related studies, or primary studies, which provide a comprehensive view on the topic. In Section 2.2, we present background information on the context of this study and information for different categories for digital sources, so as to facilitate the understanding of basic terms and concepts in the field.

A. RELATED WORK

Initial work by Capiluppi et al. [7] proposed a distribution of Internet sources, focusing on the appropriateness of each source for analyzing the profiles of candidates for positions requiring technical skills. The study recognized the importance of "social signals", i.e. measures of social reputation, which can be used by potential employers or recruiters. Another approach by Amandoru and Gambage [10] presented a general-purpose evaluation of social networks as a recruitment tool for HR professionals. More specifically, researchers conducted a study in a specific geolocation (i.e., Sri Lanka) aiming to evaluate digital content as sources of information to support the process of finding and evaluating candidates. The results of the study suggested that LinkedIn is the most well-known tool for recruiters that use the social networks. On the other hand, Kurekova et al. [11] considered digital sources as a means for labor market analytics focusing on web portals. Authors provide an insightful contribution as they identify challenges rising from the use of digital job vacancy data as a new raw source of information for labor market analytics. They analyze the advantages and disadvantages of digital data as a source of labor market knowledge.

The aforementioned studies, as depicted in Table 1, are not secondary studies but they contain the first traces for a new sector in knowledge extraction from digital content. One main problem they identify is the production process of official statistics due to the time needed to collect raw data. To the best of our knowledge, there is only one secondary study [12] which is related to the human factor in software engineering. More specifically, it reveals the term "human capital" in software engineering referring to knowledge and skills as added value in the development process. Researchers use four dimensions to categorize their results: (a) capacity; (b) deployment; (c) development; and (d) know-how. From the perspective of digital sources, their results focus on software engineering sources such as Jazz repository, GitHub or version control system (VCS) etc.

Recently, the trend of online labor market analytics clearly showed a separation between the two actors of the workforce; employers (demand) and employees (supply). Employers' needs are represented by job portals. Mezzanzanica & Mercorio [13] address the raw information of job advertisements as continuously increasing due to volume, variety and veloc-

TABLE 1. Comparison table for studies

Study	Software engineering sources	Web portals	(professional) Social networks	Hard or/and soft skills	Methods for data analysis	Secondary study
[7]	YES	NO	YES	NO	NO	NO
[10]	NO	NO	YES	NO	NO	NO
[11]	NO	YES	NO	NO	NO	NO
[12]	YES	NO	NO	NO	YES	YES
[13]	NO	YES	NO	YES	YES	NO
[8]	NO	YES	NO	YES	YES	NO
[15]	NO	YES	NO	YES	NO	NO
[14]	NO	YES	NO	YES	YES	NO
[16]	YES	NO	NO	YES	NO	YES
[17]	YES	NO	NO	YES	YES	YES
This	YES	YES	YES	YES	YES	YES

ity (3V). Using the 3V, the authors consider the knowledge extraction of job portals as big data analytics process and present an architecture for data collection, analysis and visualization. Additionally, Mauro et al [14] used job advertisements to explore required skills for big data professions. In order to find the connection between skills and job titles, they used latent direct allocation, i.e. a topic modeling approach is appropriate for extracting knowledge from the raw text of job advertisements [15].

Taking into consideration that the collected data from job portals need data analysis methods, we can understand that several existing Artificial Intelligence algorithms can be used to extract knowledge from digital sources for labor market purposes [8]. We have to note that Matturro et al [16] showed that the most appropriate sources to identify soft skills for software engineering sector are job advertisements, surveys (online or by e-mail) and interviews. Furthermore, in term of soft-skills detection, Sánchez-Gordón & Colomo-Palacios [17] conducted a secondary study related to employees' digital sources analysis. More specifically, they investigate emotional intelligence and soft skills in software engineering. The authors classify primary studies into different levels of emotional intelligence of sentiment (positive or negative), affect (e.g. love, joy or anger) and personality traits.

Using the comparison of Table 1 and in combination with the aforementioned research approaches, we can conclude that labor market analytics focusing on software engineering is a new challenging trend. To the best of our knowledge, there is no other secondary study identifying the profusion of different digital sources available for software engineering labor market analytics.

B. LABOR MARKET ANALYTICS

Nowadays, the web contains a variety of sources with information related to labor market analytics. On the one hand, there are sources that provide one-way information to employees, such as online web job portals where individuals only screen existing job advertisements (e.g. CareerBuilder). On the other hand, other sources allow interaction, i.e. page/s can be set to include a variety of information (e.g. company profile and details of people working in it) and the viewer can communicate with the owner of the page or profile (e.g. LinkedIn). Sources can be categorized as traditional, or as social and collaborative software engineering networks or sites.

Traditional sources are known as job portals. The most popular are: (i) Indeed,[6] (ii) CareerBuilder,[7] and (iii) Monster.[8] Monster is a leading private Internet recruitment site which has been used as a web source for several research tasks to analyze online job vacancy data. It has been used since 2004 [18] as an online source for analyzing job advertisements for company branding identification. Of particular interest is a recent study [19], where the researchers used a "web spider" as a data collector engine from Monster so as to extract IT requirements for skills in the UK market in comparison to those offered by universities. Debortoli et al. [20] also used Monster as source for skills related to business intelligence. Two recent studies [21], [22] used CareerBuilder to identify the appropriate skills required by each offered job.

The second category of sources contains other digital sources such as professional social networks, social coding platforms, Q&A web sites etc. In contrast to job portals, in this kind of digital sources, it is possible to find information for individuals and identify personal skills. Every digital source in this category captures daily user activity as the user interacts with the community (e.g. a participant in GitHub), or it is possible to leverage labor market statistics from his/her professional profile (e.g. LinkedIn). The human factor, as one of the main pillars in the success of software engineering projects [23], is crucial for team building and HR processes. A successful software engineer profile consists of three basic skills: (a) the knowledge (hard skills) which stems from educational and hands on experience, (b) cognitive, and (c) interpersonal skills, which are related to soft skills. The two last types of skills are becoming more important, since software development evokes continuous interaction among the team members. Due to the collaborative nature of activity that is recorded, there is a growing trend for research on social signals and emotions [24], [25]. Driven by the need to assess such parameters, many different digital sources capture developers' daily activity and signal traces. The most commonly used sources for such purposes are GitHub and Stack Overflow. As a result of the aforementioned, we can subcategorize this main category into the following two: (i) software engineering websites (e.g. GitHub, Stack Overflow) and (ii) social networks (professional social networks and social networks).

A disadvantage of job portals as a source of analysis for human capital data is that the job seeker, even companies in the same industry, cannot see the social profile of a company. A professional social profile of a company may include, for example, information related to job holders, their skills and competences, their interests, the connections of the new job with the current experience of employees, etc. In brief, the gap in traditional web sources for labor data analytics is

[6] https://www.indeed.jobs/career/SearchJobs.

[7] https://www.careerbuilder.com/.

[8] https://www.monster.com/.

essentially the lack of knowledge sharing.

This gap tends to be covered by web sources that contain social content. Such sources include social networks and media, blogs, collaborative projects, the world of virtual games and content communities. All the above sources promote knowledge sharing with a particular interest in showing professional content. The two most popular professional networks are LinkedIn and Xing. According to a jobvite[9] survey in 2015, 92% of recruiters used social media in the recruitment process with the first network being LinkedIn. Professional networks are different from social media such as Facebook, and have a focus on the structure of their content and are business-oriented.

We can understand that labor market analytics need to adopt information science techniques to fulfill their needs. More specifically, aiming to provide analytics almost in real time, labor market stakeholders retrieve content from digital sources such as social networks, online job portals or Q&A and collaborative sites. In order to handle the collected digital content practitioners need to adopt techniques such as data analysis and artificial intelligence methods [8]. We can note that the well-known approaches of sentiment analysis could find implication to labor market analytics, as a need of labor market is the detection of soft skills which can relate to emotional intelligence/sentiment analysis [17]. Through the current section we identified some studies which implicitly approach the online labor market analytics issue. However, these studies focus into specific actors of labor market demand and supply, some of them focus into employers and other into employee's perspective. The aim of the current secondary study is to introduce to information science community an emerging area which leverage digital content and implement data analysis techniques to capture the needs of a continuous evolving sector which is software engineering labor market. To the best of our knowledge, this current study can cover the existing gap between the mapping of labor market goals and skills categorization with digital sources and data analysis techniques. As software engineering labor analytics is an unmapped information science area this study will help practitioners understand the basic connections. Additionally, it will help them to find new ways to implement data analysis techniques, or how existing digital sources can be used for a new sector, or to go further into sentiment analysis utilizing sentiment and skills lexicons.

III. SYSTEMATIC MAPPING STUDY METHODOLOGY

In this section, we present the protocol of our systematic mapping study; the steps followed according to the guidelines of Petersen et al. [9].

A. IDENTIFICATION OF THE NEED FOR A SYSTEMATIC REVIEW

The research goal of this study, based on the Goal-Question-Metric (GQM) formulation [26] , can be defined as follows: **Analyze** software engineering literature, **for the purpose of** identifying: (a) what digital labor market data exist; (b) how they can be analyzed; (c) from where such data can be extracted; (d) when is/was the peak of this research field; (e) who is interested in such analysis; and (f) why is this analysis interesting; **from the point of view** of researchers and practitioners, **in the context of** software engineering labor market. Accordingly, we formulated the 6 research questions (RQs) described in the introduction.

B. SEARCH STRATEGY

Sources Selection. Since the interest in the field started to grow during the recent years, there are no dedicated conferences or journals (at a mature level), which could be considered as a targeted source for this SMS. As a result, we decided to perform the automated search procedure in digital libraries, aiming to identify the leading journals or conferences in this new area of data analysis. Specifically, we adopted the inclusion criteria by Dieste et al. [27] : content update (dynamic update with new publications); availability (provide access to the full text of every research article); quality of results (test the accuracy of results returned from the query process using a small list of expected publications which are set by our team from empirical search); and versatility export (since there is a lot of noise in the retrieved results, there is need to remove it so as to extract the related primary studies).

Taking into consideration the selection of well-known digital libraries used from other similar secondary studies [9], [28], we used: ACM Digital Library, IEEE Xplore, Science Direct, and Springer Link. Additionally, we also included the general source Scopus, which enabled us to ensure the identification of as many primary studies as possible.

Search String. We developed a systematic strategy for constructing the search query. The search string should identify studies combining the two areas of interest. The first part contains general words related to labor market sector. We selected the term "skill" to collect primary studies related to hard- or soft-skills. Additionally to "skill", other synonyms are "competency" or "competence"; that is why we included the term "competenc*" in the search string. Furthermore, we added "labor market" or "labour market" as we identified some selected target primary studies using both terms interchangeably.

The second part of the search string concerns digital content. We included general terms such as "social media" or "job portals". Also we added the names of popular sites. Finally, there are primary studies which use as source for information job portals but they use alternative terminology to express it so we added alternative terms with respect to the content, such as "job advertisement/offer/vacancy". Thus, the final search string is: ("skill" OR "competenc*" OR "labor market" OR "labour market") AND ("social network" OR "social media" OR "job portal" OR "LinkedIn" OR "Twitter" OR "GitHub" OR "StackOverflow" OR "Stack Overflow" OR "Facebook" OR "XING" OR "Career Builder" OR "Ca-

[9] https://www.jobvite.com/wp-content/uploads/2015/09/jobvite_recruiter_nation_2015.pdf.

reerBuilder" OR "job advert*" OR "job vacanc*" OR "job offer")).

We have to note here that since the study aimed to software engineering labor market and given that the related terms are too many, the exclusion of unrelated papers was conducted at a next stage manually.

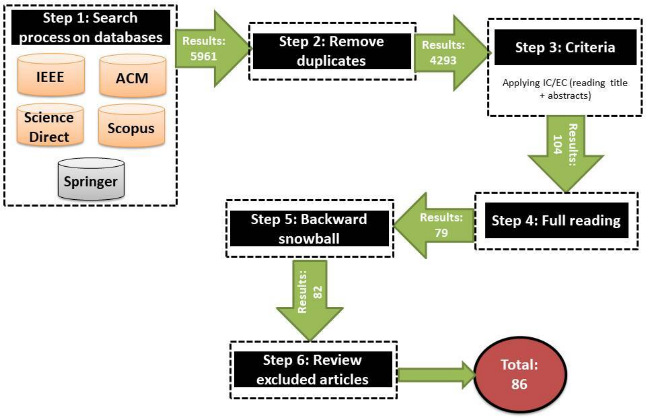
Step 1: Search process on databases
IEEE
ACM
Science Direct
Scopus
Springer
Results: 5961
Step 2: Remove duplicates
Results: 4293
Step 3: Criteria
Applying IC/EC (reading title + abstracts)
Results: 104
Step 4: Full reading
Results: 79
Step 5: Backward snowball
Results: 82
Step 6: Review excluded articles
Total: 86

FIGURE 2. Selection process

C. STUDY SELECTION

Figure 2 gives an overview of the selection process. More specifically, the steps are:

1) Search process on databases.
2) Remove duplicated articles.
3) Apply inclusion/exclusion criteria.
4) Full reading process.
5) Backward snowball sampling, i.e. checking the references of each paper from Step 4, based on inclusion and exclusion criteria.
6) Review excluded articles: Step 3 excludes massively a large number of publications. During this step the excluded articles were reviewed again so as to avoid missing any important information.

The above process is essentially a two-stage screening procedure. In the first stage (Level 1) only the abstracts and titles of the research papers were considered. More specifically, a pre-processing stage removed duplicate articles, either from the same database or from different databases. At this stage, the resulting number of studies was 94. The second stage (Level 2) consisted of applying the inclusion and exclusion selection criteria to the full text of the papers from Level 1. The resulting number of selected papers from this process was 86 and can be found in Table 2 categorized based on bibliographic database.

Regarding the snowball step, there are two options: The first is the forward procedure, where we search new articles that cite the articles of the previous step. The second option is the backward snowball, where we search the list of references of the papers from the previous step so as to find new relevant papers. We chose to apply the backward snowball because the field is relatively new and the bulk of the articles found are from the last two years only.

TABLE 2. Selected primary studies categorized by retrived source

Database	Primary Studies
ACM	[30]–[47]
Elsevier	[14], [48]–[53]
IEEE	[54]–[72] [22], [73]–[84]
Scopus	[21], [85]–[100]
Springer	[20], [101]–[111]

The inclusion criteria were:

- Studies must be written in English.
- Studies must be accessible online document files.
- Studies must be peer-reviewed and published in journals, conferences, workshops and poster sessions of conferences.
- Studies must present experimental results related to data extracted from digital sources and the purpose of their analysis must be relevant to a labor market issue.

The exclusion criteria were:

- Studies that are not related to the field of software engineering.
- Studies that are not written in English.
- Studies that do not provide accessibility to the online full-text.
- Books and gray literature (not published in journals, conferences etc.).
- Studies that present extended abstracts or summaries of conferences/editorials.

D. DATA EXTRACTION

From each primary study a set of variables were collected in order to support the answer of every RQ. In order to ensure that the selection of variables is accurate, two researchers discussed the process and every possible conflict was resolved with all authors. The variables are: V1: Authors V2: Year V3: Title V4: Source V5: Venue V6: Author keywords V7: Type of paper (conference/journal) V8: Used digital source V9: Issue related to labor market V10: A method to analyze the collected data V11: Interested parties/ Stake holders V12: Goal

E. KEYWORDING

Petersen et al. (2008) [29] proposed a classification scheme, which is based on the keywording of abstracts to maintain consistency and optimize the collection. In our case we expected that it could not be possible to find all selected keywords only in titles so we extend the approach to the abstracts.

IV. RESULTS

In this section, we present the results of this study. Section 4.1 presents findings pertaining to digital information sources, Section 4.2 focuses on the types of skills, Section 4.3 discusses the methods used to process data, sections 4.4 and 4.5 focuses on the goals and the interested stakeholders, and finally, Section 4.6 investigates whether the aforementioned results follow a time trend.

A. RQ1: WHERE DIGITAL LABOR MARKET DATA CAN BE EXTRACTED FROM? (SOURCES)

Based on our dataset, we identified that the selected studies have used 51 different digital sources. To detect the most frequently used sources, we mapped them into a generic categorization. Based on Capiluppi et al. [7] classification of social networking sites, we classified six digital sources into three categories (see Table 3). According to our empirical observation there are three ways to collect data for online labor market sources: i) use the official API of site ii) create a customized collector or iii) use a tool for data retrieval.

The first category comprises digital sources related to software development–one of the two subcategories mentioned in Section 2.2. More specifically, it contains the most popular content sharing sites for software developers, i.e., GitHub and the most well-known Q&A Stack Overflow. Job seekers in the software engineering sector are often asked whether they have a Stack Overflow or GitHub account. [103] Having an account could be an asset for a candidate, as it helps hiring managers to evaluate an applicant's expertise level.

Several people do not consider software development sites as providers of raw digital content for labor market analysis. However, as we can see from the results in Table 3, such sites not only provide content, but they are also the leading sources for labor market analytics. Thus, someone could argue that job portals or professional social networks are the most appropriate digital sources for labor market analytics. Upon closer investigation, it becomes apparent that Stack Overflow is not merely a Q&A site but also includes several other features, such as job advertisements, developers' stories, and technological trends, which make it a useful source for online labor market analytics' purposes.

The second category comprises social networks serving professional networking purposes. In this category, LinkedIn is the most frequently used source. Another network in this category is XING, but it exhibits low usage, possibly because the site is only available in German.

The final category comprises job portals. Employers make new positions available to the public through job portals. The most frequently used sources include CareerBuilder, Indeed, and Monster. If we observe the structure of job portals and as opposed to the previous two categories, they only contain job offer posts. Table 3 shows that job portals were the earliest used sources, and they remain popular in online labor market analytics.

Overall, analyzing the labor market structure of the Stack Overflow community, we find many commonalities with LinkedIn. First, Stack Overflow provides every community member with the chance to post or answer questions related to hard skills and achieve knowledge sharing. LinkedIn groups are similar to the ones in Stack Overflow, allowing users to discuss specific interests, but in this case, the groups are general. Second, Stack Overflow provides each user with the chance to create a "developer's story," i.e., a profile where every user can add their curriculum information, such as education or job path, to create a free online CV for job search. It is not mandatory to create an online CV profile to use the Q&A service and participate in the community. However, a LinkedIn user must create a CV profile, even with basic information, to use the network's services. Third, Stack Overflow contains a job section where every user, without having to log in, can find job advertisements related to the ICT sector. Although LinkedIn provides a similar section with many offers from different sectors, it requires a user to log in. Fourth, both sites provide employees the opportunity to create their own business page to attract job seekers who are interested in submitting their resume for open job offers.

Observing the annual performance of these two sources (see Table 3), it becomes apparent that Stack Overflow started gaining popularity since 2017. This increase in popularity might be associated with LinkedIn's decision to provide only general information through its API, whereas Stack Overflow makes everything available through its API. Consequently, LinkedIn has made available limited information for research purposes to compete with Stack Overflow. The software engineering sector, as described in the introduction section, is one of the leading sectors in the labor market and, thus, it is necessary to identify trends related to the audience of this sector.

B. RQ2: WHAT DIGITAL LABOR MARKET DATA CAN BE IDENTIFIED? (SKILLS)

It is not only necessary to categorize skills into hard and soft in labor market analytics, but also explore both categories in depth, as they require different textual data extraction methods. On the one hand, hard skills are acquired through learning/training processes. However, we need to capture new trends in hard skills to gain a competitive advantage in labor supply and market demand. On the other hand, soft skills are difficult to acquire, but they can match a specific candidate to a particular job. Figure 4, shows that in recent years, there has been an increase in research interest in skills analytics. More specifically, in this figure, we have divided skill analytics into the analysis of only hard or soft skills, or both. We aim to detect when the two skills' categories are analyzed together and identify whether soft skills have attracted research interest in the recent years.

The most widespread web sources to export hard-skills are software development sites (see Table 3), such as Stack Overflow and GitHub, which make it is easy to trace hard-skills, as they capture users' activity. More specifically, these sources capture activities and user preferences in programming languages and other software development tools. These sources usually attract programmers, students or researchers who use technology as a tool in their daily work or research activities. Consequently, we can export the hard skills of

TABLE 3. Digital sources

general category	source	2008	2009	2012	2013	2014	2015	2016	2017	2018	2019
Software Development Sites	Stack Overflow	0	0	2	2	1	0	3	5	6	2
	GitHub	0	0	0	1	0	2	2	1	5	0
	jazz repository	0	0	0	0	1	0	0	0	0	0
	wordpress mailing list	0	0	0	0	1	0	0	0	0	0
	drupal mailing list	0	0	0	0	1	0	0	0	0	0
	Oss	0	0	0	0	0	0	1	0	0	0
Job Portals	Indeed	0	0	0	0	0	2	0	1	5	1
	Monster	1	1	1	0	1	0	0	3	3	0
	CareerBuilder	0	0	1	0	0	0	1	1	2	0
Professional Social Networks	LinkedIn	0	0	0	0	2	3	3	6	3	0
	Researchgate	0	0	0	0	0	0	1	1	0	0
Social Networks	Twitter	0	0	0	0	1	0	0	0	1	0
	Facebook	0	0	0	0	1	0	0	0	0	0
Computer Science Bibliographic Database	dblp	0	0	1	0	0	0	0	1	0	0
Other	other job portals and sites	0	0	0	2	0	1	3	4	7	1
	Fortune 500	0	0	0	1	0	0	0	0	0	0

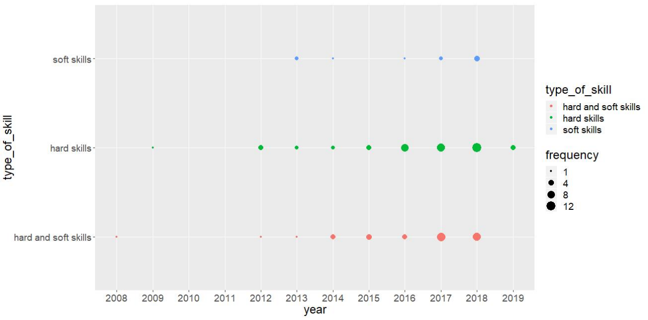

FIGURE 3. Type of skill across the year

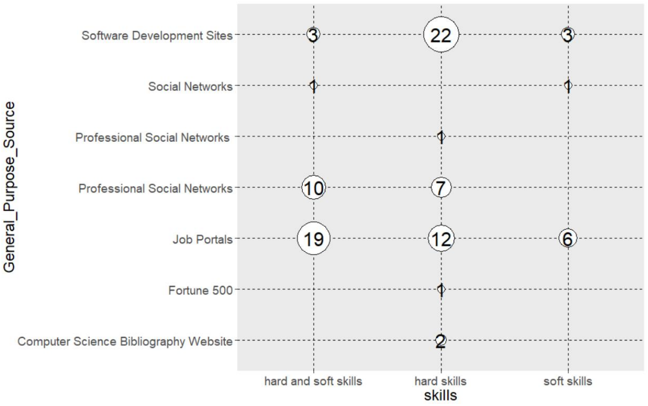

FIGURE 4. Combination of type of skills per digital source

each user from these sites as they tend to tag their expertise. For example, in Stack Overflow, for every programming language, there is a specific label called a tag, which is used by members to group their question and attract other experts in the relevant technology. As it has been already mentioned, although hard skills are necessary, as the knowledge of new technologies contributes to creating competitive advantage in the labor market, there is an increasing interest in soft skills.

Soft skills have received the attention of primary studies during the last five years (see Figure 3). Exporting and analyzing soft skills from web data sources is harder than analyzing and detecting hard skills. Moreover, as there is no specific definition of soft skills in the free text, it is difficult to detect them. The extraction of soft skills can be aided by the use of textual analysis methods and dictionaries or taxonomies related to labor market analytics available from public organizations (e.g. ESCO or O*NET).

In the last half of this decade, researchers started analyzing soft skills using hybrid methods, combining text and numerical/quantitative analysis. More specifically, Figure 3 shows that job portals are the most common web sources to export soft skills. As job portals provide job advertisements, raw data comes from free text. Managing raw text requires several steps to obtain quality results. As we will discuss in the next subsection, techniques to reduce the noise for the text are implemented. Figure 4 indicates that the first attempt to detect hard and soft skills was made in 2008 by McGill [40] via "Monster," which is a job portal. Job portals are appropriate sources not only to detect soft, but also hard skills. Companies describe the technologies they use in their job offers and expect their potential candidates to know or have experience with them.

Considering the increasing research interest in the extraction of soft skills, as shown in Figures 4 and 5, we have maintained records from all primary studies in this SMS regarding which soft skills appear most frequently and accordingly for hard skills (see Table 4). As depicted in Figure 5, "communication skills" are the most demanded soft skills. The primary studies we selected for the current SMS comprise research approaches related to the ICT sector, as it has the highest labor supply and demand. As we know from software technology [5], production has several stages, and every stage requires communication among partners. More specifically, employees need to communicate the outcome of their work to proceed to the next stage or to fix bugs in the testing phase of the production process.

As a complement to communication skills, we identified

Type of skill	Primary studies
Soft	[32], [38], [64], [88], [92], [95], [96], [109], [31], [33], [33], [36], [37], [40], [44], [45], [57], [66], [67], [69]–[72], [75], [76], [79]–[81], [90], [97], [98]
Hard	[14], [20]–[22], [31]–[34], [36], [38], [40]–[43], [45]–[48], [50]–[55], [60], [61], [63], [65], [67], [68], [70], [71], [76], [78], [82]–[84], [87]–[89], [91], [92], [95]–[100], [103], [105], [106], [110], [111]

TABLE 4. Primary studies' separation for hard and soft skills

the second most frequently demanded soft skill–"teamwork". People need to be able to work and interact in teams to obtain the best results. However, if a team wants to be productive and effective, they need to utilize "management skills", which are the third most common soft skills. People who work as a team and communicate must have knowledge of management skills such as time management, planning or networking. Figure 5 presents two other soft skills, which are categorized under management skills–"leadership" and "problem-solving". Additionally, team members need to know how to write, for example technical reports, and have knowledge of foreign languages to communicate with other team members who might not have the same mother tongue. Particularly interesting is the demand for interpersonal skills and personality traits from people. With respect to detecting the two aforementioned skills, some specific lexicons will be discussed extensively in the following subsection.

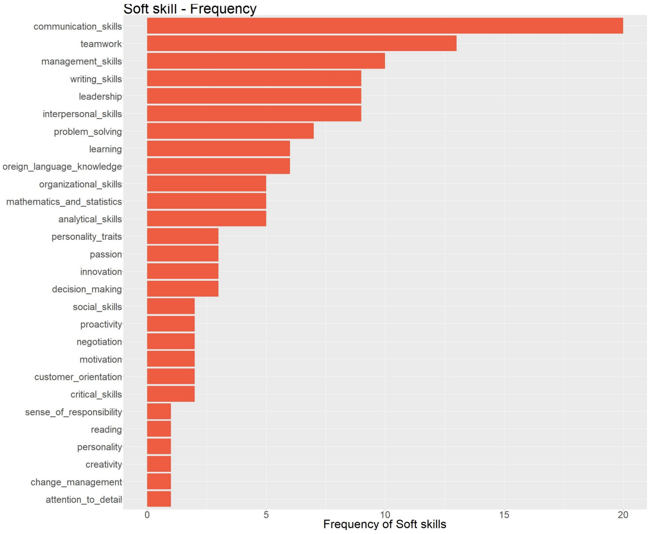

FIGURE 5. Frequency of soft skill

Figure 6 captures the most frequent hard skills (which are detected in more than 10 studies) based on the finding of our SMS. We can observe the existence of five dominant categories of programming languages, namely: i) Object oriented (i.e. "java", "c++"); ii) Data analysis (i.e. "python", "r"); iii) Web development (i.e. "javascript", "html", "php"); iv) procedural computer programming language (i.e. "c"). Other general-purpose hard-skills categories depicted in Figure 6 are: i) operation systems (i.e. "linux") ii) Databases (i.e. "sql", "mysql", "oracle"); and iii) Big data analysis framework (i.e. "hadoop").

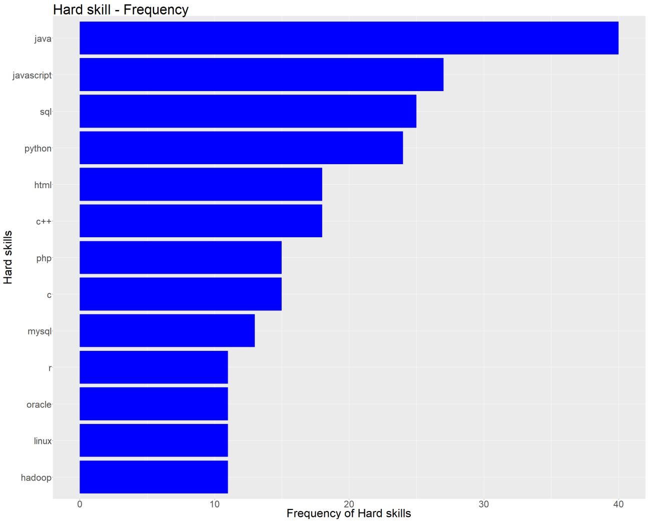

FIGURE 6. Frequency of hard skill

C. RQ3: HOW CAN LABOR MARKET DATA BE ANALYZED? (METHODS)

To extract meaningful results from the raw online labor market data, we adopted Nassirtoussi et al. [112] and Abbe et al. [113] approach: They created an analysis flow process that comprises three stages (see Figure 7): (a) the selection of appropriate web sources for data collection, (b) the cleaning/preprocessing of raw data, and (c) the selection of the analysis methods for the final results.

By analyzing the process step by step, we can outline two types of datasets: the textual and numerical datasets. Textual datasets contain raw texts such as job advertisements collected from online job portals or social networks. The flow arrows in Figure 7 demonstrate that these sources produce only one type of data, i.e., textual data. Professional social networks and software development sites are two other sources of raw text, which provide both textual and numerical datasets, as depicted by the bidirectional flow of the corresponding arrows in Figure 7. Finally, our SMS revealed two more sources, which only provide numerical datasets: "Fortune 500"[10] and DBLP.[11]

After specifying the sources and types of datasets production, another important issue is the preparation of data for analysis, as the raw content may contain a high level of noise due to the existence of too many useless variables. More specifically, textual data may contain stop-words or words

[10] https://fortune.com/fortune500.

[11] DBLP is a website for computer science bibliography: https://dblp.uni-trier.de/.

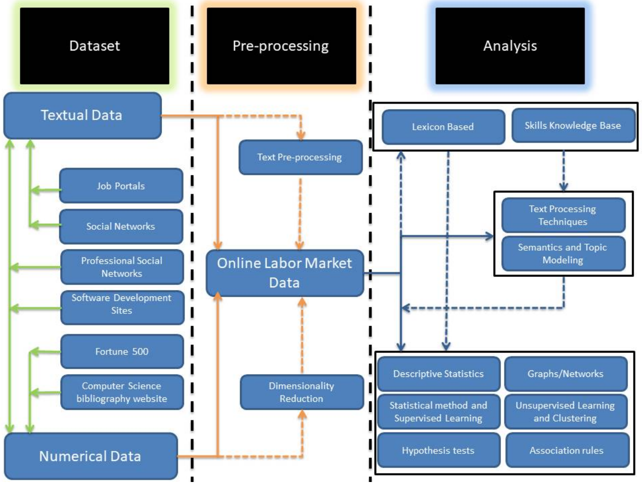

FIGURE 7. Methods flow

with the same root meaning or punctuation. These do not provide meaningful information and reduce the accuracy of the results. Therefore, there is a preprocessing stage, which involves removing stop-words, punctuations or number symbols or illegal characters as part of speech tagging (POS) or tokenization of text. By removing this noise from the text, it is possible to obtain a clean and usable dataset. However, this step is not mandatory. Some research approaches use a predefined list of words to detect whether these words are in the raw text. The cleaning step is necessary when it is not known what is to be leveraged from raw text; it is required to reduce noisy information to improve data quality. Another option of the preprocessing stage is dimensionality reduction. However, this step is also not mandatory. Researchers use it for numerical data with many and highly correlated variables to categorize the information and extract meaningful results.

The final stage of the analysis involves utilizing textual or numerical data analysis methods to extract knowledge depending on the needs of the study (research questions) and the nature of data. From the perspective of textual analysis, if researchers do not use a predefined list of words, as discussed in the previous paragraph, knowledge extraction can be based on lexicons. This approach analyzes specific variables related to human factors from the raw text, such as emotional intelligence items (e.g., personality traits) or hard and soft skills terms. Dictionaries for emotional intelligence aspects (e.g., LIWC or SentiWordNet) and skills (e.g., ESCO, ISCO, e-CF 3.0 or O*NET) are used in such an analysis. This process can produce results that can serve as input for other text or numerical data analysis.

Alternative text analysis methods involve text processing techniques, semantics, and topic modeling. On the one hand, these methods make it possible to obtain results from lexicon-based techniques as input. On the other hand, they directly receive as input the raw text from the preprocessing step. Text processing techniques are characterized by certain techniques such as term frequencies or/and invert term frequencies (tf-idf), n-gram, co-word analysis, etc. A well-known example of semantics and topic modeling methods is Latent Semantic Analysis (LSA) [20]. LSA is a widely used method of topic modeling, which extracts groups of words and specific topics from raw text [20], [108]. The results from such methods can be directly interpreted and presented as findings. Some research approaches use the vector space model [59], [107], e.q., numerical datasets created from the text process, as input for further quantitative methods (see Figure 7) appropriate for numerical data.

Numerical methods contain different data processing techniques such as descriptive statistics, graphs, hypothesis tests, clustering techniques, association rules, statistical models, and machine learning techniques. As already mentioned and depicted in Figure 7, numerical methods receive input in the following forms: (a) direct results from the preprocessing step, (b) numerical tables from lexicon raw text analysis, and (c) the vector model as the result of text processing techniques.

D. RQ4: WHO IS INTERESTED IN THE ANALYSIS OF LABOR MARKET DATA? (STAKEHOLDERS)

To identify the parties that are interested in capitalizing the existing online labor market data, stakeholders are recorded as a variable in our study. We can recognize four categories of stakeholders as depicted in Table 5. The first, corresponding to the majority, comprises companies or organizations that utilize various data sources in relation to job tasks. For example, studies [77], [82] that propose Stack Overflow (stakeholder: firm) should establish an improved recommendation system for answers using data. Another example of stakeholders in the same category is the human resource department (HR managers) who aim to save time by shortening the candidate selection and the suitable matching between the position and a candidate. These studies aim to indicate procedures that can enhance the existing ones. The result is circular, implying that stakeholders themselves benefit from using data, as in the example of CareerBuilder [22]. In general, the stakeholders in this category are legal entities who either wish to hire employees or analyze the trends in skills to obtain a better matching between candidates and the job position, or to update their information systems with labor market data.

The next category of stakeholders includes individual employees/users. More specifically, job seekers search for vacancies that better fit their own characteristics. The employees/users can be members of a community where their activity is monitored and used to produce personal recommendations. For example, developers who are members of the GitHub or Stack Overflow community are offered various facilities such as suggestions of job vacancies based on the users' profile or personality traits capturing their activities; personality traits are part of soft skills.

The third category of stakeholders includes educational institutions such as universities. Some works focus on keeping track of graduates' careers or documenting students' skills.

TABLE 5. Frequency of stakeholders and goals in collected primary studies

Stakeholder	Frq	Goal	Frq
companies and organizations	56	skills analytics	22
employees/users	27	matching job and candidates	16
educational institutions	20	team formation	3
policy makers	7	expert detection	16
		education and learning purposes	5
		cold start problem	2
		women in technology	4
		other	13

Universities primarily aim to recognize contemporary trends in the job market. In this way, the administration and the staff of universities can update its curriculum based on job market data analysis. It is apparent that data can also help researchers, as new problems can result in new techniques of data analysis.

Policymakers comprise the last category of stakeholders. Several works utilize data from job portals to fulfill the needs of European projects managed by CEDEFOP. The contribution of these works is related to monitoring trends in skills and evolving demands of the job market. By accessing data in real time, a policymaker can evaluate training processes and propose improvements in existing policies.

E. RQ5: WHY IS THE ANALYSIS OF LABOR MARKET DATA IMPORTANT? (GOAL)

It is important to specifically recognize the goal of a primary study to identify the needs of stakeholders, which are associated with general trends and requirements of the global labor market. The goals identified from our SMS were divided into eight categories (see Table 5): (a) skills analytics, (b) matching job and candidates, (c) team formation, (d) expert detection, (e) education and learning purposes, (f) cold start problem, (g) women in technology, and (h) other.

"Skills analytics" refers to primary studies that focus on detecting and analyzing hard and/or soft skills. The main aim of this category is to specify types of skills and investigate issues based on these types. This category includes primary studies, which try identifying skill gaps, skill mismatch, and skills in demand or even recommended skills. The subcategory "skills in demand" appears more frequently than the other subcategories, as it is of major importance for different stakeholders. Employees need to know the demands of their future employers to train themselves accordingly and make their CV competitive for the position. Furthermore, universities need to capture the skills on demand to update their courses and sometimes, if mandatory, to change their study programs. Finally, employers want to know the trend in skills demand to plan technological updates and provide new training to their current employees.

The second most frequent goal is to match the demand for specific skills from employers with the skills of a job seeker. In contrast to the previous category, this one involves a specific interaction between employers and employees. The third goal category is team formation. Researchers use digital sources to detect users who will contribute to forming teams with specific skills, complementing each other. Some primary studies aim to detect the expert for a team (the fourth goal category) or a specific job. Expert tracking refers to people who have experience with a specific skill. These candidates are the most appropriate option for a specific project need, such as specialized knowledge in a programming language. In the labor market field, expertise identification is called "talent detection".

A remarkable finding is that digital labor market sources are used for education and learning purposes (the fifth goal category). To achieve this goal, researchers try to compare skills in demand by industry with those taught at a university. Additionally, skills detected as required in the labor market are used as input to update study guides.

The sixth interesting goal found by our SMS was the "cold start problem" which, in the ICT labor market sector, is defined as follows: "the phenomenon that a developer doesn't show any activities on some skills" [46]. In the software engineering sector, technology evolves rapidly, and the skills associated with every technological change also evolve quickly. Due to this rapid evolution, traditional recommendation techniques, such as collaborative filtering, are not suitable for a cold start problem when new recommendations are required. A cold start problem occurs when no one knows how to handle new skills that are introduced.

Regarding "women in technology" (the seventh category), i.e., analyzing women's involvement in technical aspects, we observe that the primary studies aim to analyze the gender gap capturing the participation in software engineering sources. Observing the related studies, we can identify these digital sources to be GitHub and Stack Overflow. Finally, we employed a general (eighth) category, labeled "other", to categorize every primary study where the goal appears only once. Some examples of these goals include "classifying job titles using standard taxonomy of occupations," "duplicate job advertisements detection," etc.

F. RQ6: WHEN ARE ONLINE LABOR MARKET ANALYTICS CHANGING OVER TIME?(TIME)

Online labor market analytics can be applied in every industrial sector, which demands the work force to use digital sources. However, the software engineering sector is a leading sector in the demand and supply of online labor market and revealed to be a valuable implementation to investigate fundamental aspects of online labor market analytics. Leveraging primary studies from the aforementioned field we detect that the research interest started in the last 10 years. Capturing the outlook between 2008-2013 we can observe from Table 3 that the use of digital sources started to rise after 2012. Additionally, in the second half of the decade between 2013-2018 the domain attracted the interest of the research community. During the last two years researchers started to give special interest to the investigation of soft-skills.

From the aforementioned observation, we suggest that one important expectation for software engineers is the continuous training and evolution of their soft-skills. An additional expectation would be the simplification of the detection of developers' hard- and soft-skills is their active participation in software engineering communities (e.g. Stack Overflow).

To answer RQ6, we first plot the number of studies that use digital information sources for labor market analysis (see Figure 8). Although we identified publications in 2019, we excluded them from the specific plot, as we cannot yet include data for the entire year. The number of published papers is increasing rapidly since 2012 with an annual percentage growth rate of 49.53%. The publications are divided into conference proceedings (67.85%) and journal papers (32.15%). In 2018, the annual growth rate of journal papers was 16.67%, signifying an evolving and promising trend of academic interest. The rest of this subsection is organized based on the aspects of each research question.

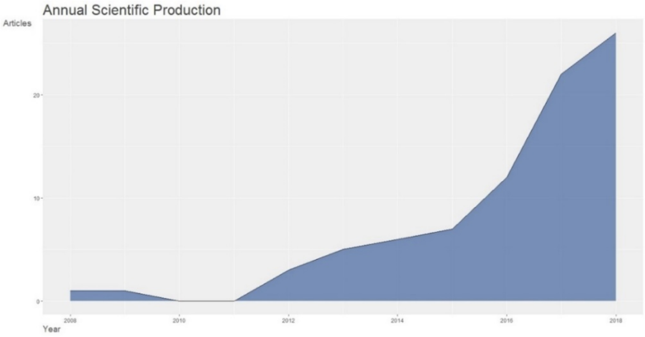

FIGURE 8. Annual Scientific Production

Sources over the years: As we can see in Table 3, "job portals" (e.g., CareerBuilder) are the sources which initially attracted the interest, starting from 2008. Gradually, their use declined but they remain a reliable source as their content can be updated in real time. Another important finding is the growing interest in the content of "software development sites" since 2012 and especially since the last three year as their use is increasing steadily. This happens probably due to the increasing needs and complexity of the software engineering industry. Domain-specific sites such as Stack Overflow, containing communities, have been recognized as wealthy sources of information, which can provide insights into specific labor market needs. We can note that digital sources which focus in communities from specific sectors, such as software development sites attract more interest. Sources provide information related to the software engineering sector and possibly attracted more research interest as they provide the desired content. Digital sources such as professional or social networks (e.g. LinkedIn, Twitter) provide information for many sectors making the extraction of focused results harder due to the noisy data. Finally, it's worth mentioning that "computer science bibliography" sites (e.g., DBLP) attracted research interest during the early years, but over the last five years, interest declined, as they only focus on academic publications and provide information about bibliometrics or scientific collaborations.

Skills over the years: Analyzing skills is directly related to the suitability of specific employees for specific jobs. In Figure 3, we can see that interest in hard skills emerged in 2009. However, a year ago, the first study identified by our SMS combined both hard and soft skills, showing the emerging importance of both categories in labor market analytics. From 2013 onwards, research approaches gradually started to focus on detecting soft factors. Detecting soft skills is related to individual personalities, and it is a complicated and controversial process. From 2018, LinkedIn started categorizing skills into categories such as industry knowledge, tools and technologies, interpersonal skills, etc. As stated in the introduction, LinkedIn has captured the importance of soft skills. Consequently, we can empirically observe that LinkedIn has started distinguishing factors belonging to soft skills, such as interpersonal skills.

Methods over the years: By observing Figure 6, which depicts the time trends of different methods for online labor market analytics, we can extract some interesting trends. The most frequently used methods include "text techniques" and "statistical analysis methods". Text analysis techniques began being used in 2008, and their use has remained continuous throughout the years. More specifically, Table 3 shows that the most frequently used digital source is a "job portal" (initially used in 2008), which contains free text; thus, the appropriate methods to extract knowledge from raw text are text-mining techniques. The raw text from job advertisements is the most suitable to extract soft skills from job offers and, as we can observe from Figure 3, research interest in this field began in 2008. Additionally, it is interesting to illustrate the evolution of lexicon-based, semantics, skills knowledge base, text preprocessing, and topic modeling methods which are recorded to have a low level of utilization. The reason of their low use (see Figure 7) is that such methods usually support other text analysis techniques (by cleaning text in the preprocessing step or by receiving input from text analysis techniques), and they are considered auxiliary, as they are used to optimize other more standard text analysis techniques. The lexicon-based (e.g., LIWC) and skills-knowledge-based (e.g., ESCO) methods support extracting vector data matrices from the raw text as input for quantitative (numerical) methods. Again, these are considered auxiliary and have garnered low interest over the years. Regarding the second most frequently used methods, the statistical and machine learning techniques, Figure 6 shows that their use started in 2012. This is related to the digital content from software engineering sites (see also Table 3), which became a trend that same year. These sources provide both numerical and textual data; as most primary studies collect numerical raw data from them, quantitative analysis techniques such as statistics and machine learning were applied.

Stakeholders over the years: Figure 9 depicts the time evolution of beneficiaries (stakeholders) of the selected primary studies. Low frequencies are especially reported for

TABLE 6. Methods over the years

method	2008	2009	2012	2013	2014	2015	2016	2017	2018	2019
association rules	0	0	0	0	0	0	0	1	0	1
clustering	0	0	0	1	1	0	1	1	3	0
descriptive statistics	0	0	1	1	2	1	1	3	4	0
dimensionality reduction	0	0	0	0	1	0	0	2	0	0
graphs networks	0	0	0	1	0	1	2	4	3	0
hypothesis tests	0	0	1	2	2	0	0	1	4	0
lexicon based	0	0	0	1	2	0	0	0	2	0
neural networks	0	0	0	0	0	0	0	1	2	0
semantics	0	0	0	0	0	1	2	2	2	0
similarity distance measures	0	0	0	0	0	2	1	2	1	0
skills knowledge base	0	0	0	0	0	0	3	5	3	0
statistical and machine learning methods	0	0	1	1	3	0	3	5	8	3
text preprocessing	0	0	0	0	0	0	2	1	4	1
text techniques	1	1	1	3	2	4	8	9	11	0
topic modeling	0	0	0	0	2	0	1	3	1	0

policymakers who appeared only recently. This trend is related to the Cedefop project for labor market analytics (described earlier), which has been publishing primary studies since 2017. The most frequently mentioned stakeholders include employees/users, organizations, and universities. Figure 9 shows that a university is the stakeholder of the first primary study, which used job portals to detect hard and soft skills. Additionally, we can see early interest in results related to employees/users, as job portals are the main sources to extract information related to skills. Firms and organizations are found to be the beneficiaries of the results of the primary studies from 2012. By observing Table 3, we can connect this finding to the start of utilizing digital sources related to employees' profiles.

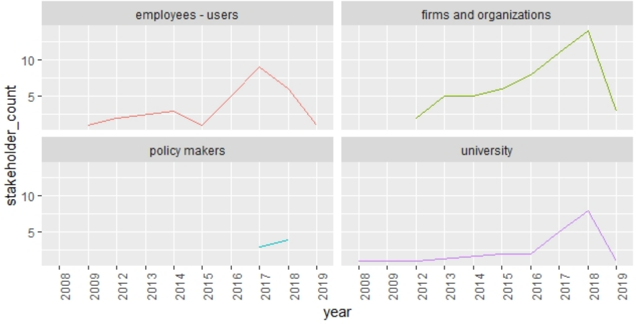

FIGURE 9. Stakeholders over the years

Goals over the years: Figure 10 shows that skills analytics started being reported as the goal of primary studies in 2008. The first primary study tried identifying skills' demand in job advertisements using text analysis techniques. Skills analytics has remained a research goal, garnering constant interest, as the related source type (job portals) also remains popular due to rapid technological development. Team formation and expertise detection began appearing as the goal of many primary studies from 2012. They are related to the appearance of firms and organizations as stakeholders. Furthermore, information regarding team formation and developers' expertise has been found in digital sources containing profile characteristics, CVs or developers' activity. These sources were first utilized in 2012. Finally, what's most interesting is the time trend regarding the role of women in software engineering. In 2012, interest in studying the activity of women in software engineering sites emerged and it became popular again in 2018 due to generic interest in the gender gap. Another finding concerns education as a goal, which attracted high interest early in 2008, but later this interest began declining.

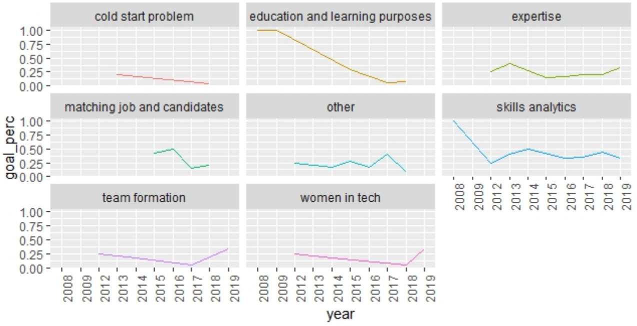

FIGURE 10. Goals over the years

V. DISCUSSION

In the current SMS, we examined studies related to labor market analytics utilizing digital content. We identified a profusion of publications, in total 86 primary studies. An interesting observation is that the large majority of primary studies were published in conferences; this finding might be an indicator of a possible lack of mature studies in the field. Only in the last two years, journals publications appeared, possibly due to the increasing interest for utilization of digital content for labor market analytics. As an interpretation of the results of the current study presenting four indicative usage scenarios, based on the stakeholders that we have identified in Section 4.4 (see Table 5). Every stakeholder has a specific need(s), which are retrieved based on the categorization of goals (see Section 4.5 - Table 5). Through these scenarios we present how a stakeholder could utilize this mapping study for real-word cases:

Guidance for Early Stage Researchers (stakeholder: Employees/Users): An early stage researcher or a PhD student could utilize the results of this secondary study to shape his/her research directions without having to perform time-consuming reviews for the discovery of new challenges and research gaps. Let's assume that the researcher finds interesting to study the "women in tech" goal, indicating that the gender gap in the software engineering sector is an active field of research. Researcher chooses to deal with both hard and soft skills. The most appropriate choice for digital sources is the software engineering sites where related information regarding women's profiles and participation activities can be found. Since the content is both textual (text comments in GitHub, question and answers in Stack Overflow community) and numerical (data related to frequency and quality of participation and also binary data for having or not a hard skill), the appropriate methods is the cleaning/noise removal preprocessing stage. Hence, statistical and machine learning methods are useful for numerical data, while lexicon-based and topic modeling techniques can be applied for soft skills.

Industrial Hiring Processes (stakeholder: Companies/Organizations): A CEO working in a software engineering company in the field of serious games decides that the company needs to train the existing staff and hire new people. The CEO in collaboration with the experts of the human resources department set the goals of "skills analytics" and "expert detection". They decide to detect hard and soft skills and the appropriate digital source according to the set goals are job portals and software engineering sites. It is necessary to preprocess the raw free text to reduce the noise and later to analyze the text using term frequency and topic modeling. Additionally for expertize identification "network/graph" methods can be implemented.

Improvement in Higher Education (stakeholder: Educational Institutions): The head of the "Statistics" lab of a department of Mathematics decided that there is a need to update the current teaching plan of the courses that lab supports and they are related to statistical analysis. As a result, s/he conducts an investigation aiming to identify what skills their graduates use in their current work, or record in their online curriculum profiles. So, the goal here is clearly "education and learning purposes" as the lab needs to identify which new technologies their graduates use in their working environment. This means that hard skills need to be considered. The appropriate source for hard skills of specific people is decided to be a professional social network (e.g. LinkedIn) containing either brief profiles or even detailed CVs. The analysis can start from simple descriptive statistics so as to identify the most frequent technologies related to the subjects of the lab, and more interestingly, to detect technologies that graduates had to learn after their graduation. Further hypothesis tests can be performed in order to find relations among hard skills or between hard skills and other demographic characteristics.

Adoption by Policy Makers (stakeholder: Policy Makers): A policy maker who works in the continuous update of the eCF 3.0 framework finds an interest in this secondary study and identifies that there are some specific primary studies which use the ESCO framework. The goal here is "skills analytics" and "matching job openings and candidate profiles". To support these goals both types of skills are considered. According the current mapping review the most appropriate digital source is the job portals. Leveraging the free text content of job advertisements, the policy maker can identify the job titles and the skills included in the eCF 3.0 taxonomy, and how these can be mapped to the job titles and skills description in the content of job offers. This map requires preprocess of the text before the analysis which involves both textual and numerical methods. Again, textual methods such as term frequencies or text semantic similarities can identify the soft skills, which are required by employers and map them to the recorded skills of eCF 3.0 taxonomy. Additionally s/he can categorize the different job titles again mapping them to the current structure of taxonomy and using descriptive statistics and statistical and machine learning methods to classify and match job titles and skills.

Against these usage scenarios we can identify some key scientific dimensions of the problems that need to be tackled:

Cross platform analysis: The results of this map showed the structure of digital sources depending on the labor market needs. A gap exists in the usage of multiple sources to provide deeper analysis. For example, cross platform identification techniques which will detect different social profiles of a developer could provide a better understanding for her soft skills. Using cross platform approaches will affect the use of methods, as more complex ones such as deep learning will be needed. Also, cross platform analysis between sources of employer and employee could be a challenge. Interestingly, very few works (e.g., Gousios et al (2015)) tried to match developer profiles and job offers. Further research is needed in the detection of developer's expanding of skillset trends and job market demand.

Smart job mobility: During the last decade, smart cities attracted the interest of the research community. Different solutions have been implemented in smart cities such as transportation, traffic or environmental metrics. However, people who live in these cities seek for new job opportunities or gain new knowledge. Smart job mobility could be a new challenge where smart dimensions would be expanded to support the training of city's workforce through open lab meetings. Hence, smart applications will capture the job trends pulse of the city.

Multimedia digital content: While we noted that primary studies use textual and numeric data there is no utilization of multimedia sources. Empirical observations can reveal that companies post new job vacancies on Instagram using photos, or job candidates post their CV video on YouTube. The uptake of multimedia digital labor content will arise the need of different sources and methods for online labor market analytics.

Evolutionary data analytics: In job advertisements analysis there is a gap as history of job posts doesn't exist. Leveraging historic data as open source datasets could help researchers analyze new skill trends or emerging sectors. Historic record of job offers is a challenging task as it requires implementation of big data storage and analytics methods, but it can help towards evolving labor market analytics and predictions.

Ontology: Capturing labor market trends from the content of online job advertisements is an important aspect. In this paper, we identified primary studies which utilize existing labor market ontologies such as ESCO [101], [107] or O*NET [49] to map job titles from collected advertisements with existing titles in these ontologies. However, a new challenge is the creation of a new ontology which describes the whole digital labor content. As the current ontologies were created with different empirical process they should be tested and updated based on the new digital era of labor market.

Learning analytics: we identified that the goal of a set of primary studies is related to education purposes. Most approaches analyze digital content to provide skill trend analytics for possible updates of courses, or to capture alumni network. The challenge here is to connect emerged skills with recommendations in learning processes through the identification of different learning styles.

VI. THREATS TO VALIDITY

In the context of systematic mapping reviews, some of the implementation steps of the framework can be characterized by subjectivity, meaning that decisions have been influenced by specific viewpoints of the researchers. As a result, changes may be required in future replications of the study in order to generalize the findings. The current section aims to analyze possible threats to validity with respect to the following issues: (i) study selection (ii) data validity and (iii) research validity. [114]

A. STUDY SELECTION VALIDITY

To ensure that our search strategy approach is consistent to the principles of relevant studies, we selected to follow the strict guidelines of a well-defined protocol [9] for SMS. Specifically:

- The automated collection and identification process of primary studies involved searching in the most well-known digital libraries containing publications of highly reputed journals and conferences.
- The generic terms of the query string ensured the collection of a high number of related publications, reducing the bias. On the other hand, by construction, using the boolean operator (AND) with terms related to digital sources, we reduced the noise.
- The list of inclusion and exclusion criteria has been extensively discussed among all authors to guarantee clarity and to prohibit misinterpretations.
- To mitigate the risk of missing (not including) relevant studies we adopted a "gold standard", consisting of 8 papers that we knew from our experience that they are relevant to the subject of the secondary study. We used them as benchmark to evaluate whether the results of the constructed string query were able to find these papers. All selected papers have been successfully returned. Additionally, in the collection process we mitigated the threat of grey literature by excluding massive databases, such as Google Scholar in the collection step.
- We did a very systematic work regarding the removal of duplicates. The same articles appearing more than once were found and removed. Furthermore, we applied a similarity detection procedure for the abstracts using an automated procedure. In case of high similarity, the papers were inspected manually by the first author. In this way, preliminary conference publications, prior to the main journal publication were found and removed.
- No paper was missed due to lack of access, since our research institutions provide full access databases.
- We applied language detection using R and in case of non-English text, the publication was removed.
- The threat of falsely excluding relevant articles was mitigated. by discussion between two of the researchers on controversies of interpreting the list of inclusion/exclusion criteria. Also, a third author screened manually a subset of randomly selected primary studies to validate the selection based on inclusion criteria.

B. DATA VALIDITY

The main threat here is the extraction bias, related to the threat of subjectivity. Although the data handling was manually accomplished by the first author, a revalidation process conducted by pairs of researchers was applied so as to mitigate bias. After this internal validation, all authors discussed the results and resolved conflicts. Furthermore, the selection of the digital libraries with publications undergone peer review is a guarantee of reduced publication bias. Regarding data validity there are some more issues which in this study are not considered as threats:

- Small sample size: The 86 primary studies offer a considerable amount of data
- Lack of relationship: There was no intention to prove relationship between variables
- Low quality of primary studies: The sources of the selected primary studies guarantee scientific quality.
- Selection of variables to be extracted: All authors discussed and identified the needs of the current research; the variables were mapped to the RQs which were based on the 5W + 1H questions model for problem solving.
- Researcher's bias in interpretation and analysis: All authors discussed thoroughly the findings of every RQ and provided reasonable explanations avoiding speculation and unnecessary generalizations.

C. RESEARCH VALIDITY

Regarding research validity there are two possible threats: The first is the experience of the author team. The last three authors are very experienced researchers in the field of empirical software engineering and they have been involved as authors and reviewers in a large number of empirical studies, including secondary studies. Additionally, all important steps of the study followed a strict protocol and all decisions were taken after extensive discussions and complete agreement. The second threat is the generalizability. This threat does not have any meaning in the current study as our efforts were focused on depicting the trends in literature with respect to a certain topic. The same research can be repeated after some time, so as to update the knowledge on how the interest on the topic evolves over time.

VII. CONCLUSIONS

During the previous decades, software engineering has evolved to a sector of major importance in the labor market as it offers every year several thousands of jobs worldwide. Labor market is a continuously developing research area from various perspectives, with great potentials for new directions and challenging problems. It is directly connected with data analysis and since the related data sources become exclusively electronic and web-based, there is a growing interest for knowledge extraction from these data sources via labor market analytics.

The current paper deals with the field of labour market analytics. We focused on the software engineering labor market and on studies appearing in the scientific literature. These primary studies use electronic data sources for extracting knowledge for various goals, for the benefit of different stakeholders and by using different data analysis methods. The characteristics of the human factor, especially the skills constitute the main content of the data and the subject of the subsequent analysis. To the best of our knowledge the systematic mapping study we present here is the first in this field and aims to contribute to the development of a relatively new field of research which has a special interest both for academia and industry.

The lessons learnt from this study involve identification and categorization of: 1) potential electronic data sources, 2) skills contained in these sources in various forms, 3) methods used for the manipulation and the analysis of data, 4) goals of such studies and 5) stakeholders who are benefited from market analytics. Furthermore, the aforementioned categorizations were investigated with respect to time, i.e. we examined how all these components are addressed by researchers year by year, starting from the year of the first publication, up to the current year. Having categorized all these components of the primary studies, we showed via hypothetical scenarios how the current study could facilitate the planning of new studies on the subject, either by academia or industry. It is expected that the interest in this field will be growing over the following years, so the current study has the potentials to stimulate further research.

ACKNOWLEDGMENT

This research is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme "Human Resources Development, Education and Lifelong Learning" in the context of the project "Strengthening Human Resources Research Potential via Doctorate Research" (MIS-5000432), implemented by the State Scholarships Foundation (IKY)

REFERENCES

[1] C. Hammer, M. D. C. Kostroch, and M. G. Quiros, Big Data: Potential, Challenges and Statistical Implications. International Monetary Fund, 2017.

[2] R. Boselli, M. Cesarini, F. Mercorio, and M. Mezzanzanica, “Classifying online job advertisements through machine learning,” Future Generation Computer Systems, vol. 86, pp. 319–328, 2018.

[3] P. Tissot and C. européen pour le développement de la formation professionnelle, Terminology of vocational training policy: a multilingual glossary for an enlarged Europe. Office for official publications of the European Communities Luxembourg, 2004.

[4] M. Blázquez, “Skills-based profiling and matching in pes,” Publications Office of the European Union, Luxembourg, 2014.

[5] D. Spinellis, “Being a software developer,” IEEE Software, vol. 35, no. 4, pp. 4–7, 2018.

[6] M. Zhang, H. Chen, and A. Luo, “A systematic review of business-it alignment research with enterprise architecture,” IEEE Access, vol. 6, pp. 18 933–18 944, 2018.

[7] A. Capiluppi, A. Serebrenik, and L. Singer, “Assessing technical candidates on the social web,” IEEE software, vol. 30, no. 1, pp. 45–51, 2012.

[8] E. Colombo, F. Mercorio, and M. Mezzanzanica, “Ai meets labor market: exploring the link between automation and skills,” Information Economics and Policy, 2019.

[9] K. Petersen, S. Vakkalanka, and L. Kuzniarz, “Guidelines for conducting systematic mapping studies in software engineering: An update,” Information and Software Technology, vol. 64, pp. 1–18, 2015.

[10] M. Amadoru and C. Gamage, “Evaluating effective use of social networks for recruitment,” in Proceedings of the 2016 ACM SIGMIS Conference on Computers and People Research. ACM, 2016, pp. 125–133.

[11] L. M. Kureková, M. Beblavy, and A.-E. Thum, “Using internet data to analyse the labour market: a methodological enquiry,” IZA Discussion Papers, Tech. Rep., 2014.

[12] S. Onoue, H. Hata, R. G. Kula, and K. Matsumoto, “Human capital in software engineering: A systematic mapping of reconceptualized human aspect studies,” arXiv preprint arXiv:1805.03844, 2018.

[13] M. Mezzanzanica and F. Mercorio, “Big data for labour market intelligence: an introductory guide,” 2019.

[14] A. De Mauro, M. Greco, M. Grimaldi, and P. Ritala, “Human resources for big data professions: A systematic classification of job roles and required skill sets,” Information Processing & Management, vol. 54, no. 5, pp. 807–817, 2018.

[15] F. Colace, M. De Santo, M. Lombardi, F. Mercorio, M. Mezzanzanica, and F. Pascale, “Towards labour market intelligence through topic modelling,” in Proceedings of the 52nd Hawaii International Conference on System Sciences, 2019.

[16] G. Matturro, F. Raschetti, and C. Fontán, “A systematic mapping study on soft skills in software engineering.” J. UCS, vol. 25, no. 1, pp. 16–41, 2019.

[17] M. Sánchez-Gordón and R. Colomo-Palacios, “Taking the emotional pulse of software engineeringŮa systematic literature review of empirical studiestaking the emotional pulse of software engineeringŮa systematic literature review of empirical studies,” Information and Software Technology, 2019.

[18] K. B. Backhaus, “An exploration of corporate recruitment descriptions on monster. com,” The Journal of Business Communication (1973), vol. 41, no. 2, pp. 115–136, 2004.

[19] A. Capiluppi and A. Baravalle, “Matching demand and offer in on-line provision: A longitudinal study of monster. com,” in 2010 12th IEEE International Symposium on Web Systems Evolution (WSE). IEEE, 2010, pp. 13–21.

[20] S. Debortoli, O. Müller, and J. vom Brocke, "Comparing business intelligence and big data skills," Business & Information Systems Engineering, vol. 6, no. 5, pp. 289–300, 2014.

[21] F. Javed, P. Hoang, T. Mahoney, and M. McNair, "Large-scale occupational skills normalization for online recruitment," in Twenty-Ninth IAAI Conference, 2017.

[22] W. Zhou, Y. Zhu, F. Javed, M. Rahman, J. Balaji, and M. McNair, "Quantifying skill relevance to job titles," in 2016 IEEE International Conference on Big Data (Big Data). IEEE, 2016, pp. 1532–1541.

[23] B. Curtis, Tutorial, human factors in software development. IEEE Computer Society Press, 1990.

[24] B. Vasilescu, Y. Yu, H. Wang, P. Devanbu, and V. Filkov, "Quality and productivity outcomes relating to continuous integration in github," in Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering. ACM, 2015, pp. 805–816.

[25] A. Murgia, M. Ortu, P. Tourani, B. Adams, and S. Demeyer, "An exploratory qualitative and quantitative analysis of emotions in issue report comments of open source systems," Empirical Software Engineering, vol. 23, no. 1, pp. 521–564, 2018.

[26] V. R. B. G. Caldiera and H. D. Rombach, "The goal question metric approach," Encyclopedia of software engineering, pp. 528–532, 1994.

[27] O. Dieste, N. Juristo, and M. D. Martínez, "Software industry experiments: A systematic literature review," in Proceedings of the 1st International Workshop on Conducting Empirical Studies in Industry. IEEE Press, 2013, pp. 2–8.

[28] F. Salo, M. Injadat, A. B. Nassif, A. Shami, and A. Essex, "Data mining techniques in intrusion detection systems: A systematic literature review," IEEE Access, vol. 6, pp. 56 046–56 058, 2018.

[29] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, "Systematic mapping studies in software engineering." in Ease, vol. 8, 2008, pp. 68–77.

[30] I. Adaji and J. Vassileva, "Personalizing social influence strategies in a q&a social network," in Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization. ACM, 2017, pp. 215–220.

[31] M. Bastian, M. Hayes, W. Vaughan, S. Shah, P. Skomoroch, H. Kim, S. Uryasev, and C. Lloyd, "Linkedin skills: large-scale topic extraction and inference," in Proceedings of the 8th ACM Conference on Recommender systems. ACM, 2014, pp. 1–8.

[32] S. Chelaru, E. Herder, K. D. Naini, and P. Siehndel, "Recognizing skill networks and their specific communication and connection practices," in Proceedings of the 25th ACM conference on Hypertext and social media. ACM, 2014, pp. 13–23.

[33] J. Ehlers, "Socialness in the recruiting of software engineers," in Proceedings of the 12th ACM international conference on computing frontiers. ACM, 2015, p. 33.

[34] G. J. Greene and B. Fischer, "Cvexplorer: Identifying candidate developers by mining and exploring their open source contributions," in Proceedings of the 31st IEEE/ACM International Conference on Automated Software Engineering. ACM, 2016, pp. 804–809.

[35] B. V. Hanrahan, G. Convertino, and L. Nelson, "Modeling problem difficulty and expertise in stackoverflow," in Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion. ACM, 2012, pp. 91–94.

[36] C. Hiranrat and A. Harncharnchai, "Using text mining to discover skills demanded in software development jobs in thailand," in Proceedings of the 2nd International Conference on Education and Multimedia Technology. ACM, 2018, pp. 112–116.

[37] S. A. Licorish and S. G. MacDonell, "Personality profiles of global software developers," in Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering. ACM, 2014, p. 45.

[38] C. Litecky, A. J. Igou, and A. Aken, "Skills in the management oriented is and enterprise system job markets," in Proceedings of the 50th annual conference on Computers and People Research. ACM, 2012, pp. 35–44.

[39] S. Marrara, G. Pasi, M. Viviani, M. Cesarini, F. Mercorio, M. Mezzanzanica, and M. Pappagallo, "A language modelling approach for discovering novel labour market occupations from the web," in Proceedings of the International Conference on Web Intelligence. ACM, 2017, pp. 1026–1034.

[40] M. McGill, "Critical skills for game developers: an analysis of skills sought by industry," in Proceedings of the 2008 conference on future play: Research, play, share. ACM, 2008, pp. 89–96.

[41] O. Odiete, T. Jain, I. Adaji, J. Vassileva, and R. Deters, "Recommending programming languages by identifying skill gaps using analysis of experts. a study of stack overflow," in Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization. ACM, 2017, pp. 159–164.

[42] D. M. Shankaralingappa, G. De Fransicsi Morales, and A. Gionis, "Extracting skill endorsements from personal communication data," in Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. ACM, 2016, pp. 1961–1964.

[43] E. M. Sibarani, S. Scerri, C. Morales, S. Auer, and D. Collarana, "Ontology-guided job market demand analysis: a cross-sectional study for the data science field," in Proceedings of the 13th International Conference on Semantic Systems. ACM, 2017, pp. 25–32.

[44] M. Stevens and R. Norman, "Industry expectations of soft skills in it graduates: a regional survey," in Proceedings of the Australasian Computer Science Week Multiconference. ACM, 2016, p. 13.

[45] J. Wan, B. Chen, and H. Si, "Mining and measurement of vocational skills and their association rules based on big data," in Proceedings of the International Conference on Digital Technology in Education. ACM, 2017, pp. 59–63.

[46] J. Yan, H. Sun, X. Wang, X. Liu, and X. Song, "Profiling developer expertise across software communities with heterogeneous information network analysis," in Proceedings of the Tenth Asia-Pacific Symposium on Internetware. ACM, 2018, p. 2.

[47] A. Santos, M. Souza, J. Oliveira, and E. Figueiredo, "Mining software repositories to identify library experts," in Proceedings of the VII Brazilian Symposium on Software Components, Architectures, and Reuse. ACM, 2018, pp. 83–91.

[48] C. Huang, L. Yao, X. Wang, B. Benatallah, and X. Zhang, "Software expert discovery via knowledge domain embeddings in a collaborative network," Pattern Recognition Letters, 2018.

[49] I. Karakatsanis, W. AlKhader, F. MacCrory, A. Alibasic, M. A. Omar, Z. Aung, and W. L. Woon, "Data mining approach to monitoring the requirements of the job market: A case study," Information Systems, vol. 65, pp. 1–6, 2017.

[50] U. P. K. Kethavarapu and S. Saraswathi, "Concept based dynamic ontology creation for job recommendation system," Procedia computer science, vol. 85, pp. 915–921, 2016.

[51] M. Neshati, Z. Fallahnejad, and H. Beigy, "On dynamicity of expert finding in community question answering," Information Processing & Management, vol. 53, no. 5, pp. 1026–1042, 2017.

[52] H. Pérez-Rosés, F. Sebé, and J. M. Ribó, "Endorsement deduction and ranking in social networks," Computer Communications, vol. 73, pp. 200–210, 2016.

[53] P. Rostami and M. Neshati, "T-shaped grouping: Expert finding models to agile software teams retrieval," Expert Systems with Applications, vol. 118, pp. 231–245, 2019.

[54] F. Gurcan and N. E. Cagiltay, "Big data software engineering: Analysis of knowledge domains and skill sets using lda-based topic modeling," IEEE Access, vol. 7, pp. 82 541–82 552, 2019.

[55] R. Arora, S. Goel, and R. Mittal, "Supporting collaborative software development in academic learning environment: A collaborative pair and quadruple programming based approach," in 2017 Tenth International Conference on Contemporary Computing (IC3). IEEE, 2017, pp. 1–7.

[56] Y. Bachrach, "Human judgments in hiring decisions based on online social network profiles," in 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA). IEEE, 2015, pp. 1–10.

[57] B. Bazelli, A. Hindle, and E. Stroulia, "On the personality traits of stackoverflow users," in 2013 IEEE international conference on software maintenance. IEEE, 2013, pp. 460–463.

[58] H. Burk, F. Javed, and J. Balaji, "Apollo: Near-duplicate detection for job ads in the online recruitment domain," in 2017 IEEE International Conference on Data Mining Workshops (ICDMW). IEEE, 2017, pp. 177–182.

[59] S. Chala and M. Fathi, "Job seeker to vacancy matching using social network analysis," in 2017 IEEE International Conference on Industrial Technology (ICIT). IEEE, 2017, pp. 1250–1255.

[60] E. Constantinou and G. M. Kapitsaki, "Developers expertise and roles on software technologies," in 2016 23rd Asia-Pacific Software Engineering Conference (APSEC). IEEE, 2016, pp. 365–368.

[61] C. Constantinov, P. Ş. Popescu, C. M. Poteraş, and M. L. Mocanu, "Preliminary results of a curriculum adjuster based on professional network analysis," in 2015 19th International Conference on System Theory, Control and Computing (ICSTCC). IEEE, 2015, pp. 860–865.

[62] D. Fang, K. R. Varshney, J. Wang, K. N. Ramamurthy, A. Mojsilovic, and J. H. Bauer, "Quantifying and recommending expertise when new skills

emerge," in 2013 IEEE 13th International Conference on Data Mining Workshops. IEEE, 2013, pp. 672–679.

[63] R. Gupta and P. K. Reddy, "Towards question improvement on knowledge sharing platforms: A stack overflow case study," in 2017 IEEE International Conference on Big Knowledge (ICBK). IEEE, 2017, pp. 41–48.

[64] J. Jia, Z. Chen, and X. Du, "Understanding soft skills requirements for mobile applications developers," in 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), vol. 1. IEEE, 2017, pp. 108–115.

[65] S. Kabicher, R. Motschnig-Pitrik, and K. Figl, "What competences do employers, staff and students expect from a computer science graduate?" in 2009 39th IEEE Frontiers in Education Conference. IEEE, 2009, pp. 1–6.

[66] G. Matturro, "Soft skills in software engineering: A study of its demand by software companies in uruguay," in 2013 6th international workshop on cooperative and human aspects of software engineering (CHASE). IEEE, 2013, pp. 133–136.

[67] A. Maurya and R. Telang, "Bayesian multi-view models for member-job matching and personalized skill recommendations," in 2017 IEEE International Conference on Big Data (Big Data). IEEE, 2017, pp. 1193–1202.

[68] W. Mo, B. Shen, Y. Chen, and J. Zhu, "Tbil: A tagging-based approach to identity linkage across software communities," in 2015 Asia-Pacific Software Engineering Conference (APSEC). IEEE, 2015, pp. 56–63.

[69] M. Papoutoglou, G. M. Kapitsaki, and N. Mittas, "Linking personality traits and interpersonal skills to gamification awards," in 2018 44th Euromicro Conference on Software Engineering and Advanced Applications (SEAA). IEEE, 2018, pp. 214–221.

[70] M. Papoutsoglou, N. Mittas, and L. Angelis, "Mining people analytics from stackoverflow job advertisements," in 2017 43rd Euromicro Conference on Software Engineering and Advanced Applications (SEAA). IEEE, 2017, pp. 108–115.

[71] F. Pires, J. Barbosa, and P. Leitão, "Data scientist under the da. re perspective: analysis of training offers, skills and challenges," in 2018 IEEE 16th International Conference on Industrial Informatics (INDIN). IEEE, 2018, pp. 523–528.

[72] A. K. Roundtree, "From engineers' tweets: Text mining social media for perspectives on engineering communication," in 2018 IEEE International Professional Communication Conference (ProComm). IEEE, 2018, pp. 6–15.

[73] A. Sambir, V. Yakovyna, and M. Seniv, "Recruiting software architecture using user generated data," in 2017 XIIIth International Conference on Perspective Technologies and Methods in MEMS Design (MEMSTECH). IEEE, 2017, pp. 161–163.

[74] V. Shankararaman and S. Gottipati, "Mapping information systems student skills to industry skills framework," in 2016 IEEE Global Engineering Education Conference (EDUCON). IEEE, 2016, pp. 248–253.

[75] F. Steinmann, K.-I. Voigt, and T. Schaeffler, "Engineering competences: A content analysis of job advertisements," in 2013 Proceedings of PICMET'13: Technology Management in the IT-Driven Services (PICMET). IEEE, 2013, pp. 1925–1929.

[76] B. Vasilescu, A. Capiluppi, and A. Serebrenik, "Gender, representation and online participation: A quantitative study of stackoverflow," in 2012 International Conference on Social Informatics. IEEE, 2012, pp. 332–338.

[77] B. Vasilescu, V. Filkov, and A. Serebrenik, "Stackoverflow and github: Associations between software development and crowdsourced knowledge," in 2013 International Conference on Social Computing. IEEE, 2013, pp. 188–195.

[78] Z. Wang, Y. Wang, and D. Redmiles, "Competence-confidence gap: A threat to female developers' contribution on github," in 2018 IEEE/ACM 40th International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS). IEEE, 2018, pp. 81–90.

[79] A. Ward, A. Gbadebo, and B. Baruah, "Using job advertisements to inform curricula design for the key global technical challenges," in 2015 International Conference on Information Technology Based Higher Education and Training (ITHET). IEEE, 2015, pp. 1–6.

[80] C. Zhou, S. K. Kuttal, and I. Ahmed, "What makes a good developer? an empirical study of developers' technical and social competencies," in 2018 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC). IEEE, 2018, pp. 319–321.

[81] M. Daneva, C. Wang, and P. Hoener, "What the job market wants from requirements engineers?: an empirical analysis of online job ads from the netherlands," in Proceedings of the 11th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement. IEEE Press, 2017, pp. 448–453.

[82] C. Hauff and G. Gousios, "Matching github developer profiles to job advertisements," in Proceedings of the 12th Working Conference on Mining Software Repositories. IEEE Press, 2015, pp. 362–366.

[83] E. Malherbe and M.-A. Aufaure, "Bridge the terminology gap between recruiters and candidates: A multilingual skills base built from social media and linked data," in Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. IEEE Press, 2016, pp. 583–590.

[84] H. Hu, Y. Luo, Y. Wen, Y.-S. Ong, and X. Zhang, "How to find a perfect data scientist: A distance-metric learning approach," IEEE Access, vol. 6, pp. 60 380–60 395, 2018.

[85] S. A. Chala, F. Ansari, M. Fathi, and K. Tijdens, "Semantic matching of job seeker to vacancy: a bidirectional approach," International Journal of Manpower, vol. 39, no. 8, pp. 1047–1063, 2018.

[86] K. Tijdens, M. Beblavỳ, and A. Thum-Thysen, "Skill mismatch comparing educational requirements vs attainments by occupation," International Journal of Manpower, vol. 39, no. 8, pp. 996–1009, 2018.

[87] J. M. Álvarez-Rodríguez, R. Colomo-Palacios, and V. Stantchev, "Skillrank: Towards a hybrid method to assess quality and confidence of professional skills in social networks," Scientific Programming, vol. 2015, p. 3, 2015.

[88] L. A. Leon, K. C. Seal, Z. H. Przasnyski, and I. Wiedenman, "Skills and competencies required for jobs in business analytics: A content analysis of job advertisements using text mining," in Operations and Service Management: Concepts, Methodologies, Tools, and Applications. IGI Global, 2018, pp. 880–904.

[89] P. Tambe, "Big data investment, skills, and firm value," Management Science, vol. 60, no. 6, pp. 1452–1469, 2014.

[90] G. Domeniconi, G. Moro, A. Pagliarani, K. Pasini, and R. Pasolini, "Job recommendation from semantic similarity of linkedin users' skills." in ICPRAM, 2016, pp. 270–277.

[91] T. P. Sahu, N. K. Nagwani, and S. Verma, "An empirical analysis on reducing open source software development tasks using so," Indian J. Sci. Technol, vol. 9, 2016.

[92] J. Salminen, M. Milenković, B. J. Jansen, and D. Dubai, "Problems of data science in organizations: an explorative qualitative analysis of business professionalsŠ concerns," in Proceedings of International Conference on Electronic Business (ICEB 2017). Dubai, 2017.

[93] M. Zihayat, A. An, L. Golab, M. Kargar, and J. Szlichta, "Authority-based team discovery in social networks," arXiv preprint arXiv:1611.02992, 2016.

[94] B. Vasilescu, A. Capiluppi, and A. Serebrenik, "Gender, representation and online participation: A quantitative study," Interacting with Computers, vol. 26, no. 5, pp. 488–511, 2013.

[95] N. G. Brooks, T. H. Greer, and S. A. Morris, "Information systems security job advertisement analysis: Skills review and implications for information systems curriculum," Journal of Education for Business, vol. 93, no. 5, pp. 213–221, 2018.

[96] A. Gardiner, C. Aasheim, P. Rutner, and S. Williams, "Skill requirements in big data: A content analysis of job advertisements," Journal of Computer Information Systems, vol. 58, no. 4, pp. 374–384, 2018.

[97] A. Verma, K. M. Yurov, P. L. Lane, and Y. V. Yurova, "An investigation of skill requirements for business and data analytics positions: A content analysis of job advertisements," Journal of Education for Business, vol. 94, no. 4, pp. 243–250, 2019.

[98] P. G. Lovaglio, M. Cesarini, F. Mercorio, and M. Mezzanzanica, "Skills in demand for ict and statistical occupations: Evidence from web-based job vacancies," Statistical Analysis and Data Mining: The ASA Data Science Journal, vol. 11, no. 2, pp. 78–91, 2018.

[99] D. Bañeres Besora and J. Conesa Caralt, "A life-long learning recommender system to promote employability," 2017.

[100] U. Sathyapriya, "Revealing hidden profile information and ranking job seekers on big data," Indian Journal of Science and Technology, vol. 9, no. 19, 2016.

[101] R. Boselli, M. Cesarini, S. Marrara, F. Mercorio, M. Mezzanzanica, G. Pasi, and M. Viviani, "Wolmis: a labor market intelligence system for classifying web job vacancies," Journal of Intelligent Information Systems, vol. 51, no. 3, pp. 477–502, 2018.

[102] F. Farhadi, M. Sorkhi, S. Hashemi, and A. Hamzeh, "An effective framework for fast expert mining in collaboration networks: a group-oriented

and cost-based method," Journal of Computer Science and Technology, vol. 27, no. 3, pp. 577–590, 2012.

[103] A. May, J. Wachs, and A. Hannák, "Gender differences in participation and reward on stack overflow," Empirical Software Engineering, pp. 1–23, 2019.

[104] I. Mewburn, W. J. Grant, H. Suominen, and S. Kizimchuk, "A machine learning analysis of the non-academic employment opportunities for ph. d. graduates in australia," Higher Education Policy, pp. 1–15, 2018.

[105] H. Pérez-Rosés and F. Sebé, "Iterated endorsement deduction and ranking," Computing, vol. 99, no. 5, pp. 431–446, 2017.

[106] Y. Wan, L. Chen, G. Xu, Z. Zhao, J. Tang, and J. Wu, "Scsminer: mining social coding sites for software developer recommendation with relevance propagation," World Wide Web, vol. 21, no. 6, pp. 1523–1543, 2018.

[107] R. Boselli, M. Cesarini, F. Mercorio, and M. Mezzanzanica, "Using machine learning for labour market intelligence," in Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, 2017, pp. 330–342.

[108] B. Coelho, F. Costa, and G. M. Gonçalves, "Hyre-me–hybrid architecture for recommendation and matchmaking in employment," in International Conference on Information and Software Technologies. Springer, 2015, pp. 208–224.

[109] R. Florea and V. Stray, "Software tester, we want to hire you! an analysis of the demand for soft skills," in International Conference on Agile Software Development. Springer, 2018, pp. 54–67.

[110] A. Menshikova, "Evaluation of expertise in a virtual community of practice: The case of stack overflow," in International Conference on Digital Transformation and Global Society. Springer, 2018, pp. 483–491.

[111] W. Poonnawat, E. Pacharawongsakda, and N. Henchareonlert, "Jobs analysis for business intelligence skills requirements in the asean region: A text mining study," in The Joint International Symposium on Artificial Intelligence and Natural Language Processing. Springer, 2017, pp. 187–195.

[112] A. K. Nassirtoussi, S. Aghabozorgi, T. Y. Wah, and D. C. L. Ngo, "Text mining for market prediction: A systematic review," Expert Systems with Applications, vol. 41, no. 16, pp. 7653–7670, 2014.

[113] A. Abbe, C. Grouin, P. Zweigenbaum, and B. Falissard, "Text mining applications in psychiatry: a systematic literature review," International journal of methods in psychiatric research, vol. 25, no. 2, pp. 86–100, 2016.

[114] A. Ampatzoglou, S. Bibi, P. Avgeriou, M. Ver-beek, and A. Chatzigeorgiou, "Identifying, categorizing and mitigating threats to validity in software engineering secondary studies," Information and Software Technology, 2018.

MARIA PAPOUTSOGLOU holds a BSc in Information Technology from the University of Macedonia and an MSc in Information Systems from the Aristotle University of Thessaloniki. Currently is a Phd candidate and an associate researcher and developer in STAINS group of the Aristotle University. In the past, she has worked in the Europass project infrastructure, at the European agency CEDEFOP, and for SEN2SOC experiment under SmartSantander FP7. Also, she has worked as a technical and data analyst in the private sector.

APOSTOLOS AMPATZOGLOU is an elected Assistant Professor in the Department of Computer Science of the University of Macedonia, where he carries out research and teaching in the area of software engineering. In the period 2013-2016 he was an Assistant Professor at the Department of Computer Science in the University of Groningen (Netherlands). He holds a BSc on Information Systems (2003), a MSc on Computer Systems (2005) and a PhD in Software Engineering by the Aristotle University of Thessaloniki (2012). His current research interests are focused on technical debt, reverse engineering, software maintainability, software quality management, open source software engineering and software design.

He has published more than 70 articles in international journals and conferences. He is/was involved in 15 research and development projects in Information Communication Technologies with funding from national and international organizations. He serves as a reviewer in numerous leading journals of the software engineering domain, as part of the organizing committee of five prestigious conferences, and as a member of various international conference program committees. Finally, he has been acknowledged as being among the top scholars in the field of design patterns, and among the top early stage researchers in the field of software engineering for the period 2010-2017.

NIKOLAOS MITTAS is an elected Assistant Professor in the Department of Chemistry of the International Hellenic University. He received a BSc degree in Mathematics from University of Crete. He also received a MSc degree in Informatics from Aristotle University of Thessaloniki (AUTh) with a specialty in Information Systems. His doctoral dissertation has the title "Statistical and Computational Methods for Development, Improvement and Comparison of Software Cost Estimation Models" covering the wider area of Software Engineering and Statistics. He was an Assistant Professor at the Department of Petroleum, Natural Gas and Mechanical Engineering, School of Technological Engineering, Eastern Macedonia and Thrace Institute of Technology. He was an adjunct faculty member at the Technological Educational Institute of Kavala, Computer Science Department of Aristotle University of Thessaloniki, Hellenic Open University and Open University of Cyprus. He has participated in several international and national conferences.

LEFTERIS ANGELIS studied Mathematics and received his Ph.D. degree in Statistics (Experimental Designs) from Aristotle University of Thessaloniki (A.U.Th.). He is currently Professor at the School of Informatics, A.U.Th and coordinator of the STAINS research group. He has served as Deputy Head and Head of the School of Informatics. He is also member of the Board of Directors and Treasurer of the Greek Statistical Institute. His research interests involve: statistical methods, especially with applications to information systems and software engineering (SE) focusing on the research on the human factor; computational methods in mathematics and statistics; planning of experiments; biostatistics - bioinformatics and simulation techniques. He has published more than 200 papers in journals, book chapters and conference proceedings.