



Navigating technical, legal, and ethical hurdles to scraping LinkedIn data for academic research

Navegando pelas barreiras técnicas, legais e éticas na extração de dados do LinkedIn para pesquisas acadêmicas

André José de Queiroz Padilha^a 

Jesús Pascual Mena Chalco^{a,*} 

ABSTRACT: In an era where professional career data is critical for analyzing occupational trends and organizational dynamics, LinkedIn data offers a rich corpus for academic research due to its expansive user base and frequent updates. This paper examines technical, legal, and ethical challenges associated with scraping LinkedIn profiles for research, arguing that scraping is the most effective method for acquiring comprehensive LinkedIn data compared to direct cooperation, purchasing data, or APIs. Despite prohibitive measures and potential legal issues outlined by LinkedIn, recent court decisions provide favorable precedents for the lawful scraping of public profiles. The paper also compiles prior research studies that leveraged LinkedIn data, highlighting various acquisition methods and their applicability to academic research. It explores strategies to ethically and legally navigate scraping, providing recommendations on how researchers can responsibly collect LinkedIn data, ensuring compliance with evolving privacy laws and ethical standards. Finally, technical considerations are discussed, emphasizing the use of tools like Selenium to overcome LinkedIn's sophisticated anti-scraping measures.

Keywords: LinkedIn Data Scraping; Data Acquisition; Legal and Ethical Challenges; Public Data Research; Scraping.


RESUMO: Na era em que dados de carreiras profissionais são críticos para a análise de tendências ocupacionais e dinâmicas organizacionais, o LinkedIn oferece um rico corpus para pesquisas acadêmicas devido à sua ampla base de usuários e atualizações frequentes. Este artigo examina os desafios técnicos, legais e éticos associados ao scraping de perfis do LinkedIn para fins de pesquisa, argumentando que o scraping é o método mais eficaz para adquirir dados abrangentes do LinkedIn em comparação com cooperação direta, compra de dados ou uso de APIs. Apesar das medidas proibitivas e possíveis questões legais estabelecidas pelo LinkedIn, decisões judiciais recentes oferecem precedentes favoráveis para a coleta lícita de perfis públicos. O artigo também compila estudos anteriores que utilizaram dados do LinkedIn, destacando vários métodos de aquisição e sua aplicabilidade à pesquisa acadêmica. Ele explora estratégias para navegar de forma ética e legal o scraping de dados, fornecendo recomendações sobre como os pesquisadores podem coletar dados do LinkedIn de maneira responsável, garantindo conformidade com leis de privacidade em evolução e padrões éticos. Finalmente, são discutidas considerações técnicas, enfatizando o uso de ferramentas como o Selenium para superar as medidas sofisticadas de proteção contra scraping do LinkedIn.

Palavras-chave: Raspagem de Dados do LinkedIn; Aquisição de Dados; Desafios Legais e Éticos; Pesquisa de Dados Públicos; Raspagem

^a Programa de pós-graduação em Ciência da Computação, Universidade Federal do ABC, Santo André, SP, Brasil.

* Correspondência para/Correspondence to Jesús Pascual Mena Chalco. E-mail: Jesus.mena@ufabc.edu.br.

Recebido em/Received: 05/05/2024; Aprovado em/Approved: 29/07/2024.

Artigo publicado em acesso aberto sob licença [CC BY 4.0 Internacional](https://creativecommons.org/licenses/by/4.0/)  

INTRODUCTION

The professional landscape is constantly evolving, leaving behind a wealth of data trails generated by individuals. This data, encompassing elements like geographical work locations, organizational affiliations, job tenure, positions held, and anonymized demographics such as gender, forms a valuable corpus for academic inquiry. By analyzing this multifaceted information, researchers can gain insights into individual career trajectories, identify macro-level trends within specific professions, and understand organizational dynamics. This data serves pivotal roles in various research areas. For instance, universities can leverage it to assess the professional outcomes of their graduates, allowing them to measure and enhance the impact of their educational programs.

Directly acquiring career data from students and alumni presents a significant challenge. While a one-time survey can provide a snapshot, it's crucial to gather data frequently to capture career trends and analyze time-dependent phenomena (PUCPR 2022; Pereira, Simon, Pacheco 2021; Coll, Liana 2021). One exception in Brazil is the platform *Alumni USP*, from *Universidade de São Paulo*, with 37% of alumni registered (USP 2024). Nevertheless, there are no details about how frequently alumni update their information on that platform.

There are also research initiatives that used online forms and achieved a higher adhesion rate—30% in Jones et al. (2017) and Bista et al. (2021). Still, it was a small alumni population (570 and 2,155, respectively). Furthermore, the whole acquisition process—preparing the questions in the form, creating the form, getting a list of emails, and emailing every individual more than once—would need to be repeated every time to assess the professional outcome of students and alumni outside academia.

A few research groups have utilized LinkedIn data to avoid the pitfalls of online forms and proprietary platforms. LinkedIn is the largest professional social network in the world, with over 1 billion users in more than 200 countries (LinkedIn 2024a) and over 59 million Brazilian users (Lisboa 2023). LinkedIn users update their information constantly to remain relevant on the platform, a feature necessary for any research that constantly assesses something about a population, like universities wanting to assess their alumni.

LinkedIn's privacy settings allow users to control the information visible on their profiles. Non-connected logged-in users have limited access, while public profile settings determine what logged-out users and search engines see. Users have granular control over the information displayed in their public profiles, including profile photo, headline, work experience, education, and certifications. However, the Skills section functions differently and remains hidden from public view regardless of privacy settings.

In this work, we explore how to acquire public LinkedIn profiles for research purposes, focusing on scraping techniques. Web scraping is the process of automatically

gathering online data using computer software without relying on an official interface that the website could provide (Mitchell 2018). We will argue why scraping is the best option for acquiring LinkedIn data for research purposes and compare it to other methods.

In 2022, Luscombe, Dick, and Walby explored the technical, legal, and ethical challenges of web scraping any website in social sciences. Although web scraping has a significant potential to acquire data, these challenges are often overseen and not discussed (Luscombe, Dick, Walby 2022), even when the research author proactively circumvented defensive mechanisms against scraping. Building on their work and in light of recent court decisions, we will outline strategies to address these challenges when scraping LinkedIn.

Ultimately, this work aims to compile and summarize the most critical studies on LinkedIn data utilization and acquisition, particularly scraping.

METHODOLOGY

This study was conducted through an ad hoc bibliographic research between March 1, 2023, and April 10, 2024, allowing for the collection of scientific articles related to the use and implications of LinkedIn data. The keywords utilized in the bibliographic research were: "LinkedIn scraping", "LinkedIn scrape", "alumni monitoring LinkedIn", "LinkedIn data", "scraping overview", "web scraping techniques", "web scraping avoid blocking", and "web scraping ethics". This approach ensured a comprehensive review of the existing literature, encompassing diverse perspectives and insights into the acquisition and utilization of LinkedIn data.

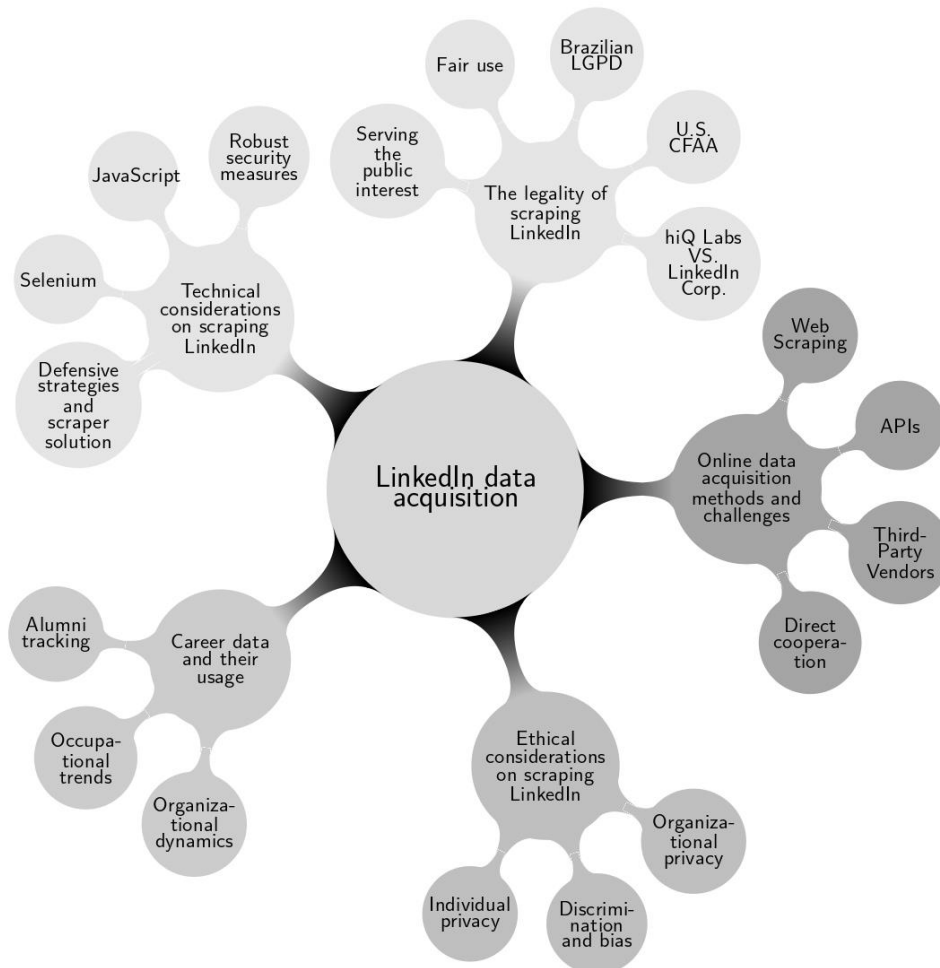
Drawing upon the foundational work of Luscombe, Dick, and Walby (2022), which explored the technical, legal, and ethical challenges of web scraping in the social sciences, this study extends the discourse by focusing specifically on the acquisition of LinkedIn data. By building on their insights and considering recent court decisions, strategies to address the challenges inherent in scraping LinkedIn are outlined. These strategies encompass both technical approaches to overcome defensive mechanisms and ethical considerations to ensure responsible data acquisition practices.

Moreover, this study critically evaluates the suitability of web scraping as a method for acquiring LinkedIn data for research purposes. By comparing scraping techniques with alternative methods and assessing their respective advantages and limitations, a comprehensive understanding of the efficacy and ethical implications of scraping LinkedIn is achieved. This methodological approach enables a detailed examination of the complexities surrounding the acquisition and utilization of LinkedIn data within the academic research context.

RESULTS AND DISCUSSION

In the subsequent sections, we will examine various aspects pertinent to the acquisition of data from LinkedIn. Refer to Figure 1 for a conceptual mind map illustrating the primary topics addressed in this discussion.

Figure 1. Mind Map illustrating key aspects explored in this study.



Carrer data and their usage

Using LinkedIn data allows different research questions to be pursued and answered, not only assessing university alumni. This section highlights various studies employing LinkedIn data, detailing their objectives and acquisition methods, including whether they relied on third-party providers or scraped data from logged-in sessions.

Goncalves et al. (2014) implemented a web scraper to extract LinkedIn data pertaining to alumni from various universities in Brazil. They did not utilize a third-party provider; however, it remains unclear whether the data acquisition occurred during logged-in sessions. In contrast, Yutao Zhang et al. (2015) used LinkedIn data, alongside

information from other social networks, to create a cross-platform profile of individuals. This data was acquired from a third-party provider.

Almeida (2018) developed a tool to extract LinkedIn data, utilizing it to analyze alumni from PUC-Rio. This tool scraped data from logged-in sessions. Similarly, Agostinho (2021) created a tool to extract LinkedIn data, which also required logged-in sessions.

Additionally, Wang et al. (2021) utilized LinkedIn data to model career trajectories and applied deep learning to predict career paths, although they did not specify the source of their data, while Yamashita et al. (2022) used career data from a third-party provider with similar purposes. Finally, Abel et al. (2023) used LinkedIn data to analyze overseas alumni populations from 106 Chinese universities, while Chaparala et al. (2023) used LinkedIn data to create an Alumni platform for their institution, but that data was acquired from a third-party provider. Contreras (2023) extracted LinkedIn data using a third-party provider to expand on previous works in Open Source Intelligence.

Recent studies used data from LinkedIn via pre-built datasets from *Revelio Labs*, a company specializing in selling workforce intelligence data. Agarwal et al. (2023) studied the relationship between the early career experience of bank regulators during a crisis and their later careers. Eisfeldt et al. (2023) used their data to investigate the effects of recent Generative AI developments and the value of firms. Liang et al. (2023) used it to study companies' voluntary disclosure of workforce gender diversity. Ahn et al. (2023) used their data to research career advancement amongst female and racial minority auditors. Lin et al. (2023) used it to study migration patterns among women auditors who live in states that restrict access to abortion.

Te et al. (2023) used LinkedIn data to analyze the success of Series A venture capital funding on startups, but it does not specify how it acquired the data. Other works utilize data outside of LinkedIn. Lungu et al. (2012) and Lungu et al. (2014) used data from a national survey to model the occupational mobility of Romanian graduates. Zamfir et al. (2013) also used data from a national survey to study the relationship between education-job mismatch¹ and wages. Del Rio-Chanona et al. (2021) used data from a national survey to study the impact of automation on the labor market.

Table 1 lists all the aforementioned works, detailing their data sources, acquisition dates, acquisition methods, and dataset sizes. As shown, all studies that utilized LinkedIn data relied either on a third-party provider or on scraping data from logged-in sessions.

¹ For example over-education for a given job position.

Table 1. Methods of data acquisition employed across different studies.

| Work | Data source | Acquisition date | Acquisition method | Size |
|--|--------------------------|---|---|----------------------|
| (Lungu, Zamfir, Militaru, Mocanu 2012) | A national survey | 2008 - 2009 | Manual | 2,194 alumni |
| (Zamfir, Matei, Lungu 2013) | A national survey | 2005 | Manual | 70,000 alumni |
| (Lungu, Zamfir, Mocanu, Pîrciog 2014) | A national survey | 2008 - 2009 | Manual | 2,194 alumni |
| (Goncalves, Ferreira, Tavares De Assis, Tavares 2014) | LinkedIn | Recurrent | Automatic scraping, doesn't inform if they were logged in | 6,092 alumni |
| (Zhang, Tang, Yang, Pei, Yu 2015) | LinkedIn | Not informed, probably around 2014 - 2015 | Automatic, doesn't inform if they used 3rd party provider or if they were logged in | 2,985,414 profiles |
| (Almeida 2018) | LinkedIn | Recurrent | Automatic scraping, logged in | 57,901 alumni |
| (Zhang, Zhu, Xu, Zhu, Qin, Xiong, Chen 2019) | Not informed | Not informed | Not informed | 2,176,157 resumes |
| (Agostinho 2021) | LinkedIn | Recurrent | Automatic scraping, logged in | 149 alumni |
| (Wang, Zhu, Hao, Xiao, Xiong 2021) | LinkedIn | Not informed | Not informed | 414,266 profiles |
| (Del Rio-Chanona, Mealy, Beguerisse-Díaz, Lafond, Farmer 2021) | A national survey | 2020 | Manual | Not informed |
| (Yamashita, Li, Tran, Zhang, Lee 2022) | Career platform Future | Not informed | Pre-built | +300,000 |
| (Chaparala, Jukuntla, Reddy, Vinayak, Sudha 2023) | LinkedIn | Recurrent | Scraping using Phantom Buster | Not informed |
| (Abel, Zhu, Huang 2023) | LinkedIn aggregated data | 2021 | Manual from LinkedIn advertising platform | 106 universities |
| (Agarwal, Lin, Shen, Wu 2023) | LinkedIn | Not informed | Pre-built from Revelio Labs | 16,570 profiles |
| (Eisfeldt, Schubert, Zhang 2023) | LinkedIn | Not informed, probably 2022 | Pre-built from Revelio Labs | Not informed |
| (Liang, Lourie, Nekrasov, Yoo 2023) | LinkedIn | Not informed | Pre-built from Revelio Labs | 322,410 companyyears |

| | | | | |
|--|----------|-----------------------------|-----------------------------|-----------------|
| (Ahn, Hoitash, Hoitash, Krause 2023) | LinkedIn | Not informed, probably 2023 | Pre-built from Revelio Labs | 449,655 users |
| (Lin, Shen, Shi, Zeng 2023) | LinkedIn | Not informed | Pre-built from Revelio Labs | 76,916 profiles |
| (Te, Wieland, Frey, Pyatigorskaya, Schiffer, Grabner 2023) | LinkedIn | Not informed | Not informed | 3,204 profiles |

We included studies that did not utilize LinkedIn data to illustrate that inquiries regarding jobs and careers can be addressed without its use. Nevertheless, these studies frequently depended on national surveys, which pose numerous challenges such as low participation rates, restricted datasets, and high costs, thereby making frequent replication arduous. In contrast, LinkedIn data presents several advantages: (i) It is readily available, though it must be procured; (ii) Users frequently update it, ensuring its relevance and accuracy; and (iii) It can be cost-effective to obtain. Given these benefits, the question arises: how can researchers effectively acquire LinkedIn data?

Online data acquisition methods and challenges

The acquisition of online data, such as information from LinkedIn, poses distinct challenges and opportunities that can significantly influence the scope and quality of research. To navigate this landscape effectively, researchers employ various data acquisition methods, each with unique advantages and limitations. These methods range from direct cooperation with data owners² to more technical approaches like web scraping and Application Programming Interfaces (APIs)—tools that allow automated access to a platform's data under specific conditions—as well as purchasing data or building dedicated user panels to track user behavior. Understanding these methods provides valuable insights into their practical applications and potential drawbacks.

The following sections detail five strategies for data acquisition identified by Possler et al. (2019) and Metaxas et al. (2014), exploring their implications for accessibility, data quality, legal and ethical considerations, and the overall impact on the research process. Legal and ethical considerations associated with these methods are significant and will be further explored in a dedicated section, given their impact on the research process and compliance requirements.

² Although data ownership is a controversial topic in legal doctrine (Lohsse, Schulze, Staudenmayer 2017). LinkedIn itself, in its User Agreement, disclaims that users own all the information they post online, only giving the company a "non-exclusive license to it". We will use *data owner* to refer to the organizations that control who can access publicly available data.

Direct cooperation with data owners such as social media platforms or digital service providers is highly effective for acquiring large, detailed datasets. Possler et al. (2019) highlight that this method allows researchers to access both publicly viewable data and proprietary data that might not be available through other means. Legal coverage is often clear due to predefined terms and conditions of the formal partnerships, but often these partnerships are highly dependent on personal connections within the industry and being geographically close to the company. While this method can provide rich, high-quality data, it often comes with strict stipulations regarding the use of data, including limitations on the scope of research questions and potential publication restrictions. The dependence on the goodwill of these corporations can pose risks, as access can be abruptly revoked or limited, with the data owners having the final say on the permissible use of their data, potentially leading to conflicts of interest or censorship issues.

Purchasing data from data owners or third-party resellers is a straightforward method but often involves substantial financial costs (Possler, Bruns, Niemann-Lenz 2019) that can limit accessibility for underfunded researchers (Bruns 2013). This method is beneficial as it enables researchers to rapidly acquire substantial data volumes without requiring advanced data collection skills. However, it often necessitates purchasing larger datasets due to minimum spending thresholds and the data purchased may not always fit the specific needs of the research, leading to possible gaps in the data, which could affect the comprehensiveness of the research findings. Furthermore, there is often little transparency regarding how the data was collected and processed before sale, which can affect the reliability and validity of the research.

APIs represent a standardized approach for systematically accessing data from platforms that offer them, allowing researchers to efficiently acquire data in a structured format, without requiring specialized technical skills (Possler, Bruns, Niemann-Lenz 2019). Similar to direct cooperation with data owners, the use of APIs is subject to the terms and limitations imposed by the data providers. These constraints can restrict the amount of available data, offer only specific portions of it, and impose caps on the number of requests. Consequently, this can severely curtail the scope of research projects, especially those necessitating extensive data analysis over prolonged periods. Additionally, changes in API access policies can disrupt ongoing research projects, as noted by Metaxas (2014), who notes that platforms often revise their API terms in response to commercial priorities or privacy considerations, abruptly limiting access to previously available data streams.

Web scraping and crawling encompass programmatically accessing and extracting data from websites. This method offers high flexibility and can be tailored to gather a diverse array of data types. Possler et al. (2019) emphasize that while scraping offers access to data that might not be available through APIs, it demands significant technical expertise. Additionally, data owners continually endeavor to thwart scraping efforts. Moreover, scraping is susceptible to errors due to the necessity for custom scripts tailored to the HTML structure of each website. Importantly, scraping data from websites without permission may contravene terms of service or copyright laws,

potentially exposing researchers to legal repercussions. Furthermore, the data acquired through scraping may lack structure, often necessitating extensive cleaning and preprocessing to render it suitable for research purposes.

Constructing user panels to track their behavior entails recruiting participants to directly provide data, often through software installed on their devices. This approach enables researchers to gather highly detailed data on user behavior, encompassing information generated by users during routine interactions with digital platforms. According to Possler et al. (2019), this method can yield profound insights into individual and group behaviors. However, it is resource-intensive, necessitating intricate software development and meticulous user consent procedures. Recruiting participants poses a challenge, resulting in small, self-selecting samples that may not be representative of the broader population. Ethical considerations also arise as tracking could potentially alter user behavior due to their awareness of being monitored.

Now, we assess the effectiveness of each approach in collecting data from LinkedIn.

Direct cooperation with LinkedIn seemed highly improbable, as no studies listed in Table 1 reported establishing an official partnership with LinkedIn for data access. Furthermore, LinkedIn does not offer an option to directly purchase data.

Acquiring data from third-party vendors such as Bright Data or Revelio Labs often proves costly, lacks reliability, and presents challenges in method replication. Our investigation of a data sample from Bright Data unveiled its failure to specify the dates of data collection for each profile, thereby undermining the systematic organization of profiles or employment histories for analysis. Additionally, this data was not anonymized.

Bright Data also promotes *The Bright Initiative*, described as "a global program providing selected non-profits, academic institutions, and public bodies with pro-bono access to *Bright Data's* leading web data technology, expertise, and support to drive positive global change" (Bright Data 2024a). We attempted to access this data by initiating contact via email on July 25th, 2023, and providing the requested details. Two days later, they responded, informing us that due to the volume of inquiries, our request was placed on a waiting list. Subsequent communications were limited to promotional emails encouraging us to purchase their dataset.

Using LinkedIn's official Profiles API requires being part of the "Compliance API Partner Program" (LinkedIn 2023; 2024b), approval for which is at the discretion of LinkedIn. Furthermore, it mandates registration with either FINRA or the SEC (Financial Industry Regulatory Authority and Securities and Exchange Commission, respectively), a requirement that is nearly unattainable for a research group, particularly one outside the U.S. Additionally, LinkedIn does not publicly disclose the API's pricing, and no studies in Table 1 reported utilizing this official API.

Constructing a user panel would entail developing a tracking mechanism within the browsers of participating users, akin to the challenge encountered with proprietary alumni platforms: persuading users to engage, often resulting in low participation rates. Moreover, this method is more suitable for researchers interested in studying user behavior within a specific website, diverging from the goal of acquiring public data.

The last method under consideration was web scraping, which is technically intricate owing to LinkedIn's ongoing efforts to thwart such activities, alongside legal and ethical considerations that must be addressed. Despite these drawbacks, it remains the sole option for accessing public profiles, thereby placing control of the data acquisition process in the hands of researchers (Possler, Bruns, Niemann-Lenz 2019), including those from underfunded institutions.

In Table 2, we present a comparative summary of five methods for acquiring online public data, drawing from the studies of Metaxas et al. (2014), Possler et al. (2019) and Luscombe et al. (2022). This table provides a qualitative assessment of the costs, advantages, disadvantages, and applicability of each method, focusing specifically on collecting data from LinkedIn.

Table 2. Description of data acquisition methods.

| Method | Legal? | Cost | Advantages | Disadvantages | Can be used on LinkedIn? |
|--------------------------------|--------------|----------------------------|--|--|--|
| Cooperation with data owner | Yes | Free | Easier data collection, high-quality formatted data, no cost | Often comes with a lot of bureaucracy, restricted access and only gives a partial view of the data | No work mentioned a cooperation with LinkedIn |
| Buying from data owner | Yes | Up to thousands of dollars | Easier data collection, high-quality formatted data | Hard to reproduce, cost, little transparency regarding how data was processed | No |
| Buying from 3rd party provider | Inconclusive | Up to thousands of dollars | Easier data collection, high-quality formatted data | Hard to reproduce, cost, little transparency regarding how data was acquired and processed | Yes, for example using Proxycurl, Bright Data or Revelio Labs datasets |
| Using official API | Yes | Unknown | High-quality formatted data | Needs some technical knowledge, APIs can change or be | No work mentioned using LinkedIn's API |

| | | | | | |
|--------------|---|---------------------|--|--|-----|
| | | | | discontinued, little transparency regarding how data was processed | |
| Web scraping | Gray area or illegal, depending on how it is executed | Potentially free | More control over what is accessed, lower cost, can acquire more data, reproducibility | Needs a lot of technical knowledge, website structure can change, prone to errors, can be illegal | Yes |

Based on the analysis of the different data acquisition methods available for gathering data from LinkedIn, it becomes apparent that web scraping emerges as the most advantageous approach. The primary rationale for selecting web scraping is its unmatched capability to grant researchers control over data collection, allowing them to tailor their data collection strategies to precisely meet their research requirements. Unlike other methods, web scraping does not hinge on third-party terms, access limitations, or cost barriers. This is especially advantageous for underfunded research groups requiring access to comprehensive data without incurring exorbitant expenses.

Web scraping, when conducted while logged out, enables the acquisition of data solely from publicly available LinkedIn profiles, accessing only the information that users have chosen to share. LinkedIn imposes restrictions on accessing certain profiles and limits the data visible when viewing profiles while logged out of the platform. A comparative examination of profiles accessible both when logged in and logged out is outlined in Table 3. This comparative analysis aids in ascertaining whether third-party providers are improperly accessing or selling data that should be restricted, particularly data exclusively available to logged-in users.

Table 2. Information availability based on user login status.

| Field | Logged in | Logged out |
|---|---|--|
| Profile picture; Cover picture; Full name; Headline; City/State/Country; Number of followers; About | Yes | Yes |
| Contact info (email, birthday, account creation) | Yes | No |
| Number of connections | Yes (and you can see who are the connections) | Yes (but you cannot see who are the connections) |
| Highlights (things that you and the person you are looking at have in common) | Yes | No |
| Activity (posts and comments) | Yes | Very limited |

| | | |
|---|---------------------|----------------|
| Experience - Role (title); Company name; Start date; End date; Location; Description | Yes | Yes |
| Experience - Location type (remote, on-site or hybrid) | Yes | No |
| Experience - Employment type (Full-time, part-time, self-employed, internship, etc) | Yes | No |
| Experience – Skills; Uploaded media | Yes | No |
| Education – School name; Degree; Field of study; Description; Grade (GPA); Activities and societies | Yes | Yes |
| Education – Start date and end date | Yes, month and year | Yes, year only |
| Education – Skills; Uploaded media | Yes | No |
| Volunteering – Organization name, Duration, Role, Cause | Yes | Yes |
| Licenses and certifications – Certificate name, Certification agency, Issue date, Certificate ID, Certificate URL | Yes | Yes |
| Skills and endorsements | Yes | No |
| Projects – Name; URL; Duration; Description | Yes | Yes |
| Projects – Associated with (people that also worked in the same project) | Yes | No |
| Recommendations | Yes | No |
| Publications | Yes | Yes |
| Courses - Course name and location | Yes | Yes |
| Honors & Awards - Honor name; Issued by; Issue date; Description | Yes | Yes |
| Languages and level of experience | Yes | Yes |
| Interests | Yes | No |
| Organizations (that the person follow) | Yes | Yes |
| People also viewed (profiles that other people also visited after the current profile) | Yes | Yes |
| Causes (that the person support) | Yes | No |

Despite its advantages, web scraping is a method that demands continuous technical adjustments because of LinkedIn's endeavors to thwart scraping activities and its daily user interface updates. Furthermore, the scraping techniques utilized must comply with legal and ethical standards, as elucidated by Luscombe et al. (2022). This will be the focus of our subsequent discussion.

The legality of scraping LinkedIn

A key challenge for research projects aiming to utilize LinkedIn public profiles is determining the legality of doing so. Most studies that have used LinkedIn data either depend on a third-party provider to ensure compliance or lack thorough consideration of potential legal ramifications³.

LinkedIn's User Agreement explicitly prohibits web scraping, even for unregistered visitors (LinkedIn 2022). Violating this agreement could result in contractual liability. However, Kerr (2015), in an essay for U.S. courts about web scraping, notes that if scraping is conducted while logged out, it should not result in criminal charges. The most relevant case on this issue is *hiQ Labs, Inc. v. LinkedIn Corp.*, a U.S. case that specifically addressed scraping LinkedIn. This case is crucial for understanding the legal implications of algorithmic decisions and data usage.

hiQ Labs, a California-based for-profit company focused on human capital data (Crunchbase 2024), offered two primary products: "Keeper," which analyzed employee retention trends, and "Skill Mapper," which assessed workforce skills. The latter directly competed with LinkedIn's "Talent Insights" (U.S. District Court, N.D. California 2022a).

hiQ Labs primarily used LinkedIn data obtained through methods such as simulating user input, employing mechanical turkers⁴ to create fake accounts, and scraping data while logged in (Neuburger 2022a). Despite knowing about hiQ's practices since 2014, LinkedIn only began to take action in 2017, sending a cease-and-desist letter and blocking hiQ's IP addresses. LinkedIn accused hiQ of violating the federal Computer Fraud and Abuse Act (CFAA), the Digital Millennium Copyright Act (DMCA), California Penal Code, and California's common law of trespass. The CFAA, dating back to 1986, aimed to prevent hacking of computer systems, but has often been used successfully against web scrapers (Reese, Alex, Quesenberry, Raven 2022).

In response, hiQ filed a preliminary injunction against LinkedIn, demanding the removal of IP blocks. The District Court of Northern California ruled in favor of hiQ, mandating LinkedIn to cease blocking measures. The court argued that public profiles were expected to be searchable and analyzed, and giving private entities like LinkedIn

³ It is important to notice that some authors would not be able to analyze court ruling decisions because those decisions were not made at the time of the author's published papers.

⁴ Turkers are individual contractors that perform all sorts of simple tasks in the computer. In this case, they used turkers to manually extract the URLs of hiQ customers' employees. hiQ explicitly instructed, "It is a good idea to make a fake account with a fake email, to deal with the possibility of being banned on LinkedIn" (CHEN, EDWARD M., 2022).

blanket authority to block access to public information could jeopardize public discourse:

It is likely that those who opt for the public view setting expect their public profile will be subject to searches, data mining, aggregation, and analysis. On the other hand, conferring on private entities such as LinkedIn, the blanket authority to block viewers from accessing information publicly available on its website for any reason, backed by sanctions of the CFAA, could pose an ominous threat to public discourse and the free flow of information promised by the Internet.

(U.S. District Court, N.D. California 2017)

LinkedIn appealed, but the Ninth Circuit Court of Appeals ruled in favor of hiQ in 2019 (U.S. Court of Appeals, 9th Cir. 2019). The Supreme Court vacated this ruling in 2021, remanding it for further consideration (U.S. Supreme Court 2021). However, the Ninth Circuit reaffirmed its initial decision in 2022, noting that giving companies unrestricted control over publicly available data could lead to monopolistic practices harmful to the public interest:

We agree with the district court that giving companies like LinkedIn free rein to decide, on any basis, who can collect and use data—data that the companies do not own, that they otherwise make publicly available to viewers, and that the companies themselves collect and use—risks the possible creation of information monopolies that would disserve the public interest

(U.S. Court of Appeals, 9th Cir. 2022)

The situation shifted in favor of LinkedIn in October 2022, as both parties filed a motion for summary judgment⁵ in the District Court of Northern California (U.S. District Court, N.D. California 2022b). LinkedIn accused hiQ of violating the User Agreement by creating fake accounts and scraping data while logged in. They estimated hiQ had made over 50 billion server requests in 18 months, with 92 million daily requests.

The court issued a mixed ruling, favoring LinkedIn only concerning the creation of fake accounts. Scraping publicly available data, despite violating LinkedIn's User Agreement, remained uncertain because LinkedIn had not enforced its agreement until 2017. The litigation concluded in December 2022, with a confidential settlement that required hiQ to pay LinkedIn \$500,000 in damages (Neuburger 2022b).

This case marked a pivotal moment in the legal landscape for web scraping. Neuburger (2022a) emphasized that the Ninth Circuit's decision was the most favorable for web scraping in technology law, providing a precedent for lawful scraping of public data without fear of CFAA liability. Reese and Quesenberry (2022) similarly noted that scraping publicly available data is unlikely to violate the CFAA unless fake accounts are used. The intense financial competition between LinkedIn and hiQ Labs, coupled with

⁵ A summary judgment is a court decision without the need of a full trial, which would be more expensive for both parties.

the massive scale of data scraping—over 50 billion requests in 18 months—prompted LinkedIn to take legal action against hiQ Labs.

The legal landscape around web scraping remains uncertain, but recent case law suggests that researchers scraping publicly available data for research purposes should generally face no major legal issues. According to Krotov (2020), a breach of contract claim generally requires explicit agreement to the "Terms of Use" and proven damages (Krotov, Johnson, Silva 2020), making it unlikely to succeed against researchers. In copyright infringement cases, the "fair use" principle allows researchers to transform copyrighted material in innovative ways, reducing the risk of legal repercussions (Krotov, Johnson, Silva 2020). Also, data points are facts, which are not subject to U.S. copyright.

LinkedIn's restrictive stance on scraping reflects the broader trend among tech companies seeking to protect user data and control access. However, researchers should critically assess these policies when scraping LinkedIn, especially when the data is publicly available, and the research serves a broader public interest (Luscombe, Dick, Walby 2022). By developing strategies that balance compliance with legal and ethical guidelines, researchers can responsibly navigate LinkedIn's terms and conditions. They must consider the public value of their work while challenging any undue restrictions that limit their ability to analyze societal trends effectively.

The Brazilian General Data Protection Law (*Lei Geral de Proteção de Dados* or LGPD) sets the framework for data protection in Brazil. It allows the processing of publicly accessible personal data without explicit consent if the data subject made it public themselves (Article 7, Paragraph 4). However, this must be done in good faith, considering the original purpose and public interest that justified making the data public (Article 7, Paragraph 3). Researchers must still respect the data subject's rights and the law's core principles, ensuring transparency and minimizing misuse. This provides a clear guideline for responsibly scraping public LinkedIn profiles in Brazil while balancing legal compliance and user privacy.

In the subsequent section, we will explore the ethical considerations associated with scraping LinkedIn profiles, focusing on how researchers can conduct data collection responsibly and respect user privacy while adhering to these evolving legal standards.

Ethical considerations on scraping LinkedIn

Krotov, Johnson, and Silva (2020) conducted a comprehensive review of the legal and ethical aspects of web scraping. They identify several potential harmful consequences of web scraping, including:

- *Web Crawling Restrictions:* Websites use a *robots.txt* file to block web scraping. This file specifies which parts of the site can be accessed by automated bots. The purpose is often to protect server capacity and ensure user privacy.
- *Individual Privacy:* Scraping data from a website may compromise user privacy.
- *Discrimination and Bias:* Preexisting biases in data sources can be reflected in the final product of the scraping process.

- *Organizational Privacy*: Web scraping can potentially reveal confidential information about organizations.
- *Diminishing Value for the Organization*: Products created with scraped data might lead to financial losses for the original data owner.
- *Data Quality and Impact on Decision-Making*: Scraped data may contain inaccuracies or fake information, leading to incorrect decisions.

As Luscombe et al. (2022) state, the answer to the ethical concerns in web scraping is that *it depends*. The potential consequences should be carefully considered when designing and using web scrapers. Ethical (and legal) concerns must be addressed, even if it makes the technical implementation more challenging. Additionally, researchers should ensure that their work serves the public interest.

To address these ethical challenges, researchers should adhere to the following guidelines when scraping LinkedIn or similar platforms:

- *Web crawling restrictions provided*: Although LinkedIn prohibits scraping via robots.txt, it allows large tech companies to access its data. This practice raises ethical concerns about potential information monopolies (U.S. District Court, N.D. California 2022a). Public interest should therefore take precedence to ensure fair data access. Still, researchers should implement a delay in their scraper to limit the frequency of requests and run the scraper outside peak working hours to minimize server load. Luscombe et al. (2022) recommends a delay between 3 and 10 seconds.
- *Individual privacy*: All scraping activities should be performed while logged out, ensuring the collection of publicly available data only. No fake accounts should be used, and all data should be anonymized before publication to protect individual privacy.
- *Discrimination and bias*: Posts containing discriminatory information are protected by login authentication and should not be included in data collection.
- *Organizational privacy*: The data collected should be a small portion of LinkedIn's overall user base and should focus on public user information. Researchers should not infer sensitive details about business operations or infrastructure from the data.
- *Diminishing Value for the Organization*: Researchers should provide insights without directly competing with LinkedIn's products. *LinkedIn Learning for Higher Education* offers universities video courses to help students develop in-demand skills. When evaluating university alumni, for example, researchers will not compete with LinkedIn, as "Skills" is not a field available on public profiles.
- *Data quality and impact on decision-making*: Researchers should assess the quality of LinkedIn data and account for artifacts like fake or duplicate accounts that could distort findings. They should ensure that the data used is reliable and accurate.

Technical considerations on scraping LinkedIn

Scraping LinkedIn presents numerous technical challenges due to the platform's robust security measures. LinkedIn employs a combination of IP address tracking, header analysis, browser fingerprinting, and TCP fingerprinting to block automated scraping tools (ScrapeOps 2023). These defensive strategies can complicate the process for researchers who lack expertise in computer science or software development.

LinkedIn's security measures include:

- **IP Address Tracking:** Monitoring and potentially blocking IP addresses that make too many requests.
- **Header Analysis:** Examining HTTP headers to identify non-human behavior.
- **Browser Fingerprinting:** Collecting data on the browser's environment to distinguish between legitimate users and bots.
- **TCP Fingerprinting:** Analyzing TCP packets to identify automated tools.

Most available content on scraping LinkedIn comes from blog posts and tutorials written by third-party data providers or proxy companies (Kaspr 2023; Bright Data 2024b; lemlist 2024; Scrapin 2023; Scrapperapi 2022; Octoparse 2023; ScrapeOps 2023; Proxycurl 2022; PhantomBuster 2024). These sources provide some useful information but often lack comprehensive details, as their primary goal is to promote and sell their own services.

Modern websites, including LinkedIn, increasingly use dynamic content and JavaScript, which makes scraping more complex (Mitchell 2018). JavaScript can dynamically generate content after the initial page load, requiring scraping algorithms to execute JavaScript code to retrieve the desired data. Every modern browser can run JavaScript, so companies utilize this fact to identify if and which browser is being used before sending the response. Websites may also use obfuscated JavaScript to hinder scraping efforts.

To scrape such websites, tools like Selenium can be employed. Selenium is a tool that automates web browsers, mimicking a real user, allowing researchers to interact with web pages programmatically (Selenium 2024). This functionality aids in circumventing detection based on navigation patterns and blocking based on user-agents.

Researchers can learn from previous works that have successfully scraped LinkedIn profiles, typically following a three-step process: profile search, profile access, and data extraction. Nevertheless, it is imperative to enhance these methods to circumvent the requirement of being logged in:

- **Goncalves et al. (2014):**
 - **Profile search:** Uses an alumni list to create name variations and searches for candidate pages using the Google Search API.

- Profile access: Compares candidate pages to reference pages using cosine similarity to select relevant profiles.
- Parsing: Extracts academic, professional, and personal data and stores it in a database.
- Almeida (2018):
 - Profile search: Searches for LinkedIn IDs using a list of familiar names and specific URL queries.
 - Profile access: Constructs LinkedIn profile URLs and accesses them directly using Selenium.
 - Parsing: Extracts data using a combination of DOM tree traversal and regular expressions.
- Agostinho (2021):
 - Profile search: Creates name variations from an alumni list and searches for matching profiles within LinkedIn search engine using Selenium.
 - Profile access: Accesses matching profiles via Selenium and saves the HTML.
 - Parsing: Compares institution names, course names, and entry/graduation years to confirm alumni identities.

From our research, all authors who successfully accessed public LinkedIn data on their own used Selenium, so it seems a promising technique to seek.

Luscombe et al. (2022) identified 8 defensive strategies that websites use to block web scraping and their respective workarounds. Table 4 summarizes these and identify if our ethical and legal concerns allow us to circumvent each strategy. Before presenting the defensive strategies, it's important to understand some key terms in web scraping (Mitchell 2018):

- **Proxy:** An intermediary server that routes client requests to other servers. It enables the rotation of IP addresses to avoid being blocked, as the server sees traffic coming from multiple different sources.
- **User-Agent:** A string that a browser sends to a web server to identify the type of browser and operating system in use. By modifying this string, a scraper can mimic a legitimate browser and avoid detection.
- **Cookies:** Small data files stored in the browser, remembering information about a user's session. Scrapers can copy cookies from legitimate sessions to circumvent restrictions or simulate human behavior.
- **CAPTCHA:** A challenge-response test designed to differentiate between humans and bots. CAPTCHA-solving services and tools like Selenium can help bypass this protection.
- **API Key:** An authentication token used to identify and authorize API requests. By rotating API keys among different users or services, researchers can simulate multiple identities to avoid triggering rate limits.

Table 4. Defensive strategies and scraper solutions.

| Defensive strategy | Scraper solution | Is circumventing it aligned with ethical and legal concerns? |
|--|---|---|
| Defining a robots.txt file | Respect rating limitations and if no access is allowed, analyze if public interest should take precedence | Yes |
| Banning IP | Rotating IP address from a pool of IP (i.e., using a proxy) | Yes |
| Rate-limiting IP requests | Implement delays on the requests to avoid being blocked | Yes |
| User-agent blocking | Modify user-agent to mimic a given browser and device | Yes |
| Banning by navigation-based detection, e.g., reCAPTCHA | Copy cookies from a human session, pay CAPTCHA solving services, use a human-mimicking web browser tool like Selenium | Yes, as long as no cookies are copied from a logged in session (which would be the same as being logged in) |
| Requiring email verification | Use a pool of email addresses or temporary/"burner" emails | No, if it means that would need to be logged in |
| Requiring mobile verification | Use "burner" mobile phone number services (like Twilo) | No, if it means that would need to be logged in |
| Requiring an API key | Rotate a pool of API keys to simulate different users | Yes, as long as the API is free, and it does not need a logged in user to utilize it |

CONCLUSIONS

Scraping LinkedIn data remains a technically challenging yet viable approach for acquiring public career data, offering researchers unparalleled control over their data acquisition strategy. Although LinkedIn's User Agreement prohibits scraping, recent court decisions and privacy laws favor responsible acquisition of public data, suggesting that scraping can be conducted without significant legal risk. Researchers must navigate technical, legal, and ethical challenges, ensuring that their scraping practices respect privacy and align with the public interest.

By adhering to best practices, such as anonymizing data, implementing rate limits, and avoiding competition with LinkedIn's commercial products, researchers can responsibly harness LinkedIn data for academic purposes. This paper provides comprehensive guidance for academics seeking to navigate these challenges effectively, emphasizing the importance of using advanced tools like Selenium to circumvent technical barriers and ensuring ethical compliance in all scraping activities.

REFERENCES

ABEL, Guy J., ZHU, XiaoXia and HUANG, Ziyue, 2023. Exploring Chinese human capital flight using university alumni data. *Asian Population Studies*. 2023. Vol. 0, no. 0, p. 1–23. DOI 10.1080/17441730.2023.2289705.

AGARWAL, Sumit, LIN, Yupeng, SHEN, Michael and WU, Sirui, 2023. *Banking Crisis Regulator..* Online. SSRN Scholarly Paper. 12 January 2023. Rochester, NY. 4385103. [Accessed 16 March 2024].

AGOSTINHO, Jackson Willian Silva, 2021. COLETA DE DADOS DE EGRESSOS VIA WEB SCRAPING DO LINKEDIN E DO ESCAVADOR. . 2021.

AHN, Jaehan, HOITASH, Rani, HOITASH, Udi and KRAUSE, Eric, 2023. *The Turnover, Retention, and Career Advancement of Female and Racial Minority Auditors: Evidence from Individual LinkedIn Data..* Online. SSRN Scholarly Paper. 22 June 2023. Rochester, NY. 4488379. [Accessed 10 March 2024].

ALMEIDA, 2018. ALUMNI TOOL: RECUPERAÇÃO DE DADOS PESSOAIS NA WEB EM REDES SOCIAIS AUTENTICADAS. Online. MESTRE EM INFORMÁTICA. Rio de Janeiro, Brazil: PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO. [Accessed 16 March 2024].

BISTA, Baibhav, SHAKYA, Aman, JOSHI, Basanta, POKHREL, Anusandhan, DANGOL, Lumanti, KEDIA, Mohit and BARAL, Daya Sagar, 2021. An Alumni Portal and Tracking System. *Journal of the Institute of Engineering*. 12 April 2021. Vol. 16, no. 1, p. 7–14. DOI 10.3126/jje.v16i1.36529.

BRIGHT DATA, 2024a. Bright Initiative - a caring home for data-centric initiatives. *Bright Initiative*. Online. 2024. Available from: <https://brightinitiative.com/> [Accessed 1 April 2024].

BRIGHT DATA, 2024b. Success! The ultimate guide to scraping LinkedIn. *Bright Data*. Online. 2024. Available from: <https://brightdata.com/blog/how-tos/linkedin-scraping-guide> [Accessed 29 March 2024].

BRUNS, Axel, 2013. Faster than the speed of print: Reconciling 'big data' social media analysis and academic scholarship. *First Monday*. Online. 3 October 2013. DOI 10.5210/fm.v18i10.4879. [Accessed 2 April 2024].

CHAPARALA, Pushya, JUKUNTLA, Amar, REDDY, V Sasidhar, VINAYAK, V Vishnu and SUDHA, T Pavani, 2023. Extraction and Upadation of Alumni Information from Web Profiles Using Web Scraping. In: *2023 International Conference on Quantum Technologies, Communications, Computing, Hardware and Embedded Systems Security (iQ-CHESS)*. Online. September 2023. p. 1–7. DOI 10.1109/iQ-CHESS56596.2023.10391404. [Accessed 10 March 2024].

COLL, LIANA, 2021. Alumni platform already has 5 thousand members. *Unicamp*. Online. 16 March 2021. Available from: <https://www.unicamp.br/en/unicamp/noticias/2021/03/16/plataforma-alumni-ja-tem-5-mil-membros> [Accessed 1 April 2024].

CONTRERAS, Jonathan, 2023. *Location-based Open Source Intelligence to Infer Information in LoRa Networks*. Online. Available from: <https://www.merlin.uzh.ch/publication/show/23905> [Accessed 16 March 2024].

CRUNCHBASE, 2024. hiQ Labs - Crunchbase Company Profile & Funding. *Crunchbase*. Online. 2024. Available from: <https://www.crunchbase.com/organization/hiq-labs> [Accessed 2 April 2024].

DEL RIO-CHANONA, R. Maria, MEALY, Penny, BEGUERISSE-DÍAZ, Mariano, LAFOND, François and FARMER, J. Doyne, 2021. Occupational mobility and automation: a data-driven network model. *Journal of The Royal Society Interface*. January 2021. Vol. 18, no. 174, p. 20200898. DOI 10.1098/rsif.2020.0898.

EISFELDT, Andrea L., SCHUBERT, Gregor and ZHANG, Miao Ben, 2023. *Generative AI and Firm Values..* Online. Working Paper. May 2023. National Bureau of Economic Research. 31222. Working Paper Series. [Accessed 16 March 2024]. DOI: 10.3386/w31222

GONCALVES, Gabriel Resende, FERREIRA, Anderson Almeida, TAVARES DE ASSIS, Guilherme and TAVARES, Andrea labrudi, 2014. Gathering Alumni Information from a Web Social Network. In: *2014 9th Latin American Web Congress*. Online. Ouro Preto: IEEE. October 2014. p. 100–108. ISBN 978-1-4799-6953-1. DOI 10.1109/LAWeb.2014.17. [Accessed 16 March 2024].

JONES, Faye R, MARDIS, Marcia A, MCCLURE, Charles M and RANDEREE, Ebrahim, 2017. ALUMNI TRACKING: PROMISING PRACTICES FOR COLLECTING, ANALYZING, AND REPORTING EMPLOYMENT DATA. . 2017.

KASPR, 2023. How to Scrape Data From LinkedIn [Guide Step-By-Step]. . Online. 6 December 2023. Available from: <https://www.kaspr.io/blog/how-to-scrape-data-from-linkedin> [Accessed 2 April 2024].

- KERR, Orin S., 2015. *Norms of Computer Trespass*.. Online. SSRN Scholarly Paper. 2 May 2015. Rochester, NY. 2601707. Available from: <https://papers.ssrn.com/abstract=2601707> [Accessed 15 April 2024].
- KROTOV, Vlad, JOHNSON, Leigh and SILVA, Leiser, 2020. Tutorial: Legality and Ethics of Web Scraping. *Faculty & Staff Research and Creative Activity*. Online. 15 December 2020. DOI <https://doi.org/10.17705/1CAIS.04724>.
- LEMLIST, 2024. LinkedIn Scraping: How to do it? [Step by Step Guide]. *lemlist*. Online. March 2024. Available from: <https://www.lemlist.com/blog/linkedin-scraping> [Accessed 29 March 2024].
- LIANG, Chuchu, LOURIE, Ben, NEKRASOV, Alex and YOO, Il Sun, 2023. *Voluntary Disclosure of Workforce Gender Diversity*.. Online. SSRN Scholarly Paper. 3 May 2023. Rochester, NY. 3971818. [Accessed 16 March 2024].
- LIN, Yupeng, SHEN, Michael, SHI, Rui and ZENG, Jean (Jieyin), 2023. *The Falling Roe and Relocation of Skilled Women: Evidence from a Large Sample of Auditors*.. Online. SSRN Scholarly Paper. 25 September 2023. Rochester, NY. 4324172. Available from: <https://papers.ssrn.com/abstract=4324172> [Accessed 16 March 2024].
- LINKEDIN, 2022. User Agreement | LinkedIn. *LinkedIn*. Online. February 2022. Available from: <https://www.linkedin.com/legal/user-agreement> [Accessed 2 April 2024].
- LINKEDIN, 2023. Profile API - LinkedIn. . Online. 8 May 2023. Available from: <https://learn.microsoft.com/en-us/linkedin/shared/integrations/people/profile-api> [Accessed 17 March 2024].
- LINKEDIN, 2024a. About LinkedIn. . Online. 2024. Available from: <https://about.linkedin.com/> [Accessed 1 April 2024].
- LINKEDIN, 2024b. Compliance FAQ - LinkedIn. . Online. 2024. Available from: <https://learn.microsoft.com/en-us/linkedin/compliance/compliance-api/compliance-faq> [Accessed 17 March 2024].
- LISBOA, Alveni, 2023. LinkedIn supera rivais de peso e é a rede social preferida dos brasileiros. *Canaltech*. Online. 21 March 2023. Available from: <https://canaltech.com.br/redes-sociais/linkedin-supera-rivais-de-peso-e-e-a-rede-social-preferida-dos-brasileiros-242502/> [Accessed 1 April 2024].
- LOHSSE, Sebastian, SCHULZE, Reiner and STAUDENMAYER, Dirk, 2017. *Trading data in the digital economy: legal concepts and tools: Münster colloquia on EU law and the digital economy III*. 1st edition. Oxford: Hart Publishing. ISBN 978-3-8487-4565-4.
- LUNGU, Eliza Olivia, ZAMFIR, Ana Maria, MILITARU, Eva and MOCANU, Cristina, 2012. *Occupational mobility network of the Romanian higher education graduates*.. Online. 2 February 2012. arXiv. arXiv:1202.0404. Available from: <http://arxiv.org/abs/1202.0404> [Accessed 15 November 2023]. arXiv:1202.0404 [physics]
- LUNGU, Eliza Olivia, ZAMFIR, Ana Maria, MOCANU, Cristina and PÎRCIOG, Speranța,

2014. Gravitational Model of the Occupational Mobility of the Higher Education Graduates. *Procedia - Social and Behavioral Sciences*. January 2014. Vol. 109, p. 417–421. DOI 10.1016/j.sbspro.2013.12.483.

LUSCOMBE, Alex, DICK, Kevin and WALBY, Kevin, 2022. Algorithmic thinking in the public interest: navigating technical, legal, and ethical hurdles to web scraping in the social sciences. *Quality & Quantity*. 1 June 2022. Vol. 56, no. 3, p. 1023–1044. DOI 10.1007/s11135-021-01164-0.

METAXAS, Panagiotis and MUSTAFARAJ, Eni, 2014. Sifting the sand on the river bank: Social media as a source for research data. *it - Information Technology*. 28 October 2014. Vol. 56, no. 5, p. 230–239. DOI 10.1515/itit-2014-1047.

MITCHELL, Ryan, 2018. *Web Scraping with Python*. ISBN 978-1-4919-8557-1.

NEUBURGER, Jeffrey D., 2022a. Mixed Ruling in hiQ Labs v. LinkedIn. *The National Law Review*. Online. November 2022. Available from: <https://www.natlawreview.com/article/court-finds-hiq-breached-linkedin-s-terms-prohibiting-scraping-mixed-ruling-declines> [Accessed 2 April 2024].

NEUBURGER, Jeffrey D., 2022b. hiQ and LinkedIn Reach Settlement in Data Scraping Lawsuit. *The National Law Review*. Online. December 2022. Available from: <https://www.natlawreview.com/article/hiq-and-linkedin-reach-proposed-settlement-landmark-scraping-case> [Accessed 10 March 2024].

OCTOPARSE, 2023. How to Scrape LinkedIn Data Without Coding | Octoparse. *Octoparse*. Online. September 2023. Available from: <https://www.octoparse.com/blog/scrape-linkedin-public-data> [Accessed 29 March 2024].

PEREIRA, Jéssica Rocha De Souza, SIMON, Lilian Wrzesinski and PACHECO, Andressa Sasaki Vasques, 2021. A GESTÃO DO ACOMPANHAMENTO DE EGRESSOS EM UMA UNIVERSIDADE FEDERAL. *Revista Interdisciplinar Científica Aplicada*. 1 October 2021. Vol. 15, no. 4, p. 101–125.

PHANTOMBUSTER, 2024. LinkedIn Job Scraper tutorial | PhantomBuster. *PhantomBuster*. Online. 2024. Available from: <https://phantombuster.com/automations/linkedin/6772788738377011/linkedin-job-scraper/tutorial> [Accessed 29 March 2024].

POSSLER, Daniel, BRUNS, Sophie and NIEMANN-LENZ, Julia, 2019. Data Is the New Oil—But How Do We Drill It? Pathways to Access and Acquire Large Data Sets in Communication Science. *International Journal of Communication*. 8 September 2019. Vol. 13, no. 0, p. 18.

PROXYCURL, 2022. The definitive guide to build your own LinkedIn Profile Scraper for 1M profiles (2022). *Proxycurl Blog*. Online. 2022. Available from: <https://nubela.co/blog/tutorial-how-to-build-your-own-linkedin-profile-scraper-2020/> [Accessed 29 March 2024].

PUCPR, 2022. Alumni - PUCPR. *PUCPR*. Online. 2022. Available from: <https://www.pucpr.br/alumni-2/> [Accessed 17 March 2024].

REESE, ALEX and QUESENBERRY, RAVEN, 2022. What Recent Rulings in ‘hiQ v. LinkedIn’ and Other Cases Say About the Legality of Data Scraping. *Farella Braun + Martel LLP*. Online. December 2022. Available from: <https://www.fbm.com/publications/what-recent-rulings-in-hiq-v-linkedin-and-other-cases-say-about-the-legality-of-data-scraping/> [Accessed 10 March 2024].

SCRAPEOPS, 2023. Python Scrapy - Build A LinkedIn People Profile Scraper [2023] | ScrapeOps. *ScrapeOps*. Online. 2023. Available from: <https://scrapeops.io/python-scrapy-playbook/python-scrapy-linkedin-people-scraper/> [Accessed 29 March 2024].

SCRAPERAPI, 2022. Easy Guide on Scraping LinkedIn With Python + Full Code! *ScrapAPI*. Online. 27 June 2022. Available from: <https://www.scrapAPI.com/blog/linkedin-scraper-python/> [Accessed 29 March 2024].

SCRAPIN, 2023. How to Scrape LinkedIn Using a LinkedIn Scraper: 5 methods | ScrapIn. *ScrapIn*. Online. November 2023. Available from: <https://www.scrapin.io/blog/linkedin-scraper> [Accessed 29 March 2024].

SELENIUM, 2024. Selenium. *Selenium*. Online. 2024. Available from: <https://www.selenium.dev/> [Accessed 29 March 2024].

TE, Yiea-Funk, WIELAND, Michèle, FREY, Martin, PYATIGORSKAYA, Asya, SCHIFFER, Penny and GRABNER, Helmut, 2023. Making it into a successful series a funding: An analysis of Crunchbase and LinkedIn data. *The Journal of Finance and Data Science*. 1 November 2023. Vol. 9, p. 100099. DOI 10.1016/j.jfds.2023.100099.

U.S. COURT OF APPEALS, 9TH CIR., 2019. *HIQ LABS, INC. v. LINKEDIN CORPORATION*. September 2019. United States Court of Appeals for the Ninth Circuit 17-16783 D.C. No. 3:17-cv-03301-EMC.

U.S. COURT OF APPEALS, 9TH CIR., 2022. *HIQ LABS, INC. v. LINKEDIN CORPORATION*. April 2022. United States Court of Appeals for the Ninth Circuit 17-16783 D.C. No. 3:17-cv-03301-EMC.

U.S. DISTRICT COURT, N.D. CALIFORNIA, 2017. *HIQ LABS, INC. v. LINKEDIN CORPORATION*. August 2017. District Court of Northern California 17-cv-03301-EMC Docket No. 23.

U.S. DISTRICT COURT, N.D. CALIFORNIA, 2022a. *HIQ LABS, INC. v. LINKEDIN CORPORATION*. October 2022. District Court of Northern California 17-cv-03301-EMC, Document 404, Docket Nos. 336-339 355.

U.S. DISTRICT COURT, N.D. CALIFORNIA, 2022b. *HIQ LABS v. LINKEDIN CORPORATION*. December 2022. District Court of Northern California .

U.S. SUPREME COURT, 2021. *HIQ LABS, INC v. LINKEDIN CORPORATION (Supreme Court)*. 14 June 2021. Supreme Court .

USP, 2024. Dados Analíticos - Alumni USP. *Alumni USP*. Online. 2024. Available from: <https://www.alumni.usp.br/alumniemnumeros/> [Accessed 17 March 2024].

WANG, Chao, ZHU, Hengshu, HAO, Qiming, XIAO, Keli and XIONG, Hui, 2021. Variable Interval Time Sequence Modeling for Career Trajectory Prediction: Deep Collaborative Perspective. In: *Proceedings of the Web Conference 2021*. Online. New York, NY, USA: Association for Computing Machinery. 3 June 2021. p. 612–623. WWW '21. ISBN 978-1-4503-8312-7. DOI 10.1145/3442381.3449959. [Accessed 15 November 2023].

YAMASHITA, Michiharu, LI, Yunqi, TRAN, Thanh, ZHANG, Yongfeng and LEE, Dongwon, 2022. Looking Further into the Future: Career Pathway Prediction. . 2022.

ZAMFIR, Ana-Maria, MATEI, Monica Mihaela and LUNGU, Eliza Olivia, 2013. Influence of Education-job Mismatch on Wages among Higher Education Graduates. *Procedia - Social and Behavioral Sciences*. 10 October 2013. Vol. 89, p. 293–297. DOI 10.1016/j.sbspro.2013.08.849.

ZHANG, Le, ZHU, Hengshu, XU, Tong, ZHU, Chen, QIN, Chuan, XIONG, Hui and CHEN, Enhong, 2019. Large-Scale Talent Flow Forecast with Dynamic Latent Factor Model? In: *The World Wide Web Conference*. Online. San Francisco CA USA: ACM. 13 May 2019. p. 2312–2322. ISBN 978-1-4503-6674-8. DOI 10.1145/3308558.3313525. [Accessed 16 March 2024].

ZHANG, Yutao, TANG, Jie, YANG, Zhilin, PEI, Jian and YU, Philip S., 2015. COSNET: Connecting Heterogeneous Social Networks with Local and Global Consistency. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Online. New York, NY, USA: Association for Computing Machinery. 10 August 2015. p. 1485–1494. KDD '15. ISBN 978-1-4503-3664-2. DOI 10.1145/2783258.2783268. [Accessed 15 November 2023].