12-15-2020

Tutorial: Legality and Ethics of Web Scraping

Vlad Krotov

Leigh Johnson
Murray State University

Leiser Silva
University of Houston

Recommended Citation

Legality and Ethics of Web Scraping, Communications of the Association for Information Systems (forthcoming)

3 authors, including:

Vlad Krotov
Murray State University
42 PUBLICATIONS **374** CITATIONS

Some of the authors of this publication are also working on these related projects:

Addressing barriers to big data View project

Web Scraping Framework: An Integrated Approach to Retrieving Big Qualitative Data from the Web View project

Tutorial: Legality and Ethics of Web Scraping

Vlad Krotov
Murray State University
vkrotov@murraystate.edu

Leigh Redd Johnson
Murray State University

Leiser Silva
University of Houston

Tutorial: Legality and Ethics of Web Scraping

Vlad Krotov

Murray State University
vkrotov@murraystate.edu

Leigh Redd Johnson   **Leiser Silva**

Murray State University   University of Houston

Abstract:

Automatic retrieval of data from the Web (often referred to as Web Scraping) for industry and academic research projects is becoming a common practice. A variety of tools and technologies have been developed to facilitate Web Scraping. Unfortunately, the legality and ethics of using these tools for collecting data are often overlooked. Failure to pay due attention to these aspects of Web Scraping can result in serious ethical controversies and lawsuits. This paper reviews legal literature together with the literature on ethics and privacy to identify broad areas of concern together with a list of specific questions that need to be addressed by researchers and practitioners engaged in Web Scraping. Reflecting on these questions and concerns can potentially help the researchers decrease the likelihood of ethical and legal controversies in their work.

Keywords: big data, web data, web scraping, web crawling, law, legality, ethics, privacy.

[Department statements, if appropriate, will be added by the editors. Teaching cases and panel reports will have a statement, which is also added by the editors.]

[Note: this page has no footnotes.]

This manuscript underwent [editorial/peer] review. It was received xx/xx/20xx and was with the authors for XX months for XX revisions. [firstname lastname] served as Associate Editor. **or** The Associate Editor chose to remain anonymous.]

1   Introduction

In the past, finding data for industry and academic research projects was difficult and costly (Munzert et al. 2015). Today, the increasing digitalization and virtualization of social processes have resulted in zettabytes (billions of gigabytes) of data available on the Web (Cisco Systems 2016). This data provides a granular, real-time representation of numerous processes, relationships, and interactions in the socio-material space (Krotov and Tennyson 2018). Thus, these vast volumes of Big Web Data present academic researchers with ample opportunities for answering new and old research questions with more rigor, precision, and timelines (Constantiou and Kallinikos 2015). Practitioners can leverage this data for developing a better understanding of their customers, formulating strategies based on these findings, and, ultimately, improving organizational performance (Ives et al. 2016).

Unfortunately, harnessing these vast volumes of Web data presents serious technical, legal, and ethical challenges. While there has been a proliferation in tools and technologies that can be used for Web Scraping (Munzert et al. 2015), legality and ethics of data collection from the Web are still "grey areas" (Snell and Menaldo 2016). While existing legal frameworks can be applied, to some extent, to the emerging practice of Web Scraping, the ethical issues surrounding Web scraping have largely been ignored. This paper reviews the legal literature together with the general Information Systems (IS) literature related to Web data, ethics, and privacy to identify broad areas of concern together with specific issues that need to be addressed when collecting data from the Web using automated tools. Compliance with these legal and ethical requirements can help industry and academic researchers decrease the likelihood of legal problems and ethical controversies in their work and, overall, foster research relying on Web data.

2   Web Scraping Explained

2.1   Big Web Data

The data available on the Web is comprised of structured, semi-structured, and unstructured quantitative and qualitative data in the form of Web pages, HTML tables, Web databases, emails, tweets, blog posts, photos, videos, etc. (Watson 2014). Harnessing Big Web Data requires addressing a number of technical issues related to volume, variety, velocity, and veracity of data on the Web (IBM 2018).

First, the data on the Web is often characterized by vast volume measured in Zettabytes (billions of gigabytes) (Cisco Systems 2016). Second, these vast data repositories available on the Web come in a variety of formats and rely on a variety of technological and regulatory standards (Basoglu and White 2015). Third, the data on the Web is not static; it is generated and modified with extreme velocity. Another characteristic of Big Web Data is its veracity (IBM 2018). Due to the open, voluntary, and often anonymous interactions on the Web, there is an inherent uncertainty associated with availability and quality of Web data. A researcher can never be completely sure whether the needed data is or will be available on the Web and whether this data is valid and reliable enough to be used in research (IBM 2018).

2.2   Web Scraping Processes, Technologies, and Tools

Given the volume, variety, and velocity of Big Web Data, collection and organization of this data can hardly be done manually by individual researchers or even large teams of academic researchers or industry data specialists (Krotov and Tennyson 2018; Krotov and Silva 2018). Because of that, researchers often resort to various technologies and tools to automate some or all aspects of Web data collection and organization. This emerging practice of automatic extraction and organization of data from the Web for the purpose of further analysis of this data is often referred to as Web Scraping (Krotov and Silva 2018; Krotov and Tennyson 2018).

Web Scraping consists of the following main, intertwined phases: website analysis, website crawling, and data organization (see Figure 1) (Krotov and Silva 2018; Krotov and Tennyson 2018). Each of the three phases of Web Scraping requires understanding of several Web technologies and at least one popular programming language used in Data Science, such as R or Python. The problem is that, oftentimes, these three phases of Web Scraping cannot be fully automated and often require at least some degree of human involvement. Each of these three phases is further discussed below.
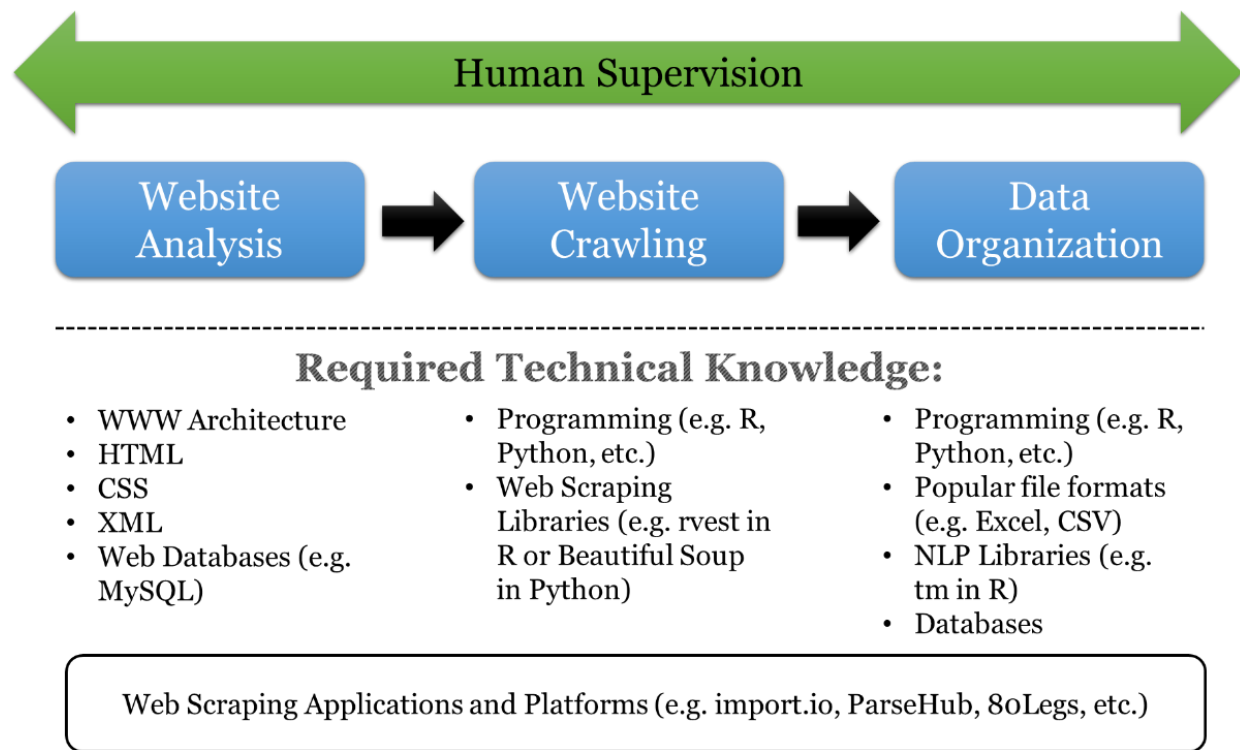
Figure 1. Web Scraping (Adapted from Krotov and Tennyson 2018; Krotov and Silva 2018)

Website analysis requires examining the underlying structure of a website (or websites) or a Web repository (e.g. an online database) for the purpose of understanding how the needed data is stored. This requires a basic understanding of the World Wide Web architecture; mark-up languages (e.g. HTML, CSS, XML, XBRL, etc.); and various Web databases (e.g. MySQL).

Website crawling involves developing and running a script that automatically browses the website and retrieves the needed data. These crawling applications (or scripts) are often developed using such programming languages as R and Python. This has to do with the overall popularity of these languages in Data Science and availability of libraries (e.g. "rvest" package in R or Beautiful Soup library in Python) for automatic crawling and parsing of Web data.

After the necessary data is parsed from the selected Web repository, it needs to be cleaned, pre-processed, and organized in a way that enables further analysis of this data. Given the volume of data involved, a programmatic approach may also be necessary to save time. Many programming languages, such as R and Python, contain Natural Language Processing (NLP) libraries and data manipulation functions that are useful for cleaning and organizing data.

In recent years, there has been a proliferation of "point and click" Web Scraping tools, that allow one to automate at least some of the steps of the Web Scraping process without deep understanding of the Web technologies discussed previously. These tools are available in the form of Cloud-based platforms or stand-alone desktop applications. A representative list of these tools is provided in Table 1 (Sagina 2018).

Table 1. Popular Web Scraping Tools (Adapted from Sagina 2018)

Tool	Description	Notable Features
import.io	A visual, Cloud-based tool for Web data extraction that can be used by individuals and organizations	Web-based, user-friendly interface Easy to set up automatic interactions with web forms and authentication Data extraction jobs can be scheduled Extracted data can be stored in the Cloud Data insights can be generated via reports, charts, and other data visualization techniques Ability to automate data extraction workflows
Dexi.io	Visual, Cloud-based tool positioned as an enterprise	Robots or agents can be created for data extraction and data cleaning

	solution	Intelligent extraction that makes choices and suggestions to resolve issues related to images, pagination, etc.
Octaparse	Advanced Web Scraping tool with workflow automation based on a Web interface	A mature tool with 7 revised versions Intuitive point-and-click interface Supports web authentication Allows exporting data in multiple formats Support scheduled crawling
ParseHub	A desktop application with a graphical interface that can access "tricky" web repositories	Can crawl and scrape web pages that are problematic to typical scrapers; these include web pages with nested comments, images, calendars, pop-up windows, AJAX or JavaScript code, etc. Available for multiple platforms, such as Windows, Mac OS X, and Linux
OutwitHub	A desktop tool with an intuitive graphical interface for extracting contacts from the Web.	Can be used to extract links, email addresses, RSS feeds, data tables, or data from databases
FMiner	A Python-based, visual web scraping tool for designing Web Scraping macros	Can be run on Windows and MacOS machines Supports embedding chunks of Python code for describing actions to be performed on a website Supports advanced web scraping tasks related to captcha solving and data cleaning
80Legs	Cloud-based tool that overcomes typical deficiencies of Web Scrapers related to Web Servers limiting automated access to the content	Overcomes access rate limiting by rotating among numerous IP addresses Contains a set of pre-built applications for various Web Scraping tasks.

One of the main problems with these "point and click" Web Scraping tools is that they do not always work as intended. Oftentimes, they miss certain web pages or are simply not "smart enough" to figure out how to access the needed data points from a website. This is quite understandable. The Web is an open system with a variety of standards and technologies being used to deliver Web content and support user interactions. Moreover, these standards evolve all the time and are often used and interpreted differently. Overall, the content on the Web is very fluid: it is generated and modified at a rapid speed. Because of that, it is not surprising that even some advanced visual Web Scraping tools require custom programming code to be inserted to make them work as intended.

All these technical issues and the fact that many of these tools require hefty subscription fees force many academic and industry researchers to develop their own custom Web Scraping tools using such popular languages as Python and R. Moreover, providing a detailed and unambiguous account of how web data was collected, cleaned, and organized via R or Python computer code can enhance replicability of the entire research protocol used by the study relying on this Web Data (Peng 2011).

2.3   Addressing Legal and Ethical Issues

In this paper we argue that being well-versed in various tools and technologies for retrieving and organizing data from the Web is not sufficient for producing publishable research or data products. Prior to commencing a Web Scraping project, researchers need to reflect on the degree to which their use of Web data is both legal and ethical. Several legal frameworks or broad principles haven been applied in court cases involving disputes over the use of Web data (see Figure 2). These legal frameworks include: illegal access and use of data, breach of contract, copyright, trespass to chattels, and trade secrets. Researchers should also be mindful of the possible ethical consequences of their actions in relation to Web data (see Figure 2). For example, a finding derived from a research project relying on data collected from a website may unintentionally compromise privacy of individuals, violate their rights as research subject, lead to erroneous decisions, or contribute to bias and discrimination. Web data can also reveal confidential information about organizations participating in the activities afforded by the website or the organization that owns the website. It is also possible that some uses of data may reduce perceived value of the website in the eyes of the intended audience.

It is important for researchers to be aware of the legal and ethical issues surrounding Web Scraping. Being mindful of potential legal and ethical issues arising from the collection of Web data and taking proactive measures to address these issues can help researchers avoid costly lawsuits or widely

publicized ethical controversies that can be damaging to their reputation. All the time and resources used to collect Big Data from the Web can quickly become a "big waste" if this data is not collected in accordance with legal and ethical standards and, thus, cannot be used or published by the researchers. This paper is an attempt to propose a framework that can guide industry and academic researchers on the legality and ethics of Web data collection. This framework is called "Legality and Ethics Framework for Web Scraping" (see Figure 2).

3   Web Scraping Literature

A search for journal articles containing "web scraping" or "data scraping" in the title indicates that automatic retrieval and organization of data from Web is still a new, emerging phenomenon. First, most papers containing these two phrases go only a few years back. Second, the total number of journal and conference papers explicitly devoted to Web Scraping found in several popular databases (e.g. EBSCO Business Source Complete, IEEE Xplore Digital Library, JSTOR, etc.) is less than two hundred (see Table 2).

Table 2. Web Scraping Literature

Database	Technical Papers (n)		Technical Total	Non-Tech Papers (n)		Non-Tech. Total	Total Articles
	Tutorials	Solutions		Legal	Ethical		
arXiv	4	3	7	0	0	0	7
EBSCO Academic Search Complete	11	5	16	2	0	2	18
EBSCO Business Source Complete	13	1	14	3	0	3	17
IEEE Xplore	17	86	103	0	0	0	103
JSTOR	0	0	0	2	0	2	2
	Technical Papers Subtotal (n):		140	Non-Tech Papers Subtotal (n):		7	
	Technical Papers (%):		95	Non-Tech Papers (%):		5	
					Grand Total:		**147**

These Web Scraping papers fall into two broad categories: (1) technical, tutorial-like papers on Web Scraping and related tools and technologies; (2) non-technical papers touching on the legal and ethical aspects of Web Scraping. The second category of papers is less numerous. This may suggest that social and ethical issues related to Web Scraping are often overlooked by researchers. Each of these broad literature categories together with some representative papers is discussed in the sections below. The goal of this literature review section is not to provide a rigorous or comprehensive review of Web Scraping literature, but rather to make an analytic argument that there is a need for an integrated, socio-technical approach to Web Scraping that takes into account not only technical, but also ethical and legal issues surrounding this emerging practice.

3.1   Technical Papers

The first sub-category of technical papers on Web Scraping is comprised of tutorial-like, technical papers. These papers either discuss technologies and tools that can be used for Web Scraping in a particular context or discipline or outline potential opportunities that Web Scraping can offer to researchers in various disciplines. For example, the research note by Krotov and Tennyson (2018) discusses popular mark-up languages that are used for storing financial data on the Web and shows how R language, together with its libraries, can be used for gathering, organizing, and pre-processing financial data from the Web. Another tutorial-like paper by Neumann et al. (2017) explains how IXPath technology can be used for metadata harvesting from digital library repositories by researchers interested in scientometric research. The paper by Beoing and Waddell (2017) shows how the data obtained from Craigslist can be retrieved and analyzed to gain a better understanding of the rental market in the United States. The paper by Kirkpatrik (2015) discusses specific ways in which Web Scraping, data mining, and data visualizations

can be used in the field of journalism. Collectively, these papers indicate that there is growing interest in Web Scraping across many scientific disciplines and industries.

The second sub-category of technical papers related to Web Scraping is comprised of mostly conference papers discussing specific technical solutions related to data acquisition from the Web. Most of these papers are authored by Computer Science researchers. For example, a paper by Hernandez-Suarez at el. (2018) discusses ways in which the Twitter API use restrictions (mostly related to the volume and frequency of data retrieval from Twitter) can be bypassed. A conference paper by Chaulagain et al. (2017) proposes an architecture for a Cloud-based tool that can be used for retrieving Big Data from the Web. Another conference paper by Ujwal et al. (2017) proposes a concept of an adaptive Web Scraping tool that adapts to the structural changes of a Web page containing the needed data. There are many other interesting solutions to Web Scraping problems and tools for specific Web Scraping applications that can be found in recent conference papers related to Web Scraping. Collectively, these papers show that Web Scraping is an important and growing area in the field of Computer Science (e.g. Ujwal et al. 2017).

3.2   Papers Related to Legal and Ethical Issues

Another broad and less numerous category of Web Scraping papers is related to legal and ethical Web Scraping (Dreyer and Stockton 2013; Hirschey 2014; Snell and Menaldo 2016; Buchanan 2017; Krotov and Silva 2018). One conclusion emerging from this literature is that there is an inherent paradox surrounding Web data, which makes its retrieval and analysis complex from legal and ethical standpoints. On the one hand, the Web was meant to be open and easily accessible for the public. The same openness principle drives many online business models: website owners benefit from a wider user base accessing the data available on their websites. On the other hand, this Web data is a critical asset for the website owners and, therefore, it needs to be protected. Ideally, many website owners would like this data to be viewed as propriety, meaning owned by the individual or entity behind the website containing this data. This ownership is hard to solidify as the website owner does not necessarily own the data generated by the website's users from a legal standpoint (Dryer and Stockton, 2013).

Either due to all these complexities surrounding Web data ownership or, perhaps, due to the novelty of the Web Scraping phenomenon no legislation specific to Web Scraping has been developed as of now. Instead, researchers and practitioners are often guided by a set of related legal principles and frameworks developed in other eras and different contexts. As of now, Web Scraping is guided by such fundamental legal theories and laws, such as "illegal access and use of data", "breach of contract," "copyright infringement," and "trespass to chattels" (Dreyer and Stockton 2013; Snell and Menaldo 2016).

Ethics and the law are distinct but complimentary (Mingers and Walsham 2010), as legal frameworks usually cover the most obvious ethical issues surrounding a particular practice that can be codified in the form of a relatively unambiguous law (Light and McGrath 2010). But Web Scraping is often plagued by more complex and subtle ethical controversies (Krotov and Silva 2018). While some legal papers touch on the ethical issues associated with Web Scraping, we found no academic article explicitly devoted to more subtle ethical issues surrounding this emerging practice of Web data collection.

3.3   Addressing the Gaps in the Literature

Based on our literature review, we conclude that Web Scraping is often viewed by researchers as a technical phenomenon. Most of the papers that we have discovered through our literature review tackle technical issues surrounding Web Scraping. Very few papers touch on the "softer issues" surrounding this practice of Web data collection. While there are a quite a few legal papers devoted to legality of web scraping, most of these papers are rather exploratory in nature. They attempt to outline some applicable legal frameworks that have been used in relation to Web Scraping and provide some examples of court cases that illustrate application of these legal frameworks (Gold and Latonero 2018; Sellars 2018; Zamora 2019). This is quite understandable. Web Scraping is still a relatively new, emerging practice and the application of various legal theories to Web Scraping has been inconsistent. This makes it difficult, if not impossible, for those engaged in the practice of Web Scraping to determine the legality of their actions. Further, as the social acceptability of Web Scraping emerges, more subtle legal issues may come into play due to a wide variety of research projects that are relying on Web data. Depending on the goals or research questions of a particular industry or scholarly research project, these broad legal frameworks and principles may be applied differently.

To address this gap, we conducted a thorough review of the court cases involving disputes over Web Scraping for the past two decades, analyzed the dispositive facts of each case, and grouped them by similar findings and precedents. The following procedure was used to identify cases relevant to Web Scraping. First, all law journal articles related to Web Scraping were identified through a database called HeinOnline. Second, these legal articles were examined to identify the most recent and relevant articles related to Web Scraping. After that, court cased discussed in this narrow set of articles were identified. These cases identified through "backwards" analysis were researched using Google search engine. Both the case history and recent updates to each case were examined using publicly available online sources found via Google. This review and analysis of court cases allowed us not only to identify and refine the list of broad legal issues and frameworks applicable to Web Scraping, but also to summarize the current status and application of the various legal theories to Web Scraping practices.

It was also discovered through this review of literature and cases that more subtle ethical issues surrounding Web Scraping are often touched upon in the legal literature too, but are rarely explicitly discussed. To the best of our knowledge, there is not a single paper explicitly and wholly devoted to the ethical issues surrounding Web Scraping. To address this gap, we reviewed IS literature related to Web data collection and ethics. While the topic of ethics is not "native" to the IS field, we believe that the field is in a unique position to adopt a holistic, socio-technical perspective on what we view is a complex, socio-technical phenomenon (i.e. Web Scraping). To identify applicable literature, we did a backward citation analysis for one of the early and seminal IS papers devoted to ethics in the Information Age by Mason (1986). The literature was scanned to identify several papers that touch on the topic of Web data collection and use. The list of such papers is not very long or exhaustive. It is rather used to strengthen and expand our own arguments in relation to the ethical issues of Web Scraping that originated from our own practical experiences in relation to automated Web data collection.

The review of legal and IS literature were used to identify the most fundamental legal and ethical principles pertaining to Web Scraping. These principles together with arising implications are discussed in the sections below.

4   Legality of Web Scraping

Currently, there is no legislation that addresses Web Scraping directly; those who engage in Web Scraping activities do so among an uncertain legal landscape. As of now, Web Scraping is guided by a set of related, fundamental legal theories and laws, such as "illegal access and use of data", "breach of contract," "copyright infringement," and "trespass to chattels" (Dreyer and Stockton 2013; Snell and Menaldo 2016). Specific details of how these fundamental legal theories apply to Web Scraping are provided below.

4.1   Illegal Access and Use of Data

Any illegal or fraudulent use of data obtained through Web Scraping is prohibited by several laws. The Computer Fraud and Abuse Act (CFAA) and its comparable state laws are the predominant legal basis for claims involving Web Scraping disputes. Among other provisions, the CFAA prohibits the intentional unauthorized access of a computer or access that exceeds authorization (18 U.S.C. § 1030(a) (2008)) and provides for both civil and criminal penalties (18 U.S.C. § 1030(c) (2008)). As of now, approximately sixty legal opinions have addressed the application of the CFAA to Web Scraping. The majority of these opinions address the question of what constitutes unauthorized access under the law. This variety of opinions suggests that courts remain divided on this issue and little consensus has been reached over the last two decades regarding the application of CFAA to Web Scraping (Sellars 2018). Still, recent cases provide some guidance to researchers employing Web Scraping for data collection regarding their potential civil and criminal liability under the CFAA.

Initially, many courts focused on whether a website's "Terms of Use" or "Terms of Service" policy prohibited Web Scraping activities, and, if so, whether the website user accessed the website in a way that is "unauthorized" (e.g. see EF Cultural Travel BV v. Zefer Corp. 2003; Southwest Airlines Co. v. Farechase, Inc. 2004; EarthCam, Inc. v. OxBlue, Inc. 2017). Since then, other courts have generally required some affirmative action on the website user to become a party to the "Terms of Use" or "Terms of Service," and, consequently, for such access to be deemed unauthorized (see e.g. Alan Ross Machinery Corp. v. Machinio Corp 2018). In 2016, the Ninth Circuit provided that violation of the "Terms of Use" or

"Terms of Service," without more, cannot be the basis for liability under the CFAA (Facebook, Inc. v. Power Ventures, Inc. 2016).

However, access to a website can be revoked and become unauthorized, when the website owner sends a cease and desist letter to a party crawling or scraping the website (Craigslist Inc. v. 3Taps Inc. 2013; Facebook, Inc. v. Power Ventures, Inc. 2016). Still, at least one court found that a cease and desist letter, without more, was insufficient to hold the Web Scraper liable under the CFAA (Ticketmaster L.L.C. v. Prestige Entertainment, Inc. 2018). In 2017, the District Court for the Northern District of California required a website owner to allow Web Scraping on its website, despite a cease and desist letter and an IP block, when the user was accessing publicly available data (hiQ Labs, Inc. v. LinkedIn Corp. 2017). In September of 2019, the Ninth Circuit Court of Appeals upheld the lower court's decision in the hiQ case, noting that Web Scraping publicly available data likely does not violate the CFAA (hiQ Labs, Inc. v. LinkedIn Corp 2019).

4.2   Breach of Contract

As noted above, it has been often argued in the legal field that a website owner can effectively prevent programmatic access to a website by explicitly prohibiting this in the "Terms of Use" or "Terms of Service" policy posted on the website. In addition to unauthorized access and use, failure to comply with these terms may lead to a "breach of contract" on the side of the website's user (Dryer and Stockton 2013). But to hold someone liable for violating the "Terms of Use" or "Terms of Service," the website user generally needs to enter an explicit agreement to comply with the "Terms of Use" or "Terms of Service" policy (e.g. by clicking on a checkbox) (Alan Ross Machinery Corp. v. Machinio Corp 2018; Facebook, Inc. v. Power Ventures, Inc. 2016). Thus, simply prohibiting Web Scraping on the website may not preclude someone from crawling the website from a legal standpoint. Further, the website must establish that it incurred material damages as the result of the breach of the "Terms of Use" or "Terms of Service" in order to be successful on a breach of contract claim.

In some jurisdictions, web users have been enjoined or prevented from (or held liable for) Web Scraping activities due to a contract between the website and a third party, who owns some or part of the content on the website. In such cases, web scrapers have been held to the terms of the contract between the website and the third party, generally because the "Terms of Use" and "Terms of Service" on the website require compliance with contracts between the website and third parties (e.g. QVC, Inc. v. Resultly, LLC 2016).

4.3   Copyrighted Material

Scraping and republishing data that is owned and explicitly copyrighted by the website owner can lead to a "copyright infringement" case, especially when the party uses the scraped data for financial gain (Dryer and Stockton 2013). However, copyright law does not prevent the collection of data itself, nor does a website necessarily own the data on its website, particularly when it is user-generated content (e.g. content from social media sites). Moreover, ideas cannot be copyrighted – only the specific form or representation of those ideas. Thus, one can use copyrighted data to create summaries of this data, for example.

Also, one can still use copyrighted material on a limited scale under the "fair use" principle when the copyrighted material is transformed in a new or original way. However, this is also a "grey area". While some courts have found that scrapers made fair use of scraped material (Kelly v. Arriba Soft Corp. 2003), others have concluded that the use of such data did not constitute fair use (Associated Press v. Meltwater U.S. Holdings, Inc. 2013). In the Meltwater case, the court indicated that the use of even a small percentage of scraped data, "as little as 4.5%", could be enough to not fall under the fair use exception. It is noteworthy that the data in the Meltwater case was available for purchase; if such data were publicly available, the court may have reached a different conclusion.

Courts have also began to consider whether authorization mechanisms, or lack thereof, constitutes a license to copy scraped data. One such mechanism is the robots.txt protocol. The instructions contained within a robot.txt file uploaded to a web server can inform scrapers about any website restrictions or limitations on Web Scraping, including which areas of the site may be accessed and expectations regarding crawl rate (Sellars 2018). In Parker v. Yahoo, the court suggested that a failure to include a robots.txt file, with specific instructions, could potentially create an implied license under copyright law (2008). In Field v. Google, Inc., another court noted that the instructions provided by a website owner's

use of the robots.txt file prevented the owner from arguing that a scraper's use of the website data was copyright infringement (2006). However, when a website owner does not own the data on the website, an implied license cannot be created by the robots.txt file (Associated Press v. Meltwater U.S. Holdings, Inc. 2013).

4.4   Trespass to Chattels

If Web Scraping overloads or damages a website or a Web server, then the person responsible for the damage can be held liable under the "trespass to chattels" theory (Dryer and Stockton 2013). In 2000, a court awarded an injunction based on trespass against a web scraper when a website was prevented from using a small amount of server resources for other uses (eBay, Inc. v. Bidder's Edge, Inc. 2000). However, since then, courts have ruled that the damage needs to be material and easy to prove in court in order for the owner of the Web server to be eligible for financial compensation (Intel Corp. v. Hamidi 2003). Since the damage is often hard to prove, the "trespass to chattels" theory is not widely used in Web Scraping cases (Gold and Latonero 2018).

4.5   Trade Secrets

Web Scraping should not be used as a deliberate surveillance mechanism that intends to reveal trade secrets of a competing organization. For example, Uber has been accused of the use of Web Scraping to "spy" on competing companies and individual drivers (Rosenblatt 2017). A "trade secret" is a legal concept that, in accordance with the definition of the United States Patent and Trademark Office (2020), constitutes a type of intellectual property that can be protected in court. According to the United States Patent and Trademark Office (2020) definition:

> *Trade secrets consist of information and can include a formula, pattern, compilation, program, device, method, technique or process. To meet the most common definition of a trade secret, it must be used in business, and give an opportunity to obtain an economic advantage over competitors who do not know or use it.*

Thus, the confidential aspect of a company's operations must meet the above definition and criteria in order to be legally protected as a "trade secret". If a certain aspect of a company's operations does not meet this definition (and, yet, the company still wants to keep it confidential), then the issue becomes related to the concept of organization privacy discussed in the next section devoted to ethics.

5   Ethics of Web Scraping

While existing laws and legal theories have been applied to Web Scraping in both courts and legal literature, the ethics of Web Scraping has been limitedly addressed. While there are many perspectives on ethics, ethical principles from the Association of Internet Researchers "Internet Research: Ethical Guidelines 3.0" (the IRE) are most applicable to Web Scraping and, hence, adopted in this paper (Association of Internet Researchers 2019). The IRE promotes "ethical pluralism and cross-cultural awareness," noting that European approaches to ethics and privacy are often deontological in nature (focusing on means or duties rather than the end result), while the United States and United Kingdom focus on a more utilitarian approach (emphasizing the greatest good for the greatest number). With respect to real world problems in the field of Information Systems, Mingers and Newman (2010) recognize that there are different ethical approaches to address various issues and advocate for the use of discourse ethics, where those affected by decisions debate the issues and define universal norms, synthesizing both deontological and utilitarian approaches. The IRE suggests the utilization of a similar pluralistic approach, where a range of ethically defensible positions to a problem (e.g., deontological and utilitarian approaches, among others) should be acknowledged and explored while recognizing that some basic norms emerge among different ethical approaches.

Although applied differently, in both deontological and utilitarian approaches, researchers must generally consider the harms caused as a result of one's actions. In addition to violating existing laws, Web Scraping can result in unintended harm to others that are associated with a particular website, such as the website's owners or customers. These harmful consequences are often hard to predict (Light and McGrath 2010). Yet some possible harmful consequences of Web Scraping are discussed below.

5.1   Web Crawling Restrictions Provided

In the context of Web Scraping, perhaps the first ethical consideration that needs to be addressed is whether a website includes a robots.txt file prohibiting automated web crawling. A web user is not required by any legal or technical restrictions to follow the instructions provided by the robots.txt protocol. Still, social norms generally dictate that web scrapers follow these instructions (Gold and Latonero 2018). There can be a variety of legitimate reasons (e.g. privacy concerns) that can lead website creators to prohibit robots from automatically crawling the website. Failure to obey these instructions prohibiting web crawling can cause unintended harm to the users and owners of the website.

5.2   Individual Privacy and Rights of Research Subjects

A research project relying on data collected from a website may unintentionally compromise privacy of individuals participating in the activities afforded by the website (Mason 1986). For example, by matching the data collected from a website with other online and offline sources, a researcher can unintentionally reveal the identity of those who created the data (Ives and Krotov 2006). This is what happened during the infamous release of search data by America Online (AOL) in 2006, when the company decided to release approximately 500,000 search queries submitted to AOL search engine by its users. Although names of users were replaced by random numbers, the Internet community was quickly able to identify some of the users due to "ego searches" involving their own names and specific geographical locations. The findings derived from this dataset lead to a number of ethical controversies. Some members of AOL performed very disturbing searches indicative of possible use of drugs, violence, and illegal pornographic materials.

Although website users may have no legal or technical protection for data shared on certain websites, they *expect* that such data is somehow protected and private, raising concerns, especially through the lens of deontological ethics. Broader and serious implications arise once data is published, especially when "personal and sensitive data…could be used directly or indirectly against individuals in [certain] countries and political systems" (Association of Internet Researchers 2019). For example, identification of individuals through Web Scraping activities may also be used by security and law enforcement officials for the prosecution of criminal activity (Zuboff 2015), creating issues with a user's assumptions regarding privacy and an unjust violation of rights. Additionally, Web Scraping of certain DNA ancestry databases may not only reveal private genetic and health-related information of a user, it could also be used by various governments and organizations to create a "biometric database useful for identifying nearly any American from a DNA sample" (Regalado 2019).

Even if individual privacy is not violated, the problem is that a website's customers may not have consented to any third party use of their data. Thus, using this data without a consent is a violation of the deontological rights of autonomy and equality of research subjects (Buchanan 2017). Given that informed consent is essentially impossible in certain Web Scraping projects, researchers must consider the steps necessary to protect confidentiality, including deleting certain identifiable information or "pseudonymiz[ing] their data separating keys from the actual data set" (Association of Internet Researchers 2019). These privacy and rights violations can lead to serious consequences for the website owner given the heightened concern with online privacy in the light of the recent privacy scandals involving such organizations as Facebook and Cambridge Analytica.

Further, users disclose information with the understanding that it will be used in a certain context. When it is used for purposes other than those expected by the user, additional issues arise regarding privacy and consent. For example, when booking a flight on a travel website, the user knows that information may be used to offer hotel or other travel-related accommodations. On the other hand, the user would not expect that data to be used later to make pricing decisions as a result of Web Scraping activities (Martin 2015). In the social media context, users understand that the use of the service is connected to acceptance of advertisements in their browsers, but they do not always appreciate how their personal information may be collected, used and disclosed through Web Scraping projects, allowing for exploitation and trafficking of personal data (Wigan and Clarke 2013).

5.3   Discrimination and Bias

Information from Web Scraping activities can contribute to discriminatory practices, inferences drawn on preexisting biases, and prejudicial labeling. In addition to the denial of due process, decisions based on Web Scraping data could lead to new forms of financial and social discrimination (Wigan and Clarke 2013) or unfair profiling practices (Someh et al. 2019). Predictive algorithms based on prior data patterns may

lead to learned prejudice built on institutionalized prejudice (e.g., in the process of college admissions) (Martin 2015). Businesses may target products and services to certain groups based on past behavior (limiting consumer choice) and charge different amounts for such goods and services based on gender, age, race, ethnicity, socioeconomic status, among others (e.g., charging young men higher rates for car insurance) (Newell and Marabelli 2015). Such potential uses of Web data should be foreseen and avoided by researchers.

5.4   Organizational Privacy

Just like individuals have the right to privacy, organizations also have the right to maintain certain aspects of their operations confidential (Mason 1986). Automatic Web Scraping can unintentionally reveal confidential information about the operations of the organization who owns a website that is being scraped. For example, by automatically crawling and counting employment ads on an online recruitment website one can approximate the website's target audience, market share, and revenues. It can also reveal some details and, possibly, flaws in the way the data is stored by the website (Ives and Krotov 2006). Revelations of privacy breaches of an organization's data, employment of discriminatory practices, lack of due process for users, among others, can damage the reputation of the company behind the website, create legal issues, and lead to material financial losses (Markus 2017).

5.5   Diminishing Value for the Organization

If one accesses the website omitting the Web interface made for humans, then the person will not be exposed to the advertisements that the website is using to monetize its content. Moreover, a Web Scraping project can lead to the creation of a data product (e.g. a report) that, without infringing on the copyright, makes it less likely for a customer to purchase a data product from the original owner of the data. In other words, the data product created with the help of Web Scraping can directly or indirectly compete with the business of the website's owner (Hirschey 2014). All this may lead to financial losses to the owner of the website or, at a minimum, an unfair distribution of value from data ownership (Mason 1986).

On the other hand, some website owners partner with researchers in attempt to find ways to monetize these Web Scraping projects or for better brand recognition. Naturally, researchers also should carefully consider how such partnerships affect the independence of their research (Association of Internet Researchers 2019).

5.6   Data Quality and Impact on Decision-Making

Organizations and public officials often make strategic decisions based on data amassed through Web Scraping activities. This can lead to faulty decision-making due veracity of Web data. Due to the anonymous, pluralistic, and "at will" mechanisms used to generate data on the Web, this data is often incomplete, inaccurate, irrelevant, or simply "fake" (Clarke 2016). Despite these issues, the need for consumer analytics can create destructive demand for such data and may encourage Web Scraping activities that collect and sell low quality Web data (Martin 2015). Further, inappropriate use of Web data by one group can have a negative impact on all users of that data in the value chain, magnifying the problem (Someh et al. 2019).   To compound these issues, users often lack the ability to modify aggregated and shared data for accuracy – something that contributes to a likely exponential growth of what this data misrepresents (Someh et al. 2019). Some Internet users deliberately create and spread false information on the Web, contributing to the problem of Web data quality even further and lowering trust of the public to any conclusions derived from this data.

Because of that, important decisions may be based on haphazard, user-generated Web data that lacks quality, relevance to the organization, and compliance with strict academic or professional standards (Constantiou and Kallinikos 2015). This can lead to flawed decisions and financial losses for the organization employing this data or flawed conclusions from research that informs policy makers. Low-quality Web data that is difficult, if not impossible, to interpret properly may be treated as authoritative, allowing organizations to manipulate consumer behavior and public opinion, limit consumer and voter choices, and create artificial markets and political agendas (Wigan and Clarke 2013). These erroneous moves and decisions pose unintended consequences at the level of entire economies and societies.

6 Legality and Ethics Framework for Web Scraping

Based on the analysis of literature presented in this paper, we generate a list of questions that need to be addressed by both researchers and practitioners in order to make a Web Scraping project legal (Dreyer and Stockton 2013; Hirschey 2014; Snell and Menaldo 2016) and ethical (Mason 1986; Ives and Krotov 2006; Buchanan 2017). These questions are as follows:

- Is Web Crawling or Web Scraping explicitly prohibited by the website's "Terms of Use" policy?
- Is the website's data explicitly copyrighted?
- Does the project involve illegal or fraudulent use of the data?
- Can crawling and scraping potentially cause material damage to the website or Web server hosting the website?
- Has the website sent the user a cease and desist letter, blocked the user's IP address, or closed access to data in some other way?
- Does a website's robots.txt protocol instructions significantly limit or prevent Web Scraping activities?
- Can the data obtained from the website compromise individual privacy, rights of research subjects, or non-discrimination principles?
- Can the data obtained from the website reveal confidential information about organizations affiliated with the website?
- Can the project requiring the Web data potentially diminish the value of the service provided by the website?
- Does the quality of the data obtained from the Web have the potential to lead to ill-informed decision-making?

These questions are derived from the broad legal frameworks and ethical issues surrounding the practice of Web Scraping and discussed in the previous section. The legal frameworks or principles upon which the framework is based include: Illegal Access of Data, Illegal Use of Data, Breach of Contract, Copyright, Trespass to Chattels, and Trade Secrets. The ethical dimension includes such concepts as: Web Crawling Restrictions (imposed by a website's owner), Individual Privacy, Rights of Research Subjects, Organizational Privacy, Diminishing Value for the Organization, Discrimination and Bias, Data Quality and Decision Making. These ethical and legal concepts are described in more detail in the previous section.

A visual depiction of these questions, legal frameworks, and ethical concerns is provided in Figure 2, which we call "Legality and Ethics Framework for Web Scraping". No specific links between questions and legal and ethical elements are drawn as there is a great degree of overlap between these elements and the formulated questions; moreover, the links are "loosely coupled" and, therefore, "non-propositional", which is a defining characteristic of a theoretical framework that does not claim to be a "theory" (Bacharach 1989).

Figure 2. Legality and Ethics Framework for Web Scraping

A positive answer to any of the questions listed in Figure 2 may suggest that the Web Scraping project can potentially result in lawsuits or ethical controversies. It may not be necessary to halt a research project involving a potential violation of one or more principles discussed in this paper. For example, copyrighted data can still be used in accordance with the "fair use" principle. Even if "terms of use" prohibit crawling, a permission for automatic collection of data can still be obtained from the website's owner. Still, researchers behind the projects involving a positive answer to any of these questions should reflect on how they will deal with the arising issues in order to avoid ethical controversies or even lawsuits. An example of a study relying on Web Scraping and reflecting on some of the questions and concern areas provided in Figure 2 is provided in Appendix B.

It should also be noted that the list of questions provided in Figure 2 may not be exhaustive of all the legal and ethical controversies that a particular Web Scraping project can produce. The data available on the Web is growing at a rapid pace and so do various tools, technologies, and applications related to this data. Thus, predicting all areas of possible future concern in relation to the use of Web data is hard (if not impossible) (Light and McGrath 2010). We do recommend that researchers always ask themselves the following two overreaching questions that targets both legal and ethical considerations:

- Can my actions in relation to Web data produce harm to individuals, organizations, or communities?
- What can we do to provide reasonable assurance that this unintended harm does not happen?

Reflecting on these issues and topics is especially important for academic researchers who are required by their respective Institutional Review Board or Human Subjects Protection committees to utilize all possible mechanisms for protecting privacy and well-being of their research subjects. We recommend that that academic journals should also incorporate these questions as a formal requirement for their manuscript submission process. This can provide an additional layer of protection for human subjects used in research and interests of other related stakeholders, such as publishers of academic journals. We also hope that these questions will be incorporated into policies and procedures of various academic and industry research organizations to provide protection for human subjects, individual researchers, and organizations conducting research from harm resulting from ethical and legal controversies. We believe that this can foster research relying on data from the Web, which is now a vast repository of valuable datasets spanning both social and material realms.

7   Conclusion

The Big Data available on the Web presents researchers and practitioners with numerous lucrative opportunities. For researchers, these opportunities include leveraging this data for developing a more granular understanding of various old and new social phenomena with better timeliness and precision. Practitioners can leverage Web data for developing a better understanding of their customers and developing data products that help with decisions and strategic planning. But leveraging Big Data from the Web presents both researchers and practitioners with big challenges as well. Apart from the need to learn and deploy new tools and technologies capable of accommodating Big Data, researchers and practitioners intending to use Web Scraping in their research projects need to comply with a number of legal and ethical requirements. Unfortunately, due to the relative novelty of the Web Scraping phenomenon, legality and ethics of Web Scraping are still "grey areas". This paper reflects on some of the legal and ethical issues surrounding Web Scraping. A list of specific questions that need to be addressed by researchers employing Web Scraping is formulated in this paper. A negative answer to all these questions does not necessarily give a clearance to proceed with the research project. This list of questions should rather be used as a starting point for reflecting on the legality and ethics of a research project relying on Web Scraping for data acquisition. A researcher should always be focused on identifying potential harm to individuals and organizations resulting from his or her project and implementing reasonable precautions to prevent this harm from happening. Finally, we would like to stress that the information provided in this paper does not, and is not intended in any way, to constitute legal advice. Both researchers and practitioners are urged to seek qualified legal help when in doubt regarding legality of their Web Scraping projects.

References

Association of Internet Researchers. 2019. "Internet Research: Ethical Guidelines 3.0," Retrieved from https://aoir.org/reports/ethics3.pdf.

Bacharach, S. B. (1989). Organizational theories: Some criteria for evaluation. Academy of management review, 14(4), pp. 496-515.

Basoglu, K. A., and White, Jr. C. E. 2015. "Inline XBRL versus XBRL for SEC Reporting," Journal of Emerging Technologies in Accounting (12:1), pp. 189-199.

Buchanan, E. 2017. "Internet Research Ethics: Twenty Years Later," in Internet Research Ethics for the Social Age: New Challenges, Cases, and Contexts, M. Zimmer and K. Kinder-Kurlanda (eds.), Bern, Switzerland: Peter Lang International Academic Publishers, pp. xxix-xxxiii.

Clarke, R. 2016. "Big Data, Big Risks." Information Systems Journal (26:1).

Constantiou, I. D., and Kallinikos, J. 2015. "New Games, New Rules: Big Data and the Changing Context of Strategy," Journal of Information Technology (30:1), pp. 44-57.

Cisco Systems. 2016. "Cisco Visual Networking Index: Forecast and Methodology, 2014-2019," White Paper. Retrieved from http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.html

DHI Group (2017). About Us. Retrieved from http://www.dhigroupinc.com/home-page/default.aspx

Dryer, A.J., and Stockton, J. 2013. "Internet 'Data Scraping': A Primer for Counseling Clients," New York Law Journal. Retrieved from https://www.law.com/newyorklawjournal/almID/1202610687621

Gold, Z. and Latonero, M. "Robots Welcome? Ethical and Legal Considerations for Web Crawling and Scraping." Washington Journal of Law, Technology & Arts. (13), pp. 275-312.

Hirschey, J. K. 2014. "Symbiotic Relationships: Pragmatic Acceptance of Data Scraping," Berkeley Technology Law Journal (29), pp. 897-927.

Ives, B., and Krotov, V. 2006. "Anything You Search Can Be Used Against You in a Court Of Law: Data Mining in Search Archives," Communications of the Association for Information Systems (18:1), pp. 593-611.

Ives, B., Palese, B., and Rodriguez, J. A. 2016. "Enhancing Customer Service through the Internet of Things and Digital Data Streams," MIS Quarterly Executive (15:4).

IBM. 2018. "The Four V's of Big Data," Retrieved from http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Krotov, V. and Silva, L. 2018. "Legality and Ethics of Web Scraping," The Twenty Fourth Americas Conference on Information Systems.

Krotov, V., and Tennyson, M. 2018. "Scraping Financial Data from the Web Using the R Language," Journal of Emerging Technologies in Accounting, Forthcoming

Lee, I. (2007). An architecture for a next-generation holistic e-recruiting system. Communications of the ACM, 50(7), 81-85.

Light, B., & McGrath, K. 2010. "Ethics and Social Networking Sites: A Disclosive Analysis of Facebook," Information Technology & People (23:44), pp. 290-311.

Markus, M. (2017). "Datification, Organizational Strategy, and IS Research: What's the Score ?" Journal of Strategic Information Systems (26).

Martin, K. 2015. "Ethical Issues in the Big Data Industry." MIS Quarterly Executive (14:2).

Mingers, J., and Walsham, G. 2010. "Towards Ethical Information Systems: The Contribution of Discourse Ethics." MIS Quarterly (34:4).

Mason, R. O. 1986. "Four Ethical Issues of the Information Age," MIS Quarterly, (10:1), pp. 5-12.

Munzert, S., Rubba, C., Meißner, P., and Nyhuis, D. 2015. Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining, Chichester, UK: John Wiley & Sons, Ltd.

Newell, S., and Marabelli, M. 2015. "Strategic Opportunities (and Challenges) of Algorithmic Decision-making: A Call for Action on the Long-term Societal Effects of 'Datification.'" The Journal of Strategic Information Systems (24:1).

Peng, R. D. 2011. Reproducible research in computational science. Science, 334(6060), 1226-1227.

Regalado, A. 2019. "The DNA Database Used to Find the Golden State Killer is a National Security Leak Waiting to Happen." MIT Technology Review,

Retrieved from https://www.tech nologyreview.com/2019/10/30/132142/dna-database-gedmatch-golden-state-killer-security-risk-hack/.

Rosenblatt J. 2017. "Uber Data-Scraping, Surveillance Detailed by Ex-Manager," Bloomberg, Retrieved from https://www.bloomberg.com/news/articles/2017-12-15/uber-data-scraping-surveillance-detailed-in-ex-manager-s-letter

Sagina, I.J. 2018. "The Ultimate List of Web Scraping Tools and Software," KDnuggets, Retrieved from https://www.kdnuggets.com/2018/07/ultimate-list-web-scraping-tools-software.html

Sellars, A. 2018. "Twenty Years of Web Scraping and the Computer Fraud and Abuse Act," Boston University Journal of Science & Technology (24), pp. 372-415.

Snell, J., and Menaldo, N. 2016. "Web Scraping in an Era of Big Data 2.0," Bloomberg BNA. Retrieved from https://www.bna.com/web-scraping-era-n57982073780/

Someh, I., Davern, M., Breidbach, C., and Shanks, G. 2019. "Ethical Issues in Big Data Analytics: A Stakeholder Perspective." Communications of the Association for Information Systems (44).

United States Patent and Trademark Office. 2020. "Trade Secret Policy," Retrieved from https://www.uspto.gov/ip-policy/trade-secret-policy

Watson, H. J. 2014. "Tutorial: Big Data Analytics: Concepts, Technologies, and Applications," Communications of the Association for Information Systems (34:1), pp. 1247-1268.

Wigan, M. R., & Clarke, R. 2013. "Big data's big unintended consequences." IEEE Computer, (46:6).

Zamora, A. 2019. "Making Room for Big Data: Web Scraping and an Affirmative Right to Access Publicly Available Information Online." The Journal of Business, Entrepreneurship & the Law (12), pp. 203-226.

Zuboff, S. 2015. "Big Other: Surveillance Capitalism and the Prospects of an Information Civilization." Journal of Information Technology (30:1).

Appendix A: Court Cases

Alan Ross Machinery Corp. v. Machinio Corp, 2018 WL 6018603 (N.D. Ill. Nov. 16, 2018).

Associated Press v. Meltwater U.S. Holdings, Inc. 931 F. Supp. 2d 537 (S.D.NY 2013).

Computer Fraud and Abuse Act of 1984, 18 U.S.C. § 1030 (2008).

Craigslist Inc. v. 3Taps Inc., 942 F.Supp 2d 962 (N.D. Cal. 2013).

EarthCam, Inc. v. OxBlue, 703 Fed. Appx. 803 (11th Cir. 2017).

eBay, Inc. v. Bidder's Edge, Inc., 100 F. Supp. 2d 1058 (N.D. Cal. 2000).

EF Cultural Travel BV v. Zefer Corp., 318 F.3d 58 (1st Cir. 2003).

Facebook, Inc. v. Power Ventures, Inc. et. al., 844 F.3d 1058 (9th Cir. 2016).

Field v. Google Inc., 412 F. Supp. 2d 1066 (D. Nev 2006).

hiQ Labs, Inc. v, LinkedIn Corp., 273 F.Supp. 3d 1099 (N.D. Cal. 2017).

hiQ Labs, Inc. v. LinkedIn Corp., 938 F. 3d 985 (9th Cir. 2019).

Intel Corp. v. Hamidi, 71 P. 3d 296 (2003).

Kelly v. Arriba Soft Corp., 336 F.3d 811 (9th Cir. 2003).

QVC, Inc. v. Resultly, LLC, 159 F. Supp. 3d 576 (E.D. Pa 2016).

Parker v. Yahoo!, Inc. No. 07-2757, 2008 WL 4410095 (E.D. Pa. September 25, 2008).

Southwest Airlines Co. v. Farechase, Inc. 318 F. Supp. 2d 435 (N.D. Tex. Mar. 19, 2004).

Ticketmaster L.L.C. v. Prestige Entertainment, Inc. et al., No. 17-cv-07232, 2018 WL 654410 (C.D. Cal. Jan. 31, 2018).

Appendix B: The Case of Dice.com

In this appendix we illustrate application of the proposed Legality and Ethics Framework for Web Scraping to Dice (www.dice.com) - an employment website for IT and Engineering professionals (DHI Group 2017). This research project aims to answer the following research question:

> *RQ: What are the most in-demand skills or requirements for the role of a Systems Analyst?*

This question can be potentially answered using the data available on Dice. Dice is one of the leading, specialized recruitment websites (Lee 2007). Historically, the website has specialized in job search and recruitment solutions for IT and Engineering workforce (DHI Group 2017). At a given moment, the website contains close to 80,000 active job listings that are publicly accessible, and has a collection of 2.1 million resumes (not in public access). The website also reports more than 2 million unique visitors to the website every month.

At its most basic level, Dice.com can be viewed as an online bulletin board, where potential job seekers can post their resumes and recruiters representing various companies can post their job ads. Job seekers can then apply to posted jobs, while recruiters can browse through the resumes of IT professionals using the website. Dice.com offers a number of other research and data solutions related to recruitment in IT for companies in Finance, Energy, Healthcare, and Hospitality (e.g. security clearance of job applicants). The company also publishes various types of research in relation to IT labor and recruitment.

Phase 1: Website Analysis

A technical analysis of Dice starts with studying the API manual provided by the website (no longer available). From this manual, one can see that the website allows for one domain to be used by the public for accessing job data: "service.dice.com" The data therein can be obtained using the API in three formats: HTML, XML, and JSON. The website's database of job listings can be queried by supplying the URL devoted to one of these data formats with parameters. Among other things, these parameters allow to narrow down the search to specific keywords or pre-format the resulting output from the Web server. A sample output from an API search query formed to retrieve job listings related to the role of a "systems analyst" is provided in Appendix C.

As one can see from the output in JSON format (see Appendix C), the API does not allow for actual job descriptions to be retrieved. The API provides only for URLs pointing to actual job descriptions. We open one of these URLs in a browser (Google Chrome) for further analysis. In Google Chrome, the "Inspect" feature can be used to analyze the underlying code of a job description page in order to determine which elements contain the needed job description. We need this to retrieve the actual job description for each of the job listings with the help of API programmatically, using an R script (which is discussed in more detail below).

Reflecting on Legality and Ethics of Crawling and Scraping Dice.com

Before commencing the actual data collection process based on the technical analysis described in the previous section, we reflect on data quality as well as legality and ethics of this Web data collection project.

First, we conclude that the data on job descriptions available through Dice.com is of high quality. Job ads are posted mostly by technology and hiring professionals and are further moderated by the website administrators. Thus, we believe that we are not likely to see glaring problems with data quality that can possibly lead to flawed decision making.

Then we examine the "Terms of Use Agreement" for the website. The "Terms of Use Agreement" explicitly allows data access using the API created by the company. This API is available to developers at no charge. The API from Dice can be used to programmatically search through jobs ads and, if necessary, download this information. Thus, we conclude that the company encourages automatic crawling and scraping by the user community.Programmatic retrieval of job descriptions outside of the data returned by the official API raises a slight ethical dilemma. Most web sites, including Dice.com, have an item in their "terms of use" that prohibits accessing website data using "unauthorized means". Thus, we need to reflect whether we are accessing the job descriptions in a way that is authorized. After some deliberations, we conclude that we not doing anything illegal or unethical by going outside of the data returned by the API

(see Appendix C). First, the URLs that point to actual job descriptions are returned by API. This means that the organization does not want to keep those URLs secret. Secondly, the URLs point to publicly available pages. One does not even need an account with Dice.com to view these pages. Thus, we conclude that we can go ahead and retrieve the job descriptions that these URLs returned via the API point to.

A quick examination of the "robots.txt" file does not reveal any restrictions for automatically crawling the areas of the website containing job ads either. Thus, we conclude that we can go ahead and develop an R script that automatically crawls the pages and downloads data.

The rate at which the job descriptions are retrieved from the website (approximately one job description every 1-2 seconds, as it was determined through some "trial" runs using smaller sets) and the relatively small size of the dataset (1105 job ads) does not make it likely that automatic crawling and retrieval of the needed data can somehow damage or slow down the servers where Dice is hosted.

We also see that the website contains a statement that the data provided on the website is copyrighted. Since we do not intent to republish the data we collect "verbatim" (we only intent to publish a high level summary of the most frequent keywords in those job ads), we conclude that our use of this data in compliance with the "fair use" principle. We also determine that if the collected data is somehow aggregated at the company level it may provide unwanted exposure of certain aspects of the operations of the companies which posted these job ads to Dice. Again, since we don't intend to republish this data "verbatim" or group it using company names, we feel like organizational privacy and trade secrets are properly respected.

Overall, we believe that we use the data for a legitimate, non-fraudulent purpose, which is research. We do not believe that results of our research can somehow diminish the value of data for Dice or organizations posting job ads. If anything, such research can provide additional exposure to the company's brand and provide a case for the quality and usefulness of its data.

Thus, we conclude that our use of the website's data is both legal and ethical.

Phase 2: Website Crawling

In this phase, step one involves developing and debugging an R script for crawling the website and downloading the data related to Systems Analyst job listings. Step two involves running the script with some degree of human supervision to retrieve the data. The main outcome of this phase is a data frame (a popular data structure in R similar to a table in Excel) filled with data and meta-data related to Systems Analyst job listings.

R was chosen as a Web Scraping tool for a variety of reasons. First R is free. Using some "ready-made" scraping tools would probably involve paying a hefty subscription fee for a Web Scraping task of this magnitude. Second, R contains several useful libraries (e.g. rvest) for simulating Web sessions and parsing data in various Web formats. Third, R code can be fine-tuned to do a Web Scraping task of any complexity or granularity.

The developed R script is first applied to a moderately-sized data set of 1,104 observations first (23 Web pages containing job descriptions). It took approximately 19 minutes to download data related to 1104 jobs on an office computer with the following configuration:
- Processor: Intel® Core™ i5-4590 CPU @ 3.30 GHz
- RAM: 16GB
- Upload/Download Speed: ~100 Mbps

It took approximately 12 hours to download data for 44,889 job listings that contained words "systems" and "analyst" (not necessarily as a single phrase) using the computer with the configuration provided above.

We also observe that we query the server at rate of approximately one job description per second. We conclude that such crawling rate is not likely to overload the company's Web servers, preventing other users from accessing its resources.

Phase 3: Data Organization

Once all the scraped data is saved in the job_table data frame (data frame is a popular data structure in R – similar to a table in Excel), we can use R to further process and organize data in a Microsoft Excel ® sheet. Using the "xlsx" R package functions for working with Excel files, the entire content of the data frame is saved into one single Excel file. Excel files were used for data storage due to the universality and popularity of this format. Excel is often used for manual data analysis. Most people are quite comfortable viewing and analyzing data in Excel. Moreover, Excel files can be easily imported into virtually any type of software (including R and NVivo).

The file name contains the additional metadata: date and time. This is done so that by simply looking at a file, one can quickly determine when a particular scraping job was performed. Some research projects may require scraping data on multiple dates to come up with a substantially large data sample.

Figure 2B shows how data is organized in an Excel sheet. The file in Figure 2B contains information about Systems Analyst jobs listed on Dice. Each row corresponds to a particular job listing, so that job ads are clearly delineated. Each column represents a particular attribute of each job listing. The specific columns include the URL used to obtain a job description (JobURL), the title of the job listed (JobTitle), the name of the company that posted the job (Company), the date when the job was posted (JobDate), and the full description of the job listings (JobDescription).

Figure 2B. Excel File Content

In addition to having one single Excel file that contains all scraped job listings, there is a separate Excel file for each page of job listings being scraped. By default, the API returns information about 50 job listings at a time. Thus, each of the 50 job listings are stored in a separate Excel file as well. These saved pages
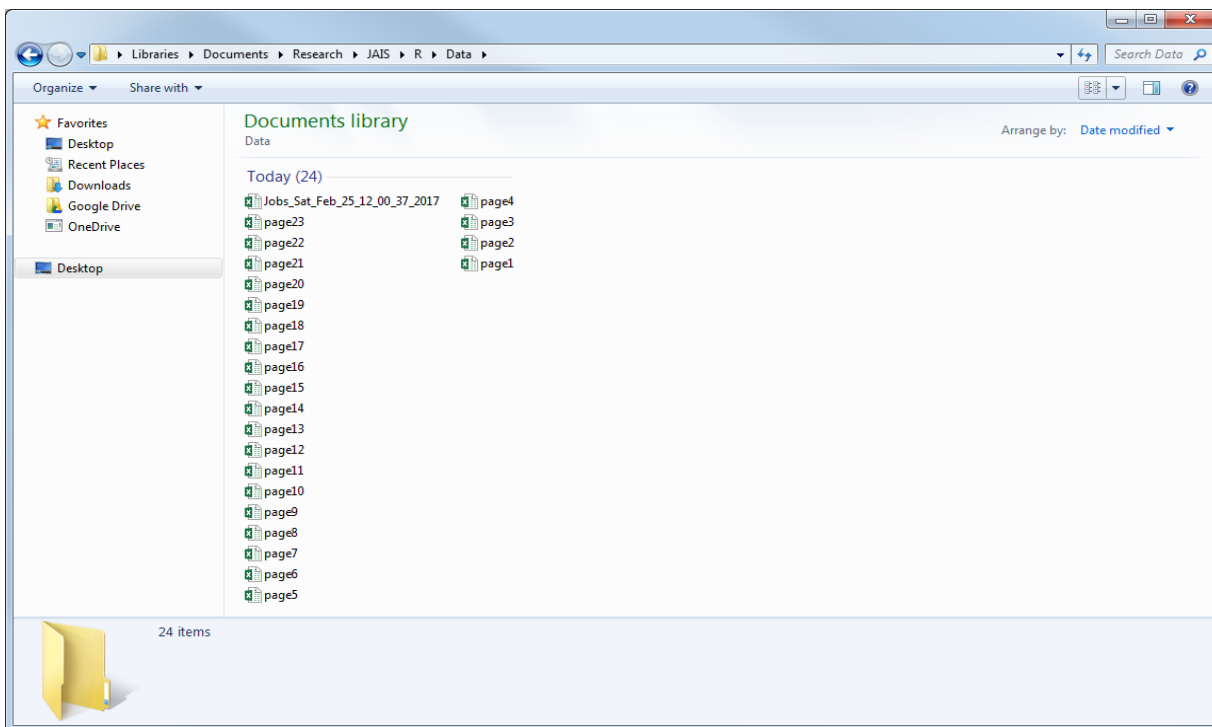
Figure 3B. File Structure Created

collectively duplicate the information saved in the main Excel sheet. The duplication is done to avoid complete data loss during the scaping process. Obtaining a dataset that contains thousands of observations may take a computer hours, or even days, of non-stop work. If script execution halts (e.g. Web server becomes unavailable, computer runs out of memory or freezes, there is a power outage in the building, etc.) researchers still have access to the data collected prior to the failure point. These data will be saved as Excel files up to the point when the script execution was halted. If script execution is successful, the entire dataset is saved in the main, time-stamped Excel file (e.g. Jobs_Sat_Feb_25_12_00_37_2017.xlsx). The file name means that the data were saved on Feb 25, 2017 at 12:00:37pm. The entire file structure produced as a result of running the Web Scraping task is provided in Figure 3B.

These files were stored on a password-protected computer belonging to one of the researchers. Thus, it was concluded that adequate measures were taken to protect what might constitute copyrighted data from unauthorized access. The researchers agreed that this data will not be shared with anyone outside of the research group.

What Happened Next

Some interesting developments related to the legality and ethics of the Web Scraping project described in this appendix occurred several months after the execution of the data collection task. One of the authors of the paper developed an R package that can automatically crawl and retrieve data from Dice.com based on supplied parameters. This was done in hope to stimulate further research in this area, save time of other researchers, and provide a valuable extension of the utilities available for the website. As far as the author was concerned, he was doing something valuable to the broader community using his time and expertise at no charge.

The package was eventually published on CRAN – a leading peer-reviewed repository of user-provided extensions of R language. The author of this package also wrote a manual on how to use this package and advertised availability of this package to a number of Web communities, including the ISWorld mailing list. The package started gaining popularity. Several researchers approached the author of the package for additional instructions and clarifications on how to use the package.

A few months after the release of the R package, the author of the package was informed by several users of the package that it was no longer working. The author of the package quickly determined that

there were no errors associated with R code of the package. It was the API supplied by Dice and upon which the package was based that was not responding. The author of the package wrote an email to Dice asking why the API was not working. He also encouraged users of the package to make similar inquiries with Dice. At first, Dice would send standard, "template" responses to each of these inquiries along the lines "Thank you for your inquiry, if you are interested in any of our data products please visit dice.com". Eventually, the author of the package was able to receive a more detailed response "from a human". The Dice representative stated that the API was no longer supported by the company. It was also added that the API was never meant to be used by the public and was rather intended for their own, internal developers; anyone interested in Dice data should use the main Web interface or subscribe to relevant data products.

At first, the author of the package wanted to modify it so that it could obtain data directly from the website, bypassing the APIs. But then the whole situation was interpreted by him as Dice imposing a restriction on crawling and scraping their data. Thus, he decided that it would be unethical and, perhaps, illegal for him to continue developing tools for scraping data from Dice. This was quite disappointing, as the author viewed himself as a volunteer who contributed his time and expertise at no charge for making the website more useful to the research community.

A more recent examination of the "robots.txt" file confirmed correctness of his interpretation of the situation. It looks like the website does not want automated web crawlers to programmatically access the job listings area. Perhaps, this was done to make sure that potential users of the website access job listings from the main "entryway", which is Dice.com. Because of that, the author decided to abandon this project. He still hopes that one day he can modify the R code that it works with some other website containing IT job listings in order to stimulate further research on in-demand IT skills and competencies.

Appendix C: JSON Output from Dice.com

The output below is generated using the following URL being sent to the server:

http://service.dice.com/api/rest/jobsearch/v1/simple.json?text=%22systems%20analyst%22

{"count":1105,"firstDocument":1,"lastDocument":50,"nextUrl":"/api/rest/jobsearch/v1/simple.json?areacode\u003d\u0026country\u003d\u0026state\u003d\u0026skill\u003d\u0026city\u003d\u0026text\u003d%22systems+analyst%22\u0026ip\u003d\u0026diceid\u003d\u0026page\u003d2","resultItemList":[{"detailUrl":"http://www.dice.com/job/result/RTX16c2f4/it-systems-analyst?src\u003d19","jobTitle":"IT Systems Analyst in Wichita, KS at CURO Financial Technologies Corp","company":"CURO Financial Technologies Corp","location":"Wichita, KS","date":"2017-02-25"},{"detailUrl":"http://www.dice.com/job/result/appblok/146_171500?src\u003d19","jobTitle":"Senior Business Systems Analyst","company":"Citizens Bank","location":"Cranston, RI","date":"2017-02-25"},{"detailUrl":"http://www.dice.com/job/result/appblok/146_171502?src\u003d19","jobTitle":"Senior Business Systems Analyst","company":"Citizens Bank","location":"Cranston, RI","date":"2017-02-25"},{"detailUrl":"http://www.dice.com/job/result/appblok/2151_5000178091506?src\u003d19","jobTitle":"Clinical Applications Systems Analyst - Ambulatory","company":"Connecticut Children\u0027s Medical Center","location":"Rocky Hill, CT","date":"2017-02-25"},{"detailUrl":"http://www.dice.com/job/result/appblok/146_171501?src\u003d19","jobTitle":"Senior Business Systems Analyst","company":"Citizens Bank","location":"Cranston, RI","date":"2017-02-25"},{"detailUrl":"http://www.dice.com/job/result/RTX1be463/4891134?src\u003d19","jobTitle":"Senior Systems Analyst","company":"Boise State University","location":"Boise, ID","date":"2017-02-23"},{"detailUrl":"http://www.dice.com/job/result/uhgbot/696317?src\u003d19","jobTitle":"Sr. IT Systems Analyst - Eden Prairie, MN","company":"UnitedHealth Group","location":"Eden Prairie, MN","date":"2017-02-24"},{"detailUrl":"http://www.dice.com/job/result/uhgbot/695890?src\u003d19","jobTitle":"Senior IT Systems Analyst - WellMed - San Antonio, TX","company":"UnitedHealth Group","location":"San Antonio, TX","date":"2017-02-24"},{"detailUrl":"http://www.dice.com/job/result/uhgbot/700495?src\u003d19","jobTitle":"Sr. Business Systems Analyst - WellMed - San Antonio, TX","company":"UnitedHealth Group","location":"San Antonio, TX","date":"2017-02-24"},{"detailUrl":"http://www.dice.com/job/result/uhgbot/700494?src\u003d19","jobTitle":"Estimating Systems Analyst - Phoenix, AZ or Eden Prairie, MN","company":"UnitedHealth Group","location":"Phoenix, AZ","date":"2017-02-24"},{"detailUrl":"http://www.dice.com/job/result/uhgbot/701950?src\u003d19","jobTitle":"Senior Business Systems Analyst - WellMed - San Antonio, TX","company":"UnitedHealth Group","location":"San Antonio, TX","date":"2017-02-24"},{"detailUrl":"http://www.dice.com/job/result/uhgbot/702044?src\u003d19","jobTitle":"Business / Systems Analyst - Basking Ridge, NJ","company":"UnitedHealth Group","location":"Basking Ridge, NJ","date":"2017-02-24"},{"detailUrl":"http://www.dice.com/job/result/10449280/844229?src\u003d19","jobTitle":"Technical Systems Analyst","company":"QuantRes Asset Management","location":"Nassau","date":"2017-02-24"},{"detailUrl":"http://www.dice.com/job/result/uhgbot/701249?src\u003d19","jobTitle":"IT Systems Analyst - Las Vegas, NV","company":"UnitedHealth Group","location":"Las Vegas, NV","date":"2017-02-24"},{"detailUrl":"http://www.dice.com/job/result/uhgbot/696804?src\u003d19","jobTitle":"IT Systems Analyst - Schaumburg, IL","company":"UnitedHealth Group","location":"Schaumburg, IL","date":"2017-02-24"},{"detailUrl":"http://www.dice.com/job/result/jobtblok/96871132?src\u003d19","jobTitle":"Financial Aid Systems Analyst","company":"Baylor University","location":"Waco, TX","date":"2017-02-17"},{"detailUrl":"http://www.dice.com/job/result/uhgbot/696554?src\u003d19","jobTitle":"Senior Business Systems Analyst - WellMed - San Antonio, TX","company":"UnitedHealth Group","location":"San Antonio, TX","date":"2017-02-24"},{"detailUrl":"http://www.dice.com/job/result/uhgbot/699843?src\u003d19","jobTitle":"Windows Systems Analyst - Warminster, PA","company":"UnitedHealth Group","location":"Warminster, PA","date":"2017-02-24"},{"detailUrl":"http://www.dice.com/job/result/uhgbot/692098?src\u003d19","jobTitle":"IT Systems Analyst - Eden Prairie, MN","company":"UnitedHealth Group","location":"Eden Prairie, MN","date":"2017-

02-24"},{"detailUrl":"http://www.dice.com/job/result/uhgbot/663371?src\u003d19","jobTitle":"IT Systems Analyst - Las Vegas, NV","company":"UnitedHealth Group","location":"Las Vegas, NV","date":"2017-02-24"},{"detailUrl":"http://www.dice.com/job/result/uhgbot/700855?src\u003d19","jobTitle":"IT Business / Systems Analyst - Rocky Hill, CT","company":"UnitedHealth Group","location":"Rocky Hill, CT","date":"2017-02-24"},{"detailUrl":"http://www.dice.com/job/result/uhgbot/702997?src\u003d19","jobTitle":"Systems Analyst - Schaumburg, IL","company":"UnitedHealth Group","location":"Schaumburg, IL","date":"2017-02-24"},{"detailUrl":"http://www.dice.com/job/result/10127643A/9698?src\u003d19","jobTitle":"Clinical Systems Analyst II - Ambulatory","company":"Allscripts","location":"Whittier, CA","date":"2017-02-25"},{"detailUrl":"http://www.dice.com/job/result/RTX15c476/853002?src\u003d19","jobTitle":"Sr. Data Systems Analyst","company":"MagNet","location":"Lake Worth, FL","date":"2017-02-10"},{"detailUrl":"http://www.dice.com/job/result/10231917/4853133?src\u003d19","jobTitle":"Systems Analyst","company":"Great River Energy","location":"Elk River, MN","date":"2017-02-23"},{"detailUrl":"http://www.dice.com/job/result/RTL159177/830108?src\u003d19","jobTitle":"EMS Systems Analyst","company":"Associated Electric Cooperative, Inc.","location":"Springfield, MO","date":"2017-02-22"},{"detailUrl":"http://www.dice.com/job/result/10127643A/9677?src\u003d19","jobTitle":"Sr Clinical Systems Analyst","company":"Allscripts","location":"Whittier, CA","date":"2017-02-25"},{"detailUrl":"http://www.dice.com/job/result/10127643A/9657?src\u003d19","jobTitle":"MS Systems Analyst-Clinical Applications","company":"Allscripts","location":"Whittier, CA","date":"2017-02-25"},{"detailUrl":"http://www.dice.com/job/result/10436270/890183?src\u003d19","jobTitle":"Lead Business Systems Analyst","company":"DeWinter Group","location":"San Francisco, CA","date":"2017-02-23"},{"detailUrl":"http://www.dice.com/job/result/jobeblok/957965?src\u003d19","jobTitle":"Business Systems Analyst 3 (7584U) #22697","company":"University of California, Berkeley","location":"Berkeley, CA","date":"2017-02-03"},{"detailUrl":"http://www.dice.com/job/result/10116872/885080?src\u003d19","jobTitle":"Sr Systems Analyst - Financial Applications","company":"Vitamin Shoppe","location":"Secaucus, NJ","date":"2017-02-17"},{"detailUrl":"http://www.dice.com/job/result/10415537/309?src\u003d19","jobTitle":"Technical Business Systems Analyst","company":"Sila Solutions Group","location":"Shelton, CT","date":"2017-02-23"},{"detailUrl":"http://www.dice.com/job/result/10210023D/836527?src\u003d19","jobTitle":"Sr. Business Systems Analyst - Business Intelligence","company":"Crescent Solutions Inc","location":"Irvine, CA","date":"2017-02-06"},{"detailUrl":"http://www.dice.com/job/result/10519627/126207?src\u003d19","jobTitle":"Jr. Systems Analyst","company":"4M Research","location":"Huntsville, AL","date":"2017-01-26"},{"detailUrl":"http://www.dice.com/job/result/10428868/RT-103082?src\u003d19","jobTitle":"HR-Systems Analyst-Workday-Long Term-CTH-Pittsburgh-Immediate","company":"Accion Labs","location":"Pittsburgh, PA","date":"2017-02-24"},{"detailUrl":"http://www.dice.com/job/result/cxelisen/261194?src\u003d19","jobTitle":"Business Systems Analyst","company":"Eliassen Group","location":"Boston, MA","date":"2017-02-24"},{"detailUrl":"http://www.dice.com/job/result/10235494/880449?src\u003d19","jobTitle":"Sr. Systems Analyst","company":"Stellargy Services, LLC","location":"Reston, VA","date":"2017-02-15"},{"detailUrl":"http://www.dice.com/job/result/gatpa001/137896?src\u003d19","jobTitle":"Applications Systems Analyst/Programmer (Senior)","company":"Mastech","location":"Washington, DC","date":"2017-02-25"},{"detailUrl":"http://www.dice.com/job/result/10124935/SP944956511375582?src\u003d19","jobTitle":"Senior Business Systems Analyst","company":"BIAS Corporation","location":"Atlanta, GA","date":"2017-02-25"},{"detailUrl":"http://www.dice.com/job/result/10204742/JO170111043-869?src\u003d19","jobTitle":"Sr. Systems Analyst","company":"Strategic IT Staffing","location":"Reston, VA","date":"2017-02-25"},{"detailUrl":"http://www.dice.com/job/result/10124488/886260?src\u003d19","jobTitle":"Business Systems Analyst","company":"Synergy Computer Solutions","location":"Mettawa, IL","date":"2017-02-20"},{"detailUrl":"http://www.dice.com/job/result/accperny/ERPEPIC-BB?src\u003d19","jobTitle":"ERP Systems Analyst - Epicor","company":"Access Staffing","location":"Bay Shore, NY","date":"2017-02-16"},{"detailUrl":"http://www.dice.com/job/result/cxtcml/17-00324?src\u003d19","jobTitle":"HRIS Business Systems Analyst","company":"Computer Merchant, Ltd., The","location":"Mentor, OH","date":"2017-02-07"},{"detailUrl":"http://www.dice.com/job/result/10124935/CC31226521679157?src\u003d19","jobTitle":"BA/Systems Analyst","company":"BIAS Corporation","location":"Brentwood, TN","date":"2017-02-25"},{"detailUrl":"http://www.dice.com/job/result/10117123/890430?src\u003d19","jobTitle":"Data Analyst / Business Systems Analyst (Data Warehouse)","company":"Agile Global Solutions, Inc","location":"Denver,

CO","date":"2017-02-25"},{"detailUrl":"http://www.dice.com/job/result/10441189/27592?src\u003d19","jobTitle":"Business Systems Analyst III - Claims","company":"Fidelis Care New York","location":"Buffalo, NY","date":"2017-02-24"},{"detailUrl":"http://www.dice.com/job/result/10105937/696702?src\u003d19","jobTitle":"Business Systems Analyst","company":"DEEGIT, INC.","location":"Greenville, SC","date":"2017-02-22"},{"detailUrl":"http://www.dice.com/job/result/10124587/2203?src\u003d19","jobTitle":"Senior Systems Analyst","company":"xScion Solutions","location":"Washington, DC","date":"2017-02-24"},{"detailUrl":"http://www.dice.com/job/result/10113579/580597?src\u003d19","jobTitle":"Systems Analyst","company":"Matlen Silver Group","location":"Charlotte, NC","date":"2017-02-24"},{"detailUrl":"http://www.dice.com/job/result/cxelisen/260944?src\u003d19","jobTitle":"Senior Systems Analyst","company":"Eliassen Group","location":"Dulles, VA","date":"2017-02-24"}]}

About the Authors

Vlad Krotov is an Associate Professor and MSIS Program Director at the Department of Computer Science and Information Systems, Arthur J. Bauernfeind College of Business, Murray State University. He received his PhD in Management Information Systems from the Department of Decision and Information Sciences, University of Houston (USA). Prior to joining Murray State University, he served as an Associate Professor and Assistant Dean of Graduate Programs at Abu Dhabi University (UAE). His teaching, research, and consulting work is devoted to helping managers use data and technology for creating organizational value. His research has appeared in a number of academic and practitioner-oriented journals: *CIO Magazine*, *Journal of Theoretical and Applied E-Commerce*, *Communications of the Association of Information Systems*, *Business Horizons*, *Journal of Competitiveness Studies*, *Polish Journal of Management Studies*, *Journal of Cases in Educational Leadership*, *Journal of Emerging Technologies in Accounting*, *The Electronic Journal of Information Systems Evaluation*, *Journal of Education for Business*, *Polish Journal of Management Studies*, etc. His research was recognized at Murray State University by the "Outstanding Research Award" in 2016 the "Emerging Scholar Award" in 2017.

Leigh Redd Johnson, JD, is a Professor of Business Ethics and Law in the Accounting Department of Murray State University where she has taught since 2006. Prior to joining Murray State, Ms. Johnson was a corporate and securities associate at Womble Carlyle Sandridge & Rice, PLLC in its Research Triangle Park, North Carolina office. She has published in *Issues in Accounting Education, Journal of Accounting Education, Research on Professional Responsibility and Ethics in Accounting, The U.C. Davis Business Law Journal, The Journal of Corporate Accounting and Finance, The Journal of Corporate Taxation, The American Journal of Business Education, and Journal of Legal, Ethical and Regulatory Issues*, among others.

Leiser Silva has served as a faculty member at the Bauer College of Business at the University of Houston since 2002 and serves as the Associate Dean for Bauer's Graduate and Professional Programs. Silva holds a Ph.D. in Information Systems from the London School of Economics and Political Science, a M.Sc. in Systems Analysis and Design from the same institution, and a B.Sc. in Computer Sciences from the Universidad del Valle in Guatemala. His current research examines issues of power and politics in the adoption and implementation of information systems, particularly in the context of public organizations. Silva has made contributions to the body of literature in Information Systems that studies the relationship between strategy and Information Technology. His work has been published in leading academic journals such as *MIS Quarterly*, *Journal of the Association of Information Systems (JAIS)*, *Journal of Information Technology (JIT)*, *European Journal of Information Systems* and *Information Systems Journal*.