UNDERSTANDING ONLINE JOB ADS DATA

A TECHNICAL REPORT

APRIL 2014

ANTHONY P. CARNEVALE
TAMARA JAYASUNDERA
DMITRI REPNIKOV

GEORGETOWN UNIVERSITY
Center on Education and the Workforce
McCourt School of Public Policy

UNDERSTANDING ONLINE JOB ADS DATA

A TECHNICAL REPORT

UNDERSTANDING ONLINE JOB ADS DATA: A TECHNICAL REPORT

ABSTRACT

As the use of online job ads has proliferated beyond the simple job-search model, the research community is increasingly experimenting with the detailed breakdown of online job ads — referred to as online job ads data — to study labor market dynamics. Despite increased usage, there has been limited research assessing the usefulness of this data source. In this report, we shed light on the emergence of online job ads data and analyze their properties, particularly as they relate to traditional, survey-based sources. We estimate that between 60 and 70 percent of job openings are now posted on the Internet, but these job ads are biased toward industries and occupations that seek high-skilled, white-collar workers. While useful in measuring labor demand and honing in on previously inaccessible variables, online job ads data come with limitations. Thus, we urge data users to exercise caution and utilize this tool in conjunction with traditional data sources.

ACKNOWLEDGMENTS

We would like to express our gratitude to our funders, the Bill & Melinda Gates Foundation, Lumina Foundation, and the Joyce Foundation, for their support of our research. We thank Burning Glass Technologies for providing the data for the report. We are grateful to our research analysts, Andrew Hanson and Artem Gulish, for their excellent research and writing support. Special thanks are due to Ban Cheah for imputing the missing education information in the data. Our thanks also go to our colleagues, Jeff Strohl, Nicole Smith and Stephen J. Rose, and to John Dorrer, the external reviewer, for comments on an earlier version. We would also like to thank Tracy Thompson, Nancy Lewis and Jim McNeill, the report's editors; Ryan Clennan and his team at Studiografik, the report's designers; and everyone at ALLIEDmedia, the report's printer.

The views expressed in this publication are those of the authors and do not necessarily represent those of Burning Glass Technologies or our funders, the Bill & Melinda Gates Foundation, Lumina Foundation, or the Joyce Foundation, their officers, or employees.

Table of Contents

Introduction

The amount of time Americans spend online has grown sixfold over the past two decades. Today, more than 85 percent of American adults are online, up from 14 percent in 1995.[1] We rely on the Internet for our day-to-day needs, from personal communications and news to shopping, banking, applying for jobs, and entertainment.[2] This growth has also revolutionized the way online data are tracked, stored, and analyzed. As a result, massive new digital data systems are being used in sectors ranging from business and finance to science and research.

These trends have dramatically changed the employer-employee job matching process. Despite recent high unemployment levels, one of the major problems that U.S. employers face is the difficulty of finding workers with the needed skill set to fill their vacancies. The asymmetry of information about the requirements of the buyer (the employer) and quality (skill set) of the supplier (the job seeker) results in inefficient matches that have been costly for both parties. When job ads moved online in the mid-1990s, the costs of advertising plunged compared to newspaper advertising. Traditional geographic boundaries became irrelevant for the job search, as did the space constraints necessitated by the high cost of traditional classified ads, enabling employers to provide detailed information about the company and the position. Applicants' response time declined significantly, lowering transition times between jobs. Overall, online labor markets have the potential to increase efficiency of job matching, boosting employee job satisfaction and increasing worker productivity.[3]

More recently, the job opening history recorded on the web has begun to morph into something much more multi-dimensional. In the aggregate, it is part of a big data revolution that holds much promise for labor market research in its ability to fill gaps in government survey-collected data. More importantly, with the failure of numerous efforts to expand the Bureau of Labor Statistics (BLS) Job Openings and Labor Turnover Survey (JOLTS) to include more detailed data, alternative sources like online job ads data are gaining influence within labor market and education circles. This report explores the promise and current issues inherent in these trends.

Job seekers, employers, students, researchers, policymakers, higher education institutions, career advisors, and curriculum developers now view online job ads data as a practical source to explore the nature of today's dynamic labor market. Compared to point-in-time snapshots provided by survey-based labor market data, which rely on random sampling, these data are cost-effective and provide the ability to improve the accuracy of labor market forecasts while

1. Much of the increase in the expansion of Internet access happened between 1995 and 2005, rising from 14 percent to 72 percent, according to data from the Pew Internet and American Life Project. Zickuhr, Kathryn, *Who's Not Online and Why*, Pew Research Center, 2013 http://pewinternet.org/Reports/2013/Non-internet-users.aspx.
2. U.S. Department of Commerce, *Exploring the Digital Nation: America's Emerging Online Experience*. Washington, D.C.: U.S. Department of Commerce, 2012, 17, http://www.ntia.doc.gov/files/ntia/publications/exploring_the_digital_nation_-_americas_emerging_online_experience.pdf.
3. This report explores only one aspect of the online labor market — the shift of the talent search process to the Internet as a result of job ads being posted online. Employer-initiated employee searches based on resume data and the growth of telecommuting is not explored in this report. With regards to the effect of the Internet on labor market outcomes, only a few studies exist to date and they report mixed outcomes. However, some of the more recent empirical investigations found positive outcomes: Kuhn and Mansour (2011) found Internet job searches reduce unemployment durations by 25 percent; Bagues and Labini (2007) using a quasi-experimental approach found the Internet reduces the individual unemployment probability and improves match quality. On the other hand, Kroft and Pope (2010) found that the rapid expansion of Craigslist between 2005 and 2007 had no effect on local unemployment rates and Kuhn and Skuterud (2004) found that the Internet had no effect or had a negative effect on unemployment duration.

producing supplemental estimates of demand within detailed occupations, industries, and geographies. It can show the relative demand for different types of skills and levels of education. The real-time nature of job ads data also allows for the early detection of labor demand trends, which gives job seekers, employers, and policymakers a forward-looking analytical tool. Real-time labor market indicators can be particularly useful in aligning education and training curricula with workforce needs in emerging or rapidly changing industries, such as healthcare and information technology.

Online job ads data show great promise, especially in combination with other educational and labor market data. In its current state, however, it has several limitations. The data are subject to systematic errors introduced by how employers utilize the Internet for their talent search, the vendor data collection processes, and the effectiveness of the artificial intelligence used to collect and piece out the information from the ads. If left untreated, systematic errors can undermine the predictive power of the data and skew public policy decisions.

Another limitation is that, although there are analyses that examine the role of online job ads, a well-defined relationship between online job ads and traditional employment data has not been established.[4] According to our back-of-the-envelope calculation, discussed in more detail later in the report, between 60 and 70 percent of job openings are currently posted online, the majority for high-skilled white-collar occupations that require at least a Bachelor's degree. There are differences in coverage from one vendor to another based on their approach used to collect online job ads.[5] Universal coverage of job openings, however, remains elusive even at this day and age of Internet use, since not all job openings are posted online. We estimate that 80 to 90 percent of openings that require at least a Bachelor's degree get posted online. By contrast, just 30 to 40 percent of openings for candidates with some college or an Associate's degree, and only 40 to 60 percent of openings for high school diploma holders appear online. It is critical for job seekers, researchers, and decision makers, then, to understand better the strengths and limitations of this emerging tool before relying on its predictive power. For example, job seekers with some college or an Associate's degree who restrict their job search efforts to online sources will see only a fraction of the available employment prospects.

Burning Glass Technologies (BGT) is one of the leading vendors of online job ads data. BGT is at the forefront of improving this quickly evolving data source; BGT browses more than 15,000 job-related websites.[6] While our analyses are based on BGT data, some of the limitations that we outline in this report have external validity and may apply to other data providers, such as Monster, CareerBuilder, and Wanted Analytics.[7] But because we have not explored competing data sources to the same extent, we are not able to discuss the limitations in other sources or make comparisons between sources.[8] We suspect that many of the concerns addressed in this report will fade over time as the country achieves universal Internet access and employers increasingly use the Internet to fill job vacancies.

4. There is some research that explores the trends in employment, job openings, and job ads series and their lags, yet the trends don't show the strong consistency needed to establish a reliable relationship between series and requires further research. See page 9 of this report and Upjohn Institute (http://www.upjohn.org/node/678).
5. Help Wanted Online (HWOL) has 28 percent more job ads than BLS's JOLTS data, the official data source of job openings. However, we have not had the opportunity to analyze HWOL data.
6. We are grateful to BGT for its transparency and willingness to allow us to examine its data. Few vendors have been so open and responsive about key issues such as field consistency and reliability, de-duplicating ads, and geographic accuracy.
7. Using online job ads data from CareerBuilder Inc., Wright (2012) reports similar concerns.
8. For example, HWOL's data series includes seasonal adjustment to its ads data and this might make the series less volatile than it would otherwise be.

There are more than 70 data fields in one online job ad.

Online job ads are collected via a web crawling technique that sends out spiders[9] to browse online job boards before systematically converting, or parsing, each job ad into usable data elements.[10] However, job ads only represent a subset of job openings, since not all openings are posted online. They are not job openings in the same way that real estate listings are not sales. Just as real estate listings do not directly translate into sales figures, the distribution of occupations and industries in the online job ads data does not perfectly reflect the distribution of job openings by occupation and industry across the real economy. Yet, just as real estate sales and listings taken together tell us something about the state of the real estate market, looking at both job openings and job ads can tell us more about the health of the labor market.

To create the data used for this study, BGT spiders online job boards, employer sites, online newspapers, etc., for job ads before saving the ads in a searchable database. These data are referred to as "job postings data," "real-time data," or "real-time labor market information" (LMI), as they provide information as soon as employers post the ads. Just as search engines are able to effectively filter a flood of disorganized content online, BGT uses algorithms to best eliminate duplicate ads from the data. As seen in Figure 1, a single job ad comprises the employer name, job title, salary, education requirements, certifications, and skills, among some 70 other data elements. Each variable is stored in a separate column in a spreadsheet — in essence, deconstructing a job ad into variables that can be analyzed.

The following description provides additional details about the crucial elements — spidering, parsing, and de-duplication — of the data collection process.

The "spidering" process: Vendors employ Internet bots (also referred to as spiders) to crawl across the web and collect job ad information. This process typically follows a fixed schedule, spidering a pre-determined basket of websites. While there is not necessarily an optimal frequency, routine spidering can make the data more volatile and susceptible to artificial spikes. For example, if spiders only collect data biweekly, an artificial spike in job ads will appear every two weeks. Similarly, it's no longer a random process if all human resource departments uniformly announce job openings every first Monday of the month. Further, a non-random spike can occur if a small website starts to copy job ads from a large job board. To avoid large fluctuations and the loss of job ads, large job boards are given more weight than individual employer sites, which are updated less frequently. Nevertheless, the basket of sites is carefully monitored and updated to ensure the most current and complete stream of online job ads.

Extraction and parsing: Once the data are located, they are extracted, parsed, and coded into specific data elements. Some vendors use systems that require words to be in somewhat rigid and predetermined sequences or particular formats, such as lexical systems. This can create inadvertent errors. For example, if a vendor's library contains "Bachelor of

9. "Spiders" are computer programs used to search for and collect information from the Internet. The word "spidering" is used to describe this process. See glossary for more information.
10. See the Glossary on page 19 for a list of online job ads data terminology and their definitions.

FIGURE 1.

Job ads provide informative data elements such as employer, industry, occupation, salary, and education and skill requirements.

Requisition Number:	FS86446
Interest Category:	Business Operations/Admin/IT
Interest Sub Category:	Administration
Job Title :	Senior Logistics Technician ← Job title, SOC code
Employment Category/ Status:	Full-time
Type of Position:	Regular Hire
Country:	U.S.
State:	Indiana ← Location
City:	Linton
Minimum Requirements:	• Bachelor's degree from a four-year college or university; or one to two years related experience and/or training; or equivalent combination of education or experience. • Must have computer skills, database knowledge. • Individual must be able to read, analyze, and interpret general business periodicals, professional journals, technical procedures, or governmental regulations.
	...
Job Description:	XXX Corporation is looking for a Senior Logistics Technician to join our team in Linton, Indiana.
	Responsibilities:
	• Reviews requisitions and negotiates within budgetary limitations and scope of authority. • Obtains material from supplier at the lowest cost consistent with considerations of quality, reliability of source, and urgency of need. • Confers with vendors to obtain product or service information such as price, availability, and delivery schedule.
	...
	XXX Corporation is a leading provider of engineering, construction, and technical services for public agencies and private sector companies around the world. The Company offers a full range of program management; planning, design and engineering; systems engineering and technical assistance; construction and construction management; operations and maintenance; information technology; and decommissioning and closure services. XXX Corporation provides services for power, infrastructure, industrial, oil and gas, and federal projects and programs. Headquartered in San Francisco, XXX Corporation has more than 57,000 employees in a network of offices in nearly 50 countries.

Education, experience, skills

Additional skills, qualifications

Employer name & industry

Science" but omits "Bachelor of Nursing", educational requirements in job ads that seek the latter will be miscoded. Other vendors, like BGT, rely on a more flexible model to sort through and process each ad. It focuses on the context and sentence structure to determine the form, subject, and meaning of each job ad. This contextual approach is largely dependent on a continuously expanding taxonomy of key words and variables, which serves as the bridge that translates job ads into the coded data elements. As a result, it is able to correct for confounding factors, such as the distinction between working for a university and attending one.

When context in a job ad is unclear or missing, BGT augments the coding process via text mining and semantic analyses, both of which contribute to the final interpretation. These techniques assume specific values for omitted variables based on preexisting information in the job ad. That is, industry and occupational information is often implied but not explicitly stated in job advertisements. Keywords like "assembly line" suggest that the job ad is likely in manufacturing, whereas "discharge planning" or "patient care" refers to the healthcare sector. In such cases, text-mining methods can correctly infer values for missing data elements. Similarly, semantic analysis studies word patterns, which by inference can also enrich the dataset. This technique has been successful due to the processing of millions of job ads, from which BGT has effectively trained its artificial intelligence to learn what to look for in a job ad, rather than focusing on information in specific formats or sequences.

De-duplication: The common practice of posting the same job ad on multiple websites, unless accounted for, can result in significant over-counting of job ads. The process of identifying and removing duplicate job ads is known as "de-duplication." Given the excess scraping,[11] duplication, and at times a lack of well-defined standards on the Internet, vendors face a growing challenge to separate original job ads from all the reproduced duplicates. This can occur on any job board that allows third-party recruiters, but is particularly prevalent on sites that are more reliant on scraped data. BGT uses algorithms to identify a series of identically parsed variables in job ads, such as location, employer, and job title. Time frames are important, too. Based on industry research, BGT employs a 60-day window for job ads, meaning that after 60 days, the same job ad would be considered new.[12] According to BGT, focusing on the content of the ad rather than its basic fields removes more than half of all ads collected by spiders as duplicates.

Job seekers, employers, state and local workforce agencies, and community colleges value online job ads data.

Job seekers. The primary use of online jobs ads traditionally has been to connect job seekers with available openings. In a market of heterogeneous jobs and job seekers that has imperfect and asymmetric information, matching the right job to the right candidate is very difficult. Online job ads have increased the number of openings that a job seeker can browse through; at the same time, job seekers are able to narrow the number based on

11. Scraping refers to the industry practice of copying identical job ads or information from one job board or website to another.
12. Also, 60 days is widely cited as the average time it takes to fill a vacancy. For example: http://online.wsj.com/article/BT-CO-20110517-711948.html

specific search criteria. Unlike newspaper ads, which have limited job and employer descriptions because of space constraints, online job ads have detailed information, making it easier for job seekers to find jobs that match their skills and interests. The ability to apply for those jobs online has considerably reduced the wait time for both parties. A better match means increased employee retention rates, reducing employers' hiring costs. At the same time, since employers are usually willing to pay more for the most attractive candidate, better matching means higher wages for the worker, which sets the stage for greater productivity and company profits. Further, the increased information in the ads has made it easier for the job seekers to identify the set of skills they need to possess to find and/or move ahead in their careers.

Employers. Employers are major users of online job ads data. They are using online job ads data to determine what to include in job ads — such as phrases, skills, education requirements, and wages — to make ads more searchable and to attract desirable candidates. Employers are also faced with questions of where and when to place their job ads to get the biggest bang for the buck,[13] since there are thousands of places where an employer can advertise a job. To keep the cost of advertising down and increase the quality and size of the pool of applicants, employers can use targeted sites based on the occupation and location of the advertised job.

State and local workforce agencies. As the majority of employers have turned to the Internet for their staffing needs, more researchers have begun monitoring aggregate job ad trends. Almost all statewide workforce agencies and several local workforce agencies around the country currently use job ads data to match the unemployed with available openings more efficiently.[14] Because this tool can quickly hone in on emerging occupations, new education requirements, and industry-specific skills that are in demand, it complements labor market information (LMI) driven by standard occupational and industry classifications.

Community colleges. Community colleges have jumped on the online job ads data bandwagon[15] in their efforts to stay current on training and curricula needs and to guide the job searches of their students.[16] The detailed information available in the online job ads is being used by planners and curriculum developers at higher education institutions to identify the knowledge and skills that are in demand in the labor market and to tailor curricula and determine seat capacity. These data also provide career advisors at higher education institutions with a tool to help students land quality jobs faster.

Unlike the static snapshot of the labor market provided by traditional survey-based data, the wealth of information in online job ads data provides a detailed illustration of a more dynamic labor market. As an indicator of future employment trends, online job ads data use has quickly spread from public institutions, such as workforce development agencies and community colleges, to much wider audiences, as seen in recent analyses in *The Wall Street Journal*[17] and *The New York Times*.[18] The Conference Board's *Help Wanted Online*, which aims to provide a

13. This requires an additional set of data that captures the performance of the ad, such as the number of times the ad appeared in searches, number of views, and the number of times applied. A few vendors of online job ads data even gather and share the less available outcome data for each job ad - the interviews and hire information.
14. For example, 67 percent of New York unemployment insurance claimants report receiving job matches they had not found elsewhere.
15. Wright (2012) finds that more than 200 community colleges are using online job ads data.
16. See "Credentials That Work" project by Jobs For The Future for institutions that are using job ads data at http://www.jff.org/projects/current/workforce/credentials-work/1222.

current analytical picture of labor supply and demand, is one of the more established real-time tools in the market.[19] Researchers are also using job ads data to address much deeper labor market questions, such as the skill mismatch in the post-recession labor market.[20]

However, there are several limitations in using these data for research and policy making. First, it is important to distinguish between job ads data and other labor market data. Second, job ads data overrepresent high-skilled, high-wage job openings and underrepresent less-skilled, low-wage job openings. It is important to understand these limitations, which are discussed in more detail in the following sections, when making use of the data for research and policy purposes.

Job ads provide an incomplete picture of labor demand.

As mentioned above, job ads are different from job openings. Job ads data show trends, but provide an incomplete picture of current labor demand. Some employers still continue to use traditional avenues of advertising such as newspapers, career fairs, word of mouth, "help wanted" notices outside store windows, and, more recently, via employees' social networks. While openings advertised on social network sites such as LinkedIn or Facebook are being captured, openings on other sources are more difficult to capture and quantify. On the other hand, sampled data such as JOLTS data cannot provide in a timely manner the detailed employment demand (by industry, occupation, or state and sub-state level) available in online job ads data. Thus, the two sources (job ads data and traditional data sources) complement one another.

Online job ads data strongly correlate with job openings data.

Given their relatively high frequency, large volume of observations, and real-time nature, online job ads are a useful source for tracking employment demand over time and for detecting changes in demand earlier than using traditional methods. Figure 2 illustrates the relationship between online job ads, JOLTS job openings, and new hires going back to 2006. While job ads data are inherently more volatile, the two series followed a similar trend until the end of 2011.[21] We found that lagging the JOLTS data improves the alignment between the two series, despite a growing divergence in recent months. Templin and Hirsch (2013) arrive at stronger conclusions using job ads data from a different vendor and new hires data from Quarterly Workforce Indicators (QWI) of the U.S. Census Bureau.

17. http://blogs.wsj.com/atwork/2013/08/30/how-liberal-arts-majors-can-get-hired/.
18. We argue later in the report that the use of real-time data for future employment trend predictions can be flawed due to biases in its coverage.
19. See Dorrer and Milfort (2012) for a more comprehensive list of vendors.
20. Sahin, Song, Topa, and Violante (2012): Lazear and Splitzer (2012); Faberman and Mazumder (2012); and Rothwell (2012). A detailed discussion of the current and potential uses of job ads data can be found in the recent report by Jobs for the Future, "Using Real-time Labor Market Information on a National Scale" (2013). See Reamer (2013).
21. During the second half of 2009, when BGT took over the real-time data collection process from EmployOn, there were a few data glitches that led to two months of low ads numbers. However, adjustments can be made to correct for such outliers in long-term series.

More specifically, they show that online job ads served as a reliable leading indicator of ensuing hiring trends. However, despite correlation coefficients of nearly .75, we have been unable to establish a direct causal link between the job ads and job openings, hiring, or employment change.

There are key differences between job ads data and JOLTS data. Online job ads data offer insight into a job market that is strictly bounded by the vacancies posted online, which can produce biases. JOLTS data offer a snapshot of labor demand at a point in time of the reference month, the last business day of the month, derived from a stratified survey of 16,400 establishments, whereas job ads data capture the online ads of the entire reference month (the flow dynamics). Further, JOLTS excludes openings with a start date more than 30 days into the future, as well as openings for independent contractors, while job ads data, after correcting for duplicate ads and other noise, attempt to gather only new job ads. In the context of BGT, this is defined by a 60-day posting window (among other things), regardless of the party posting the opening. Given the growing presence of recruiters and headhunters in the job market, this periodicity difference alone could lead to a substantial divergence between the two sources.

The concerns for job ads data that lie ahead are consistency and volatility.

Trend analyses require consistency in the form of stable data, which can be undermined given the current spidering techniques programmed to capture as many ads as possible. As spidering gets more sophisticated, increases in the number of job openings posted online may reflect the power of the technology as well as actual changes in the labor market. To increase the data's stability and predictive power, it is best to stick with a fixed basket of major job boards like Monster.com and CareerBuilder.com for tracking changes in employment demand.

The major limitation for longitudinal analysis is the high volatility of the data. This is largely a result of month-to-month variations in the number of source websites visited and data-imposed volatility in the number of ads collected from those sources. These irregularities in the data collection process make it difficult to devise a tool to correct for systemic data anomalies. As a result, it is hard to discern between changes in employment demand and fluctuations in the job ads.

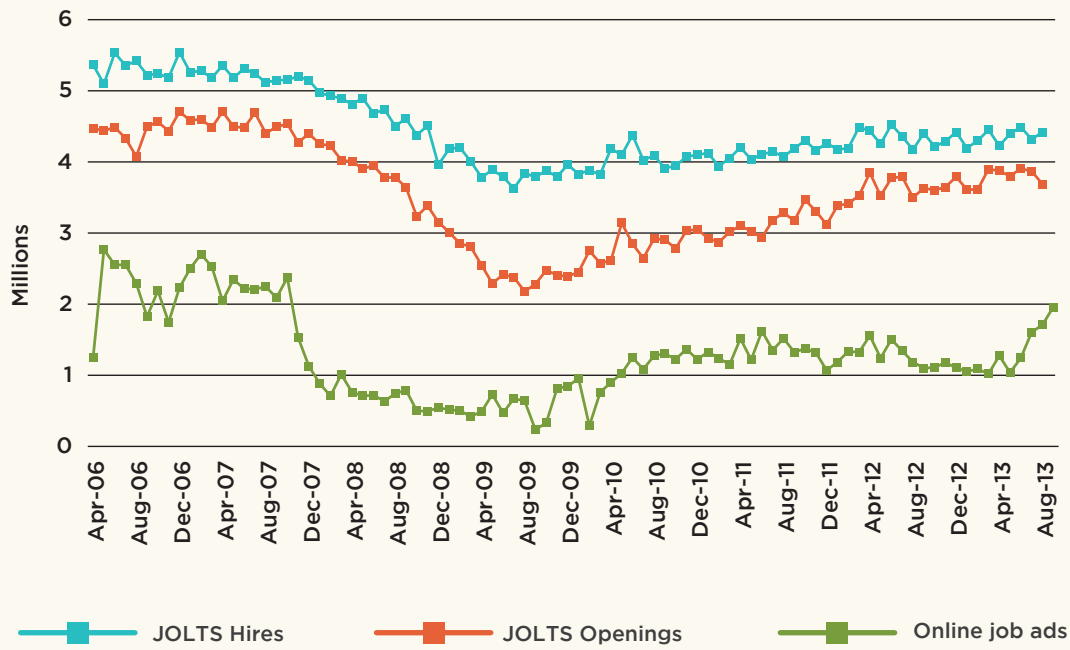
A continually expanding and diversifying database is beneficial for job seekers. For longitudinal analysis, however, it "undermines representativeness and consistency of the data over time."[22] Furthermore, observations reveal that 91 percent of the job ads collected since 2010 are from just 501 sources. Alternatively, the top 1,000 sources provided nearly 95 percent of the ads, making the contribution of the remaining 14,000 sources to the database minimal.

But since the vendor targets two different audiences (data analysts and job seekers), we believe it is beneficial to have two different data collection processes in place: one that spiders the top 500 or 1,000 sources in consistent intervals to achieve a stable series for longitudinal analysis, and another that spiders as many sources as possible to ensure the highest coverage. This second process can be used to track changes to the

22. Northeast Research Consortium, "Skills Analysis of Real-time Data" (2011).

FIGURE 2.

Online job ads data correlate closely with BLS's JOLTS data.



Source: Georgetown University Center on Education and the Workforce analysis of Burning Glass online job ads data and JOLTS job openings and hires (seasonally adjusted), 2006-2013.

top 500 or 1,000 sources over time and to make any updates to the list on an annual or biannual basis to ensure that the series maintains the same coverage. To further minimize the effect of improved spiders on the month-to-month changes in the number of job ads, preference should be given to the big job boards when de-duplicating.

Making the series stable requires careful monitoring of each individual job board for spikes and other aberrations. Such aberrations, when detected, need to be corrected for on a case-by-case basis. This not only requires powerful computers that can handle statistical analyses on large data, but also requires individuals with strong analytical and programming skills. Though maintaining a stable big data series can be expensive and time-consuming, extensive data editing is needed to ensure that the series is stable and reliable. These measures are not unique to big data; similar steps are taken even on survey data to ensure robustness.

Another limitation of the job ads data is the lack of historical data, especially compared to traditional labor market data, which have been collected for decades with the same set of survey questions to maintain consistency. This is a limitation that time will overcome. While we patiently collect the job ads data, we should also begin to direct our attention to achieving long-term consistency and reliability needed for longitudinal analysis.

Online job ads data overrepresent job openings for college graduates.

Despite the large and growing number of observations, job ads data do not represent the universe of job openings. Employers post openings online to attract those who are more likely to conduct their job searches online. These candidates also tend to be more educated relative to those who conduct their job searches offline.[23] Thus, openings for less-educated workers, like those in blue-collar construction jobs, are underrepresented in the data, which naturally leads to the undercoverage of less-skilled industries. Understanding this limitation, vendors have increased their focus on sites like Craigslist and job bank sites of public employment services. Craigslist consists of ads from the casual labor market, and, based on casual observations, public job banks tend to favor low-wage, low-skilled jobs. While states are working to change this, employer participation rates in public job banks remain low, especially among the smallest employers. Over- and underrepresentation is natural in government survey data, but carefully constructed survey data come with sampling weights or adjustment factors to correct the biases introduced through the data collection process. Establishing similar adjustment factors for job ads data has so far proven difficult.

To complicate matters further, not all openings are posted online and not all ads reflect an actual vacancy. There is anecdotal evidence that a small share of ads, especially in information technology occupations, is used to collect information about the potential applicant pool. For example, an employer may know that it will need three software developers with

23. Einav and Levin (2013) refer to them as "convenience samples."

specific skills for a project that will start in a year. Its human resources department may publish an online job ad for a software developer to gauge the available talent pool so that it can plan accordingly. On the other hand, one ad may also be used to fill multiple vacancies — an employer might intend to hire five sales representatives from one job advertisement. Further research is needed to determine the magnitude of these issues.

Given the detailed information on the employer-demanded skills and credentials, job ads data can be a valuable tool in guiding curriculum developers. However, under- and overrepresentation and other fluctuations in the ad counts can be misleading, especially when determining the seat capacity for a specific program. Hence, it is important to complement such decision making with established labor market data sources.

Between 60 and 70 percent of job openings are posted online.

With 1.6 million *new online job ads* in June 2013, the estimated *total active online job ads* in BGT data for the same month was around 2.7 million ads.[24] New job ads are the ads that did not appear online in the previous month. Total active job ads is the number of job ads that are active in a given month — that is, both new ads and ads that are still active from the previous month. Estimates indicate that two out of every five job ads active in any given month are from the previous month; the remaining three are openings that have come up since then. When we compare online job ads (2.7 million ads) to JOLTS data (3.9 million job openings) in June 2013, we estimate that roughly 70 percent of the total number of job openings were posted online. This is an upper-bound estimate compared to previous months. Generally, we estimate that total *active online job ads* captured each month to be roughly 60 to 70 percent of the total *job openings* in the labor market.[25]

More than 80 percent of jobs for those with Bachelor's degrees or better are posted online.

We believe that the main source of bias in the job ads data is due to differences in Internet access among job applicants, which varies by education level. Yet, as the only official source for job openings, the JOLTS data do not provide a breakdown by education, so quantifying the bias using JOLTS is not possible. Complicating matters further, only half of the online ads report an education requirement.[26] However, using a back-of-the-envelope approach, we estimate that 30 to 40 percent of job openings for workers with some college or an Associate's degree, and 40 to 60 percent of job openings for workers with a high school diploma, get posted online. By contrast, we estimate that 80 to 90 percent of openings for job candidates with at least a Bachelor's degree are posted online.

To investigate the education bias rooted in job ads data, we make use of the Current Population Survey (CPS) data to break down the educational composition of the current labor market and compare it to that of the ads data.[27] Even though online job ads to employment is not an accurate apples-to-apples comparison, it can provide further insight to the level of misrepresentation in the ads data. The level of education required in online

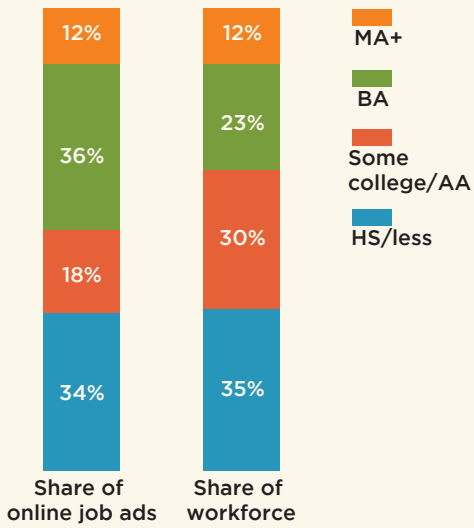
24. According to the Conference Board, about 60 percent of total active ads are new ads in its HWOL data. We assume that the ratio of new to active ads is similar in both sources.
25. As we discuss later, this number varies by education, industry, and occupational clusters.
26. Almost half of the job ads data do not report an education requirement. The job ads with a missing education requirement were imputed using a combination of the education distribution of the job ads data and the American Community Survey (2006-2010) education distribution of employment based on the occupation and the specific state the job opening was available. See Appendix for a discussion of our imputation method.
27. Lack of data does not allow for an apples-to-apples comparison, since JOLTS does not give a breakdown by education. The employment share in a particular occupation is not a direct measure of the level of turnover. Due to the lack of a better data source that provides the educational composition of the job openings, that of employment is used. By using our judgment and knowledge about the turnover and job creation rates in the local markets, these ratios are expected to suggest possible over- and underrepresentation in job ads data.

job ads significantly deviates from the educational attainment of the current labor force (Figure 3).[28] Over half of all online job ads are seeking workers with a Bachelor's degree or better. This has been the trend in the job ads data over the past few years that data have been available. Eighteen percent of all job ads are for those with some postsecondary education but less than a four-year college degree.

As seen in Figure 3, the underrepresentation is greatest for those with some college education or an Associate's degree: the share of employment with some college or an Associate's degree is greater than what is observed in job ads data. This could be due to several reasons. As stated above, jobs advertised online still fail to cover the universe of job openings. Among those advertised online, nearly half of the ads do not specify the level of education required.[29] And even when education requirements are present, it is not always a foolproof strategy, as job ads don't always explicitly read "some postsecondary education required." Instead, for example, an ad for a customer service representative can declare "a minimum of high school diploma or equivalent required," and it provides a list of skills that suggests more education or training is highly preferred. This is all the more reason to use these data in conjunction with established LMI sources.

FIGURE 3.

Job ads data overwhelmingly favor college graduates, showing limited opportunities for job seekers with some college or an Associate's degree.



Source: Georgetown University Center on Education and the Workforce analysis of Burning Glass online job ads data, 2013 Q2 and March Current Population Survey, 2013.
Note: Missing education requirements in job ads data were imputed using a combination of the education distribution of the ads data and the American Community Survey 2006-2010 education distribution. See appendix for more detail on the imputation method.

28. The QWI new hires data from the U.S. Census Bureau now provides the education breakdown of the new hires. Comparison of the online job ads data to the QWI new hires data suggests even greater overrepresentation of job openings for Bachelor's or better in online job ads.
29. Quality control tests indicated that when available in the ad, the parser fairly accurately identifies the education requirement (85 percent of the time).

Job ads overrepresent industries that demand high-skilled workers.

Another topic of interest to job ads data users is how well job ads match up to the employment demand in each industry.

Using JOLTS, we examine how job ads stack up against the distribution of JOLTS job openings by industry. This approximates the extent to which each industry is over- or underrepresented relative to the actual job market. Compared to the job openings distribution in JOLTS in Figure 4, education services, real estate and rental and leasing, and manufacturing are overrepresented in the job ads data. Professional and business services and transportation, warehousing, and utilities are among the most consistently represented industries, as the relative share of job ads is more closely aligned with that of the openings. Meanwhile, job ads in the government and construction industry are among the most underrepresented relative to actual job openings in these industries.

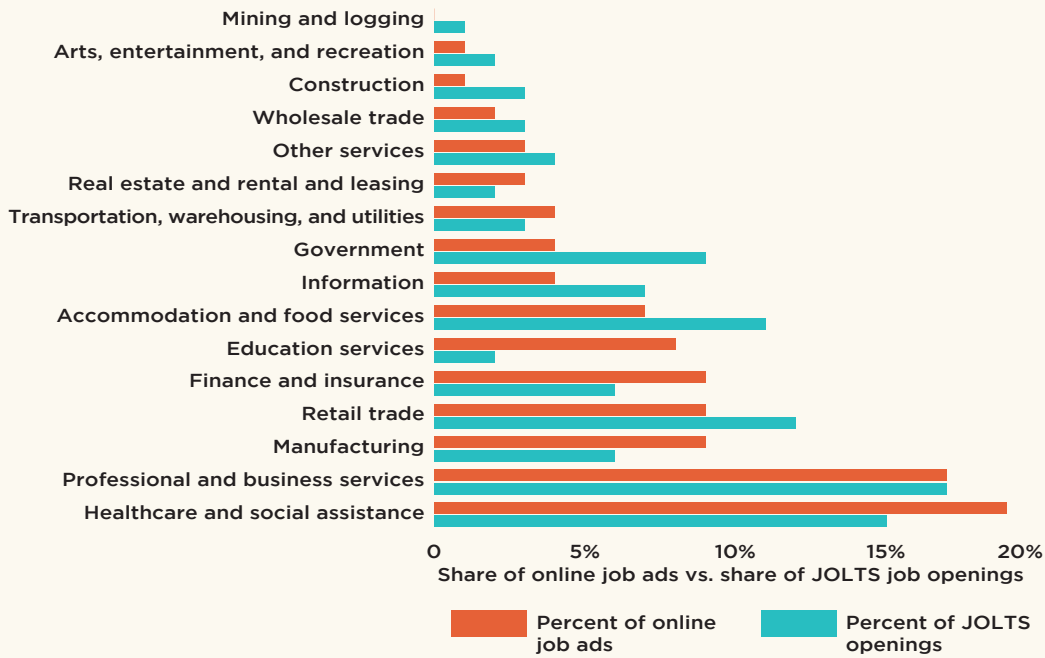
The over- and underrepresentation can occur for several reasons, but it is mostly a function of what is posted on the Internet. Recruitment methods, such as word of mouth and referrals, may still be dominant in select industries like construction and food services, which would explain their low online profile and the underrepresentation of online job ads relative to actual job openings. Conversely, the education, real estate, and manufacturing sectors have an outsized online presence relative to employment. It may be that employers in these sectors are keen on recruiting online since it greatly expands the talent pool, or it could indicate a need for a more qualified workforce. Manufacturing, for example, is a sector that is frequently cited for its skills shortage. Some of the differences may also be attributed to job churn.

Sometimes industry coding is based on the description of the employer rather than the employer name itself. However, a fifth of the job ads do not carry the employer name, so there is better accuracy in identifying the industry sectors (two-digit NAICS[30] codes) than industry sub sectors or industry groups (four-digit NAICS). The omission of employer names is common, since recruiters are known to conceal specific firms when soliciting applications to avoid applicants' bypassing the recruiter and contacting the employers directly.

30. North American Industry Classification System

FIGURE 4.

The distribution of job ads within the professional and business services industry is the most consistent with today's job market.



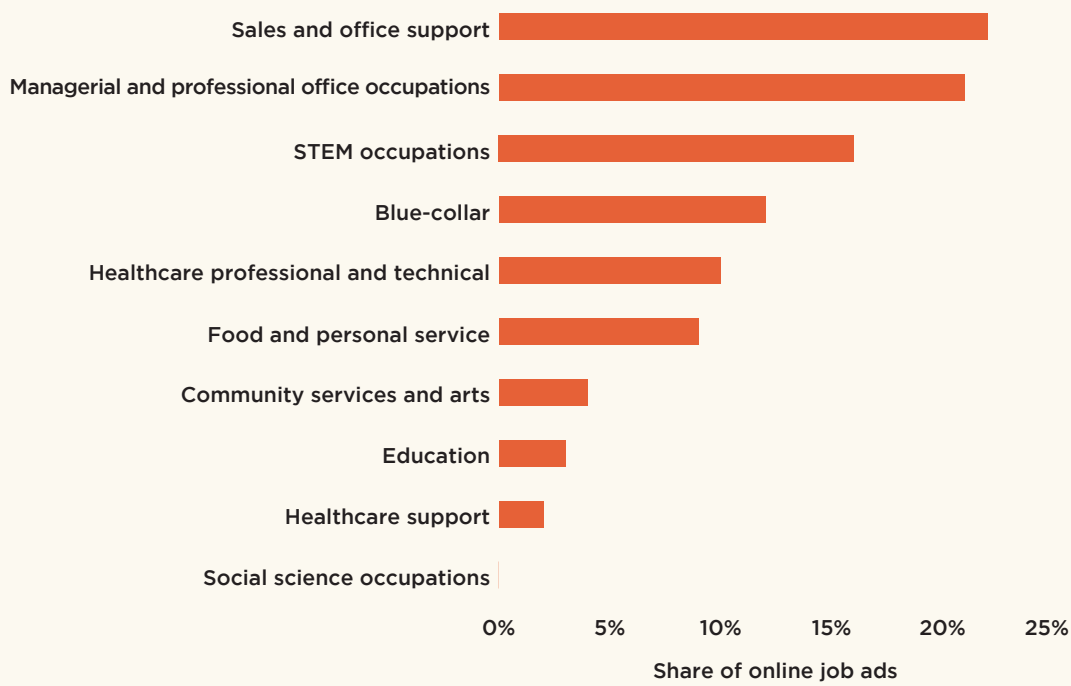
Source: Georgetown University Center on Education and the Workforce analysis of Burning Glass online job ads data, 2013 Q2 and JOLTS job openings (not seasonally adjusted), 2013.

White-collar office and STEM occupations account for the majority of job ads.

Given the close association between education, skills, and occupations, it is not surprising to find overrepresentation of high-skilled occupations in the job ads data (see Figure 5). White-collar office jobs like sales and office support, managerial and professional office, and STEM occupations make up the majority of ads.

FIGURE 5.

White-collar office jobs account for three out of five online job ads, 2013 Q2.



Source: Georgetown University Center on Education and the Workforce analysis of Burning Glass online job ads data, 2013 Q2.

The accuracy of extracted labor market information varies across data fields.

Government statistics are known for their lag, yet their validity is rarely questioned. Online ads arrive via a daily feed, but are also crafted for recruitment purposes and not necessarily for analysis, thereby often omitting critical information of interest to users. To understand the degree of accuracy of the parser, we carried out extensive quality testing of the parsed data from the job ads.[31] This involved carefully reading the full job text from a stratified random sample of ads and comparing it to the parsed information.

Despite the limitations imposed by how firms and recruiters post job openings online, our findings indicate that the parser shows considerable accuracy in populating the data fields. However, the level of accuracy varies from field to field. In fields like state, city, occupation title, major occupation group (two-digit occupation codes), and skills, we observed greater than 80 percent accuracy. The ability of the parser to produce detailed occupation coding declined with the level of detail — the accuracy rate of six-digit occupation coding was about 73 percent.

Since job ads almost always carry a job title, occupational coding in job ads data has greater accuracy than industry coding. As many recruiters and some employers conceal the employer name in the ad, it was harder to identify industries accurately. The parser was better at identifying major industries (two-digit NAICS codes) at 76 percent accuracy as the parser was relying on a brief description of the company when employer name is not available. Accuracy declined considerably in identifying detailed industries.

Only half of the ads carried education requirements. When available, the parser identified the education requirement with 85 percent accuracy.

There are thousands of sites where employers can post job ads. They can be mainly categorized into five sources: company websites, paid job boards, free job ad sites, recruiters, and others. Sites such as Monster.com and CareerBuilder.com are paid job boards and Craigslist and Snagajob are examples of free job ad sites. Our research found the lowest accuracy rate in free job ad sites and other sources. The parser mainly failed at identifying the employer name and the industry code of job ads from these sources.

While job ads data have recently made considerable improvements in collection methods and accuracy, there is still room for improvement. Aside from continuing to refine the use of artificial intelligence, greater standardization of how employers enter information and state job requirements would greatly improve the accuracy and statistical power of this tool.

31. The quality testing was carried out as part of our work for a Green Jobs Project funded under a DOL-ARRA grant. Quality control testing was carried out with considerable support from the seven states that were part of a Northeast Green Jobs Research Consortium. Each state reviewed a random sample of ads from their state.

Conclusion

Job ads data allow for the analysis of changing labor market conditions to be implemented with greater detail, efficiency, and, more importantly, in real time, without the delays and revisions of government statistics. The ability to hone in on variables that were previously inaccessible or prohibitively expensive to parse has led to a fundamental shift in both the production and use of labor market data. This data holds promise in reducing unemployment spells by increasing job-matching efficiency. It also has important implications for higher education and workforce training providers, by giving them the tools to structure curricula to help students and clients meet job ads needs of employers.

Yet despite its growing use and vast improvements in coverage and accuracy in recent years, job ads data should still be used to complement traditional LMI. The tool has great potential in the aggregate, as well as at state and sub-state levels, but its limitations and usage must be better understood. While useful as a leading proxy for labor demand, as argued in this report, several issues arise due to the systematic variation in the demanded level of educational attainment. Similarly, the online coverage by occupation and industry does not always represent the occupation and industry distributions in the actual job market. Currently, online job ads data may not meet the standards of LMI analysts or government statistics, but there is merit in carefully implementing this tool in conjunction with traditional data, particularly as it becomes more robust and viable. Used carefully, these data offer insight and suggest rich new areas for the development of employment and educational services.

References

Altstadt, David. *Aligning Community Colleges to Their Local Labor Markets*. Boston, Mass.: Jobs for the Future, 2011, http://www.jff.org/publications/workforce/aligning-community-colleges-their-local-/1303.

Auguste, Byron, Kihn Paul, and Matt Miller. *Closing the Talent Gap: Attracting and Retaining Top-third Graduates to Careers in Teaching*. Washington, D.C.: McKinsey & Company, 2010, http://mckinseyonsociety.com/downloads/reports/Education/Closing_the_talent_gap.pdf.

Bagues, Manuel F. and Mauro S. Labini. *Do On-line Labor Market Intermediaries Matter? The Impact of Almalaurea on the University-to-Work Transition*. NBER Working Paper No. 13621, 2007.

Dorrer, John and Myriam Milfort. *Vendor Product Review: A Consumer's Guide to Real-Time Labor Market Information*. Boston, Mass.: Jobs for the Future, 2012, http://www.jff.org/sites/default/files/VendorProductReview_041712.pdf.

Einav, Liran and Jonathan Levin. *Big Data and Economic Analysis*. NBER Working Paper No. 19035, 2013.

Faberman, Jason and Bhash Mazumder. *Is There a Skills Mismatch in the Labor Market?* Federal Reserve Bank of Chicago, Working Paper No. 300, 2012.

Hofschneider, Anita. "How Liberal Arts Majors Can Get Hired," *The Wall Street Journal*. Aug. 30, 2013. http:// blogs.wsj.com/atwork/2013/08/30/ how-liberal-arts-majors-can-get-hired/.

Klafter, Ben, "Best 50 Niche Job Boards," *SmartRecruiters.com* (blog), http://www.smartrecruiters.com/ blog/best-50-niche-job-boards/.

Kroft, Kory and Devin Pope. *Does Online Search Crowd Out Traditional Search and Improve Matching Efficiency? Evidence from Craigslist*. Working paper, 2010.

Kuhn, Peter J. and Hani Mansour. *Is Internet Job Search Still Ineffective?* Discussion Paper Series. Forschungsinstitut zur Zukunft der Arbeit, No. 5955. http://hdl. handle.net/10419/55102.

Kuhn, Peter J and Mikal Skuterud. "Internet Job Search and Unemployment Durations." *American Economic Review 94(1)* (March 2004): 218-232

Lazear, Edward and James Spletzer. *The United States Labor Market: Status Quo or A New Normal?* NBER Working Paper No. 18386, 2012.

Lohr, Steve. "More Data Can Mean Less Guessing About the Economy," *The New York Times*, Sept 7, 2013. http://www.nytimes.com/2013/09/08/business/more-data-can-mean-less-guessing-about-the-economy.html.

Northeast Research Consortium. *Skill Analysis of Real-time Data*, Unpublished paper prepared in fulfillment of the U.S. Department of Labor, Labor Market Information Improvement Grant, 2012. http://public.greencareercentral.org/Skills%20Analysis.pdf.

Reamer, Andrew, *Using Real-time Labor Market Information on a National Scale*, Boston, Mass.: Jobs for the Future, 2013, http://www.jff.org/publications/education/using-real-time-labor-market-information/1520.

Rothwell, Jonathan. *Education, Job Openings, and Unemployment in Metropolitan America.* Washington: Brooking Institution, 2012.

Sahin, Aysegul, J. Song, G. Topa, and G. Violante. *Mismatch Unemployment*. CEPR Discussion Paper No. DP9093, 2012.http://www.newyorkfed.org/research/economists/sahin/USmismatch.pdf.

Templin, Thomas and Lesley Hirsch. *Do Online Job Ads Predict Hiring?* New York: New York City Labor Market Information Services, 2013, http://www.gc.cuny.edu/CUNY_GC/media/CUNY-Graduate-Center/PDF/Centers/Center%20for%20Urban%20Research/LMIS/NYCLMIS-RESEARCH-BRIEF-Do-Online-Ads-Predict-Hiring.pdf.

Wright, Joshua. *Making a Key Distinction: Real-time LMI and Traditional Labor Market Data*, EMSI (Economic Modeling Specialists International) (blog). Feb. 12, 2012, http://www.economicmodeling.com/2012/02/07/making-a-key-distinction-real-time-lmi-traditional-labor-market-data/.

Glossary

Active job ads – The number of job ads that are active in a given month, composed of both brand new announcements as well as those still active from the previous month.

Canon – Method used to standardize the information in raw ads data. As a result, different entries in a variable field that represent the same entity are combined into one specific standardized entry. For instance, G.E. and General Electric represent the same company and would have the same Canon Employer title. This is done to improve the efficiency of the algorithm and is still a work in progress.

De-duplication – Process by which duplicate ad copies are removed from the dataset so only one copy of each unique ad is represented. Duplicates are caused by employers using multiple websites to post their job openings.

Entry-level Job – Job typically designated for recent high school or college graduates in a given discipline that does notnecessarily require experience in the field.

Job board – Website where employers post job announcements and job seekers deposit their resumes in hopes of a successful match.

Job ad – Actual advertisement for a potential job opportunity that appears online on a job board or a firm's career site. However, a job ad does not necessarily reflect an actual job opening in the market because various factors (testing the market, gathering a pool of candidates) make it difficult to distinguish the exact nature of an ad.

Job scraping – Industry technique that captures job openings posted on individual employer sites and reposts them across a wide range of larger job boards.

Labor exchange – Interactive websites designed to help job seekers and employers find industry and occupation information in their area.

New job ads – Job announcements that did not appear online in the previous month.

Parser – Artificial intelligence tool that analyzes a job ad's content and separates it into variable fields for the dataset. The use of artificial intelligence allows for continual improvement in the field accuracy through a learning process.

Real-time LMI – Labor market intelligence derived from the analysis of job ads and resumes placed into public and private labor exchanges. It is real time because it can be based on data pulled from the Internet on a daily basis. It is labor market intelligence because it can provide indications of supply and demand trends, emerging

occupations, current and emerging skill requirements, and market-based demand for education and certifications.[32]

Job ads data – Universe of online job ads that is used as an analytical tool to produce information on upcoming and past labor trends at multiple levels of detail. It offers potential benefits not offered by traditional Labor Market Information (LMI), but is inhibited by the inherent nature of the data.

Recruiter – Intermediary who facilitates the job match process by soliciting qualified candidates to fill openings.

Spider – Internet bot that systematically browses the web for specific web pages and copies the pages it visits for later processing. The term spider is similar to webcrawler because spiders crawl the web.

Spidering – Process of browsing the web for specific URLs and saving the web pages visited for later processing.

Appendix

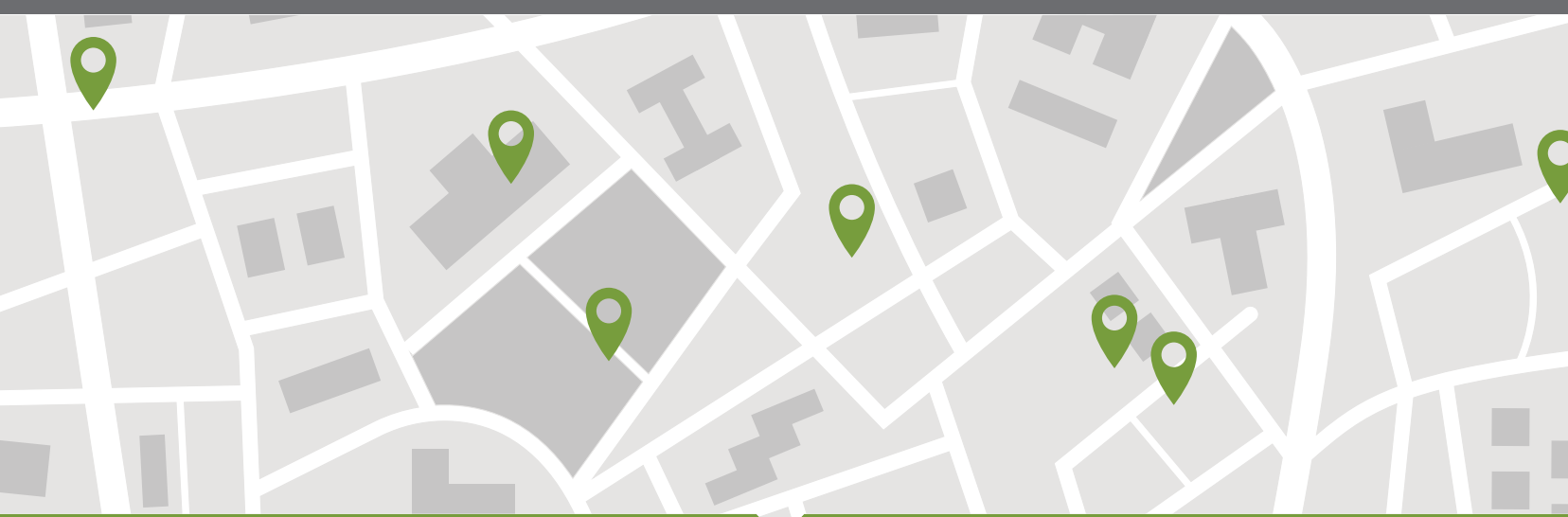
A Note on the Imputation Method

Since education requirements are only available for roughly half of the job ads data, we imputed education requirements for job ads with missing education values. This reduced a selection bias toward job ads that require a Bachelor's degree or better. The imputation approach uses a combination of hot-deck and cold-deck imputation methods. The imputation method is a cross-fertilization of the education distribution of current employment (as seen in the American Community Survey (ACS) 25 to 44 age cohort), and that of identified job ads. More specifically, suppose the education requirements are missing from 60 percent of the ads for a given occupation in a given state — we would keep the 40 percent of the ads with education information as they are. However, to avoid the 40 percent dictating the imputation process, one half of the missing values are imputed with the education distribution of the current employment in that state as given by ACS. Then, the second half is divided 60-40 — the percentage of ads with missing education requirements for that particular occupation in the given state. It follows that the 60 percent of the second half is imputed again through the ACS and the remainder is assigned through the job ads distribution. In this example, 80 percent of the weight is determined using the ACS distribution and the remaining 20 percent is made up from the job ads data.

32. As defined at the Brookings Institution LMI Forum, 2010.

Your career is loading...please wait

The Online College Labor Market comprises a full report, an executive summary, and a technical report. All reports can be accessed online at cew.georgetown.edu/onlinejobmarket.



GEORGETOWN UNIVERSITY

Center on Education and the Workforce

McCourt School of Public Policy

3300 Whitehaven Street NW,
Suite 5000
Washington, DC 20007
cew.georgetown.edu