APPENDIX A: OVERVIEW AND REPRESENTATIVENESS OF THE DATA

Throughout this report, we use data obtained from Burning Glass Technologies (Burning Glass). Burning Glass is a private labour market analytics company that uses web crawlers to scrape job posting level data from online job boards, recruiter websites, and business websites on the internet. The company has a database of more than one billion current and historical job postings worldwide. It is one of the most comprehensive sources of data on job openings. Burning Glass parses the raw text of job postings to extract key attributes, such as the occupational group the posting belongs to, the skill composition required, as well as qualifications and experience requested. The coding process was deemed to be more than 80 percent accurate by an independent audit conducted in 2016 for the US data.[32]

But what is a job posting? A job posting is a device used by employers to find potential employees; it lists the qualifications and skills an employer desires in a candidate, as well as information about the job being hired for and about the company doing the hiring.

Importantly, job postings are used strategically by employers to not only signal needs but to target specific talent that could be a good fit for the position. Employers, as a result, may use specialized or coded language in the job description that can be interpreted easily by those with the desired background.

Furthermore, though a job posting likely reflects what employers think they need in a candidate, this may not correspond to what is actually needed at the firm, or reflect the actual day-to-day tasks performed by the employee if that job is filled.

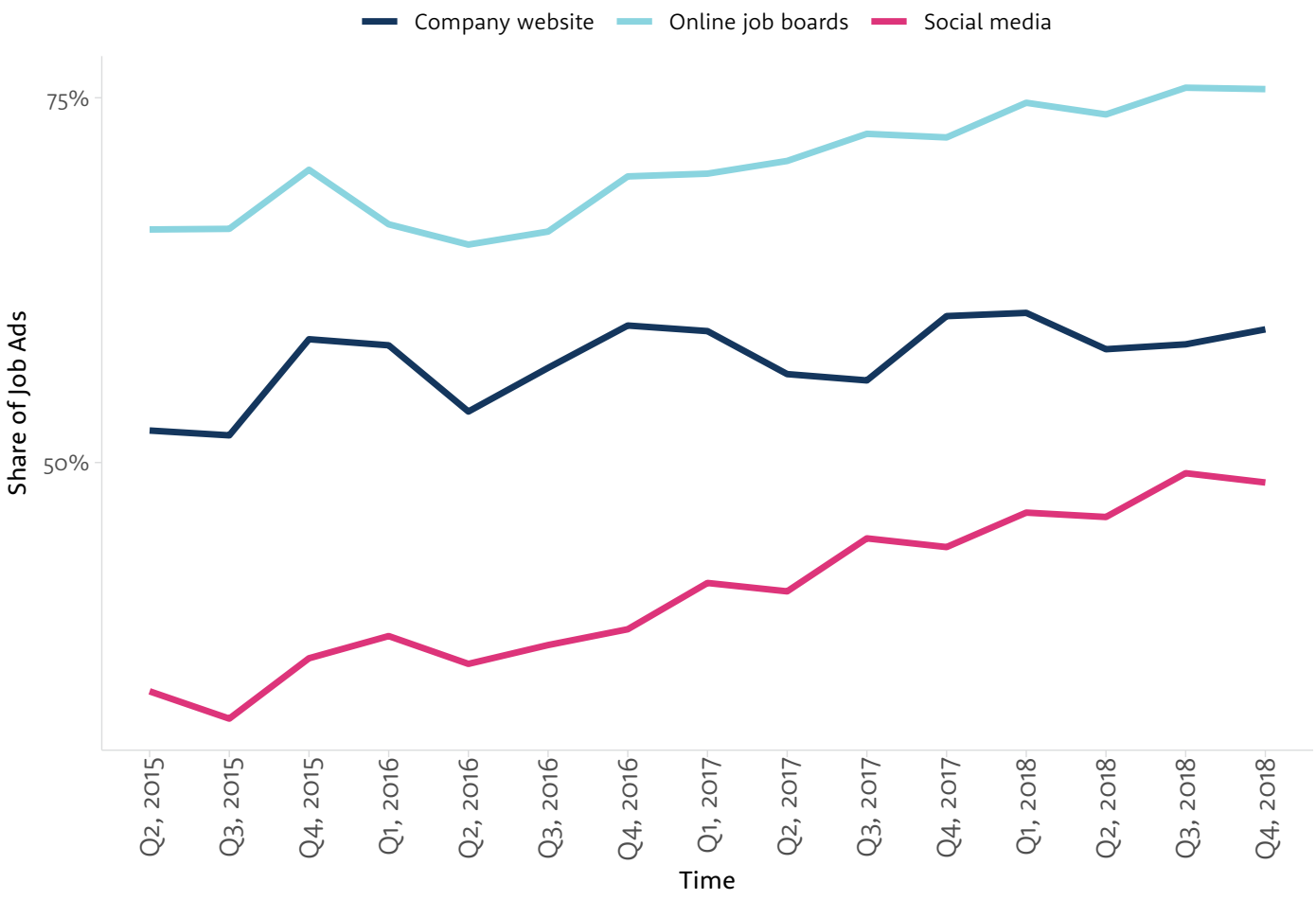
Given how job postings are generated, then, what does this represent? Firstly, job postings give us a signal about a firm's intention to hire. Though posting a job may be a relatively inexpensive process, developing a particular job description is not. As a result, a job posting represents a genuine signal of employers' intention to hire for that particular position.

Secondly, looking at job postings, and job vacancies data in general, allows us to better understand the flow of employment and the dynamics of where and how job transitions occur. This gives us a different set of insights from analyzing the stock of employment. For example, a particular occupation may have a relatively fixed stock of workers in it, but may involve short term contracts and thus a high level of dynamism in hiring and jobs being posted.

Using job postings for labour market insights is not new. Researchers have utilized newspaper job adverts as early as 1987 in the United States to understand labour dynamics.[33] However, the medium through which job advertisements have been communicated has shifted over time. Statistics Canada's Job Vacancy and Wage Survey (JVWS) shows that between 2015 and 2018, around 70 percent of vacancies were posted on an online job board, with an increasing trend. This is consistent with trends observed in the United States.[34] This implies that Burning Glass data coverage improves with time.

Figure A.1

Job Recruitment Strategy in Canada



Source: Job Vacancy and Wages Survey

However, this also shows that online job postings do not exhaustively cover the economy. It is then important to understand how well Burning Glass aligns with and deviates from other sources on job openings, as well as the current stock of workers working in a particular occupation. To understand these trends, we compared the Burning Glass data sample with two main sources: the JVWS and the 2016 Canadian long form census data.

Overview of Burning Glass data

The data we analyzed for this report spans the period between January 5th, 2012 and December 31st, 2018. This represents 7,192,983 job postings in total. Job postings were collected for all 13 provinces and territories in Canada, though only postings in English were collected, due to the platform being optimized for processing English-language job postings.

The majority of job postings were concentrated in Ontario — unsurprising given the fact that Ontario is the largest province in Canada. Though Quebec is the second largest, the lack of ability to parse job data in French means that the number of jobs recorded here was also lower than the province's overall share of employment.

On average, 85,631 postings were captured across Canada every month:

Figure A.2

Distribution of Burning Glass Job Postings by Province

Source: Burning Glass

Table A.1: **Distribution of Burning Glass Job Postings by Province**

Province/Territory	Number of Job Postings (cumulative 2012-2018)	Share of Burning Glass Job Postings	Share of job postings in JVWS (Q2 2018)	Share of Employment (2019)[35]
British Columbia	1,071,217	14.9%	19.6%	13.5%
Ontario	2,946,740	41%	38.4%	39.0%
Manitoba	193,378	2.7%	2.8%	3.4%
Alberta	1,103,817	15.3%	10.6%	12.3%
Nova Scotia	203,508	2.8%	2.1%	2.4%
Quebec	1,044,653	14.5%	21.1%	22.8%
Saskatchewan	363,185	5%	1.9%	3%
New Brunswick	129,694	1.8%	1.7%	1.9%
Newfoundland and Labrador	85,515	1.2%	0.7%	1.2%
Northwest Territories	9,117	0.1%	0.1%	<0.1%
Yukon Territories	7,510	0.1%	0.2%	<0.1%
Prince Edward Islands	28,395	0.4%	0.4%	0.4%
Nunavut	6,254	0.1%	0.1%	<0.1%

To further examine the representativeness of Burning Glass data, we compared it to other sources of labour market information. The first source of data we consulted was the JVWS. We chose this specific dataset as it attempts to capture similar job openings data to Burning Glass. Importantly, the JVWS captures job openings regardless of whether they are posted online or offline. However, some sample variation is inevitable in the JVWS, as it has a smaller (albeit random) sample size compared to Burning Glass.

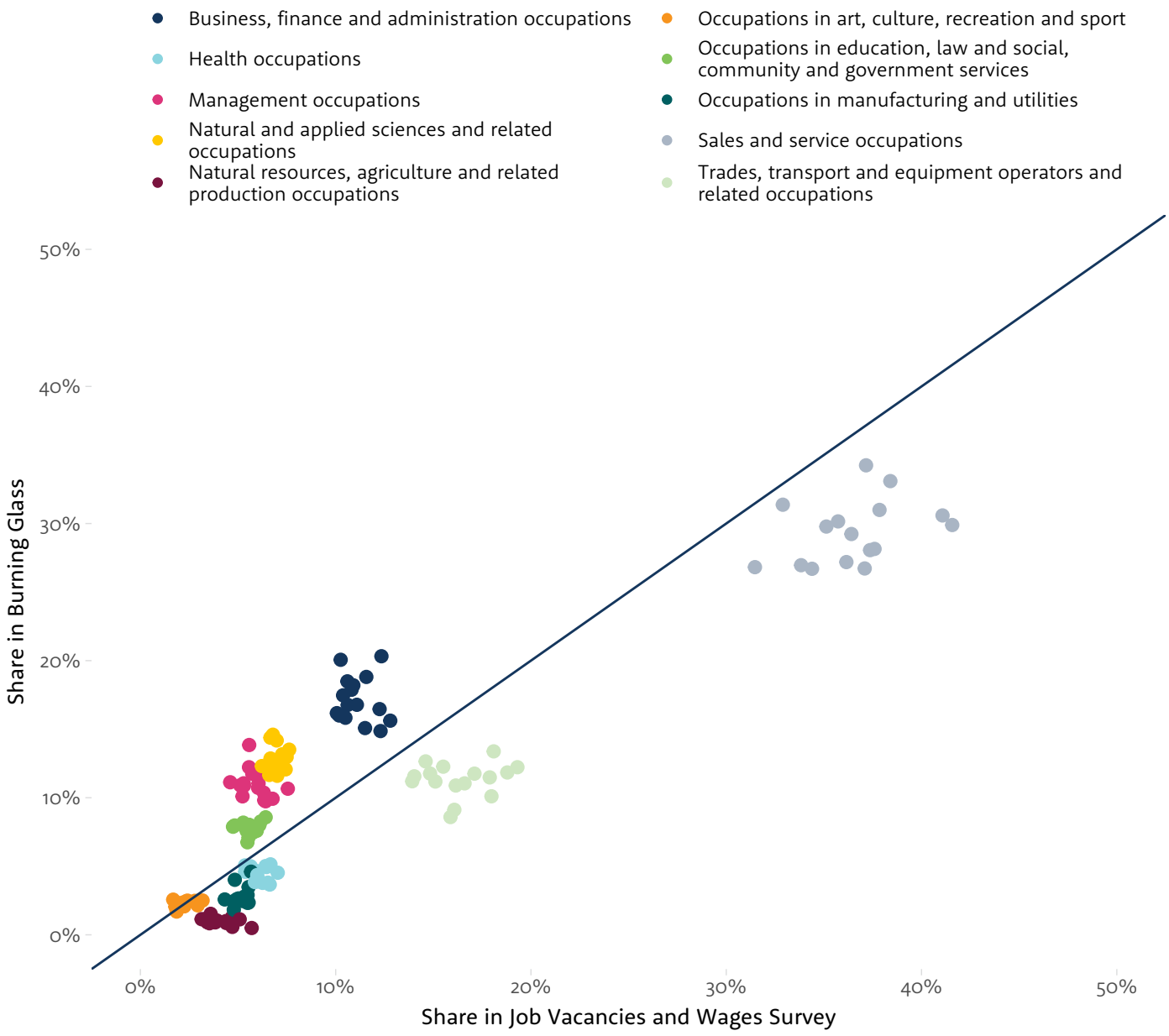
Table A.2: Data Quality in JVWS for Different NOC Levels

Quality	NOC-1	NOC-2	NOC-3	NOC-4
Excellent	73.1%	32.0%	4.8%	0.8%
Very Good	19.4%	48.1%	32.1%	10.2%
Good	1.3%	11.1%	25.4%	14.7%
Acceptable	6.3%	2.3%	20.9%	26.2%
Use with Caution	0.0%	4.5%	6.5%	13.6%
Too Unreliable	0.0%	1.9%	6.5%	18.5%
Suppressed	0.0%	0.0%	3.7%	14.1%
Not Applicable	0.0%	0.0%	0.1%	2.0%

The JVWS was introduced in 2015 and is collected on a quarterly basis — every January, April, July, and October. As such, we also compiled monthly job postings data captured by Burning Glass in those four months, for four years: 2015, 2016, 2017, and 2018. Specifically, we looked at the share of job postings in both Burning Glass and JVWS that came from a particular major occupational group (NOC-1). Further disaggregation of occupational group is possible, though not advisable, as JVWS becomes less precise. The highest aggregation of NOC is the only level with a majority of data points being deemed 'excellent'.

Figure A.3

Comparison between Burning Glass and JVWS, by Occupation

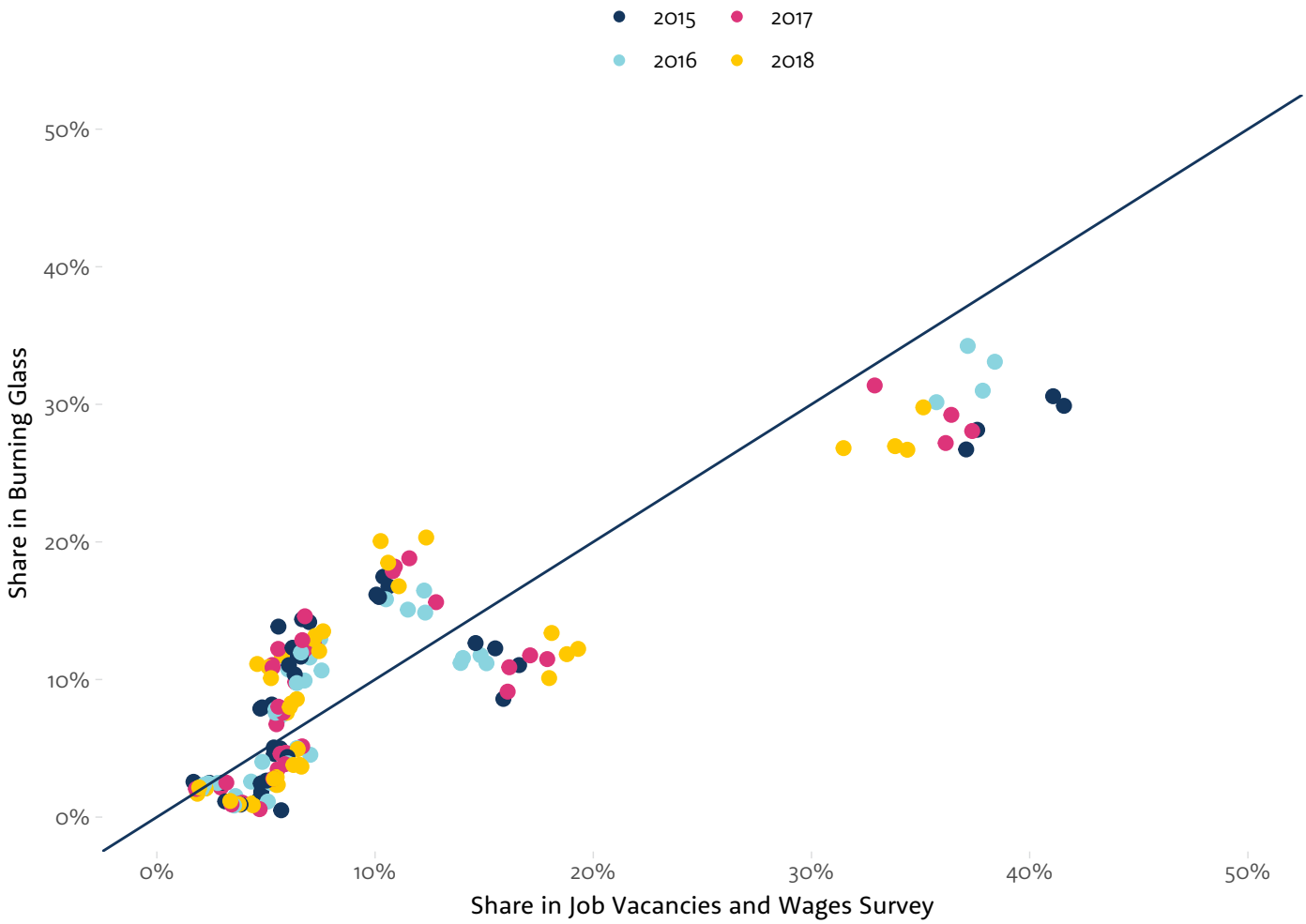


Source: Job Vacancies and Wages Survey, Burning Glass.

It can be observed that, broadly, job postings in Burning Glass match well with JVWS. There are some particular discrepancies: for example, Burning Glass has considerably less postings, in terms of proportion, from sales and service occupations, while having a higher share of manufacturing, natural and applied sciences and related occupations, and management occupations. The correlation between the two series was 0.8775.

Figure A.4

Comparison between Burning Glass and JVWS, by Year

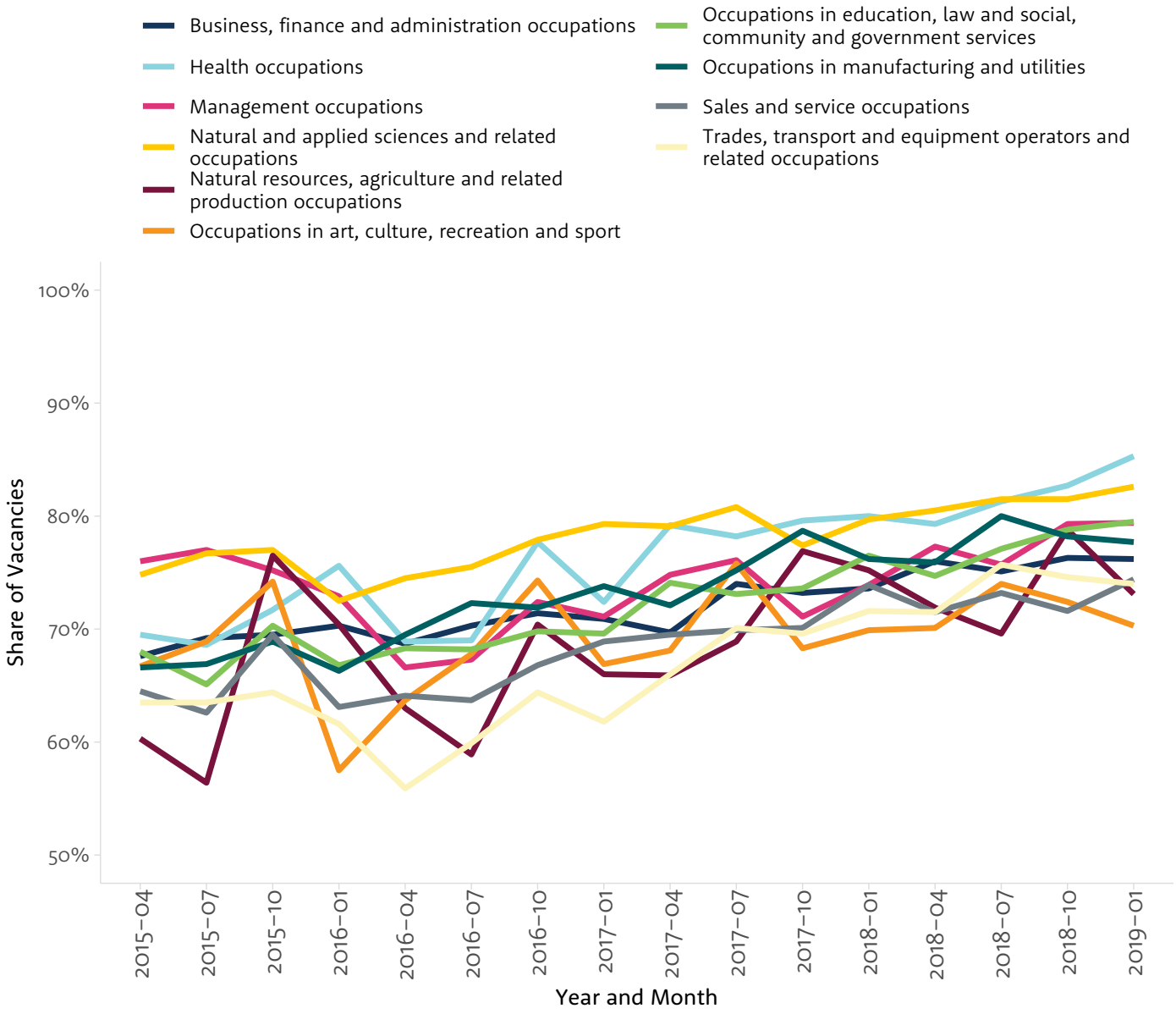


Source: Job Vacancies and Wages Survey, Burning Glass.

The deviation between the two data sources was relatively stable across the four years, while the mean of the squared differences tended to be lower in later years. 2016 had the lowest deviation between the two data sources. One potential explanation for this deviation could relate to the change in propensity for employers to advertise job postings online. The data shows that there are substantial variations in both temporal and occupational dimensions in the share of job advertisements posted online.[36]

Figure A.5

Change in Share of Vacancies Advertised Online, by Occupation



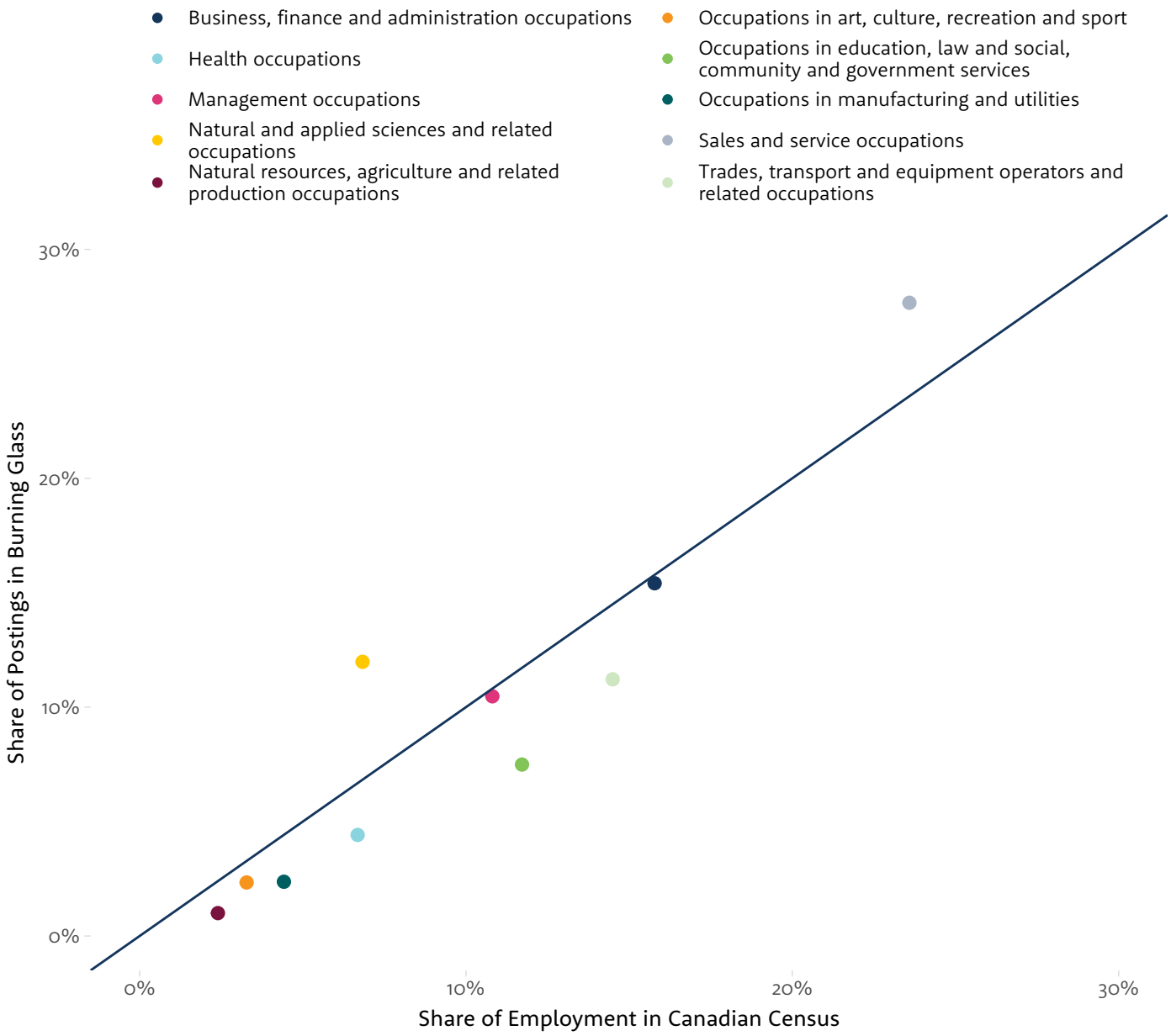
Source: Job Vacancies and Wages Survey

What this implies for Burning Glass's data is that temporal variation in job postings for an occupation can be decomposed into two main forces: increased use of online job advertisements for that occupation, and increased demand for that occupation overall.[37]

Another comparison that can characterize Burning Glass's ability to summarize labour conditions is in understanding how well it matches up to the current stock of workers. When we compared the postings by the share of workers working in major occupations in 2015 (as collected by the 2016 census), it was clear that Burning Glass postings have higher shares of posting in sales and service occupations,[38] as well as natural and applied sciences and related occupations. Apart from that, the two sources are remarkably similar.

Figure A.6

Comparison Between Burning Glass and the 2016 Census, by Occupation



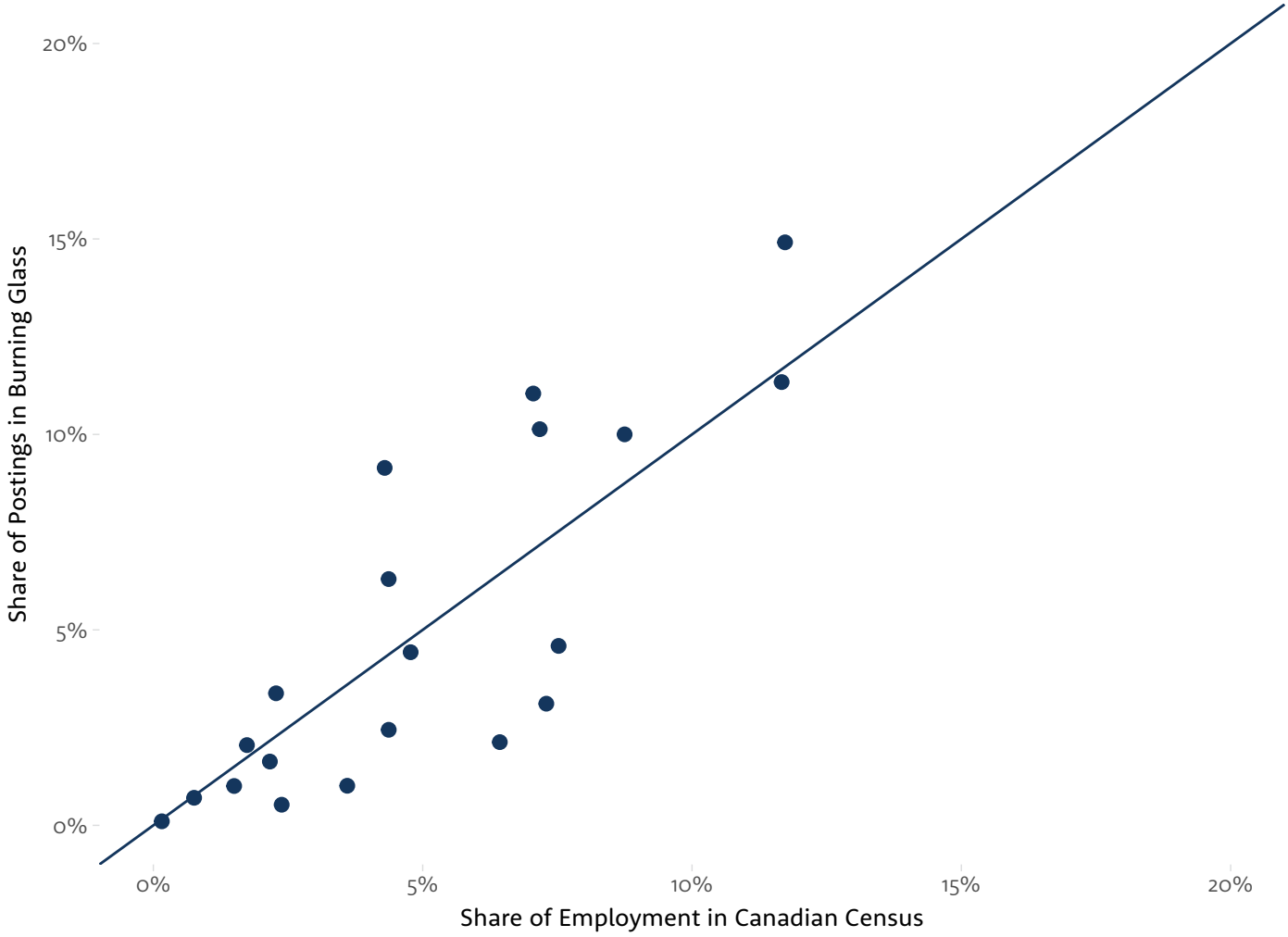
Source: Burning Glass, and 2016 Canadian Census

When we compared the share of postings that came from a particular industry in Burning Glass to the share of workers in a particular industry across Canada (again, for 2015, using the 2016 Canadian census), it was clear that, on average, Burning Glass is a fairly good representation of the broader labour market. The correlation in this case was 0.824.

When examining representativeness, we also wanted to examine whether job postings disproportionately represent less- or more-skilled workers. As a proxy, we used the education credentials asked for by employers in Burning Glass job postings data, and compared them to the educational attainment of individuals working in those same occupations. As a blunt instrument, we calculated the share of workers (and job postings) with a bachelor's degree or above.

Figure A.7

Comparison Between Burning Glass, and the 2016 Census, by Industry



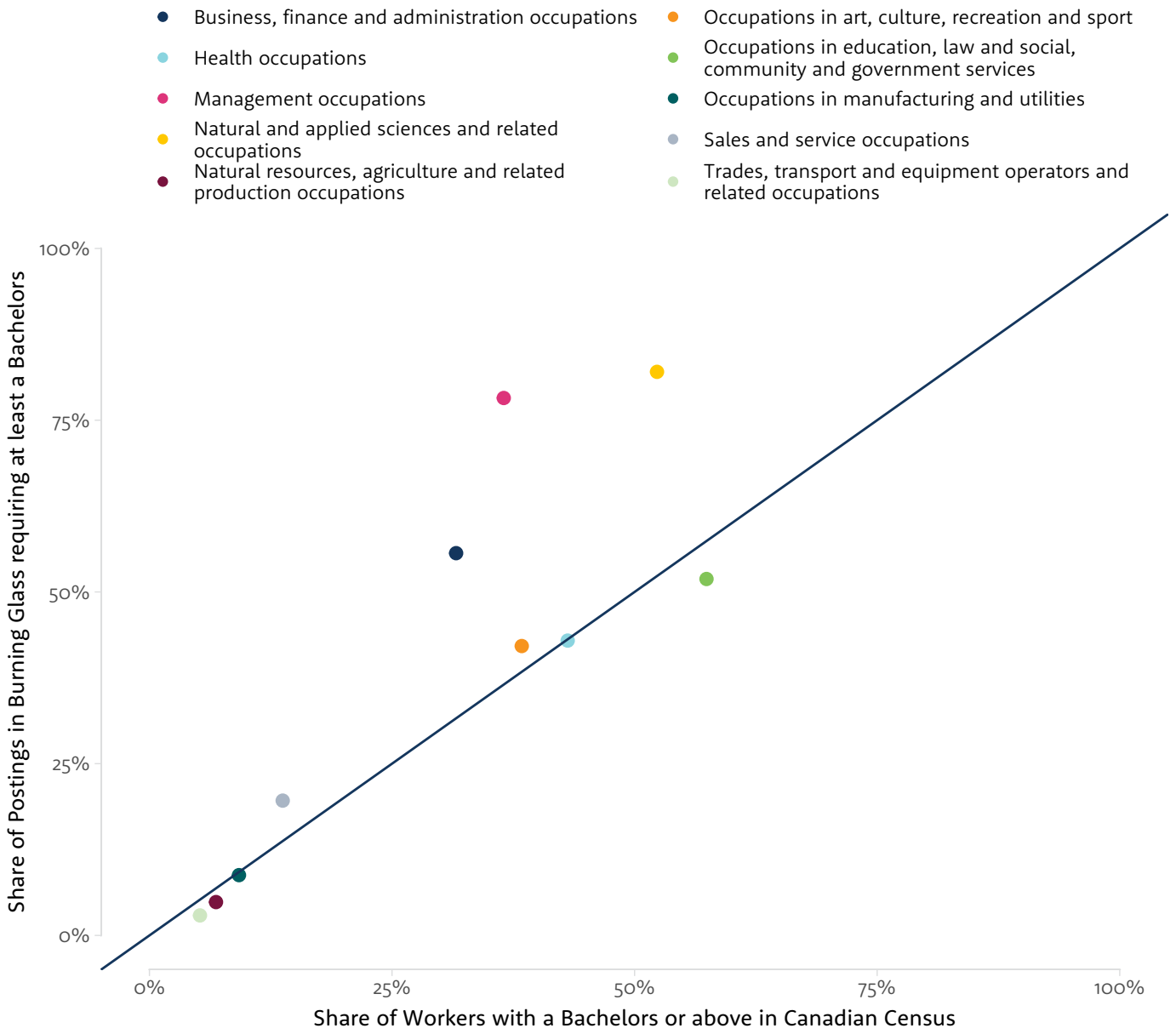
Source: Burning Glass, and 2016 Canadian Census

In Burning Glass, just below one third of job postings (31.3 percent) in 2015 listed degree requirements.

Looking at the distribution of degrees, the two data sources are similar across most occupational groups. Three occupational groups stand out in this comparison, all of which have a higher share of job postings asking for a bachelor's degree or above compared to actual occupational distribution. These are management occupations, natural and applied sciences and related occupations, as well as business, finance, and administration occupations. This could reflect changing credential requirements, or the existence of internal promotion mechanisms that might not be captured in job vacancies and postings.

Figure A.8

Comparison Between Burning Glass, and the 2016 Census, by Education and Occupation



Source: Burning Glass, 2016 Canadian Census

From this series of representativeness checks, we can conclude that, overall, Burning Glass data provides a fairly good snapshot of the Canadian labour market. There are important points to consider, however, detailed below:

+ Burning Glass underrepresents job postings coming from Quebec, though this does not appear to affect its overall representativeness of the Canadian labour market.

+ Data coverage increases over time, making temporal comparisons difficult in our data.

APPENDIX B: DEFINITION OF DIGITAL SKILLS

To define digital skills, we took inspiration from the framework established in Djumalieva and Sleeman (2018), which examined the demand for digital skills in the UK using Burning Glass software markers as a starting set. They then utilized a word-embedding model to elicit other skills that are commonly listed alongside software skills to define their set of digital skills.

For this report, we started with that same set of software skills. In our Canadian sample, there were 1,753 unique software skills. We further augmented this list by manually examining all 29 skill clusters and 650 skill cluster families to identify broad clusters that identify software and digital skills. This resulted in the following clusters (highest skill hierarchy) being included in our analysis: "Information Technology", "Analysis", "E-Commerce", "Web Analytics", and "Bioinformatics". We call these "base digital skills".

Using occupation tech intensity score to define software+

However, over 6,200 skills (representing 48% of the skills space) were not categorized into a skill family or a skill cluster family. This group might contain software and technology skills that are important to our discussion. Additionally, as our study also focuses on identifying non-digital skills that appear alongside digital skills, any form of clustering or distance-based metric between skills may not be effective for identifying digital skills, as certain non-digital skills we want to identify as unique may be classified as a digital skill instead.

As a result, to independently identify important digital skills, we focussed on augmenting the base set of skills included in the Burning Glass taxonomy by also identifying skills that consistently show up in digitally-intensive occupations. To identify the digital intensity of an occupation, we utilized a methodology outlined in Vu, Zafar, and Lamb (2019), using the US O*Net database. Unlike in the previous report, which generated the digital intensity for Canadian National Occupation Classifications (NOCs), since Burning Glass maps directly to O*Net classifications, we generated the rankings directly for O*Net occupations.

Since each job posting was assigned a corresponding O*Net occupation, we were able to assign to each posting, and all the skills listed in each posting, the corresponding digital intensity score. We then took the average rank for each skill across all job postings. Intuitively, the resulting rank for a skill will be high if that skill is consistently listed in occupations with a high digital ranking. Conversely, the rank for a skill will be low if that skill is more likely to show up in job postings associated with occupations with low digital ranking. For example, the skill for Objective C (a programming language) is 18.75 (111th highest tech skill out of 12833 skills), whereas the skill for Companionship (a dance technique) has a rank of 822.68, being one of the lowest-ranked skills.

To examine whether our approach yields useful results, we tested the following hypothesis: the probability that a particular skill is a base digital skill (listed in one of Burning Glass's previously defined digital skills classifications) increases

as the digital intensity ranking increases. As the construction for the Burning Glass digital skills classifications are binaries (i.e., a skill either does or does not exist within the digital skill category), we estimated the probability of a skill with a particular ranking of being a base digital skill using a logistic regression. A logistic regression estimates the conditional density function of a particular event happening. In this case, it estimates the probability that a particular skill, with a particular ranking, is a base digital skill. We expected the coefficient on the ranking to be negative.[39] The estimated equation shows that this is the case. More importantly, it illuminates a potential cut-off for defining digital skills, which we discuss later.

Table A.3: **Regression Results**

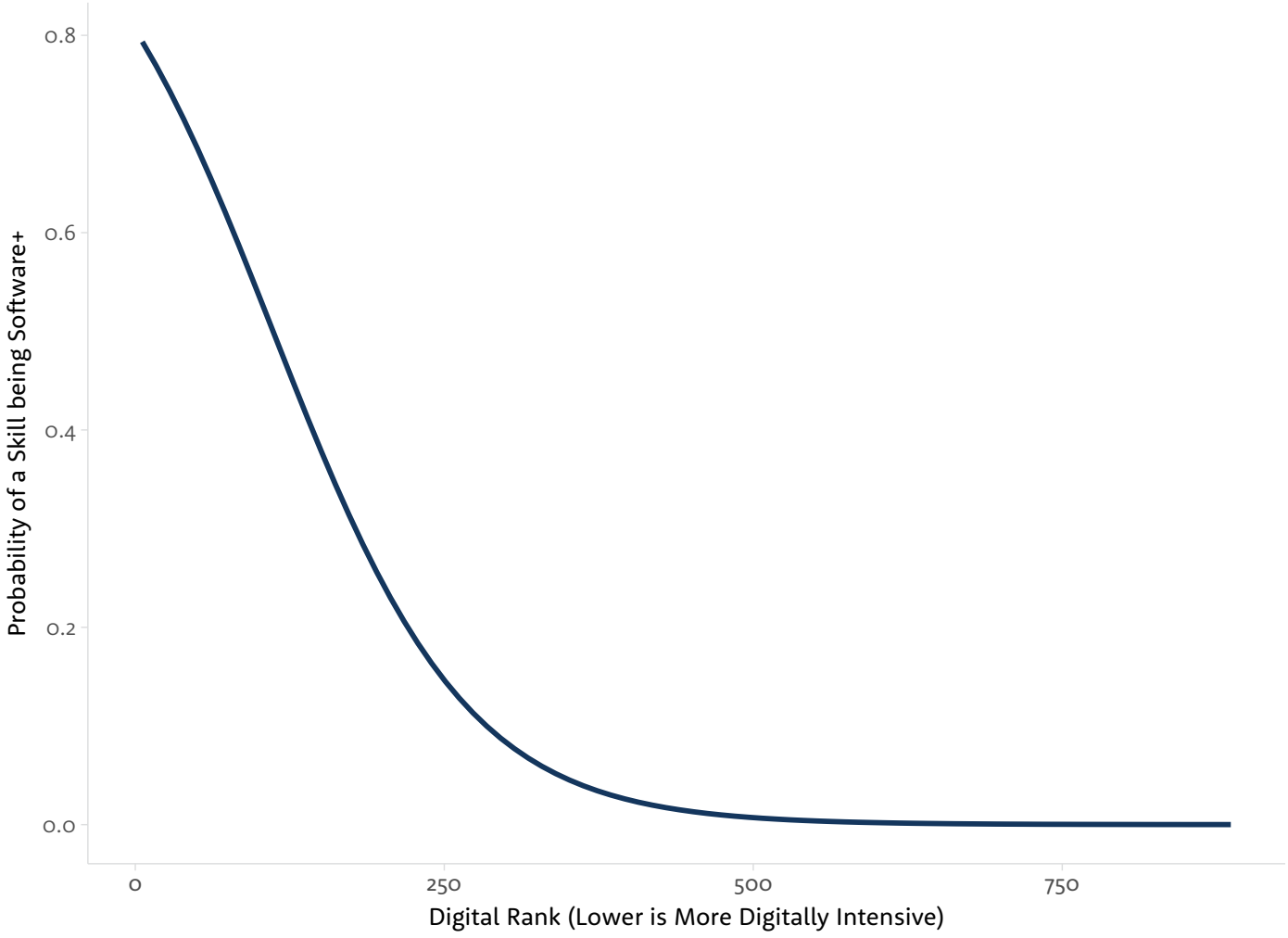
	Estimate	Z value	p-value
Intercept	**1.4181***** (0.0552)	25.71	0
Digital Intensity	**-0.0127***** (0.0002)	-43.97	0

Dependent variable is whether a skill is a base digital skill or not. Logistics regression using Maximum Likelihood Estimation.

Degrees of Freedom: 9,509

Figure A.9

Estimated Logistics Regression for Digital Intensity Score



Source: Author Calculations

Assumption verification

One obvious limitation of using this method is that it fails to capture digital skills that consistently show up in job postings where the associated occupations have low digital rankings. However, since Burning Glass also captures whether a skill is a software skill, we can circumvent part of this issue by also including in our definition of digital skills those that are labelled as software skills.

But there is still the challenge of classifying potential skills as digital skills that do not appear in highly digital occupations, and also do not fit within one of Burning Glass's base digital or software skills. For this method to be able to identify most digital skills, the following identification assumption must hold: when a digital skill is listed for a job posting, the likelihood of that digital skill being labeled as software increases as the digital intensity of an occupation declines. This assumption is motivated by previous research from Huynh, Do (2017) who distinguish between baseline digital skills, workforce digital skills, and professional digital skills. The more digitally-intensive the skill, the more one is required to not just be proficient in a specific software, but to be well-versed in computational thinking.

To test this assumption, we took the base set of 1,753 software skills and combined it with 1,578 base digital skills, and estimated the conditional probability of a skill in a job posting being a software skill[40] and how it changed with the tech intensity ranking of a job posting. Using Bayes' rule,[41] this simply became:

$$\frac{P(software)}{P(digital)}$$

We then estimated the equation:

$$\frac{P(software)}{P(digital)} = \beta_0 + \beta_1 h_1 + \epsilon_i$$

Where h is the harmonic rank or digital intensity of a particular occupation (which varies at the occupation level), and the error term varies at the individual job postings level. As the ranking increased for less digitally-intense occupations, we expected the coefficient on the ranking to be positive. In addition, there is no reason to believe in a homoskedastic error structure here (as the variance of the probability may vary according to the tech intensity of an occupation), so we computed heteroskedasticity-robust standard errors for our estimate:

Table A.4: **Regression Results**

	Estimate	t value	p-value
Intercept	6.504×10^{-1} *** (3.946×10^{-4})	1648	0
Digital Intensity	2.419×10^{-4} *** (1.102×10^{-6})	219.5	0

Dependent variable is the conditional probability of a digital skill in a job posting being a software, conditional on it being a digital skill. Linear Probability model using Ordinary Least Squares. Standard errors are heteroskedasticity-robust.

Degrees of Freedom: 2,252,332
R^2:0.021
F-statistics: 4.82×10^4

The estimate shows that a positive relationship exists, supporting the validity of our approach in identifying the right set of skills. Even if some important digital skills were not captured using this methodology, the method we developed in identifying hybrid jobs, which involves distance-based algorithms or clustering-type algorithms, should be able to identify the residual digital skills. For robustness, we also tested some common non-linear specifications, none of which invalidated this assumption in the effective range of the function [0,1].

Software+ as a definition of digital skills

Finally, we discuss the cut-off point we propose in choosing digital skills. The central tension in choosing the cut-off points for our definition of digital skills was in maximizing the number of unidentified digital skills being identified, while minimizing the number of non-digital skills being misclassified as a digital skill. As our method involved choosing a single cut-off point, the two were in direct conflict. We posit several plausible cut-off points, and discuss the value of each:

1. **The point where the estimated probability of a skill being a base digital skill is 50 percent.** The first cut-off point we propose is the digital intensity score, where roughly half of the skills present are not included in the base digital skills. For our sample, this occurs between rank 109 and 110. This means that we will end up with 1,113 additional skills being identified as digital skills, for a total of 3,651 digital skills.

2. **The point where the rate of change in the slope of the estimated function becomes increasingly negative.** In our sample, this point lies at around rank 213 and 214. This will identify 2,519 additional digital skills, for a total of 5,037 digital skills.[42]

For our study, we erred on the side of caution, where the number of additional skills identified did not exceed the number of skills already classified as base digital skills. A spot check also confirmed these results. Most skills that fell between the digital intensity rank of 109 and 110 could be reasonably classified as digital skills, while skills that fell around rank 213 were much less likely to be clearly digital skills. We chose these cut-offs as our definitions.

Table A.5: Skills Around the Strict Cut-off

Skill	Score
Bitcoin	109.59
Cognitive Science	109.64
Change data capture	109.68
General Packet Radio Service (GPRS)	109.68
Global Organizational Development	109.69
Network Installation	109.69
Quick Test Professional (QTP)	109.7
NOVELL	109.7
Engineering Design	109.71
Clinical Trial Progress Monitoring	109.71
Engineering Design and Installation	109.73
Raspberry Pi	109.80
Risk Based Testing	109.9
Group policy	109.97
Enzyme Function	110.02
Efficiency Estimation	110.05
Medical Device Design	110.05
Joomla	110.077
Rapid Prototyping	110.087
Sustainable Engineering	110.11

Table A.6: Skills Around the Generous Cut-off

Skill	Score
Biomedical Engineering	213.45
Facility and Site Construction Layout	213.47
Virtual Agents	213.48
Cell Phone Industry Knowledge	213.54
Corrective Action Planning	213.59
Mortgage-Backed Security (MBS)	213.62
Multiple Regression	213.69
RNA Isolation	213.69
Blogger	213.8
Site Assessments	213.8
Liquidity Risk Models	213.84
Long-Only	213.91
Open End Wrenches	214.05
Restoration Strategy	214.14
PPM Tools	214.2
Electrical Diagrams / Schematics	214.28
Production Part Approval Process (PPAP)	214.3
Bill Preparation	214.34
Frozen Shoulder	214.43
Fit/gap analysis	214.48

We then took a random sample of 100 skills from the taxonomy, hand-classified them into software+ skills, and assessed type 1 and type 2 errors using the two threshold heuristics. In the stricter threshold, 29 out of 100 skills were identified as software+ skills. The hand classification and the strict threshold agreed 92 times. There were seven instances where the hand classification classified a skill as software+ and the stricter threshold did not. There was one instance where the hand classification did not classify a skill as software+ and the stricter threshold did.

Type 1 error: 20%; type 2 error: 1.5%

Under the second (more generous) threshold, 37 skills were identified to be software+. The hand classification and the second threshold agreed 92 times. There were three instances where the hand classification classified a skill as software+ and the wider threshold did not. There were five instances where the hand classification did not classify a skill as software+, while the threshold categorized the skill as software+.

Type 1 error: 8.5%; type 2 error: 7.7%

APPENDIX C: NETWORK ANALYTICAL FRAMEWORK

The central focus of this report is identifying how skills interact with each other. In particular, we are interested in how digital skills interact with non-digital skills. Our contribution to this literature involves conceptualizing skills and job postings as a network (or, more formally, a graph). This is a natural conceptualization given Burning Glass's data, where different skills are connected to each other by appearing together in the same job posting. Our purpose, then, is to understand the patterns of communities that might exist within these skills.

Within a graph theory framework, there are many community-detection algorithms, each with different graph metrics as the objective function. We must prioritize the most desirable community characteristics for our research purposes. For our purposes, it is important to distinguish between two types of skills: general skills and specialized skills. General skills are skills such as 'communications' and 'leadership'. Almost all job postings contain a set of general skills, as these skills are applicable across most occupations. Specific skills are those that appear less often, and are usually confined to specific occupational groups or tasks. As our study focuses on digital skills, many of which are non-general, we ideally would place less weight on general skills in defining skills clusters. In addition, general skills should not form an important role within any cluster, and should ideally connect different clusters together.

Given these constraints, some community detection algorithms, such as label propagation, where nodes with large weights have the most influence in determining the community, were ruled out. We instead focused on two particular graph clustering techniques: modularity-based and edge-betweenness-based.

Given a community structure, the modularity of a graph is, intuitively, how well-connected nodes within each community are compared to the likelihood of these nodes connecting if such connections are formed randomly. Modularity-based algorithms in community detection are a class of algorithm that divides a network into communities that maximize modularity in the network. Many popular and well-implemented versions of these algorithms operate on a hierarchical basis: each node in the network starts in its own community and is iteratively aggregated, until all nodes belong to the same community. Communities that when grouped together increase global modularity the most are grouped, and the aggregation level that maximizes the modularity is then chosen.

$$Q = \frac{1}{2m} \sum_{vw} \left[A_{vw} - \frac{k_v k_w}{2m} \right] \delta(c_v, c_w)$$

Modularity of a Network

Where Q is the modularity

A_{vw} is the adjacency indicator for nodes v and w

k_i is the degree (sum of weights of edges connected to) of node i

m is the number of nodes in the network

$\delta(c_v, c_w)$ is a function equal to 1 when v and w belongs to the same community, otherwise 0.

$$g(e) = \sum_{s \neq e \neq t} \frac{\sigma_{st}(e)}{\sigma_{st}}$$

Betweenness Centrality of an Edge

Where $g(e)$ is the betweenness centrality of edge e

σ_{st} is the number of shortest paths that connect between nodes s and t

$\sigma_{st}(e)$is the number of shortest paths that connect between nodes s and t that go through edge e

As a result, these algorithms tend to group nodes with lower degrees first, and thus have a lower chance of being connected to each other in a

random network. Nodes with large degrees, or nodes with the highest chance of being connected to each other in a random network (which, in our context, corresponds to basic skills), do not act as pivotal nodes in defining a community. However, as modularity improves greatly with a higher number of edges in the network than expected, these large nodes are more likely to have many 'within-community' edges, as opposed to 'across-community' edges. This is due to the fact that any connection to nodes outside the community from these central nodes will not contribute toward the global modularity, while non-connections within the same community will hurt global modularity.

Edge-betweenness algorithms rely on detecting community structures through another graph metric: betweenness centrality. Betweenness centrality is measured for each edge, where edges that are on the highest number of shortest paths connecting any two nodes will have the highest betweenness. Intuitively, if such an edge is removed, the graph will be more divided than it was previously. After removing enough edges with high betweenness, the graph will break into two distinct communities, then three; eventually, each node will occupy its own community (after all edges have been deleted).

Finding community structure using edge-betweenness centrality is the reverse of a modularity-based algorithm, where edges are removed from the main graph until every node belongs to its own community. Such an algorithm, then, will likely delete the edges associated with nodes with a high degree (in our context, basic skills) first, where these nodes then act as nodes that connect between communities. However, as these nodes are likely to be isolated first, they will only determine initial community divisions, and not an inherent community structure.

However, edge-betweenness algorithms are computationally demanding, and do not scale to a large number of nodes. For this reason, we favoured modularity-based algorithms for this report.

Due to the graph having a vastly higher number of job nodes than skill nodes, when running a community detection algorithm, a graph that aggregates the bipartite network into a network just involving skills was beneficial for our purpose. An important issue that is worth discussing involves how to manage weights of edges between skills. In the full bipartite network, skills can be connected multiple times through multiple jobs, and, intuitively, edges between skills connected by many jobs should receive higher weights than edges between skills connected by fewer jobs. However, our choice of weights and their relative importance will also affect our choice of the objective function used in any clustering or community detection algorithm, so a discussion on ideal weights is warranted here.

In this quadrant of analytics, an ideal weight would, as stated in the previous paragraph, put more weights onto edges where skills co-occur more often. However, in our full sample, the edge with the highest weight is 804,756 times stronger than the edges with the lowest weight. More importantly, the edges at the 75th percentile are six times higher than the edges at the 25th percentile, highlighting a huge difference between co-occurrences. We likely want to preserve the severity of this disparity.

In transforming these weights, two considerations should be given: whether to preserve the ordinal nature of these weights, and how to transform the cardinal nature of these weights. The first question surrounding ordinality is the most important, as altering the ordering of edge weights will change the graph structure, and any such decision needs to be made with thoughtful consideration. We

Table A.7: Summary Statistics of Counts of Skills

Minimum	1st Quadrant	Median	Mean	3rd Quadrant	Maximum
1	2	5	100.2	18	80475

should only consider alternating the ordinality of the weights if we believe that weights based on the number of co-occurrences have fundamental flaws in representing the structure of how skills are related to each other.

One such flaw can relate to average job length of an occupation, given the data we are currently considering. For example, jobs in some occupations tend to be short-term in nature (sales associates, for example), due to both employee and employer reasons. In these instances, occupations with high turnover will likely post job openings more frequently, increasing the instances where two skills in that occupation co-occur (as we expect job postings for a similar, or, in some circumstances, the exact same role, to be similar). However, in Appendix A, when we explore the representativeness of Burning Glass's sample, we learn that the difference between the share of postings and the number of people working in those occupations is not particularly high (even in occupations where we expect there to be a lot of employment flow, such as sales and service occupations), negating these concerns.

One of the more intuitive ways that we can rescale these weights is by normalizing them. However, given the structure of these weights, assuming a normal distribution in re-weighting is not prudent (due to the inherent lower bound of 0 in the number of possible co-occurrences). Given these considerations, we chose not to re-weight the edges in analyzing our network.

Aggregating job postings at the occupation level

After we implemented the clustering algorithms on the skills network, we ended up with eight distinct communities of skills (characterized in the main text). The next step of the analysis was in proposing methods to translate insights gained from partitioning skills into these eight clusters onto job postings. With almost six million nodes, the full jobs space network is large and implies high levels of computational complexity, especially for algorithms that require the full adjacency matrix in assigning nodes to communities.

Though we attempted to implement a community detection algorithm for the full network using several different implementations (on a variety of different data structures), performing analytics on the full network was deemed impractical.

As a result, we decided to aggregate job postings at the O*Net Occupation level, where job postings associated with a particular O*Net occupation were aggregated, at the skills level and at the skill community level. In particular, we aimed to characterize a "representative" job posting for each of the almost 1,000 occupational groups. This approach, though useful, has several disadvantages. The main disadvantage of following such an approach is that we take the occupational partition as a given, and may potentially miss distinct occupational groups that could be identified in a skills sense, as roles in this distinct group may be distributed across several defined occupational groups.

To implement our analysis, however, we calculated, for each skill, the probability of that skill showing up in a job posting for a particular occupation using the full data. We then aggregated the individual skill probability at the skill community level. As a result, we generated the share of skills belonging to a particular community in an occupation. As a final step of our implementation, we calculated the average number of skills listed for each occupation, and calculated the implied number of skills in such a "representative job posting" belonging to each skill community.

We also used the probability that each skill shows up in a job posting to generate a hypothetical job posting for some characterized occupations.

Finally, we defined an occupation's **hybridness** by measuring the variance of shares of skills that came from eight skill clusters (four digital and four non-digital) for each occupation. Intuitively, the most hybrid occupation will have the lowest variance in how skills are distributed across different skill clusters (as no single skill cluster will dominate skill listings from one domain).

APPENDIX D: NON-DIGITAL SKILLS CLUSTERS

Using a similar methodology, we applied the community detection algorithms to non-digital skills. This allowed us to identify four broad clusters of non-digital skills in our data:

1. **Skills associated with having a bachelor's degree:** This skill cluster consisted of 2, 755 unique skills, and included specialized skills and knowledge that span domains and are typically associated with workers holding a bachelor's degree or higher. Prominent skills included in the cluster were budgeting (490,923 mentions), project management (398,143 mentions), change management (90,794 mentions), mechanical engineering (44,394 mentions), economics (42,461 mentions), and chemistry (25,839 mentions).

2. **Skills associated with not having a bachelor's degree:** This skill cluster consisted of 1,745 unique skills commonly utilized in work associated with less formal education, such as the trades, manufacturing, personal services, and administrative functions. Many skills in this cluster are those commonly thought of as more routine-oriented skills, which are more susceptible to automation. However, also included in this cluster were many specialized, non-routine manual skills. Prominent skills included communications skills (2,208,324 mentions), repair (447,334 mentions), administrative support (274,119 mentions), machinery (120,648 mentions), childcare (93,800 mentions), and welding (92,749 mentions).

3. **Communications, marketing, and public relations skills:** This skill cluster consisted of 1,627 unique skills primarily related to communications and marketing for a specific business, product, and/or service for external and internal stakeholders. Some prominent skills in this cluster included teamwork/collaboration (1,044,326 mentions), customer service (989,886 mentions), sales (784,899 mentions), written communications (369,190 mentions), and bilingual abilities (295,406 mentions).

4. **Healthcare and medicine skills:** This skill cluster consisted of 3,065 skills pertaining to healthcare and human services. Prominent skills in this cluster included those related to patient care (71,458 mentions), mental health (51,833 mentions), social services (40,150 mentions), public health and safety (36,176 mentions), and long-term care (33,605 mentions).

In particular, the names assigned to the first two clusters were guided by a high level of association between credential requirements (listing a bachelor's degree as a minimum requirement or not) associated with job postings that have a higher share of skills coming from each of the two clusters. We tested such an association in a linear probability setting and a logistics setting, obtaining similar results.

Table A.8: **Regression Results**

	Estimate	p-value
Intercept	0.94 *** (4.3×10^{-4})	0
Share of skills associated with not having a bachelor's degree	-1.00 *** (8.6×10^{-4})	0

Dependent variable is a dummy variable indicating whether a job posting lists at least a bachelor's degree as a minimum requirement. Linear Probability model using Ordinary Least Squares.
Degrees of Freedom: 2,089,791

R^2:0.021

F-statistics: 1.46×10^6

Table A.9: **Regression Results**

	Estimate	p-value
Intercept	2.42 *** (0.003)	0
Share of skills associated with not having a bachelor's degree	-6.03 *** (0.008)	0

Dependent variable is a dummy variable indicating whether a job posting lists at least a bachelor's degree as a minimum requirement. Logistics Regression using Maximum Likelihood Estimation.
Degrees of Freedom: 2,089,791

ENDNOTES

32. Hershbein, B. J., & Kahn, L. B. (2018). Do Recessions Accelerate Routine-Biased Technological Change ? Evidence from Vacancy Postings. *American Economic Review*, 108(7), 1737–1772.

33. Abraham, K. G. (1987). Help-Wanted Advertising, Job Vacancies, and Unemployment. *Brookings Papers on Economic Activity*, 18(1), 207–248.

34. Carnevale, A. P., Jayasundera, T., & Repnikov, D. (2014). Understanding Online Job Ads Data. *Georgetown University, Centre on Education and the Workforce.*

35. Taken from employment number from the seasonally adjusted July 2019 Labour Force Survey

36. As before, the share used here is restricted to the share of job advertisements advertised on "online job boards" which is a lower bound estimate of the share of job ads advertised online overall.

37. Following Cortes, Jaimovich, Siu (2018), we can theoretically decompose the change in the number of job ads captured into those component parts. However, we lack reliable sources of data on both the probability that a job vacancy for a particular occupation is advertised online (due to data quality issues associated with the JVWS), in addition to the JVWS not being available between 2012 and 2014. As a result, we refrain from interpreting temporal changes in the number of job ads in Burning Glass.

38. Likely representing the velocity within these occupations, especially when this insight is combined from the under-representation of Burning Glass posting compared to JVWS for this occupation.

39. For this regression, we also restrict the skills we test into skills that were mentioned at least 18 times in all job postings - this allows for reduction of bias in considering very rare skills that may have a biased estimate of ranking. 18 times was chosen as it is the first quadrant of the distribution of number of times skills are mentioned.

40. Digital skill as used here includes both software and the base tech skills.

41. The conditional probability of a skill being digital conditional on it being software is 1, as we defined digital as being inclusive of both software and base tech skills.

42. For exactness, we found the point at which the third derivative of the likelihood function is 0.