Research Note

The Stanford IT Tables: A Suite of Firm and Industry Metrics for Technology Use

Erik Brynjolfsson, Wang Jin, and Sebastian Steffen

WORKING PAPER - JULY, 2022

Abstract: In light of the increased prevalence of new information technologies, such as cloud computing and machine learning, traditional IT measures based on physical IT capital have become less reliable, while IT complementary skills, tools, and human capital have become the new bottleneck. New IT technologies have thus, somewhat paradoxically, made the measurement of industry and firm-level IT sophistication and productivity significantly harder than they already were. We, therefore, build a novel set of industry-level IT metrics based on demands for IT skills and occupations in job postings from 2010 until 2022. By making these data publicly available to the IT research community, we believe that we can breathe new life into research using IT metrics to address various research questions. Our methodology to define these measures is general and simple enough to allow for future, and backward-compatible, extensions. We plan to build and release future versions in correspondence with the IT community. Strong correlations with the 'official' productivity measures validate our approach at the industry level and suggest their usefulness at the firm level, where no official measures for the US economy currently exist.

1 Introduction

Information technologies (IT) are among the most important and productive technologies of our era and have fundamentally transformed how firms operate. A large body of academic work has connected the adoption of such technologies to improved firm productivity (Tambe and Hitt 2012), customer reach, stock market returns, and supply chain management, among many other aspects. IT is so ubiquitous across industry sectors that it has been argued to be a general purpose technology (GPT) (Bresnahan and Trajtenberg 1995), in particular technologies such as machine learning (Brynjolfsson et al. 2018), data-driven decision-making (Brynjolfsson and McElheran 2016) and cloud computing (Jin and McElheran 2017).

However, measuring the 'technological stack' of firms is very challenging - never mind causally identify the impact of IT on firm productivity. One major challenge is that IT requires complementary physical and human capital (Brynjolfsson and Milgrom 2012), as well as organizational changes and managerial practices (Bloom et al. 2019), in order to be used productively.

In the past, counting physical IT equipment, such as PCs or central processor computing capacity, served as a reasonable proxy for firm-level IT data (Brynjolfsson et al. 2002, Bresnahan et al. 2002). With the rise of machine learning and computing, these types of measures break down and may even indicate a lack of IT sophistication as firms rent software (SaaS), computer infrastructure (IaaS), or nearly anything else (XaaS) as a service. If a profitable firm can rent nearly anything, what is left to be measured? We argue that human capital, in particular 'high-skill' human capital in IT, has become a crucial factor for a firms' success and measured based on it can serve as a reasonable proxy for firms' IT sophistication.

The inability to properly count (physical) IT capital or to account for the vast quality improvements in such capital over time are likely a large factor in the 'IT productivity puzzle', i.e. why IT does not show up in any productivity or GDP measures.

But there is a second accounting problem. IT is commonly used to provide free services, for example as platforms for communication, advertising or purchasing. A price of zero means no (direct) profit and therefore leads to these services not showing up in GDP measurements.[1]

[1]See Brynjolfsson et al. (2019) for an in-depth discussion and a proposed alternative,

One reason why IT productivity increases may not have shown up in official statistics is that there may be significant lags in adoption or unlocking the technology's potential. Previous GPTs, such as electricity, required a decades-long learning and adjustment period before manufacturers were able to take significant advantage of them (Brynjolfsson and McAfee 2014). Could it be the same for recent information technologies? Given that there are significant IT productivity differential across countries and industries, learning and adjusting seem to play an important role (Bloom et al. 2012).

But, in many ways, recent information technologies are different from previous innovations (Agrawal et al. 2019). These technologies have achieved super-human performance on a wide range of (albeit usually quite narrow and domain-specific) tasks and enable superstar effects that can lead to indirect displacements of workers. Routine-biased technological change suggests that routine tasks will be most easily and quickly replaced, though the large cost savings associated with displacing more expensive and complex tasks and workers implies higher returns for investing in IT that can displace 'high-skill' work.

Another novel aspect of IT is that its adoption is focused on stealing market share from competitors and may not increase overall profits. While improvement for selfish reasons and competition is in the vein of Adam Smith's 'invisible hand', IT may have significant negative externalities that turn competition into a zero-sum game instead of a win-win. For example, many IT innovations are aimed at grabbing consumers attention and time and lead to distraction and reduced worker productivity. IT may therefore be relatively less productive compared to past innovations, such as electrification, and infrastructure developments.

Part of the difficulty reflects the inherent nature of the technology: it is rapidly changing, at a pace virtually unmatched by earlier technologies, and much of its value and impact derives from complementary, often intangible investments in skills, business processes, and organizational capital. In the age of AI and cheap and easily accessible cloud-based services for anything, also referred to as XaaS to encompass software (SaaS), platforms (PaaS), and infrastructure (IaaS) as a service, the physical number of computers that a firm or an industry owns are no longer (and perhaps never were) a meaningful measure of IT sophistication (Brynjolfsson et al. 2002). Even in 1987 Robert Solow summarized that 'you can see the computer age everywhere

GDP-B.

except in the productivity statistics'. The rise of and productivity-enhancing effects of intangible IT have been documented through indirect measurement exercises of firm productivity residuals (Brynjolfsson et al. 2002, Bartel et al. 2007, Brynjolfsson and McElheran 2016) as well as market-based estimates of *Tobin's q*.

2 Data & Methodology

We create three different types of industry-level IT metrics, which we describe in more detail in this section. The first set of metrics is directly calculated from occupational demands of U.S. online job postings. This data, which we acquire from Burning Glass Technologies (BGT), consists of roughly 200 million job postings and contains information on each job posting's occupational code (based on the Standard Occupational Classification (SOC) system), industry code (based on the North American Industry Classification System (NAICS)), posting date, skill demands, and firm, among many others. We then aggregate the job postings' occupational demands to the industry-level by taking a weighted average, with the weights being defined by the industry-level shares of job postings with a corresponding occupational code in a given year.

The second set of metrics is similarly derived from BGT, but leverages the skill demands instead of the occupational demands associated with each job postings. The BGT data relies on a skill taxonomy which contains over 16,000 skills. These are mapped into roughly 800 *skill clusters*, which themselves are mapped into 27 *skill cluster families*. For example, the skill Python falls into the skill cluster 'Coding and Scripting Languages', which itself falls into the skill cluster family 'Information Technologies'. The 'Information Technologies' skill cluster family defined by the BGT taxonomy naturally lends itself as a definition to identify information technology skills within job postings, and thus industries.

Some additional care is required to count and deal with skills, as some firms and industries appear to engage in strategic signaling and demand significantly more skills for some job postings than can reasonably be supplied be job applicants.[2]

[2]Some job postings demand up to 400 skills according to BGT. This highlights a potential weakness of these data: it does not contain information on how important or how frequently used each skill will be for the actual job. Employers may also list substitute

The third, and final, set of metrics is derived by combining the BGT data with occupation-level metrics defined by prior literature, such as the Suitability for Machine Learning (SML) (Brynjolfsson et al. 2018) or the Work-from-Home (WFH) Feasibility (Dingel and Neiman 2020) of occupations, among others. Using the number of job postings for each SOC code demanded by each industry, we are then able to aggregate these occupation-level metrics to industry-level metrics using a weighted average.

Before explaining these three sets of metrics, we give an overview of the underlying job postings data.

2.1 Millions of Job Postings from Burning Glass Technologies (BGT)

We obtain detailed, annotated job postings data from Burning Glass Technologies (BGT) - a high-quality data source with comprehensive coverage of job posting portals beginning in 2010 and with increasing popularity in recent economic research (Hershbein and Kahn 2018, Deming and Noray 2020, Azar et al. 2020, Das et al. 2020, Acemoglu et al. 2020).[3] The BGT data covers about 200 million online job vacancy postings posted on over 40,000 distinct online job platforms in the United States between 2010 and 2020 and arguably covers the near-universe of job postings (Figure 1). Each vacancy posting is parsed, de-duplicated, and annotated with the posting date, the SOC occupational code, the NAICS industry code, and which skills were demanded among several other variables.

Using these data we can create simple, yet insightful (and computationally-expensive) aggregations. For this study we focus on industries as defined by the 2-digit NAICS codes. We classify each job posting according to different categorical definitions described below and then, for each industry, calculate its share of all job postings that fall within each category.

One particular issue we attempt to address is the considerable bias towards 'popular' skills such as AI - skills demanded in job postings may not fully reflect actual skill demands, in particular if firms engage in strategic signaling behavior. Since some job postings request over 100 different skills,

skills along with the skill they actually require - for example, they may demand R even though their codebase relies on Python.

[3]BGT provides partial coverage for 2007, but does not provide any coverage of job postings in 2008 or 2009. We therefore use BGT data starting from 2010.

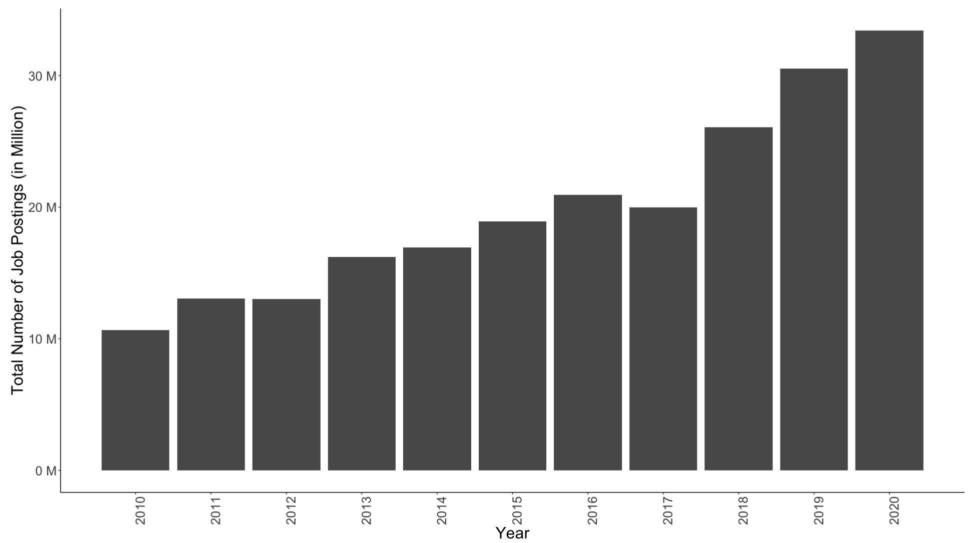

Figure 1: Number of Job Postings per Year (2010-2020) captured by Burning Glass Technologies (BGT)

we downweight

To do so we first count job postings in different ways, based on their classification into different SOC codes or their demand for different skills, and then aggregate to the industry level.

2.2 Metrics derived from Job Postings' Occupational Codes

Using the SOC taxonomy, we derive the following occupation-based IT metrics:

Definition 1 ***All Computer Occupations (IT-Occ-1)*** *All job postings whose annotated SOC code begins with 15 ('Computer and Mathematical Occupations').*

Definition 2 ***Cybersecurity Occupations (IT-Occ-2)*** *All job postings whose annotated SOC code falls within the following: 15-1122 (Information Security Analysts), 15-1121 (Computer Systems Analysts), 15-1152 (Computer Network Support Specialists), 15-1141 (Database Administrators), 15-1142 (Network and Computer Systems Administrators), 15-1143 (Computer Network Architects).*

While many other definitions are possible, all are ultimately subjective judgment calls. For now we focus on these more basic definitions and plan to study additional ones in the future.

Equipped with these occupation-based definition, we derive the share of each 2-digit industry's job postings that fall within each category:

$$\text{it}(i,t) = \frac{\sum_{p_j} \mathbb{1}[\text{soc}(p_j) \in \{\text{IT Def}\}, \text{ind}(p_j) = i, \text{year}(p_j) = t]}{\sum_{p_j} \mathbb{1}[\text{ind}(p_j) = i, \text{year}(p_j) = t]} \tag{1}$$

, where p_j is the jth job posting, $t \in \mathcal{T} \equiv \{2010, \ldots, 2020\}$ and industry $i \in \mathcal{I} \equiv \{11, 21, 22, 23, 31{-}33, 42, 44{-}45, 48{-}49, 51, 52, 53, 54, 55, 56, 61, 62, 71, 72, 81, 92\}$, as defined by NAICS.

$$\text{it}(i,t) = \frac{\sum_{p_j} \mathbb{1}[\text{soc}(p_j) \in \{\text{IT Def}\}, \text{firm}(p_j) = i, \text{year}(p_j) = t]}{\sum_{p_j} \mathbb{1}[\text{firm}(p_j) = i, \text{year}(p_j) = t]} \tag{2}$$

2.3 Metrics derived from Job Postings' Skill Demands

The BGT skills data is annotated via BGT's industry-leading skill parser, which identifies skills in the job posting's text and maps them into a detailed skills taxonomy. The skill taxonomy covers over 16,000 *skills*, which are grouped into over 600 *skill clusters*, which themselves are further grouped into 28 *skill cluster families*.

For example, *Python* is a skill within the *Scripting Languages* skill cluster, which itself falls into the *Information Technology* skill cluster family. The taxonomy was initially assembled from online resumes and is continuously updated through client feedback, research, and forums. K-Means clustering along with additional qualitative checks were employed to create meaningful skill clusters. Whenever new skills are added to the taxonomy, the labels are refit to the entire history of job postings data, which minimizes potential biases which may arise through BGT's time-varying ability to capture new skills.

Notably, the BGT taxonomy is significantly more detailed than other skill taxonomies, and thus allows micro-level insights into which *actual* skills employers demand.[4] The O*Net skill taxonomy, which contains 2 levels, with 35 skills grouped into 6 skill groups, is unable to provide such detail.

[4]While the distinction between skill, ability, knowledge, tool, work context, or experience is not entirely clear, we believe that the BGT taxonomy gets closest to the layman

The methodology to derive industry-level IT measures based on job-posting-level skill demands is similar to the that of the previous section but requires an additional aggregation step. Specifically, we first create occupation-level skill demands as in Brynjolfsson and Steffen (2021) This aggregation accounts for the large heterogeneity in the number of skills that each job posting demands. This way, job postings that demand a very large number of skills do not outweigh job postings that demand fewer skills, which reduces potential biases due to job postings that may cast too wide of a net. Thus for each job posting, p_j, we start with the binary BGT skill vectors over the entire set of skills, $S = S_1, \ldots, S_K$, which we call $S(p_j) = (\mathbb{1}[S_1], \ldots, \mathbb{1}[S_K])$, and normalize them to derive the weighted skill demands of each job posting:

$$s(p_j) = \left(\frac{S_1(p_j)}{\sum_k S_k(p_j)}, \ldots, \frac{S_K(p_j)}{\sum_k S_k(p_j)} \right) \tag{3}$$

To put it in words, for our preferred skill-based measure, we first derive a weighted skill demand vector for each job posting. Before we can aggregate these job-posting-level skill demands to the industry-year level, just like before, we define the following skill-based IT metrics:

Definition 3 ***Information Technology Skills (IT-Skill-1)*** *The share of all job postings' (weighted) skill demands, which fall into the skill cluster family 'Information Technology'.*

Definition 4 ***Cloud Skills (IT-Skill-2)*** *The share of all job postings' (weighted) skill demands, which fall into the following skill clusters: 'Cloud Solutions', 'Cloud Storage', 'Cloud Computing', 'Big Data', 'Data Warehousing', 'Database Management Systems', 'NoSQL Databases', 'Data Wrangling', 'Graph Databases', 'Remote Desktop Software'.*

Definition 5 ***Cybersecurity Skills (IT-Skill-3)*** *The share of all job postings' (weighted) skill demands, which fall into the following skill clusters: 'Cybersecurity', 'Network Security', 'Technical Support', 'Database Administration', 'Data Management', 'Information Security', 'Application Security', 'Internet Security'.*

---

understanding of skill and certainly closer than education or wage, which were commonly used in the past due to data issues

Again, there are many other possible, subjective definitions - we also follow **?** for additional definitions for Database, Computer, Data Analysis, and ML & AI skills.

Equipped with these different definitions of IT metrics, just like before, we aggregate from the job-posting to the corresponding industry-year level, for each (i, t) combination, by summing the relevant skills:

$$it(i,t) = \frac{\sum_{p_j} \sum_{k \in \text{IT Deg}} s_k(p_j) \mathbb{1}[\text{ind}(p_j) = i, \text{year}(p_j) = t]}{\sum_{p_j} \sum_{k} s_k(p_j) \mathbb{1}[\text{ind}(p_j) = i, \text{year}(p_j) = t]} \tag{4}$$

occupation-level skill demand vector by taking a weighted average over the skill demand vectors of the corresponding job postings, by weighting each job posting's skill demands by the total number of skills that each job posting demands. These occupational skill demand vectors, $\vec{s}(o_i) = (s_1(o_i), \ldots, s_K(o_i))$ measure the total (weighted) demand for each skill, and could still be driven by the large, heterogeneous increase in the number of job postings over time. We therefore normalize these vectors to sum to 1 to derive our final aggregate measure of average occupational skill demands:

$$\vec{s}(o_i) = \frac{1}{\sum_{k=1}^{K} s_k(o_i)} (s_1(o_i), \ldots, s_K(o_i)) \tag{Norm. Skill Demand Vec. for Occ. i}$$

2.4 Metrics derived from Scholarly Work

Our final set of IT metrics leverages measures defined by prior scholarly work. Specifically, we take the Work-from-Home (WFH) scores from Bai, Brynjolfsson, Chi, Jin, Steffen, which are based on Dingel and Neiman (2020); the Suitability for Machine Learning (SML) scores from Brynjolfsson et al. (2018) ; the abstract, routine, and manual task scores from Autor and Dorn (2013), and the Artificial Intelligence (AI), Software, and Robot technology scores from *Webb2020*. All of these scores are defined at the occupational, 6-digit SOC level, which allows us to merge them with our industry-year-level panel of occupational counts. To aggregate to the industry-year level we then take a weighted average of the occupation-level scores, where the weights are defined by the occupational counts.

Going forward, we plan on running the same metrics we created at the industry level, for the firm-level as well. We already built a prototype and continue to improve it by updating the firm-name matching across the BGT and Compustat data. We also validated our Industry-level metrics against 'official' productivity numbers from the BLS and reached high correlations across several measures. In the future, our metrics will be publicly available on http://digitaleconomy.stanford.edu/ITTables.

3 Conclusions

In light of the increased prevalence of new information technologies, such as cloud computing and machine learning, traditional IT metrics based on physical IT capital have become less reliable. The measurement of firm's IT sophistication and IT productivity has thus become even harder than it already was. But perhaps it has also become less meaningful. IT-complementary skills and human capital have become more important as well and we argue that firms' demands for them is an important component of firms' IT sophistication. Therefore, in this research note we leverage large and granular data from online job postings to derive novel industry-, and firm-level IT metrics. We plan to make these metrics publicly available[5] as we believe that they will be of significant value to shed light on the IT productivity paradox, firm productivity, and a variety of related important questions in IT. Our methodology to define these measures is general and simple enough to allow for future, and backwards-compatible, extensions and we plan to build and release future versions in correspondence with the academic community.

References

Acemoglu, D., Autor, D., Hazell, J., and Restrepo, P. (2020). AI and Jobs: Evidence from Online Vacancies. *National Bureau of Economic Research Working Paper Series*, No. 28257.

Agrawal, A., Gans, J. S., and Goldfarb, A. (2019). Artificial Intelligence: The Ambiguous Labor Market Impact of Automating Prediction. *Journal of Economic Perspectives*, 33(2):31–50.

Autor, D. H. and Dorn, D. (2013). The growth of low-skill service jobs and the polarization of the US Labor Market. *American Economic Review*, 103(5):1553–1597.

[5]The latest version can be found under http://digitaleconomy.stanford.edu/ITTables.

Azar, J., Marinescu, I., Steinbaum, M., and Taska, B. (2020). Concentration in US labor markets: Evidence from online vacancy data. *Labour Economics*, 66.

Bartel, A., Ichniowski, C., and Shaw, K. (2007). How Does Information Technology Affect Productivity? Plant-Level Comparisons of Product Innovation, Process Improvement, and Worker Skills. *Quarterly Journal of Economics*, 122(4):1721–1758.

Bloom, N., Brynjolfsson, E., Foster, L., Jarmin, R., Patnaik, M., Saporta-Eksten, I., and Reenen, J. V. (2019). What Drives Differences in Management Practices? *American Economic Review*, 109(5):1648–1683.

Bloom, N., Sadun, R., and Van Reenen, J. (2012). Americans do IT better: US multinationals and the productivity miracle.

Bresnahan, T. F., Brynjolfsson, E., and Hitt, L. M. (2002). Information Technology, Workplace Organization, and the Demand for Skilled Labor: Firm-Level Evidence. *The Quarterly Journal of Economics*, 117(1):339–376.

Bresnahan, T. F. and Trajtenberg, M. (1995). General purpose technologies 'Engines of growth'? *Journal of Econometrics*, 65(1):83–108.

Brynjolfsson, E., Collis, A., Diewert, W. E., Eggers, F., and Fox, K. J. (2019). GDP-B: Accounting for the Value of New and Free Goods in the Digital Economy. *National Bureau of Economic Research Working Paper Series*, No. 25695.

Brynjolfsson, E., Hitt, L. M. L. M., and Yang, S. (2002). Intangible Assets: Computers and Organizational Capital. *Brookings Papers on Economic Activity*, 2002(1):137–198.

Brynjolfsson, E. and McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. WW Norton & Company.

Brynjolfsson, E. and McElheran, K. (2016). The Rapid Adoption of Data-Driven Decision-Making. *American Economic Review*, 106(5):133–139.

Brynjolfsson, E. and Milgrom, P. (2012). Complementarity in Organizations. In Gibbons, R. and Roberts, J., editors, *The Handbook of Organizational Economics*, chapter 1, pages 11–55. Princeton University Press.

Brynjolfsson, E., Mitchell, T., and Rock, D. (2018). What Can Machines Learn and What Does It Mean for Occupations and the Economy? *AEA Papers and Proceedings*, 108:43–47.

Brynjolfsson, E. and Steffen, S. (2021). Occupational Skill Compositions and the Values of Skills: Panel Estimates from 200 million Job Postings. *Working Paper*.

Das, S., Steffen, S., Clarke, W., Reddy, P., Brynjolfsson, E., and Fleming, M. (2020). Learning Occupational Task-Shares Dynamics for the Future of Work. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pages 36–42, New York, NY, USA. Association for Computing Machinery.

Deming, D. J. and Noray, K. (2020). Earnings Dynamics, Changing Job Skills, and STEM Careers*. *The Quarterly Journal of Economics*, 135(4):1965–2005.

Dingel, J. I. and Neiman, B. (2020). How many jobs can be done at home? *Journal of Public Economics*, 189.

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

Jin, W. and McElheran, K. (2017). Economies Before Scale: Survival and Performance of Young Plants in the Age of Cloud Computing. *SSRN Electronic Journal*, pages 1–74.

Tambe, P. and Hitt, L. M. (2012). The Productivity of Information Technology Investments: New Evidence from IT Labor Data. *Information Systems Research*, 23(3):599–617.

A Tables

Please see http://digitaleconomy.stanford.edu/ITTables for the latest version of these data.