Measuring Professional Skill Development in U.S. Cities Using Internet Search Queries

Shagun Jhaver
Georgia Tech
jhaver.shagun@gatech.edu

Justin Cranshaw
Microsoft Research
justincr@microsoft.com

Scott Counts
Microsoft Research
counts@microsoft.com

Abstract

Using a sample of 10 million skill development-related queries from a popular internet search engine, drawn from almost 400 U.S. cities and over a 5 year period (2012 - 2016), we characterize the skills people search for across these cities, and relate them to measures of economic output, including GDP and unemployment. Findings show that differences in the amount, type, and specialization of skill development searches distinguish economically thriving cities from those that are struggling, including explaining variance in future GDP growth after accounting for socioeconomic demographics such as income and poverty levels. Overall, internet-based skill development appears to contribute to rich cities getting richer, with initially well-off cities seeing more and more specialized skill-related searching that in turn aligns with future economic growth.

Introduction

Ongoing, self-directed development of skills for professional advancement has become a hallmark of contemporary employment, and increasingly is critical to success in the current economy (Scholarios et al. 2008). In the aggregate, the amount and manner of such skill development behavior may align with different economic conditions and output levels across geographies. People in hotbed technology cities such as Austin and Seattle, for instance, may be developing more "new economy" skills than people in cities dominated by traditional industries. Over time, such differences could contribute to an economic divide between large segments of the population. In this paper, we target the question of (in)equality in the use of the internet to develop skills: Is internet-based skill development an equalizing force for economic opportunity or are there concentrations of people developing skills that yield greater economic output, with corresponding groups missing out on the opportunities to learn high value skills?

We focus on online skill development given its ubiquity, with options ranging from formal online universities to boot camp style courses to online tutorials and technical references. Online resources can even serve simply as a mechanism to research and locate offline employment related credentialing programs. By taking advantage of these resources, individuals can build expertise and improve employment prospects, and in the aggregate, internet-based skill development should contribute positively to the collective productivity of communities of people.

Those using the internet to develop or otherwise learn about skills are likely to do so, at least in part, through a search engine. Search engines are a primary tool for information seeking online (Purcell, Rainie, and Brenner 2012), including for tasks related to skill development such as job searching (Kuhn and Mansour 2014), and thus skill related search queries should capture a considerable portion of skill development intent and behavior. In the current work, we utilize a sample of more than 10 million skill-related search queries from Bing, a search engine with significant share in the U.S. internet search marketplace (comScore 2016). This sample of skill-related search query data is based on the U.S. Department of Labor's categorization of thousands of different skills and certifications.

We align these searches for skills with macroeconomic measures such as GDP and unemployment over time and over geography. This allows us to examine skill searching as both a leading and lagging indicator, and as a metric for comparison across regions. Geographically, we do so at Metropolitan Statistical Area (MSA) scale, which in the U.S. includes almost 90% of the population (Bureau 2018). Our search query sample is drawn from every MSA in the United States and over a time period of 5 years, from 2012 - 2016. Generally we see that tech, business, and medical skill searches dominate, and that when aligned with city economic data, the per-capita number of skill searches regardless of skill type correlates at .46 with GDP per capita and at -.25 with the per-city unemployment rate, suggesting that internet-based skill acquisition and development generally aligns with a healthy economy.

Not all cities are equally economically healthy, however. Since the 2008 economic recession, recovery trends in the US have shown increasing growth gaps between prosperous and distressed regions, raising concerns over growing economic disparities (Group 2017; Kneebone 2014). We use our sample of internet-based skill acquisition searches to address the primary question of whether, on balance, the internet serves as a skill equalizer. Our findings indicate that already privileged cities see more skill searching that in turn is associated with greater economic growth, supporting a rich

get richer effect of growing inequality in economically valuable skills in American cities.

We refine this result in two ways. First, we examine the role of core city demographics to show that our measure of internet-based skill searching explains variance in economic output growth among cities beyond factors like population size and unemployment levels. Second, we create a metric of specialization in skill searching that can be used to identify the economically strongest cities. People in these cities do search for the same skills searched for in all cities across the country, but they also search for more unique skills that differentiate these cities in terms of future economic growth.

In summary, we utilized search queries for skill development to characterize differences in the amount and type of skills likely being developed in different cities in the U.S., and then quantified how those differences align with measures of economic value and growth. The primary finding is that already well-off cities see more as well as more specialized searches for skill development, which in turn accounts for significant variance in future economic growth.

Related Work

Skill Development

Continuous and Self-directed Skill Development Employability and success in today's economy depend on the ability to learn continuously in order to maintain flexibility for evolving job demands and to demonstrate the capacity to acquire skills in varied organizational contexts (Scholarios et al. 2008). In a seminal study on the development of workers in organizational settings, Hall and Mirvis noted a shift from the organizational career to the protean career, "a career based on self-direction in the pursuit of psychological success in one's own work" (Hall and Mirvis 1995). They suggested that this person-centered emphasis on employability coincides with the notable shift in responsibility for career management and advancement from employers to employees. Hall and Mirvis also argued that contemporary high-paced work environments call for a new view of career stages that includes many cycles of learning stages or continuous learning, "rather than a single lifelong career stage cycle" (Hall and Mirvis 1995). Increasingly, it is now workers' responsibility to acquire new skills, knowledge, certifications, abilities and other characteristics that are valued by current and prospective employers (Fugate, Kinicki, and Ashforth 2004). Internet search engines provide an important entry for individuals to develop new skills on their own. Therefore, studying how search engines are used for skill development can highlight major trends in evolving characteristics of the labor force.

Skill Development in Light of Rapid Technological Changes Recent economic trends and forecasts further highlight the importance of continuous skill development, especially for individuals in developed countries. A recent examination of potential labor market disruptions from automation by the McKinsey Global Institute predicts enormous workforce transitions in the years ahead, estimating that in the event of rapid automation adoption, by 2030, "as many as 375 million workers globally (14 percent of the global workforce) will likely need to transition to new occupational categories and learn new skills (Manyika et al. 2017)," with other estimates putting that number even higher (Frey and Osborne 2017). Brynjolfsson and McAfee predict that as technology continues to replace low-skilled workers, wages for jobs with those skills will reduce and demand for high-skilled workers will increase (Brynjolfsson and McAfee 2014).

These papers highlight the importance of skill development in the context of rapid technological changes. At the same time, some researchers have raised concerns over the ability of US education and job training system to produce workers with skills that are relevant for future jobs (David 2015; Goldin and Katz 2008). By studying trends in skill development, our work could help shed light on potential gaps in labor force advancement.

Learning through Online Resources Prior research has explored the role of online resources in facilitating learning experiences (DiSalvo, Khanipour Roshan, and Morrison 2016; Greenhow, Robelia, and Hughes 2009; Reich, Murnane, and Willett 2012). Empirical evidence suggests that inequities in usage of online tools for learning across socioeconomic divides has contributed to growing disparity in educational outcomes in the US. Reich et al. studied the use of wiki learning environments in schools, finding that wiki use in schools with more affluent students persisted longer, providing more opportunities for skill development than wikis in schools with less affluent students (Reich, Murnane, and Willett 2012). Disalvo et al. investigated how parents use online social networks to find learning opportunities for their children, and found that parents with lower socio-economic status face greater challenges in using social networks for this task (DiSalvo, Khanipour Roshan, and Morrison 2016). Attewell characterized the gap between how learners from different social classes use technologies in different ways, calling it the "second digital divide" (Attewell 2001). We add to this literature by exploring how individuals in cities with varying economic strengths use search engines for skill development differently.

Growing Economic Disparities in the US

Recent economic trends emphasize growing economic inequalities in different parts of the US. A recent report by Economic Innovation Group noted that the US economy contains a fragmented landscape of economic well-being with stark growth gaps between prosperous and distressed zip codes (Group 2017). Another report on economic recovery from the 2008 recession by the same group found that employment gains from 2010 to 2014 were far more geographically concentrated than in previous recoveries (Group 2016). In a similar vein, Kneebone found that over the last decade, poverty has become more concentrated in distressed neighborhoods affecting the ability of those areas to grow in inclusive and sustainable ways (Kneebone 2014).

This literature suggests an economic future in which growing geographic disparities and diminished business dynamism would become increasingly urgent concerns (Kneebone 2014; Group 2016). At the same time, prior research

has shown that economic inequality and inequitable access to goods increase health inequalities, impede productivity and economic growth, reduce opportunities for social cohesion, and provoke antisocial behavior, violence and homicides (Braithwaite 1976; Kawachi and Kennedy 1997; Wilkinson 2002; Wilson and Daly 1997). Therefore, it is crucial to study local ecosystems of employability, skill development, entrepreneurship and investments throughout the US, so that we can develop policies and economic solutions that allow places on the wrong side of the trajectory of economic growth to cultivate competitive advantages. Our work examines the role online skill development plays in bridging or perpetuating these disparities.

Related to the concept of growing economic disparities is the Matthew effect of accumulated advantage (Kilpi-Jakonen et al. 2014), or the "rich-get-richer (poor-get-poorer)" phenomenon where individuals who are already well-off have higher chances of success, thereby exacerbating preexisting disparities. Prior research has shown this effect to be a contributing factor for inequalities in a variety of contexts such as participation in sports (Söderström et al. 2018), adult education (Kundu and Matthews 2019), funding in academic research (Kundu and Matthews 2019) and wealth (Burda, Wojcieszak, and Zuchniak 2018). However, as far as we know, the Matthew effect has not been evaluated as a contributing factor for skill acquisition in prior research. Our analysis evaluates whether there is a "rich-get-richer" phenomenon at play in skill acquisitions on a macro scale.

Societal-Scale Measurement with Search Data

The near-ubiquitous use of online platforms such as social media websites, search engines and recommendation sites has created a new source of data about people and the world. As individuals use these platforms for communication, information seeking, self-presentation and promotion (DiMicco et al. 2008), they leave behind a wide variety of rich digital traces known as *social data* (Olteanu et al. 2016). These online social data provide information about users' lives and interests at a scale and level of detail that were impractical to attain with conventional data collection techniques like surveys and user studies (Olteanu et al. 2016; Richardson 2008). Today, companies, researchers and governmental and non-governmental organizations are using these social data to create new products and services as well as make policy decisions (Olteanu et al. 2016). Social data has also led to significant progress in many areas of computing such as object recognition (Gao et al. 2013), online activism (De Choudhury et al. 2016; Jhaver, Chan, and Bruckman 2018), crisis informatics (Reuter and Kaufhold 2018), digital health (Yom-Tov 2016), job satisfaction (Hickman et al. 2019) and computational social science (Gilbert 2010; Saha, Weber, and De Choudhury 2018).

Search query logs are a category of social data that have been frequently used for understanding human behaviors and conditions. In Human-Computer Interaction (HCI), search data is most commonly employed to study health and wellness (De Choudhury, Morris, and White 2014; Fourney, White, and Horvitz 2015; White et al. 2014). However, there is a growing economics literature that incorporates search query logs. Choi and Varian used data on Google searches to improve predictions for a range of economic time series (Choi and Varian 2012). A number of studies have shown that search query logs are associated with volatility and returns in the financial and commodity markets (Basistha, Kurov, and Wolfe 2017; Dimpfl and Jank 2016). Other researchers have used search query logs to predict unemployment rates (D'Amuri and Marcucci 2010; Askitas and Zimmermann 2009). Chancellor and Counts used internet search data to estimate employment demand in the US (Chancellor and Counts 2018).

We add to this rich literature by leveraging internet search data first to provide a detailed picture of skill searches across different cities, and then show how this metric can explain variance in the measurements of GDP and unemployment.

Data

Developing a Skills List

We begin by collecting a corpus of skills of interest from the CareerOneStop website (car 2015), a non-profit organization sponsored by the US Department of Labor to serve job seekers, workers and employers by providing them information and resources related to employment trends in the US. We use two datasets available on this website that offer a thorough characterization of the skills that are used in different occupations in the US:

1. **Technologies Dataset:** a collection of 8,717 common technologies and software used by workers in specific occupations.
2. **Certifications Dataset:** a collection of 8,589 professional certifications that workers may earn to show proficiency in an occupational skill or knowledge.

We combine these two datasets into a single list of skills, which we use to identify skill development search queries. Figure 1 shows examples of commonly searched for skills in our data. Note that despite its name, the Technologies Dataset contains many skills such as typing and project management that are not explicitly related to the tech industry.

Identifying Skill Search Queries

Starting with US-based, English language search queries from a popular search engine between 2012 and 2016, we extracted a corpus of skill development queries by selecting queries that met both of the following two conditions:

1. They contain at least one of the words *course(s)*, *tutorial(s)*, *certificate(s)*, and *certification(s)*, which signal an interest in skill development.
2. They contain at least one skill from our skills list described above. We looked for an exact match for each skill in the queries.

Our list of skill development intent words was iteratively developed to ensure the final data collection maintained high precision. Other words like *program(s)*, *school(s)*, *department(s)*, and *degree* were considered, but resulted in too many false positive queries not actually about skill development. Additional data cleaning on the skills list was performed to improve data quality. We removed skills that had

Example Skill Search Queries
javascript video tutorials; blender tutorials; medical assistant certification; social media marketing course; google spreadsheet tutorial; hvac certification; cpa review courses; android development tutorial; human resources certificate; estate planning courses; how do i get a typing certificate; online biochemistry course; dental assistant courses; adobe photoshop 7.0 tutorials; small business management course

Table 1: Common skill development search queries.

a matching homonym with another commonly used English language word (e.g., *cat*, and *basic*), or any 2-letter state acronyms (e.g., *CT*, and *NC*), again in order to minimize false positives. Similarly, we removed terms like *youtube* and *instagram*, which can be considered skills and were on our initial skills list, but were also very commonly searched for terms. Finally, we removed search queries with URLs. This data collection and filtering resulted in 10,944,824 skills-related search queries matching to 6,990 skills. Table 1 provides examples of search queries occurring frequently (100 times or more) in the data sample.

Evaluation of Data Quality In order to evaluate the quality of our search queries data, two of the authors independently rated a randomly generated sample of 500 search queries. For each query, the coders labelled whether or not that query indicated an interest in developing a new skill. The two coders showed agreement in 95% of the cases, with Cohen's kappa of 0.366. 490 out of 500 queries (98%) were coded by at least one coder as showing as interest in developing new skills and 457 queries (91%) were coded by both coders as skill-development related. These results indicate that search queries in our data are related to skill development with high precision.

Mapping of Search Queries to MSAs We aggregated our search query data and conducted all analyses at the Metropolitan Statistical Area (MSA) level. We selected this geographic unit because urban areas are important units of national economic activity, and they exhibit considerable quantitative and qualitative variability in their local economies (Cohen and Simet 2018). Additionally, many US federal government agencies like the Bureau of Economic Analysis and Bureau of Labor Statistics collect and provide statistical information about economic health of MSAs, and thus we could align with government-based measures of economic output. Given the almost 400 MSAs in the US, this aggregation, as opposed to state level, also allowed us to account for appropriate levels of granularity in the interactions between the search query and economic activity measures.

Each of our search queries was tagged with a latitude and longitude, generally accurate to within about one mile. We mapped this location first to a US county and then to the MSA containing that county. MSAs are strictly defined by counties in that each MSA is comprised by only a set of counties and each county lies in only one MSA (BEA 2018). We excluded queries that originated in counties that were not part of any MSA. We note here that MSAs in the US show a lot of diversity in their degree of urbanization. For example, according to the 2017 US census data, populations across MSAs range from 55K (in Carson City, NV Metro area) to 19.5M (in New York-Newark-Jersey City Metro area) (Bureau 2018).

Ethics and Data Privacy This work was reviewed and approved by our institution's Ethics Review Board, which is also an IRB. We never obtained any form of user identification. All raw data contained only a date stamp, the query text, and a location stamp in the form of a latitude and longitude so that the skill mentioned in the query could be mapped to an MSA. Thus all data were anonymous. With the exception of randomly selecting a small number of very common queries to show as examples in Table 1, raw query text was discarded at the end of data collection and validation so that only the skills mentioned in the queries remained and were aggregated to yearly MSA-level for analysis (e.g., the number of 'dental assistant' skill queries from Los Angeles in 2015). All use and storage of any data was in agreement with the search engine's End User License Agreement and Privacy Policy, as well as GDPR regulations.

Economic Indicators

We collected the following data on the economic health of different MSAs between 2012 and 2016:

1. Unemployment Rate, *Bureau of Labor Statistics*[1]
2. GDP per capita, *Bureau of Economic Analysis*[2]
3. Population, *United States Census Bureau*[3]
4. Poverty, *United States Department of Agriculture*[4]

We attempted to investigate education also, however yearly metrics were unavailable for our period of study (ERS 2018).

These economic indicators were available over the time period for 373 US MSAs. We removed one (Sierra Vista-Douglas, AZ) due to outlying skill search query data, leaving 372 MSAs used throughout the analyses.

Findings

We break our findings into three sections. In the first section, we provide a general characterization of skills that are searched most frequently across US cities. This characterization provides a descriptive sense of the way skill searches vary between MSAs. The second section introduces our skill specialization metric and reports quantitative relationships between skill searches, demographics, and measures of economic health, such as GDP and unemployment rates. In the final section, we present evidence that skill searches are associated with both the prior health of cities and future economic growth.

[1] https://www.bls.gov/web/metro/laummtrk.htm

[2] https://www.bea.gov/regional/

[3] https://www.census.gov/en.html

[4] https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/

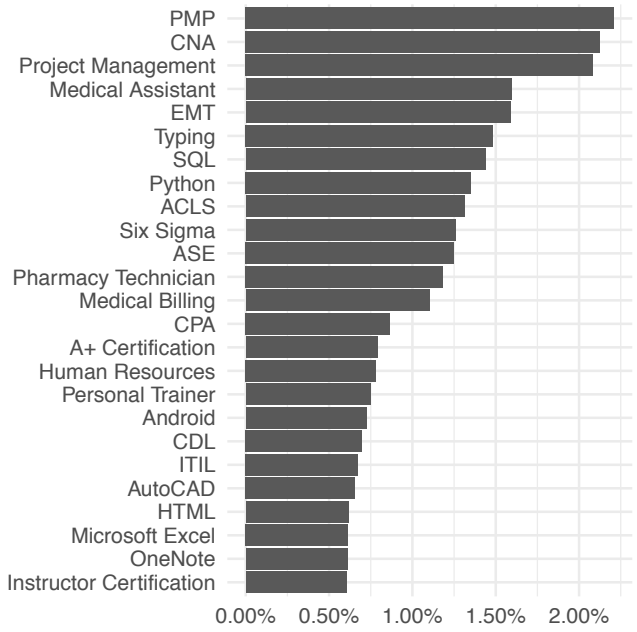

Figure 1: Top 25 most searched skills in our data.

Overview of Skill Searches in US Cities

We begin with a high-level descriptive overview of online skill searches in the United States. Figure 1 shows a breakdown of the 25 most searched for skills in our data, showing the head of a very long-tailed distribution, with the top 5 skills accounting for 10% of all searches. Nearly 70% of all skill searches in our data matched to a skill in the certifications dataset, with *Certified Nursing Assistant (CNA)*, *Project Management Professional (PMP)*, *Emergency Medical Technician (EMT)*, and *Pharmacy Tech.* receiving the most searches among all certifications. Among the remaining 30% of searches in our data that correspond to technology skills, business related searches, such as *Project Management*, *Human Resources*, and *Six Sigma*, and information technology related searches, like *SQL*, *Python*, and *Android* were popular.

Turning to differences across MSAs, Table 2 shows the top and bottom of a ranked list of MSAs based on GDP per capita, with examples of the top searched skills from each MSA. This ranking reveals some notable qualitative differences in the skill searching between the MSAs with highest and lowest GDP per capita. The top skills from the bottom 10 MSAs are overwhelmingly dominated by the medical professions, including *CNA (Certified Nursing Assistant)*, *Medical Billing*, *Medical Assistant*, and *Pharmacy Tech*, all of which appear much less frequently as top skills in MSAs with highest GDP. Similarly, *ASE (Automotive Service Excellence)* appears seven times among the top five skills in cities with the lowest GDP, while appearing only once in high GDP cities. On the other hand, skills searches in high GDP MSAs appear to have more information technology-related searches, like *SQL*, *Python*, and *Javascript*, and business related searches, like *Project Management* and *PMP (Project Management Professional)*, none of which appear as top skills in the bottom 10 MSAs. These differences suggest that the types of skills people seek to develop in a city may shed light on underlying socioeconomic factors determining the city's success.

Skill Searches and Socioeconomic Measures

Building on the descriptive differences in high and low GDP per capita MSAs, we turn to numerical relationships between skill searching and relevant economic measures. We introduce two measures, which we use to characterize the skill searches in MSAs in relation to their economic outcome measures. The first is the yearly volume of **skill searches per capita** per MSA. This is the sum of all the skill searches in each MSA per year, regardless of the skill being searched, divided by the population of the MSA in the given year. The second is a measure of **skill search specialization** to capture in a single index the degree to which people in each MSA searched for more unique, and presumably therefore more valuable, skills, relative to skill searches in other MSAs. Intuitively, we expect people everywhere to search for a core set of skills such as typing and many of the medical skills, while more specialized skills, particularly those in the tech sector, might signal investment in skill development that will pay off in the form of higher economic value in the future.

Skill Specialization Measure In order to quantify the specialization level of each skill, we designed a metric that could assess whether the searches for that skill are equally distributed across all cities or whether they are concentrated in one or few cities. To measure this property, for each skill $s \in S$ we first create a vector $V_s = (C_{s,1}, C_{s,2}, ..., C_{s,n})$, where the i^{th} component $C_{s,i}$ measures the number of skill searches mentioning skill s in city i. Next we define the specialization index of each skill s by calculating Gini coefficient of the above vector: $Specialization(s) = Gini(V_s)$. The Gini coefficient is a statistical measure of the dispersion of a distribution, on a scale of 0 to 1 (Cowell 2011). Intuitively, if a skill s is concentrated in one or a few cities, then $Specialization(s)$ will be high whereas if searches for s are spread evenly across all cities, $Specialization(s)$ will be low.

We can extend this measure to create an index of specialization for each city m as follows:

$$Specialization(m) = \sum_{s \in S} p_m(s) \cdot Specialization(s),$$

where $p_m(s) = \frac{C_{s,m}}{\sum_{s' \in S} C_{s',m}}$ is the proportion of searches for skill s in city m, relative to other skill searches in m. Thus, $Specialization$ index is a weighted average of the specialization scores for skills searched in a city, weighted by the probability mass of that skill in that city. This measure is again bounded between 0 and 1, with higher scoring cities devoting a greater percentage of their skill searching probability mass towards more specialized skills. Note that the $Specialization$ index is population relative to each city.

Relationship to Economic Measures The correlations in Table 3 reflect the degree to which our two measures of skill

GDP Rank	Search Rank	Spec. Rank	Top 10 MSAs in Per Capita GDP	Top Five Most Searched Skills
1	235	311	Midland, TX	CDL, CNA, Pharmacy Tech., Medical Asst., Typing
2	6	1	San Jose-Sunnyvale-Santa Clara, CA	Python, Android, SQL, Typing, Javascript
3	83	19	Bridgeport-Stamford-Norwalk, CT	EMT, CNA, SQL, PMP, Python
4	49	5	San Francisco-Oakland-Hayward, CA	Typing, Python, SQL, PMP, Project Management
5	48	25	Boston-Cambridge-Newton, MA-NH	EMT, CNA, SQL, PMP, Project Management
6	1	12	Seattle-Tacoma-Bellevue, WA	SQL, PMP, Python, Project Management, ACLS
7	9	3	Washington-Arlington-Alexandria, DC-VA	PMP, Project Management, SQL, ITIL, CNA
8	26	16	Trenton, NJ	EMT, PMP, SQL, Python, Project Management
9	185	367	Casper, WY	Operator, CNA, Medical Asst., ASE, ACLS
10	36	56	Durham-Chapel Hill, NC	CNA, Project Management, PMP, ACLS, Python
GDP Rank	**Search Rank**	**Spec. Rank**	**Bottom 10 MSAs in Per Capita GDP**	**Top Five Most Searched Skills**
363	368	169	Grants Pass, OR	Blender, Medical Asst., Medical Billing, Flagger, Typing
364	320	324	Prescott, AZ	CNA, ASE, EMT, Android, Medical Asst.
365	331	331	Brownsville-Harlingen, TX	CNA, Medical Asst., Pharmacy Tech., ASE, ACLS
366	303	316	Ocala, FL	CNA, ASE, Typing, Medical Billing, Medical Asst.
367	342	357	Homosassa Springs, FL	CNA, Medical Asst., Medical Billing, Pharmacy Tech., EMT
368	360	297	McAllen-Edinburg-Mission, TX	CNA, Medical Asst., ACLS, ASE, Typing
369	354	352	Punta Gorda, FL	CNA, Personal Trainer, ASE, Medical Asst., ACLS
370	294	54	The Villages, FL	CIW, CNA, Medical Asst., ASE, Medical Billing
371	302	364	Lake Havasu City-Kingman, AZ	CNA, Typing, EMT, Medical Billing, ASE
372	367	317	Sebring, FL	CNA, Operator, Medical Billing, Typing, Pharmacy Tech.

Table 2: Search Rank, Specialization Rank and Top Skills for MSAs with the highest and lowest GDP per capita.

searching relate to economic measures of MSAs over the five year time period of study. These correlations generally are medium to large effects (using Cohen's guidelines for effect sizes for correlations in the social sciences of .1, .3, and .5 for small, medium, and large effect sizes) and in expected directions, with both searches per capita and specialization correlating positively and moderately strongly with GDP per capita. These effects are also reflected in the generally low searches per capita and specialization rank of cities in the top 10 GDP per capita (Table 2). In fact, the top 40 MSAs with respect to skill search specialization, representing approximately the top 10% of MSAs, include economically strong cities such as San Jose (specialization rank #1), Washington D.C. (#3), San Francisco (#5), Seattle (#12), Austin (#14), Boston (#25), New York City (#26), Chicago (#35), and Los Angeles (#37). Many of these same cities also land in the top 40 in terms of skill searches per capita.

Both skill search measures correlated negatively with unemployment and poverty rates, suggesting again that people in healthy economies are searching more for skills. This is not necessarily intuitive, as one might imagine the unemployed or individuals otherwise down on their luck as those most likely to be developing skills in the hopes of regaining employment. This positive relationship between skill searching and economic activity foreshadows the 'rich get richer' phenomenon we explore in the next section.

Finally, we note that while skill search activity correlates positively with population, this is not a perfect correlation. There are many smaller cities with plenty of skill searching, including for specialized skills. For instance, Bloomington, IL, Hunstville, AL, and Hinesville, GA are all small to medium sized MSAs that are in the top 10 in either skill searches per capita or skill search specialization.

	Searches Per Cap	Specialization
Population	0.25	0.36
GDP Per Cap	0.46	0.43
Unemployment	-0.25	-0.23
Poverty	-0.30	-0.35
Searches Per Cap	1.00	0.65
Specialization	0.65	1.00

Table 3: Correlation across MSAs of skill searches per capita and skill search specialization with socioeconomic measures, averaged over 2012–2016.

Rich Cities Get Richer

The previous section looked at correlational relationships between skill search behavior and economic measures contemporaneously and over the entire 5-year time period. We turn now to before and after relationships: how does economic health relate to future skill searching, and how does that skill searching in turn relate to continued economic growth in the future? To be clear, we are not making causal claims that skill searching leads to economic growth. Rather we are providing evidence that skill searching plays some role in the development of growing economic regions.

Skill Searching, Past and Future Economic Strength To begin, we aggregated skill searches in the years 2012, 2013

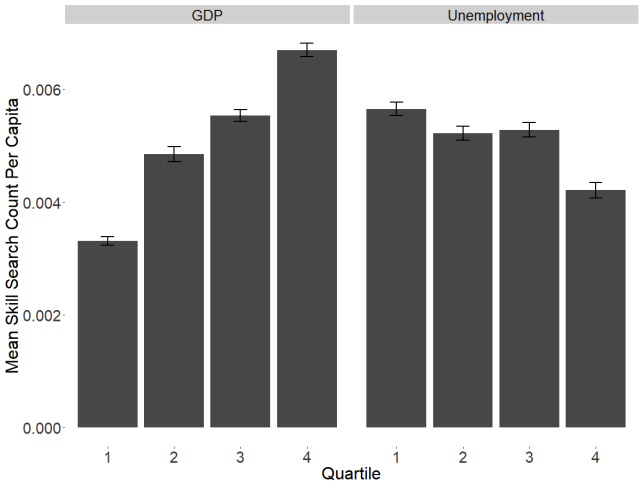

Figure 2: Mean skill searches per capita between 2012-2014 by MSA quartiles based on GDP and unemployment in 2012

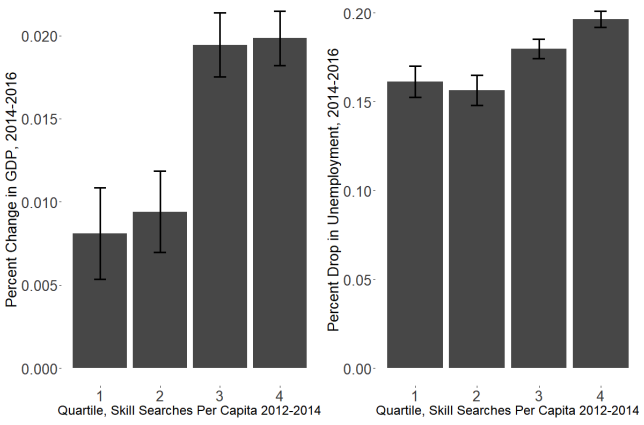

Figure 3: Mean percent change in GDP and percent drop in unemployment (2014–2016), by skill searches per capita quartile (2012–2014)

and 2014, and analyzed how economic outputs for the year 2012 are associated with aggregated skill searches in 2012–2014. This gives us a sense of how cities with high and low initial economic conditions differ in their near term skill search behavior. We grouped MSAs into quartiles based on GDP per capita values and separately on unemployment rate in 2012. Figure 2 shows how the quartiles vary in skill searches per capita between 2012 and 2014. Skill searches per capita increase quite linearly from the first to fourth quartiles of MSAs based on GDP, indicating that people in economically stronger cities at the outset of the time period engaged in more skill searching over the subsequent two years. Splitting at the median shows that MSAs in the lower half of 2012 GDP per capita generated significantly fewer skill searches per capita during 2012–2014 ($t = 8.90$, $p < 0.001$).

Similarly, MSAs with higher unemployment quartiles have lower skill searches per capita on average. Although the trend is not as linear as that for the GDP-based MSA quartiles, again this indicates that already economically healthy MSAs see more skill searches per capita going forward. Again, splitting at the median reveals that MSAs with higher unemployment generate significantly fewer skill searches than do lower unemployment MSAs ($t = 2.95$, $p = .003$).

Stepping forward in time, we next examine the relationship between skill searches and future economic growth by aligning 2012–2014 aggregated skill searches per capita with change in GDP 2014–2016. We expect that more skill searching in an MSA suggests that individuals in that MSA will have developed more skills in the near future and subsequently will contribute to greater economic output. We grouped the MSAs in four quartiles based on aggregated search queries per person and compared their GDP change percentage in 2014–2016 (see Figure 3). Splitting at the median, MSAs lower in skill searches per capita from 2012–2014 were significantly lower in GDP growth from 2014–2016 ($t = 2.46$, $p = 0.015$). While unemployment dropped almost across the board over this time period, MSAs in the top two quadrants of skill searches per capita in 2012–2014 saw a significantly greater percent change drop in unemployment from 2014 to 2016 ($t = 2.08$, $p = 0.038$), though this difference was not as pronounced as that for GDP.

Regression on Future Economic Growth Expanding this analysis, we conducted a regression analysis on GDP growth percentage from 2014–2016, using a variety of economic and skill search measures as inputs. Conceptually, this regressions tests the degree to which economic starting conditions (GDP, unemployment, population, and poverty in 2012), early economic growth (percent change in GDP, 2012–2014), and early skill searching impact future economic growth. We compare four models. The first is a baseline model for each city where only the economic factors were used. We consider this a strong baseline given that it contains economic factors very likely related to future GDP growth, even including GDP levels and growth immediately prior to the test period. We compare this baseline model to three additional models, one that adds skill searches per capita from 2012–2014, one that adds skill search specialization from 2012–2014, and a final model that adds both skill search measures.

Table 4 shows the results. In the baseline model, all factors except GDP 2012 significantly affect GDP growth. GDP growth, as a measure of percent change in GDP, is a rigorous outcome measure, with the baseline model explaining only just under 5% of the variance in future GDP growth (adjusted $R^2 = 0.047$).

The three subsequent models shown in Table 4 account for additional variance in future GDP growth. Adding just 2012–2014 skill searches per capita increases the R^2 value by more than 30% to 0.063, adding just 2012–2014 skill specialization to the baseline pushes R^2 to 0.076 (a relative increase over the baseline of more than 60%), and adding both generated the best fitting model, bringing R^2 to 0.08. Skill search specialization in particular was a strong predictor, stronger in fact than any of the economic variables.

We emphasize here that while the overall R^2 values are fairly low, explaining any variance in the *percent change* in

Model	Variables	Adj. R^2	F Stat.	P Value
Baseline	GDP Change 2012–2014* GDP 2012 Unemployment 2012* Population 2012* Poverty 2012*	0.047	4.69 (5, 366)	0.0004
Baseline+Searches	Skill Searches Per Capita 2012–2014**	0.063	5.15 (6, 365)	< 0.0001
Baseline+Specialization	Skill Search Specialization 2012–2014***	0.076	6.09 (6, 365)	< 0.0001
Baseline+Both	Skill Searches Per Capita 2012–2014 Skill Search Specialization 2012–2014**	0.080	5.58 (7, 364)	< 0.0001

Table 4: Regression analyses for GDP Change 2014–2016. Asterisk levels denote $p<0.05$, $p<0.01$, and $p<0.001$.

GDP of an MSA with only a few years lead time, especially when autoregressive components prove minimally explanatory, could be extremely valuable in forecasting the future growth of a region. Thus, while not causal, skill search behaviors appear to play a meaningful role, if nothing else as a proxy for some latent variable not captured by macroeconomic measures such as GDP and unemployment rates, that bear on the future economic growth of an MSA.

Role of Internet Access Although internet access is quite prevalent in the United States, rates of internet access in U.S. cities do vary, and thus may account for some of the effect of our measures. If one does not have internet access, searching for skill development is unlikely. Differences in internet access rates would not negate the effects of differences in skill developing searches, but rather suggest an explanation for these skill development search differences. In a theoretical world where all citizens had internet access or where access rates were equal in all geographies, would we still see effects for skill search differences across cities?

To test this, we ran an additional regression model that also included internet access rates from 2013 as a baseline predictor, noting that 2013 lies in the middle of our period of 2012 - 2014 for predicting 2014 - 2016 GDP growth[5]. Internet access rates do add variance explained, bringing the baseline R^2 to 0.07. Adding *Skill Searches per Capita* raises the R^2 to 0.077, with *Skill Searches per Capita* significant at $p < .05$. Adding *Skill Search Specialization* raises the R^2 to 0.086, with *Skill Search Specialization* significant at $p < .01$. The full model improves on this only minimally ($R^2 = .087$), though *Skill Search Specialization* remains significant ($p < .05$).

Discussion

Our primary finding highlights that the intent to develop skills via the internet aligns with a "rich gets richer" phenomenon across cities in the US, in which people in already thriving MSAs engage in more and more specialized searching for skill development online, which in turn explains variance in future economic growth. This suggests that the Matthew effect of accumulated advantage plays a role in the context of skill acquisition. The quartiles shown in Figures 2 and 3 illustrate this effect by dividing the five years of study into early and late periods, showing that the amount of skill searching early in the time period reflects initial economic conditions, as well as greater increases in future GDP and greater drops in future unemployment.

The full regression model then shows that the volume and degree of specialization of skills searched for account for variance in future economic growth beyond a number of baseline economic variables, including percent change in GDP in the years immediately prior. Specialization in particular added notable variance explained, and generally was associated with technology related skill searching in high productivity MSAs (Table 2). While we once again stress that this is not causal evidence, we do see a correlational alignment between use of the internet to develop skills and growing economic inequality at the city level in the US. This goes against the ideal of the internet as an information platform driving socioeconomic equality.

One possibility is that this correlational alignment is a sign of richer cities attracting individuals interested in skill development. Since richer cities often have larger populations and a wide variety of jobs (GDP per capita and population correlated at 0.35 through 2012-2016), they may influence inter-city migration patterns such that people who are highly motivated to advance in their careers, particularly those who want to develop specialized skills, move to these cities where they continue to develop high-value skills.

More broadly, one area for further investigation on the connection between internet-based skill development and economic growth is the role of the digital divide. Prior research shows that differences in use of online resources for learning between users with different socioeconomic status can increase existing inequalities (Reich, Murnane, and Willett 2012; DiSalvo, Khanipour Roshan, and Morrison 2016; Attewell 2001). Our research adds to this by highlighting how even general-purpose search engines may contribute to economic disparities. To be able to use search engines for gainful skill development, individuals must first know what information they should acquire to advance their careers, and they need to be aware of how to search effectively for that information. Furthermore, individuals who cannot

[5]US Census internet access data at MSA-level is only available for the year 2013 in our time period. Source: https://www2.census.gov/library/publications/2014/acs/acs-28/

afford to or do not have time to take courses or get certifications may never search for such skills even if those skills can provide them substantial career benefits.

Our finding that skill searches per capita varied across cities reflects differences in internet utilization rates of people in different regions. Differences in our specialization metric suggest a digital divide in terms of the type and value of information seeking for professional development. Further, we showed that skill search differences remain significantly associated with differences in GDP growth even when we control for variations in internet access. An important next step then is to further disentangle internet access issues from internet use issues at local levels and investigate how to overcome socioeconomic factors such as the digital divide to help people in economically struggling areas utilize internet-based resources to bolster employment prospects.

Novel Measurement of Online Skill Development

This work is, we believe, the first large-scale overview of intent to develop skills online across all the major population centers of the US. Given the correlations of our measures with economic outputs, internet-based data can be used in collaboration with similarly large scale macroeconomic statistics from traditional sources like BEA and BLS as a sensor for the economic health of different parts of the US.

Unlike measures from government sources that can lag by months, these data have the advantage of providing near real-time measurement of skill development. Furthermore, individuals' search behaviors usually reflect their real, honest and unbiased intentions (Dumais et al. 2014). Therefore, measuring aggregate demand for skill development through internet search data can highlight information about the skill gaining strategies of individuals that may not be available through traditional approaches. Through company payroll reporting, government measurement agencies do an admirable job tabulating the types of jobs people work, but skills are more nuanced, often transcending job types and titles, and the development of skills is a forward looking process, all of which suggests the need for novel skill development assessment techniques.

Our measures of skill development can also be used at a geographically granular level to understand the shifts in skill search behaviors in specific cities. This can provide important insights into how events such as the opening of a new university or closing down of a steel plant in a city affects the skill learning needs of its communities. Such information can allow relevant organizations to better respond to changing economic trends and offer them more agency. These measures can also inform the retraining and reemployment efforts of local communities by providing them insights about which skills are in demand. Social organizations and educational institutions can use such information to offer courses or training programs for teaching those skills as well as attract companies that hire individuals having those skills. However, we warn that internet search data may not contain information offered by community knowledge and macroeconomic measures provided by government websites (Lazer et al. 2014), and therefore, skill search data should be seen as a supplemental, rather than a substitute, data source.

Limitations and Future Work

We used internet skill searches as a sensor for interest in skill development throughout the US. However, this necessarily leaves out many alternate ways in which individuals begin to develop skills. For example, people may be influenced by their social network to take classes or otherwise develop skills. Additionally, our work focuses on professional skills, but not manual skills that are generally sought after by lower-wage workers. Our data did not allow us to distinguish between skill searches that were driven by individual motivations as opposed to searches driven by corporate-level innovation and growth. Generally, although our design controlled for many factors that could affect economic growth such as initial economic conditions and early economic growth, we cannot eliminate alternative explanations, as certain extraneous factors such as education could not be controlled because they were unavailable for the time period of study.

We only looked at what skills people in different regions of the US are looking to develop. In future work, it would be fruitful to incorporate the employers' demand for skills in different parts of the US, and capture how well the users' online skill searches correspond to employers' demand.

Conclusion

We studied the use of a general purpose search engine in the US for the purpose of professional skill development. By collecting over ten million skill searches over the course of five years and aggregating them to the cities in which they occurred, we found that cities with higher economic output are likely to see more skill searches. Just as importantly, cities with greater number of skill searches and more specialized searches are associated with greater economic growth in the future even after controlling for several important socioeconomic indicators. These findings indicate that the use of general purpose search engines for skill development may be contributing to growing economic disparities in different parts of the US. Building upon our findings, we argue that skill search queries can be used in conjunction with other socioeconomic variables as a sensor for economic health of different regions. Finally, our work opens new avenues for further research on improving the use of internet tools for skill development with an eye towards mitigating inter-regional socioeconomic disparities.

References

Askitas, N., and Zimmermann, K. F. 2009. Google econometrics and unemployment forecasting. *Applied Economics Quarterly* 55(2):107–120.

Attewell, P. 2001. Comment: The first and second digital divides. *Sociology of education* 74(3):252–259.

Basistha, A.; Kurov, A.; and Wolfe, M. 2017. Volatility forecasting: The role of internet search activity and implied volatility.

BEA. 2018. Bureau of economic analysis: Statistical areas.

Braithwaite, J. 1976. *Inequality, Crime, and Public Policy*. Routledge.

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

Burda, Z.; Wojcieszak, P.; and Zuchniak, K. 2018. Dynamics of wealth inequality. *arXiv preprint arXiv:1802.01991*.

Bureau, U. C. 2018. Metropolitan and micropolitan statistical areas population totals: 2010-2017.

2015. Careeronestop.

Chancellor, S., and Counts, S. 2018. Measuring employment demand using internet search data. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 122. ACM.

Choi, H., and Varian, H. 2012. Predicting the present with google trends. *Economic Record* 88:2–9.

Cohen, M., and Simet, L. 2018. Macroeconomy and urban productivity. *The Urban Planet: Knowledge Towards Sustainable Cities* 130.

comScore. 2016. comscore releases february 2016 u.s. desktop search engine rankings.

Cowell, F. 2011. *Measuring inequality*. Oxford University Press.

D'Amuri, F., and Marcucci, J. 2010. 'google it!'forecasting the us unemployment rate with a google job search index.

David, H. 2015. Why are there still so many jobs? the history and future of workplace automation. *Journal of Economic Perspectives* 29(3):3–30.

De Choudhury, M.; Jhaver, S.; Sugar, B.; and Weber, I. 2016. Social media participation in an activist movement for racial equality. In *ICWSM*, 92–101.

De Choudhury, M.; Morris, M. R.; and White, R. W. 2014. Seeking and sharing health information online: comparing search engines and social media. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, 1365–1376. ACM.

DiMicco, J.; Millen, D. R.; Geyer, W.; Dugan, C.; Brownholtz, B.; and Muller, M. 2008. Motivations for social networking at work. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, 711–720. ACM.

Dimpfl, T., and Jank, S. 2016. Can internet search queries help to predict stock market volatility? *European Financial Management* 22(2):171–192.

DiSalvo, B.; Khanipour Roshan, P.; and Morrison, B. 2016. Information seeking practices of parents: Exploring skills, face threats and social networks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, 623–634. New York, NY, USA: ACM.

Dumais, S.; Jeffries, R.; Russell, D. M.; Tang, D.; and Teevan, J. 2014. Understanding user behavior through log data and analysis. In *Ways of Knowing in HCI*. Springer. 349–372.

ERS, U. 2018. Usda ers - rural education.

Fourney, A.; White, R. W.; and Horvitz, E. 2015. Exploring time-dependent concerns about pregnancy and childbirth from search logs. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 737–746. ACM.

Frey, C. B., and Osborne, M. A. 2017. The future of employment: how susceptible are jobs to computerisation? *Technological forecasting and social change* 114:254–280.

Fugate, M.; Kinicki, A. J.; and Ashforth, B. E. 2004. Employability: A psycho-social construct, its dimensions, and applications. *Journal of Vocational behavior* 65(1):14–38.

Gao, Y.; Wang, M.; Zha, Z.-J.; Shen, J.; Li, X.; and Wu, X. 2013. Visual-textual joint relevance learning for tag-based social image search. *IEEE Transactions on Image Processing* 22(1):363–376.

Gilbert, G. N. 2010. *Computational social science*, volume 21. Sage.

Goldin, C. D., and Katz, L. F. 2008. *The Race between Education and Technology*. Harvard University Press.

Greenhow, C.; Robelia, B.; and Hughes, J. E. 2009. Learning, teaching, and scholarship in a digital age: Web 2.0 and classroom research: What path should we take now? *Educational researcher* 38(4):246–259.

Group, E. I. 2016. THE NEW MAP OF ECONOMIC GROWTH AND RECOVERY. Technical report.

Group, E. I. 2017. 2017 Distressed Communities Index - Economic Innovation Group.

Hall, D. T., and Mirvis, P. H. 1995. The new career contract: Developing the whole person at midlife and beyond. *Journal of vocational behavior* 47(3):269–289.

Hickman, L.; Saha, K.; De Choudhury, M.; and Tay, L. 2019. Automated tracking of components of job satisfaction via text mining of twitter data. In *ML Symposium, SIOP*.

Jhaver, S.; Chan, L.; and Bruckman, A. 2018. The view from the other side: The border between controversial speech and harassment on kotaku in action. *First Monday* 23(2).

Kawachi, I., and Kennedy, B. P. 1997. Socioeconomic determinants of health: Health and social cohesion: why care about income inequality? *Bmj* 314(7086):1037.

Kilpi-Jakonen, E.; Buchholz, S.; Dämmrich, J.; McMullin, P.; and Blossfeld, H.-P. 2014. Adult learning, labor market outcomes, and social inequalities in modern societies. *Adult learning in modern societies. An international comparison from a life-course perspective* 3–28.

Kneebone, E. 2014. The growth and spread of concentrated poverty, 2000 to 2008-2012. *The Brookings*.

Kuhn, P., and Mansour, H. 2014. Is internet job search still ineffective? *The Economic Journal* 124(581):1213–1233.

Kundu, O., and Matthews, N. E. 2019. The role of charitable funding in university research. *Science and Public Policy*.

Lazer, D.; Kennedy, R.; King, G.; and Vespignani, A. 2014. The parable of google flu: traps in big data analysis. *Science* 343(6176):1203–1205.

Manyika, J.; Lund, S.; Chui, M.; Bughin, J.; Woetzel, J.; Batra, P.; Ko, R.; and Sanghvi, S. 2017. Jobs lost, jobs gained: Workforce transitions in a time of automation. *McKinsey Global Institute*.

Olteanu, A.; Castillo, C.; Diaz, F.; and Kiciman, E. 2016. Social data: Biases, methodological pitfalls, and ethical boundaries.

Purcell, K.; Rainie, L.; and Brenner, J. 2012. Search engine use 2012.

Reich, J.; Murnane, R.; and Willett, J. 2012. The state of wiki usage in us k–12 schools: Leveraging web 2.0 data warehouses to assess quality and equity in online learning environments. *Educational Researcher* 41(1):7–15.

Reuter, C., and Kaufhold, M.-A. 2018. Fifteen years of social media in emergencies: a retrospective review and future directions for crisis informatics. *Journal of Contingencies and Crisis Management* 26(1):41–57.

Richardson, M. 2008. Learning about the world through long-term query logs. *ACM Transactions on the Web (TWEB)* 2(4):21.

Saha, K.; Weber, I.; and De Choudhury, M. 2018. A social media based examination of the effects of counseling recommendations after student deaths on college campuses. In *ICWSM*.

Scholarios, D.; Van der Heijden, B. I.; Van der Schoot, E.; Bozionelos, N.; Epitropaki, O.; Jedrzejowicz, P.; Knauth, P.; Marzec, I.; Mikkelsen, A.; and Van der Heijde, C. M. 2008. Employability and the psychological contract in european ict sector smes. *The International Journal of Human Resource Management* 19(6):1035–1055.

Söderström, T.; Fahlén, J.; Ferry, M.; and Yu, J. 2018. Athletic ability in childhood and adolescence as a predictor of participation in non-elite sports in young adulthood. *Sport in Society* 21(11):1686–1703.

White, R. W.; Harpaz, R.; Shah, N. H.; DuMouchel, W.; and Horvitz, E. 2014. Toward enhanced pharmacovigilance using patient-generated data on the internet. *Clinical Pharmacology & Therapeutics* 96(2):239–246.

Wilkinson, R. G. 2002. *Unhealthy societies: the afflictions of inequality*. Routledge.

Wilson, M., and Daly, M. 1997. Life expectancy, economic inequality, homicide, and reproductive timing in chicago neighbourhoods. *BMJ: British Medical Journal* 314(7089):1271.

Yom-Tov, E. 2016. *Crowdsourced Health: How What You Do on the Internet Will Improve Medicine*. Mit Press.