

Individual-level evidence of IT skill specialization and formation in the advanced industrial economies*

Mark Huberty[†]

April 2, 2012

Abstract

Typologies of capitalism emphasize the role of skill formation as both a shaper of and a response to firm incentives for specialization in different forms of technology. However, actual worker skill is usually difficult to measure and costly to observe. We present novel estimates of IT worker skill based on actual observation of skills in practice. We exploit a new data source that allows observation of tens of thousands of users across over fifty countries and three years. Based on these data, and subject to several caveats about the worker population, we find ().

1 Introduction

Theories of comparative political economy have put substantial emphasis on the interaction of welfare state incentives, worker skill, and firm choices about how and where to specialize. However, measurements of worker skill are often quite coarse. The difficulty of observing individuals at work, under similar conditions, across countries has led researchers to rely on aggregate statistics such as the level of vocational training, the share

*DRAFT NOT FOR DISTRIBUTION

[†]Travers Department of Political Science, University of California, Berkeley. Contact: markhuberty@berkeley.edu.

of university-educated individuals in the workforce, and the presence of apprenticeship programs.

We present novel measurements of highly granular skill in a large information technology (IT) worker population across many countries. Using data from an open, online IT question-and-answer forum, we are able to specify exact definitions for skill, skill diversity, and expertise. Using those definitions, we can construct very detailed comparisons of comparative IT worker skill specialization across both advanced industrial and emerging market economies. We believe this is the first use of this dataset to measure worker skill.

Within the resolution of the data, and subject to several caveats on the selection process, we find ().

2 Institutions and skill development

Theories of comparative political economy predict that institutional variation in the advanced industrial economies will create diverging patterns of skill formation in workers, and in technology adoption and exploitation among firms. The Varieties of Capitalism literature in particular ([Hall and Soskice, 2001](#)) makes two predictions about technology and skills: first, that workers in more liberal economies will adopt more general skills than their counterparts in coordinated economies; and second, that firms in liberal economies will be more likely to engage in radical innovation and adopt leading-edge technology.

These claims rest on an implicit microeconomic model of firm and worker behavior. In more liberal economies, the lack of employment protection laws means that workers must consider the possibility of losing their jobs without recourse. Likewise, firms must consider that workers who have no protection will feel little loyalty to any given firm. In that context, both workers and firms will converge on an equilibrium that values general skills—the workers, so that they can easily find jobs at other firms, and the firms, so that

workers are easily replaced. Conversely, countries with strong employment laws will tend to favor specific skills, because firms must extract high value from employees that cannot be easily dismissed, and because employees believe that investment in firm-specific skills will be rewarded by long employment tenure.

Empirically, the evidence used to support these claims has typically relied on macro-level measures of skill formation, such as the share of workers in vocational or continuing education (taken to be representative of firm-specific skills), the share of university-educated workers (taken to represent more general skills), and the role of apprenticeship programs in skill formation.¹ But these metrics do not directly measure worker-level behavior. Moreover, they may mis-identify skill specificity. For instance, we may believe that a manager with a liberal arts degree from a university has cultivated very general skills. But a university-educated statistician working in natural language processing in Silicon Valley has a highly specific skillset useful for a relatively small set of highly specialized firms. Whether that individual should be lumped into the same category as a manager, or instead treated like a highly skilled machinist graduate of an *Azubi* program in Baden-Wurtemberg is unclear.

3 Pursuing new data for measurement

I exploit a new dataset on technology adoption and skill formation in the information technology industry. That dataset provides individual-level evidence of cross-sectional and intertemporal patterns of skill formation among workers worldwide, and augments that data with measures of expertise and evidence of patterns of worker interaction.

StackExchange represents a collection of community-created and maintained internet sites that field questions and answers on topics related to information technology. Begun

¹For a full set of metrics, see [Hall and Gingerich \(2009\)](#).

in 2008 by Joel Spolsky and colleagues, it has grown into one of the most comprehensive communities of technical information in the world. As of 2011, it reported approximately 725,000 users, 1.2 million site visits per day, and 4.2 million separate answers covering 82% of the 1.9 million submitted questions. It had grown from its initial site, StackOverflow, to sites covering everything from programming to data security to cooking and philosophy. Anecdotal evidence suggested that the quality of the site had led recruiters to ask for samples of Stackoverflow questions and answers from potential hires, as one measure of worker skill.²

The format of StackExchange interactions can be stylized as follows: a user poses a question and tags the question with metadata related to the specific technologies or technological domains the question pertains to. Other users respond to the question. Those responses, in turn, are rated by respondents, the person who posed the original question, and other community members, on the basis of accuracy, solution elegance, and completeness. Questions and answers can be retrieved by querying metadata, the user ID of the questioner or answerer, and other boolean search terms.

Users contribute to StackExchange through user accounts that contain a range of information about the user. Users may, but are not required to, provide information about their geographic location, interests, website address, and other personal characteristics. There is no evidence that this data is validated, and avatars or similarly obscured or stylized user names and identities are permitted.

All StackExchange data is collected and published under a Creative Commons license. In keeping with the license, StackExchange releases a bi-monthly dump of its entire dataset, including the full record of questions and answers, user profiles, record tagging and metadata, and time stamps. The data are anonymized with respect to user names but

²cite

not other user metadata, including geographic location, web site, the questions posed or answered by each user, and the user’s community-determined “reputation”.³

These data therefore represent a rich trove of information about the interaction of IT workers worldwide. The dataset contains not only the location and other information about the community members themselves, but also, through the information on questions posed and answered and their reputation, the kinds of technology they work with and are expert in, and their rating by a community of their peers. As such, the data provide an opportunity to directly observe patterns of specialization of knowledge and interaction among IT workers across both countries and time.

4 Research Design

To employ the StackOverflow data in the study of skill and task specificity, we first need to define exactly what is meant by the terms task and skill, and how specificity will be defined.

4.1 Definitions

4.1.1 Skill

Skills will refer to knowledge of the workings of specific technologies or technological concepts. These may include, for instance, language-specific knowledge in *C*, knowledge of algorithmic implementations of specific programming or computer science concepts, or operational knowledge of how to set up and administer specific kinds of software or soft-

³For more information on the StackExchange data dump, see <http://blog.stackoverflow.com/category/cc-wiki-dump/>. For information on the terms of use for the Creative Commons license, see <http://creativecommons.org/licenses/>. For academic use of the data, including studies of social interaction and exchange markets, see <http://blog.stackoverflow.com/2010/05/academic-papers-using-stack-overflow-data/>.

ware systems. These skills may, in fact, be independent from one another. Just as someone might be very knowledgeable about fuel injection systems, but not necessarily ever need to rebuild them, so could an IT worker be very knowledgeable about the *C* programming language without every using it to, say, optimize highly parallelized programs used for scientific computing.

We can define skill on the basis of the metadata attached to each question. We treat each unique metadata tag as a member of a skill domain. Skill domains in turn are constructed from patterns of co-occurrence of metadata tags in questions. Formally, we construct the proximity matrix P from a set of questions Q , each of which contains a subset of tags $t_Q \in T$. We can construct the proximity of tag t_i to tag t_j as the conditional probability that a question containing t_i also contains t_j . In cases where $P(t_i, t_j) \neq P(t_j, t_i)$, we take the maximum conditional probability.⁴

4.1.2 IT workers

We define an IT worker as a contributor to and participant in the question-and-answer process on Stackoverflow. The April 2011 data dump provides information on approximately 700,000 individual users. Within this group, we identify two separate groups of users: those who provide geographic information in their user profiles (approximately 130,000 individuals) and those who do not. For those users who provide geography information, we disambiguate that data using the Yahoo GeoDict webservice.⁵ This service provides a scriptable API which translates unstructured geographic information into a

⁴The intuition here is straightforward. Consider the `bash` shell environment for `unix`. $P(\text{bash}|\text{unix}) \approx 1$, since `bash` is almost exclusively a `unix` tool. But $P(\text{unix}|\text{bash}) \approx 0$, since there are many, many `unix` utilities. Taking the minimum conditional probability would wildly understate the relationship between `bash` and the `Unix` operating system ecosystem.

⁵<http://developer.yahoo.com/geo/placemaker/>

standardized format.⁶

4.1.3 Individual-level skill and task specialization

Given these definitions of task and skill specificity, we can now define what we mean by those terms for any given individual. For a user $u \in U$, their skill specificity is the mean specificity in all questions $q \in Q$ that they have answered. Task specificity is defined the same way. Each user thus receives both a skill and a task specificity score. Conceptually, then, each user falls into one of four possible categories: tackling specific tasks with general skills, general tasks with general skills, specific tasks with specific skills, or specific tasks with general skills. Notice, here, however, that we are not bound by discrete categorization. Rather, we can measure both task and skill specialization as continuous quantities.

Using the alternative definition of task proximity, we can define specialization as the number of different technology clusters in which a given user displays expertise. This eliminates the (potentially artificial) distinction between tasks and skills.

4.2 Technical design

4.2.1 Option 1: continuous measures of tasks and skills

1. For the full set of users U , select those with geo-coded information in their profiles, U_g (DONE)
2. For U_g , disambiguate the geo-coded data (DONE, via the Yahoo Placemaker service)
3. For all questions Q , get two subsets of questions:

⁶For instance, variants like “Karlsruhe”, “Karlsruhe, Germany”, and “Karlsruhe, Deutschland” are all translated into a common format specifying city and country in a parsable JSON format.

- (a) All questions for which the submitter and AT LEAST one answerer are in U_g
 - (b) All questions for which AT LEAST one answerer are in U_g
4. Get all tags T in the question set Q . Subset for only those tags T^* that occur more than N times in the dataset. Construct a proximity matrix M_{T^*} with the conditional probability of tag co-occurrence as defined in equation ??
 5. For all tags T^* , categorize as either skills or tasks
 6. For all users U_g , get all tags in all questions answered by U_g and the community-generated score of the answer. Compute the weighted task and skill specificity measures as described in equation ?? and section 4.1.3

4.2.2 Option 2: clusters of tags

1. Get the users and tags as described in section 4.2.1.
2. For tags T , construct the graph G_t of tags wherein each tag is a node and the edges are defined as the conditional probability of co-occurrence. Only use edges for which $p > threshold$.
3. For G_t , find the graph k-cores (see <http://networkx.lanl.gov/reference/algorithms/core.html> and associated references).
4. For each k-core, get the tags associated with it
5. For each user U_g , determine which tags $T' \in T^*$ apply to their answers
6. Count the number of tags they answer, in the number of k-cores, for each user in U_g

5 Hypotheses

5.1 Approach 1: continuous measures

H1 Users from CME countries will have higher mean specificity, but in fewer skills or tasks (i.e., will display strong but relatively narrow expertise)

H2 Users from LME countries will have lower mean specificity scores, but will display expertise in more skills or tasks (i.e., will display weaker but broader expertise)

5.2 Approach 2: clustered measures

H1 Users from LME countries will display expertise in more technology “clusters” than users from CME countries. This is equivalent to saying that LME users will have a more general skill set, as a function of more liquid labor markets.

H2 Users from CME countries will display expertise in more tags in any given cluster than users from LME countries, even if they are present in fewer technology clusters

6 Results


7 Discussion

8 Conclusions

Appendices

A MCL cluster examples

How can a child process return two values to the parent when using pipe()?



Love Stack Overflow? Love building stuff?

Use our API to build your own mobile, web, and desktop Q&A apps — your platform, your rules!

I have my child process counting the frequency of words from a text file. I am using `pipe()` for IPC. How can the child process return both the word name and the word frequency to the parent process? My source code is in C and I am executing it in a UNIX environment.

[c](#) [unix](#) [ipc](#) [pipe](#)

[link](#) | [improve this question](#)

[edited 11 mins ago](#)
[Jonathan Grynspan](#)
12k • 1 • 8 • 26

[asked 19 mins ago](#)
[het](#)
11 • 2

1 What have you tried so far? Do you have any code that you cannot get to work? — [Vlad Lazarenko](#) 17 mins ago

Do you know how to create the pipe? Do you know how to send anything? — [Beta](#) 9 mins ago

[feedback](#)

Hello World!

This is a collaboratively edited question and answer site for **professional and enthusiast programmers**. It's 100% free, no registration required.

[about »](#) [faq »](#)

tagged

[c](#) × 40918

[unix](#) × 6069

[ipc](#) × 773

[pipe](#) × 439

asked **today**
viewed **18 times**
active **today**

2 Answers


[active](#) [oldest](#) [votes](#)

1

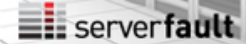
Writes to a pipe up to the size of `PIPE_BUF` are atomic (included in `limits.h`), therefore you can easily pack your information into some type of struct, and write that to the pipe in your child process for the parent process to read. For instance, you could setup your struct to look like:

```
struct message
{
    int word_freq;
    char word[256];
};
```


Then simply do a read from your pipe with a buffer that is equal to `sizeof(struct message)`. That being said, keep in mind that it is best to only have either a single reader/writer to the pipe, or you can have multiple writers (because writes are atomic), but again, only a single reader. While multiple readers can be managed with pipes, the fact that reads are *not* atomic means that you could end up with scenarios where messages either get missed due to the non-deterministic nature of process scheduling, or you get garbled messages because a process doesn't complete a read and leaves part of a message in the pipe.



GOT SERVER QUESTIONS? TRY:



serverfault



CAREERS 2.0

Join the Expedition! Explore Front-End UX at the Bleeding... Occipital

Builder NY San Francisco CA

Figure 1: Sample StackOverflow question and answer

References

Hall, P. and Gingerich, D. (2009). Varieties of capitalism and institutional complementarities in the political economy: An empirical analysis. *British Journal of Political Science*, 39(03):449–482.

Hall, P. and Soskice, D. (2001). *The Varieties of Capitalism*. Oxford University Press, Oxford.