Working Paper 24-001

Learning to Use: Stack Overflow and Technology Adoption

Daniel Jay Brown
Maria Roche

Harvard Business School

Learning to Use: Stack Overflow and Technology Adoption

Daniel Jay Brown
Harvard Business School

Maria Roche
Harvard Business School

Working Paper 24-001

Funding for this research was provided in part by Harvard Business School.

Learning to Use: Stack Overflow and Technology Adoption

Brown, Daniel Jay
danielbrown@hbs.edu

Roche, Maria
mroche@hbs.edu

July 13, 2023

Abstract

In this paper, we examine the potential impact of Q&A websites on the adoption of technologies. Using data from Stack Overflow – one of the most popular Q&A websites worldwide – and implementing an instrumental-variable approach, we find that users whose questions are answered within 24 hours are significantly more likely to adopt the technologies that they ask about in their next job than users whose questions are answered later or not at all. In analyzing heterogeneous effects, we detect that this relationship is driven entirely by users located in or in close proximity to technological hubs, is stronger for more established technologies, and for users who have already asked more questions about a focal technology. Our findings suggest that the feedback provided by virtual communities can assist with the adoption of new technologies by supporting individuals that are learning to use a technology rather than in the discovery of new ones. Critically, our results suggest that such online activity functions as a complement to physical proximity, rather than as a substitute.

Both authors' affiliation: Harvard University, Harvard Business School, Boston, MA 02163. We are grateful for the many helpful suggestions we received from Hong Luo, Olimpija Zaevska, Andrea Morrison, Stefano Brusoni and Ron Adner. This manuscript benefited from many comments provided at DRUID, the HBS Digital Workshop, and the Open and User Innovation Conference. We gratefully acknowledge funding from the Harvard Business School Division of Research and Faculty Development. All errors and omissions are our own.

1 Introduction

In the last several decades, the default way to organize workers was co-location in a city, building, or office. A major benefit of such physical proximity is attributed to the ability to encourage the serendipitous flow of information and ideas (Kabo et al. 2020, Lane et al. 2021, Lee 2019, Roche et al. 2022). Most recently, however, traditional work formats have been called into question, as it seems that working from anywhere may become more feasible than ever before (Choudhury et al. 2020). As a consequence, we may find knowledge workers become more distributed across space in the future. This potential redistribution of workers makes it critical to understand how to organize work for a highly digitized and global workforce. Not surprisingly, leading technology companies have created units, such as Google's People Innovation Lab, Meta's Global Workplace Research Group, and Microsoft's Future of Work Group to address questions along these lines.

One potential substitute for workplace interactions, which have been long-established to play a critical role in promoting knowledge exchange among workers (Allen 1997, Roche 2020, Roche et al. 2022), may be found in the virtual space. Increasingly, knowledge workers, especially software developers, seek help on online community platforms. Not only do they seek help, they also provide assistance to community members in need (Xu et al. 2020). Anecdotal evidence suggests that many rely heavily on these platforms to learn about new technologies and how to implement them. Frequently, copying and pasting code snippets from Q&A platforms, such as Stack Overflow, has become "a top method of troubleshooting among developers" (Feldman 2017).

In this paper, we seek to empirically assess whether these types of online community platforms do, in fact, aid in the transfer of knowledge, particularly in the *use* of a new technology. Moreover, provided shifts in the organization of work, our goal is to further shine light on whether the availability of online resources may, to some extent, substitute for in-person interactions by reducing frictions associated with technology adoption.

The context of our study is the virtual community Stack Overflow. Stack Overflow is a widely popular Q&A website for software programmers (Patel 2022). As of 2022, Stack Overflow has over 18 million users from around the world, and more than 23 million questions have been posted to the site. Any user can post a question, which may be responded to by any other user from around the world. We compile our data set based on activity from the virtual community platform using observations starting in 2010 and ending in 2019. Our final dataset consists of 116,976 observations of 66,006 questions from 7,669 unique users located in the US and Canada.

In examining our main research question, we first assess the baseline relationship between adopting a technology and having a question answered (within a certain time period). We measure adoption by assessing whether the asker of the question uses the focal technology in their next job, meaning their next position of employment.[1] We include a host of fixed effects to address concerns associated with omitted variable bias. Across all specifications, we find a positive association between having a question answered and adopting a technology related to the question. Compared to the average likelihood of adoption, the magnitude of the relationship for our preferred and most stringent specification lies at a 6% increase in the probability of adoption.

To address concerns associated with endogeneity, we implement an instrumental variable approach. Without such an approach, the worry may be that our more 'naive' results are merely capturing a selection into treatment or that other unobservable features, which we do not absorb with the controls or fixed effects we include, may be driving the results. For instance, users with a higher likelihood of adoption may post questions that are more likely to be answered.

Our instrument for having a question answered is constructed using a Bartik-like measure, based around the timing of the user's question and technologies that are related to but separate from the technologies that are the focus of the user's question. Using this instrument, we find that having a question answered, e.g., within a three hour window, is associated with a 5.11 percentage point higher probability of adopting the technology in the asker's next job. This represents over a doubling from the mean, confirming that having a question answered on a Q&A website may, substantially impact the adoption of new technologies.

Prior literature indicates that proximity is important for knowledge spillovers (Almeida and Kogut 1999, Jaffe et al. 1993), and especially critical for the adoption of new technologies (Tambe 2014). Because Q&A websites can be posted to from anywhere in the world, it's possible they can serve as a substitute for physical proximity. To test the interaction between proximity and use of the platform, we apply the distance of a user from a technology hub and examine the differential relationship with technology adoption. For our purposes, hubs are defined as cities with the top fifteen highest employment levels for either "Computer Programmers" (Occupational Code 15-1251)

[1]This could be a different job with the same employer or with a new employer.

or "Software Developers and Software Quality Assurance Analysts and Testers" (Occupational Code 15-1256). From these analyses, we find that rather than serving as a substitute, feedback received from the virtual community complements the benefits of physical proximity for technology adoption.

In addition to finding that a question promptly answered increases the likelihood that the user will adopt the related technology in their next job, and that this effect is stronger for users located in technological hubs, we test for other heterogeneous impacts. Exploiting heterogeneity among the users and technologies, we find evidence that this relationship is also stronger for users who have asked more questions related to the focal technology and for more established technologies. These results indicate that Q&A websites may play a particularly important role in the adoption of established technologies by serious users with specific problems.

The external validity of our approach may be limited given that we focus on a specific virtual Q&A platform. However, Stack Overflow is an especially important Q&A website - as of November 2022, it was one of the 250 most visited websites in the world, according to data from SimiliarWeb. Although our results are based on particular activities on a specific online platform, we believe they have broader implications, especially for knowledge workers.

The results of this paper contribute to three main bodies of work. For one, we contribute to the literature assessing who benefits from contributions to online communities, and the factors influencing the extent of these benefits. Much prior research in this space has focused on the implications of contributions – especially to open source – for firms (Nagle 2018, Seo et al. 2021, Huang et al. 2022, Mollick 2016) or for the contributor, in the context of reviews (Forman et al. 2008, Mudambi and Schuff 2010). Another stream of work deepens the analysis regarding the individual contributor and how their activity may impact career outcomes, finding a strong relationship between career advancement motivations and answering questions posted on online communities (Xu et al. 2020, Hertel et al. 2003, Huang and Zhang 2016). Finally, our work speaks to the literature on the drivers of technology diffusion by examining the role of online communities in the spread of technologies, where thus far, physical, social, and existing technology capabilities have been suggested to be important factors promoting adoption (Dewan et al. 2010, Tambe 2014, Angst et al. 2010, Fichman and Kemerer 1997, Roche et al. 2022).

Taken together, our results provide crucial insight to the current discussion around organizing work and the flow of knowledge. Critically, our findings highlight that frictions associated with learning to use new technologies are still bound by physical space where virtual communities aid in the adoption of new technologies by helping solve specific problems, but cannot fully substitute for physical proximity to technology hubs.

Moreover, our results may have significant implications for firm performance. The ability to not only identify, but actually adopt new technologies is a key firm capability (Cohen and Levinthal 1990). Our results indicate that virtual communities can help workers adopt new technologies, but that these tools work best for those located in technology hubs and for established technologies. This implies that firms and their knowledge workers that seek to adopt cutting-edge technologies may still need to maintain a substantial presence in locations where there is a critical mass of workers using these tools.

2 Background

Learning to use new knowledge can be a critical driver of firm success, especially in industries featuring rapid technological change (Cohen and Levinthal 1990, Teece et al. 1997). Firms with the ability to quickly adopt new technologies are more likely to succeed in the face of new developments (Karimi and Walter 2015). Differences in the ability to adapt and commercialize new knowledge can thus significantly affect firm performance (Wales et al. 2013).

The notion that the production and diffusion of knowledge is spatially driven has been studied extensively over the past decades (Acs et al. 2002, Carlino and Kerr 2014, Jaffe et al. 1993, Porter 1996, Rosenthal and Strange 2003, Scott and Storper 2003, Roche et al. 2022). Numerous studies have provided empirical evidence that physical proximity, even down to a building, street or bridge level, can increase innovation and entrepreneurship (Audretsch et al. 2006, Roche 2020, Dutta et al. 2022, Roche et al. 2022). Reasons for the benefits of proximity that have been suggested encompass: reductions in transportation costs, cost reduction through labor market pooling, cost reduction through shared inputs, and a reduction in search effort (Marshall 1890). A major theme in this line of work is that the physical environment influences the costs associated with the search for problems and solutions, and costs associated with access to available resources, which often occur through face-to-face interactions (Sorenson and Audia 2000, Sørensen and Sorenson 2003).

Although we know the physical environment matters in many instances (Jacobs 1969, Porter

1996), we have less information at hand regarding how these mechanisms translate into more digital and data-driven environments where many of the rational benefits of proximity may carry less weight. For example, the use of an open source digital tool should incur close to zero transportation and access costs, and its adoption thus be fairly agnostic to location. However, recent work suggests that location still matters in such contexts where transportation and distribution costs are near zero. These contexts include online purchasing (Blum and Goldfarb 2006, Forman et al. 2009) and online crowdfunding (Lin and Viswanathan 2016). That we still observe deep localization of software and other digital-first industry seems to present a puzzle from the perspective of rational, utility-maximizing agents (Lewis 1999). This is a puzzle that is critical for both managers and policy-makers alike to understand.

To explore why this may be the case and when the virtual space may be able to compensate for the physical, a closer examination of potential virtual substitutes for offline interactions may provide critical insight. One such potential substitute may be found in online community platforms, where knowledge workers, especially software developers, increasingly seek help. Not only do they seek help, they also provide assistance to community members in need (Xu et al. 2020). Anecdotal evidence indicates that copying and pasting code snippets from Q&A platforms, such as Stack Overflow, has become "a top method of troubleshooting among developers"(Feldman 2017), suggesting that many rely on these virtual spaces to learn about new technologies and how to implement them. As such, these platforms may substantially reduce the costs associated with access to knowledge – particularly knowledge pertaining to *using* a technology.

Yet while these digital platforms may help workers learn to use technological knowledge, it is unclear whether they serve as a substitute to other resources, such as spatial proximity to practitioners, or as a complement to these resources. If these platforms can serve as a partial substitute for physical proximity, then they may reduce the advantage that workers and firms in technological hubs possess in quickly learning to implement new technologies. Fundamentally, platforms like Stack Overflow are oriented towards problem solving - they are intended to fill in the user-specific gaps in technical knowledge that are not readily served by other resources. In this sense, they are not intended to substitute for the formal knowledge provided by classes or textbooks, or even readme files which provide a basis for the general usage of a technology. But they may provide an alternative to the type of tacit knowledge that would be gained from proximity to existing practitioners. Transfer of this type of "know-how," which is difficult to articulate or codify ex ante because it relies on situational needs, is particularly associated with physical proximity and interaction (Gray et al. 2015, Lee 2019). Because participants in digital platforms can respond to specific problems and demonstrate solutions, they may be particularly well-suited to handling this type of knowledge transfer. In which case, they may serve as a partial substitute for spatial proximity for users learning to use a new technology.

Contrarily, it is possible that these platforms may be most helpful as a complement to proximity to a technological hub. Digital platforms may provide access to a different set of benefits than those provided by physical interaction. For example, virtual communities might be helping workers solve problems more quickly or conveniently. The digital platform may thus lower the cost of adoption, without substituting for the assistance provided by physical interaction. If this is the case, virtual communities may actually amplify the advantages workers located in hubs have over workers located elsewhere.

3 Estimation Strategy

3.1 Empirical Setting

To assess the role of online communities in providing access to critical knowledge, our study focuses on Stack Overflow, a widely popular Q&A website for software programmers. As of 2022, Stack Overflow has over 18 million users from around the world, and more than 23 million questions have been posted to site. It covers a broad set of programming technologies, including languages and software packages. Any user can post a question on Stack Overflow, and any other user from around the world can answer the question. If the asker considers the answer helpful, they can then mark the answer as "accepted," which benefits the answering user's reputation score on Stack Overflow.

Answering questions on Stack Overflow can be a way for users to display their expertise and help advance their career (Xu et al. 2020). Until April 2022, [2] Stack Overflow had a feature called Stack Overflow Developer Story, which functioned as a digital resume, helping users display both their job history and their history of answering questions on the site. As we will describe in the

[2]Stack Overflow discontinued its entire Talent business, which included the Developer Story, because the revenue generated from job postings and ads was insufficient (Mulchandani 2022).

next section, the combination of Stack Overflow Q&A data and data from the Stack Overflow Developer Stories provides us with an opportunity to examine the relationship between having a question answered and adopting a new technology.

3.2 Data

Our dataset for this analysis is based on the Stack Overflow question pages, user profiles, and Stack Overflow Developer Stories. Using the user profile data, we identify users who are explicitly located in the United States or Canada, because Developer Stories was primarily popular with users based in these countries. We collect Stack Overflow Developer Stories for each of these users, with 5% (11,830/254,184) of them available and including job history. We also include the history of all the questions these users have asked in our data collection effort.

Stack Overflow provides data on all Stack Overflow questions and answers via Google BigQuery. These data include when the question was asked, whether and when it received an accepted answer and a variety of other metrics. For each question, we identify the technologies involved using the "tags" that the user applied to the questions. Tags are standardized labels for the topics a question involves. There are 68,040 tags in Stack Overflow, and their coverage ranges from broad topics ("server") to programming languages ("r") and specific packages ("pandas"). Questions can have up to five tags.

We organize our dataset on a question-tag basis. For each question, the key independent variable is whether the question received an answer within a specified number of hours that the asker marked as "accepted". Since our assumption is that quick responses are critical for learning to use to take place, we examine the time from one hour to 24 hours after asking. Organizing on a question-tag basis allows us to include tag fixed effects and a variety of user-tag related controls.

To measure adoption, we rely on data from the user's Stack Overflow Developer Story. For each question, we find the next job that the user began after asking the question. Jobs in developer stories are marked with tags to denote which technologies the jobs involved, and these tags are the same tags used for questions in Stack Overflow. Each job is a person's position of employment. For each question-tag observation, a tag is marked as "adopted" if the next job has the tag on it. If the user did not begin a new job after asking the question, that question is dropped from the dataset. Additionally, if the user had the tag on a job they started before the question, we drop the question-tag observation. Finally, we only include question-tag observations if they were the first question the user asked about a tag before their next job, in order to avoid multiple tag observations for the same job.

Table 1 provides a comparison of the 7,669 users included in our sample to the the total population of users who asked a question and lived in the United States or Canada. Compared to the population, our sample users ask and answer more questions across a larger number of unique tags. This may indicate that our sample disproportionately contains a set of users who are highly proficient, full-time software developers with extensive experience, rather than beginners or individuals whose programming abilities are not a central requirement for their work. The quick adoption of new software technologies may be particularly important to full-time developers, making our sample especially relevant to our research focus.

Insert Table 1 about here.

Our final dataset starts in January 2010 and ends in December 2019 and consists of 116,976 observations of 7,669 unique users, 66,006 questions, and 5,827 tags. As displayed on Table 2, the average probability of adoption is .048. The likelihood of a question getting an accepted answer within one hour is 37.7%, within three hours is 46.1%, within six hours is 50%, within twelve hours is 53.4%, and within twenty-four hours is 56.9%. Table 2 reports further summary statistics.

Insert Table 2 about here.

4 Answering and Adoption

4.1 The Ideal Experiment

In considering our estimation strategy, it may be useful to conduct a thought experiment first. Intuitively, we would like to run an experiment that would randomly assign questions to a treatment group or a control group. The questions the treatment group asks would all be answered within a specific amount of time, while the questions in the control group would not be answered, or would be answered later. We would have the full record of the askers' usage of the related technology and we could simply compare the average rate of technology adoption in the treatment group to the average rate of adoption in the control group. The difference in the averages would equal the effect of having a question answered on technology adoption. To assess the role of location and other features, we would further need to stratify our sample based on these distinct conditions and estimate the differential effects.

While we are limited in the extent to which we can run the experiment described above, it provides a useful guideline for designing an empirical test and addressing possible threats to identification. For instance, we want to avoid the risk of selection, where askers who are more likely to adopt the technology select into the treatment group. Moreover, we want to avoid the threat of omitted variable bias, where a variable that we do not control for is driving the apparent effect of having a question answered.

In the absence of the ideal experiment, we rely on archival data. We first apply a set of stringent controls and fixed effects in our "naive" approach. We then test if our estimate of the relationship between answering and technology adoption is robust by applying an IV approach, which further deals with the concern of selection into treatment.

4.2 OLS Regression

In order to test the relationship of having a question answered with adopting a technology, we first run a series of OLS regressions using the following specification:

$$\begin{aligned} Prob(Adoption)_{ijkt} = \alpha &+ \beta Answered_j \\ &+ \gamma LogTagQuestionNumber_{ijk} + \delta LogTagNumber_{jk} + \zeta LogTagCount_{jk} \\ &+ \theta SOReactionControls_j + \eta QuestionTraitControls_j \\ &+ \iota LogOtherTagQuestionNumber_{ijk} \\ &+ \upsilon_i + \phi_{kt} + \chi_t + \psi_t + \epsilon_{ijkt} \end{aligned} \tag{1}$$

where i indexes askers, j indexes questions, k indexes tags, and t indexes the day and time the question was asked. The key coefficient of interest is β, providing the estimate of how having a question answered within three hours influences technology adoption. As we discuss below, we include a variety of controls for Stack Overflow activity around the question (*SO Reaction Controls*) and question characteristics (*Question Trait Controls*), while υ_i, ϕ_{kt}, χ_t, and ψ_t represent user, tag-year, date, and hour fixed effects, respectively. We cluster the standard errors at the question level, since answering occurs at the level of the question.

Table 3 provides the results of our more "naive" specification with having a question answered within three hours as the main independent variable of interest. In the Appendix, Table A1, we include the results of our most stringent model for other time cutoffs. Figure 1, presents the coefficients from estimating Equation 1 for one to 24 hours. The dependent variable in these regressions, *Adopted*, indicates that the asker's next job included the focal tagged technology. The key independent variable, *Answered*, is a binary variable indicating whether the question received an accepted answer within three hours of the question being posted. Across all specifications, we find a positive, statistically significant association between having a question answered and technology adoption. Each specification includes controls for *Log Tag Question Number*, *Log Tag Number*, and *Log Tag Count*. In column (1) of Table 3, where these are the only controls, the coefficient on *Answered* is .0104, indicating that having a question answered within three hours is associated with a 1.04 percentage point higher likelihood of adoption. Compared to an average likelihood of adoption of .048, this represents a 21.7% (.0104/.048) increase in the probability of adoption. This result is statistically significant at the 1% level.

In column (2), we add user fixed effects, to control for user-specific propensities to adopt technologies. The coefficient on *Answered* is .0107, indicating that having a question answered within three hours is associated with a 1.07 percentage point higher likelihood of adoption. Compared to an average likelihood of adoption of .048, this result represents a 22.3% (.0107/.048) increase in the probability of adoption.

In column (3), we add tag-year fixed effects, to control for the propensity to adopt specific technologies in a given year. The coefficient on *Answered* is reduced substantially to .00295, indicating that tag-year trends are critical to include in the specification. The magnitude of the coefficient suggests that having a question answered within three hours is associated with a .295 percentage point higher likelihood of adoption. Compared to an average likelihood of adoption of .048, this represents a 6.1% (.00295/.048) increase in the probability of adoption.

In column (4), we add date fixed effects, to control for the propensity to adopt and to have questions answered on a specific date. The coefficient on *Answered* remains largely unchanged. In column (5), we add hour fixed effects, to control for the propensity of users who post their questions at specific hours of the day to adopt and to have questions answered. Again, the coefficient on *Answered* is similar to the previous columns.

Finally, in column (6), we add a variety of additional controls relating to attributes of the questions and reactions they received on Stack Overflow. *SO Reaction Controls* include *Log Answer Count*, *Log View Count*, *Log Score*, *Log Score*, *Log Comment Count*, *Log Favorite Count*, *Log Questions in Hour*, and *Log Answers In Hour*. *Question Trait Controls* include *Log Body Length* and and *Log Title Length*. *Log Answer Count* is the inverse hyperbolic sine ("IHS") transformation of the number of answers the question received. *Log View Count* is the IHS transformation of the number of times the question was viewed. *Log Score* is the IHS transformation of the number of upvotes the question has received minus the number of downvotes it has received. *Log Comment Count* is the IHS transformation of the number of comments the question has received. *Log Favorite Count* is IHS transformation of the number of times the question has been marked as a favorite. *Log Questions in Hour* is the IHS transformation of the number of questions that were posted to Stack Overflow in the same hour as the focal question. *Log Answers in Hour* is the IHS transformation of the number of answers that were posted to Stack Overflow in the same hour as the question. *Log Title Length* is the natural log of the number of characters in the title of the question. *Log Body Length* is the natural log of the number of characters in the body of the question. *Log Other Tag Question Number* is the IHS transformation of the number of questions the asker has previously posted that did not include the focal tag. Including all the above controls does not reduce the coefficient on *Answered* substantially. This most conservative specification suggests that compared to an average likelihood of adoption of .048, there is a 6.3% (.003/.048) increase in the probability of adoption.

Across a wide variety of specifications, our results indicate that having a question answered within three hours is associated with a higher likelihood of adopting a related technology. Importantly, as displayed in Figure 1, this result stays consistent as we increase the time cutoff, with the effect essentially peaking at six hours and staying consistent through twenty-four. This does not imply that promptness is not important - each answering metric is cumulative of earlier measures, as any question that was answered within three hours was also answered within six hours. Table A2 provides evidence that prompt answers matter for adoption, as users whose questions are answered more than twenty-four hours after being posted are no more likely to adopt the related technology than users whose questions are never answered. In the next section, we use an IV estimation strategy to address concerns around omitted variable bias.

Insert Table 3 and Figure 1 about here.

4.3 Addressing Endogeneity: Instrumental Variable Analysis

Our results thus far provide evidence that having a question answered on Stack Overflow can increase the likelihood that the asker will adopt a technology related to the question. However, it is possible that there are a set of omitted variables that are driving both answering and adoption. For instance, users with a higher likelihood of adoption may post questions that are more likely to be answered. We therefore adopt an instrumental variable approach in order to provide a more stringent test of feasible causality.

Our instrument for *Answered* is built around a Bartik-like measure called *Predicted Answer Rate*. For each tag associated with a given question, we determine the tag that is most commonly combined with that tag, but is not tagged in the focal question. We refer to this as the *Alternative Tag*. For each question-*Alternative Tag* pairing, we calculate the rate at which a question with that tag which was asked in the same hour of the day in the same year received an accepted answer within three hours, which we refer to as the *Alternative Tag Answer Rate*. We then take the simple average across tags for the question to get *Predicted Answer Rate*. Finally, we winsorize at the 1st and 99th percentiles to handle outliers.

The instrument relies on the idea that there is a group of users who are able to answer questions about certain technologies. These users are not always present on Stack Overflow, but if other questions about the focal technology are being answered in the same hour, it indicates that these users are currently active.

We do not want to use the rates at which questions with the same tags as the focal question are being answered as an instrument, since users may choose to post because they see other questions with the same tags being answered. To avoid this, we rely on answer rates for tags that are often associated with the tags in the question, yet do not appear in this question. Askers are less likely to notice answer rates for these tags, and it is likely that the same users answer questions across related tags.

After controlling for the total amount of answers and questions in the same hour, along with controls for the hour of the day, the date, and the tag-year, this predicted answer rate should not have any significant association with technology adoption except through the channel of question answering. This is in accordance with the exclusion restriction, which states that the instrument should only affect technology adoption through the channel of question answering, conidtional on controls.

Table 4 provides the results of first stage regressions. The dependent variable in column (1) of these regressions, *Answered*, is a binary variable indicating whether the question received an accepted answer within three hours of the question being posted, while the key independent variable is the instrument *Predicted Answer Rate*. In column (2) the dependent variable is *Answered - 6 Hour*, a binary variable indicating whether the question received an accepted answer within six hours of the question being posted, and the key independent variable is *Predicted Answer Rate - 6 Hour*, a version of the instrument based on six hour answer rates instead of three hour answer rates. In both cases, a higher predicted answer rate is associated with a higher rate of receiving an accepted answer, with a statistical significance of 1%. The F-statistics are 162.39 and 160.73, respectively.

Insert Table 4 about here.

Table 5 displays the results of running IV regressions, with a *Predicted Answer Rate* as an instrument for the *Answered*. Column (1) displays result for the three-hour version of *Answered*, and indicates that having a question answered within a three hours is associated with a 5.11 percentage point higher probability of adopting the technology in the user's next job. Column (2) displays result for the one-hour version of *Answered*, and indicates that having a question answered within a 1 hour is associated with a 5.71 percentage point higher probability of adopting the technology in the user's next job. Both of these results are significant at the ten percent level.

Our results indicate that having a question answered on a Q&A website may assist users in adopting new technologies. Noticeably, our IV estimates for this effect are substantially higher than the results of our naive regression. IV regressions provide estimates of Local Average Treatment Effects (LATE), in this case, meaning that the estimate is specific to users whose questions would only be answered if they had a higher *Predicted Answer Rate*. Given that 46% of questions are answered within three hours, its possible that many of them would be answered regardless of the hour at which they were posted. Questions that may have not been answered may be especially difficult questions, and having these questions answered may have an outsize effect on adoption. In this vein, in the next section, we run a series of tests to examine which factors influence the impact of a question being answered on a user's adoption of a technology.

Insert Table 5 about here.

5 Answers, Adoption, and Heterogeneous Effects

So far, our evidence indicates that having a question promptly answered increases the likelihood that the user will adopt the related technology in their next job. This suggests that Q&A websites can play an important role in the adoption of new technologies. In what follows, we examine potential characteristics that may impact the effect of answers on adoption.

5.1 The Role of Proximity to Technology Hubs

Prior literature indicates that proximity is important for knowledge spillovers (Almeida and Kogut 1999, Jaffe et al. 1993), and especially important for the adoption of new technologies (Tambe 2014). Because Q&A websites can be posted to from anywhere in the world, it's possible that they can serve as a substitute for physical proximity. Rather than relying on nearby people for help, users can ask anyone on a Q&A website a question. This could indicate that the Q&A websites should be especially important for users who are located outside of technology hubs. Contrarily, Q&A websites may function as a complement to physical proximity. It is possible that answers from online platforms are most helpful when combined with the assistance of nearby users. In this section, we test how the proximity of users to a technology hub interacts with the effect of answering on technology adoption.

To calculate the distance of users from technology hubs, we use the location that they note in their developer story, and take the minimum distance of this location from San Francisco, San Jose, New York City, Boston, Los Angeles, Austin, Houston, Seattle, Chicago, Atlanta, San Diego, Washington, D.C, Denver, Dallas, Minneapolis, Detroit, Phoenix, and Philadelphia. These are the principal cities of the metropolitan statistical areas with the top fifteen highest employment of either "Computer Programmers" (Occupational Code 15-1251) or "Software Developers and Software Quality Assurance Analysts and Testers" (Occupational Code 15-1256) according to the US Bureau of Labor Statistics' May 2019 edition of the Occupational Employment and Wage Statistics.[3] Only users located in the United States are included in this analysis. We then calculate *Hub* as equal to 1 if the asker is located within 30 kilometers of a technology hub, and 0 if they are more than 30 kilometers from a hub. For robustness, we also test other distances as cutoffs, as described below.

To determine whether the adoption effects differ based on hub proximity, we run OLS regressions using the following specification:

$$\begin{aligned} Prob(Adoption)_{ijkt} = {} & \alpha + \beta Answered_j \times Hub_i + \gamma Answered_j \\ & + \delta Hub_i + \zeta LogTagQuestionNumber_{ijk} + \\ & + \theta SOReactionControls_j + \eta QuestionTraitControls_j \\ & + \iota LogOtherTagQuestionNumber_{ijk} + \upsilon_i + \phi_{kt} + \chi_t + \psi_t + \epsilon_{ijkt} \end{aligned} \tag{2}$$

where i indexes askers, j indexes question, k indexes tags, and t indexes the day and time the question was asked. The coefficient β is the key figure, providing the estimate of how being farther from a hub impacts the effect answering on technology adoption. As in our prior specifications, we include a variety of controls for Stack Overflow activity around the question (*SO Reaction Controls*) and question characteristics (*Question Trait Controls*), while υ_i, ϕ_{kt}, χ_t, and ψ_t represent user, tag-year, date, and hour fixed effects, respectively. We cluster the standard errors at the question level.

Table 6 provides the results from examining the differential impact of answering on technology adoption within thirty kilometers of a hub. Effectively, we seek to test whether Q&A websites function as a substitute or a complement for physical proximity. Across specifications, we find there is no relationship with having a question answered absent proximity to a hub. The positive relationship with between having a question answered and adoption are driven by those observations in close proximity to a hub.

[3] This excludes cities in several large metropolitan areas, such as Miami, Baltimore, and St.Louis.

In columns (1) and (2) we present the results of running the regressions with three and six hour answer times, respectively. Three hours represents our main measure because we are specifically interested in quick answers, whereas six hours serves as our alternative because the coefficient represents the local maximum in our naive regression, see Table A1. Our results reveal that being located in a hub increases the likelihood of adoption after having a question answered by *Answered* by .86 and .91 percentage points, respectively. These results are both statistically significant at the five percent levels. The technology adoption of users appears to be more heavily influenced by having their question answered when they are located inside of a technological hub. This finding is consistent with the notion that Q&A websites function best as a complement to physical proximity, rather than as a substitute.

Figure 2 displays the results using different levels of distance from a technology hub as the cutoff, ranging from ten to a hundred kilometers. Note that cutoff coefficients are cumulative - if a user is located within ten kilometers of a hub, they are also located within forty kilometers of a hub. Using a cutoff of thirty through forty kilometers results in a positive interaction that is significant at the five percent level. Meanwhile, using ten, or twenty kilometers as a cutoff results in a smaller positive interaction that is significant at the ten percent level. Past forty kilometers,[4] the results are smaller and statistically insignificant. This indicates that the effects are strongest for those who are located quite close to a technology hub. See Table A3 for exact coefficients.

Insert Table 6 and Figure 2 about here.

5.2 The Role of Technology Nascency

The adoption of new technologies can be constrained based on the number of avenues through which knowledge about them can be obtained. Nascent technologies are less likely to be taught through codified sources, such as university classes and textbooks, which makes adoption of these technologies more dependent on physical proximity and hiring (Tambe 2014, Desmet and Rossi-Hansberg 2009). By providing an alternative avenue for obtaining knowledge about a technology, Q&A websites might play an especially important role in the adoption of nascent technologies. Contrarily, Q&A websites might be most be useful for adoption when combined with other sources

[4] Note that users are located within forty kilometers of a technology hub in 49.7% of US observations.

of knowledge, and thus might be most useful for the adoption of established technologies. Merely having a question answered on a Q&A website might not be sufficient to drive adoption if the user does not have other resources to consult in order to facilitate their adoption of the focal technology.

To measure the extent to which a tagged technology has alternative avenues of knowledge, we measure the number of answers that have previously been provided for questions with a focal tagged technology, which we label *Previous Tag Answers*. Previous answers provide a base of knowledge that the user can utilize for their adoption of the technology, while also serving as a proxy for the other codified sources of knowledge about the technology. Because this measure is skewed rightward and includes zeroes, we calculate the IHS transformation of *Previous Tag Answers* as *Log Previous Tag Answers* and use this as our key operating variable for this portion of the analysis. As an alternative, for simpler interpretation, we also generate the binary variable *High Previous Tag Answers*, which is equal to 1 if *Previous Tag Answers* is in the top decile.

To determine whether the effects of having a question answered differs based on the alternative avenues of knowledge, we run OLS regressions using the same specification as in Equation 2, but replacing *Hub* with *Log Previous Tag Answers* or *High Previous Tag Answers*.

Table 7 below provides the results from examining the differential impact of answering for technologies with greater *Previous Tag Answers*. Technologies with more previous answers tend to be older and have more alternative avenues to knowledge. Across specifications, we find that adoption of technologies with greater *Previous Tag Answers* is more heavily influenced by answering on Stack Overflow.

In columns (1) and (3) we present the results of running the regressions interacting *Log Previous Tag Answers* with three and six hour answer measures, respectively. We find a positive interaction between *Log Previous Tag Answers* and *Answered* in both specifications. These results are both statistically significant at the one percent level. Additionally, in columns (2) and (4) we present the results of running the regressions interacting *High Previous Tag Answers* with three and six hour answer measures, respectively, and find a positive interaction between *Log Previous Tag Answers* and *Answered* in both specifications. Taken together, these results suggest that having a question answered on Stack Overflow is primarily helpful for adoption when a technology is more established. Overall, these findings indicate that Q&A websites likely best serve as a complement to other knowledge sources for facilitating technology adoption.

Insert Table 7 about here.

5.3 The Role of Technology Choice and Cumulative Effort to Learn

For many tasks, users have the ability to choose from a wide range of technologies in order to accomplish the same goal. In addition to assisting with users' ability to use technologies in which they are already interested, Q&A websites may play a key role in determining users' choice of technologies. If a user's first question about a technology is answered, this may encourage them to focus on that technology rather than explore other technologies that they might use for the same task. Contrarily, if Q&A websites do not play a significant role in technological choice, we may expect that having a question answered will have a larger effect on adoption when the user has asked previous questions about the same technology, since this indicates a greater cumulative effort to learn to use the technology.

To measure the influence of Stack Overflow on the user's technological search, we count the number of previous questions the user has asked with the same tag, which we label *Tag Question Number*. Because this measure is right-skewed and includes zeroes, we take the IHS transformation to generate the variable *Log Tag Question Number*. To determine whether the effects of having a question answered on adoption differs based on the user's previous questions about the technology, we run OLS regressions using the same specification as in Equation 2, but replacing *Hub* with *Log Tag Question Number*.

Table 8 below provides the results from examining the differential impact of answering for technologies with a greater *Log Tag Question Number*. Observations with a higher *Log Tag Question Number* involve users that have previously expressed stronger interest in the technology. Our results provide evidence that adoption of technologies with a greater *Log Tag Question Number* is more heavily influenced by answering on Stack Overflow.

In columns (1) and (2) we present the results of running the regressions with three and six hour answer times, respectively. With the three hour answer time, we find that having a *Log Tag Question Number* that is ten percent higher increases the adoption effect associated with *Answered* by .07 percentage points. In other words, every additional ten percent higher in *Log Tag Question*

Number raises the adoption increase associated with *Answered* by 36% (.0007/.00192). This result is statistically significant at the ten percent level. However, our result with the six hour answer time is not statistically significant on conventional levels. It is possible that quicker answers are especially important to users who have already been making efforts to adopt the technology and whose questions may come at a more critical juncture in their attempt to adopt. Still, our results provide suggestive evidence that having a question answered on Stack Overflow is more helpful for adoption when the user already has a strong interest in the technology. This indicates that Q&A websites serve less of a role in technological search, but instead help users adopt technologies on which they are already focused. In other words, having a question answered helps askers learn to use technologies in which they are interested.

Insert Table 8 about here.

---

6 Mechanism

Taken together, our results suggest that there is a positive association between having a question answered (within 24 hours) and adopting a technology in the next job. Moreover, important heterogeneity exists: a) the closer to technology hubs, the more likely adoption will occur, b) the more questions have been answered (a repertoire of knowledge exists), the more likely having a question answered leads to subsequent adoption, and c) the more questions an individual poses about a specific technology, the more likely adoption will occur after having a question answered. These findings suggest that the underlying mechanism is unlikely to be a case of technology "discovery," where a worker finds out that a technology exists and then adopts it. It is more likely, based on the results from the previous sections, that the user has already identified a technology and is seeking help in learning how to use it in real time.

Our interpretation of the set of findings we present is that a question answered on Stack Overflow may help a user "learn to use" a technology and develop relevant skills rather then discover new tools. Understanding the use of such online help platforms in enabling technology adoption is critical from a firm perspective. Prior work suggests that the ability of workers to develop new skills and learn to use new technologies can be a key driver of firm performance (Cohen and Levinthal 1990).

Our results provide evidence that virtual communities like Stack Overflow may provide an avenue to promote this, though with important boundary conditions related to factors such as location and the existence of an established knowledge basis. Online communities such as Stack Overflow support, but do not appear to substitute for, conventional avenues of technology adoption.

7 Conclusion

In this paper, we examine the potential impact of Stack Overflow, a popular Q&A website, on the adoption of technologies. Using information on 7,669 users located in the US and Canada who posted 66,006 questions on Stack Overflow, we find that users whose questions are answered within three hours are significantly more likely to adopt the technologies that are "tagged" in their questions in their next job than users whose questions are not answered or are answered after 24 hours. The magnitude of the effect is substantial, suggesting a 6.3% increase in the probability of adoption from the mean. We address potential endogeneity to the extent possible concerns by implementing an instrumental variable approach.

Critically, we find that users located in technological hubs, more established technologies, and users who have already asked more questions about a focal technology experience the largest impact. These findings provide suggestive evidence that Q&A websites can play a particularly important role in the adoption of technologies by sophisticated users with specific problems. In addition, our findings suggest that the feedback provided by virtual communities can assist with the adoption of new technologies across distance, yet function more strongly as a complement to physical proximity, rather than as a substitute.

Our findings suggest that assistance from virtual communities can help users learn to use new technologies. Users in our setting have already tagged their questions with the technologies they want to learn, and our results indicate that the relationship between having questions answered and adoption is higher when the user has already asked questions about the technology. This suggests the use of Stack Overflow is not a case of technology discovery, but rather of skill development. Feedback from virtual communities may help workers adopt technologies they want to learn, which prior literature indicates is a critical challenge (Tambe 2014).

The core findings of our paper contribute to three main streams of literature. The first is a large stream of literature emphasizing the importance of physical proximity for the diffusion of knowledge (Almeida and Kogut 1999, Jaffe et al. 1993). Because digital tools can be distributed across distances at near zero transportation costs, it may seem that their diffusion should be less sensitive to geography. Despite the reduction in costs, studies have provided evidence that benefits of digital adoption has flowed disproportionately to geographic clusters (Forman et al. 2012), suggesting that physical proximity remains important for the diffusion of digital tools, especially for emerging technologies (Tambe 2014). We provide additional support, but highlight important nuances in terms of user-heterogeneity.

A second stream of work suggests that online communities could serve as a potential channel for technological diffusion across a widely distributed workforce, democratizing technology access. Users around the world participate in online Q&A platforms such as the SAP Community Network (Huang et al. 2022) and Stack Overflow (Xu et al. 2020) posting questions and providing answers to other community members from virtually anywhere on the globe. Much of this literature examines the beneficiaries of online communities, and the factors driving these benefits (Forman et al. 2008, Mudambi and Schuff 2010, Nagle 2018, Seo et al. 2021, Huang et al. 2022, Mollick 2016). Our focus is directed towards understanding important interactions with other features of the users, their locations, and the existing knowledge base, highlighting important limits to such potential reach and expected benefits.

A third body of work dives deeper into career motivations and outcomes for online community contributors (Xu et al. 2020, Hertel et al. 2003, Huang and Zhang 2016). We complement this work by providing insight on how online communities may provide an additional channel for learning to use technologies that may be needed or helpful for a new job.

Taken together, our results provide critical insights into the role that virtual communities can play in the diffusion of technologies. Our findings provide evidence that online communities can assist with the later stages of the process of adopting new technologies. We find that these communities are most helpful as a complement to other location- and knowledge-based resources, rather than serving as a substitute.

References

Acs ZJ, Anselin L, Varga A (2002) Patents and innovation counts as measures of regional production of new knowledge. *Research Policy* 31(7):1069–1085.

Allen TJ (1997) *Managing the flow of technology: Technology transfer and the dissemination of technological information within the R&D organization* (Cambridge, Massachusetts: MIT Press).

Almeida P, Kogut B (1999) Localization of knowledge and the mobility of engineers in regional networks. *Management Science* 45(7):905–917.

Angst CM, Agarwal R, Sambamurthy V, Kelley K (2010) Social contagion and information technology diffusion: The adoption of electronic medical records in U.S. hospitals. *Management Science* 56(8):1219–1241.

Audretsch DB, Keilbach MC, Lehmann EE (2006) *Entrepreneurship and Economic Growth* (Oxford: Oxford University Press), ISBN 0195183517.

Blum BS, Goldfarb A (2006) Does the internet defy the law of gravity? *Journal of International Economics* 70(2):384–405.

Carlino G, Kerr WR (2014) Agglomeration and innovation. Working Paper 20367, National Bureau of Economic Research.

Choudhury P, Foroughi C, Larson B (2020) Work-from-anywhere: The productivity effects of geographic flexibility. *Strategic Mangement Journal* 42(4):655–683.

Cohen WM, Levinthal DA (1990) Absorptive capacity: A new perspective on learning and innovation. *Administrative Science Quarterly* 35(1):128–152.

Desmet K, Rossi-Hansberg E (2009) Spatial growth and industry age. *Journal of Economic Theory* 144(6):2477–2502.

Dewan S, Ganley D, Kraemer KL (2010) Complementarities in the diffusion of personal computers and the internet: Implications for the global digital divide. *Information Systems Research* 21(4):925–940.

Dutta S, Armanios DE, Desai JD (2022) Beyond spatial proximity: The impact of enhanced spatial connectedness from new bridges on entrepreneurship. *Organization Science* 33(4):1620–1644.

Feldman B (2017) The hidden power of Stack Overflow: How a website you've never heard of is holding the web together. *New York Magazine* URL `https://nymag.com/intelligencer/2017/03/the-hidden-power-of-stack-overflow.html`.

Fichman RG, Kemerer CF (1997) The assimilation of software process innovations: An organizational learning perspective. *Management Science* 43(10):1345–1363.

Forman C, Ghose A, Goldfarb A (2009) Competition between local and electronic markets: How the benefit of buying online depends on where you live. *Management Science* 55(1):47–57.

Forman C, Ghose A, Wiesenfeld B (2008) Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic market. *Information Systems Research* 19(3):291–313.

Forman C, Goldfarb A, Greenstein S (2012) The internet and local wages: A puzzle. *American Economic Review* 102(1):556–575.

Gray JV, Siemsen E, Vasudeva G (2015) Colocation still matters: Conformance quality and the interdependence of R&D and manufacturing in the pharmaceutical industry. *Management Science* 61(11):2760–2781.

Hertel G, Niedner S, Herrmann S (2003) Motivation of software developers in open source projects: an Internet-based survey of contributors to the linux kernel. *Research Policy* 32(7):1159–1177.

Huang P, Ceccagnoli M, Forman C, Wu D (2022) IT knowledge spillovers, absorptive capacity, and productivity: Evidence from enterprise software. *Information Systems Research* 33(3):765–1118.

Huang P, Zhang Z (2016) Participation in open knowledge communities and job-hopping: Evidence from enterprise software. *MIS Quarterly* 40(3):785–806.

Jacobs J (1969) *The Economy of Cities* (New York: Vintage Books).

Jaffe AB, Trajtenberg M, Henderson R (1993) Geographic localization of knowledge spillovers as evidenced by patent citations. *The Quarterly Journal of Economics* 108(3):577–598.

Kabo FW, Cotton-Nessler N, Hwang Y, Levenstein MC, Owen-Smith J (2020) Proximity effects on the dynamics and outcomes of scientific collaborations. *Research Policy* 43(9):1469–1485.

Karimi J, Walter Z (2015) The role of dynamic capabilities in responding to digital disruption: A factor-based study of the newspaper industry. *Journal of Economics and Management Strategy* 32(1):39–81.

Lane JN, Ganguli I, Gaule P, Guinan E, Lakhani KR (2021) Engineering serendipity: When does knowledge sharing lead to knowledge production? *Strategic Management Journal* 42(6):1215–1244.

Lee S (2019) Learning-by-moving: Can reconfiguring spatial proximity between organizational members promote individual-level exploration? *Organization Science* 30(3):467–488.

Lewis KK (1999) Trying to explain home bias in equities and consumption. *Journal of Economic Literature* 37(2):571–608.

Lin M, Viswanathan S (2016) Home bias in online investments: An empirical study of an online crowdfunding market. *Management Science* 62(5):1393–1414.

Marshall A (1890) *Principles of Economics* (London: Macmillan).

Mollick E (2016) Filthy lucre? Innovative communities, identity, and commercialization. *Organization Science* 27(6):1472–1487.

Mudambi SM, Schuff D (2010) What makes a helpful online review? A study of customer reviews on amazon.com. *MIS Quarterly* 34(1):185–200.

Mulchandani P (2022) Sunsetting jobs & developer story. `https://meta.stackoverflow.com/questions/415293/sunsetting-jobs-developer-story` [Accessed: 06/27/2023].

Nagle F (2018) Learning by contributing: Gaining competitive advantage through contribution to crowdsourced public goods. *Organization Science* 29(4):569–587.

Patel N (2022) The people who make your apps go to Stack Overflow for answers. Here's how it works. *The Verge* URL `https://www.theverge.com/23421320/stack-overflow-ceo-interview-prashanth-chandrasekar-software-engineering-microsoft`.

Porter ME (1996) Competitive advantage, agglomeration economies, and regional policy. *International Regional Science Review* 19(1-2):85–90.

Roche M, Oettl A, Catalini C (2022) (Co-)working in close proximity: Knowledge spillovers and social interactions. Working paper, Harvard Business School.

Roche MP (2020) Taking innovation to the streets: Microgeography, physical structure, and innovation. *Review of Economics and Statistics* 102(5):912–928.

Rosenthal SS, Strange WC (2003) Geography, industrial organization, and agglomeration. *Review of Economics and Statistics* 85(2):377–393.

Scott A, Storper M (2003) Regions, globalization, development. *Regional Studies* 37(6-7):579–593, ISSN 0034-3404.

Seo E, Nagle F, Shah SK (2021) Does who helps you impact your behavior? Examining the effects of social interactions on knowledge sharing in online communities. Working paper, Harvard Business School.

Sørensen JB, Sorenson O (2003) From conception to birth: Opportunity perception and resource mobilization in entrepreneurship. *Geography and Strategy*, volume 20, 89–117 (Emerald Group Publishing Limited).

Sorenson O, Audia PG (2000) The social structure of entrepreneurial activity: Geographic concentration of footwear production in the United States, 1940–1989. *American Journal of Sociology* 106(2):424–462.

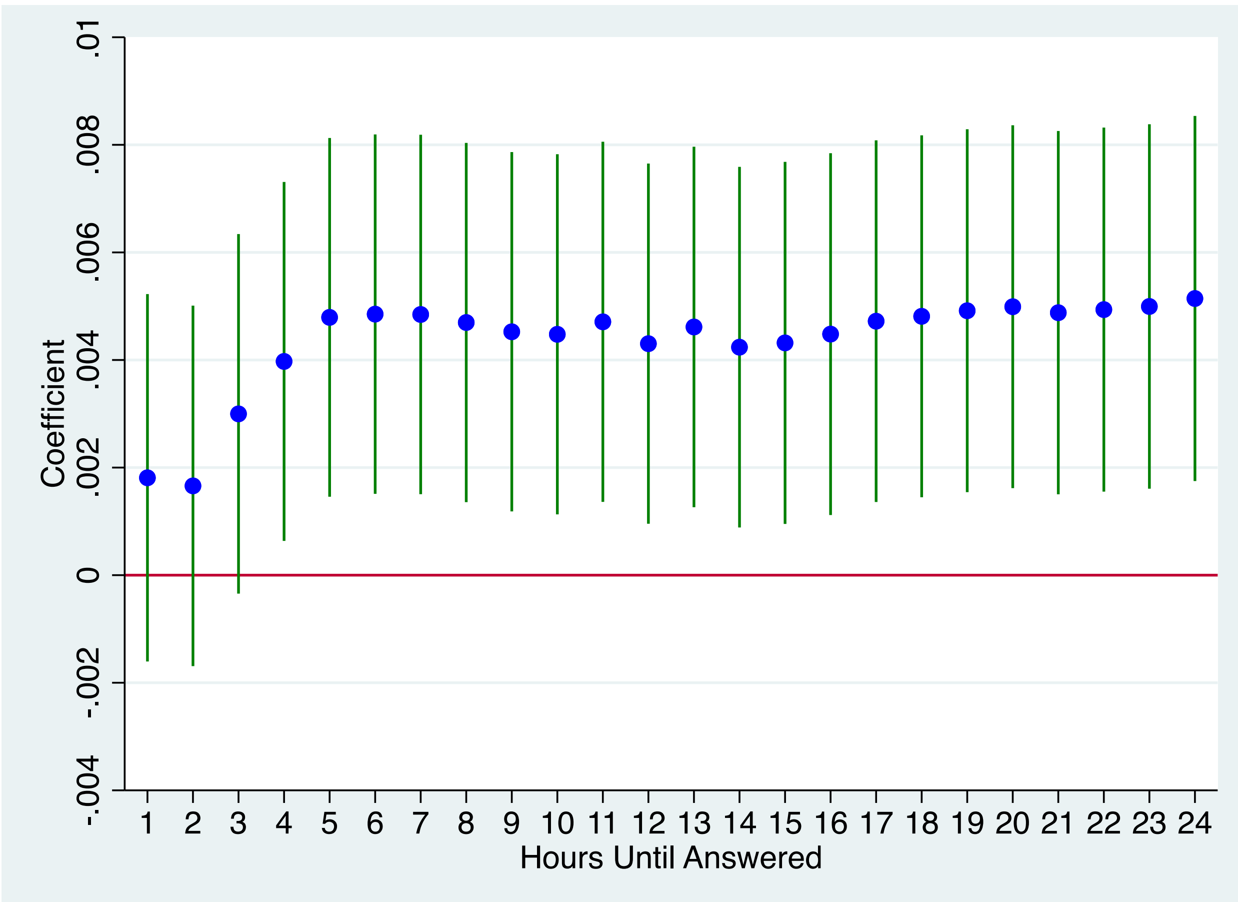
Tambe P (2014) Big data investment, skills, and firm value. *Management Science* 60(6):1452–1469.

Teece DJ, Pisano G, Shuen A (1997) Dynamic capabilities and strategic management. *Strategic Management Journal* 18(7):509–533.

Wales WJ, Parida V, Patel PC (2013) Too much of a good thing? Absorptive capacity, firm performance, and the moderating role of entrepreneurial orientation. *Strategic Management Journal* 34(5):622–633.

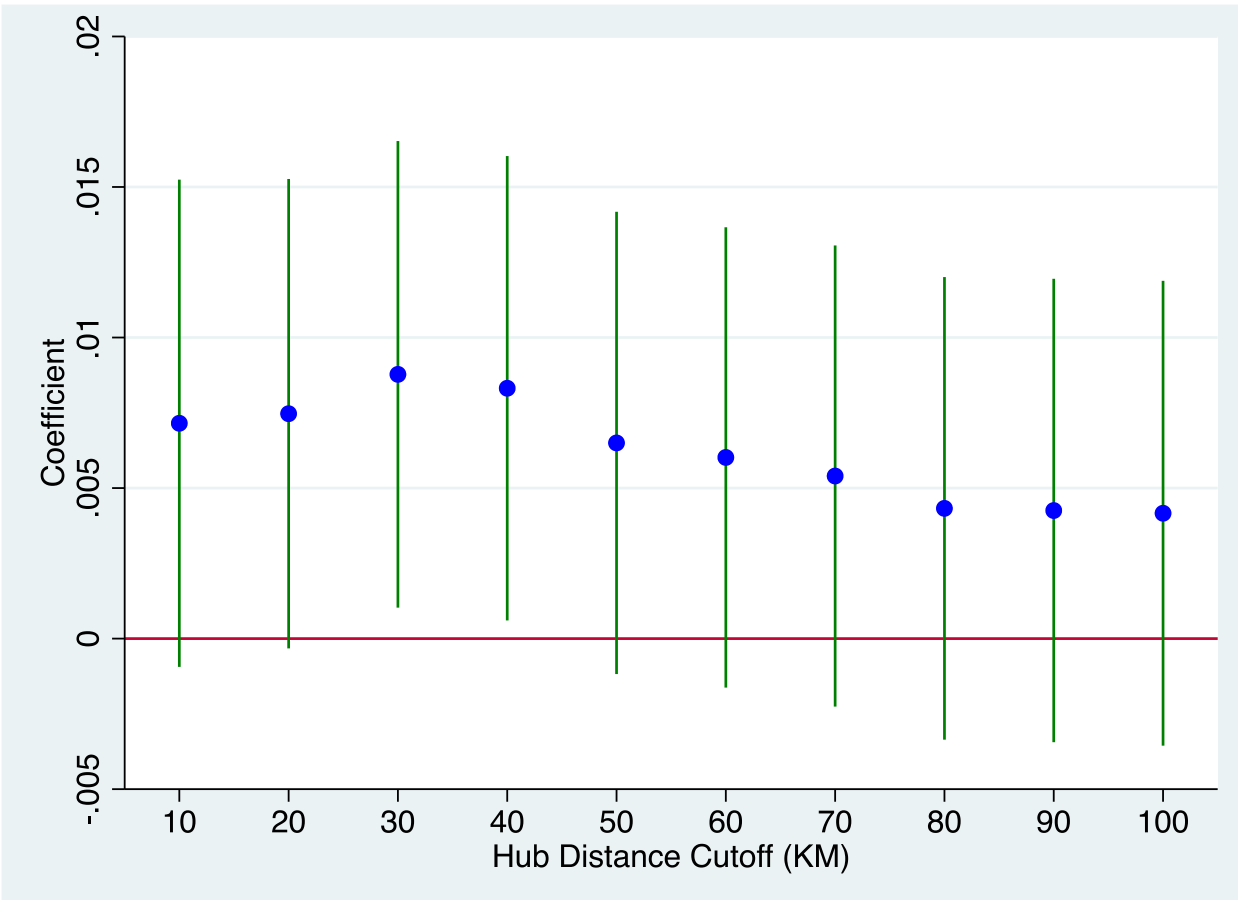
Xu L, Nian T, Cabral L (2020) What makes geeks tick? A study of Stack Overflow Careers. *Management Science* 66(2):587–604.

Figure 1: "Naive" Coefficients By Hour



Notes: This figure displays the coefficients and 95% confidence intervals resulting from re-running Equation 1 with different versions of *Answered*, where each version varies in the amount of time allowed before an accepted answer. For instance, *Answered - 6 Hour* denotes a binary variable indicating whether a question received an accepted answered within 6 hours of being posted. Note each answering metric is cumulative of earlier measures, as any question that was answered within three hours was also answered within six hours. Standard errors are clustered by question.

Figure 2: Interaction Coefficients by Hub Distance Cutoff



Notes: This figure displays the coefficients and 95% confidence intervals resulting from by interacting *Answered* with a binary variable indicating the user is located within a certain number of kilometers of a technology hub. Each coefficient uses a different distance cutoff. *Hub - 60 KM*, for instance, denotes a binary variable indicating whether the user was located within 60 kilometers of a technology hub. The coefficient above 60 KM in this figure is associated with the interaction of *Answered* with this binary variable. *Answered* is a binary variable indicating whether a question received an accepted answer within three hours of being posted. Each coefficient is generated in a separate regression. The dependent variable, *Adopted*, is a binary variable indicating that the user tagged the technology in their next job.

Table 1: Sample and Population Statistics

	Mean	SD	p10	p50	p90	Obs.
Panel A: Sample Statistics						
Questions	21.721	50.062	2.000	8.000	50.000	7,669
Answers	86.457	323.151	1.000	17.000	175.000	7,669
Answers Accepted	35.607	174.711	0.000	5.000	65.000	7,669
Question Tags	35.755	52.006	4.000	19.000	82.000	7,669
Accepted Answer Tags	35.887	77.257	0.000	13.000	85.000	7,669
Question Answered Share	0.639	0.293	0.167	0.674	1.000	7,669
Panel B: Population Statistics						
Questions	8.595	26.906	1.000	2.000	18.000	254,184
Answers	19.691	161.984	0.000	1.000	30.000	254,184
Answers Accepted	7.653	86.560	0.000	0.000	10.000	254,184
Question Tags	14.852	28.412	2.000	6.000	34.000	254,184
Accepted Answer Tags	8.816	36.319	0.000	0.000	20.000	254,184
Question Answered Share	0.472	0.395	0.000	0.500	1.000	254,184

Notes: This table reports summary statistics for the sample of users in our dataset (Panel A) and the total population of users in the United States and Canada who asked a question (Panel B). *Question* refers to the number of questions the user asked. *Answers* refers to the number of answers the user supplied. *Answers Accepted* refers to the number of accepted answers the user provided. *Question Tags* refers to the number of different tags the user attached to their questions. *Accepted Answer Tags* indicates the number of different tags attached to questions where the user supplied the accepted answer. *Question Answered Share* is defined as the share of the user's questions that received an accepted answer.

Table 2: Summary Statistics

	Obs.	Mean	SD	p10	p50	p90
Adopted	116,976	0.048	0.214	0.000	0.000	0.000
Answered	116,976	0.461	0.498	0.000	0.000	1.000
Answered - 1 Hour	116,976	0.377	0.485	0.000	0.000	1.000
Answered - 6 Hour	116,976	0.500	0.500	0.000	0.000	1.000
Answered - 12 Hour	116,976	0.534	0.499	0.000	1.000	1.000
Answered - 24 Hour	116,976	0.569	0.495	0.000	1.000	1.000
Predicted Answer Rate	108,720	0.436	0.106	0.296	0.434	0.579
Predicted Answered Rate - 6 Hour	108,720	0.464	0.106	0.326	0.462	0.607
Tag Question Number	116,976	0.254	2.005	0.000	0.000	0.000
Tag Number	116,976	2.458	1.101	1.000	2.000	4.000
Tag Count	116,976	3.050	1.149	2.000	3.000	5.000
Answer Count	116,976	1.757	1.970	1.000	1.000	3.000
View Count	116,976	4758.842	30415.875	86.000	750.000	7196.000
Score	116,976	5.082	43.047	0.000	1.000	8.000
Comment Count	116,976	1.871	2.678	0.000	1.000	5.000
Favorite Count	116,976	1.455	16.107	0.000	0.000	2.000
Answers in Hour	116,976	352.710	155.155	159.000	346.000	562.000
Questions in Hour	116,976	225.998	107.392	95.000	209.000	378.000
Hub	81,841	0.474	0.499	0.000	0.000	1.000
Other Tag Question Number	116,976	39.399	88.361	0.000	11.000	99.000
Title Length	116,912	55.354	20.485	32.000	53.000	83.000
Body Length	116,912	1563.232	1816.524	387.000	1071.000	3058.000
Previous Tag Answers	116,976	37558.618	101524.000	197.000	4169.500	94574.000
Question	66,006					
User	7,669					
Tag	5,827					

Notes: This table reports summary statistics on an Question-Tag basis. *Adopted* refers to whether the user attached the relevant tag to their next job. *Answered* indicates whether the question received an answer within three hours of the question being asked that was marked as accepted. *Answered - 6 Hour* indicates whether the question received an answer within six hours of the question being asked that was marked as accepted. *Predicted Answer Rate* indicates the predicted likelihood that the question would receive an accepted answer within three hours, based on answer rates for related tags. *Predicted Answer Rate - 6 Hour* indicates the predicted likelihood that the question would receive an accepted answer within six hours. *Tag Question Number* refers to how many questions the user had previously asked about the tagged technology. *Tag Number* indicates the order in which the tag was attached to the question, with one indicating first and five indicating fifth. *Tag Count* is the total number of tags associated with the question. *Answer Count* is the number of answers the question received. *View Count* is the number of times the question was viewed. *Score* is the number of upvotes the question has received minus the number of downvotes it has received. *Comment Count* is the number of comments the question has received. *Favorite Count* is the number of times the question has been marked as a favorite. *Answers in Hour* is the number of answers that were posted to Stack Overflow in the same hour as the question. *Questions in Hour* is the number of questions that were posted to Stack Overflow in the same hour as the focal question. *Hub* is equal to 1 if the asker is located within thirty kilometers of a technology hub, and 0 if they are more than thirty kilometers from a hub. *Other Tag Question Number* is the number of questions the asker has previously posted that did not include the focal tag. *Title Length* is the number of characters in the title of the question. *Body Length* is the number of characters in the body of the question. *Previous Tag Answers* is the number of answers that have have previously been provided for questions with the same tag.

Table 3: Naive Regression Results

DV: Adopted	(1)	(2)	(3)	(4)	(5)	(6)
Answered	0.0104***	0.0107***	0.00295*	0.00304*	0.00307*	0.00300*
	(0.00134)	(0.00133)	(0.00166)	(0.00166)	(0.00167)	(0.00170)
Log Tag Question Number	0.00918***	0.0182***	0.00307	0.00332	0.00334	0.00348
	(0.00183)	(0.00191)	(0.00216)	(0.00215)	(0.00215)	(0.00215)
Tag Number	-0.0370***	-0.0346***	-0.00209***	-0.00163*	-0.00158*	-0.00154*
	(0.000680)	(0.000702)	(0.000806)	(0.000833)	(0.000834)	(0.000844)
Tag Count	0.0148***	0.0104***	0.00206**	0.00187**	0.00186**	0.00154*
	(0.000673)	(0.000703)	(0.000830)	(0.000837)	(0.000837)	(0.000839)
Log Answers in Hour						0.0103
						(0.00851)
Log Questions in Hour						-0.00213
						(0.00895)
Log Score						0.00207**
						(0.00101)
Log Answer Count						-0.00147
						(0.00157)
Log Comment Count						-0.000606
						(0.000809)
Log Favorite Count						-0.00147
						(0.00129)
Log View Count						-0.000555
						(0.000666)
Log Body Length						0.00227**
						(0.00111)
Title Length						0.0000116
						(0.0000432)
Log Other Tag Question Number						-0.0101***
						(0.00122)
User FE	No	Yes	Yes	Yes	Yes	Yes
Tag-Year FE	No	No	Yes	Yes	Yes	Yes
Date FE	No	No	No	Yes	Yes	Yes
Hour FE	No	No	No	No	Yes	Yes
Adjusted R-Squared	0.03482	0.09978	0.1541	0.1592	0.1593	0.1603
Observations	116,976	116,362	106,665	106,617	106,617	106,560

Notes: The unit of analysis is Question-Tag. Standard errors are clustered at the question level. The dependent variable, *Adopted*, is a binary variable indicating that the user tagged the technology in their next job. *Log Tag Question Number* is the IHS transformation of number of questions the user had previously asked about the tagged technology. *Log Tag Number* is the natural log of the order in which the tag was attached to the question, with one indicating first and five indicating fifth. *Log Tag Count* is the natural log of the total number of tags associated with the question. *Log Answer Count* is the IHS transformation of the number of answers the question received. *Log View Count* is the IHS transformation of the number of times the question was viewed. *Log Score* is the IHS transformation of the number of upvotes the question has received minus the number of downvotes it has received. *Log Comment Count* is the IHS transformation of the number of comments the question has received. *Log Favorite Count* is IHS transformation of the number of times the question has been marked as a favorite. *Log Questions in Hour* is the IHS transformation of the number of questions that were posted to Stack Overflow in the same hour as the focal question. *Log Answers in Hour* is the IHS transformation of the number of answers that were posted to Stack Overflow in the same hour as the question. *Log Title Length* is the natural log of the number of characters in the title of the question. *Log Body Length* is the natural log of the number of characters in the body of the question. *Log Other Tag Question Number* is the IHS transformation of the number of questions the asker has previously posted that did not include the focal tag.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 4: First Stage Regression Results

	(1) Answered	(2) Answered - 6 Hour
Predicted Answer Rate	0.495***	
	(0.0388)	
Predicted Answered Rate - 6 Hour		0.506***
		(0.0399)
User FE	Yes	Yes
Tag-Year FE	Yes	Yes
Date FE	Yes	Yes
Hour FE	Yes	Yes
Order Controls	Yes	Yes
SO Reaction Controls	Yes	Yes
Question Trait Controls	Yes	Yes
F-Statistic	162.39	160.73
Observations	98,896	98,896

Notes: The unit of analysis is Question-Tag. Standard errors are clustered at the question level. *Answered* is a binary variable indicating that the question received an accepted answer within three hours of being asked. *Answered - 6 Hour* is a binary variable indicating that the question received an accepted answer within six hours of being asked. *Predicted Answer Rate* indicates the predicted likelihood that the question would receive an accepted answer within three hours, based on answer rates for related tags. *Predicted Answer Rate - 6 Hour* indicates the predicted likelihood that the question would receive an accepted answer within six hours, based on answer rates for related tags. Order Controls include include *Log Tag Question Number*, *Log Tag Number*, *Log Tag Count*, and *Log Other Tag Question Number*. SO Reaction Controls include *Log Answer Count*, *Log View Count*, *Log Score*, *Log Comment Count*, *Log Favorite Count*, *Log Questions in Hour*, and *Log Answers in Hour*. Question Trait Controls include *Log Body Length* and *Log Title Length*.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 5: IV Regression Results

	(1)	(2)
DV: Adopted		
Answered	0.0511*	
	(0.0306)	
Answered - 6 Hour		0.0571*
		(0.0307)
User FE	Yes	Yes
Tag-Year FE	Yes	Yes
Year-Month FE	Yes	Yes
Hour FE	Yes	Yes
Order Controls	Yes	Yes
SO Reaction Controls	Yes	Yes
Question Trait Controls	Yes	Yes
Observations	98,896	98,896

Notes: The unit of analysis is Question-Tag. Standard errors are clustered at the question level. *Answered* is a binary variable indicating that the question received an accepted answer within three hours of being asked. *Answered - 6 Hour* is a binary variable indicating that the question received an accepted answer within six hours of being asked. The dependent variable, *Adopted*, is a binary variable indicating that the user tagged the technology in their next job. Order Controls include include *Log Tag Question Number*, *Log Tag Number*, *Log Tag Count*, and *Log Other Tag Question Number*. SO Reaction Controls include *Log Answer Count*, *Log View Count*, *Log Score*, *Log Comment Count*, *Log Favorite Count*, *Log Questions in Hour*, and *Log Answers in Hour*. Question Trait Controls include *Log Body Length* and *Log Title Length*.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 6: Hub Regression Results

DV: Adopted	(1)	(2)
Answered=1	-0.000694	
	(0.00279)	
Answered=1 × Hub	0.00861**	
	(0.00394)	
Answered - 6 Hour=1		0.000957
		(0.00281)
Answered - 6 Hour=1 × Hub=1		0.00911**
		(0.00395)
User FE	Yes	Yes
Tag-Year FE	Yes	Yes
Date FE	Yes	Yes
Hour FE	Yes	Yes
Order Controls	Yes	Yes
SO Reaction Controls	Yes	Yes
Question Trait Controls	Yes	Yes
Observations	72,542	72,542

Notes: The unit of analysis is Question-Tag. Standard errors are clustered at the question level. *Answered* is a binary variable indicating that the question received an accepted answer within three hours of being asked. *Answered - 6 Hour* is a binary variable indicating that the question received an accepted answer within six hours of being asked. The dependent variable, *Adopted*, is a binary variable indicating that the user tagged the technology in their next job. *Hub* is equal to 1 if the asker is located within 30 kilometers of a technology hub, and 0 if they are more than 30 kilometers from a hub. For the purposes of this calculation, hubs include San Francisco, San Jose, New York City, Boston, Los Angeles, Austin, Houston, Seattle, Chicago, Atlanta, San Diego, Washington, D.C, Denver, Dallas, Minneapolis, Detroit, Phoenix, and Philadelphia. Only users located in the United States are included in this analysis. Order Controls include include *Log Tag Question Number*, *Log Tag Count*, *Log Tag Number*, and *Log Other Tag Question Number*. SO Reaction Controls include *Log Answer Count*, *Log View Count*, *Log Score*, *Log Comment Count*, *Log Favorite Count*, *Log Questions in Hour*, and *Log Answers in Hour*. Question Trait Controls include *Log Body Length* and *Log Title Length*.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 7: Previous Tag Answer Regression Results

DV: Adopted	(1)	(2)	(3)	(4)
Answered=1	-0.0242***	0.00108		
	(0.00820)	(0.00160)		
Answered=1 × Log Previous Tag Answers	0.00289***			
	(0.000958)			
Answered=1 × High Previous Tag Answers=1		0.0171**		
		(0.00802)		
Answered - 6 Hour=1			-0.0299***	0.00212
			(0.00812)	(0.00160)
Answered - 6 Hour=1 × Log Previous Tag Answers			0.00370***	
			(0.000950)	
Answered - 6 Hour=1 × High Previous Tag Answers=1				0.0247***
				(0.00802)
User FE	Yes	Yes	Yes	Yes
Tag-Year FE	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes
Hour FE	Yes	Yes	Yes	Yes
Order Controls	Yes	Yes	Yes	Yes
SO Reaction Controls	Yes	Yes	Yes	Yes
Question Trait Controls	Yes	Yes	Yes	Yes
Observations	106,560	106,560	106,560	106,560

Notes: The unit of analysis is Question-Tag. Standard errors are clustered at the question level. *Answered* is a binary variable indicating that the question received an accepted answer within three hours of being asked. *Answered - 6 Hour* is a binary variable indicating that the question received an accepted answer within six hours of being asked. The dependent variable, *Adopted*, is a binary variable indicating that the user tagged the technology in their next job. *Log Previous Tag Answers* is the IHS transformation of the number of questions with the focal tag that received accepted answers before the date the question was posted. Only users located in the United States are included in this analysis. *High Previous Tag Answers* is a binary variable indicating whether the the number of questions with the focal tag that received accepted answers before the date the question was posted is above the median. Order Controls include include *Log Tag Question Number*, *Log Tag Number*, *Log Tag Count*, and *Log Other Tag Question Number*. SO Reaction Controls include *Log Answer Count*, *Log View Count*, *Log Score*, *Log Comment Count*, *Log Favorite Count*, *Log Questions in Hour*, and *Log Answers in Hour*. Question Trait Controls include *Log Body Length* and *Log Title Length*.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 8: Question Number Results

	(1)	(2)
DV: Adopted		
Answered=1	0.00192	
	(0.00175)	
Answered=1 × Log Tag Question Number	0.00685*	
	(0.00397)	
Answered - 6 Hour=1		0.00386**
		(0.00175)
Answered - 6 Hour=1 × Log Tag Question Number		0.00625
		(0.00394)
User FE	Yes	Yes
Tag-Year FE	Yes	Yes
Date FE	Yes	Yes
Hour FE	Yes	Yes
Order Controls	Yes	Yes
SO Reaction Controls	Yes	Yes
Question Trait Controls	Yes	Yes
Observations	106,560	106,560

Notes: The unit of analysis is Question-Tag. Standard errors are clustered at the question level. *Answered* is a binary variable indicating that the question received an accepted answer within three hours of being asked. *Answered - 6 Hour* is a binary variable indicating that the question received an accepted answer within six hours of being asked. The dependent variable, *Adopted*, is a binary variable indicating that the user tagged the technology in their next job. *Log Tag Question Number* is the IHS transformation of the number of questions the user had previously asked about the tagged technology. Only users located in the United States are included in this analysis. Order Controls include include *Log Tag Question Number*, *Log Tag Count*, *Log Tag Number*, and *Log Other Tag Question Number*. SO Reaction Controls include *Log Answer Count*, *Log View Count*, *Log Score*, *Log Comment Count*, *Log Favorite Count*, *Log Questions in Hour*, and *Log Answers in Hour*. Question Trait Controls include *Log Body Length* and *Log Title Length*.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Online Appendix for:

Learning to Use: Stack Overflow and Technology Adoption

Table A1: Answer Speed Regression Results

Answer Measure	Coefficient	(Standard Error)
Answered - 1 Hour	0.00181	(0.00174)
Answered - 2 Hour	0.00166	(0.00171)
Answered - 3 Hour	0.00300*	(0.00170)
Answered - 4 Hour	0.00397**	(0.00170)
Answered - 5 Hour	0.00479***	(0.00170)
Answered - 6 Hour	0.00485***	(0.00170)
Answered - 7 Hour	0.00485***	(0.00170)
Answered - 8 Hour	0.00470***	(0.00170)
Answered - 9 Hour	0.00452***	(0.00170)
Answered - 10 Hour	0.00448***	(0.00171)
Answered - 11 Hour	0.00471***	(0.00171)
Answered - 12 Hour	0.00430**	(0.00171)
Answered - 13 Hour	0.00461***	(0.00171)
Answered - 14 Hour	0.00424**	(0.00171)
Answered - 15 Hour	0.00432**	(0.00172)
Answered - 16 Hour	0.00448***	(0.00172)
Answered - 17 Hour	0.00472***	(0.00172)
Answered - 18 Hour	0.00481***	(0.00172)
Answered - 19 Hour	0.00492***	(0.00172)
Answered - 20 Hour	0.00499***	(0.00172)
Answered - 21 Hour	0.00488***	(0.00172)
Answered - 22 Hour	0.00494***	(0.00173)
Answered - 23 Hour	0.00499***	(0.00173)
Answered - 24 Hour	0.00514***	(0.00173)
User FE	Yes	
Tag-Year FE	Yes	
Date FE	Yes	
Hour FE	Yes	
Order Controls	Yes	
SO Reaction Controls	Yes	
Question Trait Controls	Yes	

Notes: The unit of analysis is Question-Tag. Each row is a separate regression, with the regressions varying the number of hours before a question received an accepted answer. *Answered - 6 Hour*, for instance, indicates whether the question received an accepted answer within six hours of being posted. Note each answering metric is cumulative of earlier measures, as any question that was answered within three hours was also answered within six hours. The standard errors displayed in parentheses are clustered at the question level. The dependent variable, *Adopted*, is a binary variable indicating that the user tagged the technology in their next job. *Log Tag Question Number* is the IHS transformation of the number of questions the user had previously asked about the tagged technology. Order Controls include include *Log Tag Question Number*, *Log Tag Number*, *Log Tag Count*, and *Log Other Tag Question Number*. SO Reaction Controls include *Log Answer Count*, *Log View Count*, *Log Score*, *Log Comment Count*, *Log Favorite Count*, *Log Questions in Hour*, and *Log Answers in Hour*. Question Trait Controls include *Log Body Length* and *Log Title Length*.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A2: Late Answer Regression Results

	(1) Adopted
Answered - Over 24 Hours	-0.00523**
	(0.00256)
Never Answered	-0.00511***
	(0.00196)
User FE	Yes
Tag-Year FE	Yes
Date FE	Yes
Hour FE	Yes
Order Controls	Yes
SO Reaction Controls	Yes
Question Trait Controls	Yes
Observations	106,560

Notes: The unit of analysis is Question-Tag. *Answered - Over 24 Hours* indicates that the question received an accepted answer, but it was provided more than 24 hours after the question was posted. *Never Answered* indicates the question never received an accepted answer. Standard errors are clustered at the question level. The dependent variable, *Adopted*, is a binary variable indicating that the user tagged the technology in their next job. *Log Tag Question Number* is the IHS transformation of the number of questions the user had previously asked about the tagged technology. Order Controls include include *Log Tag Question Number*, *Log Tag Number*, *Log Tag Count*, and *Log Other Tag Question Number*. SO Reaction Controls include *Log Answer Count*, *Log View Count*, *Log Score*, *Log Comment Count*, *Log Favorite Count*, *Log Questions in Hour*, and *Log Answers in Hour*. Question Trait Controls include *Log Body Length* and *Log Title Length*.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A3: Varying Hub Distance Regression Results

Variable	Coefficient	(Standard Error)
Answered=1 × Hub - 10 KM=1	0.00757*	(0.00407)
Answered=1 × Hub - 20 KM=1	0.00891**	(0.00396)
Answered=1 × Hub - 30 KM=1	0.00861**	(0.00394)
Answered=1 × Hub - 40 KM=1	0.00675*	(0.00394)
Answered=1 × Hub - 50 KM=1	0.00466	(0.00393)
Answered=1 × Hub - 60 KM=1	0.00375	(0.00395)
Answered=1 × Hub - 70 KM=1	0.00271	(0.00399)
Answered=1 × Hub - 80 KM=1	0.00201	(0.00401)
Answered=1 × Hub - 90 KM=1	0.00193	(0.00402)
Answered=1 × Hub - 100 KM=1	0.00197	(0.00404)
User FE	Yes	
Tag-Year FE	Yes	
Date FE	Yes	
Hour FE	Yes	
Order Controls	Yes	
SO Reaction Controls	Yes	
Question Trait Controls	Yes	

Notes: The unit of analysis is Question-Tag. Each row is a separate regression, with the regressions varying the cutoff in the number of kilometers the user is from a technology hub. *Hub - 60 KM*, for instance, denotes a binary variable indicating whether the user was located within 60 kilometers of a technology hub. This part of the analysis only includes users located in the United States. Note each distance variable is cumulative of earlier measures, as any user that was located within ten kilometers of a hub was also located within fifty kilometers of a hub. The standard errors displayed in parentheses are clustered at the question level. The dependent variable, *Adopted*, is a binary variable indicating that the user tagged the technology in their next job. *Log Tag Question Number* is the IHS transformation of the number of questions the user had previously asked about the tagged technology. Order Controls include include *Log Tag Question Number*, *Log Tag Number*, *Log Tag Count*, and *Log Other Tag Question Number*. SO Reaction Controls include *Log Answer Count*, *Log View Count*, *Log Score*, *Log Comment Count*, *Log Favorite Count*, *Log Questions in Hour*, and *Log Answers in Hour*. Question Trait Controls include *Log Body Length* and *Log Title Length*.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.