Reputation and Contribution in Online Question-Answering Communities

Theodoros Lappas
Stevens Institute of Technology, tlappas@stevens.edu

Chrysanthos Dellarocas
Boston University, dell@bu.edu

Seyyedeh Neda Derakhshan
Boston University, nedaa@bu.edu

Question-Answering (QA) communities have emerged as a rich source of information by allowing users to ask questions, contribute responses, and evaluate responses submitted by others. QA platforms encourage the contribution of useful content by establishing a measure of reputation, based on the volume and quality of each user's contributions. While previous work has verified the positive relationship of reputation with response volume, its relationships of with other important aspects of contribution have been overlooked. Our work addresses this gap by studying how reputation relates to users' risk-taking propensity (difficulty of questions tackled), performance (answer quality), and topical interests. First, we introduce a novel technique for measuring user ability and question difficulty and find that users tend to tackle harder questions as their ability grows. Interestingly, we also find that, while reputation is positively associated with risk-taking for users with very low reputation, the association becomes increasingly negative once the user's reputation surpasses the community's mean. Our third finding reveals the duality of the association between reputation and performance, which is positive when users tackle questions within their ability and negative when they reach beyond their expertise. Finally, we use a novel method to reveal the diamond-like shape of the correlation pattern between a user's reputation and interests. We find that low-reputation users focus on a narrow set of introductory topics, medium-tier users have a wider spectrum of interests, and top-tier experts specialize on a small set of advanced topics. Our findings have important implications for online communities and reputation-based systems.

1. Motivation

Online Question-Answering (QA) communities have emerged as the standard source of information for users seeking advice on different topics of interest. This new learning and collaboration paradigm is exemplified by *Stack Exchange* (`StackExchange`), a network of over 150 QA communities that covers a diverse array of topics (e.g. technology, science, arts, business) and received a total of 8 billion page views in 2015 alone (Ericson 2015) The network's flagship platform is `StackOverflow` [1],

[1] `http://stackoverflow.com/company/about`

a vibrant community focused on computer programming. In 2015 alone, `StackOverflow` users asked 2.5 million questions and contributed 3.2 million responses (Ericson 2015).

In order to manage such vast amounts of user-generated content (UGC), platforms utilize their user base to crowdsource the evaluation of response quality (Liu et al. 2008). Once a new response is posted, community members can assign a negative or positive vote to it, depending on whether or not they feel it properly responds to the question. Responses are then sorted according to their aggregate vote count (i.e. the difference between positive and negative votes), in descending order. In addition, the person who originally asked the question has the option to mark a response as *accepted* and further improve its rank and visibility. This feedback process allows the platform to prominently display high-quality answers at the top, while pushing low-quality responses and spam to the bottom. Given that previous work has repeatedly verified that users heavily favor top-ranked options, this ranking mechanism makes it easier to obtain useful information and improves the user's experience (Ghose et al. 2014, Guan and Cutrell 2007, Pan et al. 2007). The increased visibility offered to highly-ranked responses motivates attention-seeking users to submit high quality content (Ghosh and Hummel 2014). In fact, given that QA platforms are based on voluntary participation, the promise of an elevated status as a verified expert within the community is one of the few meaningful motivations for a user (Raban and Harper 2008).

Previous work has repeatedly verified the positive connection between a user's reputation and the volume of her contributions (Anderson et al. 2013, Grant and Betts 2013, Cavusoglu et al. 2015, Li et al. 2012, Movshovitz-Attias et al. 2013). One of the most characteristic findings of these connection is that users tend to increase their activity level as they approach a reputation milestone, typically attached to a reward mechanism such as a community badge (von Rechenberg et al. 2016, Goes et al. 2016). Recent work has also shown that users tend to adjust the volume of both their reputation-seeking and non-reputation-seeking actions after they find a new job (Xu et al. 2014). Despite these encouraging findings, contribution volume is one only *one* of many interesting aspects of a user's behavior that could be associated with her reputation. In this work, we extend the relevant literature by presenting and evaluating a theoretical framework for the connections between a user's reputation and:

1. Her propensity to be a risk taker and respond to difficult questions.
2. Her performance, which we measure via the quality of her contributions.
3. The topics of questions that she chooses to respond to.

We utilize a large dataset from `StackOverflow` to study the relationships between these elements in the context of the highly-cited theoretical frameworks on self-efficacy (Bandura 1977) and prospect theory (Kahneman and Tversky 1979). The rich relevant literature guides our hypothesis development and informs our efforts to account for the various factors that contribute to a QA

user's behavioral patterns. Our findings provide new insight on how reputation affects user contribution patterns and have immediate applications for QA communities, as well as for any platform that utilizes reputation-based mechanisms to motivate its users.

2. Background and Hypothesis Development

In his highly cited work, Bandura studied self-efficacy and its effects on behavioral patterns (Bandura 1977). According to Self Efficacy Theory (SET), a person's belief in their own efficacy affects their behavior, as it determines the amount of resources that they are willing to devote toward the completion of a task. Research based on Bandura's work focuses on two dimensions of user behavior: effort (Bandura 1986, 1997) and persistence in the face of adversity (Cervone and Peake 1986, Bandura 1986, Strecher et al. 1995). As we discuss next, self-efficacy has since been linked to multiple fundamental concepts related to human behavior, such as risk-taking and performance.

In online QA communities, a user's self-efficacy is shaped by the positive or negative feedback that she receives for her contributions. The platform aggregates these accolades into a measure of reputation, which represents the user's authority and expert status within the community (Bosu et al. 2013, Movshovitz-Attias et al. 2013). Previous research has linked self-efficacy with the volume of a user's contribution (Jin et al. 2013). Our own work builds on SET theory and other relevant theoretical constructs to frame and study the connection between reputation and three different aspects of a user's activity within the community: (i) their risk taking, (ii) their performance, and (iii) their topics of interest. Next, we present the theoretical foundation of our study on each aspect in the context of the relevant literature.

2.1. Reputation and Risk-Taking

In QA communities, users build their reputation by receiving positive feedback for their contributions. For instance, the `StackOverflow` platform computes the user's reputation at a specific point in time based on the positive and negative votes that the she has received for her previous responses up to that point, as well as on the number of her previous responses that have been marked as *accepted* by the users who asked the corresponding questions [2]. Previous work has repeatedly verified the role of reputation as a motivator for contribution and has studied the influence of different reputation mechanisms on the user's activity and contribution levels (Anderson et al. 2013, Grant and Betts 2013, Bosu et al. 2013, Movshovitz-Attias et al. 2013, Yu et al. 2007, Constant et al. 1996, Jin et al. 2013, Kankanhalli et al. 2005).

The simplest strategy for a user trying to build their reputation is to submit as many high-quality responses as possible. If these responses are well-received by the community, the user will

[2] `http://stackoverflow.com/help/whats-reputation`

have multiple sources of positive votes and acceptances to build her reputation. However, this strategy ignores the existence of *competition*. For instance, if a question receives many responses, the body of votes awarded by the community is more likely to be fragmented. In addition, in a highly competitive setting with many responders, a user has fewer chances to submit the first response that gets accepted by the asker (Anderson et al. 2012). The obvious strategy for reducing competition is to respond to harder questions, which are less likely to receive many high-quality responses (Bosu et al. 2013). Even though the strategy of responding to harder questions can be effective in practice, it is not without risk. In QA communities, any contribution that a user makes is openly available and subject to evaluation by the entire community: if the response is satisfactory, it could be rewarded with positive votes and even with an acceptance by the asker. However, a sub-par response is likely to be ignored or even attract criticism, expressed via harsh comments and negative votes that hurt the user's reputation. For instance, in the `StackOverflow` platform, a negative vote has a direct effect on the responder's reputation, which it reduces by 2 points.

The tradeoff between risky questions that reduce competition and low-hanging fruit with limited expected returns creates an adaptive explore-or-exploit setting, similar to the one studied by the extensive literature on organizational learning (Benner and Tushman 2003). As stated by March (1991), there is *a delicate tradeoff between the risky exploration of new possibilities and the exploitation of old certainties*. We formalize this tradeoff into a theoretical framework that models the user's risk propensity in the context of two factors: (i) her ability and (ii) her reputation within the community. These two factors shape the user's self-efficacy and, through that, her goal-setting and risk-taking behavior.

2.1.1. Ability

First, we hypothesize that the promise of reduced competition and higher reputation gains motivates users to tackle the hardest possible questions *within their ability*. Even though a user can sharpen her skills via her involvement with the community, her ability is primarily shaped by exogenous factors, such as the nature of her profession and her ongoing education. These exogenous experiences also allow the user to build her perception of her own ability (Mabe and West 1982) and, based on that perception, set goals and make decisions (Nicholls 1984). In their seminal work on ambiguity and competence in choice under uncertainty, Heath and Tversky (1991) define and investigate the *competence hypothesis*. This challenges the belief that a decision-maker's willingness to bet on an uncertain event depends only on the estimated likelihood of that event and the precision of that estimate. Instead, the competence hypothesis posits that willingness also depends on the extent to which the decision maker considers himself knowledgeable or competent in the

decision's context. The authors verify this hypothesis via a sequence of controlled experiments that eliminate alternative explanations.

Vancouver et al. (2008) leverage concepts from self-regulation theory to study the effects of self-efficacy on goal acceptance. They posit that an individual will only accept a goal if their estimate of the resources required to complete the goal does not exceed a certain threshold. In this setting, higher self-efficacy is likely to lead to lower estimates and thus has a positive relationship with goal acceptance. Sitkin and Weingart (1995) hypothesized and verified that the more successful the outcomes of a decision-maker's past decisions the higher their risk propensity. Chen et al. (1998) study the effects of self-efficacy on risk-taking in the context of entrepreneurship. Their studies on both students and business executives revealed a significant and consistent positive effect of self-efficacy on the likelihood of being an entrepreneur. Similar findings were reported by Zhao et al. (2005), whose study on 265 MBA students revealed that an individual's propensity to take risks and become an entrepreneur is fully mediated by their self-efficacy. Schunk (1990) studies the connection between self-efficacy and risk-taking in an educational context, by studying the students' practice of observing their own performance and evaluating their goal progress. The study finds that, as goals are attained and their self-efficacy increases, students are motivated to set new, more challenging goals. Finally, in their studies on sports-related activities, Slanger and Rudestam (1997) find that self-efficacy is the factor most responsible for the disinhibition associated with risk-taking.

Based on our review of the relevant literature, we expect users with higher ability to tackle harder questions. Formally:

HYPOTHESIS 1. *A user's ability is positively associated with their risk propensity, as represented by the difficulty of the questions that they choose to respond to.*

2.1.2. Reputation

The second factor that we expect to have an association with a user's risk propensity is her reputation *within the community*, which serves as her reward for her contributions a symbol of status. A possible hypothesis could be that increased reputation could enhance the user's risk-taking behavior by further boosting their self-efficacy. While we expect this to be true for newly registered users and users with very low reputation, the role of reputation as a symbol of status within the community motivates us to consider an alternative hypothesis. We expect that, as the user's reputation grows, it becomes a hard-earned asset that the user will justifiably want to protect. Seeking harder questions becomes increasingly less appealing in this stage, as it can lead to failed responses that hurt the user's status via the criticism and negative votes that they are likely to attract. Intuitively, users with higher reputations have more to lose and are thus more

likely to be risk-averse. For instance, a new community member with no reputation has nothing to lose by responding to what they consider a hard question. On the other hand, we expect that a person who has already established a solid reputation will be less motivated to take risks that could hurt their hard-earned status. Thus, as a user's reputation within the community grows, we expect them to become less likely to take risks by responding to harder questions.

The extensive literature on decision-making and risk propensity provides a solid theoretical foundation for this hypothesis. A long line of literature has build on the seminal work of Kahneman and Tversky, who presented prospect theory as a model of how people choose between probabilistic alternatives that include risk (Kahneman and Tversky 1979, Tversky and Kahneman 1985, Kahneman and Tversky 1984, Tversky and Kahneman 1991). According to prospect theory, individuals in favorable states tend to be risk-averse because they feel they have more to lose than to gain. Conversely, individuals who are in unfavorable circumstances are risk-seeking, because they feel they have little to lose. Numerous works have explored this loss-aversion phenomenon in a controlled decision-making setting (Nygren et al. 1996, Arkes et al. 1988, Isen and Patrick 1983, Isen et al. 1988). Previous work has reported similar findings in different domains. In politics, research has shown that decision makers who currently hold favor or power and, therefore, have more to lose from a wrong decision, tend to be risk-averse (Ross and Stittinger 1991). In the banking sector, the shareholder of a poorly capitalized bank will prefer a riskier investment than the shareholder of a well-capitalized bank (Green and Talmor 1986). Poorly capitalized banks have little to lose by bankruptcy, so they maximize the option value of deposit insurance by gambling in riskier assets. On the other hand, well-capitalized banks prefer less risky investments, because they have more to lose in case of bankruptcy. In industry, successful entrepreneurs may be less inclined to take risks because they have more to lose (Stewart Jr and Roth 2001).

Finally, our hypothesis is consistent with the highly-cited expected utility framework, which attributes risk aversion to diminishing marginal utility (Rabin 2000, Von Neumann and Morgenstern 2007). Conceptually, a person has lower marginal utility for additional wealth when she is wealthy than when she is poor. In our QA setting, low-reputation users are simply more motivated to prove their worth than established community members. Therefore, as a user's reputation grows, their motivation to take risks and devote the increased effort required by harder questions is likely to wane, simply because the promise of gaining even more reputation becomes decreasingly appealing.

We formalize our hypothesis as follows:

HYPOTHESIS 2. *A user's reputation is initially positively associated with their risk propensity. The association becomes increasingly negative as the user's reputation grows and they become risk-averse.*

2.2. Reputation and Performance

Previous work on QA communities has focused on predicting the performance of a response, using features related to the question, the responder, and the response itself (Harper et al. 2008, Agichtein et al. 2009). Such predictive models have often leveraged reputation as one of their predictive feature. Our work extends the relevant literature by revealing the nature of the association between performance and reputation, as well as the factors that moderate this association. Our hypothesis development builds on the rich self-efficacy literature and extends it to the domain of QA communities.

Our study of the self-efficacy literature reveals an ongoing debate on whether increased self-efficacy is positively or negatively related to performance, as well as on whether the relationship is causal. The seminal work of Bandura, which served as the foundation for a long line of relevant research efforts, advocates a causal and positive relationship that channels self-efficacy into motivation, increased effort, and better performance (Bandura 1977, 1986, 1997). According to Bandura, self-efficacy determines the amount of effort that people expend, their perseverance in the face of difficulties, and their resilience to failures. When faced with hard challenges, people who harbor self-doubts about their capabilities slacken their efforts or give up quickly. On the other hand, those who have a strong belief in their capabilities exert greater effort in the face of adversity. Hence, strong perseverance contributes to performance accomplishments. Motivated by Bandura's work, several researchers have delivered findings in favor of a positive connection. For instance, Stajkovic and Luthans (1998) found a strong positive relationship between self-efficacy and performance based on 109 studies conducted on work-related tasks. Caprara et al. (2011) showed that academic self-efficacy led to increased grades for high school and junior high school students. Lee and Ko (2010) found a positive correlation between self-efficacy and performance for nurses. Finally, in a study on about 450 employees of US banking organizations, Walumbwa et al. (2008) found self-efficacy to be positively related with supervisor-rated performance.

While numerous works support the positive relationship between self-efficacy and performance, others have challenged its causal nature and have even reported contradictory findings in favor of a *negative* relationship. For instance, in his work on perceptual control theory, Powers (1973) suggests that an individual's motivation derives from the comparison of their current state with desired states. In this context, self-efficacy is used to estimate one's current state, especially when information on that state is ambiguous (Powers 1991). Hence, higher self-efficacy can cause a person to become overconfident and overestimate their current state or, equivalently, underestimate the distance between their current and desired state. This leads to decreased effort and, ultimately, reduced performance.

The research by Vancouver et al. (Vancouver et al. 2001, 2002, Vancouver and Kendall 2006) is arguably the most characteristic of this line of literature. The authors hypothesized that the positive correlation between self-efficacy and performance reported by previous works might be due to the combined effect of two causal relationships: (1) a strong positive relationship, which is the result of the influence of past performance's on self-efficacy, as described by Bandura (Bandura 1986, 1977), and (2) a weaker negative relationship, which is the result of self-efficacy's negative influence on resource allocation and performance, as described by Powers (1991).

In order to evaluate this hypothesis, they designed a study that examined self-efficacy and performance across time. Their study, which utilized a controlled experiment based on consecutive rounds of the popular Mastermind game, provided evidence that supported the hypothesis on the dual (positive/negative) relationship between self-efficacy and performance. Specifically, the authors found that an individual's self-efficacy was (1) positively related to their performance in the previous game, and (2) negatively related to their performance on the following game. In order to validate their findings and further a support a causal connection between self-efficacy and performance, Vancouver et al. presented a deeper analysis in a follow-up paper (Vancouver et al. 2002). Their new also introduced an experimental manipulation designed to artificially induce high self-efficacy. They found that increased self-efficacy led to decreased performance future tasks, confirming the hypothesis of a causal and *negative* effect. Similar findings were reported by a follow-up study focused on student learning, in which performance was measured via exam grades (Vancouver and Kendall 2006). The study revealed self-efficacy to be negatively related to planned and reported study time (effort), as well as performance.

Our review of the literature reveals that the research community is divided on whether self-efficacy is positively or negatively associated with performance. Our first study on risk propensity in Section 2.1 focuses on two primary factors: *ability* and *reputation*. For our study on performance, it is imperative to consider a third factor: the *difficulty* of the questions that the user chooses to respond to, which serves as the dependent variable for our first study. This third factor allows us to formulate a hypothesis that reconciles the two opposing views on the effects of self-efficacy on performance.

According to Bandura, individuals who are confident about their abilities are more likely to succeed in the face of adversity, by persevering and devoting extra effort and resources (Bandura 1977). On the other hand, the work by Powers and Vancouver suggests that very high confidence can lead to the underestimation of a task's difficulty and, ultimately, to reduced effort and performance (Powers 1973, Vancouver et al. 2001). We achieve a compromise of the two theories in the context of QA communities by hypothesizing and modeling a state of *overconfidence*. We posit that, while a reasonable amount of reputation-induced confidence can be beneficial, overconfidence

can be detrimental to a user's performance. Even though this is an intuitive assumption, the need for a formal hypothesis raises the following question: *How much reputation is too much?*

An obvious approach would be to assume that every user can become overconfident after reaching a certain level of reputation, and that this "turning point" level is the same for all users. This, however, would be an oversimplifying assumption that would ignore the ability of each user, as well as the difficulty of the question that she is trying to address. Instead, we posit that the polarity of the association between reputation and performance depends on the gap between the user's ability and the difficulty of the task at hand. It is reasonable to expect that a user's reputation can lead to overconfidence and underestimation of a question's difficulty. However, this underestimation will only hurt performance if the question is beyond the user's ability. A successful response to such a challenging question would take a lot of effort. A prudent user would try to bridge the gap between her ability and the question's difficulty by carefully considering possible responses and even by doing research to enhance her grasp of the problem. If the gap is trivial, then the user will likely be able to deliver a good response with little effort. However, as the gap that the user needs to bridge becomes larger, then the added effort that she will have to devote will also become more substantial. An overconfident user would underestimate the required effort, leading to decreased performance.

Our hypothesis is motivated by previous work that has tried to reconcile the two stances on the nature of relationship between self-efficacy and performance. In a study conducted on a student population, Moores and Chang (2009) found that, while self-efficacy was positively and significantly related to performance when the entire population was considered, the relationship changed when the *level* of self-efficacy is taken into consideration. Specifically, the study showed that overconfidence (defined as the difference between self-efficacy and actual performance) leads to a significant negative relationship between self-efficacy and subsequent performance. Hmieleski and Baron (2008) study the connection of self-efficacy and performance in an entrepreneurial context. In dynamic environments, the connection was found to be positive when combined with moderate optimism, but negative when combined with high optimism. In their own study, Stone (1994) found that overconfident individuals tend to allocate less resources toward the completion of a task and be less attentive than individuals with low self-efficacy.

We formalize our hypothesis as follows:

HYPOTHESIS 3. *Reputation has a positive association with performance when a user tries to respond to a question that is within her ability. The association becomes negative if the user tackles a question that is beyond their ability.*

2.3. Reputation and User Interests

Finally, we complete our theoretical framework with a hypothesis on the connection between reputation and user interests. Even though the SET literature is rich with studies on risk-taking and performance, our review revealed very limited evidence on the connection between self-efficacy and user interests. Previous work has reported a positive correlation between self-efficacy and course interests in different academic disciplines, including math (Waller 2006) and medicine (So 2008) In a similar study on a different domain, Adachi (2004) report a strong positive connection between self-efficacy and vocational interests. In their work on information seeking, Theng and Sin (2012) found a statistically significant relationship between a user's self-efficacy and her preferences on the various features of the information system that can influence the seeking task. Finally, Lent et al. (1994) found both self-efficacy and outcome expectations to be accurate predictors of career interests which, in turn, lead to career-choice behaviors. In short, while there exists some evidence of a strong connection between self-efficacy and different types of individual preferences or interests, the relevant literature in this area does not provide a sufficient theoretical background to allow for detailed hypotheses on the nature of this connection.

Previous work on QA communities has used topic-modeling techniques to analyze question content and pair new questions with experts that are identified as relevant according to their response history (Tian et al. 2013, Pal et al. 2012b). However, the association of question topics with user reputation remains unexplored. Intuitively, we expect that users at different reputation levels are likely be attracted to different types of questions. This can be due to the fact that some topics are more challenging and, thus, tend to attract high-reputation users. In addition, we expect that novice, low-reputation users are more likely to focus on mainstream topics, while high-reputation experts are more likely to be interested in specialized topics that resonate with only a smaller fraction of the community's user base. In general, we posit that a user's interests evolve along with their reputation. Formally:

HYPOTHESIS 4. *As a user's reputation within the community evolves, the topical focus of the questions that she chooses to respond to is also likely to change.*

Even though our primary goal is to verify a strong connection between reputation and topical interests in QA communities, we also extend self-efficacy theory by revealing the exact nature of this connection. As we discuss in Section 6, our methodology allows us to accomplish both of these goals in a single study.

3. Data

We conduct our analysis on a large dataset from Stack Overflow (`StackOverflow`), a booming QA community focused on computer programming. Our dataset includes approximately 18 million

responses to 9.8 million questions that were posted on the platform between July 31, 2008 and August 6, 2016. This rich panel dataset includes the full response history of each community member. Hence, it allows us to perform a longitudinal study that examines the evolution of a user's contribution patterns as their reputation evolves with time[3]. We present a list of all the variables that we utilize in our studies in Table 1. We also report descriptive statistics on the continuous variables in Table 2.

Table 1 List of variables in the `StackOverflow` dataset.

Code	Description
`Reputation`*	The responder's reputation at the time of the response as computed by `StackOverflow`.
`Performance`	Performance is measured as the number of votes that the response received within the first 2 months after its submission. Measuring performance within a fixed-size window is necessary to control for the variance in the exposure of its response. Experiments with 3,4,5, and 6 months delivered qualitatively similar findings.
`Question Pop`*	The question's popularity, measured as the number of views that it receives per day.
`Query Exp`*	The number of previous questions by the asker. Controls for the asker's experience with question formulation.
`Ability`*	The responder's ability, measured as described in Section 3.1.
`Difficulty`*	The question's difficulty, measured as the ability of the asker.
`year`	Dummy variables that encode the year of the community's lifetime that includes the response. They control for the community's evolution and for seasonal effects.
`Resp Order`	The response's arrival order with respect to all others for the same question. Used to control for the response's visibility and competition.
`Prior Acceptance`	Binary variable. *True* if a previous response had been accepted at the time of the response. Controls for visibility and competition.
`Tags`	A collection of 200 dummy variables that correspond to the 200 most prevalent tags in `StackOverflow`. Each question can have multiple tags.
`MinToFirst`*	Number of minutes until the first response to this question. Controls for the question's visibility.
`Resp Length`	Number of characters in the response. Controls for the size of the response.

*We use the log of variables marked with a * to control for distribution skewness. Due to the small scale of the PageRank probabilities, we multiply `Difficulty` and `Ability` by 10^6 before computing the log.

The reputation variable `Reputation` is as computed according to the formula employed by `StackOverflow`. The primary contributing factors are the votes and acceptances that the user

[3] We control for time-invariant user characteristics by extending all our models with fixed effects at the user level.

Table 2 **Statistics on the continuous variables in the `StackOverflow` dataset.**

Variable	Mean	Std. Dev.
`Reputation`	4.747532	2.751619
`Performance`	1.034785	1.349477
`Question Pop`	-.5635412	1.697642
`Query Exp`	2.172372	1.60816
`Ability`	4.48902	2.306628
`Resp Order`	.9339007	1.670638
`MinToFirst`	3.38073	2.599197
`Resp Length`	857.4699	849.5706

receives for her contributions: +10 for a positive vote to response, −2 for a negative vote to a response, +15 for an accepted response, +5 for a positive vote to a question, −2 for a negative vote to a question. The full formula can be found on the `StackOverflow` website[4].

3.1. Measuring Question Difficulty

In order to evaluate our hypotheses, we need to operationalize the difficulty of a given question q. Our input for this task includes all the data that we have on both the question and the profile of the asker. The available information also includes the timing and full text of each response to q, as well as the identity and profile of every responder. Next, we discuss several alternative definitions of difficulty and identify their advantages and shortcomings. We then introduce an approach that addresses these shortcomings and takes into consideration the unique nature of online QA communities.

Question Text: Mining the full text of the question can deliver useful insight on its nature. Applying an unsupervised topic modeling algorithm (e.g. Latent Dirichlet Allocation Blei et al. (2003)) would be an example of such an approach. However, even though it is reasonable to expect that some topics are generally harder than others, it is trivial to construct counter-examples to such a general rule. For instance, a high-level scripting programming language, such as JavaScript, is arguably easier than a general purpose, performance-oriented language, such as C++. However, it is unreasonable to expect that this ranking holds for every aspect or library of these two languages; a question on an advanced JavaScript concept (e.g. closures) is likely to be easier on an entry-level aspect of C++ (e.g. constructors). Even if we dismiss the complexity of learning deep and wide topic hierarchies (Wallach 2006), a text-based operationalization of question difficulty would still need to evaluate and compare the difficulty of the included subtopics, which is an equally (if not more) challenging task. A supervised solution that tries to predict the difficulty of a question is also unlikely to be effective, as it would require a large and diverse volume of manually annotated training data.

[4] `http://stackoverflow.com/help/whats-reputation`

Response Delay: The set of responses to a question is another possible source of information that could be correlated with the question's difficulty. For instance, one could argue that harder questions are likely to exhibit a higher delay between the timestamp of their creation and the timestamp of the first response. This rule would clearly not work in practice, as it would be very sensitive to early but low-quality responses. An improved version could use the delay between the question's creation and the timestamp of the response that would ultimately be *accepted* by the asker (Hanrahan et al. 2012). While improved, this version still ignores several contributing factors. First, the delay can be influenced by the popularity or overall visibility of the question. Second, this argument assumes that all responders are equally capable or that the order of the responses is similarly correlated with the capability of the responders across questions. An example of such a correlation pattern would encode that expert users are likely to respond later than less capable community members. Even though this would indeed be promising, we found no meaningful pattern on the timing of the first acceptable response. Modeling the arrival time of an adequately proficient responder is a challenging problem in itself (Bhat et al. 2014, Berger et al. 2016). Even if we could compute an accurate model that accounts for multiple contributing factors (e.g. a survival regression model), it would still not be trivial to isolate the variance in the delay that is due to the question's difficulty.

Responder Reputation: Another alternative would be to estimate a question's difficulty based on the reputation of the responders. However, given that our goal is to evaluate whether reputation is correlated with difficulty, our regression model would include the same (or parts of the same) reputation quantity as both dependent and independent variables. This would undermine the validity and interpretability of our findings. This concern could be partially alleviated if we focused on the asker's reputation. Intuitively, a high-reputation user is more likely to be an expert and, thus, is more likely to ask difficult questions that she cannot tackle on her own. However, even this approach would naively equate a user's reputation with their ability. Even though we expect that the two are correlated, it is easy to consider counter-examples. For instance, a low-ability user can build a strong reputation by answering many easy questions. Similarly, a high-ability user may not have a high reputation, simply because she chooses not to be active.

Our Approach: A difficulty measure that does not suffer from any of the above drawbacks is the *asker's ability*. A question that perplexes an expert programmer to the point that she decides to ask the help of the community is bound to be challenging. On the other hand, questions asked by a novice programmer are more likely to be about introductory or basic concepts that are easier for the community to address. Previous work has also verified that users at higher ability levels are more likely to ask challenging questions (Wray 2010, Pal et al. 2012a).

In order to estimate the ability of all the users in the community, we use the underlying *Acceptance Graph*. The graph includes a node for each community member. There is an edge from user i to user j if i accepts a response submitted by j to one of her questions. This acceptance represents an endorsement from the asker to the responder. We observe that an endorsement from a high-ability user is more important and more indicative of the the endorsee's ability than an endorsement from a novice. This setting mirrors the link structure of the Web, in which a directed link from an important page raises the importance of the link's receiver. The importance of the nodes in the Web graph is defined in the context of the random surfer model (Brin and Page 2012). Starting from a random node, the surfer traverses the graph by randomly selecting and following one of the outgoing edges of the node that he is currently visiting. At any point in time, the surfer has a certain probability (typically set to 0.15) to teleport to a random node rather than following a link. This ensures that the surfer does not get trapped in sink nodes with no outgoing edges and makes all nodes accessible. In this setting, the highly cited PageRank algorithm can compute the importance of each node as the probability that a server arrives at that node after a large number of clicks (Page et al. 1999). Formally, the importance of a node i is defined as:

$$PR(i) = \sum_{j \in in(i)} \frac{PR(i)}{|out(i)|}, \tag{1}$$

where $in(i)$ is the set of nodes that have a link to i and $out(j)$ is the set of node that have a link from j. PageRank is an iterative algorithm that solves this recursive problem by computing the left-hand eigenvector of the normalized adjacency matrix of a given graph. The algorithm converges with few iterations even on very large graphs and has been used to estimate the authority of users in online communities and expertise networks (Ding et al. 2009, Zhang et al. 2007, Heidemann et al. 2010). The interpretation of the score depends on the nature of the input graph. In the context of our Acceptance Graph, the PageRank score of a node i represents the probability that the random surfer converges to i as the community's top expert: *the user who is most likely to have one of her responses accepted*. Recent work has successfully applied PageRank on the Acceptance graph in order to identify experts (Bouguessa et al. 2008, Zhang et al. 2007). Next, we motivate and describe our own extensions to this approach.

The straightforward approach would be to apply PageRank on the entire Acceptance Graph. However, this approach would be problematic in two ways that could result in inaccurate expertise estimates. First, the standard version of the PageRank algorithm is only applicable to unweighted graphs. In our setting, this means that we would have to falsely ignore multiple acceptances awarded from a user i to another user j. To address this, we use a generalized version of the algorithm that allows weights on the edges of the input graph (Mihalcea and Tarau 2004).

The second issue is the dismissal of the possibility that a user's ability can evolve with time. If we assume that a user's ability remains constant through time, then applying PageRank on the Acceptance Graph would be sufficient. However, this assumption is unlikely to be valid, especially since our dataset spans multiple years in the community's lifetime. Instead, we expect that the ability of at least some of the users will evolve, either via their involvement in the community or via exogenous training and experience. Similarly, a user's ability may deteriorate due to lack of practice. In order to account for such changes, we segment our data by calendar year and apply PageRank independently for each segment. For each year y, we create a separate dataset that contains only the questions and responses that appeared during y. Our motivation is that a user's ability is unlikely to change significantly over the course of a single year. It is important to note that our experiments with semester-long timeframes delivered the same findings.

The process of dividing user questions and responses according to the calendar year might affect the stability of our estimates for users that are not consistently active within the community. For instance, consider a highly active expert who made a very small number of mostly unsuccessful contributions during a given year y. Even though this user is an expert with high PageRank scores in previous years, her score for this year will be deceivingly low. In order to address this, we consider a user's performance in previous years as a prior for the computation of her current score. We achieve this by using an extended version of PageRank designed for personalized rankings (Langville and Meyer 2005). This version allows us to incorporate prior knowledge into the computation by specifying a different teleportation probability for each node. In our analysis, we utilize the PageRank scores from year y as priors for the computation of the following year $y+1$. This stabilizes our expertise estimates and reduces their sensitivity to outlier periods of atypically high or low performance.

The extended PageRank functionalities for both personalization priors and weighted edges are implemented in Python's networkx library, which we utilize in our experiments. [5].

4. Study I: Reputation and Risk-Taking

The goal of this first study is to formally evaluate Hypotheses 1 and 2, which posit that (i) users tend to take more risks by focusing on harder questions as their ability grows and (ii) even though for a user with little or no reputation the association between reputation and risk-seeking is positive, it gradually becomes negative as the user builds her reputation and becomes risk-averse in order to protect it.

[5] `https://networkx.github.io/documentation/networkx-1.10/reference/generated/networkx.algorithms.link_analysis.pagerank_alg.pagerank.html`

4.1. Methodology

We evaluate Hypotheses 1 and 2 via a regression analysis on the `StackOverflow` dataset. We refer to our model as Model I. Each response r in the dataset serves as a point in our regression. The dependent variable is set to be the `Difficulty` of the question that corresponds to each response. The primary independent variables of interest are the responder's ability (`Ability`) and reputation (`Reputation`) at the time of the response. Given that Hypothesis 2 posits a non-linear relationship between `Reputation` and `Difficulty`, we also introduce a quadratic term for `Reputation`. We note that, for completeness, we also examined the interaction between both the first-order and second-order terms of ability and reputation. However, we found that none of these terms were statistically significant, which indicates that the association of reputation with risk-taking is not moderated by the user's ability (and vice versa).

Our model controls for the order in which the response was submitted (`Resp Order`), whether or not the asker had already accepted a response when r was submitted (`Prior Acceptance`), the question's popularity and visibility (`Question Pop`, `MinToFirst`), the query experience of the asker (`Query Exp`) , the content of the question (`Tags`), and the time period during which the response was submitted (`year`). The Variance Inflation Factors (VIFs) for Model I were consistently below 3, thus addressing any collinearity concerns. We report the VIF values in Table 10 of the Appendix.

As we described in Section 3, `StackOverflow` is a panel dataset with multiple responses for each user. Therefore, we utilize a linear regression model with user fixed effects (Allison 2009, Baltagi 2008, Blackwell III et al. 2005), in order to control for the user's unobserved characteristics. We used the Hausman test to decide whether fixed or random effects were appropriate for our dataset (Hausman 1978).

Formally, the model that we fit is defined as:

$$y_{it} = \alpha + x_{it}\beta + v_i + \epsilon_{it}, \tag{2}$$

where y_{it} is the difficulty of the question that corresponds to the t_{th} response of the i_{th} user in the dataset, x_{it} represents the covariates that correspond to the same response, and β includes the coefficients that we need to estimate. Finally, $v_i + \epsilon_{it}$ is the composite error term. Here, v_i is the user-specific component. The value of this term is different for each user but remains constant for all the responses of the same user. The second component ϵ_{it} is the standard homoskedastic error term with a zero mean that is uncorrelated with itself.

Let N_i be the number of responses submitted by the i_{th} user. Then, from Equation 2, it follows that:

$$\overline{y_i} = \alpha + \overline{x}_i\beta + v_i + \overline{\epsilon}_i, \tag{3}$$

where $\overline{y}_i = \sum_{t=1}^{N_i} y_{it}/N_i$, $\overline{x}_i = \sum_{t=1}^{N_i} \overline{x}_{it}/N_i$, and $\overline{\epsilon}_i = \sum_{t=1}^{N_i} \overline{\epsilon}_{it}/N_i$. After subtracting Equation 3 from 2, we get:

$$(y_{it} - \overline{y}_i) = (x_{it} - \overline{x}_i)\beta + (\epsilon_{it} - \overline{\epsilon}_i) \tag{4}$$

Given Equations 2, 3, and 4 we can compute the coefficients in β by running standard OLS on the following fixed-effects estimator (Allison 2009):

$$(y_{it} - \overline{y_i} + \overline{y}) = \alpha + (x_{it} - \overline{x_i} + \overline{x})\beta + (\epsilon_{it} - \overline{\epsilon_i} + \overline{v}) + \overline{\epsilon} \tag{5}$$

Here, given that we have n unique users in our dataset, $\overline{y} = \sum_{i=1}^{n}\sum_{t=1}^{N_i} y_{it}/(nN_i)$, $\overline{x} = \sum_{i=1}^{n}\sum_{t=1}^{N_i} x_{it}/(nN_i)$, $\overline{\epsilon} = \sum_{i=1}^{n}\sum_{t=1}^{N_i} \epsilon_{it}/(nN_i)$, and $\overline{v} = \sum_{i=1}^{n} v_i/n$.

In order to account for potential heteroskedasticity or within-panel serial correlation in the idiosyncratic error term $e_i t$ we utilize the consistent Huber/White/sandwich VCE estimator (Stock and Watson 2008, Arellano 1987, Wooldridge 2015). While using the VCE estimator does not affect the size of the coefficients, it results in robust standard errors that are identical to those obtained by clustering on the panel variable (the id of the responder).

4.2. Results

We report the results in Table 3. For the sake of brevity, we omit the results for the 200 `Tags` variables. For each independent variable, the table includes the regression coefficient, the standard error, the p-value, and the 95% confidence interval.

Table 3 Results for Hypotheses 1 and 2: Reputation, Ability, and Risk-Taking
(Model I, Fixed-Effects Linear Regression with robust standard errors)

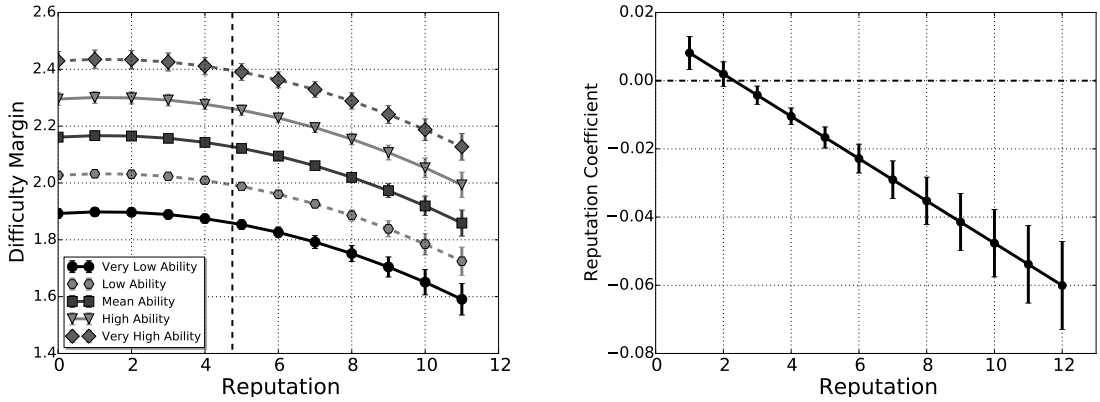
Variable	Coefficient	(Std. Err.)	t	$P>\|t\|$	[95% Conf. Interval]
`Reputation`	.0085331	.002408	3.54	0.000	[.0038134, .0132528]
`Reputation`^{2}	-.0032744	.0003767	-8.69	0.000	[-.0040127, -.0025361]
`Ability`	.0671055	.0029292	22.91	0.000	[.0613643, .0728466]
`Query Exp`	.1819425	.0011333	160.54	0.000	[.1797212, .1841638]
`MinToFirst`	.023628	.0006431	36.74	0.000	[.0223675, .0248885]
`Resp Order`	.022199	.0011385	19.50	0.000	[.0199676, .0244304]
`Prior Acceptance`	.0703971	.0036882	19.09	0.000	[.0631682, .0776259]
`year`=2010	-1.241361	.0091815	-135.20	0.000	[-1.259356, -1.223365]
`year`=2011	-2.099384	.0109089	-192.45	0.000	[-2.120765, -2.078002]
`year`=2012	-2.737978	.0123752	-221.25	0.000	[-2.762234, -2.713723]
`year`=2013	-3.249302	.0136997	-237.18	0.000	[-3.276153, -3.22245]
`year`=2014	-3.628583	.0147101	-246.67	0.000	[-3.657415, -3.599751]
`Question Pop`	.0288094	.0007979	36.11	0.000	[.0272455, .0303733]
within $R^2 = 0.31$		between $R^2 = 0.56$		overall $R^2 = 0.54$	

The first observation is the model's fit, as encoded in the reported R^2 values. The overall R^2 represents the squared correlation between the actual values and the the values that the model predicts. Formally: $corr(x_{it}\hat{\beta}, y_{it})^2 = 0.54$. The between R^2 represents the squared correlation between the actual mean difficulty value for each user and the corresponding mean that the model predicts. Formally: $corr(\overline{x}_i\hat{\beta}, \overline{y})^2 = 0.56$. Finally, the within R^2 represents the squared correlation from the mean-deviated regression. Formally: $corr\{(x_{it} - \hat{x}_i)\hat{\beta}, y_{it} - \overline{y}_i\}^2 = 0.31$. Even though the scientific literature is diverse when it comes to the evaluation of R^2 measurements, the values achieved by our model are competitive by any standard (Henseler et al. 2009, Zikmund et al. 2013).

The second observation is the large and positive coefficient for the `Ability` variable. Even though the very high significance of the coefficient could be questioned due to the large sample size (Lin et al. 2013), the extremely tight confidence interval ($0.0613643 \leq 0.0671055 \leq 0.0728466$) dismisses such concerns. The coefficient implies that, for every unit increase of the user's ability (on a log scale), we expect that the difficulty of the questions that she chooses to respond to goes up by 0.067 (on a log scale). The results provide strong evidence in favor of Hypothesis 1 and verify the positive association between ability and risk-taking.

Next, we focus on the evaluation of Hypothesis 2. We observe that, while the coefficient of the primary reputation effect `Reputation` is positive (.0085331), the coefficient of the quadratic term (`Reputation`^{2}) is negative ($-.0032744$). In both cases, the coefficients are very significant and within tight confidence intervals. The results verify that, while the effect of reputation is initially positive, it becomes weaker as the user's reputation grows. In order to establish whether or not the effect of reputation eventually becomes negative, as per Hypothesis 2, we visualize the mean predicted difficulty value for users at different ability levels for increasing reputation values, while keeping all other covariates constant (i.e. the average predictive margin). We present the plot in Figure 1(a).

Each line in the figure corresponds to one of five ability levels: Very High (`Ability`=9), High (`Ability`=7), Medium (`Ability`=5), Low (`Ability`=3), and Very Low (`Ability`=1). Note that, as we list in Table 2, the range of the `Ability` variable is $[.6132243, 11.05674]$. For all five levels, we observe a slight upward trend during the early stages of reputation. The difficulty values quickly converge and then begin to decrease, verifying that the difficulty-performance association becomes negative. The turning point is at a `Reputation` score of about 4. Interestingly, this value is very close to the mean of the `Reputation` variable for the entire community, marked by the dotted line on the plot. The figure provides strong evidence in favor of Hypothesis 2 by verifying that users tend to become risk-averse and start focusing on easier questions once their reputation surpasses a certain level.



(a) Reputation Margins

(b) Reputation Slopes

Figure 1 Predictive difficulty margins and reputation slopes for different levels of user reputation.

In order to provide additional evidence, we also compute the coefficient of the `Reputation` variable for increasing reputation values. This is the marginal effect of `Reputation`, computed as the derivative of the mean predicted value with respect to this covariate (Aiken et al. 1991). If we had not included the quadratic term $\texttt{Reputation}^2$ in the regression, the marginal effect would simply be equal to the coefficient assigned to `Reputation` via the regression. However, the presence of both the linear and quadratic terms allows us to account for the fact that the effect of increased reputation can depend on the user's current reputation level [6]. We visualize the marginal effect in Figure 1(b). The results provide further support for Hypothesis 2: while the effect of reputation is positive for users at low reputation levels, it becomes increasingly negative as reputation grows. The turning point for the coefficient is around `Reputation`$= 3$. This is slightly less than the turning point for the overall prediction, as it only consider the effect of increased reputation and not the initial value of the covariate.

Finally, we observe that the coefficients for the control variables in our model are intuitive. For instance, the positive association between risk propensity and the existence of a previous acceptance (`Prior Acceptance`) and the order of the response (`Resp Order`) is reasonable, as contributions from previous responders (and the feedback on these contributions) can help the user formulate her own response and facilitate the task of responding to a difficult question. Another facilitating factor is the quality of the question, which is (at least partially) represented by the `Query Exp` variable. The positive coefficient of `Question Pop` is also reasonable, as we expect popular questions to attract users, even if they are difficult.

[6] `https://www.ssc.wisc.edu/sscc/pubs/stata_margins.htm`

5. Study II: Reputation and Performance

The goal of this second study is to formally evaluate Hypothesis 3 which, as we discuss in detail in Section 2.2, posits that the association of reputation with performance is positive when the difficulty of the question is within the responder's ability and negative when it is beyond her ability.

5.1. Methodology

We evaluate Hypothesis 3 via a regression analysis on the `StackOverflow` response dataset. Each response r in the dataset serves as point in our regression. The dependent variable is the responder's performance for each response, as encoded in the `Performance` variable. As we discuss in detail in Section 3, `Performance` measures the number of votes that the response received within a fixed time window. We note that, while this number is primarily determined by the quality of the response, it may also be affected by other factors. These includes the difficulty, popularity, visibility, and topical focus of the question. We control for these factors via the `Difficulty`, `Question Pop`, `MinToFirst`, and `Tags` variables. The performance of a response may also be affected by the competition that it faces. We control for this by considering the number of previous responses (`Resp Order`) and whether or not a previous response has already been accepted (`Prior Acceptance`). Finally, we control for the community's evolution through time (`year`), the question experience of the asker (`Query Exp`), and the size of the response (`Resp Length`).

We consider two alternative models. For the first model, which we refer to as Model II, the primary independent variables of interest are the responder's reputation at the time of the response (`Reputation`), as well as a new variable that encodes the gap between the responder's ability and the difficulty of the question. We compute this new variable as `Abil_Diff_Gap`=`Ability`-`Difficulty`[7]. We also introduce an interaction term `Reputation#Abil_Diff_Gap` which will allow us to evaluate the role of reputation at different `Abil_Diff_Gap` values (positive and negative), as mandated by Hypothesis 3.

By introducing the `Abil_Diff_Gap` variable, Model II allows us to operationalize and evaluate the hypothesis that the effect of reputation depends on whether the question's difficulty is within or beyond the user's ability. However, by combining the two variables, this approach does not allow us to consider the effect of reputation for different levels of user ability or question difficulty. We achieve this via an alternative model that models the interaction between ability, difficulty, and reputation by introducing the 3-way interaction term `Reputation#Ability#Difficulty`, as well as the 3 possible 2-way interactions. We refer to this model as Model III. In addition to providing an alternative way to evaluate Hypothesis 3, Model III will allow us to investigate size of the

[7] The subtraction is valid because `Difficulty` is defined as the `Ability` score of the asker and thus the two variables have the same value space (See Section 3.1)

reputation effect for different types of users and questions. We report the VIF values in Table 10 of the Appendix.

Given the count nature of the dependent variable, we opt for a Poisson regression model (Cameron and Trivedi 2013). We extend our model with fixed user effects, in order to account for unobserved user characteristics. We used the Hausman test to decide whether fixed or random effects were appropriate for our dataset (Hausman 1978). Formally, let $\mathbf{y}_i = (y_i 1, ..., y_{iN_i})$ is a $N_i \times 1$ vector of counts that holds the performance achieved by user i for each of her N_i responses. Let $\mathbf{x}_i$ be a $N_i \times K$ matrix that includes the values of all K observable covariates for each response of user i ($\mathbf{x}_{it}$ is a $1] \times K$ vector, with $t = 1, 2, ..., N_i$). Finally, let ϕ_i be an unobserved, user-specific scalar effect. Then, under the Fixed Effects Poisson model, we have:

$$y_{it}|\mathbf{x}_i \sim \text{Poisson}(\phi_i \mu(\mathbf{x}_{it}, \beta)), t = 1, 2, ..., N_i,$$

where β is the vector of coefficients that we need to estimate and y_{it}, y_{ir} are independent conditional on $\mathbf{x}_i, \phi_i, t \neq r$ with $E(y_{it}|\mathbf{x}_i, \phi_i) = \phi_i \mu(\mathbf{x}_{it}, \beta)$.

In accordance with previous work, we can estimate β via the Quasi-Conditional Maximum Likelihood Estimator (QCMLE) (Wooldridge 1999, Allison 2009). We present the full specification of the model and the derivation of the log-likelihood function in the Appendix.

As we did for our study on risk propensity, we utilize the consistent Huber/White/sandwich VCE estimator to compute cluster-robust standard errors and account for potential heteroskedasticity or within-panel serial correlation in the idiosyncratic error term $e_i t$ (Stock and Watson 2008, Arellano 1987, Wooldridge 2015). Previous work has shown that this estimator is also robust to overdispersion (Cameron and Trivedi 2009, Wooldridge 1999). This is a useful property, as the Poisson estimator assumes that the mean is equal to the variance and our dependent variable exhibits a slight overdispersion. As per Table 2, `Performance` has a mean and standard deviation of 1.034785 and $1.349477^2 = 1.821088$, respectively. In order to confirm that overdispersion is not a concern for our results, we repeat the regression with a Fixed-Effects Negative Binomial Model (Hausman et al. 1984, Cameron and Trivedi 2013), which we refer to as Model IV. The results for Models II and III, which we include in Tables 8 and 9 of the Appendix, are consistent with those of the Fixed Effects Poisson Model and verify the robustness of our findings.

5.2. Results

First, we present the results for Model II in Table 4. The very low p-value and tight confidence interval for the coefficient of the interaction term verify its significance. In order to properly interpret the effect of the interaction term, we visualize the marginal effects of `Reputation` for different

values of Abil_Diff_Gap [8]. We present the plot in Figure 2(a), which provides strong evidence in support of Hypothesis 3. As stated by the hypothesis, we observe that the coefficient of reputation is positive when the user's ability surpasses the difficulty of the question (Abil_Diff_Gap> 0) and negative when it is falls short (Abil_Diff_Gap< 0). In fact, the turning point is almost exactly when ability and difficulty are equal (Abil_Diff_Gap$= 0$). Given the presence of the interaction term, the coefficient of Reputation represents the effect of reputation when the gap is equal to zero (i.e. the turning point). We observe that reputation is not significant at this level (p-value=0.472), a finding that is consistent with our hypothesis.

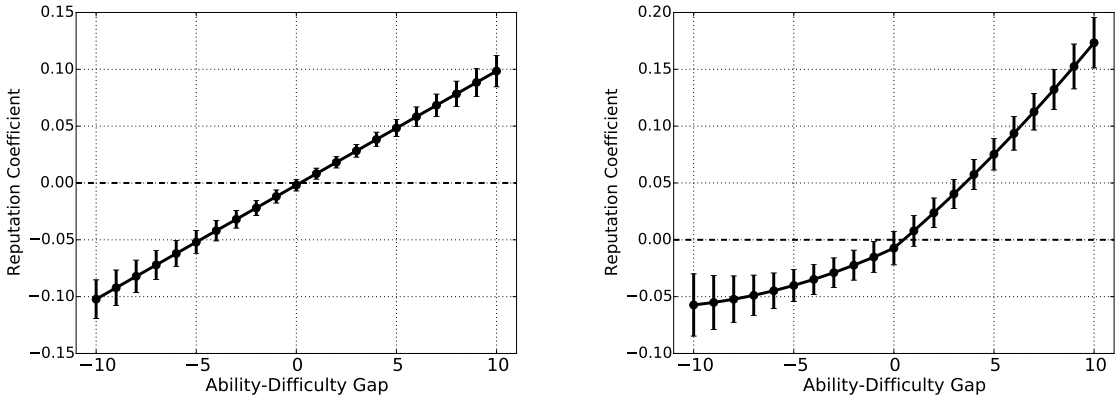
Table 4 provides some anticipated secondary findings, such as the positive association of performance with the question's popularity (Question Pop), the asker's experience with formulating questions (Query Exp), and the length of the response (Resp Length). The negative coefficients for Resp Order and MinToFirst are also intuitive: a late response is less likely to receive votes, as is a response to a low-visibility question. Finally, the negative coefficient of Prior Acceptance implies that a pre-existing accepted response is likely to attract the lion's share of the votes (due to both its quality and prominent positioning at the top of the page) and thus hurt the performance of any future response.

Table 4 Results for Hypothesis 3: Reputation and Performance
(Model II, Fixed-Effects Poisson with a (Abil_Diff_Gap) Variable and robust standard errors)

Variable	Coefficient	(Std. Err.)	t	$P>\|t\|$	[95% Conf. Interval]
Reputation	-.0018769	.0026088	-0.72	0.472	[-.0069901, .0032363]
Abil_Diff_Gap	-.0919423	.0040522	-22.69	0.000	[-.0998843, -.0840002]
Reputation#(Abil_Diff_Gap)	.0100336	.0007465	13.44	0.000	[.0085705, .0114967]
Query Exp	.0796667	.0013398	59.46	0.000	[.0770408, .0822926]
MinToFirst	-.105022	.0015779	-66.56	0.000	[-.1081146, -.1019294]
Resp Order	-.0379143	.0028529	-13.29	0.000	[-.0435058, -.0323228]
Prior Acceptance	-.3891313	.0065354	-59.54	0.000	[-.4019404, -.3763222]
year=2010	-.1009714	.0156337	-6.46	0.000	[-.131613, -.0703299]
year=2011	-.1058421	.0164551	-6.43	0.000	[-.1380935, -.0735907]
year=2012	-.070946	.0179196	-3.96	0.000	[-.1060676, -.0358243]
year=2013	-.1026408	.0193558	-5.30	0.000	[-.1405775, -.0647041]
year=2014	-.1090948	.0209103	-5.22	0.000	[-.1500781, -.0681114]
Question Pop	.0749889	.0013801	54.34	0.000	[.072284, .0776938]
Resp Length	.0001205	2.96e-06	40.77	0.000	[.0001147, .0001263]

Next, we report the results for Model III in Table 5. A first observation that the coefficients of the control variables are consistent with those reported by Model II. We also observe that the 3-way

[8] We compute the marginal effects for this 2-way interaction as per: http://www.ats.ucla.edu/stat/stata/faq/conconb12.htm.



(a) Reputation Slopes, Model II (b) Reputation Slopes, Model III

Figure 2 Predictive performance margins and reputation slopes for different levels of user reputation.

interaction term is highly significant, as evidenced by both the p-value and the tight confidence intervals. However, the presence of multiple interaction terms motivates us to visualize the behavior of the model for different values of the interacting covariates. We begin by computing the coefficient of reputation for different values of the gap between ability and reputation, as we did in Figure 2(a) for Model II [9]. However, given that Model III does not have a variable to directly encode this gap, we compute the coefficient of `Reputation` for all possible $11 * 11 = 121$ combinations of `Ability`$\in \{1,2,3,4,5,6,7,8,9,10,11\}$ and `Difficulty`$\in \{1,2,3,4,5,6,7,8,9,10,11\}$. We then compute the average coefficient for each distinct value of the `Ability-Difficulty` gap. For instance, when the gap is equal to 8, we report the average of the following (`Ability`,`Difficulty`) combinations: $(11,3)$, $(10,2)$, and $(9,1)$. We visualize the results in Figure 2(b).

We anticipated some variation between Figures 2(a) and 2(b), due to the different underlying models and the way in which we generated the plots. However, we find that the two models are consistent and fully aligned in their support of Hypothesis 3. We observe that, as in the case of Figure 2(a) the coefficient of reputation is negative when the user's ability is smaller than the difficulty of the question and positive when it is larger. Once again, the turning point comes when the two quantities become equal. The range of values of the coefficient is also highly similar between the two models.

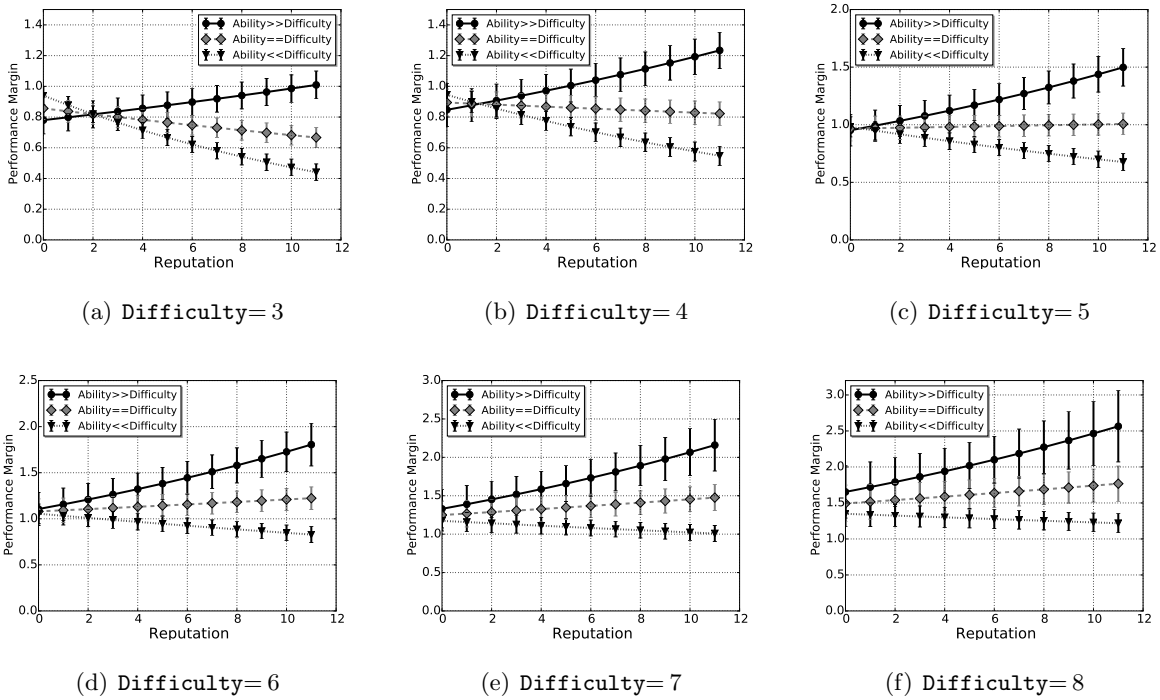
Model III allows us to obtain additional evidence in support of Hypothesis 3 by computing the predicted value of performance (i.e. the average predictive margin) for different levels `Ability` and `Difficulty` variables while tuning `Reputation` and keeping all other covariates constant. We visualize the results in Figure 3. Each plot in the figure corresponds to a different level of question

[9] We compute the marginal effects for this 3-way interaction as per: `http://www.ats.ucla.edu/stat/stata/faq/con3way12.htm`.

Table 5 Results for Hypothesis 3: Reputation and Performance

(Model III, FE Poisson with a `Reputation#Ability#Difficulty` Interaction and robust standard errors)

Variable	Coefficient	(Std. Err.)	t	$P>\|t\|$	[95% Conf. Interval]
`Reputation`	-.0904324	.0068584	-13.19	0.000	[-.1038747, -.0769901]
`Ability`	-.0912095	.0098886	-9.22	0.000	[-.1105908, -.0718283]
`Difficulty`	.0165432	.0076058	2.18	0.030	[.0016362, .0314502]
`Reputation#Ability`	.025621	.0016567	15.47	0.000	[.022374, .028868]
`Reputation#Difficulty`	.0025505	.001719	1.48	0.138	[-.0008186, .0059196]
`Ability#Difficulty`	.0169023	.0019215	8.80	0.000	[.0131362, .0206684]
`Reputation#Ability #Difficulty`	-.0018691	.0003212	-5.82	0.000	[-.0024987, -.0012395]
`Query Exp`	.0753868	.0013299	56.69	0.000	[.0727803, .0779934]
`MinToFirst`	-.1047713	.0015759	-66.49	0.000	[-.10786, -.1016827]
`Resp Order`	-.0371188	.0028277	-13.13	0.000	[-.0426611, -.0315766]
`Prior Acceptance`	-.3848028	.0065168	-59.05	0.000	[-.3975754, -.3720301]
`year=2010`	.0176871	.0184406	0.96	0.337	[-.0184558, .05383]
`year=2011`	.0682311	.0209652	3.25	0.001	[.0271401, .1093221]
`year=2012`	.1387674	.0228198	6.08	0.000	[.0940414, .1834934]
`year=2013`	.1336173	.0246241	5.43	0.000	[.085355, .1818796]
`year=2014`	.1403946	.0268052	5.24	0.000	[.0878573, .1929318]
`Question Pop`	.0750738	.0013737	54.65	0.000	[.0723813, .0777662]
`Resp Length`	.0001183	2.94e-06	40.23	0.000	[.0001125, .000124]



(a) `Difficulty`= 3 (b) `Difficulty`= 4 (c) `Difficulty`= 5

(d) `Difficulty`= 6 (e) `Difficulty`= 7 (f) `Difficulty`= 8

Figure 3 Predictive performance margins and reputation slopes for different levels of user reputation.

difficulty (i.e. 6 plots for `Difficulty`$\in \{3,4,5,6,7,8\}$) and includes three lines that corresponds to 3 different ability levels: one that is equal to the plot's `Difficulty` value (`Ability`<<`Difficulty`), one that is 1 standard deviation above `Difficulty` (`Ability`>>`Difficulty`), and one that is 1 standard deviation below `Difficulty` (`Ability`<<`Difficulty`). The consistently near-flat line for `Ability`==`Difficulty` indicates that the association of performance with reputation is trivial when the responder's ability is equal to the difficulty of the question. When the user's ability is larger (smaller) than the question's difficulty, the predicted performance becomes larger (smaller) as reputation increases, consistent with Hypothesis 3.

6. Study III: Reputation and User Interests

In this study we evaluate Hypothesis 4 which, as we discussed in detail in Section 2.3, posits the existence of a strong connection between a user's reputation and the topical focus of the questions that she chooses to respond to. Given the absence of both empirical and theoretical work in the relevant literature, our secondary goal is to provide insight on the exact nature of this connection.

6.1. Methodology

First, we use Latent Dirichlet Allocation (Blei et al. 2003) on the text of the questions in our dataset to compute the topics that are discussed by askers. We set the number of topics to $K = 100$ after using the established perplexity measure to experiment with $K \in \{25, 50, 100, 125, 150, 175\}$. LDA computes the probability distribution θ_q over the set of topics for every question q. We utilize these distributions to design our experiment as follows: let $max(\theta_q)$ be the highest-probability topic for question q. We refer to this as the question's *main topic*. Also, let $rep(u, q)$ be the reputation score that a user u had when she responded to question q. We then define R_q to be the full set of $rep(u, q)$ scores of all the users who responded to q. We refer to this set as the question's *reputation set*. Finally, let S_t be the union of the reputation sets of all questions that have topic t as their main topic.

Conceptually, S_t is the population of reputation scores that are associated with topic t. We can now compare the reputation level associated with any two topics by comparing their respective means. First, we use a standard two-sided T-test for every possible pair of topics to evaluate the null hypothesis that the two means are equal at a significance level $\alpha = 0.01$. If the null hypothesis is rejected, we use one-sided tests to determine if one of the two reputation means is significantly larger than the other. We then construct a directed unweighted graph as follows: first, we eliminate spurious topics that were consistently attached to trivial probabilities by the LDA algorithm[10]. We create a node for each of the remaining 76 topics in our dataset. We then add a directed edge from

[10] We remove every topic t that did not have $P(t|q) \geq 0.1$ for at least one question q in our data.

topic t_i to topic t_j if t_j is associated with a significantly higher reputation mean than t_i. Therefore, the ancestors of a topic t in the graph are all the topics chosen by users with a significantly lower reputation than those that choose t.

Finally, we organize the topics into levels as follows: the first level includes all topics that have no ancestors in the graph. The second level includes all topics that only have ancestors in the first level. We proceed by adding consecutive levels until all topics have been assigned. Figure 4 shows the final visualization of the topic graph. The weight on the edge from level L_i to level L_j represents the percentage of all possible edges between the two edges that actually exist in the graph. For instance, the graph includes 92% of the $6 \times 28 = 168$ possible edges between L_1 and L_3.

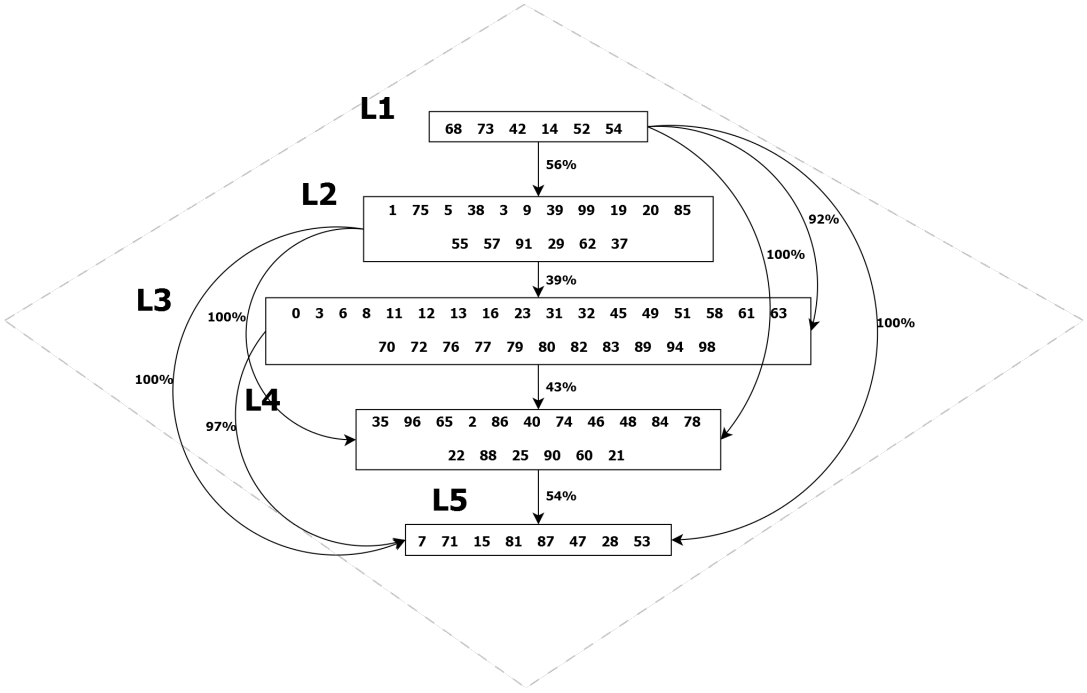


Figure 4

6.2. Results

The figure verifies the connection between self-efficacy and topical interests and provides strong evidence in support of Hypothesis 4. Specifically, the Figure reveals a clear 5-level hierarchy of topics with a diamond shape. Levels 1 and 5 are the smallest, including 6 and 8 topics, respectively. Levels 2 and 4 are noticeably larger and include the same number of topics (17). Finally, level 3 is by far the largest, including a total of 28 topics. Conceptually, level 1 includes introductory topics that attract novice, low-reputation users. As users gain experience and build their reputation, they expand their focus to cover an increasingly wider range of topics, visualized by the increasingly

larger sizes of levels 2 and 3. However, after their reputation increases even further, users shift their focus to an increasingly narrower set of elite topics in levels 4 and 5.

The directed edges connecting the graph's levels further validate its hierarchical nature and the gradual evolution of the users' topical interests. First, we observe percentages between $39\% - 56\%$ between adjacent levels. This indicates that the progression from one level to the next is gradual and the difference in reputation is not significant between all possible pairs of topics from consecutive levels. On the other hand, the percentages between non-consecutive levels are consistently $\geq 92\%$. This verifies the clear reputation gap between these levels and validates the hierarchical structure of the topic graph.

This study reveals a strong correlation between a user's reputation and her topical interests. If we also consider the clear correlation between reputation and question difficulty that we demonstrated in Section 4, it would be reasonable to consider that the levels of the topic graph represent different levels of difficulty. However, an alternative explanation is that a user's reputation shifts her focus to different and not necessarily harder topics. For instance, high-reputation users may be more attracted to questions on new and emerging topics, even if these questions are not particularly challenging. In order to obtain additional insight, we perform a secondary analysis that allows us to assign representative keywords to each of the 5 levels.

For each topic t, we first retrieve the set of questions Q_t that have a probability of at least 0.5 for t in their respective topic distributions. Formally: $Q_t = \{q : \theta_q(t) \geq 0.5\}$. These are relatively "pure" questions that are representative of this specific topic. As we discussed in Section 3, our dataset includes the set of characteristic tags $tags(q)$ that the asker assigns to her question. Therefore, given a level L_i of the diamond structure, we can compute the multiset sum M_{L_i} of all the tag-sets of all the pure questions that correspond to topics from L_i. Formally: $M_{L_i} = \biguplus_{t \in L_i, q \in Q_t} tags(q)$. Finally, we sort all the unique tags in M_{L_i} by their frequency within the multiset (i.e. the number of pure questions from L_i topics that include them in their tag-sets) and report the top 25 tags for each level in Table 6.

The table verifies that the levels of the diamond-like shape represent different genres within the broad programming domain. In addition, the terms reveal the nature of the users' transition across the levels. The first level clearly represents mobile programming (i.e. the development of mobile Apps), as all the terms included in the top 25 have a strong association with this area. An initial hypothesis could be that mobile programming tends to be easier and have a lower entry barrier than other types of software development and, thus, attracts novice users at their early reputation phases. However, there is no prior literature to support this claim. In fact, previous research has repeatedly highlighted the challenges that mobile developers have to face (Joorabchi et al. 2013, Burd et al. 2012, Lim et al. 2015). Industry experts have also dismissed this claim and have even advocated

Table 6 Most Prevalent Tags for each Topic Level of the diamond structure in Figure 4

Level 1	Level 2	Level 3	Level 4	Level 5
android	android	android	c++	java
iphone	css	java	java	c#
ios	iphone	jquery	git	c++
objective-c	ios	c#	sql	inheritance
android-layout	.htaccess	php	swing	.net
ruby-on-rails	opengl	javascript	mysql	oop
uitableview	html	html	sql-server	interface
xcode	jquery	mysql	tsql	class
ruby	c++	css	c#	jsf
layout	xcode	.net	svn	delphi
cocoa-touch	php	visual-studio-2010	c	generics
ruby-on-rails-3	apache	sql-server	regex	design-patterns
xml	java	asp.net	sql-server-2008	android
java	javascript	python	php	reflection
ipad	mod-rewrite	database	sql-server-2005	abstract-class
listview	python	visual-studio	string	jsp
android-linearlayout	c#	twitter-bootstrap	templates	jsf-2
uiviewcontroller	opengl-es	eclipse	branch	winapi
uinavigationcontroller	objective-c	sql	boost	constructor
ios5	css3	algorithm	javascript	singleton
uiview	matlab	cocos2d-iphone	github	asp.net-mvc
scrollview	listview	c++	jpanel	exception
ios4	c	android-intent	version-control	servlets
relativelayout	osx	broadcastreceiver	jquery	google-app-engine
textview	android-fragments	django	stored-procedures	eclipse

that mobile programming can be more challenging than desktop or web development Paul (2013), Cailean (2013). A more realistic hypothesis is that novice developers are attracted by the explosive growth of the mobile sector which, coupled with the well-documented talent shortage (Omojola 2013), leads to an abundance of job openings and increased salaries (PayScale 2016, Greenspan 2015). The popularity of Apps has also led to the emergence of tools that promise end-to-end mobile development with little or no coding (Maltby and Loten 2013, Bloomberg 2016). This promise is particularly appealing to novice developers, who often underestimate the effort that it takes to create a successful App (Mombrea 2016, AppPromo 2013).

The second level maintains a strong presence of mobile-related terms. In addition, the level also includes terms related to web development (e.g. *css, html, .htaccess, css3, php, apache*). We observe that most of these terms are more representative of the visual components of website design, rather than more advanced features, such as web services. Finally, the level also includes some sporadic evidence of terms related to core programming languages that have applications across domains (e.g. *c++, c, matlab, python, c#*). The third and largest level is very diverse, including terms from mobile programming, web development, core programming, and even databases (e.g. *mysql, sql, sql-server*). This intuitive finding verifies that users in this large, mid-reputation tier cover a wide range of topics within the programming universe. The fourth reputation level enhances the

focus on core programming and databases, while departing from mobile and web programming. Programming languages and database platforms are prevalent in this level, which also includes programming constructs (e.g. *string, regex*), libraries (e.g. *swing, boost*), and development tools (e.g. *svn, github*). The fifth level, which represents the interests of users at the top reputation tier, is almost exclusively dedicated to core programming languages and concepts. The list covers basic programming constructs (e.g. class, oop, exception), as well as more advanced concepts (e.g. *inheritance, generics, singleton, reflection*), indicating that the questions in this level are heavily focused on the core details of software development and not on specific applications domains.

Our study reveals significant differences in the nature of the questions included in the different reputation levels. We observe that, as users build their reputation, they tend to transition from popular application domains, such as mobile and web development, to more generic topics, such as databases and core programming concepts. One possible explanation is that application topics are easier than core programming, as they require a closed set of skills and often prioritize peripheral software components, such as interfaces. Even though the development of such components is certainly not a trivial task, it is arguably less challenging for a novice than the details and nuances of core programming. In addition, the plethora of low/no code solutions for peripheral components make their development even more approachable for novices. An alternative explanation is that users transition from specific applications to generic concepts because, as their reputation grows, both their confidence and curiosity also grow, driving them to tackle generic programming questions that go beyond the narrow bounds of a particular application domain. Even though this does not necessarily imply that such questions are harder, the two explanations are certainly compatible.

We can gain further insight if we consider our findings in the context of the popular ModelView-Controller (MVC) paradigm, which is the de facto design pattern for mobile, web, and desktop software (Krasner et al. 1988). The MVC pattern isolates business logic from the user interface (UI). The Model represents the information (the data) of the application and the business rules (e.g. algorithms) used to manipulate it. The View corresponds to elements of the user interface such as graphics, text, and checkbox items. The Controller manages the communication between the Model and View. The first level of our reputation hierarchy is mostly associated with the View, as it includes many terms relevant to the design of UIs. The second level maintains this association to some extent, although the UI-related terms are significantly less. The third level is very diverse and contains terms from all three components. The prevalence of database terms in the fourth level reveals its strong association with the Model component. Finally, The terms in the fifth level represent fundamental programming constructs that are essential for all three components. As we discuss in Section 7, our findings motivate future work on the factors that drive users to focus on different types of topics as their reputation evolves.

7. Implications and Future Work

Our study focuses on the association between reputation and key aspects of user behavior in online QA communities. We make several theoretical and methodological contributions, which we enumerate in Table 7. As we discuss next, our findings have implications both for online communities and reputation-based motivation mechanisms in general.

Table 7 The main contributions of our work.

Key findings
(1) A user's ability is positively associated with her propensity to be a risk taker, i.e. to respond to difficult questions (Section 4).
(2) A user's reputation is positively associated with her risk propensity while her reputation is low. The association becomes increasingly negative after the user's reputation surpasses a level close to the community's mean (Section 4).
(3) A user's reputation is positively associated with her performance when the difficulty of the questions that she tackles is within her ability. The association becomes increasingly negative as the user focuses on questions beyond her ability (Section 5).
(4) A user's reputation has a diamond-shaped association with her topical interests. Low-reputation users focus on a small set of introductory topics. As reputation grows, the set becomes larger. However, after a certain reputation level, the set of becomes increasingly smaller and stabilizes to a size equal to that of the introductory set. (Section 6).
Methodological Contributions
(5) A method for estimating user ability and question difficulty in QA communities (Section 3.1).
(6) A method for studying the correlation pattern between a user's topical interests and any of her time-variant characteristics, such as her reputation (Section 6).

Our first study verifies that the polarity of the association between reputation and risk-taking is not consistent; while the association is positive for users at very low reputation levels, it becomes increasingly negative as the users increase their reputation beyond the community's average level. Based on both our findings and previous theoretical work, we can attribute this change to either lack of motivation or the development of loss-aversion, as users become content with their current reputation and avoid risks that could damage it. However, at the community level, it is critical to have users that are willing to take such risks and providing valuable responses to difficult questions. Ongoing work on online communities has examined mechanisms for motivating users (Ardichvili et al. 2003, Raban and Harper 2008, Mamykina et al. 2011). Our findings can directly inform such efforts by revealing the critical turning point in the user's lifetime; the point when gaining more reputation stops being a strong motivator and might even discourage them from tackling harder challenges. This finding can help community managers time their interventions and policy changes and focus their motivational efforts on the right users at the right time.

Loss-aversion is an intuitive explanation for the finding that users tend to focus on increasingly easier questions as their reputation grows beyond a certain level. This phenomenon warrants customized policy changes for QA communities that value reputation as an asset and status symbol. Our findings provide insight on how a community can better motivate its users to take risks by tackling harder questions. The current reputation policy of `StackOverflow` and other communities provisions fixed rewards for responders, such as +10 reputation points for every positive vote to a response or +15 points for every response that the asker accepts. However, given that our study reveals that users tend to become increasingly risk-averse as their reputation grows, a fixed-reward scheme might not be appropriate. A straightforward solution could be a flexible reward policy that offers higher rewards to users at higher reputation levels. This approach could make risk-taking more attractive for high-reputation users and could also motivate users to work toward higher reputation levels with higher yields. However, such a policy would inevitably lead to rich-get-richer effects and could widen the reputation faultlines among the users. A simpler method would be to attach higher rewards to harder questions. This would require a scalable and effective way to operationalize question difficulty. Our work makes a significant methodological contribution in that direction, as we discuss in detail in Section 3.1. A third option would be to directly target the effects of loss aversion. As an example, consider a policy that rewards users with "forgiveness" badges when they reach different reputation milestones. A user could then use a forgiveness badge to protect one of her future responses. A protected response would be immune to negative votes and would thus be unable to harm the user's reputation. Future work can consider such policies and evaluate whether or not they indeed motivate users to focus on harder questions.

Our second study examines the association between reputation and performance. The self-efficacy literature, which provides the theoretical framework for our hypothesis development, has been torn on the polarity of the association between self-efficacy and performance. Our findings extend previous theoretical work by providing a compromise in the context of online QA communities. We find that, while the association is positive when users tackle questions within their ability, it becomes increasingly negative when the difficulty of the questions they are targeting is beyond their expertise. Our work is the first to make this distinction and operationalize the turning point for the effect of reputation on performance.

Our findings have direct implications for the efforts to match questions with appropriate responders, which has been the focus of an ongoing stream of research (Tian et al. 2013, Jurczyk and Agichtein 2007, Zhou et al. 2009). In QA communities, the short-term goal is to deliver high-quality responses to information seekers. However, in the long-term, a community can benefit from developing the ability of its members (Pal et al. 2011). As we also discuss in our hypothesis development in Section 2.2, users who focus on questions beyond their ability are likely to succeed only if they

allocate the time and effort required to bridge this gap. Thus, by reaching beyond their limits, users can develop their expertise and become more useful to the community. On the other hand, this is also a risky practice that could increase the number of low-quality responses and displease information seekers. Our findings on the role of reputation provide insight on how a community's management team can leverage this trade-off. QA communities like `StackOverflow` have the ability to steer responders toward specific questions, by controlling the visibility (e.g. ranking, frequency) of the questions that each user views on her personalized interface. Even though the exact algorithm employed by each platform is proprietary, we know that current recommendation engines are primarily based on the relevance between the user's interests and expertise (mined based on their contribution history) and the content of each question. [11]. Our study reveals the importance of also considering the 3-way interaction of user ability, user reputation, and question difficulty when modeling performance. By leveraging these factors, a question-recommendation engine could go beyond topical relevance and match questions with relevant users at strategically selected reputation levels. According to our findings, the practice of steering a user toward questions that surpass her ability becomes increasingly risky (i.e. is more likely to lead to low quality responses) as her reputation grows. Thus, in order to develop the ability of its members while keeping a high standard of contribution quality, a community has to employ a recommendation engine that is cognizant of both the ability-difficulty gap and the user's reputation at any point in time.

Our third study reveals a very strong association between reputation and topical interests. Even though we anticipated the existence of this association, the structured, diamond-shaped manner in which it manifests is surprising. We find that, during their early reputation stages, users tend to focus on a narrow set of introductory topics. As they develop their reputation, their set of interests grows in both diversity and size. However, after a certain reputation level, the set becomes increasingly narrower until it reaches a size comparable to that of the introductory set. Our text analysis also reveals that users at this advanced reputation stage tend to favor depth rather than breath and, thus, focus on more specialized topics. Monitoring and anticipating the user's interests is essential for an effective recommendation system (Mobasher et al. 2000). As we discussed above, current question-recommendation methods already take into consideration the topical focus of each question. However, our study verifies the need to extend current modeling efforts to also account for the user's reputation.

Another application that could benefit from a reputation-aware engine is the personalized recommendation of jobs. Large QA communities like `StackOverflow` mine the users' contributions to match them to relevant job Ads that potential employers submit to the platform. For instance,

[11] `http://stackoverflow.com/users/prediction-data`

a user who frequently responds to questions about web development is more likely to see Ads for web developer positions. Successful job recommendations are critical, as they can affect both user satisfaction and the influx of revenue from Ad placements [12]. The consideration of reputation can complement the information mined from responses and help the platform estimate the user's interests and deliver more relevant Ads.

The second implication of our third study comes from a methodological perspective. The novel graph-based method that we used to establish the shape of the correlation pattern between a continuous variable (reputation) and a set of categorical variables (user interests) can be used to perform similar studies on other important factors in the context of QA communities. For instance, a practitioner could use our approach to study the association between a user's interests and her ability or between the user's reputation and the demographics of the users that she tends to interact with.

In short, our work provides insight on the associations of reputation with key aspects of user behavior on online QA communities, as well as on how these associations can be effectively leveraged to improve critical community outcomes. It is our hope that our findings and methodological contributions will inspire and support relevant research in this domain and help QA communities maximize their potential both as sources of high-quality information and as effective learning platforms for millions of users.

Acknowledgments

Appendix.

A. The Fixed Effects Poisson Model

Let Y_{it} is the performance that corresponds to the t_{th} response of the i_{th} user in the dataset, let x_{it} represent the covariates that correspond to the same response, let α_i be the user-specific effect, and let β include the coefficients that we need to estimate. We can then formally specify the fixed-effects Poisson model (Allison 2009, Wooldridge 1999), as:

$$Pr(Y_{it} = y_{it}|x_{it}) = \exp\{\exp(\alpha_i + x_{it}\beta)\}\exp(\alpha_i + x_{it}\beta)^{y_{it}}/y_{it}! =$$

$$\frac{1}{y_{it}!}\exp\{\exp(\alpha_i)\exp(x_{it}\beta) + \alpha_i yit\}\exp(x_{it}\beta)^{y_{it}} = F_{it}$$

Because we know that the observations are independent, we may write the joint probability for the observations within a panel as

[12] http://stackoverflow.blog/2015/01/targeted-jobs-for-stack-overflow/

$$
\begin{aligned}
&Pr(Y_{i1}=y_{i1},...,Y_{in_i}=y_{in_i}|\mathbf{X}_i)\\
&\quad=\prod_{t=1}^{n_i}\frac{1}{y_{it}!}\exp\{-\exp(\alpha_i)\exp(x_{it}\beta)+\alpha_i y_{it}\}\exp(x_{it}\beta)_{it}^{y}\\
&\quad=\left(\prod_{t=1}^{n_i}\frac{\exp(x_{it}\beta)_{it}^{y}}{y_{it}!}\right)\exp\left\{-\exp(\alpha_i)\sum_t\exp(x_{it}\beta)+\alpha_i\sum_t y_{it}\right\}
\end{aligned}
$$

and we also know that the sum of n_i Poisson independent random variables, each with parameter λ_{it} for $t=1,...,n_i$, is distributed as Poisson with parameter $\sum_t\lambda_{it}$.

$$
Pr\left(\sum_t Y_{it}=\sum_t y_{it}|\mathbf{X}_i\right)=
$$

$$
\frac{1}{\sum_t y_{it}!}\exp\left\{-\exp(\alpha_i)\sum_t\exp(x_{it}\beta)+\alpha_i\sum_t y_{it}\right\}\left\{\sum_t\exp(x_{it}+\beta)\right\}^{\sum_t y_{it}}
$$

So, the conditional likelihood is conditioned on the sum of the outcomes in the set (panel). The appropriate function is given by

$$
\begin{aligned}
&Pr\left(Y_{i1}=y_{i1},...,Y_{in_i}=y_{in_i}|\mathbf{X}_i,\sum_t Y_{it}=\sum_t y_{it}\right)=\\
&\left[\left(\prod_{t=1}^{n_i}\frac{\exp(x_{it}\beta)^y it}{y_{it}!}\right)\exp\left\{-\exp(\alpha_i)\sum_t\exp(x_{it}\beta)+\alpha_i\sum_t y_{it}\right\}\right]\Big/\\
&\left[\frac{1}{(\sum_t y_{it}!}\exp\left\{-\exp(\alpha_i)\sum_t\exp(x_{it}\beta)+\alpha_i\sum_t y_{it}\right\}\left\{\sum_t\exp(x_{it}\beta)\right\}^{\sum_t y_{it}}\right]\\
&=\left(\sum_t y_{it}\right)!\prod_{t=1}^{n_i}\frac{\exp(x_{it}\beta)^{y_{it}}}{y_{it}!\{\sum_k\exp(x_{ik}\beta)\}^{y_{it}}}
\end{aligned}
\tag{6}
$$

which is free of α_i.

The conditional log likelihood is given by

$$
\begin{aligned}
L&=\log\prod_{i=1}^{n}\left[\left(\sum_{t=1}^{n_i}y_{it}\right)!\prod_{t=1}^{n_i}\frac{\exp(x_{it}\beta)^{y_{it}}}{y_{it}!\{\sum_{\ell=1}^{n_\ell}\exp(x_i\ell\beta)\}^{y_{it}}}\right]\\
&=\log\prod_{i=1}^{n}\left\{\frac{(\sum_t y_{it})!}{\prod_{t=1}^{n_i}y_{it}!}\prod_{t=1}^{n_i}p_{it}^{y_{it}}\right\}\\
&=\sum_{i=1}^{n}\left\{\log\Gamma\left(\sum_{t=1}^{n_i}y_{it}+1\right)-\sum_{t=1}^{n_i}\log\Gamma(y_{it}+1)+\sum_{t=1}^{n_i}y_{it}\log p_{it}\right\}
\end{aligned}
\tag{7}
$$

where

$$
p_{it}=e^{x_{it}\beta}\Big/\sum_\ell e^{x_{i\ell}\beta}.
$$

Table 8 Results for Hypothesis 3: Reputation and Performance

(Model IV, Fixed-Effects NegBin with a (Abil_Diff_Gap) Variable and robust standard errors)

Variable	Coefficient	(Std. Err.)	t	$P>\|t\|$	[95% Conf. Interval]
Reputation	-.0077725	.0015188	-5.12	0.000	[-.0107492, -.0047958]
Abil_Diff_Gap	-.0775172	.0024734	-31.34	0.000	[-.0823651, -.0726693]
Reputation#Abil_Diff_Gap	.01052	.0004109	25.60	0.000	[.0097146, .0113254]
Query Exp	.093173	.0008994	103.59	0.000	[.0914101, .0949358]
MinToFirst	-.0905397	.0008112	-111.62	0.000	[-.0921296, -.0889499]
Resp Order	-.0607883	.0013366	-45.48	0.000	[-.0634081, -.0581686]
Prior Acceptance	-.3268583	.0045816	-71.34	0.000	[-.3358382, -.3178785]
year=2010	-.0666523	.0093801	-7.11	0.000	[-.085037, -.0482676]
year=2011	-.0250636	.0097506	-2.57	0.010	[-.0441743, -.0059529]
year=2012	.018285	.0101604	1.80	0.072	[-.0016291, .038199]
year=2013	-.0173418	.0106291	-1.63	0.103	[-.0381745, .0034909]
year=2014	-.025567	.0114729	-2.23	0.026	[-.0480534, -.0030806]
Question Pop	.0590458	.0009259	63.77	0.000	[.057231, .0608606]
Resp Length	.8721292	.0122816	71.01	0.000	[.8480578, .8962007]
constant	.8721292	.0122816	71.01	0.000	[.8480578, .8962007]

Table 9 Results for Hypothesis 3: Reputation and Performance

(Model V, FE NegBin with a Reputation#Ability#Difficulty Interaction and robust standard errors)

Variable	Coefficient	(Std. Err.)	t	$P>\|t\|$	[95% Conf. Interval]
Reputation	-.084258	.0039658	-21.25	0.000	[-.0920308, -.0764853]
Ability	-.0378571	.0056951	-6.65	0.000	[-.0490192, -.026695]
Difficulty	.0252341	.006047	4.17	0.000	[.0133821, .037086]
Reputation#Ability	.020424	.0008644	23.63	0.000	[.0187299, .0221182]
Reputation#Difficulty	.0016923	.0013192	1.28	0.200	[-.0008933, .0042778]
Ability#Difficulty	.0150679	.0014573	10.34	0.000	[.0122118, .0179241]
Reputation#Ability #Difficulty	-.001781	.0002461	-7.24	0.000	[-.0022633 , -.0012987]
Query Exp	.0858309	.0009103	94.28	0.000	[.0840467, .0876151]
MinToFirst	-.0909003	.0008131	-111.80	0.000	[-.0924939, -.0893067]
Resp Order	-.0595484	.0013315	-44.72	0.000	[-.0621581, -.0569387]
Prior Acceptance	-.3233039	.0045783	-70.62	0.000	[-.3322772, -.3143307]
year=2010	.0899079	.010204	8.81	0.000	[.0699084, .1099074]
year=2011	.2205953	.0113937	19.36	0.000	[.1982641, .2429265]
year=2012	.3271803	.0122697	26.67	0.000	[.3031321, .3512285]
year=2013	.3407853	.0129629	26.29	0.000	[.3153786, .3661921]
year=2014	.3614311	.0137089	26.36	0.000	[.3345621, .3883001]
Question Pop	.0587612	.0009248	63.54	0.000	[.0569486, .0605738]
Resp Length	.0001164	1.45e-06	80.56	0.000	[.0001136, .0001193]
constant	.4581119	.02585	17.72	0.000	[.4074469, .5087769]

Table 10 Variance Inflation Factors

Variable	Model I	Model II	Model III
Question Pop	1.14	1.15	1.16
Reputation	2.98	2.54	3.29
Query Exp	1.01	1.05	1.07
MinToFirst	1.13	1.15	1.15
Resp Order	1.19	1.20	1.20
Resp Length	-	1.03	1.03
Ability	2.92	-	3.47
Difficulty	-	-	1.28
Abil_Diff_Gap	-	2.56	-

References

Adachi T (2004) Career self-efficacy, career outcome expectations and vocational interests among japanese university students. *Psychological reports* 95(1):89–100.

Agichtein E, Liu Y, Bian J (2009) Modeling information-seeker satisfaction in community question answering. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 3(2):10.

Aiken LS, West SG, Reno RR (1991) *Multiple regression: Testing and interpreting interactions* (Sage).

Allison PD (2009) *Fixed effects regression models*, volume 160 (SAGE publications).

Anderson A, Huttenlocher D, Kleinberg J, Leskovec J (2012) Discovering value from community activity on focused question answering sites: a case study of stack overflow. *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 850–858 (ACM).

Anderson A, Huttenlocher D, Kleinberg J, Leskovec J (2013) Steering user behavior with badges. *Proceedings of the 22nd international conference on World Wide Web*, 95–106 (International World Wide Web Conferences Steering Committee).

AppPromo (2013) The necessity of mobile app marketing. `http://app-promo.com/wp-content/uploads/2012/04/AppPromo-TheNecessityofMobileAppMarketing.pdf`, accessed: 2016-12-05.

Ardichvili A, Page V, Wentling T (2003) Motivation and barriers to participation in virtual knowledge-sharing communities of practice. *Journal of knowledge management* 7(1):64–77.

Arellano M (1987) Practitionerscorner: Computing robust standard errors for within-groups estimators. *Oxford bulletin of Economics and Statistics* 49(4):431–434.

Arkes HR, Herren LT, Isen AM (1988) The role of potential loss in the influence of affect on risk-taking behavior. *Organizational behavior and human decision processes* 42(2):181–193.

Baltagi B (2008) *Econometric analysis of panel data* (John Wiley & Sons).

Bandura A (1977) Self-efficacy: toward a unifying theory of behavioral change. *Psychological review* 84(2):191.

Bandura A (1986) *Social foundations of thought and action: A social cognitive theory.* (Prentice-Hall, Inc).

Bandura A (1997) Self-efficacy: The exercise of control.

Benner MJ, Tushman ML (2003) Exploitation, exploration, and process management: The productivity dilemma revisited. *Academy of management review* 28(2):238–256.

Berger P, Hennig P, Bocklisch T, Herold T, Meinel C (2016) A journey of bounty hunters: Analyzing the influence of reward systems on stackoverflow question response times. *Web Intelligence (WI), 2016 IEEE/WIC/ACM International Conference on*, 644–649 (IEEE).

Bhat V, Gokhale A, Jadhav R, Pudipeddi J, Akoglu L (2014) Min (e) d your tags: Analysis of question response time in stackoverflow. *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, 328–335 (IEEE).

Blackwell III JL, et al. (2005) Estimation and testing of fixed-effect panel-data systems. *Stata Journal* 5(2):202–207.

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. *the Journal of machine Learning research* 3:993–1022.

Bloomberg J (2016) OutSystems 10 Sets High Bar For Low-Code Mobile App Development. `http://www.forbes.com/sites/jasonbloomberg/2016/10/09/outsystems-10-sets-high-bar-for-low-code-mobile-app-development/#52a5b05955d5`, accessed: 2016-12-05.

Bosu A, Corley CS, Heaton D, Chatterji D, Carver JC, Kraft NA (2013) Building reputation in stackoverflow: an empirical investigation. *Proceedings of the 10th Working Conference on Mining Software Repositories*, 89–92 (IEEE Press).

Bouguessa M, Dumoulin B, Wang S (2008) Identifying authoritative actors in question-answering forums: The case of yahoo! answers. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 866–874, KDD '08 (New York, NY, USA: ACM), ISBN 978-1-60558-193-4, URL `http://dx.doi.org/10.1145/1401890.1401994`.

Brin S, Page L (2012) Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer networks* 56(18):3825–3833.

Burd B, Barros JP, Johnson C, Kurkovsky S, Rosenbloom A, Tillman N (2012) Educating for mobile computing: addressing the new challenges. *Proceedings of the final reports on Innovation and technology in computer science education 2012 working groups*, 51–63 (ACM).

Cailean I (2013) Infographic: How Hard Is It To Break Into The App Store? `http://www.trademob.com/infographic-how-hard-is-it-to-break-into-the-app-store/`, accessed: 2016-12-05.

Cameron AC, Trivedi PK (2009) *Microeconometrics using stata*, volume 5 (Stata press College Station, TX).

Cameron AC, Trivedi PK (2013) *Regression analysis of count data*, volume 53 (Cambridge university press).

Caprara GV, Vecchione M, Alessandri G, Gerbino M, Barbaranelli C (2011) The contribution of personality traits and self-efficacy beliefs to academic achievement: A longitudinal study. *British Journal of Educational Psychology* 81(1):78–96.

Cavusoglu H, Li Z, Huang KW (2015) Can gamification motivate voluntary contributions?: The case of stackoverflow q&a community. *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing*, 171–174 (ACM).

Cervone D, Peake PK (1986) Anchoring, efficacy, and action: The influence of judgmental heuristics on self-efficacy judgments and behavior. *Journal of Personality and social Psychology* 50(3):492.

Chen CC, Greene PG, Crick A (1998) Does entrepreneurial self-efficacy distinguish entrepreneurs from managers? *Journal of business venturing* 13(4):295–316.

Constant D, Sproull L, Kiesler S (1996) The kindness of strangers: The usefulness of electronic weak ties for technical advice. *Organization science* 7(2):119–135.

Ding Y, Yan E, Frazho A, Caverlee J (2009) Pagerank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Technology* 60(11):2229–2243.

Ericson J (2015) Stack Exchange Year in Review, 2015. *stackoverflow.com* Accessed: 2016-06-01.

Ghose A, Ipeirotis PG, Li B (2014) Examining the impact of ranking on consumer behavior and search engine revenue. *Management Science* 60(7):1632–1654.

Ghosh A, Hummel P (2014) A game-theoretic analysis of rank-order mechanisms for user-generated content. *Journal of Economic Theory* 154:349–374.

Goes PB, Guo C, Lin M (2016) Do incentive hierarchies induce user effort? evidence from an online knowledge exchange. *Information Systems Research* 27(3):497–516.

Grant S, Betts B (2013) Encouraging user behaviour with achievements: an empirical study. *Mining Software Repositories (MSR), 2013 10th IEEE Working Conference on*, 65–68 (IEEE).

Green RC, Talmor E (1986) Asset substitution and the agencycosts of debt financing. *Journal of Banking & Finance* 10(3):391–399.

Greenspan D (2015) Best Computer Jobs for the Future. `http://www.itcareerfinder.com/brain-food/blog/entry/best-computer-jobs-for-the-future.html`, accessed: 2016-12-05.

Guan Z, Cutrell E (2007) An eye tracking study of the effect of target rank on web search. *Proceedings of the SIGCHI conference on Human factors in computing systems*, 417–420 (ACM).

Hanrahan BV, Convertino G, Nelson L (2012) Modeling problem difficulty and expertise in stackoverflow. *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work Companion*, 91–94, CSCW '12 (New York, NY, USA: ACM), ISBN 978-1-4503-1051-2, URL `http://dx.doi.org/10.1145/2141512.2141550`.

Harper FM, Raban D, Rafaeli S, Konstan JA (2008) Predictors of answer quality in online q&a sites. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 865–874 (ACM).

Hausman JA (1978) Specification tests in econometrics. *Econometrica: Journal of the Econometric Society* 1251–1271.

Hausman JA, Hall BH, Griliches Z (1984) Econometric models for count data with an application to the patents-r&d relationship.

Heath C, Tversky A (1991) Preference and belief: Ambiguity and competence in choice under uncertainty. *Journal of risk and uncertainty* 4(1):5–28.

Heidemann J, Klier M, Probst F (2010) Identifying key users in online social networks: A pagerank based approach .

Henseler J, Ringle CM, Sinkovics RR (2009) The use of partial least squares path modeling in international marketing. *Advances in international marketing* 20(1):277–319.

Hmieleski KM, Baron RA (2008) When does entrepreneurial self-efficacy enhance versus reduce firm performance? *Strategic Entrepreneurship Journal* 2(1):57–72.

Isen AM, Nygren TE, Ashby FG (1988) Influence of positive affect on the subjective utility of gains and losses: it is just not worth the risk. *Journal of personality and Social Psychology* 55(5):710.

Isen AM, Patrick R (1983) The effect of positive feelings on risk taking: When the chips are down. *Organizational behavior and human performance* 31(2):194–202.

Jin XL, Zhou Z, Lee MK, Cheung CM (2013) Why users keep answering questions in online question answering communities: A theoretical and empirical investigation. *International Journal of Information Management* 33(1):93–104.

Joorabchi ME, Mesbah A, Kruchten P (2013) Real challenges in mobile app development. *2013 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, 15–24 (IEEE).

Jurczyk P, Agichtein E (2007) Discovering authorities in question answer communities by using link analysis. *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, 919–922 (ACM).

Kahneman D, Tversky A (1979) Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the econometric society* 263–291.

Kahneman D, Tversky A (1984) Choices, values, and frames. *American psychologist* 39(4):341.

Kankanhalli A, Tan BC, Wei KK (2005) Contributing knowledge to electronic knowledge repositories: an empirical investigation. *MIS quarterly* 113–143.

Krasner GE, Pope ST, et al. (1988) A description of the model-view-controller user interface paradigm in the smalltalk-80 system. *Journal of object oriented programming* 1(3):26–49.

Langville AN, Meyer CD (2005) A survey of eigenvector methods for web information retrieval. *SIAM review* 47(1):135–161.

Lee TW, Ko YK (2010) Effects of self-efficacy, affectivity and collective efficacy on nursing performance of hospital nurses. *Journal of advanced nursing* 66(4):839–848.

Lent RW, Brown SD, Hackett G (1994) Toward a unifying social cognitive theory of career and academic interest, choice, and performance. *Journal of vocational behavior* 45(1):79–122.

Li Z, Huang KW, Cavusoglu H (2012) Quantifying the impact of badges on user engagement in online q&a communities .

Lim SL, Bentley PJ, Kanakam N, Ishikawa F, Honiden S (2015) Investigating country differences in mobile app user behavior and challenges for software engineering. *IEEE Transactions on Software Engineering* 41(1):40–64.

Lin M, Lucas Jr HC, Shmueli G (2013) Research commentary-too big to fail: large samples and the p-value problem. *Information Systems Research* 24(4):906–917.

Liu Y, Bian J, Agichtein E (2008) Predicting information seeker satisfaction in community question answering. *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 483–490 (ACM).

Mabe PA, West SG (1982) Validity of self-evaluation of ability: A review and meta-analysis. *Journal of applied Psychology* 67(3):280.

Maltby E, Loten A (2013) App Building, the Do-It-Yourself Way. `http://www.wsj.com/articles/SB10001424127887324034804578344061497735762`, accessed: 2016-12-05.

Mamykina L, Manoim B, Mittal M, Hripcsak G, Hartmann B (2011) Design lessons from the fastest q&a site in the west. *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2857–2866 (ACM).

March JG (1991) Exploration and exploitation in organizational learning. *Organization science* 2(1):71–87.

Mihalcea R, Tarau P (2004) Textrank: Bringing order into texts (Association for Computational Linguistics).

Mobasher B, Cooley R, Srivastava J (2000) Automatic personalization based on web usage mining. *Communications of the ACM* 43(8):142–151.

Mombrea M (2016) Mobile development is tougher than people think. `http://www.itworld.com/article/2701225/mobile/mobile-development-is-tougher-than-people-think.html`, accessed: 2016-12-05.

Moores TT, Chang JCJ (2009) Self-efficacy, overconfidence, and the negative effect on subsequent performance: A field study. *Information & Management* 46(2):69–76.

Movshovitz-Attias D, Movshovitz-Attias Y, Steenkiste P, Faloutsos C (2013) Analysis of the reputation system and user contributions on a question answering website: Stackoverflow. *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on*, 886–893 (IEEE).

Nicholls JG (1984) Achievement motivation: Conceptions of ability, subjective experience, task choice, and performance. *Psychological review* 91(3):328.

Nygren TE, Isen AM, Taylor PJ, Dulin J (1996) The influence of positive affect on the decision rule in risk situations: Focus on outcome (and especially avoidance of loss) rather than probability. *Organizational behavior and human decision processes* 66(1):59–72.

Omojola A (2013) The Shortage Of Developer Talent Is Crushing Mobile. `http://www.forbes.com/sites/ayoomojola/2013/07/15/the-shortage-of-developer-talent-is-crushing-mobile/y`, accessed: 2016-12-05.

Page L, Brin S, Motwani R, Winograd T (1999) The pagerank citation ranking: bringing order to the web. .

Pal A, Chang S, Konstan JA (2012a) Evolution of experts in question answering communities. *ICWSM*.

Pal A, Farzan R, Konstan JA, Kraut RE (2011) Early detection of potential experts in question answering communities. *International Conference on User Modeling, Adaptation, and Personalization*, 231–242 (Springer).

Pal A, Harper FM, Konstan JA (2012b) Exploring question selection bias to identify experts and potential experts in community question answering. *ACM Transactions on Information Systems (TOIS)* 30(2):10.

Pan B, Hembrooke H, Joachims T, Lorigo L, Gay G, Granka L (2007) In google we trust: Usersdecisions on rank, position, and relevance. *Journal of Computer-Mediated Communication* 12(3):801–823.

Paul F (2013) Autodesk: Free Apps Are Harder To Make Than Enterprise Software. `http://readwrite.com/2013/03/26/autodesk-ceo-carl-bass-making-free-apps-is-harder-than-making-enterprise-software/`, accessed: 2016-12-05.

PayScale (2016) Mobile Applications Developer Salary. `http://www.payscale.com/research/US/Job=Mobile_Applications_Developer/Salary`, accessed: 2016-12-05.

Powers WT (1973) *Behavior: The control of perception* (Aldine Chicago).

Powers WT (1991) Commentary on bandura's" human agency." .

Raban D, Harper F (2008) Motivations for answering questions online. *New media and innovative technologies* 73.

Rabin M (2000) Risk aversion and expected-utility theory: A calibration theorem. *Econometrica* 68(5):1281–1292.

Ross L, Stittinger C (1991) Barriers to conflict resolution. *Negotiation journal* 7(4):389–404.

Schunk DH (1990) Goal setting and self-efficacy during self-regulated learning. *Educational psychologist* 25(1):71–86.

Sitkin SB, Weingart LR (1995) Determinants of risky decision-making behavior: A test of the mediating role of risk perceptions and propensity. *Academy of management Journal* 38(6):1573–1592.

Slanger E, Rudestam KE (1997) Motivation and disinhibition in high risk sports: Sensation seeking and self-efficacy. *Journal of Research in Personality* 31(3):355–374.

So Y (2008) The effects of achievement goal orientation and self-efficacy on course interests and academic achievement in medical students. *Korean Journal of Medical Education* 20(1):37–49.

Stajkovic AD, Luthans F (1998) Self-efficacy and work-related performance: A meta-analysis. *Psychological bulletin* 124(2):240.

Stewart Jr WH, Roth PL (2001) Risk propensity differences between entrepreneurs and managers: a meta-analytic review. *Journal of applied psychology* 86(1):145.

Stock JH, Watson MW (2008) Heteroskedasticity-robust standard errors for fixed effects panel data regression. *Econometrica* 76(1):155–174.

Stone DN (1994) Overconfidence in initial self-efficacy judgments: Effects on decision processes and performance. *Organizational Behavior and Human Decision Processes* 59(3):452–474.

Strecher VJ, Seijts GH, Kok GJ, Latham GP, Glasgow R, DeVellis B, Meertens RM, Bulger DW (1995) Goal setting as a strategy for health behavior change. *Health Education & Behavior* 22(2):190–200.

Theng YL, Sin SCJ (2012) Analysing the effects of individual characteristics and self-efficacy on users' preferences for system features in relevance judgment. *Information Research* 17(4).

Tian Y, Kochhar PS, Lim EP, Zhu F, Lo D (2013) Predicting best answerers for new questions: An approach leveraging topic modeling and collaborative voting. *Workshops at the International Conference on Social Informatics*, 55–68 (Springer).

Tversky A, Kahneman D (1985) The framing of decisions and the psychology of choice. *Environmental Impact Assessment, Technology Assessment, and Risk Analysis*, 107–129 (Springer).

Tversky A, Kahneman D (1991) Loss aversion in riskless choice: A reference-dependent model. *The quarterly journal of economics* 1039–1061.

Vancouver JB, Kendall LN (2006) When self-efficacy negatively relates to motivation and performance in a learning context. *Journal of Applied Psychology* 91(5):1146.

Vancouver JB, More KM, Yoder RJ (2008) Self-efficacy and resource allocation: support for a nonmonotonic, discontinuous model. *Journal of Applied Psychology* 93(1):35.

Vancouver JB, Thompson CM, Tischner EC, Putka DJ (2002) Two studies examining the negative effect of self-efficacy on performance. *Journal of Applied Psychology* 87(3):506.

Vancouver JB, Thompson CM, Williams AA (2001) The changing signs in the relationships among self-efficacy, personal goals, and performance. *Journal of Applied Psychology* 86(4):605.

Von Neumann J, Morgenstern O (2007) *Theory of games and economic behavior* (Princeton university press).

von Rechenberg T, Gutt D, Kundisch D (2016) Goals as reference points: empirical evidence from a virtual reward system. *Decision Analysis* 13(2):153–171.

Wallach HM (2006) Topic modeling: beyond bag-of-words. *Proceedings of the 23rd international conference on Machine learning*, 977–984 (ACM).

Waller B (2006) Math interest and choice intentions of non-traditional african-american college students. *Journal of Vocational Behavior* 68(3):538–547.

Walumbwa FO, Avolio BJ, Zhu W (2008) How transformational leadership weaves its influence on individual job performance: The role of identification and efficacy beliefs. *Personnel Psychology* 61(4):793–825.

Wooldridge JM (1999) Distribution-free estimation of some nonlinear panel data models. *Journal of Econometrics* 90(1):77–97.

Wooldridge JM (2015) *Introductory econometrics: A modern approach* (Nelson Education).

Wray S (2010) How pair programming really works. *IEEE software* 27(1):50.

Xu L, Nian T, Cabral L (2014) What makes geeks tick? a study of stack overflow careers. Technical report, Working paper.

Yu J, Jiang Z, Chan HC (2007) Knowledge contribution in problem solving virtual communities: the mediating role of individual motivations. *Proceedings of the 2007 ACM SIGMIS CPR conference on Computer personnel research: The global information technology workforce*, 144–152 (ACM).

Zhang J, Ackerman MS, Adamic L (2007) Expertise networks in online communities: structure and algorithms. *Proceedings of the 16th international conference on World Wide Web*, 221–230 (ACM).

Zhao H, Seibert SE, Hills GE (2005) The mediating role of self-efficacy in the development of entrepreneurial intentions. *Journal of applied psychology* 90(6):1265.

Zhou Y, Cong G, Cui B, Jensen CS, Yao J (2009) Routing questions to the right users in online communities. *Data Engineering, 2009. ICDE'09. IEEE 25th International Conference on*, 700–711 (IEEE).

Zikmund WG, Babin BJ, Carr JC, Griffin M (2013) *Business research methods* (Cengage Learning).