Analyses of User Behavior and Social Incentives in Q&A Communities

ANDREW WILLIAM VARGO

Abstract

The goal of this thesis is to provide insight into the relationship between Question and Answering (Q&A) community reputation classes and socially desirable behavior within these communities. Many Q&A sites use peer evaluation systems in which contributors of information can gain or lose reputation points. These reputation points are important for two reasons: 1) They incentivize users to contribute to the community, and 2) They can be used as a mechanism to bestow moderating privileges on users who earn certain levels of points. In addition, previous research has shown that users with higher reputation points give better questions and answers. What is not understood is the relationship between higher reputation and socially desirable behavior outside of the acts of questioning and answering.

Q&A has become an important venue for the peer-production of valuable technical information. In particular, Stack Overflow (SO), the research site of this thesis, has become an invaluable resource for those interested in computer programming and has spawned many clones in other domains. SO has a community-driven reputation system that rewards users for good contributions to Q&A sessions, and relies on users with earned privileges in higher reputation classes to do a wide-array of socially desirable activities such as editing bad questions and correcting wayward users. The key concern is whether reputation class is predictive of engaging in pro-social behaviors.

Q&A communities are not monolithic, and we consider three models in our analyses: 1) An Expert-Based Q&A Model where the community relies on a group of accepted experts from the field. It is important in this model that questioners trust the abilities of the experts, 2) A Collaborative Model where users are encouraged to edit question-answer pairs so that the answer reflects the best information at the time, and 3) A Community-Based Model where the flow of

Q&A is dependent on user interaction and discussion. Each one of these models requires different types of socially desirable behaviors.

With respect to the above points, this thesis provides analysis on the following issues in Q&A:

1. *Determining identity and its relationship with performance.* In the Expert-Based Q&A Model, it seems natural that users will trust experts if they know the identity of the expert. Identities and the different roles identities play in social Q&A communities has long been discussed from both a theoretical and empirical standpoint. Identity is usually analyzed in ways that emphasizes a transaction, and many studies use third-party raters to assign value judgements to these factors, which may not be relevant to a community. In this chapter, we examine profiles in SO. These profiles are highly customizable, allowing users to choose the level of personal information they want to share: from extensive to none. We develop a categorization scheme using grounded theory to develop definitions of identity, and analyze behavior based on these definitions. We find that the choice of identity is diverse within the community and that there is a correlation with identifiers and increased reputation earning among the general population. An analysis of elite users, however, indicates that identity is closely tied to membership length, but not to performance.
2. *Understanding the relationship between reputation incentives and site maintenance.* Collaborative editing is an effective tool for improving contributions in peer production communities like Wikipedia and question-answer (Q&A) communities. However, the mechanisms behind who edits and why is not well understood. Previous studies have focused on the effectiveness of editing and emergent hierarchies in editing communities.

What is unknown is how editing is executed in a system that contains gamified motivations for contributing edits. In this chapter, we examine participants editing unfit questions on SO. The combination of SO's community and reputation system with the dynamics of unfit questions allows us to examine how different actors behave. We find that early edits come from high-reputation users who do not participate as a questioner or answerer, indicating that these users work to retain certain questions. We also find that high-reputation users actually decrease the time between their edits as time passes. The results show that the Collaborative Q&A Model is dependent on high-reputation users.

3. *Understanding interactions between users and reputation class.* In the Community-Based Model, it is important for users to be able to communicate with each other in socially desirable ways. What kind of comments do Q&A community members leave on bad comments, and is there a difference between the comments left and reputation class? We studied this question on SO which has strong reputation and privilege systems and clear guidelines on commenting. Peer-production systems often employ feedback dialogue to engage with producers of low quality content. However, dialogue is only beneficial if it is constructive, as previous work has shown the adverse effects of negative feedback on quality and production. Previous studies indicate that feedback is likely critical, but the extent, orientation, and actors within this assumption are unknown. In this paper, we contribute a basic taxonomy of commenting and perform analysis on user types and community preferences. Results indicate that the most popular and frequent comments are undesirable since they include unconstructive criticism, and that higher-reputation users do not leave more socially desirable feedback.

The results of this thesis show that reputation class is limited in predicting socially desirable behavior. While it is effective at indicating users who do the most valuable work in the Collaborative Q&A Model, it cannot be relied upon as a good indicator in other situations.

Acknowledgments

I have been very fortunate. Many wonderful people have taken time and have expended great effort to help me with this thesis. Without this support, I would not have succeeded. I am thankful, and I am hopeful that I will be able to support others in the same way.

First, I would like to thank my supervisor, Professor Shigeo Matsubara. His support and patience are extraordinary. I am blessed to have had a supervisor who let me explore different methodologies for addressing the topics in this thesis and one who gave me such valuable feedback. It is because of him that I have learned how to become a researcher.

I am very grateful to my committee members at Kyoto University, Professor Toyoaki Nishida and Professor Masatoshi Yoshikawa. Their constructive feedback and advice were essential to this thesis.

I would like to thank Professor Toru Ishida for all his help and support. I learned a lot from the Laboratory Seminars, and I always appreciated his questions. He developed a friendly and diverse environment that was an ideal place to learn how to research.

I am thankful to all the faculty of Ishida and Matsubara Laboratory, both past and present, who always provided great discussion. I would especially like to thank Donghui Lin, Hiromitsu Hattori, Rieko Inaba, Masayuki Otani, Yohei Murakami, Yuu Nakajima, and David Kinny for all of their helpful comments during my study.

I would like to express my appreciation to Terumi Kosugi, Yoko Kubota, Hiroko Yamaguchi who helped guide me through the academic procedures at Kyoto University.

The members of Ishida and Matsubara Laboratory were incredible. I learned a lot from them and always looked forward to their presentations at lab seminar. I would especially like to thank Huan Jiang and Ari Hautasaari for always listening to me and providing great discussions

about my research. I would also like to thank Linsi Xia, Xun Cao, and Shinsuke Goto for the constant encouragement during my study.

I am also grateful for my colleagues at The Kyoto College for Graduate Studies for Informatics for the help and support they provided me. I would especially like to thank Kotomi Abe for taking time out of her busy schedule to help me write emails in Japanese.

I am thankful to Robert T. Singer, Head, Pavilion for Japanese Art at the Los Angeles County Museum of Art, for all of his encouragement and support. I would like to thank Benjamin Tag (Keio University) and Taylor Martin (Keio University) for their collaboration, help, and friendship. I am grateful to Tadg and Mika McLoughlin for their constant support through the years. I could never express enough appreciation for my friends Chris Blakely, Rhys Brown, and Andrew Board for always being there to help me and listen to me talk endlessly about my work.

Finally, I would like to thank my parents, Dr. Larry and Carolyn Vargo, my brother David Vargo, and my sister and her husband, Mary Beth and Gregory Weaver. Having a family with an expert in computer programming, an expert in library science, and great proofreading skills was a tremendous help. They also provided me with unending encouragement and support, and always reminded me to cast all of my anxiety on God.

Table of Contents

List of Figures

List of Tables

Chapter One

Introduction

1.1 Objectives

Question-Answering (Q&A) websites have been repositories of information seeking and collecting since the popularization of the World-Wide Web (Adamic et al. 2008). They represent an effective form of a peer-production (or crowdsourced) information system, where users collectively create valuable repositories of information for both participants in their respective communities and for users on the web (LaToza and Hoek 2016). In the early days of Q&A many different models to incentivize contributions were explored and commercialized, from pay-for-answers sites like Google Answers (closed in 2006) and Mahalo, to generalist free sites like Yahoo! Answers where users receive reputation points instead of money. Each version has unique incentive-design issues (Adamic et al. 2008a; Chen, Ho, and Kim 2010). Over time, the free model, where contributors are rewarded with reputation points, came to dominate and they are now ubiquitous (Hsieh, Kraut, and Hudson 2010; Mamykina et al. 2011).

With this came the advent of domain-based Q&A (E. Choi, Kitzie, and Shah 2010), where one topic or area of study is discussed exclusively. In contrast to generalist sites, these systems can center around one group of users with their own "community of practice" (Mamykina et al. 2011). This facilitates two important features: 1) the Q&A site can become an authoritative corpus for the domain if there are enough questions to describe the domain, and there are enough corresponding quality answers to these questions, and 2) the members of the community can be entrusted to skillfully curate the information within the Q&A system through voting and editing.

A challenge for system designers of domain-specific sites is choosing the correct type of incentive to reach critical mass and sustain user participation. Reputation points that are rewarded through community voting have been shown to effectively incentivize quick and accurate answers (Anderson et al. 2012). Indeed, these reputation points are so effective that they can influence user behavior towards answering questions, even if the users deny the impact of reputation points on their own behavior (Tausczik and Pennebaker 2012).

While these reputation points are a powerful tool for Q&A, including domain-specific Q&A, it is still unclear how much they impact, or predict, socially desirable behavior. Many communities, both Q&A and other types of peer-production communities, have numerous important and necessary tasks, such as collaborative editing, social communication, and even question-asking.

The objective of this thesis is to study the behavior of users within the context and mechanics of a reputation system based on reputation points. In particular, this thesis uses the construct of the "Reputation Class," that is, the way users are divided based on their aggregated reputation points. This study includes analyses of behavior and self-representation, maintenance behavior in the site, and interactions with other users on the site. The results of this thesis are important for the development of Q&A communities and other peer-production communities, especially those reliant on expert users, and gives significant design implications to their incentive and community-management systems. The understanding of reputation class is important for maximizing the efficacy of a popular reputation management platform for the peer-production of information.

1.2 Approach and Issues

This thesis combines qualitative and quantitative analysis of real-world user behavior to investigate the relationship between user behavior and social incentives. The overall goal is to help designers of such systems understand the implications of reputation points beyond indicating answering prowess.

We focus on data and interactions extracted from Stack Overflow (SO), a large domain-specific Q&A community dedicated solely to computer programming with over 8 million users and 16 million questions[1]. SO is a leading Q&A site and is ranked within the Top 50 most visited websites in the United States, and Top 75 in the World ("Stackoverflow.com Traffic, Demographics and Competitors - Alexa" 2018), and has been cloned for many different domains of information across the web ("Stack Exchange Clones" 2018; Furtado et al. 2013; Tausczik, Kittur, and Kraut 2014). In this way, the lessons learned from SO are generalizable to many other Q&A communities.

SO provides an ample number of users and interactions to study, and also provides an incentivized reputation system that gives users control over the community as they gain reputation points. This means that users have concrete categorizations given to them by the system itself. In each analysis, we seek to understand how users of different reputation classes approach particular tasks and representation.

For this research we chose SO over a system like Quora, for several reasons: SO has a maintained its reputation system since its inception (Mamykina et al. 2011), it has a strong emphasis on maintain its archive (Correa and Sureka 2014; G. Li et al. 2015), and it has the site information presented in a manner which is able to be disseminated. Quora, on the other hand,

[1] http://www.stackoverflow.com/

has gone through multiple iterations of its reputation system and how it presents its information ("Why Did Quora Get Rid of Question Details? Isn't It Rather Crucial to Understand over Half of the Existing Content on the Site? Why Not Just Prevent People from Adding Details to New Questions While Retaining the Details Attached to the Old Ones? - Quora" 2017), and does not provide the same type of access to its archive as SO does (Mill 2014). This means that while SO's archive can be studied as if viewing a continuous entity, a site like Quora needs to be dissected based on its iterations.

Throughout this thesis, the goal is to describe the relationship between user behavior and reputation class in a community-run Q&A system by going beyond answering prowess and instead analyzing social behavior. In order to build a realistic understanding of the community of practices within the system, we adopted Grounded Theory to facilitate our analyses. The following three topics are discussed in this thesis:

1. **Determining identity and its relationship with performance.** A useful function for Q&A systems would be the ability to predict which users are going to be the most productive and effective within the community. This is an issue for many types of peer-production communities, such as product review sites like Yelp! or Amazon, where expertise and the disclosure of real world qualifications may signal trustworthiness. One way to analyze identity is by extracting profile features from user accounts and seeing if certain types of user identity lines up with certain types of performance. Of course, one approach would be to use a data-mining technique to see where different fields in profiles are filled in. The limitation with this method is that instead of determining what users have put inside their profile fields, it merely indicates that they have filled their profile fields with something. This means that concluding that certain types of profiles lead to certain behaviors is often an extrapolation which may not be accurate in reality.

For example, a profile picture of American President John F. Kennedy would be a valid profile picture that would be included in data sets seeking to identify users using their own profile photos on their accounts. In order to obtain detailed information of account information, we use a Grounded Theory ("Why Did Quora Get Rid of Question Details? Isn't It Rather Crucial to Understand over Half of the Existing Content on the Site? Why Not Just Prevent People from Adding Details to New Questions While Retaining the Details Attached to the Old Ones? - Quora" 2017) technique to create classifications of users. By doing this, we were able to run quantitative analysis and compare behavior based on users' statuses in the reputation class system.

2. **Understanding the relationship between reputation incentives and site maintenance.** Since reputation points are known to be a strong incentive for answering questions, it is important to understand how effective these points are in incentivizing other desired behaviors, like editing. Sites like SO and Wikipedia, which require its users base to maintain the quality of the archive through editing need to know how to create and keep a broad-base of contributing users. SO has a tiered reputation class system that provides reputation incentives for editing questions for different types of users based on their aggregated reputation. From this scenario, we categorized different types of actors on these questions to derive the incentive structures for each of them. We were then able to quantitatively compare different classes of users and gauge the effect of incentives.

3. **Understanding interactions between users and reputation class.** We consider whether reputation points not only indicate quality contributors of question and answers, but whether these points also indicate users who are more likely to contribute to socially desirable interactions. Many peer-production communities, like Wikipedia and GitHub, need to know

whether production performance can be tied to pro-social behavior. The expectation is that users with more reputation points would be more conscientious about site rules and conduct and thus contribute more socially desirable comments. On the other hand, if social incentives are strong, we might also expect users to contribute the most popular types of comments. Using the constructs of the SO reputation class system we analyze feedback left by using Grounded Theory. We then are able to apply quantitative tests to evaluate the feedback by user class.

The results of the topics show that reputation points and classes are limited in their ability to predict socially-desirable behavior. The designers of Q&A sites and peer-production systems which use reputation points to bestow power and status upon users should be aware of the limitations of reputation points in predicting socially desirable behavior.

1.3 Thesis Outline

This thesis consists of six chapters, including this introduction chapter. Chapter 2 provides a review of the state of reputation as an incentive in Q&A communities. First, we explain the nuance between different types of Q&A and information seeking in order to define where the chapters in this study have the most theoretical and practical relevance. We identify three aspects of Q&A that are applicable to the studies in this research, namely; Expert-Based, Collaborative, and Community-Based. Then we examine different types of incentive structures, in particular difference between paid incentives and virtual reputation points and how virtual reputation became to be recognized as a viable method for encouraging good contributions. We then present the different scenarios under which reputation points are proven to incentivize behavior. Finally, we discuss the assumptions about reputation and socially desirable behavior based on previous studies. This helps to clearly define the contribution of this work within the body of previous studies.

Chapter 2 also introduces the mechanics of SO and its reputation system. We briefly explain the history, context, and goals of the Q&A system. In particular, we document how the site grew rapidly and developed its own culture and reputation. We then introduce the process and reward system behind questioning and answering and its relationship with the privilege system. The goal of this section is to provide an accurate view of where SO stands within the known research of reputation incentives in Q&A.

In Chapter 3 we introduce a Grounded Theory study on identity and reputation earning on SO. The study was conducted to understand if users who give more information about themselves have better contributions. Previous studies (Adaji and Vassileva 2016; Ginsca and Popescu 2013) indicated that we should expect to find better per-contribution scores for users who give more information about their real-world identity. In addition, it has been hypothesized that giving more personal information leads to better contributions and social behavior on the Web (Friedman* and Resnick 2001). By using reflexive iterative Grounded Theory (Srivastava and Hopwood 2009), we developed a categorization scheme that allows for performance comparisons between different types of users. We constructed two samples from the entire data set. The *General* sample which took a random sample of users from the entire population, and the *Elite* sample that randomly sampled from the top 1% of reputation holders. Through the sampled profiles, we used Grounded Theory to determine the important factors for constructing a profile (profile picture, name, and weblink) and built classifications of users based on frequently observed patterns. In particular, three categorizations of users emerged: Full ID, Link ID, and Pseudonym. Full ID members provide a profile picture and web-link that confirms their real-world identity with their SO account; Link ID members provide a web-link which does the same, and Pseudonym users do not provide any information which links their SO account to their real-

world identities. Quantitative tests initially indicate that Full ID and Link ID users are better users in the system. However, normalizing the data indicates that these users are more likely to have older accounts and do not score significantly differently on performance tests. We conclude that identity may be useful for identifying original members of the community but does *not* correlate with better per-contribution performance.

In Chapter 4 we start using SO's user ranking system to categorize behavior around the editing of questions. Previous studies indicated that reputation points provided the necessary incentive to encourage good answers (Anderson et al. 2012; Tausczik and Pennebaker 2012; Tausczik, Kittur, and Kraut 2014). In this chapter, we seek to understand whether this incentive structure extends to other socially desirable activities. We look at the process of editing bad questions that are submitted to SO. The process of editing and improving these questions can be seen as a socially desirable as they help to improve the corpus for the community and world wide web. The incentive system of SO enables us to be able to construct definitions of user groups and their incentives: the question-asker has the incentive to improve the score of their question; low-reputation users can gain reputation through editing these questions; answerers who are incentivized to keep questions from being deleted; and high-reputation users which gain no reputation points via editing. The analysis shows that high-reputation users are the most likely to not only the first edit, but also contribute the most edits overall. We conclude that the reputation incentive itself is not the driving behavior behind most of the edits.

Next, in Chapter 5 we look at the process of user interaction and feedback. In chapter 4, we found that having a high-reputation account was linked to socially desirable behavior. However, these actions are limited to actions on material, rather than interactions with other users. In SO, bad questions are a problem since they make user experience of the site poorer

(Correa and Sureka 2014). As such, users were encouraged to leave helpful feedback for questioners in the form of comments that appear under the question. As is common in many social systems, comments left as feedback are more often critical rather that constructive. In addition, critical comments may invoke a positive response from other users. In order to understand the types of comments different users leave, we used Grounded Theory to develop comment types. We found that comments were either Corrective (desirable from the community rules perspective), Critical (undesirable), or Answer (undesirable). We found that Critical comments were the most common type left on bad questions and the most socially rewarded through a voting system. In addition, there was no differentiation between comments left and the reputation of the commenter. We conclude that reputation cannot be used to predict more desirable social interactions.

Finally, we conclude this thesis with Chapter 6, which summarizes the results obtained in this work and addresses possible future work.

Chapter Two

Background

Question & Answering (Q&A) sites may seem like a simple proposition; there is a question that is asked and answers that are provided. However, there are many different types of systems (E. Choi, Kitzie, and Shah 2010) and therefore many different types of scenarios in which to evaluate and understand incentives to contribute. In this chapter, we explain what makes Q&A different from peer-production systems, and then explain the different variants of Q&A systems that exist and where the target system of this study, Stack Overflow (SO), is placed within these systems. We then present an overview of literature on incentives in Q&A, especially studies that relate to the target system. Finally, we explain the development and functions of the target system. This includes background on why the site exists, why it has been widely adapted, how the Q&A exchanges work, and how its reputation and incentive systems work.

2.1 An Overview of Q&A Systems

At the time of this thesis, Q&A systems have been around for over 15 years. The first official site was AnswerBag, a generalist Q&A site which was started in 2003 (Shah, Oh, and Oh 2008). The number of Q&A sites rapidly expanded, and all of them have similar problems: how to retain users and promote content contributions that meet the goal of the system (R. Gazan 2007). In order to do this, system designers implemented a number of techniques to incentivize users; from allowing users to pay answerers, to rewarding reputation points with no redeemable aspect, and to using privilege models in which users compete for points for prestige and power.

Q&A sites are a form of a peer-production community, like Wikipedia (Warncke-Wang et al. 2015). The similarity between Q&A and other forms of peer-production is that the end goal of each system is to produce valuable information. This requires the system to find ways to

harness the community's collective intelligence. This means that peer-production systems of different types tend to have similar results when they use similar techniques, such as editing (G. Li et al. 2015). On the other hand, Q&A seems to require more expertise from individual users than peer-production sites like Wikipedia (Joo and Normatov 2013). This is likely because the process of Q&A requires an individual who is motivated to ask a question, which is often time-sensitive and achieves its greatest utility when information is provided sooner rather than later (Jain, Chen, and Parkes 2009).

The last sentence is particularly important when considering what makes Q&A different from other peer-production communities. Communities like online encyclopedias, review sites, and archival sites receive contributions from their users. These contributions are then disseminated and consumed by the community. For instance, on GitHub, on online repository of computer programming code, users will contribute code, the code is reviewed and edited by the community, and then the code is consumed by users (Marlow, Dabbish, and Herbsleb 2013). The knowledge-exchange is conceived when the knowledge contributor makes the submission. In these types of peer-production systems, experts can create the content on their own since the knowledge-exchange can be unidirectional.

Q&A, on the other hand, requires a set of users, the questioner and the answerer (Jain, Chen, and Parkes 2009), and is thus required to be bidirectional. It is the questioner that makes the initial step in the knowledge-exchange. The answerers are not creating submissions by themselves but are instead reacting to a direct request for help. Because of this, experts cannot create Q&A systems by themselves, as they do not necessarily know what beginner users or other non-domain experts need to know. For instance, expert users of SO failed to correctly identify the questions that would be asked on the other language versions of its system (Vargo et

al. 2018). Instead, the questioners drive what information is created. In this way, Q&A is effective at both facilitating direct knowledge-exchanges between questioner and answerer and in creating an archive that responds to market forces where the knowledge-exchanges are given value by the community via voting. (Anderson et al. 2012). This does not mean that Q&A is necessarily just a transaction from beginner to expert, as there is collaboration between questioners and askers (Rich Gazan 2010). It does show that reputation indicates not only expertise about the topic, but the ability of the user to collaborate with others, share concerns, and engage in the learning process (Barua, Thomas, and Hassan 2014).

There are numerous places where individuals can ask questions. For instance, question-askers can query their social network sites to receive information, go to a generalist Q&A site like Yahoo! Answers, where any question is welcome, or go to a domain-specific Q&A site like Stack Overflow (SO). In the early years of Q&A, generalist sites like Yahoo!Answers and AnswerBag dominated the market (Shah, Oh, and Oh 2008). However, these types of communities were not necessarily optimized for all Q&A types of activity. Choi (E. Choi, Kitzie, and Shah 2010) developed a taxonomy of the types of information-seeking that questions exhibit:

- Information-Seeking – Looking for a direct answer to a problem or question. Example: “How do I make a data frame in R?”
- Advice-Seeking – Asking how to do something or for methods on how to solve a problem or question. “I want to learn Data Science. What is a good software setup for a beginner?”
- Opinion-Seeking – Seeking feedback on a course of action or gauging popularity of a subject topic. “Which is better: Python or R?”

- Non-Information-Seeking – Rhetorical questions not looking for substantial additional information. “Why don’t people understand statistics?”

The differences between the questions are highly contextual and depend on how and where they are being asked (Harper, Moy, and Konstan 2009). Information-Seeking is the easiest to understand in that there is either a clear best answer or an answer approximating being best for the questioner. With Non-Information-Seeking, many rhetorical questions could be seen as being honest inquires if asked in the correct context. For instance, the example question, “Why don’t people understand statistics?” would be a valid Information-Seeking question that may be able to be asked in a Q&A community devoted to education. However, if the question is asked in a context where no one is expected to contribute meaningful insights, then the question is Non-Information-Seeking.

There is also a nuance between Advice-Seeking and Opinion-Seeking (B. Choi et al. 2010; Harper, Moy, and Konstan 2009). Advice-Seeking requires the user to have an objective that can ostensibly be achieved via answers. In the example question, the user is clearly looking for a setup for learning Data Science. In Opinion-Seeking, on the other hand, the user is asking an open-ended question.

Each type of question-asking does not necessarily belong in the same type of community. For instance, Non-Information-Seeking or Opinion-Seeking regarding personal information may be best asked on social networks (Morris, Teevan, and Panovich 2010; Jiang Yang et al. 2011). Choi (E. Choi, Kitzie, and Shah 2010) also developed a classification system of where these types of questions could be asked. They then mapped which models were the best fit for different types of questions based on the distribution of the questions in real systems:

- A Community-Based Model – These types of systems are knowledge exchanges where users can ask and answer questions to a community which is based around the exchange. The community then builds up pairs of questions and answers for the archive. These models also often include reputation systems and encourage users to vote on questions and answers and is also dependent on user interaction and discussion. In Choi's topology, this model is the best fit for Advice-Seeking (40.8% of all questions) and Opinion-Seeking (50% of all questions).
- A Collaborative Q&A Model – These types of systems encourage members to edit question-answer pairs so that the answer reflects the best information at the time. This means that Q&A sessions are in potentially constant flux and are dependent on collaborators to make them a better fit for the system. In Choi's topology, this model is the best fit for Information-Seeking (50.6% of all questions) and Advice-Seeking (38.4% of all questions).
- An Expert-Based Q&A Model – These types of systems restrict answers from the open community, and instead rely on a group of accepted experts from the field. Often experts are paid for their services. It is important in this model that questioners explicitly trust the experts to understand the domain well. In Choi's topology, this model is the best fit for Information-Seeking (87.2% of all questions).
- A Social Q&A Model – This model is where a user relies on their own personal network to answer their questions. For instance, a user may ask their Facebook network of friends a question. Unlike the other models, this is more informal and more reliant on personal connections. In Choi's topology, this model is the best fit for

Non-Information-Seeking (43.6% of all questions) and Advice-Seeking (34% of all questions).

As with many theoretical topologies, the real-world application of the taxonomy leaves many systems straddling multiple definitions. This is largely due to the concept of "community of practice" (Schwen and Hara 2003), where developers and community members iteratively choose different facets that result in a desirable outcome for users. This is especially true for a site like SO, where designers of the system did not create the platform alone, but were part of the computer programming community and received input from other members of the community on a daily basis (Mamykina et al. 2011; Ford 2012).

The types of information seeking on SO does vary, but most successful Q&A pairs revolve around questions that have a concrete answer of solution (Treude, Barzilay, and Storey 2011). That is, questioners are participating in "Information-Seeking" and are trying to get an answer for an actual coding problem that they have. The type of questioning that is asked is similar across all topics, technologies, and languages (Allamanis and Sutton 2013). In Section 2.4 we go into further detail surrounding the system rules that dictate and reinforce these patterns of question-asking.

The SO model does not follow a Collaborative or Expert-Based model even though most of the questions are Information-Seeking in a field of technical information. Instead, it uses a hybrid Community-Based model which effectively incorporates elements of Expert-Based and Collaborative models. SO has a full Community-Based model in that it has an active knowledge-exchange which includes a reputation system and voting system.

At the same time, SO also takes inspiration from Collaborative models. One of the community's goals is to build an accurate archive of programming information (Mamykina et al.

2011). The quality of Q&A pairs is of concern not only at the inception point of the exchange, but also as technology advances and changes the applicability of the information. Therefore, users are actively encouraged to edit and improve all parts of exchanges (G. Li et al. 2015).

SO also borrows from the Expert-Based model. Questioners do not need to pay for experts to answer their questions, but the system does allow for self-promotion and identification. While users are not required to give information about themselves, many of them choose to do so, often identifying themselves as well-known experts in their respective fields. In addition, the longevity of SO has allowed for the meaningful identification of experts. Finally, users are also encouraged to use their portfolio of Q&A activity on SO in Stack Exchange's proprietary job-hunting website, Careers 2.0 (Xu, Tingting, and Cabral 2014).

The combination of these models makes SO a dynamic venue to study. Applying these models allows us to better frame the social incentives in their proper contexts. In the next section, we briefly discuss the theoretical motivations users have to participate in Q&A and its relationship to this thesis.

2.2 Incentives to Participate in Q&A

A Q&A system constitutes a type of explicit crowdsourcing where the functionality of the site is based on the users' willingness and ability to share information with others (Morris, Teevan, and Panovich 2010; Jie Yang et al. 2014). There is an obvious two-part relationship here; the information-seeker and the information provider (Shah, Oh, and Oh 2008). The willingness to participate in the first part of this relationship is easy to delineate. Users will contribute a question if:

- The users are seeking information (Jain, Chen, and Parkes 2009; Mamykina et al. 2011).

- The users sufficiently trust the information provided in the community (Morris, Teevan, and Panovich 2010).

Thus, it can be assumed that as long as the community has reached a sufficient critical mass of information providers, the system will not have a shortage of questions (Adamic et al. 2008a; Rich Gazan 2011a; Mamykina et al. 2011; Oliver, Marwell, and Teixeira 1985). Therefore, it is unsurprising that most research focuses on understanding how and why users are incentivized to answer questions, especially when the field is technical and requires expertise.

The pay-for-answers Expert-Based model was tried extensively in the mid-2000s and Google Answers (closed in 2006) was a particularly well-studied example (Harper et al. 2008; Chen, Ho, and Kim 2010). In this model, experts are invited to contribute answers in competition for fees offered by questioners. Questioners were willing to pay more for answers to more difficult questions, but answers did not improve (Chen, Ho, and Kim 2010). Instead of getting *better* answers for more money, questioners were getting *longer* answers, which was clearly a non-optimal result (Chen, Ho, and Kim 2010). While money does not seem to incentivize better answers, gratis Q&A has been successful in garnering experts and answerers to contribute to their communities (Adamic et al. 2008).

The obvious question is why do answerers willingly contribute to forums without receiving or expecting payment? Recognition from peers is a very strong motivating tool for contributors. Answerers look for questions where they can have an impact on the community (Anderson et al. 2012; Dearman and Truong 2010; Pal and Konstan 2010). Nam (Nam, Ackerman, and Adamic 2009) found in a survey that users were motivated by the ability to contribute answers that had a positive effect for questioners, and Oh (Oh 2011) found that users

contributed answers because of altruistic motivations. In summary, answerers contribute when they feel they can help the questioner and when they can help the larger community.

At the same time, answerers are more likely to target questions that return more reputation via community voting. Anderson (Anderson et al. 2012) finds users in higher reputation classes focus on questions that have more reputation value and may have a special ability to predict which questions are going to be worth more. What this means is that certain questions are more likely to receive answers. For instance, Treude (Treude, Barzilay, and Storey 2011) found that answerers target questions that have concrete solutions and which can be more easily identified as being correct or not. In addition, Tausczik (Tausczik and Pennebaker 2012) found that on a mathematics Q&A community that users who asserted that reputation points were not important exhibited the same behavior in prioritizing high-value questions.

Wei (Wei, Chen, and Zhu 2015) found that users were more likely to be motivated by their relative reputation. That is, it is not just the aggregation of points which is important, but that having more points than other users are what motivates contributions. These studies seem to indicate a dissonance between what contributors say and what actually motivates them. Competition is clearly a driver of action over stated altruism.

However, reputation points are a community's way of explicitly rewarding the best contributions from users. That is, users target certain questions because they are better than other questions. This means that the difference between the stated motivation and what is observed is not incongruous. The answerers would not contribute if their actions did not help the intended audience. The desire to earn more reputation and recognition is often compatible with helping the audience.

In this study, we assume that users are motivated by both being rewarded by the community and helping the community. We also consider that these motivations may not always be compatible and may in some cases compete with each other. For instance, in Chapter 5, we consider how socially desirable behavior (from the perspective of the system and questioner) competes with social approval. In the next section we consider the meaning of reputation points in the broader research spectrum of identifying authoritative users or predicting behavior.

2.3 The Meaning of Reputation Points in Q&A

In 2.2, we show that previous research finds that reputation points are effective at incentivizing answers. In this section we look at previous research and discuss the other uses for reputation points. We do know that in system design, reputation points are used as a measure for public trust and authority ("What Is Reputation? How Do I Earn (and Lose) It? - Help Center - Stack Overflow" 2015), but how effective is it at differentiating between users?

An important assumption is that of Power Law distribution. We assume that most "valuable" content is produced by a relative few number of users (Adamic et al. 2008; Mamykina et al. 2011). In the case of Q&A, this is expected to be seen in the answer part of Q&A rather than the question part (Mamykina et al. 2011), although this is not necessarily true depending on the difficulty of the domain (Tausczik and Pennebaker 2011). Regardless, when considering reputation classes, the majority of research assumes that most users have a few contributions, while a few have many.

We can consider that there are two types of reputation a person could have; offline and online. In a technical Q&A community, users may require expertise within a field to be able to ask and answer questions. Tausczik (Tausczik and Pennebaker 2011) looked at a mathematics Q&A site and found that both types of reputation were effective at predicting Q&A quality. In

answering, the online community reputation was slightly more effective. This is confirmed by studies that find that high-reputation owners tend to produce high-quality answers (Patil and Lee 2016).

This does not mean, however, that reputation is a perfect one-to-one predictor of expertise or quality. Users can target certain times of day or use strategies to target less popular questions to aggregate reputation points (Anderson et al. 2012; Bosu et al. 2013). In general, we can expect that higher-reputation users are likely to have a wider range of Q&A than other users (Lappas, Dellarocas, and Derakhshani 2017). However, the users with the most expertise may not have the highest reputation, and may focus on a narrower range of topics (Lappas, Dellarocas, and Derakhshani 2017; Jiang Yang et al. 2011).

Previous research also indicates that higher-reputation users will typically be the ones to participate in other activities related to the Q&A community. For instance, most tags on SO are done by higher-reputation users (MacLeod 2014). This is unsurprising considering the Power Law distribution, where we would expect to see an overlap between all types of participation.

Higher reputation can be used to indicate a number of other attributes. For one, users with higher-reputation tend to user fewer negative emotions in their Q&A submissions and tend to be more extroverted (Bazelli, Hindle, and Stroulia 2013). In addition some research asserts that higher-reputation users are more likely to share personal information within the community (Adaji and Vassileva 2016; Ginsca and Popescu 2013). This would indicate that higher reputation users may have different personalities compared to the rest of the community.

Based on this previous research, we assume that reputation is a good indicator of perceived user quality when answering questions within the system. However, the understanding of what user reputation means regarding other socially desirable behavior is not well understood

and is thusly a main theme of this research. In the next section we provide an overview of the SO ecosystem.

2.4 Stack Overflow

In this section, we give a concise but thorough explanation of the basics of how SO works. We first give an overview of the system, followed with an explanation of the reputation system. We then describe the privileges that users can earn in the site and how they can be used. Finally, we describe the community's archival goals.

2.4.1 Overview

SO was started in 2008 by a community of computer programmers interested in creating an active knowledge-exchange of programming questions that would also help create a reliable archive of programming information (Mamykina et al. 2011). As of 2018, ten years after its inception, Stack Overflow (SO) has over 8.3 million registered users that have asked 15 million questions and have contributed 24 million answers. It receives over 6 million visits a day and averages about 6,000 questions per day. In total, 71% of all questions have received an answer and 55% have an answer that has been accepted by the questioner as solving the problem. The success of SO has spawned a related network of other domain-specific Q&A sites called Stack Exchange (Tausczik and Pennebaker 2012), and the site mechanics have been exported to dozens of other independent sites ("Stack Exchange Clones" 2018)

The SO community is managed by a user-driven reputation system (De Alfaro et al. 2011; Mamykina et al. 2011; Anderson et al. 2012) in which members vote on the quality of questions and answers. Upvotes result in an increase in points for the contributor, while downvotes result in a decrease. While part of the attribution for this success has been the design

of the reputation system, there is also an issue of community organization and outreach (Mamykina et al. 2011). While developing the reputation system, SO engaged with members in the target community over how to organize this system. Therefore, it is important to acknowledge that the design of the reputation system and community structures had iterations and active feedback from the target audience (Mamykina et al. 2011).

The development of the reputation system of SO also means that it has incorporated many proven elements from other peer-production systems. For instance, it boasts an effective Wikipedia-like editing system to provide information which is not a snapshot of the truth at a time, but can evolve as information changes (G. Li et al. 2015). Another element that is borrowed from other peer-production sites is the commenting system that allows users to interact with each other and discuss possible outcomes and changes to information.

2.4.2 The Stack Overflow Reputation System

The Stack Exchange reputation system was built by the founders of the system to be a "rough measurement of how much the community trusts you; it is earned by convincing your peers that you know what you're talking about" ("What Is Reputation? How Do I Earn (and Lose) It? - Help Center - Stack Overflow" 2015). SO has three general reputation classes that describes users with little power "Users", some power, "Established Users" and those with maximal powers "Trusted Users". This classification system signifies how important SO determines reputation points to be.

The community can vote on questions and answers with varying effects. Table 2.1 shows how a user can earn reputation points. In addition, a questioner can choose to give a bounty to the best answerer of a particularly difficult or urgent question. There is a bias for answers, as they are worth double in points compared to questions. Questions are also free to down vote,

where it costs a voter a reputation point to down vote an answer. This was a conscientious design choice, as one of the founders indicated in an interview (Ford, 2012). Specifically, the design choice to weigh answers more heavily than questions was influenced by the belief that "questions are easier to ask…if there were too many questions and not enough answers…the ecosystem would fail." Whether this assertion is true or not is actually an open question for consideration, as the value of a question has fallen from an average of 8 votes at the beginning to under 1 vote by the year 2014.

Action	***Points Earned/Lost***
Answer is Voted Up	+10 (Max 200 Points per day)
Question is Voted Up	+5
Answer is Accepted	+15
Question Maker Accepts Answer	+2
Question is Voted Down	-2
Answer is Voted Down	-2
Voter Votes Answer Down	-1
Edit of Another Post is Accepted by Peer Review	+2 (Maximum of 1000 points gained through editing)
Questioner Offers Bounty of Own Points for an Accepted Answer	Determined by questioner

Table 2.1 List of Actions and Corresponding Reputation Points.

The distribution of reputation on SO, as shown in Figure 2.1, is in a pyramid shape which adheres to the expected Power Law distribution (Mamykina et al. 2011). On SO, only a few hundred users have managed to aggregate more than 100,000 points. This by itself is typical of most peer-production communities, as a relatively small number of users tend to contribute the

most and best material. As Anderson et al. (2012) found, these Elite users, at least in the first few years of the system's existence, were able to identify valuable questions quickly and earned more reputation points by obtaining best answer status from a questioner. This apparent skill, combined with the expanding time horizon and reputation limits, means that the reputation distribution is stuck in an expanding pyramid shape.

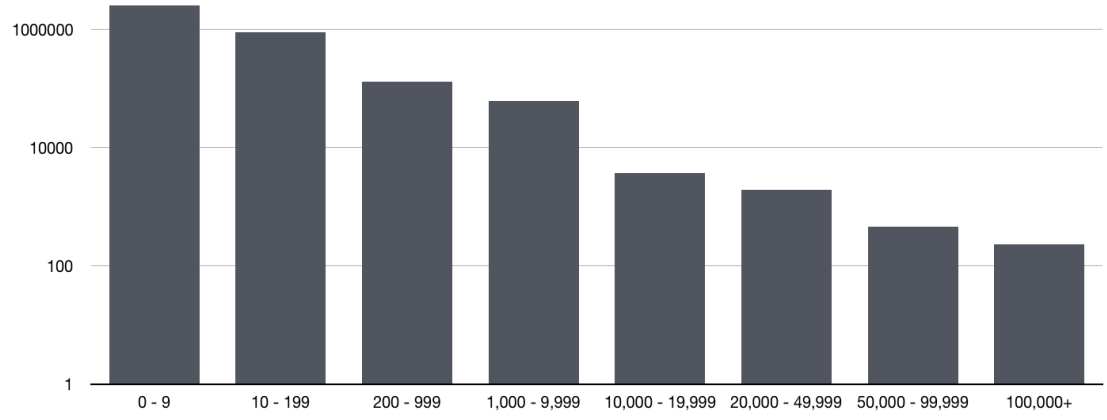

Figure 2.1. Log Distribution of Reputation of Stack Overflow: X – Axis is reputation class; Y – Axis is number of users in the class.

2.4.3 Stack Overflow Privileges

Some community members actively participate in Q&A and in the site's functions. Earning reputation points allow members access to privileges in the site and allows for a user to manage parts of the site's functions. For example, having 15 points allows a user to vote a question or answer up, while 125 points allows a user to down vote. The penultimate privilege can be earned with 25,000 points. Table 2.2 shows some of the important privileges as well as the relative breakdown of users with each respective privilege.

Earning reputation points gives users power, and as Figure 2.2 shows, many of the privileges bestow formidable amounts of power. For example, the "trusted user" privilege

essentially allows a user to have wide-ranging moderation authority over the site and its users. This also gives reputation points in SO a tangible use.

Privileges provide an interesting aspect to this study. The social approval that comes from within community when providing Q&A actions is from a selected group of users who also have privileges as well, and therefore have also received social approval.

Privilege	**Points Required for Privilege**	**Percentage of Stack Overflow Users with Privilege**
Vote Up	15 Points	22%
Leave Comments	50 Points	11%
Vote Down	125 Points	7%
Edit Posts without Approval	2000 Points	1%
Vote to Close any Questions	3000 Points	1%
Vote to Delete Questions	10000 Points	0.1%

Table 2.2 Examples of Privileges on Stack Overflow

2.4.4 Stack Overflow as an Archive of Computer Programming Information

The Q&A sessions themselves are straightforward but have many avenues for interaction. Once a question is asked, it can be edited by members of the community with that privilege, commented on by the community, and voted on by eligible members. A Q&A session can play out in different ways. A question could receive dozens of excellent competing answers, and some of the content could receive hundreds of votes. On the other hand, a question could receive many down votes and be closed from receiving further answers. Some of these questions will also be deleted

from the corpus, which takes a collaborative effort from the community (Correa and Sureka 2014).

Stack Overflow is strict as to the types of questions that can be asked; both in terms of question scope and in terms of question formation. For example, SO's general rules require questions to be "specific to computer programming or the computer programming profession.". Questions are also required to be answerable, well researched, and have transcendence beyond just the questioner. These restrictions on SO content are something that helped produce its early success in domain-specific Q&A as it was building a tightly engaged community (Mamykina et al. 2011), as well as to manage the value difference between surplus questions and scarce answers (Ford 2012).

2.5 Summary

Q&A is complex and comes in many different forms. The central theme of most previous research has focused on understanding why users answer questions, extracting the best answers from corpora, and identifying authoritative users. There are still a large number of unexplored behavioral issues in Q&A that are essential for research and designers to understand. One of these areas of research is in analyzing how reputation and reputation classes are related in predicting or identifying socially desirable behavior.

In this survey, we have laid out the general taxonomy of Q&A communities and how the assumptions of each community type are required to be considered when analyzing social incentives and user behavior. We also described the research about the motivations for asking and answering questions as well as describing the current understanding of what reputation and reputation classes indicate. Finally, we briefly described the target system's infrastructure. In the following chapters, we focus on the relationship between reputation classes and social incentives.

Chapter 3.

Identity and Reputation in Q&A: A Grounded Theory Study

3.1 Introduction

In this chapter, we consider whether reputation class correlates with identity choice. Decentralized knowledge-creation systems, like Q&A, depend on strangers from around the world choosing to interact with each other. In many systems, users can choose how to represent themselves in the systems, from giving personally identifying information to choosing to use a pseudonym. We specifically analyze identity and performance within Stack Overflow (SO). When considering the Expert-Based aspect of a technical system like SO, it would make sense that community members would desire that expertise in the domain area is validated through links to real world identities. Unsurprisingly, out of the very top 36 overall users on SO, only 2 are not identifiable in the offline world.

We investigate how users formulate identity in relation to real-world identities, whether identity prevalence changes over time, whether certain types of identities outperform other types, and whether individual identity factors are associated with better performance in reputation earning on a per contribution basis (i.e., how many points a user accrues for each question or answer posted). It is our goal that gaining insight into the role of identity in Q&A systems can help designers of these systems determine identity rules, as well as to allow individual users to optimize their experiences within these systems and understand the consequences of their identity choices.

Research on identity in the web does not have a conclusive determination, but it does tend to correlate customization of an account with better question and answering performance, as well as more contributions (Adaji and Vassileva 2016; Ginsca and Popescu 2013). One line of

research considers that there is a social cost for new or pseudonym identities (Friedman and Resnick 2001). The notion of identity in Q&A has largely dealt with authority and trust, most often from a third-party rater analysis; that is, evaluators who come from outside the community and are specifically asked to rate material (Golbeck and Fleischmann 2010). An additional method of exploring identity revolves around conducting interviews of site participants (Jeon and Rieh 2014). Results from these studies are informative but may not show how identity is related to performance in the community, and small sample sizes can limit generalizability. Finally, some research uses big data in order to categorize the behavior of participants (Anderson et al. 2012; Ginsca and Popescu 2013). These studies are valuable for identifying trends, but they do not formulate concrete definitions of identity and can omit data which is not easily categorized.

Also, in previous studies, there is a gap in the understanding of the relationship between identity and community. Ginsca and Popescu conclude that assumptions could be made that fuller disclosure of identity is important in performance and standing (Ginsca and Popescu 2013), and therefore we would expect the community to reflect that. On the other hand, Jeon and Rieh conclude that it is possible that identity does not matter in a technical community such as SO (Jeon and Rieh 2014). Therefore, it is difficult for designers to implement informed identity systems. The goal of this chapter is to add to an understanding of how a community organizes itself with regards to identity and how identity is related with reputation-point earning. We focus on reputation since it is a clear way to determine the performance of users (Anderson et al. 2012) and users are motivated to earn reputation points (Anderson et al. 2012; Tausczik and Pennebaker 2012). Based on the research listed in Chapter 2, we assume that users want to earn as many reputation points on each post as possible.

We use Grounded Theory (Glaser and Strauss 2009) to develop classifications of identity. Grounded Theory is a qualitative process that focuses on allowing themes and theories to emerge from data. Data is collected through numerous rounds until a theory is achieved, and we use it for building a classification of identity types. We look at two samples from SO: an overall "General" sample of the site, and an "Elite" sample containing those that rank within the top 50,000 reputation scores in the system. The two samples are chosen because together they represent both the entire community as well as the most active part of the community. We developed seven classifications, with three (Full ID, Link ID, and Pseudonym) appearing in large numbers in both samples.

This chapter's questions and the contributions can be summarized as follows:

1: Is there a relationship with longevity and chosen identity? In both the General and the Elite samples, longer membership in SO correlates with more complete types of identity. This means there is a link between providing more identifying features in a profile at the beginning of the community life-cycle that decreases as time passes.

2: Is there a relationship with identity type and reputation aggregation and efficiency? Full ID users are more likely to have higher point aggregation than users who have Pseudonym accounts in both samples. Full ID users also have higher efficiency scores, receiving more points per contribution divided into total reputation.

3: Is there a relationship with ratio and quality of questions and answers a user contributes and their identity? In both samples, Pseudonym account users are more likely to engage in question-asking compared to other users, suggesting a correlation between an activity and choice of identity. Regarding performance on individual actions, in the Elite sample, there is no significant difference between the ability to score highly on answering questions or asking

questions. This indicates that the reputation advantage is a benefit of longevity, rather than directly correlated to skill or bias in the system.

4: Do individual identifying features improve reputation earning efficiency? The presence of a profile picture, full name, or web link does not predict higher reputation earning efficiency in the Elite or General Samples. This indicates that the individual factors do not correlate with higher earning efficiency compared to the composite profiles. We do find that the number of months of membership for a user is predictive of higher efficiency for both samples.

5: Is identity static or fluid? There is a visible shift in the makeup of user identities as the system ages, from more complete to less complete. Also, a look at the profiles 12-18 months after they were collected indicates that most profiles do not change from one type to another.

The results show that while more identities filled with real-world identifying information are correlated with longer tenure and higher reputation aggregation, the actual information does not cause a certain type of performance. Gaining insight into the role of identity in Q&A systems can help designers of these systems determine identity rules, as well as allow individual users to optimize their experiences.

3.2 Identity in Online Communities

Studies on identity in online communities often center on trust, credibility, and a transaction. For instance, in online dating, participants are spending possible combinations of time, effort, and money, and therefore seek trust and credibility with potential partners (Couch and Liamputtong 2008). In these situations, users will try to manage identity to match their ideal selves (Ellison, Heino, and Gibbs 2006). Online auction sites require trust and credibility in order to facilitate the exchange of goods and money (Friedman and Resnick 2001). In cases such as this, the identity of

seller is formulated based on feedback left by previous customers, while the identity of the buyer does not matter as much (Zhang 2006).

Q&A communities have a similar issue with trust. Q&A sessions can be effort intensive, and users may incur prohibitive costs if they choose to interact with less than optimal counterparts (Jain, Chen, and Parkes 2009). Therefore, researchers have often sought to understand the explicit relationship between identity and trust (Golbeck and Fleischmann 2010; Jeon and Rieh 2014). However, trust may not be the most important factor for Q&A users. Instead, it could be speed of answer (Jain, Chen, and Parkes 2009) or reputation earning potential of a question (Anderson et al. 2012). In this study, we do not explicitly analyze the relationship between identity and trust, but rather use the peer-assessment of SO users to judge performance within the system. An advantage of this approach is that it broadens the applicability of the study by incorporating trust as one of many possible factors in peer-assessment.

3.2.1 Studying Identity in Q&A

A popular approach to studying identity in Q&A is to use qualitative studies. Golbeck and Fleischmann (2010) used expert and non-expert raters to judge the impact of text and visual cues of expertise when establishing trust in a Q&A system centered on dog care. In this case, having profile pictures increases the trust level for non-expert raters, but not for experts. However, when there are text cues that suggest expertise and experience, both experts and non-experts report higher trust levels.

Jeon and Rieh (2014) conducted an experiment in which participants rated answerers' credibility based on trustworthiness and expertise in a Q&A system (Yahoo! Answers), Some of the raters found that answerers who had uploaded profile pictures onto the site showed increased investment in participation and providing answers. However, these assessments were not

universally shared by all the raters. Another area Jeon looked at was whether non-answer-based cues played into the determination of answer credibility. Raters seemed to focus more on the answer itself rather than on other cues. Hart and Sarma (2014) confirms this by noting that third-party raters focus on the accepted answer regardless of the account origin. This includes ignoring visual cues such as badges, reputation points, and profile pictures.

Raban (2009) used Grounded Theory to analyze participants on Google Answers, a for-pay generalist Q&A system that closed in 2006. This study found that only implicit representation of self-presentation had an increase in the social system. This is an important finding, which indicates that explicit social cues may not have an impact, whereas social credibility can be cultivated through exhibiting knowledge in action.

Not all studies are qualitative. Ginsca and Popescu (2013) looked at identity and Q&A performance on SO. The authors of this work found in a set of 75,657 users that the more complete a user profile was (i.e., user name, avatar, etc.), the higher the probability that this user would provide a quality answer, with the number of aggregate votes being the determination of quality. This study does not take any temporal aspect into account, with the exception that answers are given a two-month time-frame to earn votes. This means that performance could simply correlate with longevity. If new accounts lack information, it is possible that an assumption that better performance is a result of providing more identifying features is not true.

Another study, Adaji and Vassileva (Adaji and Vassileva 2016), followed up Ginsca's study and concurred with the results, indicating that users who fill in account identifying features (i.e., profile picture, age, location, and url) are more likely to have higher contributing accounts and give better Q&A contributions. Like Ginsca, the study lacks a temporal aspect. In addition, no consideration is given to what users enter in the fields. That is, the study indicates that users

who fill out the identity fields are more likely to contribute and contribute well, but it does not indicate that what is in the fields is correlated with better performance.

Thus, the results of these studies show that there is a gap in the understanding of the relationship between identity and community. Previous research indicates that identity does not affect the evaluation of contributions (Raban 2009), but that having more identifying features results in better Q&A answering (Ginsca and Popescu 2013). An approach to reconciling this difference is by constructing identity classifications and measuring performance and activity type amongst similar groups. In this study, we accomplish this by constructing identity classifications using Grounded Theory.

3.3 Profiles on the Web and Stack Overflow

Each account on SO has a profile page that allows for a wide range of customization. As shown in Figure 3.1, users can choose to share a profile picture, their real name, links to off-site pages, location, current employment, and CVs. Users can instead choose to use their randomly generated default user name and a randomly generated fractal avatar and give no further details. There is no additional privilege or penalty for sharing or withholding information. An example of a SO profile space can be observed in Figure 3.1.

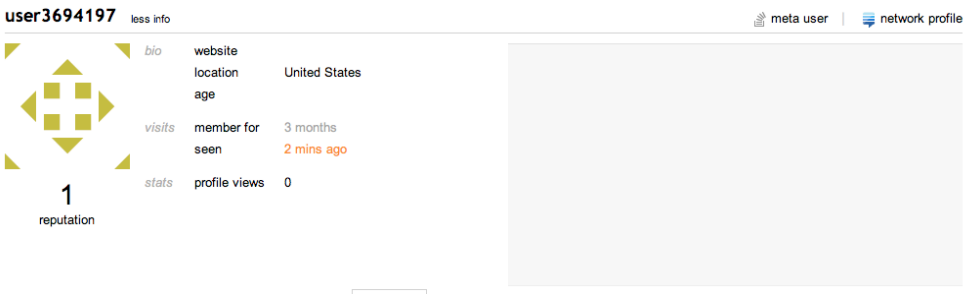

Figure 3.1. Example of a profile on Stack Overflow

There are multiple ways in which a user can sign up for an account. Since its inception in 2008, SO has supported both account creation via its own infrastructure, or through OpenID ("OpenID, One Year Later" 2010). OpenID allows users to sign-up for an account by using another account that utilizes an OpenID structure. For instance, a user could create an account with a Facebook account. A user could also initiate an OpenID account with Stack Exchange. This does mean that there are two different ways in which identifying information is given. When creating an account from scratch, identifying information is manually added by the user. When a user logs in from an OpenID provider like Facebook or Google+, the information fields are automatically populated, which can then be edited by the user.

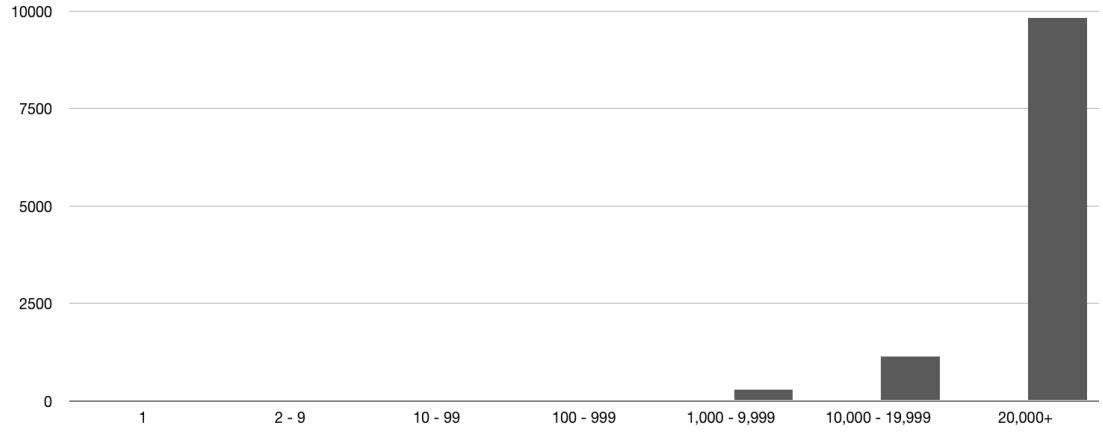

Figure 3.2 A Linear Distribution of Profile Views (y-axis) and Reputation (x-axis)

There is a close correlation (R=0.9339) between reputation and the amount of views a profile sees. As indicated in Figure 3.2, the views are under 1 for accounts with 1 point and climb to 10,000 for users who have achieved 20,000 or more points. This indicates that successful users receive attention to their profiles. Users may consider this when choosing whether to share personal information in their profiles.

Stack Exchange's profile scheme facilitates an active mix of profiles. There is a potential to gain rich interaction information as members are not required to share information but can do so easily if they so choose. As this has been a consistent rule of SO from its inception, this site allows us to view how a community can evolve when personal information sharing is decided independently.

3.4 Description of Samples

Stack Overflow (SO) has over 8 million registered members, but, as Figure 2.1 shows in chapter 2, millions of users have few points. While obtaining a breakdown of the entire community is interesting, we also wanted to look at the core community of users who have had success with the reputation system. The advantage of this method is that we are more clearly able to understand how the community interacts with users who have contributed to the site. Therefore, we sample from the entire contributing community and from the top contributors. We formulated the data sets by using the data explorer available from stack exchange (data.stackexchange.com). Using the SQL queries the we wrote, we obtained unique profiles that fit the parameters of each data set.

The first sample, which we refer to as the "General Sample", gathered data from the entire population that has contributed at least one answer or question. As of January 2018, 3.6 million users had at least one question or answer. Having at least one post is a requirement to be included in the sample because it would be impossible to measure the performance of users who have not contributed a post. In addition, previous research (Adaji and Vassileva 2016; Ginsca and Popescu 2013) has already considered the participation rates of users with different identity factors. We took a sample of 1,037 users (Confidence=99%, Margin of Error=4%), and recorded their profiles. The reputation distribution (M=229.9, SD=1053.24) shows that only a few users

have numerous points as shown in Figure 3.4. On the other hand, there are users who have only 1 default reputation point. 248 users have contributed at least one answer or question but have not been awarded any reputation points. These are users who have contributed but have not been rewarded with positive reputation points by the community. Figure 3.3 shows the reputation distribution for the General Sample.

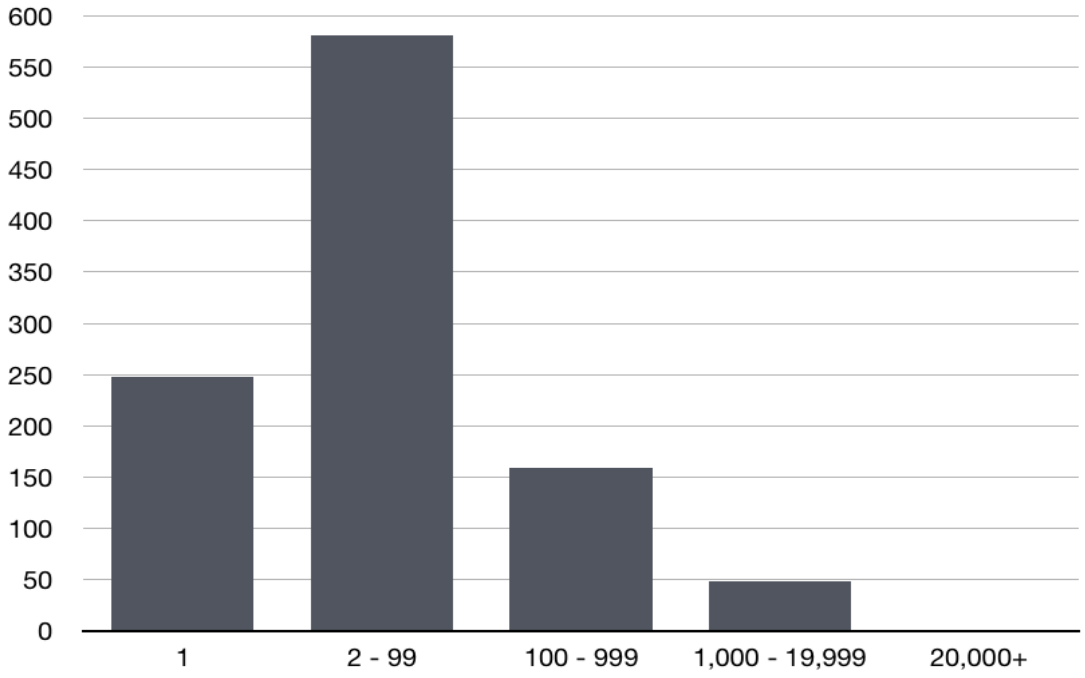

Figure 3.3. Linear Distribution of Reputation of General Sample: X – Axis is reputation class; Y – Axis is number of Users.

Next, we looked at users who comprise what we termed the "Elite Sample" as shown in Figure 3.4. We drew from users who were in the top 1% of all users (About 50,000 users). We took a random sample of 1,017 users (Confidence=99%, Margin of Error=4%), and recorded their profiles. The reputation distribution (M=6668.8, SD=15037.55) also shows a very steep pyramid. Users with 1,358 to 9,999 points make up 87.5% of the sample. The highest-ranking

user in this sample had 188,533 points. The reputation distribution of the Elite sample is shown in Figure 3.4.

It is also important to note that some users in the General sample would also be in the Elite sample. 39 users, (or 3.7%) of the General sample users would also be eligible for the elite sample. The larger than expected amount can be explained by the sampling method, which required that users make at least one contribution, reducing the population. A minimum of one contribution is required because it is impossible to judge the performance of users without any contributions.

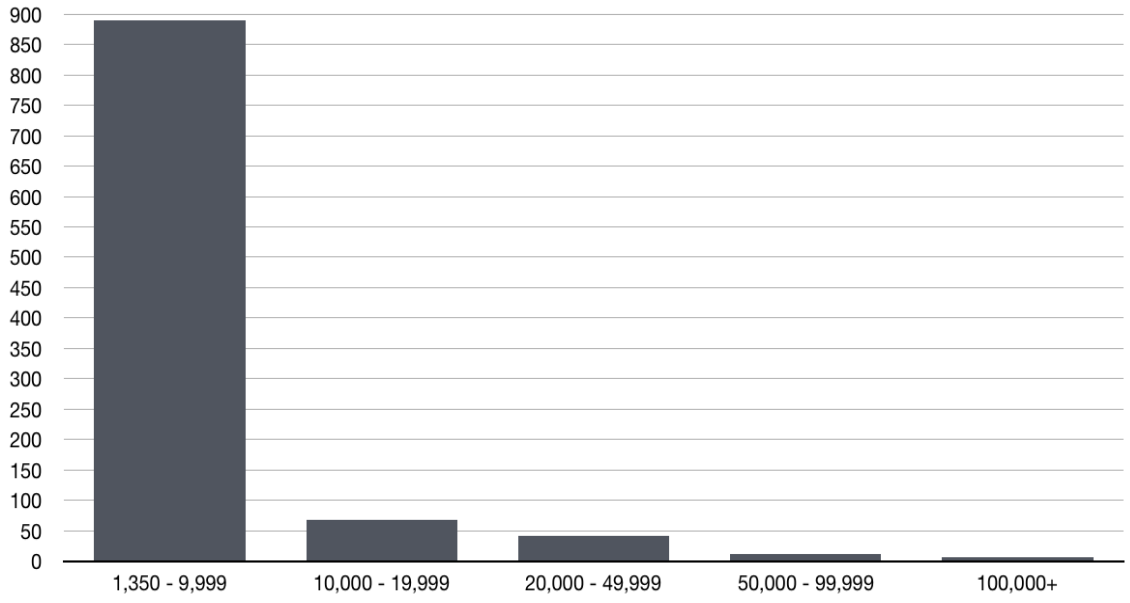

Figure 3.4. Linear Distribution of Reputation of Elite Sample: X – Axis is reputation class; Y – Axis is number of Users

3.5 Observable Data Points for Profiles

As shown in Figure 3.1, there are several areas where we can extrapolate identity information from each profile. Since this paper looks at the relationship between identification and

performance, we focused on four areas: username, profile picture/image area, website links, and the information box to the right. The information box presents an interesting challenge since almost anything can be inserted into it. Some users will write long biographies, including important parts of a CV, or insert many links to various sites.

Two areas that we did not consider for this paper were age and location. While both pieces of information are interesting, they fall out of the scope of this research. We only observe whether users are making their real-world person identifiable or not.

3.6 Research Questions

In this section, we describe four research questions that are focused on identity, longevity, and performance within the Stack Overflow (SO) community.

R1. Is there a relationship with longevity and chosen identity? SO has been an active community for over six years. There is a possibility that the idea of community has shifted over time and that impetus to share personal information with the community is different based on when the user joined the site.

R2. Is there a relationship with identity type and reputation aggregation and efficiency? Success in SO is often measured by the amount of reputation points that a user accumulates. In addition, success can be measured by examining the average amount of points that a user earns per action, which we call "efficiency". This average provides a metric for evaluating a user per contribution.

R3. Is there a relationship with ratio and quality of questions and answers a user contributes and their identity? Certain identity types may excel at certain types of activities. For instance, if profile pictures were very important in establishing trust, it would be expected that *Full ID* profiles would have a higher rate of accepted answers.

R4: Do individual identifying features improve reputation earning efficiency? The profile categorizations are created out of a combination of user names, profile pictures or images, and links to the outside websites. In R4 we look at the individual features of the profiles and test whether they predict higher reputation earning efficiency. For instance, it may be that only profile pictures have a correlative effect with better performance irrespective of the classification of the account.

R5. Is identity static or fluid? Users may or may not keep their identity type for a long time. For instance, users may change from providing little information to large amounts of information.

3.7 Methodology

Having such a diverse collection of profile types creates a challenge for categorization. There is the issue of determining how data is measured. While there might be some factors that allow for quick categorization (for example: location), it is difficult to create an accurate categorization based on multiple factors *a priori*. Indeed, while quantitative measurements are used for analysis, the categorization for this research is highly dynamic and must be calibrated for them to have meaning. Thus, the taxonomy is qualitative, focusing on what the owner of the profile "intended to do".

A common problem with the analysis of qualitative data categorization is the consideration of whether research is purely inductive (Glaser and Strauss 2009) with themes appearing on their own, or whether it is influenced by the research aim (Bruce 2007). In this article, the process of inquiry is interpreted through our research aims, indicating that the process is highly reflexive (Bruce 2007; Mauthner and Doucet 2003; Srivastava and Hopwood 2009). Users on Stack Overflow (SO) are not choosing an identity type from a limited range of choices;

indeed, they are not even specifically choosing an identity type per se. Instead, they are making autonomous choices that lead toward an information sharing level. We use a reflexive iterative process that develops connections on multiple reviews of the data. This process is closely related to Grounded Theory and includes the open coding, axial coding, and selective coding phases. A key understanding, however, is the acknowledgement that the process is inductive and driven by research questions. Thus, each review round uncovers further questions that lead to refined categorization.

3.7.1 Reflexive Iterative Sub-Categorization Process

The goal of the reflexive iterative process in Grounded Theory is to accurately reflect the intentions of the end user (Srivastava and Hopwood 2009). Profiles on SO can and have been analyzed using broad-base terms. To explore this, we performed a review of data and identified themes of information sharing. This allows us to better understand dominant patterns of information sharing within the community.

Two coders, the first author and one additional coder, who has a Master's degree in Information Science, reviewed the profiles pages that were in the sample. The coders reviewed a total of 100 profiles, delineating patterns of identity contribution from the profiles. Disagreements were resolved through discussion. The initial coding phase identified 26 open codes (9 related to avatar, 5 related to name, 8 related to off-site links, and 4 related to the description). From these codes, we created seven categories that related to information sharing. To initiate the categorization of the samples, we focused on types of information that indicated active sharing on the part of the account holder: profile pictures and linked information to a site that is outside SO or Stack Exchange. From these two points we sought to discover if these data

types could be used to identify the account holder. We performed three Yes or No (Y/N) sub-categorization tests on all the profiles in this order:

Q1. Does the account include a facial portrait of the account holder (Y - N)? In this question, we sought to determine whether the user had supplied a facial profile picture to the account. We tried to ensure the highest probability that the account did indeed belong to the account holder by running the picture through a Google image search.

Step 1: Is there a facial portrait in the profile page? Y➔Move to Step 2.

Step 2: Does the profile picture pass the Google image search? Y➔ Profile is determined to have a facial portrait.

Q2. Is there a link that points to an outside page where the account holder's name can be confirmed (Y - N)? Some profiles include what looks to be a full name. In these cases, if there was a link to an offsite page, we sought to determine if the name could be confirmed in the link. Confirmation was achieved when the full name was used in a professional project context (CV, mailing address, university page, company ownership, or company profile).

Step 1: Is there a full name on the profile page? Y➔Move to step 2

Step 2: Is there a link to an off-site webpage? Y➔Move to step 3

Step 3: Is there a confirmation of full name within two clicks in the link? Y➔Profile is determined to use a real name.

Q3. Does the link to an outside page confirm the account holder's name (Y - N)? In some cases, it is not clear whether the account holder's name is fully written in the account, but there is a link to an outside web page. In this case, we looked for the same signs of confirmation as in question 2, but we added the step of requiring a link back to the profile page to confirm ownership.

Step 1: Is there a link to an off-site webpage? Y➔Move to step 2

Step 2: Is there a confirmation of full name within two clicks in the link? Y➔Profile is determined to have a confirmation of a real name.

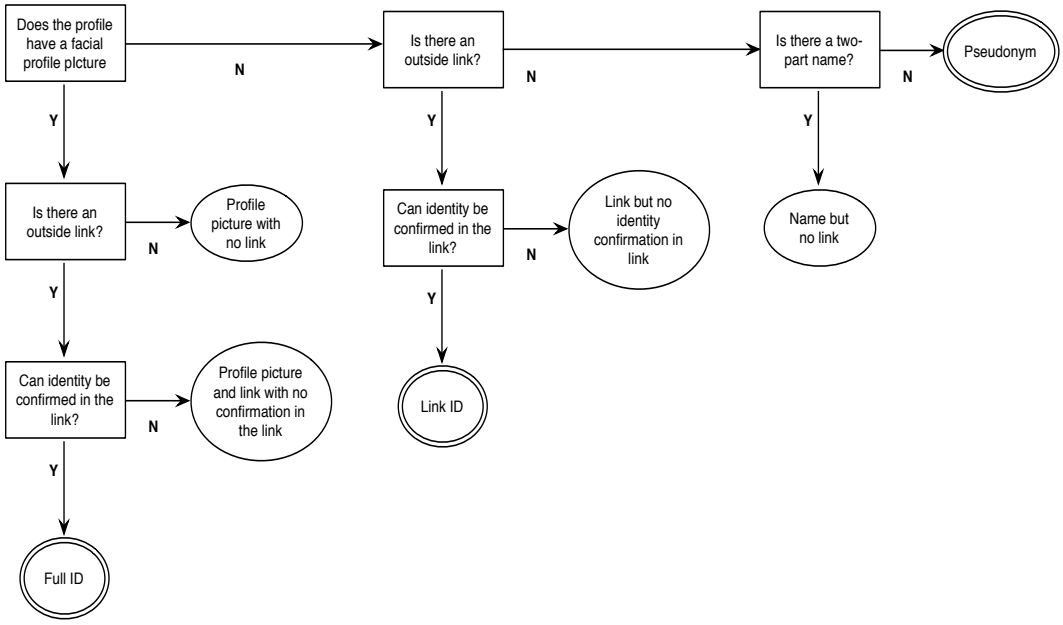

Figure 3.5. Decision Tree for Categorization of Profiles on Stack Overflow: Circles Indicate Terminal Points; Double Circles Indicate Highlighted Terminal Points.

Q4. Does the link point to an outside page where the account holder's name cannot be confirmed (Y - N)? In some cases, it is not clear if an account has a link to a real name or not. These could be sites in which there is no clear owner, or an unrelated page.

Step 1: Is there a link to an off-site webpage? Y➔Move to step 2

Step 2: Is there no confirmation of identity within two clicks? Y➔Profile is determined to have a link with no identification.

Q5. Does the account user have a two-part name that matches a name on Google, yet is unlinked (Y - N)? Accounts sometimes contain a user name that appears to be a real name yet is unlinked. In these cases, the users may not intend to engage in a Pseudonym account. There is also a consideration of culture and language; while some names are easily recognizable to the authors, others may be unknown. Therefore, we used an inclusive test. The limitation of this test is that it identifies names that could be real to a human reader, rather than definitively stating that the names are belonging to the users.

Step 1: Is there a user name made up of more than one word? Y➔Move to step 2

Step 2: Does a Google search match any names? Y➔Profile name is determined to match a real name.

3.7.2 Categorization of Profiles

The next step in the process was to use the sub-categories to create coherent categories that related to the information sharing intention of each account holder. The Elite user sample was the first data set to be reviewed, followed by the General sample. We developed seven distinct full categories, as shown in Table 3.1, based on attributes of the five sub-categories, and categorized by the decision tree found in Figure 3.5. Each profile can only belong to one category. The makeup of the General and Elite samples is different. For both samples the largest single category is the Pseudonym designation, with 59.59% for the General sample, and 29.20% for the Elite sample. These profiles have no facial profile picture, real name, or link to an outside website. The one exception for this category is that some accounts had links to search engines, such as Google, or the Stack Overflow homepage. These types of links were regarded as being non-personal and did not figure in further analysis. An important designating attribute of a

Pseudonym profile is that, without a detailed search, it would be difficult to obtain the real-world identity of the account holder.

Category	Number of Users: General	Percentage of Users: General	Number of Users: Elite	Percentage of Users: Elite	Description
Full ID	**65**	**6.27%**	**259**	**25.46%**	**Profile facial picture Real name or link to real name**
Link ID	**48**	**4.63%**	**173**	**17.01%**	**No profile facial picture Real name or link to real name**
Pseudonym	**618**	**59.59%**	**297**	**29.20%**	**No profile facial picture No link to real name or**
Link but no identity confirmation in link	36	3.47%	83	8.28%	No real name in link No profile facial picture
Profile picture but no link	52	5.01%	94	9.24%	Profile facial picture No link off site webpage
Profile picture and link with no confirmation in link	53	5.11%	19	1.86%	Profile Picture Link to off-site webpage with no identity
Name but no link	165	15.91%	92	9.04%	No profile facial picture Matches a two-part real name No link

Table 3.1. Categorizations of SO Users

The next two groups contained identification information. We call these *Full ID and Link ID.* A profile is categorized as *Full ID* when there is a profile picture and a real name, or a link to a real name. For instance, a user with a profile picture who has a nickname as user name, but includes links to a full CV, would be eligible for the *Full ID* designation. *Link ID* has the same rules, sans the profile picture. While only 6.27% and 4.63% of accounts were *Full ID* and *Link ID* respectively for the General sample, the Elite sample had *Full ID* at 25.46% and *Link ID* at 17.01%.

These first three categories make up 71.67% of the Elite sample and 70.49% of the General sample. From the iterative analysis viewpoint, these categories are complete. That is, we can understand clearly the intention of the profile owners. This is not true for the next four categories: *Link but no information; Profile picture but no link; Profile picture and link with no confirmation in the link;* and *Name but no link.* In each of these categories, there can be some confusion about what the intention of the account holder is.

In *Link but no information*, the account has a no profile picture, but does have a link that leads out of SO. This link does not contain any information on who owns the account. For instance, a link could lead to a Github page of a project that has no obvious owner.

In *Profile picture but no link*, there is a profile picture that passes an image search for uniqueness. In some cases, it is even possible to assess who likely owns the account based on this search. However, there is no direct link available on the profile page. *Profile picture and link with no confirmation in link* is a mixture of the previous two categories. There is a picture and a link, but the link contains no name or confirmation of identity.

The final category, *Name but no link*, is made up of accounts that would have otherwise met the level of being a *Pseudonym*. These accounts, however, have a two-part name that could

be a real name. We used a search engine to check any user name that consisted of two parts or initials in front of one word. Results that matched any name were separated into this category.

A chi-squared test, as shown in Table 3.2, of distribution was performed to determine the independence of the General and Elite samples, and a difference between the populations was found. $X^2(6, N = 2054) = 366.75$, $p = 0.001$. This result indicates that the General sample identities skew heavily towards the use of *Pseudonyms*, compared to the Elite sample whereas the Elite sample has more Full ID and Link ID accounts.

Sample / Category	**Full ID**	**Link ID**	**Pseudonym**	**Others**
General	65 [163.58] (59.41*)	48 [111.58] (36.23*)	618 [461.95] (52.71*)	306 [299.89] (0.12)
Elite	259 [160.42] (60.57*)	173 [109.42] (36.94*)	297 [453.05] (53.75*)	288 [294.11] (0.13)
				Chi-Squared = 299.85*

Table 3.2. Chi-Squared Test of Independence. Observed samples first, Expected Observations in Brackets, Chi-Squared Statistic in Parentheses *p<0.05

3.7.3 Summary of Categorization

The results of the categorization process provide us with an interesting picture of the community evolution. Users who are in the Elite sample do have many ways in which they present their profiles, including *Pseudonym* profiles. While the ID classes are bigger when combined, the *Pseudonym* designation is the largest single classification. However, for the General sample, we see that the *Pseudonym* profile is by far the largest category, while the ID classes are much smaller.

3.8 Analyses of Identity Performance

Since the intentions of non-complete categories are not completely decipherable through the process and the numbers of users in these categories are relatively small in the Elite and General samples, this article focuses on Pseudonym, Full ID, and Link ID only for the quantitative analyses for both the General and the Elite samples.

We developed tests to analyze the categories to answer the research questions. All test results reported at $p<0.05$ with the appropriate Bonferroni correction due to repeated tests on the same data set. Kruskal-Wallis Test was chosen, as data is non-parametric. Post-hoc tests were conducted with Dunn's Test. Corrected p-values are listed under Tables 3.3, 3.4, and 3.5.

Test		Full ID	Link ID	Pseud.	Kruskal-Wallis	Dunn	Dunn	Dunn
		A	B	C		A vs. B	A vs. C	B vs. C
Inc	M	51.87	42.98	31.92	**H=48.925***	1.5757	**6.3248***	**3.5026***
	SD	23.5	20.5	20.4				
Rep	M	1222.3	885.56	120.78	**H=80.447***	1.4446	**7.9512***	**5.0818***
	SD	3107	1860	582				
Effic	M	72.19	22.43	15.69	**H=35.794***	1.7990	**5.5996***	**2.5880***
	SD	305	26.2	32.4				
QperA	M	0.75	0.74	1.51	**H=54.777***	0.1246	**-6.6449***	**-5.9409***
	SD	0.80	0.83	0.80				

Table 3.3. Results of Tests for General Sample 1-4 (Long= Longevity, Rep= Reputation, Effic= Efficiency, QperA= Questions per Answer): Kruskal-Wallis *p<0.0125; Dunn's *p<0.016

Test		Full ID	Link ID	Pseud.	Kruskal-Wallis	Dunn	Dunn	Dunn
		A	B	C		A vs. B	A vs. C	B vs. C
Inc.	M	52.23	48.74	46.91	**H=15.431***	**2.1900***	**3.9043***	1.2222
	SD	15.77	16.62	16.35				
Rep	M	8438	7871	4519	**H=7.156**	---	---	---
	SD	17822	16939	6448				
Effic	M	47.69	36.66	39.54	**H=13.019***	1.7832	**3.605***	1.3756
	SD	56.88	25.53	60.84				
QperA	M	0.23	0.28	0.53	**H=10.621***	0.2040	**-2.7894***	**-2.6890***
	SD	0.30	0.44	0.72				

Table 3.4. Results of Tests for Elite Sample 1-4 (Inc= Incumbency, Rep= Reputation, Effic= Efficiency, QperA= Questions per Answer): Kruskal-Wallis *p<0.008; Dunn's *p<0.016

Test		Full ID	Link ID	Pseud.	Kruskal-Wallis	Dunn	Dunn	Dunn
		A	B	C		A vs. B	A vs. C	B vs. C
Ans Sc	M	35.68	34.27	34.83	H=0.9623	---	---	---
	SD	13.26	13.02	14.50				
User	No.	252	169	272				
%		97.29	97.68	91.58				
Ques. Sc	M	5.23	4.64	3.81	H=7.3757	---	---	---
	SD	5.83	5.44	4.13				
User	No.	122	90	169				
%		47.10	52.02	56.90				

Table 3.5. Results of Test 5-6 (Ans Sc= Answer Score, Ques Sc= Question Score): For Elite Sample Kruskal-Wallis *p<0.008; Dunn's *p<0.016

3.8.1 Tests and Results

R1. Is there a relationship with longevity and chosen identity?; In order to answer this

question, we measured longevity (Long) as the length of membership in months. In the General sample, we see that there is a significant difference between classes ($H=48.925$, $p<0.0125$). Dunn's Test found significant differences between the ID classes and *Pseudonym* users, as shown in Table 3.3. For the Elite sample, we also find a significant difference between classes ($H=15.431$, $p<0.008$); however, we see with Dunn's Test that *Full ID* users have longer tenure than both *Link ID* and *Pseudonym* users, while there is no difference between the *Link ID* and *Pseudonym* users. In both samples, the results indicate that users who give the fullest form of identity have been members the longest. *Link ID* users fall somewhere in the middle of the two, possibly suggesting an incremental evolution in the organization of the community.

R2. Is there a relationship with identity type and reputation aggregation and efficiency? To answer this, we measured both reputation (Rep) and efficiency (Effic). Reputation looks at the raw point aggregation. There is a significant difference between classes in the General sample ($H=80.447$, $p<0.0125$), and as Table 3.3 reports, the Dunn's Test reports a significant score between both ID classes and the *Pseudonym* class for reputation score. The Elite sample, shown in Table 3.4, does not return a significant Kruskal-Wallis result ($H=7.156$, $p<0.008$).

Next, we looked at Efficiency, which is defined by Expression 1 as the points earned per action and is a measurement of overall performance:

$$N = TotalReputation/(Answers + (Questions/2))$$

As shown in Table 2.1 in chapter 2, answers are worth 10 points while questions are worth 5 points. Therefore, 2 divides the total number of questions. This system ignores points earned from accepted answer selection and does not consider points earned from bounties. However, this rating system still gives a valid result since we do know that most points are earned through votes on answers and questions (Anderson et al. 2012).

The results for the General sample in Table 3.3 show a difference between classes, ($H=35.794$, $p<0.0125$) and Dunn's Test shows that the ID classes have significantly higher efficiency than the *Pseudonym* class. In the Elite sample, we see that there is a significant difference between classes ($H=13.019$, $p<0.008$) and Dunn's Test shows a difference between the *Full ID* class and the *Pseudonym* class, where the *Full ID* class earns higher efficiency. This, at least on a surface level, would indicate that more complete profiles do outperform profiles that do not reveal identity information.

R3. Is there a relationship with ratio and quality of questions and answers a user asks and their identity?: We measured the ratio of questions to answers (QperA), answer acceptance score (Ans. Sc), and question score (Ques. Sc.) in order to answer this question. Looking at the ratio of questions to answers, instead of using raw ratio numbers, we use a scale, since the goal of the test is to capture the balance of question asking and answering. This scale runs from 0-2. A score of "0" would indicate that the user never asks questions, while a score of "1" would indicate that a user asks equally as many questions as gives answers. Finally, a score of "2" would indicate that the user asks at least twice the number of questions to answers.

In the General Sample, we find a significant difference between classes ($H=54.777$, $p<0.0125$). As Table 3.3 shows, the *Pseudonym* users from the General sample have a higher a ratio of questions to answer than the ID classes. The Elite sample, as shown in Table 3.3, has the same result ($H=10.621$, $p<0.008$). This indicates that Elite *Pseudonym* users are more likely to ask more questions in relation to answers than other users. An important thing to emphasize, however, is that *Pseudonym* also has the largest standard deviation by a large margin. Not all *Pseudonym* users ask a high ratio of questions, but this class is more likely to do this compared to other classes.

We next look at the percentage of answers that are accepted by the question-asker. The metric shows the ability of a user to compete for the "accepted answer" designation from questioners. A high rating also signals ability to provide answers that satisfy the questioners' problem quickly. In a way, this test provides a better test of capability than measuring votes. While an answer that receives many up-votes from the community is a signal of quality, it is also bound by temporal constraints that may give an unfair advantage to users who have longer tenure. That is, answering a simple, but commonly asked, question may receive more votes than a technically brilliant answer to a more obscure question.

In this test, we only look at the percentage of users who have given more than 10 answers. The purpose of introducing this threshold is to assure that users have multiple answers on which performance can be analyzed. Over 97% of the *Full ID* and *Link ID* classes passed this threshold. The *Pseudonym* class was not far behind with over 91% of users reaching the threshold. There was no significant result in the Kruskal-Wallis ($H=0.9623$, $p<0.008$). A closer look at the data in Table 3.5 shows that different classes have very similar means and standard deviations ((*Full* ID (M=35.68, SD013.26), *Link* ID (M=34.27, SD=13.02), *Pseudonym* (M=34.83, SD14.50)).

Finally, we look at the average question score for each class. This test simply averages the total number of votes that a question receives. Unlike Test 5, this test is vulnerable to the effect of time and relative competitiveness. In other words, it is possible that certain users could receive a benefit from being active during certain periods.

Once again, the threshold for being included in the test was having more than 10 questions. Again, this threshold assures that users have multiple samples for which to measure performance. The Pseudonym class has 56.9% of users in this test, while Link ID has 52.02%

and Full ID has the least with 47.1%. The Kruskal-Wallis test, as shown in Table 3.4, is not significant ($H=7.3757$, $p<0.008$), indicating that there is no difference between subgroups.

R4: Do individual identifying features correlate with higher reputation earning efficiency?

While the tests on the classifications do not show an advantage for users with identifying features, this does not mean that the individual factors which make up the classified identities are not related to better performance. For instance, it could be possible that providing the information that makes up a *Full ID* account does not provide users with a reputation earning advantage, but that having a profile picture, two-part name, or a link to a personal website does correlate with reputation earning. To determine this, we used a Generalized Linear Model (GLM) to measure the relationship between factors and reputation earning efficiency.

The samples were run in separate models. We chose reputation earning efficiency as the dependent variable because it allows us to view users' average performance on a per-contribution basis, rather than the aggregate of contributions. The number of contributions is highly correlated with the amount reputation a user has earned (General Sample $R=0.74$, Elite Sample $R=0.86$). However, in this test we are seeking to understand if identity is related to per-contribution performance, rather than measuring who has the most aggregated points. For example, a user who has earned 100 points with 10 answers is as efficient as a user who has 1000 points with 100 answers. We defined reputation earning efficiency as Expression 1 in section 3.9.1.

The entire General and Elite samples were analyzed. Individual profile features were used following the workflow of Figure 3.5: 1) Is there a facial profile picture?, 2) Is there a link where identity can be confirmed?, and 3) Is there a two-part name? A differentiation between this work and previous studies into identity, is that our analysis looks for completed profile

sections which are or seem to point to a real-world person, rather than measuring whether the user has completed the section or not (Adaji and Vassileva 2016; Ginsca and Popescu 2013). For example, while providing a link such as www.Google.com would be counted as filling in the URL section in previous work, in this paper, it is not included as a valid URL. In addition, while Ginsca and Popescu (2013) test for a two-part name, they limit the search to the user name feature, while our test includes the entire field of the profile.

Sample	Coefficient	Estimate	Std. Error	t-Value	VIF
General	Intercept	0.5053	0.0338	14.981***	
	Profile Picture	0.0913	0.0502	1.819	1.209681
	Name	0.0353	0.0407	0.868	1.325622
	Web Link	0.0827	0.0658	1.258	1.473376
	Months of Membership (Longevity)	0.0106	0.0008	13.119***	1.062107
				Pseudo R-squared	0.1705
Elite	Intercept	1.1820	0.0259	45.782***	
	Profile Picture	0.0232	0.0188	1.237	1.195261
	Name	-0.0100	0.0259	-0.388	2.502409
	Web Link	0.0134	0.0268	0.500	2.623844
	Months of Membership (Longevity)	0.0065	0.0005	12.966***	1.015852
				Pseudo R-squared	0.1620

Table 3.6. Results of Generalized Linear Models *p<0.05, **p<0.01, ***p<0.001

The models were constructed as follows: Efficiency ~ Profile Picture, Name, Web Link, Months of Membership (which represents longevity). The models were validated with McFadden's pseudo R-squared and by testing model assumptions.

Both the General and Elite data sets have a dependent variable which is non-normal. Log-transformation failed to achieve normality for either data set. However, log-transformation of the dependent variable did reduce heteroscedasticity. The General sample was fit to a GLM with a Gaussian distribution with an identity link function as log transformation of the dependent variable made some efficiency zero values, while the Elite sample was fit to a GLM with a

Gamma distribution with an identity link function as the dependent variables were all positive and skewed to the right. Link functions were chosen with a diagnostic check by comparing plots of residuals versus the predictors. We identified outliers and influential points using the univariate approach and Cook's Distance. Both samples follow the expected power-law distribution (Adamic et al. 2008). Outliers are not removed from the models *a priori*, since the data points are not errors and can be explained by the system mechanics itself. There were no influential outliers according to Cook's Distance.

The independent variables, as shown in Table 3.6, have low VIF scores indicating that multicollinearity is not an issue. Homoscedasticity was checked by visually observing Scale-Location and Residuals vs Fitted plots, in which neither model showed heteroscedasticity. Log-transformation of Months of Membership increased heteroscedasticity, so transformation was redacted. Autocorrelation was checked visually with an Autocorrelation Plot (ACF). Both models showed decay in lags that indicates no presence of autocorrelation. Variability for all independent variables was positive, and the independent variables and the residuals were uncorrelated. Normality of the residuals was confirmed visually with a Normal q-q plot of residuals.

In the General sample model, the pseudo R-squared is around 17%, and in the elite sample it is 16%. These scores indicate that the independent variables, in this case the identity factors, are not leading predictors in reputation earning efficiency. However, we can consider that the results are still important, as we assume that the skill of the users will explain most of the variance in efficiency.

As shown in Table 3.6, Months of Membership is a significant predictor of reputation earning efficiency for both samples (General and Elite GLM: $P<0.01$). This indicates a relationship with longevity and expected per-contribution reputation earning.

We performed robustness checks on our GLMs by running the models through a general linear model and a robust linear model. When running both models under a normal distribution assumption for the log-transformed dependent variable, results were maintained. Likewise, a robust linear model, which compensates for outliers, also returned similar R-squared values and the same significant predictors.

R5 . Is Identity Static or Fluid?: The above results capture the status of the community for two samples at one moment in time. However, it does not tell us if identity type is static or moving. To understand this, we analyzed the profiles a second time; to judge we went back and re-categorized the accounts after they had been initially coded. Because of the order of the initial categorization of data, the General sample was recoded 12 months after the initial coding, and the Elite sample was recoded 18 months after initial coding.

As shown in Table 3.7, very few of the General sample users have changed their type of identity. Only 41 or 3.95% of accounts had a change of status. The biggest gainer was the LinkID account with an increase of 1.16%, and the biggest loser was the Pseudonym class with a -1.16% change.

The Elite sample, which had a longer time from original analysis, does show a larger change. 154 or 15.14% of accounts changed status. The biggest gainer was Full ID accounts with a 4.04% increase. The biggest loser was the Pseudonym account with a -3.24% decrease. However, most accounts, 85%, do not change status, and the breakdown of users is similar to the previous coding. The results of the re-categorization indicate that most accounts are static.

Category	Number of Users: General	New Percentage /Percentage Change	Number of Users: Elite	New Percentage /Percentage Change
Full ID	67	6.46% +0.19%	300	29.50% +4.04%
Link ID	60	5.79% +1.16%	198	19.50% +2.49%
Pseudonym	606	58.44% -1.15%	264	25.96% -3.24%
Link but no identity confirmation in link	36	3.47% +0.00%	80	7.87% -0.41%
Profile picture but no link	54	5.21% +0.20%	86	8.47% -0.77%
Profile picture and link with no confirmation in link	54	5.21% +0.10%	19	1.86% +0.00%
Name but no link	158	15.24% -1.04%	67	6.59% -2.45%
Deleted Accounts	2	0.001%	3	0.003%
Total Number of Accounts that Changed	**41**	**3.95%**	**154**	**15.14%**

Table 3.7. Categorizations of SO Users after 12 months for the General Sample and 18 months for the Elite Sample

3.9 Discussion and Limitations

The results of this chapter indicate that there is no additional social-reward for the provision identity within the community. For a community designer, the overall results are a positive mark for allowing identity choice. Identity diversity is represented in both the General and Elite

samples and Pseudonym users can compete with other identities on a per-answer basis in the Elite sample. This would indicate that allowing for a diversity of choice along a full spectrum of user types, from giving the most identity to using a Pseudonym, does not harm system efficacy and may allow the site to aggregate a wider range of high-quality users.

The relationship between fuller identity and longevity is also interesting for community designers. It signals that the initial relationship that was built between the community and the system (Mamykina et al. 2011) is changing and newer users are not as willing to share personal identities with the community, and do not seem to reveal them over time. This would indicate that type of membership within the community shifts as time passes and the reputation pyramid becomes more static.

There is also significant difference between the classes and the proclivity to have question-oriented accounts. *Pseudonym* users are much more likely to have an account that focuses on asking questions compared to the ID classes. This is interesting because even though these users ask good questions, they choose not to share their identity. This could be out of fear of real world implications, where asking too many questions may show a lack of expertise. When compared to answer-oriented accounts, we see that *Full ID* users are much more likely to have this type of account. This would indicate that users who provide lots of answers want their identity to be known, potentially because it indicates expertise. A Pseudonym account may allow a user to separate their real-world identity from their online information seeking identity, and thus actually improve their contributions as per Donath's (2014) conceptual model. For designers of systems like SO, facilitating the ability to create Pseudonym accounts would appear to be beneficial in gathering and maintaining question-oriented users.

For users of a Q&A community, the results indicate that the practical reality of the community's task-based technical aims means that the quality and timeliness for answers is a more important authority than identity. This trajectory aligns with previous research such as Hart and Sarma (2014), Raban (2009), and Jeon and Rieh (2014), and satisfies the theoretical framework of Jain, Chen, and Parkes (2009). The community is focused on the exchange of knowledge and does not discriminate within discrete interactions. Different types of users have different considerations when choosing to create a profile. Certain users may worry about their real-world lives intermingling with an online community, while others may be concerned that limiting exposure of their identity inhibits their ability to compete and take part in the community. The results of this study mean that users can choose to represent their identity to the extent of their comfort, without experiencing adverse consequences when contributing to the community (e.g., receiving fewer upvotes for a contribution due to an incomplete profile).

The result of R4 would seem to indicate that previous research (Adaji and Vassileva 2016; Ginsca and Popescu 2013) is somewhat wrong, but this may not be the case. In the former studies, what is being studied is whether the act of filling out identity forms indicates they will be better and more consistent users. This chapter shows that the contents of what the users put in the identity forms does not affect performance. This is an important distinction that is informative for site designers. While it may make sense to require users to fill out their profiles with "something", it is not necessary to make them fill out their profiles with personally identifying information.

This also gives a possible explanation for the result of R5. If there was an advantage to disclosing identity information, we might expect to see users change their profiles to give more

information as a function of efficiency seeking. However, as there is no advantage, users do not change profile type often.

An interesting future path of research is to closely investigate the type of information that users share with the community. There is a potential for users to use Q&A sites to show skills for professional development, as evidenced by SO's on-site job-hunting site called Careers 2.0. In addition, researchers found that users in various software development communities, such as Github, were very self-aware of presentation and presence of recruiters (Marlow, Dabbish, and Herbsleb 2013; Singer et al. 2013). Users may change behavior based on the intent of their information sharing in SO and may explain why users interested in asking questions use pseudonyms. Another path of research that needs to be analyzed is that of the role of activist. While we were unable to find any difference in performance when it came to Q&A, there could be potential differences in commenting and editing. Further qualitative research can be done to determine the quality of contributions to answer these questions.

3.10 Summary

In this chapter, we looked at the relationship between reputation and the provision of identity. An underlying assumption based on previous research was that the provision of identity had some sort of social reward attached to it. That is, providing identity was correlated with better performance in the Q&A, either as an intrinsic relationship within the user themselves, or because there is a reward from the rest of community. The results of this study show that while the provision of identity is related to longevity and reputation user class, the results on a per-contribution bases show no bias towards identity types.

A forum like Stack Overflow is run on an open platform that invites anyone to become part of the community and contribute. Because of the community aspect, understanding the

different types and roles that identities play is important for making design choices around the limitations and scope of identity. In this chapter, we created a taxonomy of identity using Grounded Theory and compared their performance against each other.

The results indicate that identity is diverse and relatively static. While accounts with identifying features do outperform accounts that withhold identity in some respects, it does not seem to affect performance on a per-contribution level amongst high-level contributors. This would indicate that a liberal policy of allowing many choices in formulating identity does not have an adverse effect for either system efficacy or the users themselves and may in fact encourage quality contributing users who do not want to divulge personal details.

The results also indicate that the correlation of aggregated reputation and identity is largely an artifact of the longer membership statuses, rather than as an inherent advantage for providing more identifying features. This would mean that identity is not causally related to reputation, but rather that the shape of the community changes, and that the norms of membership are fluid. These results are informative for many stakeholders, including community designers and users.

In the next chapter, we switch from a problem that is focused on the Expert-Based aspect of SO and look at a Collaborative Q&A model problem.

Chapter 4.

Effects of Incentives in Collaborative Editing

4.1 Introduction

In this chapter we look at whether reputation class correlates with the socially desirable behavior of editing bad questions. Stack overflow (SO) has an element of the Collaborative Q&A model in that it actively requests that users change material for the better, both for the archive and for the community. We consider two scenarios in which editing a bad question occurs: competitive editing, in which actors have a reputation-incentive for editing, and intrinsic editing, in which editors contribute with no reputation incentive. In this chapter, we seek to understand what users contribute to the editing and how they edit by analyzing these unfit questions and the actions individuals took to answer and edit them.

One of the technical difficulties in this analysis is the ability to accurately identify actors and classify data adequately. While SO allows for some automated collection of data, without meticulous review it is impossible to clearly know what kind of actor does what action. Because of this, we use a sample that has been carefully classified and reveals discrete classification of users. This allows us to achieve a highly accurate view of the editing actions and actors. We analyzed 650 questions from 2013 to 2015 that received at least 5 down votes from the community and were closed by voters. The reason for this cut-off is to separate a group of questions that are *clearly* unfit. At least 5 qualified voters had to agree to vote the question down. In addition, 5 qualified community members had to agree to close the question and mark it as being in violation of the community's question-asking policies. While it is true that the questions with fewer down votes could also be considered unfit, the cutoff creates a data set in which we can assume that most actors on the question understand the quality of the question.

We begin this chapter with a look at the background of Stack Overflow's editing system. Then, we analyze the initial answers that are found on unfit questions and analyze them for quality and completeness in order to understand the workflow of answering unfit questions. The results are that answers on these questions are of very high quality and would allow the questioner to resolve their question. Next, we explain our categorization system of users and then analyze the actions of different types of participants (Questioners, Answerers, Contributors, and Low-reputation contributors) who have varying motivations and repercussions when they contribute. For instance, while low-reputation users can earn points from editing, and questioners can save reputation points by improving their question, high-reputation users cannot earn points by editing. The results from this study show that most edits come from questioners and contributors, while answerers and low-reputation participants do not typically provide edits. This shows that editing largely depends on users with intrinsic rather than competitive motives. Next, we consider the editing-history of high-reputation contributors, and measure whether they are incentivized by earning badges related to editing. The results indicate that they are not incentivized and are acting instead because of intrinsic motivations. This would suggest that these users play the most important role in the social function of editing questions and improving them for the archive.

4.2 Collaborative Editing in Peer Production Communities

Perhaps the best known peer production community that uses collaborative editing to maintain quality is Wikipedia (Kittur and Kraut 2008). Wikipedia depends on a large community to both produce and edit encyclopedia articles for open use. This collaborative editing scheme has been dependent on a few number of power editors making most of the edits and making the biggest impact on article quality (Ortega, Gonzalez-Barahona, and Robles 2008; Panciera, Halfaker, and Terveen 2009). While there were arguments that the amount of editors were

expanding outside the power editors (Kittur and Kraut 2008), Halfaker (Halfaker et al. 2013) found that over an expanded time, Wikipedia's editing culture had hardened and had become resistant to new editors.

Editing articles in Wikipedia can be a contentious and difficult process due to the subjectivity of the content (Liu and Ram 2011). Because of this, development in limiting edits in controversial or critical articles was explored in detail (Adler and de Alfaro 2007) in order to maintain quality. Liu (Liu and Ram 2011) also explored collaboration patterns in editing, finding that encouraging article authors to edit their own articles was a key way to increase article quality.

The issue of why editors edit has often been asked about Wikipedia, the answer is not easy to understand. While it would seem that altruism plays a large role in unpaid editing service, Wikipedia developed a strict hierarchy of users (Halfaker et al. 2013; Zhu et al. 2013). This hierarchy controls and greatly restricts the access of new and inexperienced users (Halfaker et al. 2013; Zhu et al. 2013). In a study of the development of successful and unsuccessful Wikipedia projects, those that had many different contributors at the start of their lifespan were much more likely to be successful, but even these projects would devolve into the few power users (Solomon and Wash 2014). Editing in Wikipedia starts as a large communal project, but eventually evolves into a small committed group.

The value of community editing in social Q&A has also been recognized. In SO, it is known that the editing of questions and answers does improve the point earning potential of the post as well as its archival value, making editing a socially desirable action (G. Li et al. 2015). In addition, editing is a signal that a question will not be deleted and does have some value or merit (Correa and Sureka 2013, 2014). However, unlike Wikipedia, the competitive nature of the

reputation system makes it unclear whether the underlying nature of the editing is competitive or intrinsic.

4.3. Unfit Questions

Stack Overflow (SO) has a set of rules for asking a question. SO was created with the idea that a large amount of questions, or especially poor questions, would lead to an unsustainable eco-system (Ford 2012). Therefore, all potential users on SO have to agree to a set of rules. While these rules evolve, for 2013 and 2014, these rules were as follows: 1) Search and research their answer on the site, internet, and other resources before asking, 2) Ask questions relevant to the site, 3) Be specific instead of vague, and 4) Make sure the question is relevant to others and not localized.

4.3.1 Reasons for Closing a Question

Questions that violate SO's specific rule set can be "closed" by users with over 3,000 points or a moderator. While questions can be reopened if they receive enough support through editing, this is rare (Correa and Sureka 2013). Users who decide to vote to close a question have several options to do so. In 2013, the following categories were prescribed by SO: Not a real question; Not constructive; Off topic; Primarily opinion-based; Too broad; Too localized; Duplicate; and Unclear what you're asking. Prior to 2012, SO had only four tags (Off-topic, Too subjective, Not a real question, and Too localized). The "Too subjective" tag was essentially replaced by the new tags, which give voters more specific reasons for closing a question.

SO does go through an iterative process for determining why a question can be closed. In spring of 2014, the tags were narrowed to: Off topic, Unclear what you're asking, Too broad, and Primarily opinion-based. In general, the most popular reasons for closing a question were Not a real question and Duplicate (each at around 30%).

4.3.2 Unfit Questions that are Closed VS. Deleted Questions

If a question is deemed to have no archival value, it can be deleted from the corpus. This happens often to questions that have no answer to them or have nothing to do with the computer programming (Correa and Sureka 2014). On the other hand, a question could possibly be reopened if it is edited to fix its flaws and the community recognizes the improvements. Questions that remain perpetually closed but not deleted account for around 3% of the SO Q&A corpus (Correa and Sureka 2013). Questions that are deleted are much lower in quality to their closed counterparts in content. Correa (Correa and Sureka 2013) also found that deleted questions had less code and were questions that attracted very little interest from the community, receiving fewer views and votes than closed questions. While closed questions are low in quality and unfit for the system, there is tangible difference in their construction that signals some potential value for the community members or archival value for the corpus.

4.4 When Unfit Questions have Answers

When unfit questions receive an answer, are they good? In the case of good questions, measuring votes or best answer selection is a good method for determining if a question received a good answer. However, with unfit questions, simply measuring votes or best answer selection is not a good metric, since activity is frozen on closed questions. In order to answer this, two coders, the author of this thesis and a coder with over twenty years of computer programming and development experience at a professional level, analyzed the first answers left on a random sample of 384 question and answer exchanges. To better understand original intent of the answerer, the coders looked at the first iteration of the question and the first iteration of the first answer. They did not include edits unless those came after the first answer. This helps to mitigate the possibility that answerers were acting because of an edit by another user.

The coders rated the first answer on how completely it answered the question based on a five-point scale, with 0 being no answer at all, 2 being an adequate answer, and 4 being a complete answer. The coders rated all 384 answers and had an ICC2 score of 0.9 indicating high agreement. The completeness of the answers is very high (M=3.41, SD=0.733), indicating that answerers do put significant effort into answering. The quality of the answers also means that value for the community is present and lends understanding of why these questions are not deleted.

This finding is also very important for two reasons: First, it allows us to assume that the first answer is likely to be high quality, meaning that it can be used as a benchmark for comparing the order of editing and answering. Edits before the first answer may be trying to improve the question to encourage answers, while edits after the first answer are done to improve the reputation potential of the question or improve the long-term archival value. The other reason is that it allows us to assume that answerers are contributing high-levels of effort. These answers are not "joke" answers, but rather serious contributions. This informs the assumptions behind the motivation of the answerers. In accordance with the research in Chapter 2, we can assume that answerers are seeking to help the questioner and receive reputation points.

4.5 Editing

4.5.1 Editing on Stack Overflow

Editing a post on SO is a privilege with access that differs depending on the reputation of the editor. The author of the post can edit their contribution regardless of their reputation (an exception could be made if the author has made a series of bad edits) ("How Does Editing Work? - Meta Stack Exchange" 2016). For users with less than 2,000 points, editing other users' posts does not take immediate effect. These edits need to be accepted by users with higher reputation. However,

+2 points can be earned for each accepted edit with a cap of 1,000 points from edits (“How Does Editing Work? - Meta Stack Exchange” 2016). Users who have reached 2,000 reputation points are given the privilege of editing questions immediately with no review. As such, these users cannot earn reputation from their edits. All edits are kept in the history next to the original post, so the evolution of the question can be viewed and reverted if necessary (“Why Can People Edit My Posts? How Does Editing Work? - Help Center - Stack Overflow” 2016).

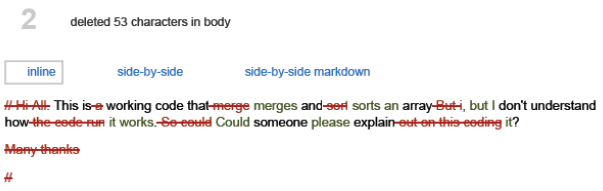

Figure 4.1. Example of an Edit on Stack Overflow

SO encourages users to create edits whenever it will improve a post. In particular, their editing guide includes a number of reasons edits are particularly made (“Why Can People Edit My Posts? How Does Editing Work? - Help Center - Stack Overflow” 2016):

- *To fix grammar and spelling mistakes*
- *To clarify the meaning of the post (without changing that meaning)*
- *To include additional information only found in comments, so all of the information relevant to the post is contained in one place*
- *To correct minor mistakes or add updates as the post ages*
- *To add related resources or hyperlinks*

An example in Figure 4.1 shows an edit to an unfit question on SO. In this case, the editor has corrected numerous grammatical errors in order to make the question easier to read. There are six types of edits that can be performed on a question post:

- **Edit Body** – The editor makes changes to the body of the question. This includes making grammatical changes or changes to the presented code.
- **Edit Title** – The editor makes a change in the title of the question.
- **Edit Tags** – The editor changes the tags for the question. Tags provide metadata for the archive.
- **Rollback Body; Rollback Title; Rollback Tags** – In these cases, the editor is undoing a previous edit.

Li (G. Li et al. 2015) found that, in general, Edit Body actions have the greatest impact on post quality, while Edit Tags have the least impact. Edit Body actions are also the most common type or action found across all types of posts and we would expect this to be the same in unfit questions.

4.5.2 Types of Editors

Why would a user choose to edit an unfit question? One reason might be for *competitive* motives. Competition in Q&A is well documented. Anderson (Anderson et al. 2012) found that users will target reputation point aggregation, and that the better the user, the better the ability to find areas where points can be earned. Tausczik (Tausczik and Pennebaker 2012) found that users claimed that reputation points did not matter as a motivation for participation, but that, in reality, these users targeted reputation earning.

This does not discount possibility of *intrinsic* motivations like altruism, a sense of duty to the efficacy to the archive, or a bond to the community. In this paper, we consider that intrinsic

motivation is defined when a user makes a socially desirable action with no reputation benefit. The rule of intrinsic motivations in Q&A is difficult to define, but Mamykina (Mamykina et al. 2011) did find that intrinsic motivations like altruism were important in motivating community action in SO. We also consider that attachment to the community is a primary motivator towards expending effort on improving another users' content.

We identified four types of editors with different motivations:

- **Questioner** – These users asked the unfit question. They have been down voted. Editing the question is a potential way to recoup reputation points, and seems likely based on Correa's work (Correa and Sureka 2014). In addition, editing the question may be a way for a user to encourage better answers. We consider the motive for editing to be **competitive.**
- **Answerer** – These users answered the unfit question. In this case, it is beneficial for the question not to be deleted, since all gained reputation points are lost if the question is deleted. We assume that the answerer is interested in earning reputation points (Lappas, Dellarocas, and Derakhshani 2017) and in helping the community (Tausczik and Pennebaker 2012). Making the question a better fit for the community likely makes question more popular in the community, increasing the number of votes the answer will receive. We consider the motive for editing to be **competitive**.
- **Contributor** – These users have more than 2,000 points and have edited the question without posting an answer. They earn no reputation from editing. We consider the motive for editing to be **intrinsic**.
- **Low-reputation Contributor** – These users have fewer than 2,000 points and can earn reputation points by making successful edits. Since this strategy can be a viable one for

earning reputation (Furtado et al. 2013), we consider the motive for editing to be **competitive**.

4.6 Edits on Unfit Questions

We took a random sample of 650 closed questions from a repository of 15,000 eligible questions (99% Confidence, 5% Margin of Error). We chose questions that received at least 5 down votes and had at least one answer and one edit. Using a self-created data query from the public data repository at data.stackexchange.com, we were able to isolate edit actions on target questions. In total there were 1531 edits on the 650 questions. We analyzed user type and actions with these edits.

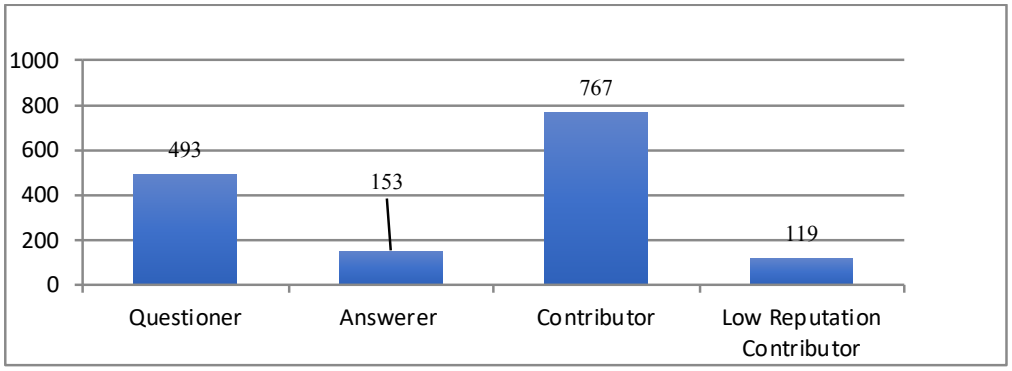

Figure 4.2. Total Edits by User Type

Who contributes edits? A quick view of the edit counts shows that contributors give half of the edits, as Figure 4.2 shows. In the data set, 767 edits came from contributors, while 765 came from the other three combined. The next biggest group was the questioners, which accounted for 32% of the edits. Surprisingly, answerers and low reputation contributors contributed small number of edits. The immediate take-away is that users without a direct incentive to edit are doing so at a much greater rate than users with vested interest in improving the question.

Who edits when and how? An answer or an edit can happen at any point once a question is posted. As Figure 4.3 shows, there is an even split between answers coming first and edits coming first. At each stage, the balance between answers and edits is evenly split until the very late stages where editing dominates. Within the editing section, we see that Edit Body actions are the most common at each stage. This result is similar to Li's (G. Li et al. 2015) distribution findings where Edit Body actions are the most common type of edit. This result indicates that these contributors are attempting to make the most impact on the quality of the question. It should be noted that there is no significant difference between editor class and edit type.

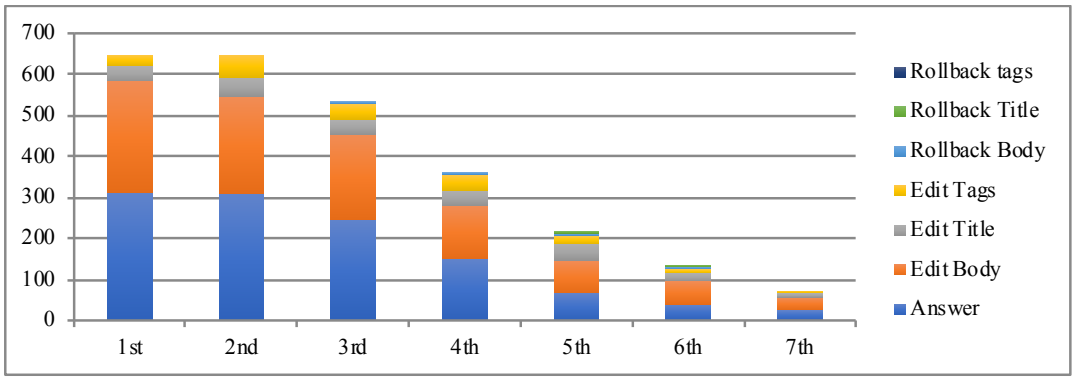

Figure 4.3. Order of Actions on an Unfit Question

The only group that contributes more than half of their edits before the first edit is the contributor group as shown in Figure 4.4. A contingency table gives a chi-squared value of 85.4 ($p<0.01$), with post-hoc Bonferroni corrected comparison validating that only contributors are skewed towards contribution before the first answer.

This shows that the contributor group identifies value in an unfit question faster than other groups. On the other hand, questioners and answerers tend to make edits after the first answer has been posted. This would seem to concur with Correa's work (Correa and Sureka

2013, 2014) that asserts that questioners are trying to edit the question in order to save reputation points. Answerers are possibly trying to assure that reputation points are not lost due to deletion.

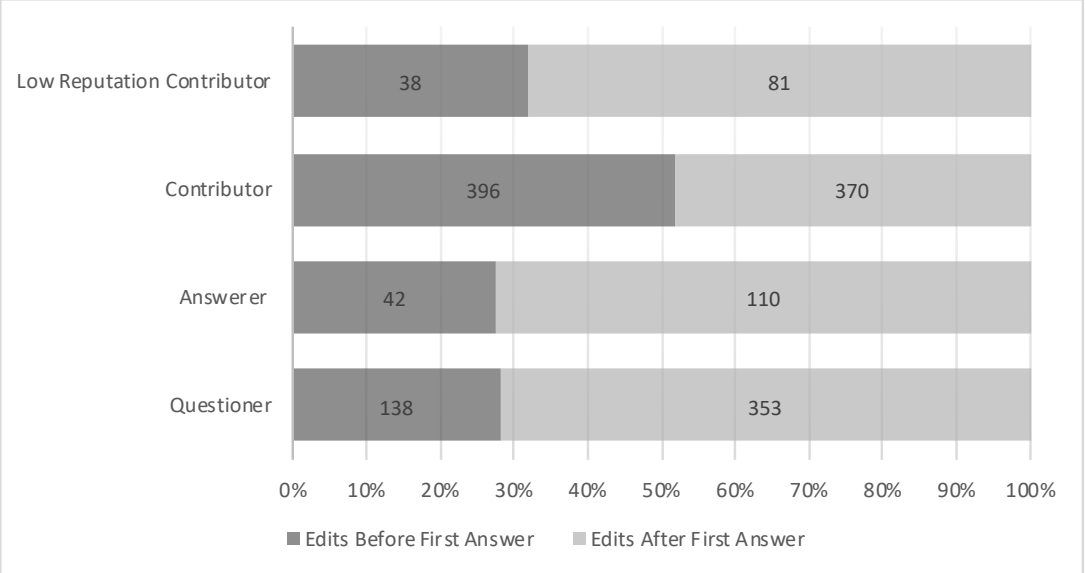

Figure 4.4 Edits Before and After the First Answer

The other notable thing is that answerers contribute so few edits. While there were 1198 answers on the 650 questions, only 60 answerers contributed edits. This is surprising given the quality of the answers they give, and potential risk that the question might be deleted. Low reputation contributors also gave few contributions to the questions, but this may be a function of the difficulty of obtaining acceptance on an edit.

Finally, questioners were the second most likely to edit the question. Questioners mostly edit once they have received an answer. This indicates that the editing is done not to induce answers, but more likely to save reputation (Correa and Sureka 2013, 2014). The fact that

questioners edit a lot is encouraging according to Liu (Liu and Ram 2011) since this could help improve the authors' future posts.

The most interesting conclusion is that the contributors by far provide the most edits before the question becomes answered. These users are able to identify value within a question before an answer occurs. This also explains why Correa found more instances of editing in retained closed questions compared to deleted questions. It is not only answers that signal archival value, but that high reputation users often identify this value prior to an answer.

4.7 Analyzing the Contributor Class

It is interesting to see how dominant contributor class users are in the editing of these questions. Despite having reputation incentives to act, these users attempt to make the question better, potentially improving both the outcome for the community and the actors who are benefiting from the question. This would mean that, like on Wikipedia (Halfaker et al. 2013), there is a core of elite users that edit and maintain the system. This idea is further solidified by analyzing the edits given by the contributor class and the number of days that passes between edits.

Badge	**Edits Required**	**Number of Users with Badge**	**% of Users with Badge**
Editor	1	2.0M	22.4%
Strunk & White	80	14.8K	0.0016%
Copy Editor	500	2.8K	0.00031%

Table 4.1. A List of Edit Related Badges and Users with Badge

Of course, a possibility is that there is an incentive to contribute that has not been considered as of yet. Users can earn badges for editing 80 and 500 posts (These badges are called

"Strunk and White" and "Copy Editor" respectively. Badges are awards for doing certain activities. Badges have shown to increase socially desirable behavior (Cavusoglu, Li, and Huang 2015; Z. Li, Huang, and Cavusoglu 2012), and SO itself saw an increase in the number of edits after implementing the badges (Z. Li, Huang, and Cavusoglu 2012). However, at the same time, very few users own theses badges as shown in Table 4.1.

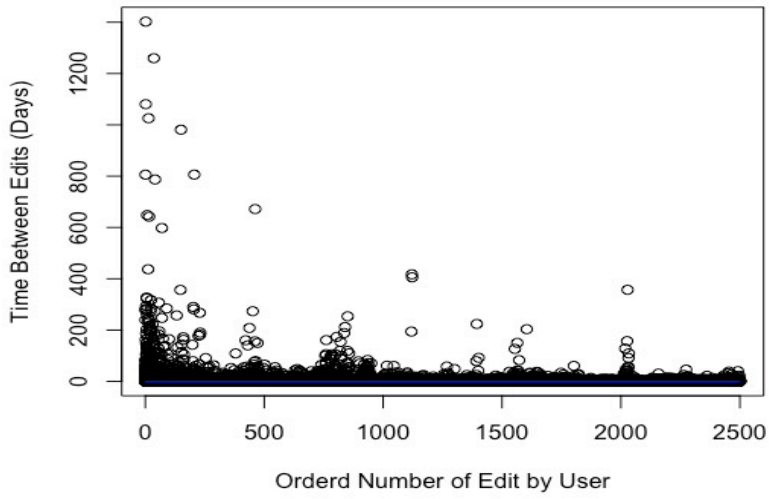

Figure 4.5. Time Distribution Between Edits and Days Passed for High-Reputation Users

A limitation to the previous research on badges and gamification, is that they often consider whether the incentive causes a spike in contributions system-wide. If the badge was truly a motivating factor in editing, we should expect to see a slow-down in frequency of edits produced once a user passes 500 edits.

We took a sample 59,854 edits from 39 users in our contributor class data set. We capped the number of edits produced to 2,500. Edits were ordered by the number of edit that was contributed by a particular user (i.e. all edits are given in order and are therefore assigned a

unique ordered number). A visualization, as shown in Figure 4.5 shows that editing frequency is mostly flat.

We ran a mixed effects linear model, where the user is a random effect. The dependent variable was the Time between edits and the fixed effect was the ordered number of the edit. The results, as shown in Table 4.2, show that the increased number of edits seems to moderately decrease the time between edits. Order of edit affected time between edits ($\chi 2(1) = 4.1757$, $p = 0.04$), decreasing it by about -0.3416 days ± 0.1674 (standard errors) per edit added.

	Time Between Edits	**Estimate**	**Std. Error**
Order of Comment	$\chi 2(1) = 4.1757$, $p = 0.04$*	-0.3416	0.1674

Table 4.2 Results of Mixed Effects Model (* significance at $p<0.05$).

The results lead us to conclude that contributors are producing edits with no ulterior motivation, and that they do represent a core of elite users who manage the site. This affirms the notion that contributors are editing due to intrinsic factors.

4.8 Conclusion and Discussion

The results show that intrinsically motivated editing plays an important role in the collaborative editing of unfit questions. Even when there are numerous users who have intrinsic incentives to improve the question, half of the edits come from users with no direct incentive. Editing cannot be completely powered by a competitive system. Competitive editing often happens once the question has received an answer. This would signal that intrinsic editing is also more likely to create value for the question asker and the community. We suggest that system designers facilitate community editing instead of relying solely on users with competitive reasons to edit.

This could be interpreted as a disappointing result. If competitive editing could completely facilitate editing, this would mean that system designers could afford to not worry about the implications of an editing class largely influencing the editing of material. For one, it limits the creation of closed exchanges between direct contributors. This is unfortunate when considering there might be situations in which designers would prefer to keep the exchanges limited. In addition, there is also the problem of community atrophy, as shown in Wikepedia (Halfaker et al. 2013). Another important implication to this is that within the Collaborative Model aspect of the system, high-reputation users are simply more adept at identifying and fixing these issues. An interesting avenue for future research would be to investigate the skill of editing by low reputation users in comparison to higher reputation users. While there might be a demonstrable difference in quality, there also might be more of an influence of ingrained community standards.

The results show that high reputation users choose "good" unfit questions before answers are posted. This finding is useful when considering expediting the deletion of truly unfit questions that have no archival value. These results suggest that capable users quickly recognize and attempt to fix questions with merits; while previous research indicates that they ignore those without merit. Finally, we found that the community gives high quality answers to unfit questions. This means that edits from high reputation users could be used to automatically allow unfit questions to stay open and prompt answers.

There are some important limitations to this study. First, this looks at one type of site in a specific setting, specialized Q&A. In addition, we only looked at a subset of questions. In future work, we plan to expand this study to other domains and question types, in order to broaden the understanding of the relationship between motivations and community editing.

4.9 Summary

In this chapter, we looked at the Collaborative Model aspect of SO and analyzed the relationship between reputation class, motive, and editing. We found that overwhelmingly, users, with high-reputation and no obvious reputation incentives, were much more likely to contribute all edits including the most valuable edits. This means that the Collaborate Model is dependent on high reputation users who choose to frequently edit with no clear reputation incentive for doing so.

This is contrasted by SO offering clear reputation point incentives to other vested parties. The implication of this chapter is that reputation incentives are limited in influencing socially desirable behavior in a Collaborative Model aspect. This does not mean that this activity does not occur, but rather, it is small compared to the intrinsic motivations of those in high reputation classes.

In the next chapter we look at a Community-Based aspect of Q&A by examining how the community comments and interacts on undesirable questions.

Chapter 5.

Reputation Classes and Communication via Commenting

5.1 Introduction

In Chapter 4, we considered the editing of unfit questions in a Collaborative scenario. In Chapter 5, we look at reputation and the type of feedback that is left for the questioner in the Community-Based aspect of Q&A. Specifically, we aim to understand the type of feedback that is left and if reputation class can predict more socially desirable feedback.

Bad questions are a problem that exists for successful sites in that the number of poor questions will start to hinder the efficacy of the archive, and an effective way of handling this problem is to let the community moderate the content itself (Correa and Sureka 2014). However, as effective as the process is at removing unwanted questions, one concern facing such systems is the quality of feedback dialogue that is given to the contributors of the question. In this chapter we analyze how a community comments on unwanted material, and how the community judges different types of feedback through a voting mechanism.

Dialogue-based feedback on contributions has been often studied in different peer-production communities. Empirical studies have shown that the type of feedback that users receive in peer-production systems can affect their following contributions. Negative feedback can decrease a user's future contributions (Zhu, Kraut, and Kittur 2012), while positive feedback can encourage new users to contribute more (B. Choi et al. 2010). In social media systems, negative feedback can create a loop in which the recipients of negative feedback produce even more poorly rated material and become negative themselves (Cheng, Danescu-Niculescu-Mizil, and Leskovec 2014).

A limitation of these studies is the focus on correlative value and impact on the receiver of the comment (Zhu et al. 2013), rather than a focus on the creators of the feedback. It is unclear

whether contributors of critical feedback are trolls, low reputation users, or the entire community. In addition, studies that draw correlations based off data-mining (Ahn et al. 2013; Zhu et al. 2011) are limited by a lack of context and nuance. It is difficult to rely on machine mining of negative comments, as kappa agreement with human evaluators is low to moderate (Zhu et al. 2011). This signifies that research can be strengthened by qualitative methods.

As mentioned in Chapter 2, SO is governed by a reputation and privilege system that is based on the idea that reputation points are synonymous with trust. SO users with enough reputation points are given moderating powers. Previous work has indicated that volunteer administrators on Wikipedia will use more positive emotions in discussion pages (Laniado et al. 2012), so there is potential for experienced community members to guide inexperienced or poor questioners to fix their errors and improve future contributions through constructive dialogue and feedback. In the same way, unconstructive criticism could potentially come from lower reputation users.

We extend the understanding of feedback and its relationship with a community by performing a qualitative review. We developed a basic taxonomy through a Grounded Theory process (Glaser and Strauss 2009), with special focus on the context of the comment with the question. This process allows the research to consider the question and context of the feedback. We then investigated whether the community preferred certain types of comments by using votes given to comments, and whether different types of users, especially users with high reputation, contribute differently from each other. Lessons from this study are useful for better understanding community orientation towards feedback in Q&A and peer-production communities, as well as informing rules and incentives for these systems.

5.2 Feedback and Learning in Peer Production Communities

Wikipedia is a standard-bearer for large-scale peer-production systems, and relies on feedback in order to create high quality articles (Laniado et al. 2012). A large number of peripheral editors play a significant role in providing leadership and feedback (Zhu et al. 2011). In order to achieve a continuous critical mass of content, educating newcomers to the community is essential (Schneider, Passant, and Decker 2012). Two studies (Halfaker, Kittur, and Riedl 2011; Zhu, Kraut, and Kittur 2012) found that negative feedback, through reversions and commenting, harms new participants and fails to improve future outcomes, as opposed to positive feedback. Zhu (Zhu et al. 2013) did find, however, that in lab controlled settings, mild negative feedback did have a corrective influence on participants.

Negative feedback can also mean more poor contributions in social computing situations. Cheng (Cheng, Danescu-Niculescu-Mizil, and Leskovec 2014) reviewed data from online news communities, and found that negative feedback actually propagated more negatively received contributions, while positive feedback did not evoke a similar response from users. On the other hand, Wohn (Wohn 2015) found in the social media site Everything2 that negative feedback could have a positive effect on established users who had developed a habit and perhaps should be used for experienced users only. These studies on Wikipedia and the other online communities do indicate that negative feedback does not necessarily have a deterministic outcome. There are some situations in which negative feedback could be useful for a community's aims. In order to understand outcomes beyond correlation, it is important to understand how the feedback is actually formulated and propagated by the community.

There has also been some research regarding feedback on SO and its network of Q&A sites, Stack Exchange. Tausczik (Tausczik, Kittur, and Kraut 2014) found that commenting could be used to help solve complex problems on the Stack Exchange site Math Overflow, which

indicates that commenting has a tangible use in this type of system. Yang (Jie Yang et al. 2014) explored the theoretical possibilities that commenting could be used as part of editing to moderate poorly formulated questions. However, Ahn (Ahn et al. 2013) found that the volume of comments had no positive effect on the future participation of a questioner. The authors concluded that content and not volume may play the more pivotal role in influencing future behavior. Li (G. Li et al. 2015) looked at constructive commenting as a function of collaborative editing, and found that commenting did have a small negative effect on the future participation of the questioner or answerer. Since the role of feedback and its effect on users is relatively clear, in this paper, we do not focus on the correlative nature of feedback on future outcomes, but rather, on the contributions and interactions themselves as part of its role in the community.

5.2.1 The Process of Commenting on Bad Questions on Stack Overflow

SO has a clear guideline for the commenting privilege, most importantly that comments should ask for clarification or give constructive feedback that helps the author improve the post. On the other hand, criticism which is not constructive or answering a question is not allowed.

Figure 5.1 shows an example of the commenting process on SO. At 10:50, five minutes after the question was posted, User A comments on the simplicity of the question and motivation of the questioner by writing, “people are so lazy to make some googlefu”. Later, at 11:02, User C comments on the question, but in a completely new direction, instead explaining to the questioner why the question was poorly received. The questioner responds to User C with thanks for the guidance.

The votes on the comments do not count towards reputation, but they do act as a form of social approval since the community clearly states its preferences (Cheshire 2007). While there has been a discussion to make these comments count towards reputation, this has been rejected

by the community. However, users can earn a participation badge for comments receiving more than 5 votes, and the voting structure does indicate the community's preferences. In this case, we can see that User A's comment was the most popular with 11 votes, while User C's comment was second with 5 votes.

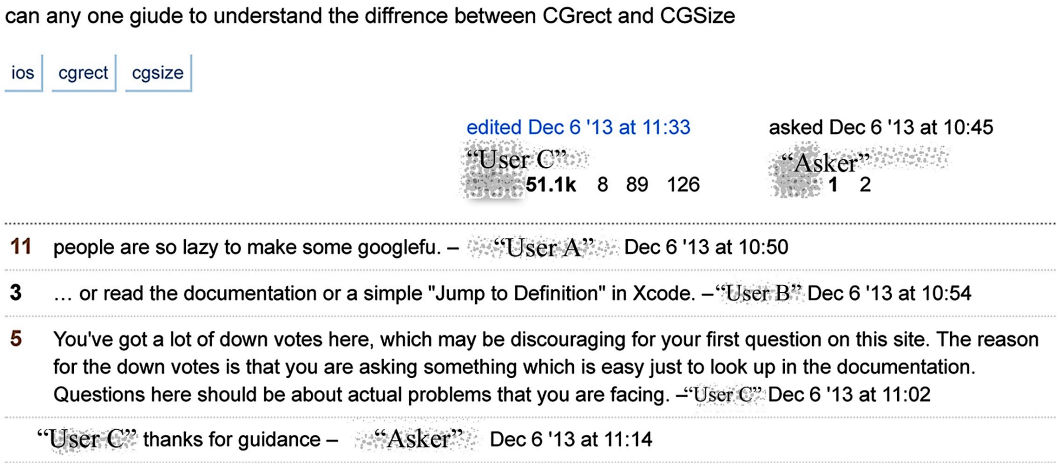

Figure 5.1. Question 20421944: An Example of a Bad Question and Comments (Anonymized).

The example in Figure 5.1 shows an interesting dynamic. The initial comment contributes a harsh criticism of the questioners' abilities and motivations. The third comment contributes guidance. This example shows how the contribution of commenters can vary greatly from user to user, and informs this research: should we expect commenters to contribute constructive criticism and guidance, or is the ratio more skewed toward criticism without guidance?

The prevalence of unfriendliness was concerning enough for SO's administration that the CEO and co-founder, Joel Spolsky, launched a campaign called the, "kicking off the summer of love" in 2012 (Spolsky 2012). (Spolsky 2012) A key passage from the blog post stated:

> Newbies will show up, make a newbie mistake, like wearing shoes indoors or forgetting to close the toilet lid, and the old-timers will look at each other, roll their eyes, and snort,

> "Typical!" At this point, if it's a normal human community, it will start to feel a little bit unfriendly to outsiders. Insular. And the newbies will say, "well, gosh, that's not a very friendly place." Not just the newbies who got scolded. Also the 100 passers-by who *saw* the newbies get scolded. Who might have been great members of the community, and who did nothing wrong, but who are not really interested in joining a community that appears to be full of smug jerks. This is very dangerous. You have to be able to recruit new members to replace the old ones that drift away. The success of the community depends on it.

From this point, the community leaders made concerted efforts to educate community members on what constitutes acceptable feedback. Specifically, tone and content were singled out, asking users to be civil and informative.

5.3 Data Set

Since the goal of this study is to look closely at interaction that occurs on *clearly* bad questions on Stack Overflow (SO), we created a number of rules to collect a sample that would allow for a close investigation: 1) Questions must have more than five down votes: The rationale behind this rule is to make sure that a larger section of the community agrees with the closers that the question is not suitable for SO, 2) Questions must be asked by a unique user in order to be included in the data set. There are no repeated questioners, and 3) Questions must have at least one comment that is made by someone other than the author.

From these rules, we were able to identify approximately 20,000 eligible questions. We took a random sample of N=581 (5% margin of error, 98% confidence). Some questioners were completely new (N=241), while other had asked before (N=340). The mean of previous

questions was 18.4 (SD=97.9, Mdn=1). The mean for the number of comments was 4.5 (SD=3.5, Mdn=4).

Reputation	Number of Commenters	% of Comments	Unique Commenters
User	108	18.6%	107
Established User	284	48.9%	254
Trusted User	182	31.3%	121

Table 5.1. Number of Commenters by Reputation Class

The commenters came from various backgrounds as well. We used the classification scheme developed by SO to classify the users. SO defines users on the following metrics: User – 0 to 999 reputation points, Established User - 1,000 to 19,999 reputation points, and Trusted User – 20,000+ reputation points. There were seven commenters whose reputation could not be identified and are not included in Table 5.1. As mentioned previously in 2.2 in chapter 2, reputation points earn privileges. Trusted Users have all privileges, including significant editing and voting powers. As we mentioned earlier, reputation in SO is supposed to be a rough measure of trust a user has in the community, and the system gives privileges based on this trust. From a system point of view, Trusted Users would be expected to closely follow the commenting rules.

An important part of understanding the interaction on these questions is having a clear view of what these questions are trying to accomplish. We used a Grounded Theory approach to distill comprehensible categories. We started with a simple question in order to orient the study, "What is the questioner trying to achieve by asking the question?" Instead of trying to understand why the question was good or answerable, we focused on the end goal of the

questioner. The open-coding session led to the identification of a small number of strong themes that quickly coalesced into four distinct categories, which align with previous research categorizations (Treude, Barzilay, and Storey 2011):

- **Process:** These questions ask for help in completing a process. These questions are often when a problem has been setup with code, and the questioner does not know the next step in order to achieve an objective. (152 questions)

- **Process without code:** These questions are essentially the same as Process questions but lack any code or fail to show any attempted work. (163 questions)

- **Debugging**: The questioner has a problem or error in code. (186 questions)

- **General Concept**: These questions deal with choosing a tool or theoretical questions such as, "what language is best to learn for a future career?" (80 questions)

5.4 How does the Community Comment on Bad Questions?

5.4.1 Methodology

A grounded theory approach (Glaser and Strauss 2009) was used to inductively categorize the type of contributions that questioners receive from commenters. The entire question set was studied in detail by the authors in order to understand the content given by the commenters. In this way, the coding was conducted from the viewpoint of a questioner. There were two types of coding: open coding, where the categories and sub-categories were discussed and mediated, and individual coding, where the same coders independently assigned values to denote strength of a category within a comment. During the open-coding session, numerous labels and sub-categories were identified. These sub-categories were then used to guide the formation of more abstract

categories. In total, we were able to distill three distinct categories of comments.

- **Corrective:** These comments explained procedural elements of why the question is bad or gave explicit help on how to improve the question. For instance: providing reference to Stack Overflow (SO) questioning procedures; giving direct guidance to asking a better question; explaining the reaction of the community.

- **Critical:** These comments point out flaws in the question and are unconstructive. For instance: giving accusations of a lack of effort; of cheating on homework; of a lack of ability; of trying to get work done for free; or mocking user's ability.

- **Answer:** These comments addressed the question. For instance: giving a direct link to answer; giving correction of error in a question; giving a full answer.
  The interactions present in the comments showed a wide range of variability.

Corrective comments often focused on the specific process that needs to be followed. For instance, in Table 5.2, the last example comment refers to the FAQ that the questioner has already agreed to. In this way, we found that corrective commenting often gave a pedagogical guide to the questioner. However, correction also occurred in other ways, when errors in the formatting or general guidance were given along with criticism.

Criticism was often given in a way that either directly accused the user of a misdeed or mocked their question. The first comment gives an example of a direct accusation of being lazy and violating the spirit of the site. Comments two and four, on the other hand, give less direct but strong criticism regarding the questioners' skill and understanding of the content. To a certain

extent, these comments would seem completely in violation of the commenting rules as they fail to be constructive at all.

Question Type	Comment	Type of comment
Process without code	You've shown no effort, are basically asking for someone to do the work for you.. and you've tagged this with tags that don't relate to SQL..	Critical: Accusation of lack of effort; Lack of skill Corrective: Gives correction on content
Process	"I have a class which represents database record with over 100 string fields" - OMG you have a much greater problem than string initialization.	Critical: Lack of Skill
Debugging	Because it prints, then increments. ++index will increment, then print. Postfix versus prefix.	Answer: Comment explains the error in a short comment
General Concept	Which century are you living in?	Critical: Lack of skill;
Process	Welcome to Stack Overflow. Please read the FAQ on how to ask questions here - you'll need to describe what you've tried so far and what didn't work, preferably with code samples. At that point we're more than happy to help.	Corrective: Gives directions on how to ask

Table 5.2. Examples of Comments and Codes

5.5 Coding Categorization of Comments

In order to better understand how an entire comment is read in the context of the question and to understand the distribution of comments, we employed a taxonomic coding scheme based on the open coding results to categorize the comments as received by the questioner. To initiate coding, the independent coders participated in open coding of the contributions. Disagreements on definitions were resolved through discussion. The coders rated a selection of the top scoring comments (N=581) from the data set. We chose this method because this study focuses not of the popularity of commenting on certain types of bad questions, but on the popularity of commenting types. Rating the top comment as opposed to all comments removes homophily that can be present in discussions and gives a clearer picture of what kind of comments are popular. When there was a tie (N=110) the earlier comment was selected for coding. Three coders, including the first author, participated in the coding. The first coder has a Master in Information

Science and the second is a Master student in Media Studies. While all of the coders have familiarity with computer programming topics, none of them has been or are active users of SO.

Comment Type	**Presence in Comments with a Score of ≥1**	**Comment Mean and Standard Deviation**	**Comments with an Isolated Score of ≥1**	**Inter-rater Agreement (ICC)**
Corrective	47.1%	M=0.92 SD=0.98	13.23%	0.78
Critical	66.1%	M=1.57 SD=1.52	35.1%	0.72
Answer	28.1%	M=0.62 SD=0.96	16.0%	0.84

Table 5.3. Comments by Classification

The coders were given the comment with the original question. In addition, the coders were allowed to follow the off-site links to view resources that were mentioned in the comments. The reputation of the commenters or questioners was not available. The coders were asked to rate the comments on the following five-point Likert scale (from extensively (4) to not at all (0)) for each of the three categories. The coders considered each contribution on these evaluation points. A comment could have the elements of multiple categories. An important concern before the coding session was that of tone and implied content in the comments could make it difficult

to obtain high agreement. However, they were not guided to ignore tone or implied content due to the design of the test.

We calculated the inter-rater agreement as a Cohen's kappa using Shrout and Fleiss ICC2 as shown in Table 5.3, and the results show that there was strong agreement. Comments tended to have strength in more than one type of category, explaining why total presence of comments exceeds 100%.

We also looked at isolated scores. Isolated scoring highlights a comment that may have more than one type of attribute but is stronger in one particular area. This is done through subtracting aggregate scores of each category. If a comment received aggregate scores of 1 for corrective, 1 for critical, and 3 for answer, it would be classified as an answer comment with an isolated score of 1. Corrective comments were the least likely to be isolated, while critical comments were the most likely.

	No. of Comments	**Score ≥ 1**	**No. of Comments**	**Isolated Score ≥ 1**
Corrective	274	3.0073	77	2.8701
Critical	384	3.6354	204	4.1029
Answer	163	2.9202	93	2.9570
F-Value		6.2126**		7.4783**
Tukey A Vs B		4.1744**		4.4114**
Tukey A Vs C		0.4625		0.2698
Tukey B Vs C		4.0207*		4.3835**

Table 5.4 ANOVA: Comments by Type and Votes Received (*p<0.05, **p<0.01)

Votes on Comments: Next, we looked at the number of votes received on different types of comments. First, we looked at all comments that received an aggregate score of ≥1 in a particular category. An ANOVA with post-hoc Tukey as shown in Table 5.4 showed that critical comments received higher scores than both corrective and answer comments. The isolated critical comments are even more popular as they gain about 4.1 votes on average. This would indicate that the community has a strong preference for criticism. An important thing to note is that this

analysis only looks at comments that received the most votes (or were tied for such status). It does not take into account comments that follow in the ratings.

Relationship with Commenter Reputation Class: The next thing we wanted to look at was the type of comments left by users from different reputation classes. An assumption could be made that Trusted and Established users would be more likely to follow the rules since they have been through the privilege process and may be more sensitive to the actions that are deemed socially desirable by the SO administration. As Table 5.5 shows, there was no significant difference between any of the classes and the type of comments left. Critical comments were almost identical between all classes.

	Corrective	**Critical**	**Answer**
User	0.8056	1.5123	0.6235
Established User	0.9437	1.5716	0.5892
Trusted User	0.9652	1.5513	0.6923
F-Value	1.0105	0.1040	0.6345

Table 5.5 ANOVA: Comments by Type and Commenter Reputation Class (*p<0.05, **p<0.01)

A limitation of this test, however, is that we only look at the top-rated comment. There is the possibility that Trusted Users give constructive comments that are not highly rated by the community members. However, it is important to recognize that while Trusted Users account for 31% of the comments, they are less than 0.1% of the community at large. In addition, Established Users are part of the top 1% of users, but 49% of all comments. The results thus can be considered to be indicative of a larger pattern of behavior.

5.6 Conclusion and Discussion

We were able to identify three common types of comments that were left as popular feedback on the bad questions. The most common was critical, followed by corrective, and comments that

helped to answer the question. We might expect that the most common type of feedback would receive the highest amount of voting, and this is true. Critical comments receive more votes than other types of comments. In addition, comments that consisted of only criticism received even more votes than other isolated types of comments. Finally, we found that comments came from all types of users, including high reputation users. The results not only confirm the suspicion from previous work (Ahn et al. 2013) that there is a critical trend in the comments, but that this trend is popular and affirmed within the community.

The results indicate, however, that social approval is given to criticism, some of which would be considered "unconstructive" by the SO administration. Since we can expect users will develop contribution patterns when they are socially affirmed by the community (Cheshire 2007), this offers an explanation for why users of all reputation classes act similarly. Rather than criticism being something that happens by individual actors, it is part of a larger social construct. This is something that is cause for concern from a system design point of view as strong negativity has a likely adverse effect on producers and potential producers (Zhu et al. 2013), which might have a critical externality effect on outside users, and is specifically banned by the community administration. At the same time, there is a large amount of corrective commenting that is present, and it does receive some social approval. This indicates that there are users that support the socially desirable actions as defined by the administration.

One observation that comes out of this research is that the community users have norms that differ from the administration. Q&A has expanded from generalist systems with little concern for archival efficacy to domain specific sites that seek to be an authoritative and useful archive of knowledge (Anderson et al. 2012; Dearman and Truong 2010). A challenge to these archives is the sheer number of bad questions that appear. In such situations, it is likely a natural

outcome that the community both writes and approves of feedback that is critical and unconstructive. It should be noted, however, that not all of the comments contain negativity, and that there is a considerable amount of corrective feedback mixed into the comments. There is potential to draw corrective feedback to the forefront of the contributions.

The driving force behind the social approval of criticism may be that it essentially accomplishes a desired goal: to firmly chastise users who post bad questions. The interactions we reviewed show a strong emphasis on effort and skill. The commenters do emphasize that learning through reading and study is an essential part of learning to be a computer programmer. In addition, those who show no effort in their questions are despised. Commenters may believe that bad questioners have no potential value to the community, so critical comments are valid from the viewpoint of the community, regardless if it is actually socially beneficial from the viewpoint of the administrators.

One possibility for ameliorating the power of social approval is to create badges that support socially beneficial behavior. While we cannot conclude that the current commenting badge is directly responsible for unconstructive criticism, we can conclude that current voting system allows for the earning of these badges. This defeats the purpose of comments in situations like bad questions. The dialogue is supposed to support the questioner, who will then benefit the community by producing better content. A way to encourage this is to create badges that are directly linked to support situations. For instance, allowing questioners to select the most helpful comment and rewarding badges for such an effort could offer a tangible incentive to leave constructive feedback. Another recommendation is to focus on Trusted Users. Since previous work has identified that users are especially effected by the actions of core members (Zhu et al. 2013), focusing on these users' comments could have a beneficial network effect.

There are some limitations and future work that need to be discussed. The observations were gathered in a single system that has a technical focus and the study analyzed a subset of questions. Future work could focus on the relationship between socially desirable behavior, what is socially rewarded by the community, and clearer reputation incentives. That is, if the dynamics change and socially undesirable behavior is penalized by the community, does behavior of the commenters significantly change? Other avenues could center around modifying social approval mechanisms to remove unwanted incentives.

A limitation of this work lies in the lack of interviews. It would be informative to get the perspective of users who have contributed bad questions as well as gain an understanding of the explicit motives of users who comment on these questions. It would also be useful to understand if there is a pattern of commenting from contributors that evolves over time as they are exposed to more content.

5.7 Summary

In this chapter, we examined an aspect of Community-Based Q&A by looking at the characteristics of bad questions and analyzed how the community of Stack Overflow (SO) comments on them. Questions usually seek help in completing some sort of task, such as creating a process or debugging code. By using Grounded Theory, we were able to determine that comments had large amounts of corrective guidance and criticism. Comments also sometimes aimed at giving an answer to question. We found that critical comments were the most popular and received the highest vote counts, and that commenters did not change their commenting behavior based on their reputation.

This means that higher reputation users cannot be trusted to be better at interaction and are likely influenced by social approval from the upvotes received for critical comments. This

shows a limitation to the effectiveness of reputation class as a predictor of socially desirable behavior, and it also highlights the potential vulnerability to social incentives which do reinforce undesirable behavior.

In the next chapter, we conclude this thesis by summarizing the contributions and providing avenues for future work.

Chapter 6

Conclusion

6.1 Contributions

In this thesis we provided analysis of user behavior and action in a Q&A community within the construct of a reputation class system. We focused on aspects of behavior that go beyond the answering of questions, and further focus on how users present themselves, collaborate on questions, and communicate with each other via public commenting. We focus on the idea of reputation class as being a possible signal for socially desirable behavior that goes beyond indicating answer-quality status. Overall, we find that reputation can be used to identify collaborative behavior, but not in the provision of identity or in the quality of communication. The main contributions of this thesis are as follows:

1. *An analysis of Identity and Reputation in Q&A*

We analyzed the provision of identity and the relationship with reputation class. Our main focus was to understand whether users with fuller identities tied to their offline persons were either better at earning reputation or were socially rewarded by the community by giving their reputation. In order to accomplish an accurate and real-world analysis, we focused on classifying identities through an exhaustive reflexive-iterative version of Grounded Theory. This classification technique allowed us to quantitatively analyze users based on frequent composites of identity types that community members see. It also allowed us to determine variables based on what users put into their identity fields rather than if they simply put "something" into their field. We found that in this study, users with fuller identities were more likely to belong to higher reputation classes. However, we found that this was also influenced by longer tenures in the system. A per-contribution analysis shows that identity does not impact Q&A performance. This

indicates that there is no social incentive for providing identity when registering a new account. The results are important to system designers who have to decide the flexibility of identity options available for users.

2. An analysis of collaborative behavior and reputation in Q&A

We analyzed the behavior of editors on bad questions in Q&A. Editing these questions is a socially desirable activity for the community, since it saves questions with potential archival value from being deleted or ignored. We developed a classification scheme of potential editors of questions, deriving their possible motivations for contributing edits: from competitive to intrinsic. We found that the users with the highest reputation and intrinsic motivations were the most likely to provide edits, were the most likely to provide the first edit, and were the most likely to provide the most valuable edits. An analysis of edits and high-reputation users show that these users are dedicated to editing material frequently. The results indicate that reputation incentives are limited in encouraging socially desirable collaborative behavior. This is an important result for use by system designers seeking to expand the number of their participants in collaborating in archival maintenance.

3. An analysis of reputation classes and communication through commenting.

In this topic we presented a study of the comments left by users on bad questions. Communication with a questioner is often necessary when they leave an unwanted poor-quality question. Ideally, this communication, in the form of comments left underneath the question, would give the questioner value feedback on improving the question or future contributions. We would expect that high-reputation users who receive moderation privileges from the community would be more likely to leave desirable comments. Using Grounded Theory, we categorized comments based on three elements: 1) how constructive they are, 2) how critical they are, and 3)

if they give an answer. The first type of comment is desirable, while the other two are not. We then compared the comments left by different reputation class users and measured their popularity with the overall community by using the vote total left on the comment. The results show that the most social approval is given to critical comments, and that there is no differentiation between comments left and the reputation status of the commenter.

Overall, the results show a clear limitation to the application of reputation class in predicting behavior. Social incentives need to be constructed within the context of the target activity in order to be successful, and the power of reputation class as an indicator of quality is not transitive to all activities.

The results of this thesis are derived from a single system, Stack Overflow (SO). The results are generalizable in several ways, however. First, we expect that results will be transferable to Q&A systems that use similar mechanics to SO or have imported the platform entirely like the SO. We also expect that similar results will be seen in Q&A and peer production communities that mirror the individual elements found in SO, such as the editing of questions or social commenting.

6.2 Future Directions

There are several avenues for future work that we hope that can be explored in the near future. The first avenue is in mapping a behavior when a user participates across domain communities. The experience of being a high-reputation class user in one community may be transferable to other sites, even if their reputation remains low. For instance, our research regarding language versioning of SO show that elite users tend to target the most popular questions, even if they are specialized and not general (Vargo et al. 2018). This would be in the vein of the socially desirable outcome (creating solid archives of computer programming Q&A) not being fulfilled in

the face of social incentives.

An important avenue for future work is the development of mechanisms or reputation systems that effectively reward and manage social behavior. Since reputation points earned from Q&A do not accurately predict whether users engage in socially desirable behavior, there is an opportunity to find new ways to measure and predict users who will engage in desirable behaviors. For example, these could consist of something like parallel reputation systems, such as a comprehensive social score running side by side with the current reputation system. This would also add a challenge by introducing network externalities that may affect (positively or negatively) the efficacy of the main reputation system.

Finally, we aim to study the impact of culture and language on outcomes regarding socially desirable behavior. Since most content is created and distributed in English, this is where the majority of research occurs. However, different cultures do not always treat information and peer-production the same (Hecht and Gergle 2010; Vargo et al. 2018). It is valuable to consider how different cultures and languages treat reputation points and socially desirable behavior.

References

Adaji, Ifeoma, and Julita Vassileva. 2016. “Towards Understanding User Participation in Stack Overflow Using Profile Data.” In *Social Informatics*, 3–13. Springer, Cham. https://doi.org/10.1007/978-3-319-47874-6_1.

Adamic, Lada A., Jun Zhang, Eytan Bakshy, and Mark S. Ackerman. 2008a. “Knowledge Sharing and Yahoo Answers: Everyone Knows Something.” In *Proceedings of the 17th International Conference on World Wide Web*, 665–74. WWW ’08. New York, NY, USA: ACM. https://doi.org/10.1145/1367497.1367587.

———. 2008b. “Knowledge Sharing and Yahoo Answers: Everyone Knows Something.” In *Proceedings of the 17th International Conference on World Wide Web*, 665–74. WWW ’08. New York, NY, USA: ACM. https://doi.org/10.1145/1367497.1367587.

Adler, B. Thomas, and Luca de Alfaro. 2007. “A Content-Driven Reputation System for the Wikipedia.” In *Proceedings of the 16th International Conference on World Wide Web*, 261–70. WWW ’07. New York, NY, USA: ACM. https://doi.org/10.1145/1242572.1242608.

Ahn, June, Brian S. Butler, Cindy Weng, and Sarah Webster. 2013. “Learning to Be a Better Q’er in Social Q&A Sites: Social Norms and Information Artifacts.” *Proceedings of the American Society for Information Science and Technology* 50 (1): 1–10. https://doi.org/10.1002/meet.14505001032.

Allamanis, Miltiadis, and Charles Sutton. 2013. “Why, When, and What: Analyzing Stack Overflow Questions by Topic, Type, and Code.” In *Proceedings of the 10th Working Conference on Mining Software Repositories*, 53–56. MSR ’13. Piscataway, NJ, USA: IEEE Press. http://dl.acm.org/citation.cfm?id=2487085.2487098.

Anderson, Ashton, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2012. “Discovering Value from Community Activity on Focused Question Answering Sites: A Case Study of Stack

Overflow.” In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 850–58. KDD ’12. New York, NY, USA: ACM. https://doi.org/10.1145/2339530.2339665.

Barua, Anton, Stephen W. Thomas, and Ahmed E. Hassan. 2014. “What Are Developers Talking About? An Analysis of Topics and Trends in Stack Overflow.” *Empirical Softw. Engg.* 19 (3): 619–54. https://doi.org/10.1007/s10664-012-9231-y.

Bazelli, B., A. Hindle, and E. Stroulia. 2013. “On the Personality Traits of StackOverflow Users.” In *2013 IEEE International Conference on Software Maintenance*, 460–63. https://doi.org/10.1109/ICSM.2013.72.

Bosu, Amiangshu, Christopher S. Corley, Dustin Heaton, Debarshi Chatterji, Jeffrey C. Carver, and Nicholas A. Kraft. 2013. “Building Reputation in StackOverflow: An Empirical Investigation.” In *Proceedings of the 10th Working Conference on Mining Software Repositories*, 89–92. MSR ’13. Piscataway, NJ, USA: IEEE Press. http://dl.acm.org/citation.cfm?id=2487085.2487107.

Bruce, Catherine D. 2007. “Questions Arising about Emergence, Data Collection, and Its Interaction with Analysis in a Grounded Theory Study.” *International Journal of Qualitative Methods* 6 (1): 51–68.

Cavusoglu, Huseyin, Zhuolun Li, and Ke-Wei Huang. 2015. “Can Gamification Motivate Voluntary Contributions?: The Case of StackOverflow Q&A Community.” In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing*, 171–74. CSCW’15 Companion. New York, NY, USA: ACM. https://doi.org/10.1145/2685553.2698999.

Cheng, Justin, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2014. “How Community

Feedback Shapes User Behavior." In *Proceedings of ICWSM*. http://arxiv.org/abs/1405.1429.

Chen, Yan, Teck-Hua Ho, and Yong-Mi Kim. 2010. "Knowledge Market Design: A Field Experiment at Google Answers." *Journal of Public Economic Theory* 12 (4): 641–64. https://doi.org/10.1111/j.1467-9779.2010.01468.x.

Cheshire, Coye. 2007. "Selective Incentives and Generalized Information Exchange." *Social Psychology Quarterly* 70 (1): 82–100. https://doi.org/10.1177/019027250707000109.

Choi, Boreum, Kira Alexander, Robert E. Kraut, and John M. Levine. 2010. "Socialization Tactics in Wikipedia and Their Effects." In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, 107–16. CSCW '10. New York, NY, USA: ACM. https://doi.org/10.1145/1718918.1718940.

Choi, Erik, Vanessa Kitzie, and Chirag Shah. 2010. "Developing a Typology of Online Q&A Models and Recommending the Right Model for Each Question Type." *Proceedings of the American Society for Information Science and Technology* 49 (1): 1–4. https://doi.org/10.1002/meet.14504901302.

Correa, Denzil, and Ashish Sureka. 2013. "Fit or Unfit : Analysis and Prediction of 'Closed Questions' on Stack Overflow." *arXiv:1307.7291 [cs]*, July. http://arxiv.org/abs/1307.7291.

———. 2014. "Chaff from the Wheat: Characterization and Modeling of Deleted Questions on Stack Overflow." In *Proceedings of the 23rd International Conference on World Wide Web*, 631–42. WWW '14. New York, NY, USA: ACM. https://doi.org/10.1145/2566486.2568036.

Couch, Danielle, and Pranee Liamputtong. 2008. "Online Dating and Mating: The Use of the Internet to Meet Sexual Partners." *Qualitative Health Research* 18 (2): 268–79. https://doi.org/10.1177/1049732307312832.

De Alfaro, Luca, Ashutosh Kulshreshtha, Ian Pye, and B. Thomas Adler. 2011. "Reputation

Systems for Open Collaboration." *Commun. ACM* 54 (8): 81–87. https://doi.org/10.1145/1978542.1978560.

Dearman, David, and Khai N. Truong. 2010. "Why Users of Yahoo!: Answers Do Not Answer Questions." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 329–32. CHI '10. New York, NY, USA: ACM. https://doi.org/10.1145/1753326.1753376.

Donath, Judith. 2014. *The Social Machine: Designs for Living Online*. Cambridge, MA: The MIT Press.

Ellison, Nicole, Rebecca Heino, and Jennifer Gibbs. 2006. "Managing Impressions Online: Self-Presentation Processes in the Online Dating Environment." *Journal of Computer-Mediated Communication* 11 (2): 415–41. https://doi.org/10.1111/j.1083-6101.2006.00020.x.

Ford, Heather. 2012. "Online Reputation: it’s Contextual." Ethnography Matters. February 24, 2012. http://ethnographymatters.net/blog/2012/02/24/online-reputation-its-contextual/.

Friedman*, Eric J., and Paul Resnick. 2001. "The Social Cost of Cheap Pseudonyms." *Journal of Economics & Management Strategy* 10 (2): 173–99. https://doi.org/10.1111/j.1430-9134.2001.00173.x.

Furtado, Adabriand, Nazareno Andrade, Nigini Oliveira, and Francisco Brasileiro. 2013. "Contributor Profiles, Their Dynamics, and Their Importance in Five Q&a Sites." In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, 1237–52. CSCW '13. New York, NY, USA: ACM. https://doi.org/10.1145/2441776.2441916.

Gazan, R. 2007. "Seekers, Sloths and Social Reference: Homework Questions Submitted to a Question-Answering Community." *New Review of Hypermedia and Multimedia* 13 (2): 239–48.

https://doi.org/10.1080/13614560701711917.

Gazan, Rich. 2010. “Microcollaborations in a Social Q&A Community.” *Information Processing & Management*, Collaborative Information Seeking, 46 (6): 693–702. https://doi.org/10.1016/j.ipm.2009.10.007.

———. 2011. “Redesign As an Act of Violence: Disrupted Interaction Patterns and the Fragmenting of a Social Q&A Community.” In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2847–56. CHI ’11. New York, NY, USA: ACM. https://doi.org/10.1145/1978942.1979365.

Ginsca, Alexandru Lucian, and Adrian Popescu. 2013. “User Profiling for Answer Quality Assessment in Q&A Communities.” In *Proceedings of the 2013 Workshop on Data-Driven User Behavioral Modelling and Mining from Social Media*, 25–28. DUBMOD ’13. New York, NY, USA: ACM. https://doi.org/10.1145/2513577.2513579.

Glaser, Barney G., and Anselm L. Strauss. 2009. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Transaction Publishers.

Golbeck, Jennifer, and Kenneth R. Fleischmann. 2010. “Trust in Social Q&A: The Impact of Text and Photo Cues of Expertise.” *Proceedings of the American Society for Information Science and Technology* 47 (1): 1–10. https://doi.org/10.1002/meet.14504701048.

Halfaker, Aaron, R. Stuart Geiger, Jonathan T. Morgan, and John Riedl. 2013a. “The Rise and Decline of an Open Collaboration System How Wikipedia’s Reaction to Popularity Is Causing Its Decline.” *American Behavioral Scientist* 57 (5): 664–88. https://doi.org/10.1177/0002764212469365.

———. 2013b. “The Rise and Decline of an Open Collaboration System How Wikipedia’s Reaction to Popularity Is Causing Its Decline.” *American Behavioral Scientist* 57 (5): 664–88.

https://doi.org/10.1177/0002764212469365.

Halfaker, Aaron, Aniket Kittur, and John Riedl. 2011. “Don’T Bite the Newbies: How Reverts Affect the Quantity and Quality of Wikipedia Work.” In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, 163–72. WikiSym ’11. New York, NY, USA: ACM. https://doi.org/10.1145/2038558.2038585.

Harper, F. Maxwell, Daniel Moy, and Joseph A. Konstan. 2009. “Facts or Friends?: Distinguishing Informational and Conversational Questions in Social Q&A Sites.” In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 759–68. CHI ’09. New York, NY, USA: ACM. https://doi.org/10.1145/1518701.1518819.

Harper, F. Maxwell, Daphne Raban, Sheizaf Rafaeli, and Joseph A. Konstan. 2008. “Predictors of Answer Quality in Online Q&A Sites.” In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 865–74. CHI ’08. New York, NY, USA: ACM. https://doi.org/10.1145/1357054.1357191.

Hart, Kerry, and Anita Sarma. 2014. “Perceptions of Answer Quality in an Online Technical Question and Answer Forum.” In *Proceedings of the 7th International Workshop on Cooperative and Human Aspects of Software Engineering*, 103–6. CHASE 2014. New York, NY, USA: ACM. https://doi.org/10.1145/2593702.2593703.

Hecht, Brent, and Darren Gergle. 2010. “The Tower of Babel Meets Web 2.0: User-Generated Content and Its Applications in a Multilingual Context.” In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 291–300. CHI ’10. New York, NY, USA: ACM. https://doi.org/10.1145/1753326.1753370.

“How Does Editing Work? - Meta Stack Exchange.” 2016. 2016. http://meta.stackexchange.com/questions/21788/how-does-editing-work.

Hsieh, Gary, Robert E. Kraut, and Scott E. Hudson. 2010. “Why Pay?: Exploring How Financial Incentives Are Used for Question & Answer.” In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 305–14. CHI ’10. New York, NY, USA: ACM. https://doi.org/10.1145/1753326.1753373.

Jain, Shaili, Yiling Chen, and David C. Parkes. 2009. “Designing Incentives for Online Question and Answer Forums.” In *Proceedings of the 10th ACM Conference on Electronic Commerce*, 129–38. EC ’09. New York, NY, USA: ACM. https://doi.org/10.1145/1566374.1566393.

Jeon, Grace YoungJoo, and Soo Young Rieh. 2014. “Answers from the Crowd: How Credible Are Strangers in Social Q&A?” *iConference 2014 Proceedings*. https://www.ideals.illinois.edu/handle/2142/47266.

Joo, Jaehun, and Ismatilla Normatov. 2013. “Determinants of Collective Intelligence Quality: Comparison between Wiki and Q&A Services in English and Korean Users.” *Service Business* 7 (4): 687–711. https://doi.org/10.1007/s11628-013-0183-0.

Kittur, Aniket, and Robert E. Kraut. 2008. “Harnessing the Wisdom of Crowds in Wikipedia: Quality Through Coordination.” In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*, 37–46. CSCW ’08. New York, NY, USA: ACM. https://doi.org/10.1145/1460563.1460572.

Laniado, David, Andreas Kaltenbrunner, Carlos Castillo, and Mayo Fuster Morell. 2012. “Emotions and Dialogue in a Peer-Production Community: The Case of Wikipedia.” In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, 9:1–9:10. WikiSym ’12. New York, NY, USA: ACM. https://doi.org/10.1145/2462932.2462944.

Lappas, Theodoros, Chrysanthos Dellarocas, and Neda Derakhshani. 2017. “Reputation and Contribution in Online Question-Answering Communities.” SSRN Scholarly Paper ID 2918913.

Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=2918913.

LaToza, T. D., and A. van der Hoek. 2016. “Crowdsourcing in Software Engineering: Models, Motivations, and Challenges.” *IEEE Software* 33 (1): 74–80. https://doi.org/10.1109/MS.2016.12.

Li, Guo, Haiyi Zhu, Tun Lu, Xianghua Ding, and Ning Gu. 2015. “Is It Good to Be Like Wikipedia?: Exploring the Trade-Offs of Introducing Collaborative Editing Model to Q&A Sites.” In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 1080–91. CSCW ’15. New York, NY, USA: ACM. https://doi.org/10.1145/2675133.2675155.

Liu, Jun, and Sudha Ram. 2011. “Who Does What: Collaboration Patterns in the Wikipedia and Their Impact on Article Quality.” *ACM Trans. Manage. Inf. Syst.* 2 (2): 11:1–11:23. https://doi.org/10.1145/1985347.1985352.

Li, Zhuolun, Ke-wei Huang, and Huseyin Cavusoglu. 2012. “Quantifying the Impact of Badges on User Engagement in Online Q&A Communities.” *ICIS 2012 Proceedings*, December. http://aisel.aisnet.org/icis2012/proceedings/ResearchInProgress/74.

MacLeod, L. 2014. “Reputation on Stack Exchange: Tag, You’re It!” In *2014 28th International Conference on Advanced Information Networking and Applications Workshops*, 670–74. https://doi.org/10.1109/WAINA.2014.108.

Mamykina, Lena, Bella Manoim, Manas Mittal, George Hripcsak, and Björn Hartmann. 2011. “Design Lessons from the Fastest Q&a Site in the West.” In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2857–66. CHI ’11. New York, NY, USA: ACM. https://doi.org/10.1145/1978942.1979366.

Marlow, Jennifer, Laura Dabbish, and Jim Herbsleb. 2013. “Impression Formation in Online

Peer Production: Activity Traces and Personal Profiles in Github." In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, 117–28. CSCW '13. New York, NY, USA: ACM. https://doi.org/10.1145/2441776.2441792.

Mauthner, Natasha S., and Andrea Doucet. 2003. "Reflexive Accounts and Accounts of Reflexivity in Qualitative Data Analysis." *Sociology* 37 (3): 413–31. https://doi.org/10.1177/00380385030373002.

Mill, Eric. 2014. "Quora Keeps the World's Knowledge For Itself." 2014. /post/quora-keeps-the-worlds-knowledge-for-itself.

Morris, Meredith Ringel, Jaime Teevan, and Katrina Panovich. 2010. "What Do People Ask Their Social Networks, and Why?: A Survey Study of Status Message Q&a Behavior." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1739–48. CHI '10. New York, NY, USA: ACM. https://doi.org/10.1145/1753326.1753587.

Nam, Kevin Kyung, Mark S. Ackerman, and Lada A. Adamic. 2009. "Questions In, Knowledge in?: A Study of Naver's Question Answering Community." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 779–88. CHI '09. New York, NY, USA: ACM. https://doi.org/10.1145/1518701.1518821.

Oh, Sanghee. 2010. *Answerers' Motivations and Strategies for Providing Information and Social Support in Social Q&A an Investigation of Health Question Answering*. ProQuest LLC.

Oliver, Pamela, Gerald Marwell, and Ruy Teixeira. 1985. "A Theory of the Critical Mass: I. Interdependence, Group Heterogeneity, and the Production of Collective Action." *American Journal of Sociology* 91 (3): 522–56. https://doi.org/10.1086/228313.

"OpenID, One Year Later." 2010. Stack Overflow Blog. 2010. /2010/04/openid-one-year-later/.

Ortega, Felipe, Jesus M. Gonzalez-Barahona, and Gregorio Robles. 2008. "On the Inequality of

Contributions to Wikipedia." In *Proceedings of the Proceedings of the 41st Annual Hawaii International Conference on System Sciences*, 304 – . HICSS '08. Washington, DC, USA: IEEE Computer Society. https://doi.org/10.1109/HICSS.2008.333.

Pal, Aditya, and Joseph A. Konstan. 2010. "Expert Identification in Community Question Answering: Exploring Question Selection Bias." In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, 1505–8. CIKM '10. New York, NY, USA: ACM. https://doi.org/10.1145/1871437.1871658.

Panciera, Katherine, Aaron Halfaker, and Loren Terveen. 2009. "Wikipedians Are Born, Not Made: A Study of Power Editors on Wikipedia." In *Proceedings of the ACM 2009 International Conference on Supporting Group Work*, 51–60. GROUP '09. New York, NY, USA: ACM. https://doi.org/10.1145/1531674.1531682.

Patil, Sumanth, and Kyumin Lee. 2016. "Detecting Experts on Quora: By Their Activity, Quality of Answers, Linguistic Characteristics and Temporal Behaviors." *Social Network Analysis and Mining* 6 (1): 5. https://doi.org/10.1007/s13278-015-0313-x.

Raban, Daphne Ruth. 2009. "Self-Presentation and the Value of Information in Q&A Websites." *Journal of the American Society for Information Science and Technology* 60 (12): 2465–73. https://doi.org/10.1002/asi.21188.

Schneider, Jodi, Alexandre Passant, and Stefan Decker. 2012. "Deletion Discussions in Wikipedia: Decision Factors and Outcomes." In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*, 17:1–17:10. WikiSym '12. New York, NY, USA: ACM. https://doi.org/10.1145/2462932.2462955.

Schwen, Thomas M., and Noriko Hara. 2003. "Community of Practice: A Metaphor for Online Design?" *The Information Society* 19 (3): 257–70. https://doi.org/10.1080/01972240309462.

Shah, Chirag, Jung Sun Oh, and Sanghee Oh. 2008. “Exploring Characteristics and Effects of User Participation in Online Social Q&A Sites.” *First Monday* 13 (9). https://doi.org/10.5210/fm.v13i9.2182.

Singer, Leif, Fernando Figueira Filho, Brendan Cleary, Christoph Treude, Margaret-Anne Storey, and Kurt Schneider. 2013. “Mutual Assessment in the Social Programmer Ecosystem: An Empirical Investigation of Developer Profile Aggregators.” In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, 103–16. CSCW ’13. New York, NY, USA: ACM. https://doi.org/10.1145/2441776.2441791.

Solomon, Jacob, and Rick Wash. 2014. “Critical Mass of What? Exploring Community Growth in WikiProjects.” In *Eighth International AAAI Conference on Weblogs and Social Media*. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8104.

Spolsky, Joel. 2012. “Kicking off the Summer of Love - Stack Overflow Blog.” 2012. https://stackoverflow.blog/2012/07/20/kicking-off-the-summer-of-love/.

Srivastava, Prachi, and Nick Hopwood. 2009. “A Practical Iterative Framework for Qualitative Data Analysis.” *International Journal of Qualitative Methods* 8 (1): 76–84.

“Stack Exchange Clones.” 2018. Meta Stack Exchange. 2018. https://meta.stackexchange.com/questions/2267/stack-exchange-clones/37953.

“Stackoverflow.com Traffic, Demographics and Competitors - Alexa.” 2018. 2018. https://www.alexa.com/siteinfo/stackoverflow.com.

Tausczik, Yla R., Aniket Kittur, and Robert E. Kraut. 2014. “Collaborative Problem Solving: A Study of MathOverflow.” In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 355–67. CSCW ’14. New York, NY, USA: ACM. https://doi.org/10.1145/2531602.2531690.

Tausczik, Yla R., and James W. Pennebaker. 2011. “Predicting the Perceived Quality of Online Mathematics Contributions from Users’ Reputations.” In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1885–88. CHI ’11. New York, NY, USA: ACM. https://doi.org/10.1145/1978942.1979215.

———. 2012. “Participation in an Online Mathematics Community: Differentiating Motivations to Add.” In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, 207–16. CSCW ’12. New York, NY, USA: ACM. https://doi.org/10.1145/2145204.2145237.

Treude, Christoph, Ohad Barzilay, and Margaret-Anne Storey. 2011. “How Do Programmers Ask and Answer Questions on the Web? (NIER Track).” In *Proceedings of the 33rd International Conference on Software Engineering*, 804–7. ICSE ’11. New York, NY, USA: ACM. https://doi.org/10.1145/1985793.1985907.

Vargo, A. W., Benjamin Tag, Kai Kunze, and Shigeo Matsubara. 2018. “Different Languages, Different Questions: Language Versioning in Q&A.” In *UK Academy of Information Systems (UKAIS 2018)*.

Warncke-Wang, Morten, Vladislav R. Ayukaev, Brent Hecht, and Loren G. Terveen. 2015. “The Success and Failure of Quality Improvement Projects in Peer Production Communities.” In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 743–56. CSCW ’15. New York, NY, USA: ACM. https://doi.org/10.1145/2675133.2675241.

Wei, X., W. Chen, and K. Zhu. 2015. “Motivating User Contributions in Online Knowledge Communities: Virtual Rewards and Reputation.” In *2015 48th Hawaii International Conference on System Sciences*, 3760–69. https://doi.org/10.1109/HICSS.2015.452.

"What Is Reputation? How Do I Earn (and Lose) It? - Help Center - Stack Overflow." 2015. 2015. http://stackoverflow.com/help/whats-reputation.

"Why Can People Edit My Posts? How Does Editing Work? - Help Center - Stack Overflow." 2016. 2016. http://stackoverflow.com/help/editing.

"Why Did Quora Get Rid of Question Details? Isn't It Rather Crucial to Understand over Half of the Existing Content on the Site? Why Not Just Prevent People from Adding Details to New Questions While Retaining the Details Attached to the Old Ones? - Quora." 2017. 2017. https://www.quora.com/Why-did-Quora-get-rid-of-question-details-Isn%E2%80%99t-it-rather-crucial-to-understand-over-half-of-the-existing-content-on-the-site-Why-not-just-prevent-people-from-adding-details-to-new-questions-while-retaining-the-details-attached-to-the-old-ones.

Wohn, Donghee Yvette. 2015. "The Effects of Feedback and Habit on Content Posting in an Online Community," March. https://www.ideals.illinois.edu/handle/2142/73646.

Xu, Lei, Nian Tingting, and Luis Cabral. 2014. "What Makes Geeks Tick? A Study of Stack Overflow Careers." *Working Paper*.

Yang, Jiang, Meredith Ringel Morris, Jaime Teevan, Lada A. Adamic, and Mark S. Ackerman. 2011. "Culture Matters: A Survey Study of Social Q&A Behavior." In . Barcelona, Spain: AAAI. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2755/3305.

Yang, Jie, Claudia Hauff, Alessandro Bozzon, and Geert-Jan Houben. 2014. "Asking the Right Question in Collaborative Q&a Systems." In *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, 179–89. HT '14. New York, NY, USA: ACM. https://doi.org/10.1145/2631775.2631809.

Zhang, Jie. 2006. "The Roles of Players and Reputation: Evidence from eBay Online Auctions." *Decision Support Systems* 42 (3): 1800–1818. https://doi.org/10.1016/j.dss.2006.03.008.

Zhu, Haiyi, Robert E. Kraut, Yi-Chia Wang, and Aniket Kittur. 2011. "Identifying Shared Leadership in Wikipedia." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 3431–34. CHI '11. New York, NY, USA: ACM. https://doi.org/10.1145/1978942.1979453.

Zhu, Haiyi, Robert Kraut, and Aniket Kittur. 2012. "Effectiveness of Shared Leadership in Online Communities." In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, 407–16. CSCW '12. New York, NY, USA: ACM. https://doi.org/10.1145/2145204.2145269.

Zhu, Haiyi, Amy Zhang, Jiping He, Robert E. Kraut, and Aniket Kittur. 2013. "Effects of Peer Feedback on Contribution: A Field Experiment in Wikipedia." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2253–62. CHI '13. New York, NY, USA: ACM. https://doi.org/10.1145/2470654.2481311.

Chapter 3 is an Accepted Manuscript of an article published by Taylor & Francis in Behaviour and Information Technology on 2018/7/1, available online: https://www.tandfonline.com/doi/full/10.1080/0144929X.2018.1474251