FABLES OF NEGATIVE BEHAVIORS IN ONLINE ENVIRONMENTS: MOTIVATIONS AND MANAGEMENT OF ONLINE CONTRIBUTIONS

A Dissertation
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Loxley Sijia Wang

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Shawn Curley, Yuqing Ren

July 2016

Acknowledgements

I would like to acknowledge the many, many people who were instrumental throughout the process of this dissertation, both directly and behind the scenes. In particular, I would like to thank:

My advisors, Ching Ren and Shawn Curley, for their dedicated support, feedback, and advice in everything academic and non-academic;

My committee, Alok Gupta and Loren Terveen, for their generous time and guidance in helping to shape this research;

The IDSc department faculty, for all of their teachings and inspiration;

The IDSc office staff and the PhD office staff, for the conversations and the administrative support;

The late, great John Riedl, for his unwavering spirit and for lessons on the key to happiness;

Jilin Chen, Christopher Lu, and Samantha Lee, for their collaboration and help in research;

Catherine Bui, for companionship in the lab and her thoughtfulness;

My fellow IDSc PhD students, for their friendship and belief that I actually know what I'm doing;

My officemates, YoungOk Kwon and Chengxin Cao, for the commiseration and for acting as my personal therapists;

Jenny Chen, for both keeping me on track and distracting me;

Sarah Looff, for showing me the light at the end of the tunnel;

My non-academic friends, for reminders to step outside every once in a while;

My family, for believing the best in me;

And last, but certainly not least, The Programme—Trey Hickman, Georg Meyer, and Miguel Velasco, for the coffee, the inside jokes, the memories, and being the best cohort anyone could hope to have.

Dedication

This dissertation is dedicated to the sister for whom I hope to set a better example.

Abstract

Online communities depend on positive contributions from productive members in order to thrive. However, many members may behave in ways that are detrimental to the health of these communities. Three studies were used to examine negative behaviors in each of three different types of online communities.

First, gaming behaviors and work quality were examined through a lab experiment and a field experiment in an online marketplace for work. In such communities, improving workers' perceptions of task significance led to increases in both quality and quantity of work. Second, withdrawal behaviors and decreased contributions were examined in an online volunteer community. Better time management and the prevention of burnout and stress could encourage both retention of members and continued productivity. Third, trolling behaviors were examined in an online social network community through interviews and an exploratory study. Allowing community members to identify and define negative trolling behaviors while defending false accused members could aid administrators in moderating these types of behaviors without wrongly banning legitimate contributors.

These three studies provide insights into the causes and effects of problematic behaviors that could hinder online contributions. These insights further help to prescribe methods to moderate and manage the negative impacts of these behaviors.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction to Three Areas of Online Contribution

I. INTRODUCTION

In the spring of 2005, shortly after Pope Benedict XVI's election to the papacy, someone changed his entry on the online encyclopedia Wikipedia from his own photo to one of the evil Emperor Palpatine from Star Wars (Barack 2005). Due to an increasing number of such incidences, Wikipedia founder, Jimmy Wales, stated that he would add time delays to the software in order to help protect the articles from such acts of vandalism. Though forms of "semi-protection" provide a degree of stability against similar acts, they may increase difficulty for casual contributors to use the site.

In 2009, within the online game World of Warcraft, one group of members held a virtual funeral for the real-life death of a fellow member (Cuddy and Nordlinger 2009). During the in-game funeral, an enemy group intentionally attacked and virtually slaughtered the members of the first group. The enemy group considered the environment to be a game first and foremost and used the event to their advantage. The community did not come to a consensus on whether or not the action was wrong, as it did not break any formal rules. However, the ambiguity surrounding proper behavior on the site caused increased conflict within the community.

In both of these incidents, legitimate contributors to online platforms may be hindered by the behaviors of deviants and the responses to control such behaviors. Although online communities allow people to connect to others, problems with these

communities can hinder progress and deter members. In various collaborative online communities, where members do anything from responding to discussion forums to actively working together, users must be constantly vigilant when interacting with others to prevent harm to themselves (Jarvenpaa and Majchrzak 2010). In this dissertation, I examine three types of sites in which online contributions are crucial to the function of the communities. Through this research, I hope to advance information systems theory and practice in the area of online communities by proposing ways to identify and manage negative online behaviors.

Specifically, this dissertation aims to explore the following through three different studies: 1) understanding workers' negative behaviors within online marketplaces for work, 2) understanding tradeoffs in negative behaviors within formal online volunteer groups, and 3) understanding disruptive trolling behaviors within online social networking communities. Though they take place in different contexts, these three studies all examine negative behaviors that occur when members, intentionally or not, disrupt an online community to reduce legitimate contributions. The studies investigate how these behaviors may be similar to or different from offline behaviors, with an eye toward reducing their effects.

II. STUDY 1: Online Marketplaces for Work

The past decade has seen growing interest in the online phenomenon of crowdsourcing, the idea that jobs can be "outsourced to the crowd." The crowdsourcing model can provide many advantages. Businesses can tap into a larger, more diverse, and comparatively

cheaper pool of labor. Workers can work on whatever tasks interest them, have fun, earn extra income, or simply kill time.

One fast-growing area of research is online marketplaces for work, in which companies and individuals can post tasks and request work to be done while other individuals can work on these tasks for payment. Such marketplaces for work are a subset of crowdsourcing communities but are unique in that they often focus on simple tasks that humans can do better than computers. Online marketplaces for work like Amazon Mechanical Turk (MTurk) enable businesses to post micro-tasks, from which thousands of online workers can select preferred tasks, often with hourly rates of $1 to $3 (Ross et al. 2010). Nonetheless, these communities can have problems with poor work quality just as there are in offline workplaces. For example, with low payment and a lack of face-to-face contact, job-posters have difficulty motivating and monitoring workers, often finding that users put in very little time and effort to earn the payment (Kittur et al. 2008).

With widespread gaming behavior, requesters who post jobs become worried about the quality of work and are more likely to withhold payment. Since it is difficult to control for job characteristics like the worker's environment, and because many online labor markets do not provide extensive support for giving feedback, there is additional ambiguity when requesters refuse a worker’s submission. This proves to be frustrating for workers, who do not understand the cause of rejection (Irani 2009). Motivating workers to work harder on an online marketplace for work is important to help provide a better experience for both requesters, who can obtain better results, and workers, who can be more consistently paid for their work.

More recent research on crowdsourcing has been largely focused on the issue of quality in these environments. Workers on sites like MTurk generally have no expertise in the area of the specific tasks (Callison-Burch and Dredze 2010), which causes great concern about the quality of the resulting work. This leads to interesting ideas about how to design tasks to better control for quality. For example, task results could be aggregated independently or workers could revise the work done by others (Little et al. 2010). More iterations of revisions generally improve the quality of work, but this is not a reliable method and could become costly as those requesting work must pay for each iteration.

Another approach to controlling for work quality is to weed out workers who are likely to game the system, reducing the need to aggregate results from a large number of workers (Ipeirotis et al. 2010). Various algorithms have been proposed to predict the quality of work that a particular worker might produce, so that those workers with low quality could be blocked from participation in the task (Huang et al. 2010). This is often used in conjunction with other methods, like well-designed qualification tasks, that filter out the workers who are expected to be poor performers (Downs et al. 2010, Akkaya et al. 2010). Rashtchian et al. (2010) found that prescreening methods are not only less-costly but also more effective in general. However, filtering workers is not always possible in anonymous communities or may not be desirable for tasks requiring more diversity, so it is important to first ensure good design of tasks to encourage high quality work.

Traditionally, job performance may improve through the use of interpersonal contact with beneficiaries, which increases perceived job meaningfulness and social impact (Grant et al. 2007). However, this is difficult to accomplish in the online context where

contact is minimal. In an overview of work design literature, Morgeson and Campion (2003) describe the conceptual background of work design elements and structure, including the Job Characteristics Theory presented by Hackman and Oldham (1976). This approach suggests five job characteristics that are likely to affect psychological states and work outcomes: 1) skill variety, 2) task identity, 3) task significance, 4) autonomy, and 5) feedback from the job. Of these five attributes, task significance is the only factor not dictated by the job itself or the specific platform. Thus, to manage quality in online marketplaces for work, task significance best lends itself to immediate improvement. This study examines ways in which task design, and especially task significance, can be used to encourage better quality contributions in online marketplaces for work like MTurk. Through discussion of how work quality may be improved, this study contributes to the literature on preventing negative gaming behaviors and promoting positive contributions to online marketplaces for work.

III. STUDY 2: Online Volunteer Groups

For online volunteer groups like Wikipedia, the free online encyclopedia, dedicated members who actively contribute to community efforts are crucial to their success. Though some negative behaviors, such as vandalism, have been studied extensively within this context, lack of participation or of contributors can cause sustainability issues and remains a major problem (Butler 2001) while high membership turnover makes knowledge retention for collaboration difficult (Ransbotham and Kane 2011). Despite its success, Wikipedia still faces challenges in recruiting and retaining active contributors, but with

unique problems. Unlike offline organizations, online groups have relatively low exit barriers, and online volunteer groups in particular have no formal processes to train new members or ensure participation. Thus, a lack of contribution and withdrawal behaviors are the major problems within these groups.

Prior research on online volunteer groups has often focused on motivating members to contribute and on the ways in which members coordinate and resolve conflict (Cosley et al. 2006, Kittur and Kraut 2010, Kittur et al. 2007a). In addition to continued contributions, volunteer groups like Wikipedia also require members to remain. Though withdrawal from volunteer groups may not be intended to harm the group, it nevertheless can create stagnation in the collaborative process, especially if new members are unable to be recruited. Understanding the causes of withdrawal can help to prevent volunteers from leaving prematurely.

Organizational science literature on offline organizations provide some insights into reasons members of a group may withdraw. For example, similar to their offline counterparts, online volunteers may suffer from burnout and feel unable to continue (Wilson 2000). Unlike many offline organizations with formal processes, online groups and projects may be unable to recover from the sudden withdrawal of a member, and a new volunteer may lack the knowledge to pick up the pieces (Robles et al. 2005). Although sites like Wikipedia encourage productivity and attempt to reduce withdrawal, there may also be a trade-off between these two goals. Some organizational research shows that larger groups may correlate with more "free-riding" among members who receive credit for work without putting in equal effort (Albanese and Van Fleet 1985). Though it may appear that

members are less likely to withdraw from these groups, not all members are productive. Furthermore, due to burnout from high productivity, some core members may subsequently decrease contributions, effectively withdrawing (Cordes and Dougherty 1993).

Wikipedia, in particular, also has a subgroup structure such that members may focus on individual projects and work in more cohesive groups. Studies have also shown a trade-off between issues within a group and within an organization as a whole. For instance, maintaining membership with multiple projects may limit the amount of time a member can contribute to any one of them (Barron et al. 1994, Becker 1965). Meanwhile, though having more communication within a subgroup generally increases contributions to the subgroup (O'Reilly et al. 1989), having communications outside of the subgroup may either improve contributions by bringing in new ideas and information (Cross and Cummings 2004) or present more opportunities for members to leave the subgroup (McPherson et al. 1992).

This study examines the major problems of membership withdrawal and a lack of contribution to online volunteer groups. Traditionally, productivity and contribution can be measured by work output. In the context of Wikipedia, productivity may be measured by the number of edits a user makes (Kittur and Kraut 2008, Kittur and Kraut 2010, Suh et al. 2009). Additionally, in traditional organizational research, the term "withdrawal" includes behaviors like "psychological withdrawal, lateness, absenteeism, and turnover" (Beehr and Gupta 1978). In Wikipedia, withdrawal may simply be indicated by a lack of contributions from a previously active user or a formal withdrawal from a subgroup. In order to understand how these negative behaviors might be managed, this study examines

potential trade-offs between productivity and withdrawal in online volunteer groups like Wikipedia. Due to the struggles between retaining productive members in projects and contributions to the organization that have been investigated in traditional organizational literature, this study also examines trade-offs between participation in subgroups versus the overall site in the online context. Thus, this study contributes to the literature on preventing withdrawal and promoting positive contributions to online volunteer groups.

IV. STUDY 3: Social Networking Communities

Finally, in social communities, intentional deviant behaviors are most likely to create disturbances that negatively affect the experiences of other community members, often undermining existing social norms. Research has shown that users, and particularly older users, are less tolerant of profanity and incendiary messages, meaning sites have a harder time attracting adults when such behavior is present (Wilson 2007). There are other problems with online deviance as well. For instance, members of an online network may be influenced by the actions and opinions of other members during the course of social interactions (Han and Kim 2008). Thus, members of a group who exhibit antisocial behaviors may influence other members who are otherwise non-deviant to follow the same behaviors, shifting the group's norms (Robinson and O'Leary-Kelly 1998). Due to the lack of physical barriers, people who may find themselves on the fringe of offline communities may gather online for a social outlet (Adler and Adler 2008). However, this also increases chances of deviants congregating in the online arena (Evans 2011).

Psychology and sociology literature provide many theories regarding deviant behaviors in offline contexts that help to understand online deviance. A social learning approach posits that deviant behaviors, like other behaviors, are learned through differential reinforcement—positive and negative rewards or punishments (Akers et al. 1979). On the other hand, some studies have shown that certain people are predisposed to deviant behaviors. For instance, children with more "callous-unemotional" traits are more likely to develop deviant behaviors, due to their aggressive and antisocial nature (Frick and White 2008). In contrast, moral violation theories suggest that deviant behaviors are often not habitual actions but rather retaliation as a response to perceived injustice (Mullen and Nadler 2008). Like their offline counterparts, in order to help prevent the alienation of rule-abiding users, online community administrators need to moderate the community's activity. However, because there is a degree of anonymity and a lack of accountability online, people who might otherwise not behave in negative ways are more likely to act out online (Doig 2008). Thus, it may be more difficult to regulate online deviance than in traditional communities.

One particularly problematic and prevalent deviant behavior in online social communities is the act of trolling, or intentionally disrupting an online community with incendiary or off-topic messages that are meant to provoke others (Phillips 2015). Trolling is especially hard to identify due to the difficulties in determining a user's intent, and moderators need to be careful not to ban legitimate members on suspicion of trolling. Furthermore, past research has highlighted additional ambiguity in the impact of trolling. For example, some types of trolling may actually be useful in negotiating social boundaries

within a community or increase cohesion through mischief (Kirman et al. 2012, Hardaker 2010). It is, therefore, important to distinguish between harmful trolling behaviors with clearly negative impact and non-harmful behaviors.

This study examines trolling in an online social networking community, deviantART.com, because of the aforementioned prevalence of such behaviors in these types of communities. In order to maintain balance and encourage participation, these communities need to prevent the negative emotional responses that trolls attempt to evoke. Specifically, this study attempts to examine how users identify a troll, under what circumstances it happens, and what can be done to help moderators recognize and resolve it, particularly in the case of discussions that are especially volatile and controversial. Because uncommon or controversial opinions may be mistaken for trolling (Kelly et al. 2006), this study focuses on ways in which a community can more accurately identify trolling and distinguish it from less disruptive behaviors. By helping to more clearly identify trolling and understanding the impacts of this behavior, this study contributes to the literature on preventing negative trolling behaviors and promoting positive contributions to online social networking communities.

V. GENERAL GOALS

Although many of the behaviors discussed in this dissertation have offline counterparts that have been extensively studied in the past, there are many differences due to the technology of online organizations that changes how people behave. Online marketplaces for work are different from traditional job environments where there is more oversight of workers.

Online volunteer groups do not have the same formal processes for members to withdraw or join. Social networking communities have increased anonymity that causes many users to behave differently than they would offline. Although, traditional literature helps us to understand the core motivations behind problematic behaviors, these differences require additional research in order to understand the same behaviors online. This dissertation attempts to contribute to theory and practice by illuminating ways to identify and moderate several negative online behaviors that have not been extensively studied in these particular contexts.

CHAPTER 2

Better Than an Automaton:

The Significance of Task Significance in Online Marketplaces for Work[1]

I. INTRODUCTION

Crowdsourcing (Howe 2006) is a new approach, and a rising trend, to getting work done by posting problems or work on the Internet in an open-call format and allowing anyone to submit solutions or work products. The types of tasks that can be accomplished range from creative ones like designing graphics at 99designs.com and solving scientific problems or generating open innovation ideas at InnoCentive.com (Lakhani and Boudreau 2009) to more specialized tasks like developing software at Rent A Coder (Gefen and Carmel 2008) to mundane tasks like annotating images and videos, writing product reviews, and tracking migration patterns of animal species (Howe 2008). While many of these platforms are still newly emerging, they have the potential to dramatically transform how work is organized in today's society. For instance, one crowdsourcing platform, oDesk, features success stories of companies who build or expand whole departments on the platform, whether it is customer support, web development, multimedia design, or administrative support. Companies are able to solve global and local problems at competitive rates, while still maintaining the flexibility of having an on-demand workforce.

Online marketplaces for work like oDesk and Amazon Mechanical Turk (MTurk hereafter) are a special type of crowdsourcing platform that focuses on simple, mundane

[1] In collaboration with Yuqing Ren, with guidance from Shawn Curley.

tasks that can be performed effectively by humans but not by computers (e.g., tagging an image). Amazon refers to these tasks as Human Intelligence Tasks, or HITs. In recent years, they have served as an increasingly popular and powerful way of getting work done, cheaply and quickly. Businesses use these platforms to perform search relevance evaluations, verify websites, and clean data sources (Feng et al. 2009). Researchers use the platforms for user studies, image labeling, natural language processing, and replicating classic economic and social science experiments (Kittur et al. 2008, Ross et al. 2010). There is continued effort to explore novel ways of leveraging the platform to, for instance, identify tumor cells in medical images (Chandler and Kapelner 2012), collect large-scale review data (Su et al. 2007), and generate question-answer pairs (Kaisser and Lowe 2008). One promising direction is human knowledge acquisition (Feng et al. 2009), that is, using the on-demand workforce to generate expert knowledge to train machine learning algorithms for both industrial applications and academic research. Many machine learning and natural language processing tasks require large amounts of manually annotated data, known as golden standards or expert ratings. Traditionally, these data have been collected from domain experts. The process is very expensive, laborious, and time consuming, with some efforts being estimated to take 16 person-years (Akkaya et al. 2010). Researchers have found these new platforms, MTurk in particular, to be a fast and cost effective way of collecting linguistic annotation for a variety of natural language tasks (Snow et al. 2008).

Meanwhile, there are clear risks in "outsourcing to the crowd," particularly to an unknown, faceless crowd. The lack of face-to-face contact and control over how online workers approach the work and how much time and effort they exert leads to high

uncertainty in work quality. There is clear, consistent evidence of quality issues and gaming behaviors. For example, Kittur et al. (2008) asked MTurk workers to rate Wikipedia articles and found only a moderate correlation between worker and Wikipedia admin ratings. Further analysis showed gaming behaviors in both the amount of time workers spent and the informativeness of their comments. Similar patterns have been reported in other studies. A study of foreign language translation found that prolific MTurk workers performed barely above a 1/3 chance of agreement with experts (Callison-Burch 2009). In another study where participants answered two questions about an email message detailing an upcoming teleconference, only 61% of workers answered both questions correctly (Downs et al. 2010). In yet another study, workers were asked to write one descriptive sentence for each of ten images (Rashtchian et al. 2010). On average, workers spent four minutes writing the ten sentences. Only half of the 5000 sentences were deemed acceptable, and the rest of the descriptions either missed salient entities in the images or were written in poor English. Furthermore, 2.5% of the 5000 sentences were empty strings, a clear sign of gaming behaviors by exerting no effort to earn the money.

A couple of factors might have contributed to the work quality issues, including lack of motivation due to low monetary payment and lack of knowledge or skills required to perform the task well. Although Amazon suggests payments should follow a reasonable hourly rate similar to the minimum wage in the US[2] ($8 per hour), in practice, MTurk workers were paid much less. Many tasks pay a few cents and can take several minutes to translate paragraphs, watch a video and provide feedback, summarize a website, describe

[2] Amazon Mechanical Turk: Best Practice Guide. http://mturkpublic.s3.amazonaws.com/docs/MTURK_BP.pdf.

a creative idea, or write product reviews (Downs et al. 2010). Several studies have found that MTurk workers are paid an average hourly rate of $1 to $3 (Ross et al. 2010). In the image description task we described earlier, workers were paid an hourly rate of $1.30 (Rashtchian et al. 2010). Low payment leads to low work morale and, inevitably, minimal effort to do the work. In some cases, the work quality of MTurk workers has been shown to be even lower than the work of pure volunteers (Downs et al. 2010).

Due to the severity and ubiquity of the issue, Amazon Mechanical Turk has introduced several mechanisms to assure work quality and discourage gaming behaviors. First, job requesters can refuse to pay the workers if they are unsatisfied with their work quality. This is not a very effective mechanism, because the cost and time of verifying work quality can be comparable to the cost and time of performing the task (Ipeirotis et al. 2010). Second, job requesters can use qualification tests or screening questions to pre-filter workers and only display the task to qualified workers. One such qualification is approval ratings, which record the percentage of a worker's accepted HITs. Other qualifications include country of residence or lifetime approved number of HITs. These qualifications are somewhat effective yet do not fully address the quality issue. For instance, in the study where workers were paid to write descriptions of ten images, the researchers restricted the HIT to workers who had a 95% approval rating. Yet half of the submissions still did not meet standards, and 2.5% were empty strings (Rashtchian et al. 2010). Another limitation of these existing measures is that they are not good at separating negligence or errors from individual bias (Ipeirotis et al. 2010).

In this study, we examine a new approach to improve work quality and reduce gaming behaviors in online marketplaces for work by informing workers of the purpose of the task and who benefits from it. The concept of task significance originates from the job motivation and design literature (Frey and Osterloh 2001, Hackman and Oldham 1980) and has been shown to lead to greater motivation and improved performance in offline settings through intrinsically motivating workers. We conducted a laboratory experiment and a field experiment using Amazon Mechanical Turk. In the laboratory experiment, participants, mostly college students, performed spell checking tasks to fix errors in Wikipedia articles to assure their accuracy and quality. In the field experiment, MTurk workers performed spell checking tasks to fix errors in digitized books and were told that these books would be made freely available to underprivileged people. Task significance improved work quality in both experiments, but only when the information properly registered with the worker. A majority of participants who received the purpose statement ignored it, shedding light on the challenge of promoting task significance in the online context.

We also experimented with three formats of delivering the purpose statement (text, female-narrated video, and male-narrated video) and found that format had no significant effects on either recall of the purpose statement or work quality. Further analysis showed new workers, workers who have heard of e-books, workers with more than minimal income, and workers with certain personality traits such as agreeableness or introversion were more likely to recall the purpose statement than other workers. Compared to task significance, increasing monetary payment by 50% or highlighting workers' self-benefits

had no significant effects on work quality. Through post hoc analysis, we also found that workers who described the HIT as being fun and enjoyable had higher work quality than those who did not. In addition, our findings suggest that different types of motivation were not independent or mutually exclusive. Increased payment and enjoyment both led to a greater level of perceived task significance.

II. THEORY AND RESEARCH QUESTIONS

Task significance is part of the Job Characteristics Theory presented by Hackman and Oldham (1976). They suggested five job characteristics that are likely to affect psychological states and work outcomes: (1) skill variety – whether a job requires different skills, (2) task identity – whether the outcome of a job is visible from beginning to end, (3) task significance – whether the worker considers the job to be important, (4) autonomy afforded the worker, and (5) feedback from the job. Meta-analyses of the work motivation literature show that many job characteristics, like autonomy, task significance, and feedback, help to increase performance and job satisfaction (Humphrey et al. 2007). In traditional work environments, interpersonal contact with beneficiaries can greatly improve perceived job meaningfulness and social impact, which in turn improves performance (Grant et al. 2007). On crowdsourcing platforms, it is more difficult to control for job characteristics like the worker's environment, and many online labor markets do not provide extensive support for giving feedback. Of the five job characteristics, task significance is the only factor not directly limited by the job itself or the specific platform

on which a task might be posted, making it an ideal factor to examine in a crowdsourcing context.

2.1 Task Significance

Task significance refers to the degree to which the job provides opportunities to positively impact the well-being of others, whether they are in the immediate organization or the world at large (Hackman and Oldham 1980). Research has shown that task significance cues increase the job dedication and performance of workers in various jobs, from fundraisers to lifeguards, partly because they experience the work as meaningful and having impact on others and partly because the job connects workers to other people, which makes it social or relational (Grant 2008b). Pro-social motivation, which measures the degree to which people are motivated to help others, not only explains increased performance but also has been shown to increase persistence over time and mediate intrinsic motivations (Grant 2008a).

There has been at least one study that examines the impact of meaningfulness in online marketplaces for work (Chandler and Kapelner 2012). Amazon Mechanical Turk workers were asked to label objects in images. Those in the meaningful treatment condition were told that they were labeling tumor cells in order to assist medical researchers. In contrast, those in the control condition were simply told to label objects of interest. Workers in both conditions watched a 3-minute video to learn how to perform the task. The meaningfulness treatment increased workers' likelihood of participating in the HIT. In the meaningful condition, 80.6% of workers labeled at least one image, compared to 76.5%

in the control condition. However, the meaningfulness treatment had no significant impact on work quality, which is puzzling. Hence, our first research question is:

Does task significance lead to higher work quality in online marketplaces for work?

2.2 Extrinsic Motivation and Monetary Payment

An alternative approach we explore to increase work quality is monetary reward. Although performance-based incentive is a common motivational tool in organizations, research provides mixed support of monetary incentives, especially in improving work quality. A meta-analysis of over 39 studies shows that financial incentives are positively and moderately related to effort and performance quantity but not performance quality (Jenkins et al. 1998). Other research in education, sports, and work settings has provided conflicting results on how extrinsic motivation like money could affect the intrinsic value of the work itself. A meta-analysis by Cameron and Pierce (1994), for instance, shows no effect of reinforcement or reward on intrinsic motivation. However, both their analysis and conclusion have been criticized as flawed and overly simplistic, while others have shown that tangible rewards do indeed have a substantial undermining effect on intrinsic motivation (Deci et al. 1999). On the other hand, some research has indicated an additive effect of intrinsic and extrinsic rewards on worker motivation to spend more time working on a task (Wiersma 1992), though this does not necessarily result in higher quality work.

This interplay is even more sophisticated in online marketplaces. To some extent, crowdsourcing inherits the openness, playfulness, and voluntary spirit of peer production and the open source movement (Lakhani and Wolf 2005). However, adding money to the

equation runs the danger of crowding out other types of motivations. Exploratory research of Amazon Mechanical Turk shows that continual increase of payment does not always translate to commensurate increase in work quality (Feng et al. 2009). The researchers asked MTurk workers to label 10 web queries to be paid $0.01, $0.02, $0.05, or $0.10 per task. As payments increased, hourly rates increased from approximately $0.72 to $9.73 and turnaround time dramatically dropped from 2 days to about 1.5 hours. Workers worked on more HITs yet spent less time on each HIT, and although increasing payment from $0.01 to $0.05 increased quality from 81% to 93%, further increase from $0.05 to $0.10 slightly decreased quality to 90%. High monetary reward seemed to have motivated workers to complete the task as quickly as possible to get paid, instead of submitting high-quality work. Hence, our second research question is:

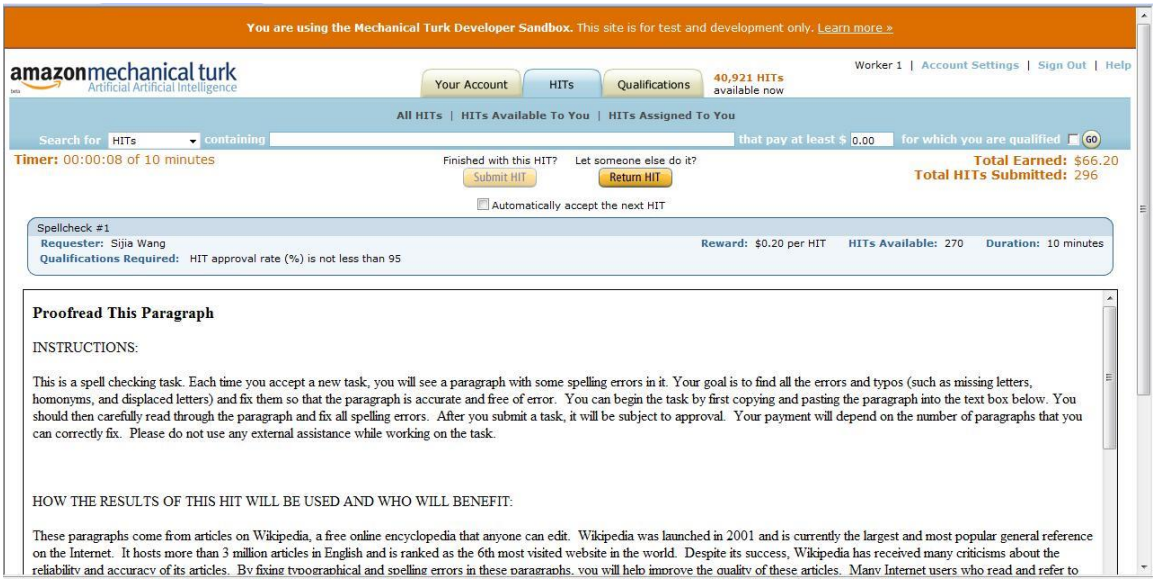
Does monetary payment lead to higher work quality in online marketplaces for work?

III. LABORATORY EXPERIMENT

We first ran a laboratory experiment to answer these questions. We used mostly college students as our participants, although they performed the tasks on the Amazon Mechanical Turk platform, similar to MTurk workers. MTurk is a crowdsourcing platform launched by Amazon in 2005. There are two types of roles: requesters and workers. Requesters can create Human Intelligence Tasks and specify how many submissions they want on each HIT and how much they will pay for each completed HIT. Workers can search and choose

to work on these HITs and submit their work to get paid. According to one data source[3], there are about 200,000 HITs at any given time, and on average, requesters pay $30,000 to $40,000 in rewards per day. We posted our task as a proofreading task on Amazon Mechanical Turk. Figure 1 shows a screenshot of what the HIT looked like to the participants.

Figure 1. Amazon Mechanical Turk HIT Screenshot



3.1 Experimental Design

3.1.1 Participants

We recruited 163 participants (100 female, 63 male) through a behavioral research lab in a large Midwest university. We posted the research as a study that examines how people perform spell checking tasks in an online environment and made the opportunity accessible

[3] http://mturk-tracker.com/general/

only to those who indicated they were a native English speaker. The majority of the participants were undergraduate (n = 120) and graduate (n = 30) students who attended the university. The average age of the participants was 23.21 (SD = 7.33).

3.1.2 Procedure

Each experimental session lasted an hour with a maximum of six participants scheduled per session. The laboratory was equipped with six computers in a row, all with Internet connection. The computers were separated by dividers so that participants were in the same room but could not see each other's screens. Two experimenters collectively conducted the experiments – one was responsible for running the experiment (primary) and the other helped with logistics and payment calculation in the end (secondary). As participants arrived, the primary experimenter greeted them, handed them the consent form, and seated them at a computer in the order of their arrival. Conditions had been previously randomized and assigned to the computers.

After all participants had arrived and signed the consent forms, the primary experimenter began the experiment by introducing herself and giving a brief description of the spell checking task. The experimenter then asked participants to complete a training task by following step-by-step instructions on paper to learn how to accept and perform the spell checking tasks in the Amazon Mechanical Turk sandbox. The sandbox replicates the real environment of MTurk and does not require payment to the participants through MTurk. The purpose of the training task was to familiarize the participants with the interface so that they understood all the steps needed to complete a HIT. As participants

worked on the training task, the experimenter walked around to make sure they were following the instructions properly.

After the training task, participants were informed that they had 30 minutes to work on as many paragraphs as they wanted and that their payment was contingent upon the number of paragraphs they correctly fixed. As participants submitted paragraphs, the second experimenter performed quick checks in real time to calculate payment by excluding paragraphs with too many unfixed errors. This is similar to the review process that requesters undertake to approve or reject work submitted by workers on MTurk, except that MTurk workers might not be paid immediately after task completion.

In the meantime, we asked participants to perform a distraction task on paper to increase external validity. Workers in online marketplaces for work rarely solely concentrate on the tasks at hand. Many people work on the tasks in between other things in their lives, such as a full-time or part-time job, student work, or household chores. The distraction task consisted of standard questions taken from Wonderlic tests (http://www.wonderlic.com/). This is similar to an IQ test that is often used to assess the aptitude of prospective employees for learning and problem solving in various occupations. The task also served as a way to assess a participant's cognitive ability. The task consisted of 30 questions about nouns, verbs, reading vocabulary, estimation, and numerical sequences. Participants were told that it was up to them how they wanted to allocate time between this task and the spell checking task as long as they answered all questions by the end of the 30 minutes. At the end of the 30 minutes, participants completed a post-questionnaire that asked about their experience of working on the spell checking tasks as

well as about some demographic information. As they worked on the questionnaire, the second experimenter calculated their payment. Their total payment included $3 for participation, $2 for completing the task on paper, plus the amount they earned in performing the spell checking tasks (ranging from $4 to $7). Participants were then paid and debriefed.

3.1.3 Experimental Task

We created and administered the spell checking tasks at Amazon Mechanical Turk. MTurk provides templates to post a broad range of HITs such as data collection (e.g., search for phone numbers of restaurants), data correction (e.g., check the spelling of search terms), image tagging, and filtering. Although we did not use a template, our tasks were structured in a similar way. We chose spell checking in the first experiment because (1) it is representative of MTurk tasks in terms of the nature of the task and the time demand and (2) the quality of the work can be objectively assessed by counting the number of errors fixed.

In each task, we presented a participant with a 100-150 word paragraph and asked them to find and correct all typographical errors in the paragraph. The paragraphs were taken from featured articles at Wikipedia, the free online encyclopedia. We collected 270 paragraphs altogether from three areas: business, sports, and music. We introduced a random number of errors, ranging from three to seven, into each paragraph. These errors included a variety of types, such as switched letters, added or removed letters, and incorrect use of homophones. We shuffled articles in different areas so that each participant saw a

mix of articles from all three areas. Here is an example paragraph, with five sample errors shown in quotations for clarity:

> In spite of popular belief, actuaries do not always "**atempt**" to predict aggregate future events. Often "**there**" work may relate to determining the cost of financial liabilities that have already "**ocurred**", called retrospective reinsurance, or the development or re-pricing of new products. Actuaries also design and maintain products and systems. They are involved in "**financal**" reporting of companies' assets and liabilities. They must communicate complex concepts to clients who may not share their language or depth of knowledge. Actuaries work under a strict code of ethics that covers their communications and work products, but their clients may not adhere to those same standards when "**enterpreting**" the data or using it within different kinds of businesses.

On top of the task page, participants saw instructions telling them that their goal was "to find all errors and typos (such as missing letters, homophones, and displaced letters) and fix them so that the paragraph is accurate and free of error." They were instructed to copy and paste the paragraph to a text box and read carefully to fix all errors. They were also informed that the paragraphs they submitted would be subject to approval and that their payment depended on the number of paragraphs that they correctly fixed. They were asked not to use any external assistance while working on the task (all participants followed this instruction except one).

3.1.4 Manipulations

We manipulated two factors: task significance and monetary payment. The experiment is between-subjects, and we randomly assigned participants into 4 conditions in a two-by-two design.

Participants in the task significance condition saw a paragraph titled "How the results of this HIT will be used and who will benefit" below the general instructions. The paragraph explained that their work through fixing errors in the paragraphs would improve the quality of Wikipedia articles and therefore benefit millions of Internet users who refer to Wikipedia. In comparison, participants in the control condition did not see the following task significance paragraph:

> These paragraphs come from articles on Wikipedia, a free online encyclopedia that anyone can edit. Wikipedia was launched in 2001 and is currently the largest and most popular general reference on the Internet. It hosts more than 3 million articles in English and is ranked as the 6th most visited website in the world. Despite its success, Wikipedia has received many criticisms about the reliability and accuracy of its articles. By fixing typographical and spelling errors in these paragraphs, you will help improve the quality of these articles. Many Internet users who read and refer to these articles will benefit from your work.

Participants in the *low* payment condition were told that they would earn $0.20 for each correctly fixed paragraph whereas participants in the *high* payment condition were told that they would earn $0.30 for each correctly fixed paragraph. The payments were higher than average MTurk payments because our participants were college students.

3.1.5 Work Quality Measures[4]

We measured work quality by the percentage of fixed errors. If a paragraph had five randomly introduced errors and a participant fixed three, the quality of the work based on accuracy is 60%. There were altogether 2408 paragraphs submitted. We wrote a Perl script to automatically calculate accuracy. In the meantime, 1055 paragraphs were manually coded for accuracy. The correlation between the manual coding and automatic coding was over 78%. We analyzed and reported our findings based on the automatic coding.

3.2 Results

3.2.1 Manipulation Checks

We analyzed individual-level responses to the post-questionnaire to check our manipulations of task significance and monetary payment. To check task significance, we asked participants their "best knowledge or best guess of how the fixed paragraphs will be used to benefit others." Surprisingly, the majority of the participants who had seen the purpose statement were not able to recall how the fixed paragraphs would be used to benefit others. Only 28.4%, or 23 out of 81 participants, mentioned improving article quality at Wikipedia. As expected, most of the participants who did not see the task significance paragraph answered "don't know" or "have no idea". We also performed manipulation checks of payment, and 95.8% of all participants correctly recalled their payment condition.

[4] We tracked two additional performance measures. We measured task quantity by counting the number of paragraphs that a participant submitted, excluding the training paragraph(s). We also tracked how much time a participant on average spent on fixing a paragraph. Neither task significance nor monetary payment had any significant effects on these measures.

3.2.2 Work Quality

Table 1. Effects of Task Significance and Monetary Payment on Work Quality

	Task Accuracy	
	Low	High
Task Significance	0.65*	0.60*
	(0.15)	(0.20)
Monetary Payment	0.62	0.62
	(0.19)	(0.16)

Note: * significant at $p < .05$ level

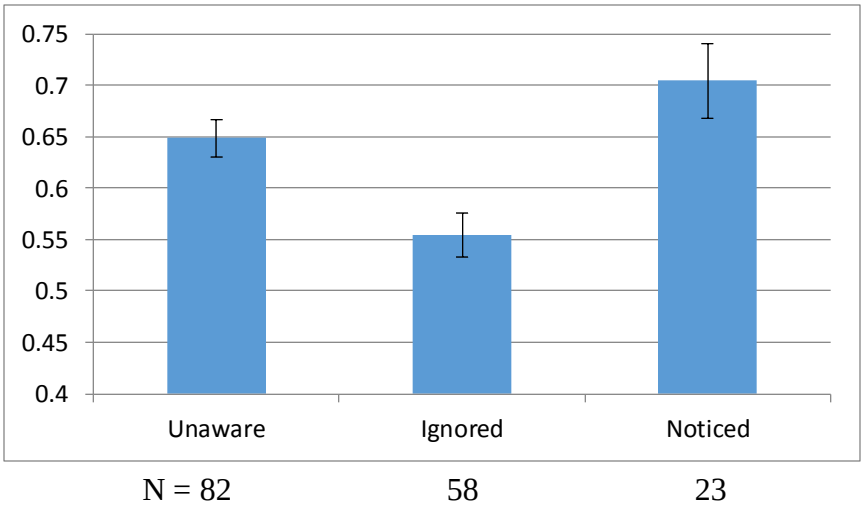
The analysis showed a significant effect of our task significance manipulation on work quality, $F(1, 169) = 4.03$, $p < 0.05$, but in the opposite direction of what we had expected. As shown in Table 1, task significance, or seeing the purpose statement of how the work would benefit others, led to lower instead of higher accuracy (60% compared to 65%). In addition, the analysis showed no significant effects of increased monetary payment, $F(1, 169) = 0.02$, $p = 0.89$.

3.2.3 Additional Analysis of the Effects of Task Significance

As mentioned earlier, our manipulation of task significance was partially effective. Only 23 out of 81 participants who received the purpose statement correctly recalled how their work output would be used to benefit others. We ran another set of ANOVA tests with three levels. Participants who didn't receive the manipulation were labeled the "*unaware*" group; participants who received the manipulation, yet chose to ignore it or forgot it, were labeled the "*ignored*" group; participants who received and correctly recalled the manipulation were labeled the "*noticed*" group. The analysis, again, showed a significant

main effect of task significance on work quality, $F(1, 161) = 8.59$, $p < 0.001$. As illustrated in Figure 2, the *noticed* group had the highest level of accuracy of 0.70, followed by the *unaware* group, 0.65, and then the *ignored* group, 0.55. The differences between the *ignored* group and the other groups were significant at the $p < 0.01$ level. The difference between the *noticed* and *unaware* groups was not significant, probably due to the small sample size of the noticed group (N = 23). Analysis of time taken to perform the tasks showed that the *ignored* group took longer to perform the tasks (149.33 seconds per task), significantly longer than the *noticed* group (127.48 seconds per task) but not significantly longer than the *unaware* group (136.64 second per task).

Figure 2. Effects of the Recall of Purpose Statement on Work Quality



3.2.4 Post-hoc Analysis

In the post-questionnaire, we also asked participants about their intrinsic and extrinsic motivations to perform the task, together with their perceptions of task difficulty on a five-point Likert scale. We ran exploratory factor analyses with oblique rotation. Several items

– one assessing intrinsic motivation ("the spell check task was boring") and one assessing extrinsic motivation ("I would perform the spell checking tasks even without payment") – loaded on more than one factor and were dropped from the analysis. Table 2 shows the loading patterns of the remaining factors. We took the average of each set of three items to measure intrinsic motivation and task difficulty.

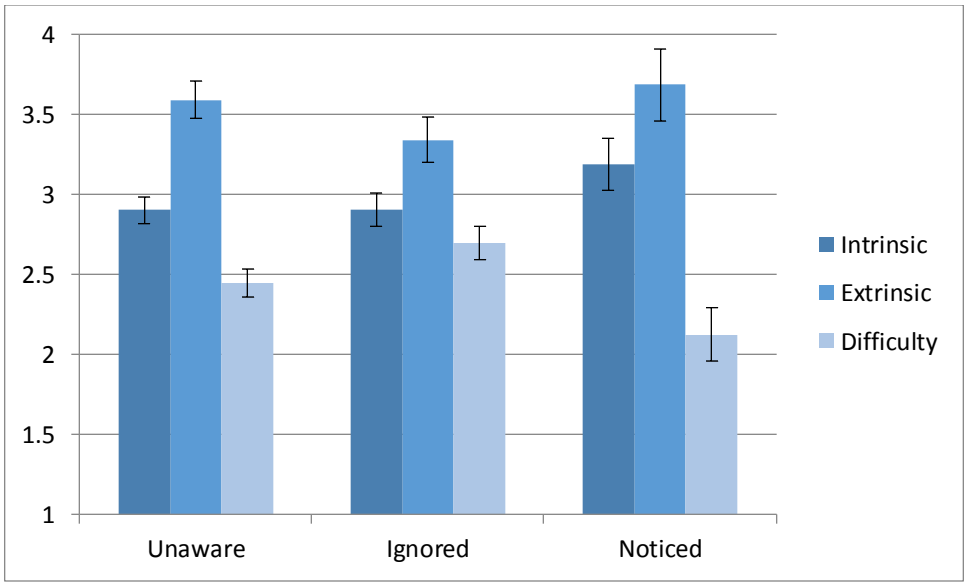
Table 2. Items on Intrinsic Motivation, Task Difficulty, and Extrinsic Motivation

F1	F2	F3	Questionnaire Items
61	-23	-28	I enjoyed performing the spell checking task.
80	6	25	I would return to MTURK to perform more of the spell checking task.
42	21	-12	I would perform the spell checking task even without the payment.
5	**65**	18	I got overwhelmed by the number of words in the spell checking paragraph.
-5	**86**	-3	The spell checking task is difficult.
8	**83**	-19	The spell checking task is challenging.
1	-1	**43**	If I got paid more on the spell checking task, I would work harder on it.

We ran another set of ANOVA tests to check the effects of task significance and monetary payment on motivation and task difficulty. Analysis of intrinsic and extrinsic motivation showed no significant effects. Paired t-tests suggested, however, that the *noticed* group reported a higher level of intrinsic motivation than the *unaware* group (p = 0.08) and a higher level of extrinsic motivation than the *ignored* group (p = 0.07). Again, it is possible that we would observe more significant effects with a larger sample size of the *noticed* group. Analysis of task difficulty showed that participants in the *noticed* group

reported significantly lower levels of task difficulty (2.13) than participants in the *ignored* group (2.7) and participants in the *unaware* group (2.45), as shown in Figure 3.

Figure 3. Effects of Task Significance on Motivations and Task Difficulty



3.3 Discussion of Lab Experiment Results

The lab experiment provides initial support for the importance of task significance in online marketplaces for work. It also highlights the challenge of communicating task significance messages in the online context. Although the Amazon Mechanical Turk Requester Best Practice Guide recommends that requesters inform their workers of the purpose of the work, our results suggest that this practice may need to be implemented carefully. Simply including a paragraph about how the work would be used to benefit others may not increase work quality. For participants who read and internalized such a purpose statement, our study shows positive effects (higher quality, higher intrinsic motivation, lower perceived task difficulty). For participants who ignored or forgot the statement, the additional

information created additional work (such as more scrolling to get to the task) which may have led to the lower work quality, lower level of intrinsic motivation, and higher level of perceived task difficulty present in our results. Furthermore, for those who noticed the purpose statement, the possible additional work did not cause a disruption since these participants were able to complete tasks faster than those who did not notice it.

Why did the majority of the participants ignore the purpose statement? What might have led to the individual differences in processing and recalling the purpose statement? We further analyzed demographic and cognitive ability data to see what predicts the divergent responses to the task significance manipulation. We calculated the number of questions completed and the percentage of correct answers in the written task, which acted as a proxy for cognitive ability. Participants in the *noticed* group answered 84% of the questions correctly, whereas participants in the *ignored* group answered 73% correctly (with 76% for the unaware group). Thus, cognitive ability seemed to be a predictor of whether participants noticed the purpose statement and performed better.

Another finding from the lab experiment is that monetary payments were not a significant predictor of work quality. It seems that being paid 20 cents per paragraph versus being paid 30 cents made no difference to our participants in the lab. However, research has shown that about 43% of MTurk workers are from lower income levels (less than $20,000 per year), and only a third are part-time or full-time students (Ross et al. 2010). Furthermore, many of the students in the study had no prior experience with these types of marketplaces for work. Therefore, college students may not be representative of typical MTurk workers in terms of how they respond to monetary rewards. Besides cognitive

ability, this study also did not account for other potential individual differences that might affect whether a participant notices the purpose statement. To address these issues, we ran a second study with real MTurk workers and used a modified version of the spell checking task to highlight task significance. Our third research question is:

What individual worker attributes lead to better recall of the purpose statement?

In addition, we explored rich media as a way to communicate the purpose statement and highlight task significance. Past research has shown that both the presentation of a message and the person receiving the message can influence the effectiveness of communicating the information (Kalyuga et al. 2000). In addition to intelligence and ability, cognitive styles and learning styles also affect individual information processing (Riding and Rayner 1998). For example, the VAK (or VARK) model suggests that use of multimedia such as visual or audio presentation and active engagement of the learners can help reach a more diverse audience (Hawk and Shah 2007). According to one source (http://www.studyingstyle.com/), approximately 65% of people are visual learners, 30% are audio learners, and the remaining 5% are kinesthetic (or tactile) learners. This suggests that a majority of the population can effectively digest information through either visual or auditory channels, and both might be necessary to effectively communicate a message online. Furthermore, media richness theory suggests that richer media is sometimes needed in cases where the task is more ambiguous, though there is also evidence that less ambiguous tasks can benefit from media that allows for more cues as well (Dennis and Kinney 1998). Hence, our fourth research question is:

Does rich media format lead to better recall of the purpose statement?

IV. FIELD EXPERIMENT

4.1 Experimental Design

The experimental design and task are similar to the lab experiment but with real MTurk workers being the participants. To increase the meaningfulness of the task, instead of Wikipedia articles, MTurk workers were told that the paragraphs are from out-of-print books that have been digitized and converted to electronic text using Optical Character Recognition, or OCR, software. They are further told that the software is not 100% accurate in its conversion, and consequently, typos and errors have been introduced and need to be fixed before the books can be made available to the public. Each HIT in this task contained only one paragraph to be proofread, and workers could complete as many as 15 HITs or stop whenever they wanted. After completing 8 HITs, and through the remaining HITs, workers were given the option to complete a survey for additional payment. The survey included questions about demographics and personality measurements like the Big Five Personality Factors (Barrick and Mount 2006).

4.1.1 Participants

Altogether, 888 MTurk workers participated in the field experiment by proofreading at least one paragraph, and 503 of them completed the survey. According to the survey, 59% of the survey respondents were male, about a third were 18-24 years old and another third 25-34 years old, and 76% had a college or graduate degree. Workers were almost evenly split between having full-time jobs, part-time jobs or being unemployed, and 78% of respondents had an individual income of less than $20,000 per year while 54% of

respondents had household incomes of less than $20,000 per year. In general, workers in our study are fairly representative of worker demographics at Amazon Mechanical Turk with a slightly higher percentage of males (Ross et al. 2010).

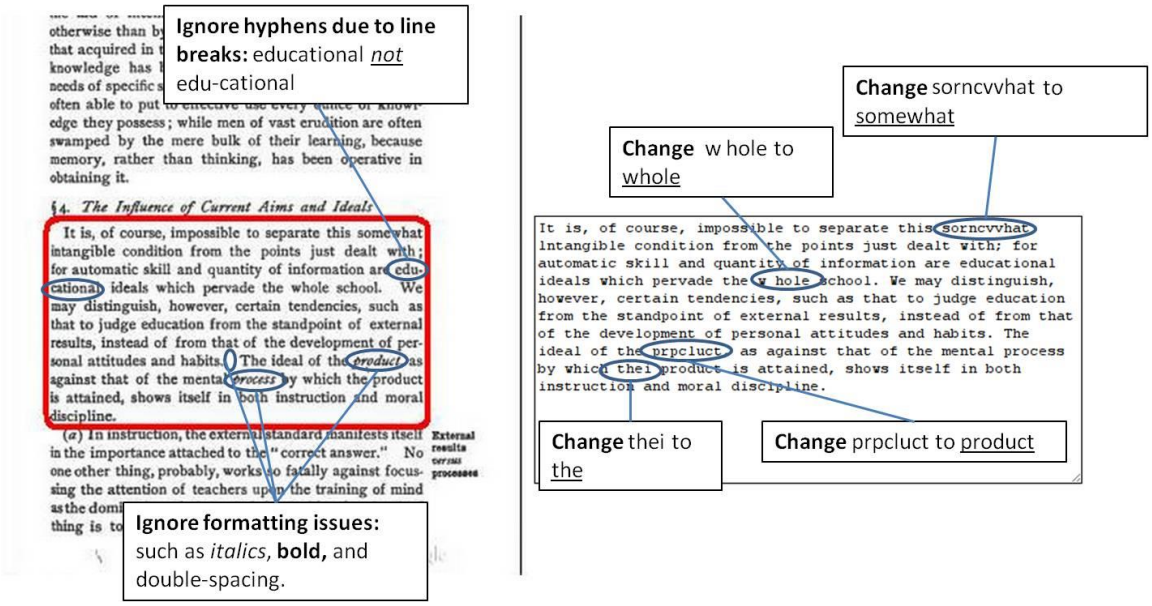
4.1.2 Procedure

We posted the HIT on Amazon Mechanical Turk where workers could read about the task, how to complete it, and the payment scheme. If a worker wanted to work on the HIT, they could click to accept it. As soon as workers accepted the HIT to begin working on it, a built-in link automatically re-directed them to a server hosted at a large Midwest university. Although the workers continued to view the Amazon Mechanical Turk website, the HIT presented to them within that frame was supplied by our server. We needed to host the HIT on our own server in order to randomly assign the workers to experimental conditions and avoid having one worker in multiple conditions. Each worker could proofread up to 15 paragraphs and could stop at any time.

The paragraphs were taken from Project Gutenberg (http://www.gutenberg.org/), a site that organizes volunteers to collectively edit digitized versions of old books that are no longer bound by copyright in order to create free e-books. We gathered pages from five books on diverse topics including public speaking, ornithology, and food preparation. The paragraphs were taken from scans of old books, many of which are blurry and hard to read. Figure 4 shows our instructions to workers with the scanned page on the left and the supposedly converted text (with errors) on the right. Workers were told the errors occurred because we used OCR to convert the scanned image to text, though in reality, we

introduced the errors ourselves. Their goal was to find and fix all the errors so that the text exactly matched the original text in the scanned image.

Figure 4. Instructions on How to Proofread a Paragraph



After completing 8 paragraphs, a link appeared under the paragraph, informing the workers that they had an option to take a short survey for additional payment. Workers were told they could stop at any point to work on the survey, or they could wait until all HITs had been completed. Of all workers, 612 completed 14 or more paragraphs and 362 subjects completed all 15 paragraphs. 550 subjects began the survey, and 503 completed the entire survey.

4.1.3 Manipulations

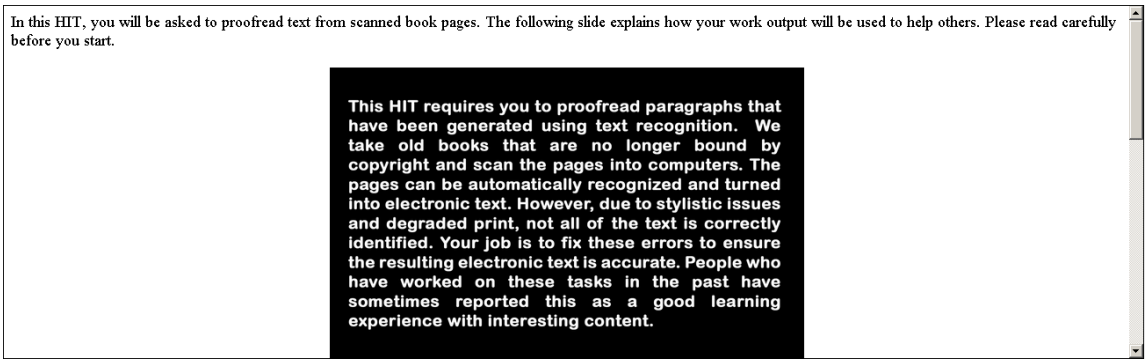
We manipulated three factors: task significance, media format, and monetary payment. There were three conditions of task significance: no statement, self-benefit statement, and

purpose statement, as shown in Figure 5. Workers in the no statement condition saw the basic instructions of how to perform the task and nothing else. Workers in the self-benefit statement condition saw the basic instructions and a sentence at the end about how "people who have worked on these tasks in the past have sometimes reported this as a good learning experience with interesting content." Workers in the purpose statement condition saw the basic instructions and a sentence at the end that said, "By proofreading the text in this HIT, you will help preserve human knowledge and produce free e-books available to underprivileged people." We included the self-benefit statement condition to tease apart the effects of seeing any statement at all versus seeing the purpose statement.

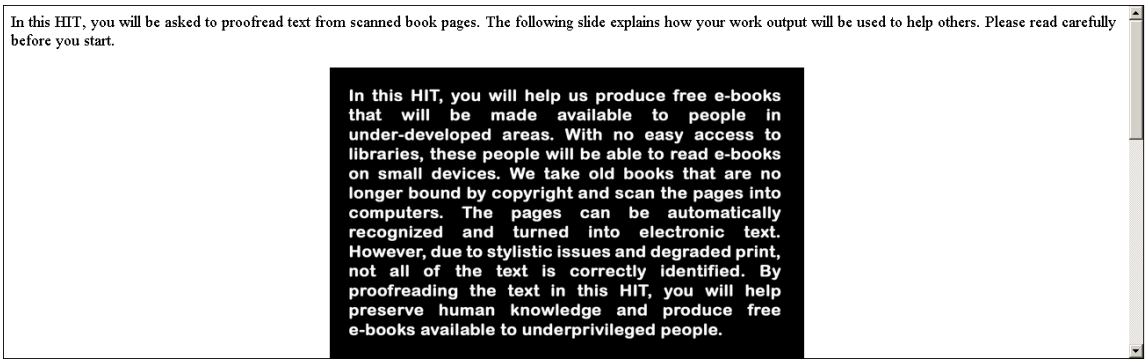
We also experimented with three different formats of presenting the purpose statement: plain text, a female-narrated video, and a male-narrated video. In the video conditions, the worker watched a captioned video with pictures of books and "underprivileged people," voiced by a female or a male narrator, as shown in Figure 5(c). We had two levels of payment: 15 cents versus 10 cents per HIT for each paragraph that was proofread. Workers who completed the survey received an additional 50 cents. Altogether, we had 5 (no statement, self-benefit, text, female video, male video) x 2 (10 vs. 15 cents) = 10 conditions. It was a between-subjects design, and workers were randomly assigned to a condition. Similar to the lab experiment, work quality was measured as the average percentage of introduced errors that were fixed. We again ran the Perl script to automatically parse and compare the paragraphs to count the number of errors fixed.

Figure 5. Screenshots of Experimental Manipulations

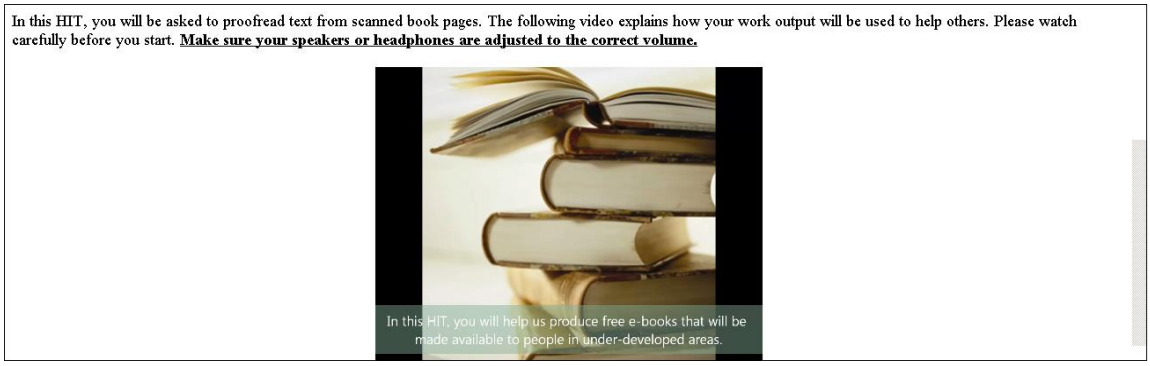
(a) Self-Benefit Statement



(b) Text Version of the Purpose Statement



(c) Video Version of the Purpose Statement



4.2 Field Experiment Results

4.2.1 Effects of Task Significance

We ran an ANOVA test of all workers who submitted at least one HIT. There were 870 workers, with 490 in the low payment condition and 380 in the high payment condition. Among them, 184 saw no statement, 192 saw the self-benefit statement, 193 saw the text version purpose statement, 155 saw the female video version, and 146 saw the male video version. We ran additional analyses to understand whether presentation format or individual attributes led to better recall of the purpose statement. An ANOVA test showed no significant effects of presentation format. 43 of 114 participants (37.7%) who saw the text statement recalled it, 37 of 99 participants (37.4%) who saw the female-narrated video recalled the statement, and 33 of 87 participants (37.9%) who saw the male-narrated video recalled it correctly. Because of the lack of a significant effect from different formats, we collapsed the three formats into a single purpose statement condition in further analysis.

ANOVA results showed no significant differences in work quality across the three task significance conditions, F (2, 867) = 0.94, p = 0.39. Workers who received either the self-benefit statement or any version of the purpose statement did not have higher work quality than workers who received no statement. Table 3 (left column) shows the least square means of the three task significance conditions. An ANOVA test of monetary payment showed no significant effects of the extra payment, F (2, 868) = 1.9, p = 0.17. Workers who were paid 15 cents per HIT actually had a slightly lower accuracy (0.885) than workers who were paid 10 cents per HIT (0.897), although the difference was not

statistically significant. In general, there did not appear to be differences in work quality across the different conditions on a surface level.

4.2.2 Manipulation Checks: Recall of Purpose Statement and Task Significance

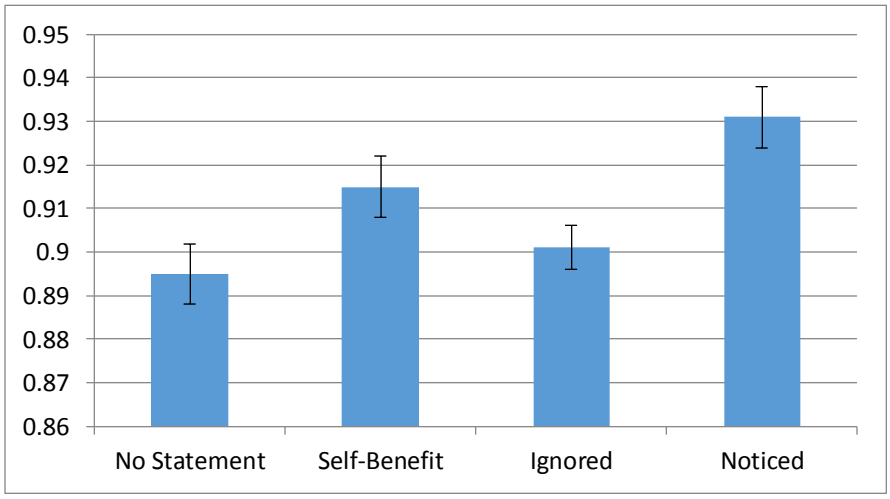
Similar to the lab experiment, we asked workers "Who do you think will benefit from your work on proofreading the paragraphs?" and "How will they benefit from your work on proofreading the paragraphs?" Two research assistants coded the responses separately, and differences in coding were discussed between the coders until a consensus was reached. We coded a worker as having noticed the purpose statement if they were able to recall either question correctly. Out of the 300 survey respondents who received the purpose statement (in any of the three formats), 113 or about 37.7% were able to recall it, slightly higher than the 23% in the laboratory experiment. We labeled these workers as being in the *noticed* condition and the others as in the *ignored* condition. An ANOVA test showed significant differences across the conditions, F (3, 516) = 5.99, p = 0.001. As shown in the right column of Table 3 and in Figure 6, workers who were able to correctly recall the purpose statement had the highest level of accuracy compared to the other three conditions. Workers in the self-benefit statement condition had higher accuracy than those in the no statement or ignored conditions although neither difference was statistically significant.

Table 3. Effects of Task Significance Manipulations on Work Quality

	Accuracy		Accuracy
No statement	0.881_a (0.009)	No statement	0.895_a (0.007)
Self-benefit statement	0.898_a (0.008)	Self-benefit statement	0.915_a (0.007)
Purpose statement	0.893_a (0.005)	Ignored statement	0.901_a (0.005)
		Noticed statement	0.931_b (0.007)
N=870		N=518	

Note: (1) Means with the same subscripts are not significantly different at $p < 0.05$.
(2) Standard errors are included in parentheses.

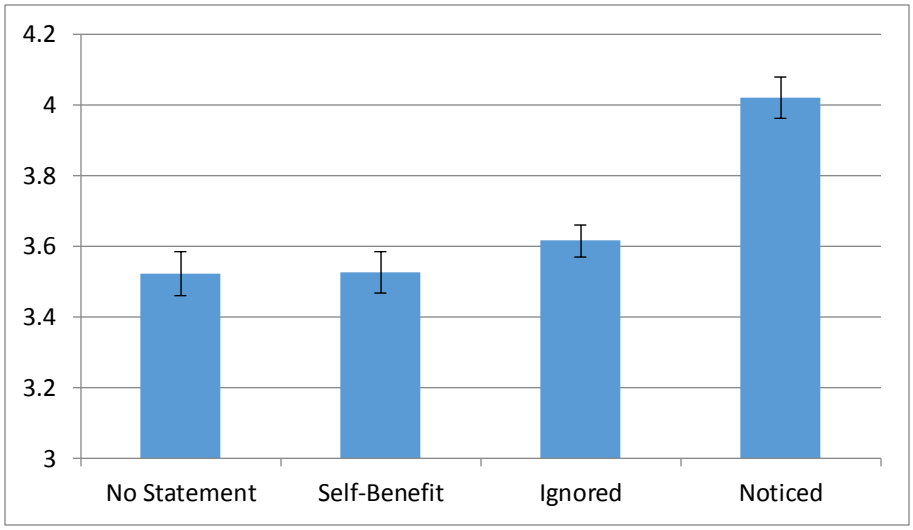
Figure 6. Effects of Purpose Statement Recall on Work Quality



Being able to recall the purpose statement does not necessarily mean that workers have internalized the message and therefore changed their perceptions of the meaningfulness of the task. We included four questions in the survey to measure task significance as follows: (1) the task provides opportunities to improve the welfare of others, (2) some people will be positively affected by how well I perform the task, (3) the task provides opportunities to have positive impact on others, and (4) the task seems trivial and

does not have positive impact on others (reversed). We averaged the responses to the four items to measure task significance. As shown in Figure 7, workers who noticed the purpose statement perceived higher task significance than workers in the other three conditions, $F(3, 516) = 15.95$, $p < 0.001$.

Figure 7. Effects of Purpose Statement Recall on Task Significance



4.2.3 Predicting the Recall of Purpose Statement

In addition, we included the Big Five Personality Questions in the survey and asked our participants to indicate, on a 5-point Likert scale, the extent to which a list of attributes could be used to describe them. We had six items for each of the five factors: conscientiousness, agreeableness, neuroticism, openness, and extraversion. We conducted exploratory factor analysis with oblique rotation, and several items did not load on the correct factors or loaded on multiple factors and were dropped from the analysis. We took the average of the remaining items to measure the factors.

Table 4 shows the results of logistic regression to predict whether a worker was able to recall the purpose statement based on demographic and personality traits. Age and Education did not have any significant effects. Instead, workers who were brand new (one to six months) were more likely to recall the purpose statement. In addition, workers who had heard of e-books were more likely to recall the purpose statement, and workers with a household income of less than $20,000 were more likely to ignore the statement. Two personality traits were marginally significant. Agreeableness increased the likelihood of noticing the purpose statement whereas extraversion decreased the likelihood of noticing the statement.

4.2.4 Ad Hoc Analysis

We included an open-ended question in the survey asking workers to provide additional comments about their experiences. Surprisingly, a good number of workers wrote about how much they enjoyed performing the HIT, so we manually coded whether a worker had mentioned fun or enjoyment in the open-ended response. On average, workers who reported enjoyment had higher accuracy, $F(1, 518) = 10.38$, $p = 0.001$, than those who did not report any enjoyment of working on the tasks (0.938 versus 0.906). The difference is comparable to the difference between workers who recalled the purpose statement and those who did not.

We also found some interesting interactions among the three types of motivations: intrinsic, extrinsic, and social. Purpose statement was not the only factor that increased task significance. Increased payment and enjoyment both led to a higher level of perceived task

significance. Although monetary payments had no impact on work quality, workers in the 15 cents condition reported greater task significance than workers in the 10 cents condition (3.735 versus 3.635, $p = 0.08$). Workers who mentioned fun and enjoyment in the survey also reported greater task significance (3.778 versus 3.592, $p < 0.01$).

Table 4. Predicting the Likelihood of Purpose Statement Recall

Variables	**Model 1**	**Model 2**	**Model 3**
Intercept	-0.261 (0.25)	-1.698 (0.624)	-3.501(1.418)
Age (18-24)	-0.305 (0.296)		
Age (25-34)	-0.154 (0.284)		
Age (35-44)	-0.485 (0.312)		
Age (45-54)	-0.481 (0.373)		
Education (high school)	-0.258 (0.34)		
Education (some college)	0.49 (0.256)		
Education (college)	0.055 (0.2)		
MTurk Tenure (1-6 months)		**0.548 (0.321)**$^{+}$	**0.54 (0.325)**$^{+}$
MTurk Tenure (6-12 months)		0.03 (0.419)	0.129 (0.428)
MTurk Tenure (1-2 years)		0.606 (0.423)	0.627 (0.426)
Heard Of E-Books		**1.176 (0.519)***	**1.188 (0.522)***
Household Income (< 20K)		**-0.727 (0.277)****	**-0.767 (0.285)****
Household Income (20-40K)		0.04 (0.302)	-0.007 (0.308)
Household Income (40-60K)		0.173 (0.402)	0.071 (0.41)
Household Income (60-80K)		0.454 (0.439)	0.421 (0.446)
Household Income (80-100K)		-0.911 (0.718)	-0.735 (0.73)
Agreeableness			**0.443 (0.267)**$^{+}$
Openness			0.233 (0.265)
Extraversion			**-0.352 (0.197)**$^{+}$
N	393	393	393
-2 Log Likelihood	365.82	339.54	329.35

Note: The base for Age is > 55 years old. The base for Education is advanced degree. The base for MTurk tenure is > 3 years. The base for Household Income is > $100,000. $^{**}p < 0.01$, $^{*}p < 0.05$, $^{+}p < 0.1$.

We also examined the impact of demographics and personality traits on work quality. Workers who self-reported having high English ability performed with greater accuracy, $F(3, 480) = 4.74$ and $p = 0.003$. Workers who could "read and understand English

extremely well" outperformed those who could "read and understand English very well" or those who could "read and understand most English." Older workers had higher accuracy than younger workers, $F(5, 480) = 2.1$ and $p = 0.06$, while workers who had heard of e-books also had higher accuracy than those who had not heard of e-books, $F(1, 480) = 6.58$ and $p = 0.01$. Of the five personality traits, conscientiousness (0.016, $p < 0.01$) and openness (0.019, $p = 0.02$) were positively associated with work quality while the other traits had no significant associations with work quality.

4.3 Discussion of Field Experimental Results

The field experiment provides further evidence for the importance of impressing upon workers the task significance of even trivial tasks in online marketplaces for work. In this study, more than one third of workers were able to correctly recall the purpose statement. They not only reported a greater level of task significance, i.e., believing that their work had a positive impact on others, but they also performed the task with higher quality work. Monetary payments, again, had no impact on work quality, although it did marginally increase perceived task significance. Meanwhile, enjoyment had a positive association with both work quality and task significance. This points to a similar effect of motivations on quality as seen in the lab experiment, with intrinsic motivation boosted by task significance.

Compared to the lab experiment, the field experiment provides more insights into the puzzling question of why some workers were better able to process and recall the purpose statement. Two possible explanations are ability and effort. Workers who were able to recall the statement may have been more cognitively capable, or they may have

exerted greater effort to perform the task correctly. Although we did not directly measure cognitive ability in the field experiment, education and English ability served as two proxies for workers' capability in doing the task. Neither factor predicted the recall of the purpose statement, suggesting that inherent ability did not have a major effect. Instead, workers who were new to MTurk, workers who had heard of e-books, workers with more than minimal household income, and workers who were more agreeable or less extraverted were more likely to recall the statement. These workers are likely to put in more effort, whether due to initial enthusiasm, greater intrinsic motivation, or personality attributes. In addition, none of these factors, except having heard of e-books, predicted work quality, which further disproves the influence of ability on recalling the purpose statement.

Overall, our findings suggest that the main difference between the workers who noticed the statement and those who did not was effort rather than ability. Workers who noticed the statement were not necessarily more capable or better educated than their counterparts. Instead, they read instructions more carefully, were not solely motivated by money, and were socially oriented to be willing to exert more effort to help others. Similar to the results of the lab experiment, this study showed the importance of recalling the purpose statement to increase work quality. The field experiment provides further information on the conditions under which recall is improved while also providing insights into worker attributes that result in greater effort being exerted. Although the improvements were marginal, there is a slight increase in work quality with the use of a self-benefit statement. However, the greatest increase in work quality was still attributed to workers who were given the purpose statement and recalled it, and who subsequently reported

higher perceived task significance. This suggests that, ceteris paribus, motivating workers by properly introducing a statement to educate them on the significance of the task is the most effective way to immediately improve the quality of the resulting work.

V. CONCLUSION

Despite our best efforts, there are limitations to this study. While we believe the workers who participated in the field experiment are comparable to prior research and representative of MTurk workers, only 56.7% of them completed the survey. Workers who completed the survey may be somewhat different from the ones who did not, and we do not have data to assess this selection bias. Also, our payment schemes were higher than the current MTurk averages, and we only experimented with a small range (20 vs. 30 cents in the lab experiment and 10 vs. 15 cents in the field experiment). It is possible that payments have a non-linear or marginally decreasing effect on work quality and that our range falls in the leveling-off stage. This may explain why our payment manipulations did not have significant effects.

To summarize, our study highlights the importance and challenges of task and incentive designs in online marketplaces for work. Providing a short statement of task significance significantly improved work quality, yet a 50% increase in monetary payment did not. Surprisingly, rich media formats had no impact on helping workers to process the information from the purpose statement or on improving work quality. Instead, new workers or workers with more than minimal household income were more likely to recall the statement, possibly because of their willingness to exert effort to carefully read the text

or watch the video. Our findings provide practical implications to both Amazon Mechanical Turk and its job requesters, who struggle to address the quality issues. Simply increasing the payment turned out to be a less fruitful approach than informing workers of the purpose. We do not interpret our results as a way of helping requesters to "squeeze" more work out of the already-exploited workers. Instead, we believe improving the morale of workers and their perceived meaningfulness of the work is a win-win solution for all parties involved in MTurk and similar online marketplaces for work.

CHAPTER 3

Searching for the Goldilocks Zone:

Trade-Offs in Managing Online Volunteer Groups[5]

I. INTRODUCTION

The past decade has observed tremendous growth in the number of online volunteer groups that self-organize on the Internet to accomplish tasks that used to be performed in traditional organizations. One successful example is Wikipedia, the free online encyclopedia. In just ten years, millions of volunteer editors collaboratively created more than three million articles in English – and nearly twenty million across all languages. Despite its success, Wikipedia and similar online collaboration efforts face challenges in recruiting and retaining active contributors.

Active contributors are crucial to the success of online volunteer efforts. Without a steady rate of participation and a critical mass of contributors, online groups may suffer sustainability issues (Butler 2001). Contributors may leave for many different reasons. Some may leave due to outside influences in their lives, such as school, work and family. Others may leave due to conflict within the community (Kittur et al, 2007b). Still others, who work on many articles or participate in multiple projects, may leave due to burnout.

Much of the existing research on Wikipedia and other online volunteer efforts has focused on understanding what motivates members to contribute and how they coordinate

[5] Originally published in collaboration with Jilin Chen, Yuqing Ren, and John Riedl.
Wang, L.S., J. Chen, Y. Ren, J. Riedl. 2012. Searching for the Goldilocks Zone: Trade-Offs in Managing Online Volunteer Groups. In *Proc. CSCW 2012*, 989–998, ACM Press. doi:10.1145/2145204.2145351

and resolve conflict (e.g., Cosley et al. 2006, Kittur and Kraut 2010, Kittur et al. 2007a). Our research, instead, focuses on the opposite problem: understanding why members stop contributing to a group and the implications of their withdrawal behaviors. Are newcomers or old-timers more likely to withdraw? Are active or inactive contributors more likely to withdraw? For what reasons? What about the impact of social connections between editors? In addition to retaining members, another crucial goal for online volunteer groups is to complete work. In the context of Wikipedia, continued edits to articles are important and have been used repeatedly to measure group activity and individual productivity (Kittur et al. 2007b, Suh et al. 2009). By comparing factors that affect individual productivity and withdrawal, we hope to explore a critical trade-off between the two: Can the goals of increasing productivity and reducing withdrawal be pursued simultaneously, or does one need to be sacrificed for the other?

The second key trade-off we investigate is between subgroups within a community and the larger community as a whole. Many online groups have a subgroup structure that allows members to participate in more intimate settings (e.g., sub-forums in web discussions and guilds in online games). In Wikipedia, these subgroups are known as WikiProjects, where members organize to work on Wikipedia articles related to specific topics. The relationship between the subgroups and the community is subtle. Some members of online groups may need to decide whether to focus their efforts on a specific subgroup or to split their efforts among multiple subgroups (Daniel and Diamant 2008).

In this paper, we examine the two trade-offs in the context of Wikipedia and WikiProjects. We analyze data from 648 WikiProjects to explore the effects of member

tenure, tenure dissimilarity, past performance, and involvement in multiple projects on members' productivity and withdrawal behaviors at the project level. We then examine the effects of these factors on member productivity and withdrawal at the Wikipedia level. Our results provide evidence for both trade-offs and highlight the challenges in balancing multiple desirable outcomes of community health within a large online volunteer effort like Wikipedia.

The rest of our paper is organized as follows: (1) We review organization science and volunteerism literature on member productivity and withdrawal and propose how they may generalize to online groups. (2) We describe our research setting of Wikipedia and WikiProjects as well as how we assembled the data set. (3) We present our main results and discuss their implications for the management of online collaboration efforts.

II. THEORIES AND HYPOTHESES

We focus on studying productivity and withdrawal, two common ways to measure participation in online volunteer groups. In traditional organizations, productivity can be measured as the quantity of work output, such as the number of papers and reports produced in a research lab. In Wikipedia, number of edits is a common measure of productivity (Kittur and Kraut 2008, Kittur and Kraut 2010, Kittur et al. 2007b, Suh et al. 2009). In traditional organizations, withdrawal includes behaviors such as "psychological withdrawal, lateness, absenteeism, and turnover" (Beehr and Gupta 1978). In Wikipedia, the primary evidence of withdrawal is for a formerly active editor to cease contributing.

Research on productivity and withdrawal behaviors in traditional groups and organizations sheds light on studying similar behaviors in online groups (Cotton and Tuttle 1986). At the same time, insights from offline groups may not be readily applicable to online groups for three reasons: 1) online groups have low entry and exit barriers. In traditional organizations, employees need to follow established procedures to enter or exit the organization or its subunits. In contrast, members of most online groups can enter and exit at their will. 2) Online volunteer groups have little leverage to mandate active participation. Members are motivated by different incentives (Robles et al. 2005). There is a wider variation in member participation (Kittur et al. 2007b), and factors that predict productivity in traditional organizations, such as tenure, may not predict productivity in online groups. 3) Compared to traditional organizations, online volunteer groups like Wikipedia often don't have formal training and socialization programs, both of which could help new members learn the ropes and form relationships with other members to increase productivity and reduce withdrawal.

Furthermore, the voluntary nature of online groups like Wikipedia poses additional challenges in encouraging productivity and discouraging withdrawal. Similar to members of offline volunteer organizations, online volunteers also may suffer from burnout and subsequently withdraw from these groups (Wilson 2000). Because withdrawal may happen suddenly without a formal process, project work may be left unattended until new volunteers take over (Robles et al. 2005).

2.1 Trade-Offs Between Productivity and Withdrawal

While Wikipedia strives to encourage productivity and reduce withdrawal, it is unclear whether the two goals can be accomplished equally well. Ideally, online volunteer groups should achieve both goals simultaneously. In reality, there may be tension between the two. For instance, research has shown that members working in larger groups tend to contribute less than members working alone or in smaller groups. There is a greater penchant for "free-riding" (Albanese and Van Fleet 1985) in larger groups. Free-riding occurs when members retain their membership in a group and receive benefits or credit without providing the same amount of work as others. When free-riding behaviors are frequent, the group may have a seemingly large membership but does not actually benefit from these extra members.

The tension between productivity and withdrawal may also arise from overburdening the core set of active contributors. It is common for a core group of productive members to be the driving force behind a successful online collaboration effort (Kittur et al. 2007b). Studies on Wikipedia have shown how important that core group is in helping to maintain the overall productivity of the free online encyclopedia, compared to the bulk of community members (Kittur and Kraut 2008, Priedhorsky et al. 2007). Although these core members are highly committed and dedicated to the effort, shouldering a heavy workload may erode their energy and enthusiasm and increase their likelihood of withdrawal. An alternative explanation for why active members leave may be that they feel they have accomplished their mission by contributing all that they know, and that further contribution would require much more research and effort (Taylor 2009).

In summary, both productivity and withdrawal are important outcome measures to study, along with potential trade-offs between the two. A successful online collaboration effort requires both a "critical mass" of participants (Markus 1987) and a reasonable level of contribution. Both the amount of activity within an online volunteer group and its size have a strong impact on the ultimate sustainability of the group (Butler 2001).

2.2 Factors That Affect Productivity and Withdrawal

In this section, we identify a set of factors that have been linked to individual productivity and withdrawal in the organization science literature. We first summarize the insights from the existing literature and then speculate how they may generalize to online volunteer groups.

Tenure. Tenure has been conceptualized and measured as the amount of time that an individual has been part of a group or organization. Organization science literature posits a curvilinear, inverted-U relationship between tenure and employee productivity (Sturman 2003). When a newcomer first joins an organization, productivity is expected to increase over time as the person acquires skills, accumulates experience, and becomes familiar with organizational routines and policies. After a number of years of effort, participants are more prone to a burnout effect. The idea of job burnout is based on the potential buildup of stress and exhaustion that may cause workers to decrease productivity (Cordes and Dougherty 1993). Research suggests that certain people may be especially prone to burnout over time (Bakker et al. 2007), and burnout especially affects members with longer tenure who have felt frustration on the job (Cordes and Dougherty 1993). Some

of the nonlinear effects may be due to the aggregation of a plateau effect (Hofmann et al. 1992). In essence, most workers improve at first but then reach a plateau, while others may start strong and begin to decline later.

Tenure has also been shown to be a strong predictor of withdrawal behavior across different professions (Arnold and Feldman 1982). Longer tenure allows for more experience, which would help increase productivity while reducing withdrawal (Arthur 1994). In addition, newcomers have been consistently shown to be more likely to leave an organization than those with longer tenure (Griffeth et al. 2000). We expect a similar effect of tenure in the context of Wikipedia collaboration, because newcomers who lack the experience of doing the work and interacting with other editors may feel frustrated or perceive a lack of fit with individual work groups (Suh et al. 2009). In addition, research has shown that when newcomers' edits are reverted by old-timers, the newcomers are more likely to leave Wikipedia permanently (Halfaker et al. 2011). We thus expect tenure to increase productivity and decrease withdrawal.

Tenure dissimilarity. Another strong predictor of productivity and withdrawal is interpersonal similarity. Studies of traditional organizations have shown that members in the minority component of a group are discouraged from making substantial contributions due to assumptions that they may be weak performers and because they may have more trouble aligning their interests with the rest of the group (Randel and Jaussi 2003). These arguments hold true for demographic attributes such as age and tenure (Wagner et al. 1984). A newbie in a group that consists of mostly experienced members may feel uncomfortable or inadequately prepared to contribute. Similarly, an experienced member

in a group with mostly newbies may have different goals and ideas that are hard to communicate to the rest of the group. Hence, we expect tenure dissimilarity to be negatively associated with productivity.

Tenure dissimilarity may also affect withdrawal (Wagner et al. 1984). The homophily literature suggests that people tend to interact with others who are similar to them on attributes like age, race, ethnicity, etc. Ties between people will dissolve if there is too much dissimilarity (McPherson et al. 2001), while groups that are more homogeneous, in terms of age and tenure, have fewer members leaving than heterogeneous groups (O'Reilly et al. 1989). We expect that members whose tenure differs more from the rest of the group are more likely to withdraw.

Past productivity. In traditional organizations, past performance has been shown to be a reliable predictor of future performance. We therefore expect a strong association between an individual's past and future productivity.

The effects of past productivity on withdrawal are more complicated. In traditional organizations, poor performers are generally more likely to leave than good performers, which implies that withdrawal may not be detrimental for the organization (McEvoy and Cascio 1987). Although some past research suggests that high performers can find alternate opportunities and therefore be enticed to leave the organization, more recent studies usually show that poor performers are more prone to voluntary turnover. In addition, these members more often consider other factors, such as job satisfaction, in evaluating whether to remain in the organization (Spencer and Steers 1981).

We expect the effects of past productivity on withdrawal to be even more complicated in online volunteer groups. There are reasons to expect poor performers to voluntarily leave the project due to lack of contribution. However, there are also reasons to expect good performers to voluntarily leave, either because they have contributed what they know and accomplished their mission or because they are burnt out from maintaining a high level of contribution.

Concurrent projects. Being involved in multiple groups within a community affects member productivity and withdrawal. Similar to traditional organizations that compete for limited resources (Barron et al. 1994), online groups that are created on the same platform or have similar functions also compete for scarce resources like members' time and effort. Online volunteer groups need dedicated members, who put in the most time and effort, in order to survive as a group (Butler 2001). However, from a resource-based view, the amount of time spent on one activity is time that cannot be spent on another (Becker 1965). Those who are involved in multiple subgroups need to decide how to allocate their time to each group, making them physically unable to spend as much time on any one group. We thus expect members who are involved in multiple projects in Wikipedia to do less work for each project. Being involved in multiple projects increases the demand for members' time and effort and, thus, may increase the likelihood of withdrawing from the projects.

An alternative view on the effects of multiple subgroups or projects is job or social embeddedness theory (Mitchell et al. 2001). Members who are affiliated with multiple projects are more socially embedded within Wikipedia, which should reduce their

likelihood of withdrawing from Wikipedia as a whole, and may reduce their likelihood of withdrawing from the individual projects as well.

Communication and social integration. Social identity has been shown to be positively correlated with performance measures (Randel and Jaussi 2003). Members who strongly identify with a group are willing to exert greater effort and make more contributions to the group than those who do not identify with the group. This leads to greater social integration of members into the group. As members become more active within a group, they are more likely to be productive and display better performance (O'Reilly et al. 1989). Interpersonal relationships or psychological contracts are especially important to increase participation and reduce withdrawal intentions in not-for-profit volunteer organizations (Farmer and Fedor 1999). Communication with others is an important way for members of online volunteer groups to be socially integrated and feel like an essential part of the group (Haythornthwaite 2009).

Communication may occur internally within subgroups or externally across subgroups within a large community. Internal and external communication may have different effects on individual behavior. Social integration and frequent communication with members of one's group is likely to focus members' attention on the group goals and needs, and thus increase their contributions to the group and reduce their likelihood of withdrawal from the group (O'Reilly et al. 1989).

In contrast, when members have many ties or frequent communication with those outside of a group, they are more likely to be pulled away from the focal group (McPherson et al. 1992). External connections have been shown to affect performance and withdrawal

differently in traditional organizations. On one hand, an individual's connections with external groups have been shown to improve performance due to access to novel and relevant information (Cross and Cummings 2004). On the other hand, social networks research has shown that employees or community members who are on the outskirts of groups and have strong external connections are more likely to leave (McPherson et al. 1992). Increased communication and socialization outside of a project is therefore likely to increase both individual productivity and withdrawal.

2.3 Trade-Offs Between the Community and its Subgroups

Online groups often consist of subgroups, such as projects within Wikipedia. Membership in multiple subgroups, on one hand, creates tension around how an individual allocates time and effort to different subgroups; on the other hand, this strengthens connections between an individual and the larger entity by enriching the web of connections tying the member to the community. It is important to understand more precisely how this trade-off works and how individual commitment and contribution to subgroups might transfer to the community as a whole.

Two mechanisms may be at work in determining how an individual's involvement in multiple subgroups affects the individual's involvement with the larger community. The first mechanism is the competition argument. As mentioned earlier, members participating in multiple subgroups must dole out their time and effort to each subgroup (Becker 1965). Because members have limited time to spend on volunteer work overall, splitting their efforts is likely to cause decreased productivity and increased withdrawal for individual

subgroups (Daniel and Diamant 2008). Reduced productivity and increased withdrawal for individual projects does not necessarily mean reduced productivity and increased withdrawal for the community. Reduced levels of contribution to multiple projects, when aggregated, may exceed the average level of contribution to a single project. Similarly, for members who are involved in multiple projects, leaving one project doesn't conclude the person's affiliation with Wikipedia.

The second mechanism is job or social embeddedness. Job embeddedness theory suggests that more investment in a job increases the quitting cost, which is negatively correlated with the likelihood of leaving the organization (Mitchell et al. 2001). Someone who is highly embedded (or involved) in an organization would be less likely to withdraw. In the context of Wikipedia, involvement in multiple projects increases the extent to which individual editors are socially embedded within Wikipedia, which makes them less likely to leave the Wikipedia community as a whole.

Furthermore, affiliation with multiple projects can lead to effective knowledge transfer across the projects, and thus increase the amount of work done within each project (Cross and Cummings 2004). In these instances, members increase their ability to draw from outside sources and gain new knowledge that is useful to their productivity within the group. Furthermore, in the online context, member participation is socially driven by perception of and interactions with other members (Butler et al. 2007), Farmer and Fedor 1999). The social embeddedness perspective implies that involvement in multiple projects may increase member contribution to Wikipedia as a whole. To explore these trade-offs, our research questions are:

R1: How do individual tenure, past productivity, project involvement, and communication affect productivity and withdrawal at the project level?

R2: How do individual tenure, past productivity, project involvement, and communication affect productivity and withdrawal at the Wikipedia level?

R3: Are there any trade-offs between productivity and withdrawal? Are there any trade-offs between these outcomes at the project level and the Wikipedia level?

III. WIKIPEDIA AND WIKIPROJECTS

In this section, we describe our research setting – Wikipedia and the WikiProjects subgroups. Wikipedia is a free online encyclopedia that anyone can edit. An edit is simply a revision, large or small, to any article, talk, or user page. If users perform edits, they may also choose whether or not to register for a user account. Those who have accounts may then interact with other registered users, or editors, to coordinate work and discuss articles. Each article page has an associated talk page enabling editors to collaborate while working on the article.

A WikiProject is defined as "a group of pages in the 'Wikipedia' article namespace which are devoted to the management of a specific topic or family of topics within Wikipedia; and, simultaneously, a group of editors who use those pages to collaborate on encyclopedic work"[6]. Since 2002, more than 20,000 Wikipedia editors have joined more than a thousand projects. WikiProjects provide a way to organize editors with the goal of improving a specific subset of articles in Wikipedia. Members may choose to join or leave

[6] http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Philosophy

a project by adding or removing their names on the project's member list. The main page of a WikiProject typically includes a brief description of the project and its scope, a list of project members, guidelines, and tasks that require member contribution.

In this study, we focus on editors who have contributed to Wikipedia as a part of one or more WikiProjects. We chose to study WikiProjects because they have clearly defined goals and boundaries, which make it easy to assess individual productivity and withdrawal at the project level.

IV. DATA AND VARIABLES

The dataset we use in this study is extracted from the January 2008 dump of the English Wikipedia, which includes the full text of all pages and their complete edit histories from the creation of Wikipedia to the end of 2007. To gather information about projects and their members, we traversed the main directory page of WikiProjects and included all projects that are topical (thus excluding projects such as WikiProject Citation Cleanup). We also excluded projects that never grew to have at least three members (the minimum size of a group), projects that do not have a member list to track membership, and projects whose scopes could not be estimated using categories. Our final data set has 648 WikiProjects and 14,464 individual editors who are or have been members of these projects.

We determined each WikiProject's membership and scope following the approach in Chen et al. (2010). We used historical edits of a project's member list to identify members of each WikiProject. We considered an editor to have joined a project when her username appeared on the member list and to have left when her username was removed

or contribution stopped. To determine the scope of a WikiProject, we first found the Wikipedia category that matched the title (like category *Computer science* for WikiProject *Computer science*). We then traversed all subcategories of the matched category down to the 4th level and considered all articles in those categories to be within the scope of the WikiProject (see Chen et al. (2010) for more detail).

We constructed a longitudinal dataset to temporally separate the independent and dependent variables. Each observation records the characteristics and activities of an individual editor as a member of a project for each quarter in that project's lifespan. Within each project, each quarter is a 90-day period in a project's lifespan, with the first quarter beginning immediately after its creation date. Within each quarter, every editor who was a member of the project in that quarter was measured once for that project. The level of the analysis is therefore *project individual quarter*, with quarters nested within individuals and individuals nested within projects. For any given WikiProject, the first quarter is the first 90-day period following its date of creation. If the project had 10 members in this quarter, the dataset would have 10 observations for the project during this quarter, with each observation measuring the activity of one editor within the project. In total we had 85,105 project individual quarters in the WikiProject dataset.

4.1 Dependent Variables

Project-Level Productivity: We measured an editor's project-level productivity as the number of edits performed by the editor on articles *within the scope of the WikiProject* during the current quarter.

Project-Level Withdrawal: We measured an editor's project-level withdrawal as a binary variable, i.e., either 1 or 0. The variable is 1 if and only if the editor was an active member of the project in the current quarter, but removed her username from the project member list or stopped contributing within the scope of the project by the end of the next quarter.[7] We considered a member to be active if the person had at least one edit during that quarter to any of the following: an article within the project scope, the talk page of such an article, any project organization page, or the user pages or user talk pages of another project member.

Wikipedia-Level Productivity: We measured an editor's Wikipedia-level productivity as the number of edits performed by the editor on *any and all Wikipedia articles* during the current quarter.

Wikipedia-Level Withdrawal: We measured an editor's Wikipedia-level withdrawal as a binary variable, i.e., either 1 or 0. The variable is 1 if and only if the editor had performed at least one edit in Wikipedia during the current quarter but made no edits in Wikipedia in the subsequent quarter.

4.2 Independent Variables

Tenure: We measured an editor's tenure by how long the editor had been a member of Wikipedia, that is, the number of days elapsed from a member's first edit in Wikipedia to the end of a quarter (Chen et al. 2010).

[7] 11% of editors returned one or more times after withdrawing, with a majority returning after one inactive period. Excluding these editors from the analysis did not significantly affect our results. In the analysis below we consider only the last and final instance of withdrawal for each editor.

Tenure Dissimilarity: We measured an editor's tenure dissimilarity from the rest of the project members using Euclidian Distance as follows (Wagner et al. 1984):

$$\sqrt{\sum_{j=1}^{n} \frac{(S_i - S_j)^2}{n}}$$

where S_i is tenure for the editor in question, S_j is the tenure of the j-th member in the project, and n is the total number of members currently in the project.

Past Productivity: We measured an editor's past productivity as the total number of edits performed by the editor on articles within the scope of the WikiProject before the current quarter.

Concurrent Projects: We measured an editor's concurrent projects by the total number of projects of which the editor is currently listed as a member. A higher number means that the editor is involved in more projects at the same time.

In-Project Communication: We measured an editor's in-project communication by the number of edits that other members of the project have made to the editor's user page and user talk page. We only counted edits by others on the editor's pages, not the editor's own edits, for two reasons. First, it is easier to track one user page than two. Second, strong ties tend to be reciprocal, so the amount of inbound communication to an editor should be a good measure of how socially integrated the editor is within Wikipedia.

Out-Project Communication: We measured an editor's out-project communication by the number of edits that non-members of the project have made to the editor's user page and user talk page.

4.3 Control Variables

Quarter Index: The index of time within the project was measured in quarters (90-day periods), starting with quarter 0 from the moment the project is created until the last full quarter before the end of 2007.

Project Scope: Measured as the number of articles falling under the project scope. Project scope was determined using the same approach as Chen et al. (2010).

Project Size: Measured as the number of project members during the current quarter.

V. RESULTS

Descriptive statistics and correlations of the variables in the dataset are displayed in Table 5. Most of the variables are highly skewed to the right so we performed base-2 logarithmic transformations for normality considerations. Because our data is nested by nature, we analyzed the data using Hierarchical Linear Models (HLM) (Bryk and Raudenbush 1992), with member productivity and withdrawal as the dependent variables and project and individual characteristics as the independent variables. HLM is an advanced form of linear regression that allows us to examine the effects of independent variables on dependent variables while taking into account potential correlations across observations that are nested within a high-level entity (e.g., individuals nested within projects). Our dataset is cross-nested between projects and individuals, meaning an editor can belong to multiple projects. Thus, we ran the analysis using the lmer function in R.

Table 5. Descriptive Statistics and Correlations of Variables

Descriptive Statistics			Correlations											
	Mean	Std Dev	1	2	3	4	5	6	7	8	9	10	11	12
1. Project-Level Productivity	34.63	162.3												
2. Project-Level Withdrawal	0.213	0.409	.02											
3. Wikipedia-Level Productivity	200.5	656.7	.52	.05										
4. Wikipedia-level withdrawal	0.262	0.440	-.01	.50	-.02									
5. Quarter	5.292	2.864	-.07	.04	-.10	.14								
6. Project scope	20681	58493	.07	.03	.01	.01	.02							
7. Project size	58.83	49.95	-.07	-.01	-.10	.07	.41	.01						
8. Tenure	527.7	341.8	.00	.06	.04	.08	.24	.01	.03					
9. Tenure dissimilarity	405.3	166.8	.00	.06	.01	.08	.21	.05	.00	.48				
10. Past productivity	239.3	832.2	.43	.07	.24	.03	.07	.08	-.02	.14	.05			
11. Concurrent projects	3.253	4.154	.02	.06	.12	.06	.00	.05	-.07	.16	.08	.07		
12. In-project communication	0.964	5.800	.34	.02	.19	-.01	-.06	-.02	-.02	-.02	-.02	.21	.03	
13. Out-project communication	14.44	36.83	.28	.08	.50	-.01	-.11	.01	-.10	.05	.00	.17	.22	.30

Our main results are displayed in Table 6. The first two columns show how project and individual characteristics relate to individual productivity and withdrawal at the project level, and the second two columns show the relationships at the Wikipedia level. We standardized all independent variables for ease of comparing coefficients across variables. We estimated our HLM models using maximum likelihood estimation, random intercepts, and unstructured covariance structure. For each dependent variable, we ran two models: the base model with the intercept and quarter as predictors, and the complete model with two project level variables and six member level variables. Due to the large size of our data set, we examined Bayesian Information Criterion (BIC) across the models. BIC punishes models with a large sample size and a large number of parameters (Burnham and Anderson

2004). We included BIC and deviance between the complete model and the base model in Table 6. In all analyses, the deviance is greater than 10, meaning the complete model fits the data better than the base model.

Table 6. Predicting Member Productivity and Withdrawal Behaviors

Variables	Project-Level Productivity	Project-Level Withdrawal	Wikipedia-Level Productivity	Wikipedia-Level Withdrawal
Intercept	2.151**	- 2.981**	5.343**	-5.495**
Quarter	- 0.191**	0.473**	- 0.567**	1.236**
Project Scope	0.464**	- 0.111	0.127**	- 0.446**
Project Size	- 0.216**	0.104*	0.254**	-0.122
Tenure	- 0.165**	- 0.129**	0.024+	-0.073**
Tenure Dissimilarity	0.085**	0.108**	0.029*	0.079**
Past Productivity	0.596**	0.362**	0.113**	0.133**
Concurrent Projects	- 0.287**	- 0.061**	- 0.116**	0.034**
In-Project Communication	0.611**	- 0.042**	0.290**	- 0.047**
Out-Project Communication	1.006**	0.350**	2.390**	0.188**
BIC	317209	56144	347630	53816
Deviance	35867**	2361**	57519**	1185**
N	N = 85105	N = 65393	N = 85105	N = 65393

We use the following notation in tables to represent p-values: ** p < .01, * p < .05, + p < .1

The first two columns of Table 6, respectively, predict member productivity, measured as the number of edits performed by an individual member within a project in the current quarter, and member withdrawal, measured as either removing one's username from a project or stopping contribution to the project in the next quarter. Our analysis revealed a negative relationship between tenure and both productivity and withdrawal ($p <$ 0.01). Compared to newcomers, old-timers contributed fewer edits and were less likely to withdraw. Our analysis also revealed a positive relationship of both tenure dissimilarity and past productivity with productivity and withdrawal in the current quarter ($p < 0.01$). Members with tenure different from the group contributed more edits than those with

similar tenure and were more likely to withdraw from the project. Members who had made more edits in the past continued contributing more but were also more likely to withdraw from the project.

Our analysis revealed a negative relationship of the number of concurrent projects with productivity and withdrawal ($p < 0.01$). Members who took on more projects contributed fewer edits than those with fewer projects. Contrary to our predictions, members involved in multiple projects were less likely to withdraw from any individual project. We also found a positive effect of communication both within and outside of the focal project ($p < 0.01$). Members who engaged in communication, either with other project members or editors outside of the project, contributed more edits than those who engaged in less communication. Communication within and outside of the focal project had opposite relationships with withdrawal from the project, with more internal communication being associated with less withdrawal and more external communication with more withdrawal. These relationships are consistent with our predictions.

The last two columns of Table 6 show how project and member characteristics relate to productivity and withdrawal at the Wikipedia level, i.e., the total number of edits within Wikipedia during the quarter and whether an editor stopped contributing to Wikipedia after the next quarter. Most of the effects on productivity at the Wikipedia level are similar to those at the project level, with tenure being the only exception. While editors with longer tenure contributed fewer edits to individual projects, they contributed more edits to Wikipedia as a whole ($p < 0.1$). Contrary to our predictions, members who belong to multiple projects contributed less work, both to an individual project and to Wikipedia

as a whole. The standard coefficient for project-level productivity is greater than the one for Wikipedia-level productivity, implying greater negative impact on the former.

The results on withdrawal at the Wikipedia level versus the project level are similar as well, with three exceptions. First, while project scope had no significant effect on project-level withdrawal, larger scope was associated with reduced withdrawal from Wikipedia as a whole. Second, members of projects with more total editors were more likely to withdraw from the project but not significantly more likely to withdraw from Wikipedia. Third, while belonging to multiple projects was associated with a lower likelihood of withdrawing from any single project, it was associated with a higher likelihood of withdrawing from Wikipedia as a whole ($p < 0.01$).

VI. DISCUSSION

We set out to understand two critical trade-offs in online volunteer groups like Wikipedia: the trade-off between productivity and withdrawal and the trade-off between subgroups within a community and the community as a whole. Our quantitative analysis of WikiProjects provides evidence for both types of trade-offs. Table 7 summarizes some of the critical trade-offs revealed in our analysis.

For the ease of illustrating the trade-off, we summarize our results as the impact on *productivity* and *retention*, where retention is the opposite of withdrawal. A trade-off exists when a factor increases one outcome while decreasing the other or increases the same outcome at one level while decreasing it at another level. The left-hand side of Table 7 refers to trade-offs between productivity and withdrawal, while the right hand side refers

to trade-offs between projects and Wikipedia as a whole. For instance, the second row in the left column shows some major trade-offs between productivity and withdrawal when tenure and past productivity are the independent predictors. The display of (–) tenure → (+) productivity and (–) retention means that shorter tenure is associated with higher productivity and lower retention.

In the rest of the section, we discuss these trade-offs, speculate on their underlying processes, and highlight implications for managing online collaboration.

Table 7. Summary of Main Findings

	Tradeoffs between productivity and retention			**Tradeoffs between projects and Wikipedia as a whole**		
Project Characteristics	**No tradeoffs**			**Mixed tradeoffs**		
	(+) project scope	→	(+) productivity (+) retention	(+) project scope	→	No effects on project retention (+) Wikipedia retention
	(–) project size			(+) project size	→	(–) project productivity (+) Wikipedia productivity
Member Attributes	**Major tradeoffs**			**No tradeoffs**		
	(–) tenure (+) past productivity	→	(+) productivity (–) retention	(+) tenure	→	(–) project productivity (+) Wikipedia productivity
Member Connections	**Mixed tradeoffs**			**Mixed tradeoffs**		
	(+) concurrent projects **(–)** out-project communication	→	(–) productivity (+) retention	(+) concurrent projects	→	(–) project productivity (–) Wikipedia productivity (+) project retention (–) Wikipedia retention
	(+) in-project communication	→	(+) productivity (+) retention			

6.1 Trade-off between Productivity and Retention

Our first main finding is a critical trade-off between how many edits an individual editor contributes to a project and her likelihood of staying with the project. A typical factor with

such opposite effects is an editor's past edits, which is associated with increased productivity but a reduced retention rate. Contrary to traditional organizations where poor performers are more likely to leave an organization (McEvoy and Cascio 1987), we find that good performers are more likely to leave or stop contributing in online volunteer groups. One possible reason is the "mission accomplished" effect, as illustrated in the following quote:

"Having done all I can on the Andorra rugby and womens sevens pages (aside from keep them up to date), I am going to see if I can help with the Shannara project."

Another possible reason is the burnout effect, as illustrated in the following quotes from conversations between active editors who have experienced "wikiburnout" or "wikistress":

"I am suffering from wikiburnout and chronic Wikistress. […] I probably won't be able to log in as frequently and contribute as much as I would like."

"On another point; I noticed your wikistress level is high, and your contributions may be dropping. […] I'd like to add something else; avoid burnout. You are a very active contributor. It is easy for highly active contributors to get caught up in burnout."

Stress management seems to be a major issue among active contributors. For dealing with stress, the meta-wiki of Wikimedia (the mother organization of Wikipedia) lists 154 tips contributed by editors[8]. The meta-wiki also suggests that stressed people leave Wikipedia for a short while so they can recover[9].

[8]http://meta.wikimedia.org/wiki/Wikistress
[9]http://meta.wikimedia.org/wiki/Wikibreak

Many active Wikipedia editors created "wikistress meters" on their user pages to indicate their stress levels to fellow editors[10]. The use of these stress meters may help alleviate the burnout effect of productive editors by increasing awareness of stress levels among editors. However, the effectiveness of such a solution remains limited due to the effort and skills required to create and update the meters. In addition, some research has shown that focusing on the negative aspect of stress may lead to unintended effects such as depression (Lyubomirsky and Nolen-Hoeksema 1995).

A design opportunity lies in improving the ease of use and functionality of stress awareness tools like stress meters. For instance, software agents can be developed to automatically estimate stress levels from an editor's recent activity and comparison with historical patterns. Stress information can be customized or hidden from the focal editor to avoid negative impact. Instead, it can be made available to other editors and project leaders who can use the information to proactively manage the stress of project members and avoid overburdening already-active members. Tools like these can be promising for alleviating the stress of productive editors, thus helping to maintain high productivity while improving member retention.

6.2 Trade-off between WikiProjects and Wikipedia

Our second main finding is a critical trade-off between members' continued contributions to individual projects versus Wikipedia as a whole. Membership in multiple projects reduces one's likelihood of leaving an individual project but increases the likelihood of

[10]http://en.wikipedia.org/wiki/Template:Wstress3d

leaving Wikipedia. In contrast, membership in multiple projects reduces the amount of work editors contribute at both levels. The negative effects of multiple project membership on productivity have wide implications because about 45% of the editors in our data were involved in more than one project, with 10% being involved in five or more projects. Our results suggest the detrimental effect is not limited to individual projects. Instead, it may have spilled over to affect other types of work an editor does for Wikipedia.

Anecdotal evidence we found on talk page conversations further highlights the challenge for active editors to take on and juggle too many projects and the risk for project leaders to over-draft from the same pool of active editors.

"Goodnes[sic]; I should have abolished this article last month but got too many projects on my plate and forgot."

"I'm working on too many projects atm. I'm going to be moving slowly here."

"Ditto. PS. Considered joining us in this fine wikiproject? :)" "Thanks. I can definitely occasionally lend a hand here and there, but I already am involved in too many projects for the limited time budget I am on."

The real trade-off lies between withdrawal behaviors at the project level and those at the Wikipedia level. Being involved in multiple projects showed decreased likelihood of leaving a specific individual project, possibly because multiple projects give members more opportunities to remain involved. At the same time, being involved in too many projects may increase their stress level, which could eventually cause them to abandon the site altogether. The challenge is how to leverage the benefits of multiple project membership while minimizing its negative impact. Resolving this challenge requires both

change to Wikipedia's policies and guidelines and the development of software tools to improve awareness of editors' activity and commitment across projects. Tools can be developed to share information across projects, such as how many projects an editor has joined, how many edits an editor has made for each of these projects in recent months, etc. Wikistress meters may also be an indication of instances in which editors have been overtaxed by too many projects. We expect such information to provide insights into an editor's workload that may help coordination efforts to avoid competition among projects for member attention. Tools can also be developed to use a combination of signs such as involvement in a large number of projects and sudden or significant drops in recent editing behaviors to generate alerts for editors or project leaders.

VII. CONCLUSION

At a high level, many of the insights from social and organizational theories still apply to the online context. Both individual productivity and withdrawal are affected by factors such as group characteristics, individual attributes like tenure and past performance, and social connections within and across projects. We were able to replicate many relationships between variables in traditional organizations, such as the negative relation between tenure and withdrawal, the positive relation between internal and external communication and productivity, and the positive relation between external communication and withdrawal. However, there are subtle differences between our findings and what we expected from reviewing old theories.

One example is the negative relationship between tenure and productivity and the positive relationship between past productivity and withdrawal. Following the organization science literature, we expected old-timers to be more productive and members with low levels of productivity to be more likely to withdraw. Yet we found the opposite of these effects. The discrepancy can be attributed to the informal and voluntary nature of online groups compared to formalization and bureaucracy in traditional organizations. Members of online groups don't have fixed roles (except those who become administrators) and self-select to take on certain tasks and responsibilities. The goal of many members may be to find or share information and contribute to a good cause, rather than sticking around and climbing corporate ladders. Therefore, members who have contributed much of their knowledge may either feel a sense of "mission accomplished" or become burnt out and leave or stop contributing.

Another discrepancy between our findings and the organizational literature is the effects of the number of concurrent projects on productivity and withdrawal. The variable itself is somewhat unique to the online context. Although employees of traditional organizations may work on multiple projects, they are limited in how many they can join, either by their billable hours or managerial oversight. The voluntary nature of online collaboration allows members to join as many projects as they wish, up to or even beyond what their time and effort allow. With greater control and less individual autonomy in choosing projects to join, we expect some differences in the effects of multiple project membership on individual behaviors. For instance, being involved in multiple projects does not necessarily increase one's likelihood of withdrawal. Instead, individuals may develop

social networks with different units within the organization, reducing their likelihood of withdrawing from these units though they may not have strong bonds with the organization itself.

Comparison with the organization science literature reveals a limitation in our measure of productivity. It does not consider the type of work or quality of work. This may help explain the interesting dilemma we found with old-timers – they stay longer with a project but do not contribute as much as newcomers. Because our measure of productivity only considers the quantity of contribution, it is possible that, with more experience, old-timers shift their focus to administrative work or more challenging tasks, which is not reflected in a simple edit count. Examining type and quality of work would be fruitful for future research.

Overall, our study confirmed the applicability of social science theories to at least one online group, while also highlighting the importance of reconsidering and modifying the assumptions and propositions to fit the online context.

CHAPTER 4

Shining a Light Under the Bridge:

The Identification of Trolling in Online Communities[11]

I. INTRODUCTION

Online communities provide a platform, with its own rules and social norms, for users to gather for a common purpose (Preece 2000). Websites like Facebook and Wikipedia allow people to meet and work with others who they may or may not know in person. These communities provide benefits that offline organizations are unable to do. Much of the existing literature on online communities examines positive behaviors, such as what motivates members to contribute to these communities (Cosley et al. 2006) or what design principles are best for creating such communities (Chaturvedi et al. 2011). Positive contributions from users and healthy discussions are necessary for social communities to thrive and endure. However, there are also users who behave in ways detrimental to the intent of the community, and these behaviors may also differ significantly from how the same individuals may behave offline. Furthermore, there is a loss of control and an inability to enforce rules in an online environment due to lack of face-to-face contact and diminished impact of sanctions.

In this research, internet trolling is studied due to its nature as a prevalent and persistent deviant behavior in online communities. *Trolling* can be defined as posting incendiary or off-topic messages with the purpose of eliciting agitated responses and

[11] With guidance from Shawn Curley and Yuqing Ren.

disrupting an online community's normal functions (Phillips 2015). *Trolls*, users who intentionally engage in trolling behaviors, target the community as a whole but also often prey on a subgroup of members who are likely to respond. A troll might, for example, post a politically controversial comment on an online forum that is not political in nature, such as a health support forum, thus disrupting the normal flow of conversation. Although many communities use the common mantra of "do not feed the troll," it is often difficult to identify when a member is trolling rather than being genuine. Because it is difficult to ascertain the intent of a suspected troll, and to avoid restricting the diversity of opinions, community administrators need to successfully identify true cases of trolling.

Research on trolling has largely focused on the intentions of trolls and the impacts of their actions with an eye toward reducing these impacts (Adler and Adler 2008, Chaturvedi et al. 2011, Shachaf and Hara 2010). For example, Shachaf and Hara (2010) examined trolls on Wikipedia and found that many of them derive enjoyment from their disruptive behaviors, often violating the terms of use. In their interviews with administrators, they found that administrators viewed trolling in a highly negative light and believed they were able to identify individual users who were trolls. The study reviewed specific cases of trolling and concluded that users with good intentions, by definition, could not be considered trolls though their actions may seem similar. There was no direct contact with these potential trolls, and judgment of their intent was made through the authors' analysis of their actions.

Although past research has largely studied behaviors like trolling from the perspective of those who want to eliminate them, according to some (Herring et al. 2002,

Kelly et al. 2006, Kirman et al. 2012), there are types of trolling that are not necessarily detrimental or intentionally harmful. This study attempts to distinguish between trolling that is harmful and behavior that may not be harmful through an exploratory study of online forum comments on the site deviantART.com. Within this study, trolling is not automatically defined as a negative behavior; instead, it is defined as a behavior that the *community* may consider disruptive. In particular, this research attempts to construct a framework to better understand how community members identify potential trolling, what the potential effects of trolling are, and what this may mean for moderators intent on mitigating the negative effects of trolling. Therefore, the major research questions posed in this study include:

How can we reliably identify trolling behavior?

What impact does trolling have on the community?

In this study, members of the community were first interviewed to gain insights into the behaviors that the community considered to be harmful and to understand how the community views trolling. These insights, in turn, were used to develop a coding scheme for analyzing forum comments. Data were then collected from community forums and coded for specific attributes, such as methods of trolling that were evident in comments and what kinds of evidence members cited to identify cases of trolling. In general, trolling was viewed as a negative behavior, but community members indicated a willingness to enjoy trolling that they considered humorous. Members used a variety of methods to identify potential trolling, but there was often no community consensus on whether a user was a troll. However, when a member was accused of trolling but did not exhibit many

trolling behaviors, the community was likely to defend the accused member and move on to continue regular discourse. In other words, although it was difficult for the community to identify clear cases of trolling, particularly trolling that had a negative impact, it was less difficult for the community to identify cases that were not trolling. This study confirms the difficulties faced by online communities in accurately identifying trolls and suggests that a community consensus may be used by moderators to make a final determination.

II. RELATED WORK

2.1 Negative Impact of Deviant Behaviors

Although there is a sparsity of research on trolling behaviors, we can draw from prior research on offline deviant behaviors and other online deviant behaviors to understand the general phenomenon. For example, there are a number of theories from traditional sociology and psychology literature regarding the causes of deviant behaviors. In an article by Frick and White (2008), the authors examine the development of deviant behavior in childhood and adolescence, especially in children with more callous-unemotional (CU) traits. People with this trait, they argue, are particularly aggressive and antisocial. Another stance considers the possibility that deviant behaviors may be learned in much the same way that other social behaviors are learned. Akers et al. (1979) suggest that positive and negative rewards or punishments may be able to reinforce deviant behaviors through differential reinforcement. Furthermore, some studies on moral violations have found that certain deviant behaviors are retaliatory behaviors in response to perceived injustices

(Mullen and Nadler 2008). Like other behaviors, deviant behaviors may be caused by a combination of personal traits as well as stimulants in the environment.

Theories on deviant behaviors that manifest in the offline world can help explain some of the deviance online, especially when we consider how such traits can be exacerbated by the lack of restraints in an online setting. For example, Evans (2011) contributes to the study of deviance in the online arena by examining how offline social deviants are able to gain acceptance and a chance to socialize with like-minded people online. Because the norms of an organization can change due to the behavior of a fraction of the members, this effect can quickly spread within an online community, causing deviant behaviors to propagate faster. On the one hand, physical barriers are removed, making it easier to form communities and close-knit groups. On the other hand, increased anonymity and decreased accountability make it difficult to enforce regulations. For instance, Clay Shirky describes how users can act out in front of a crowd without paying any personal price to their reputation due to this sense of anonymity online (Doig 2008). Furthermore, people who are naturally loners and "self-injurers" are prone to finding a social outlet online, which exacerbates the problem of online deviance (Adler and Adler 2008). In a survey of users and their attitudes toward trolling, one study found that negative personality traits, such as sadism, correlated with users' enjoyment of trolling, suggesting that online trolls are likely to already exhibit offline deviance (Buckels et al. 2014).

Regardless of the causes, deviant behaviors in online communities are problematic for several reasons. First, there is the immediate threat of disruption to the community. For instance, because of older users' low tolerance of profanity and flaming (posting incendiary

and provocative comments), social sites have trouble attracting the adult demographic. "While people between ages 30 and 49 are about as likely to be online as their younger counterparts, according to the Pew Internet and American Life Center, they are significantly less likely to read blogs or consult Wikipedia" (Wilson 2007). Second, some behaviors, such as antisocial behaviors, can actually change a group norm by influencing non-deviants to exhibit the same behaviors (Robinson and O'Leary-Kelly 1998). Social interactions within an online network can affect members' actions and opinions, with members being influenced by the actions and opinions of their peers (Han and Kim 2008). In a similar study, Dishion et al. (2008) discuss how adolescent deviant behaviors are likely to increase when deviants are able to socialize with similar people and form groups. Online communities that want to prevent negative impacts of these behaviors may need to monitor specific users to prevent the spread of similar deviance.

2.2 Prevention of Online Deviance and Trolling

Past research shows that more moderation, and specifically more timely moderation, can help to prevent alienation of rule-abiding users (Wise et al. 2006). The Slashdot community, for instance, uses a very effective peer-review method of allowing randomly selected users to help moderate messages through a voting system. Users can vote a comment up or down and use descriptive tags, such as "troll," to explain why a downvote is cast. When a comment has been too negatively rated, however, it is typically hidden from view by default, and many users might never change their default settings (Lampe et al. 2007). Thus, the cost of this method is in the small percentage of comments that might

be incorrectly rated, or that might contain interesting or valuable information. Slashdot has been estimated to have a 92-93% accuracy of rating, so this cost may be worth it in exchange for reducing the need for administrative effort (Poor 2005).

Other creative methods for dealing with trolling include *disemvoweling*, which is the process of removing vowels from comments suspected to be trolls (Thompson 2009), and *hellbanning*, a technical feature that shields members from seeing the activities of a banned user but does not prevent the user from executing any normal actions (Atwood 2011). In the latter method, a troll will not be immediately aware of the ban and can continue to post comments and view the site as a legitimate member. Meanwhile, other users will not see the troll's activities, effectively implementing a secret ban. However, these methods rely on the correct identification of a troll in order to prevent punishing legitimate members.

Some studies have attempted to identify trolling based on a generally accepted definition that trolls intentionally provoke other members. For example, one study examines how a troll is able to successfully disrupt a community through useless or pointless discussion (Herring et al. 2002). In this study, trolling is already predefined as a method of subverting normal community discussion. Another study further defined trolling as posing as a sincere member of a group while intending to create conflict, but the author notes the difficulty in deciphering intent (Hardaker 2010). Other studies (Kelly et al. 2006, Phillips 2015) have identified trolls in discussion forums through their interactions with members of the discussion groups, often highlighting the discrepancy in opinions. In

general, trolling is understood to be an intentionally disruptive behavior, but the identification of trolling may be subjective depending on the accepted norms of a group.

In addition, there is still a lot of ambiguity in being able to identify the negative impacts of trolling behaviors. In particular, Kelly et al. (2006) note that those with fringe opinions, or opinions that are uncommon and controversial within a group, can very easily be confused with trolls. Despite being unpopular, these fringe opinions do not directly disrupt the community and can actually foster healthy debate. Kirman et al. (2012) describe how some forms of trolling and general mischief may actually contribute to a community by pushing boundaries and helping to renegotiate social norms, especially as the boundary between acceptable and unacceptable behavior is often fuzzy. Engaging in mischief can even signal a higher level of comfort in more tightly knit communities. Some forms of trolling may also increase group cohesion rather than act disruptively (Hardaker 2010). Thus, the ability to identify harmful forms of trolling, or actions that appear to be trolling, is an important skill that can aid online communities in distinguishing between behaviors that cause disruptions and those that may have benefits.

III. STUDY METHOD AND THEMES

3.1 Research Site and Preliminary Interviews

In this study, similar to the approach taken by Shachaf and Hara (2010), interviews were first conducted with active members of one large online community, deviantART, in order to better ground the research. Although best known for its thriving artistic community, deviantART also encourages participation in non-artistic areas of the site, like topical

forums. The site was founded in 2000, predating many other currently active social networking sites, and is often among the top 100 visited sites in the United States.[12] The large size and older age of the community provides an interesting platform to study trolling within an online community with well-established norms.

Participants for the initial survey were recruited through the site's random member search function, and 9 out of 30 members agreed to be interviewed through either private messages or a chat service. Of these, seven were female and two were male, with ages ranging from 19 to 37 at the time of interview. Additionally, two participants had previously been administrators for the site or were holding administrative positions at the time of interview. Interviewees were asked a number of open-ended questions regarding their observations of deviant behaviors as well as their own behaviors. Examples of specific questions are listed in Appendix A.

Although interviewees were initially asked broadly about any type of deviant behavior, every interviewee identified trolling as a deviant behavior they found to be pervasive, prompting the focus of this study. The interview responses identified a few key aspects of how trolling is viewed by the community, as shown in Table 8. Interviewees clearly identified at least two reasons for trolling: malicious intent toward the site or its members and the effort to provide entertainment. Several participants indicated that they themselves have trolled in the past, but there was an obvious disconnect between their own motivations and what they perceived as the motivations of other trolls. Specifically, these participants believed their own actions to be harmless or entertaining while the actions of

[12] Alexa rank, Dec. 2014.

others were viewed as malicious. This illustrates one of the key difficulties in understanding the intent of trolls. However, there appears to be a common understanding that differing opinions should not be considered trolling, although it can be difficult or even impossible to distinguish between the two. Furthermore, if there is trolling that is meant for entertainment, it is unclear whether the behavior can be considered entirely detrimental to the community.

Table 8. Interview Excerpts

Emergent Themes of Trolling	**Quotes from Interviews**
Prevalence of trolling	- "but trolling is now apart [sic] of the internet nature" - "[trolls] are the largest and most prominent problem […] on ANY social website" [administrator]
Ambiguity of identifying trolling	- "because of that context when a normal person tries to defend a view point or speak in opposition, instead of being taken seriously they are labeled as [a troll]" - "I was instantly dubbed a 'troll' because I had a differing opinion"
Motivations of trolling	
- Entertainment	- "[trolling] just gives me a good laugh every now and then" - "people intentionall [sic] try to be trolls it makes you take situations less seriously […] because of the hilarity factors"
- Malicious intent	- "trolling is just another form of bullying too" - "usually the troll would serverly [sic] depress their target and leave them feeling like a mess" - "malicious trolling comes with the intent of making the other party feel like crap and inciting their anger" [administrator]
Trolling perspective	
- Perception of self	- "I only troll condescending, self center [sic] people" - "Have I ever trolled? No, not consciously."
- Perception of others	- "usually they have such a big ego that anything you throw back at them is useless" [administrator] - "a random stranger is doing it just because of general douche baggery" - "they'll keep switching sides to keep you baited as long as possible for their own entertainment"

3.2 Web-crawled Data

To further investigate trolling behaviors within a forum structure, a web crawler was programmed to collect comments from forum discussions on deviantART. The web crawler was written in the Python programming language and utilized Scrapy, a web crawling framework. Specifically, the web crawler collected all comments on forum threads active within a span of three weeks from three sub-forums covering controversial topics (i.e., complaints, politics, and religion/philosophy). These three weeks spanned the latter half of December 2012 through the beginning of January 2013 and were times when the forums were especially active due to the holiday season. These particular sub-forums were chosen for their culture of debate and an increased likelihood of ambiguity within those debates. For each comment, the collected data included a unique comment ID, a unique thread ID, the user who posted the comment, the unique comment ID of a previous comment to which the comment replied (if applicable), and the comment itself. In this way, the original structure of comments within a thread was preserved so that comments could be evaluated within the context of a discussion.

Using the Grounded Theory approach (Glaser and Strauss 1967), a two-stage coding process was employed to understand how trolling was viewed within the community. First, during an open-coding stage, the collected data was organized by thread and filtered by a keyword search of the word "troll" and its cognates, resulting in 97 out of over 1000 total discussion threads being used for analysis and a total of 34427 comments across the 97 threads. Within each thread, each comment identifying or defending a potential troll was labeled as an identification comment. Then, each user who was

identified as possibly being a troll or defended as not being a troll was manually coded as a potential troll, and each comment by that user within the thread was coded as a potential troll comment. Furthermore, each comment replying to a potential troll comment was labeled as a response comment. Because of the difficulty in clearly defining trolling and due to variations in accepted community norms on different sites, we identify users as potential trolls based on the community's standards. In this way, members of the community are only labeled as a potential troll if another member of the community identifies them as such.

Through several iterations, and partly based on the interview responses, common themes from these sets of comments were identified and grouped into several categories, including methods of trolling that suspected trolls may have used, evidence that users may have used in identifying a troll, and reactions to potential troll comments. These categories became a part of the coding scheme presented in Appendix B, which contains examples of actual forum comments. During the second stage of structured coding, each comment in the dataset was then coded by a primary coder and a secondary coder using the coding scheme developed in the first stage. In addition to the coding of individual comments, some factors were coded at the thread level to gauge the overall impact of potential trolls on the thread. For example, at the comment level, potential troll comments were coded based on the methods of trolling used (i.e., lack of proof, inflammatory language, all caps, and personal attacks). However, an additional method of trolling, repetition of comments, could only be coded when looking at multiple comments by the potential troll within the thread. Other constructs coded at the thread level include a potential community consensus within

the thread on whether a suspected user was a troll, based on the number of users identifying the potential troll versus the number of users who defended the potential troll or continued with normal discussion, and whether administrative action was suggested or enacted.

IV. RESULTS AND DISCUSSION

An analysis of the inter-rater reliability between the primary and the secondary coders resulted in a Cohen's kappa of 0.79. As this is generally considered a fairly high level of agreement (Dewey 1983, Fleiss 1981), and because the primary coder was more conservative in the coding process (e.g., fewer "yes" evaluations than the secondary coder), the primary coder's coding has been used in subsequent analyses.

4.1 Identification Comments

There were a total of 489 comments coded in the dataset in which a user was accused of being a potential troll or defended as not being a troll. Of these, 354 comments were firm identifications of trolling in which the accuser appeared to firmly believe the suspected member was a troll. In 109 comments, the accuser appeared uncertain as to whether the other member was a troll but tentatively identified them as such. For example, during a heated argument, one user may suggest that the other user's position is so ridiculous or illogical that it sounds like trolling (e.g., "you're either trolling or an idiot"). Furthermore, there were 23 cases in which a user firmly stated that someone was not a troll, and 3 comments in which a user tentatively defended a suspected troll. However, a closer look at the identifying comments revealed 15 comments in which users were identifying or

defending themselves. 11 of these comments were users declaring themselves to be trolls, and 4 were users firmly denying trolling.

In a majority of identification comments (322 comments), users gave at least one reason why they thought someone was a troll. The identification comments that did not give any reasons were simple statements like "you're a troll." In all 15 cases of self-identification, the users did not state any additional evidence to support their claims as it may be assumed they are certain of their own intentions. Table 9 shows the number of comments using various types of evidence to support a claim of trolling or not trolling. Out of the 474 comments in which users identified another member as a troll or defended them as not a troll, more than half (59%) pointed to some content within the comments posted by the suspected troll as evidence. For example, some users quoted specific passages from the suspected troll's comments.

More rarely, users sometimes drew conclusions from the suspected troll's profile page (4%) or a past history of trolling behavior (7%). For example, some members are widely known to be considered trolls by the community and a user may point out the suspected troll's reputation. A few users drew from past personal experiences with trolls (3%) and likened a suspected troll's behavior to those examples. Additionally, some users pointed to identification comments made by other users (2.5%) to support their position. In these cases, users may justify their identification by suggesting an agreement within the community. Approximately 8% of methods used multiple types of evidence as justification, though a majority (285 comments) used just one type of evidence. There was no significant

difference in the types of evidence used for different types of identification (i.e., identifying or defending potential trolls) based on a chi-squared goodness of fit test with $p > 0.1$.

Table 9. Evidence Used to Support Identification of Trolling or Not

Type of Evidence	Number of Comments
Content of suspected troll's posts	281 (59.28%)
Cues from the community	12 (2.53%)
Profile of suspected troll	18 (3.80%)
Past behavior of suspected troll	34 (7.17%)
Personal experience of user	16 (3.38%)

Of the 474 identification comments addressed to other users, 309 were directed to the suspected troll. In other words, most identifications were direct accusations. 124 identification comments were addressed to another member of the community. In these cases, users may have been debating amongst themselves whether a third party was a troll. Some identifications were also made to warn another user that they may have been dealing with a troll or to inform them their suspicions of trolling were incorrect. In the remaining cases, users posted comments that appeared to be addressed to the general public. For instance, a user may post a reply to a thread warning that the original threat poster was a troll and that no one should respond to the thread.

In response to these identification comments, there were 57 comments in which suspected trolls defended themselves by denying any trolling, especially in cases where the term "troll" was used as a form of attack. In addition, there were 6 comments in which suspected trolls admitted to being trolls or to having posted trolling messages. This may have been done in order to diffuse an argument or to end a conversation. However, a majority of identification comments were ignored by the suspected trolls who were

identified. In 149 cases, the suspected trolls replied directly to their accusers but made no mention of the accusation. Perhaps, by not responding to accusations, suspected trolls may continue to create uncertainty and provoke further response from the community.

4.2 Potential Troll Comments

An initial analysis based on community-identified suspected trolls resulted in a total of 158 potential cases of trolling within the dataset. Some of the 97 threads had multiple suspected trolls, and each suspected troll was counted as a separate potential case of trolling. Moreover, a user who was accused of trolling in one thread may not have been trolling in other threads, so each instance of an accused troll within a thread was coded separately from the user's activity on other threads. Thus, the 158 potential cases of trolling are unique for a specific thread and suspected troll combination. Comments a potential troll made within a thread in which the user was accused were then coded as potential troll comments, resulting in a total of 6837 potential troll comments. In this way, only comments which the community could have identified as trolling were considered for analysis.

Of the total potential troll comments, 82.42% were coded as not using any of the methods of trolling identified in the coding scheme (i.e., lack of proof, inflammatory language, all caps, and personal attacks). There was no clear method of trolling used in a majority of comments by suspected trolls. Another 3.31% used more than one method of trolling. Several reasons may account for the low number of potential trolling comments that actually used an identified trolling method. First, although attempts were made to capture all possible trolling methods, the coding scheme may not be able to completely

capture all methods used by potential trolls. However, a manual inspection of the potential troll comments did not reveal any obvious trolling methods that were missed by the coding scheme. Alternatively, although trolls may make some comments that use certain identifiable methods, they may often engage in conversation that looks similar to legitimate discussion and may not troll all the time. In addition, some users might have been misidentified as a potential troll by the community, and there may not be a clear consensus.

In order to filter out comments that may have been misidentified as trolling, comments in which at least one method of trolling was identified were then further evaluated for trolling methods. Of these 1202 out of 6837 potential troll comments, the most commonly used methods were personal attacks (640 comments) and comments that made exaggerated claims without evidence or arguments to support the claims (478 comments). Personal attacks are likely to elicit a response from another individual. Exaggerated claims without proof may similarly elicit counter arguments or comments asking for proof. Because these methods are likely to evoke a response, the users who post these types of comments may be more likely to be suspected of trolling. Inflammatory language (265 comments) may also elicit responses, but to a lesser degree due to the debate nature of these forums and the prevalence of such language in general. Typing comments in all caps (60 comments) may also be used as a way to emphasize a point. Thus, the use of inflammatory language or all caps may be seen as common enough to not be considered trolling. Furthermore, potential trolls may find that these methods do not elicit the same types of responses and are therefore ineffective as trolling methods.

Potential troll comments were then aggregated for each of the 158 instances of potential trolling by different users within different threads in order to look at the thread-level variable of repetition as a method of trolling. Repetition is coded when multiple comments by a suspected troll use the same exact phrasing. The instances of trolling were then further aggregated to look at individual users who were suspected of trolling at least once in the 97 threads, resulting in a total of 105 unique suspected trolls across all threads. Table 10 shows a distribution of the trolling methods used by suspected trolls, where each troll was counted once if they used a particular method of trolling at any point in any of their comments that were considered potential troll comments. Aggregated at the user level, only 23% of suspected trolls did not use any of the identified methods of trolling. Over half (54%) used more than one method of trolling. As seen with the analysis of individual comments, the use of all caps is a less common method of trolling (only used by 15% of suspected trolls), potentially because it is more often used as emphasis. Furthermore, repetition (11%) is uncommon among suspected trolls. Copied and pasted responses may be too obvious as troll comments and may, therefore, not be effective in eliciting a reply. Meanwhile, personal attacks were used by a majority of suspected trolls (64%), possibly due to the likelihood of provoking angry responses.

Table 10. Trolling Methods Used by Suspected Trolls

Method of Trolling	Number of Suspected Trolls
Lack of proof	45 (42.86%)
Inflammatory language	46 (43.81%)
All caps	16 (15.24%)
Personal attacks	67 (63.81%)
Repetition	12 (11.43%)

Another important aspect of potential troll comments is the intended audience of the comment. Theoretically, a troll may achieve the goal of irritating more community members by "casting a wider net," so to speak, and targeting the community in general. However, the data showed that an overwhelming majority of potential troll comments in which at least one troll method was identified appeared to be directed to an individual, rather than the community. Only 1.83% of these potential troll comments were open-ended comments that solicited responses from the community. Out of the 105 suspected trolls, 36 of them posted at least one comment directed to the community, and all of them posted multiple comments to individuals. Looking closer at the comments directed to the community, a majority of them were comments in which the suspected troll was the original thread poster. Although a troll may set up a thread to create debate and elicit emotional responses, some form of follow-up is often needed to continue to keep individuals baited. Suspected trolls may make the effort to post more comments targeting individuals who are likely to respond, and in this way, the thread continues to be active.

4.3 How Consensus of Trolling Is Reached

In order to understand how trolling may be identified, we need to look at how a community reaches a consensus on whether a user is a troll or not. To that end, coders evaluated each instance of potential trolling in a thread by looking at a combination of the identification comments accusing or defending a suspected troll and whether other users continued to hold normal discourse with that member. Because consensus was coded at the thread level for each suspected troll, the 158 cases of trolling by unique users within unique threads

was used for this analysis. Of the 158 cases of potential trolling, over half (92) resulted in no clear consensus. This may be in part due to a large number of accusations of trolling that occurred within a dialogue between two individuals where other community members had no input, and thus no ability to reach a consensus. In some instances, members debated among themselves whether a user was a troll but could not agree. In other instances, a few members may accuse a suspected troll, but other users may continue to hold normal conversations with the suspected troll or tentatively defend the suspected troll. Generally, if there was no clear consensus of trolling or not trolling, the instance of trolling was coded as no consensus.

13 cases were concluded to be a consensus of trolling, and 53 cases had a consensus of not trolling. In cases where a consensus appeared to be reached, it was more likely that the community decided a suspected troll was actually not trolling. This type of consensus may be reached by a large number of users defending the suspected troll against an accusation of trolling. It may also be reached by the community ignoring an accusation of trolling and continuing normal discourse, thus passively indicating that they believed the user was genuine. The community appeared to err on the side of caution and was more likely to defend a user from a wrong accusation than to confirm an existing accusation. However, when the community decided that a user was indeed a troll, many members typically posted accusations of trolling in quick succession and conversations with the suspected troll in the thread faded. This suggests that there are indeed cases of suspected trolling that appear to be obvious to the community.

Table 11. Consensus of Trolling Based on Evidence Used to Identify Trolling

Evidence Used in Identification of Suspected Troll		Consensus			p-value
		Troll	None	Not Troll	
Content of posts	Used	10	70	37	0.5924
	Not used	3	16	13	
Cues from the community	Used	3	5	1	0.0175
	Not used	10	81	49	
Profile of suspected troll	Used	4	5	1	0.0010
	Not used	9	81	49	
Past behavior of suspected troll	Used	5	15	7	0.1210
	Not used	8	71	43	
Personal experience of user	Used	2	7	3	0.5409
	Not used	11	79	47	

To further determine how consensus is reached, an analysis of the types of evidence used for comments identifying a suspected troll was conducted as a possible predictor of consensus. For each instance of suspected trolling, aggregate thread-level data included whether each type of evidence was used in at least one comment identifying the suspected troll and what the final consensus of the community was on whether the instance appeared to be actual trolling. Table 11 shows a comparison between the consensus distributions for each type of cited evidence. A chi-squared test was also performed to determine if the usage or lack of usage of a certain type of evidence had any bearing on the final consensus. Results show that the most common evidence used in identifying trolls (content of posts) was actually the weakest in predicting the eventual consensus. Conversely, less common evidence was more effective when actually present. In particular, cues from the community and the profile of the suspected troll appeared to be significant ($p < 0.05$) in determining what kind of consensus was reached. The use of either as evidence to support a trolling accusation led to a greater chance of reaching the consensus that a suspected troll was a

troll in comparison to the much greater chance of no consensus or a "not troll" consensus when these types of evidence were not used. This is not surprising, since cues from the community suggest a community discussion of the particular case of trolling. Also, a suspected troll's profile page may be offered as unbiased evidence of trolling, such as noting that the suspected troll is a new member with no prior history. These forms of evidence may result in a greater likelihood of reaching a firm consensus.

Table 12. Consensus of Trolling Based on Method of Trolling

Method of Trolling		Consensus			p-value
		Troll	None	Not Troll	
Lack of proof	Used	9	41	24	0.2393
	Not used	4	51	29	
Inflammatory language	Used	6	30	20	0.5779
	Not used	7	62	33	
All caps	Used	3	10	4	0.2692
	Not used	10	82	49	
Personal attacks	Used	9	57	33	0.8768
	Not used	4	35	20	
Repetition	Used	1	12	4	0.5497
	Not used	12	80	49	

Community consensus of trolling was also tested against the methods that may have been used by suspected trolls. Table 12 examines whether a troll's usage of a noticeable method of trolling is likely to lead to consensus that they are a troll. In this case, using or not using a particular method of trolling did not appear to lead to significantly different consensuses. There may not be a single method that very obviously points to trolling, and it could be difficult to reach a clear consensus based solely on what types of methods are used.

Table 13. Consensus of Trolling Based on Percent of Potential Troll Comments

Percent of Comments with One or More Troll Methods	Consensus		
	Troll	None	Not Troll
0-25%	6 (5.94%)	54 (53.47%)	41(40.59%)
25-50%	3 (9.68%)	21 (67.74%)	7 (22.58%)
50-75%	2 (11.11%)	11 (61.11%)	5 (27.78%)
75-100%	2 (25%)	6 (75%)	0 (0%)

To further investigate the usage of trolling methods, each instance of trolling was then divided based on the percentage of the suspected troll's comments that used at least one trolling method. As previously discussed, a large number of suspected troll comments did not use any identifiable method of trolling. The percentage of comments using trolling methods that a suspected troll made within a thread was divided into quarters. Table 13 indicates that suspected trolls who make more comments containing at least one method of trolling have a greater chance of being labeled a troll by the community. Suspected trolls who used trolling methods in at least 75% of their comments were not cleared by the community. Meanwhile, suspected trolls with a small percentage of comments that use trolling methods are more likely to be labeled not a troll. The more comments that included trolling methods identified in the coding, the more likely the community decided the suspected trolls were, in fact, trolling. A chi-squared test of consensus based on the percentage of comments using trolling methods resulted in a p-value of 0.1090, indicating moderate significance. However, when suspected trolls with only 0-25% of comments using trolling methods are compared to users who had at least 25% of comments with trolling methods, the difference in predicting consensus is much more significant with a chi-squared test resulting in $p = 0.0299$. This suggests that users in the 0-25% category, or

users who rarely post comments using trolling methods, are much more likely to be considered legitimate members and not trolls by the community.

4.4 Impact of Trolling

It has traditionally been assumed that trolling is a negative behavior and that trolls are harmful to an online community. In this study, several measures designed to look at the actual impact of perceived trolling were also coded.

First, the comments that attempted to identify potential trolls were coded for a tone of voice that might indicate whether the accuser viewed trolling as a positive or a negative behavior. Most identification comments did not directly indicate a specific feeling toward trolling itself. However, users occasionally, in 48 comments, described trolling in a negative light, especially in explaining why a suspected troll comment was bothersome. Additionally, 10 comments indicated an enjoyment of trolling, either in finding a potential troll comment to be humorous or in complaining that a suspected troll was not as amusing as other trolls may be. This suggests that, though trolling is typically viewed as a disturbance to the community, there are cases of trolling that can be lighthearted and do not negatively impact the community.

In looking at the community's responses to suspected trolls, there were two cases in which users announced their intentions to leave the thread due to the suspected trolling. Furthermore, coders attempted to detect a tone of voice or emotion in 8346 comments that replied to suspected troll comments. Of these, 288 comments indicated negative emotions while only 91 indicated a positive emotion. However, although more negative emotions

were detected and two members withdrew from conversation, these negative effects are possibly due to the debate nature of the forums, rather than directly resulting from actual trolling. Nonetheless, even perceived trolling could be harmful to the community. Though the comments indicating a positive emotion suggest that suspected trolls may bring value to the community, this is largely not the case when trolling is expressly identified.

In some rare cases, especially when a suspected troll was the original poster of a thread, community members may post comments calling for a moderator to take action, usually by locking the thread. If a thread is clearly not contributing to free discussion and debate, a moderator may lock the thread to prevent further comments from being posted. On this particular site, moderators appeared to avoid banning members due to forum posts, as comments may be misconstrued. In total, 16 threads included comments that called for moderator action, and 3 of these threads were locked due to suspected trolling.

Despite calls for action and the locking of a few threads, there was no significant correlation between these actions and a community consensus of whether a member was trolling. There may be a couple of reasons behind this. First, as previously explained, calling a member a "troll" is sometimes used as an attack during the course of an argument. Similarly, calling for moderator action may be used to indicate that a user no longer wishes to debate a suspected troll rather than being a genuine request for a moderator to intervene. Second, moderator actions such as locking a thread are only useful when the suspected troll is the original poster of the thread. In cases where the suspected troll is a participant within the thread, there are not a lot of options for moderators to intervene. This reinforces the

idea that management of trolling is a delicate process, and that uncertainty in a user's intentions may make it more difficult to mitigate negative aspects of trolling.

V. PRACTICAL IMPLICATIONS

At a high level, the results from this study indicate that trolling is as difficult to define and identify as previously expected. It was easier for the community to come to a consensus on users who were not trolling than on clear cases of trolling. Less used evidence of trolling, such as cues from the community and the profile of a suspected troll, were more useful in helping to come to a consensus.

As shown in the data analysis, there is not a lot of consistency in the ways in which trolling may be accurately identified. Poe's Law is famously used to suggest that, within online arenas, it is always difficult to distinguish between someone who has an extreme viewpoint and someone who is parodying an extreme viewpoint (Aikin 2013). Because it is impossible to be completely certain of a suspected troll's intentions, sites must be careful in how they decide if a member exhibits behavior that can be classified as trolling. In the particular community of deviantART, members already appear to be cautious in coming to a consensus that a user may be a troll, often defending other users against accusations of trolling.

As multiple interviewees mentioned, differing opinions may sometimes cause another member to suggest the user is a troll, despite the user not having exhibited other trolling behaviors. Because of the types of arguments held on the forums used within this study, the term "troll" may often be used as an ad hominem attack on another community

member during the course of heated debate. In these cases, accusing another member of being a troll was a form of insult, suggesting an unreasonable argument. Thus, many suspected trolls did not make comments that fell into any category of trolling method. These suspected trolls were then also more likely to be eventually cleared as not being trolls. Community moderators may be able to use the amount of comments posted by a suspected troll that actually contain some form of recognizable trolling behavior, as coded in this study, in order to help determine if the suspected troll may be innocent of wrongdoing. Suspected trolls who are defended by other community members are also more likely to be innocent, and administrators might use profile information to confirm their suspicious.

Contrary to previous fears (Herring et al. 2002, Kelly et al. 2006), the discussion community in this study seemed to allow for diverse opinions to be expressed, and members defended the wrongly accused when they believed contributions were genuine. In addition, members may bond through helping each other to identify and "battle" potential trolls, often debating amongst themselves or warning each other of potential trolling. General attitudes toward trolling were negative, as expected, but there were also a number of users who appeared to enjoy some trolling and might even encourage "successful" trolling as long as it was humorous. This suggests that trolling might indeed be a form of entertainment, particularly as a way to lend lightness to serious discussions. However, perceived trolling also elicited very negative emotions and, in a few cases, caused members to leave discussions. Thus, moderators still need to monitor these behaviors and prevent them from damaging the community. In this study, community

consensus was used to determine whether a user may be a real troll, and moderators may also use this tactic to allow a community to decide whether certain behaviors are acceptable or not. Members who post many comments that use methods of trolling, and who the community decides are trolls, may then be banned to prevent further undue disruption to conversations.

VI. LIMITATIONS AND FUTURE WORK

This study only examines the identification of trolling as it is defined within one online community, deviantART. In order to provide more generalizable results regarding trolling behaviors and responses to trolls, additional studies should be conducted in other large online communities with discussion platforms to look for similar patterns of behavior and a common consensus (or lack thereof) on what is trolling and how to manage it.

One potential limitation of this study is the inability to identify cases where trolling was not suspected by a member of the community. Because this study relies on the community to point out cases of potential trolling, it is possible that some users who intended to troll are not included in the study if no member accuses them of trolling. However, by definition, deviant behaviors deviate from accepted norms. In online communities, these norms are decided by administrators and members of the community, making the opinions and response of the community the most important aspects in deciding whether a behavior can be considered deviant. If the community is unperturbed by the actions of someone who is attempting to be a troll, those actions may not be considered "successful" trolling and are therefore immaterial from an administrator standpoint.

Because many users accuse others of being potential trolls by default in ambiguous cases, it is unlikely that instances of "successful" trolling go unnoticed by the community. In addition, administrators only find the cases of trolling that actually impact the community negatively to be relevant when moderating discussions. Rather than attempting to create an objective rule to define and identify trolls across internet communities, perhaps it is more practical to allow for subjective judgment by the communities themselves. By using a community's definition of trolling and its identification tactics, future work can focus on how to best minimize the impact of these negative behaviors.

CHAPTER 5

Summary and Concluding Remarks

Through the examination of contributions to three types of online communities, this dissertation has attempted to understand the causes and effects of various online problems that may affect those contributions. Regardless of the type of online community, there is a need for users to participate and positively contribute in order for the community to thrive. The detrimental effects of negative behaviors on communities and their members need to be curbed in an accurate manner to prevent discouraging possibly beneficial behaviors, and understanding the causes of negative behaviors in the first place is crucial to developing methods to prevent them.

Negative behaviors can include low quality work in online marketplaces for work that cause requesters to be wary and less likely to pay, lack of contribution and withdrawal behaviors in online volunteer groups that cause projects to lose momentum and valuable contributing members, and trolling behaviors in online social communities that both annoy and distract other members from legitimate conversations and participation. Through better design of tasks, both higher quality and higher quantities of work can be encouraged in online marketplaces for work. With better management of volunteers' time and involvement, continued contribution can co-exist with lower withdrawal rates in online volunteer groups. By allowing community members to debate the motives of suspected trolls and defend users who may be falsely accused, online social network communities may take countermeasures against trolling without alienating legitimate contributors.

Understanding how these different negative behaviors manifest in different types of online communities adds to the theory and practice of identifying and moderating such problems in order to improve the overall online experience. This, in turn, allows communities to focus on the many positive contributions of online interactions. Future work may continue to discover improved methods to reduce negative impacts while encouraging valuable participation within online communities.

BIBLIOGRAPHY

Adler, P. A. and P. Adler. 2008. The Cyber Worlds of Self-Injurers: Deviant Communities, Relationships, and Selves. *Symbolic Interaction*, 31(1), 35-56.

Aikin, S. F. 2013. Poe's Law, Group Polarization, and Argumentative Failure in Religious and Political Discourse. *Social Semiotics*, 23(3), 301-317.

Akers, R. L., M. D. Krohn, L. Lanza-Kaduce, M. Radosevich. 1979. Social Learning and Deviant Behavior: A Specific Test of a General Theory. *American Sociological Review*, 44(4), 636-655.

Akkaya, C., A. Conrad, J. Wiebe, R. Mihalcea. 2010. Amazon Mechanical Turk for Subjectivity Word Sense Disambiguation. In *Proc. NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 195-203.

Albanese, R. and D. D. Van Fleet. 1985. Rational Behavior in Groups: The Free-Riding Tendency. *Academy of Management Review*, 10(2), 244-255.

Arnold, H. J. and D. C. Feldman. 1982. A Multivariate Analysis of the Determinants of Job Turnover. *Journal of Applied Psychology*, 67(3), 350-360.

Arthur, J. B. 1994. Effects of Human Resource Systems on Manufacturing Performance and Turnover. *Academy of Management Journal*, 37(3), 670-687.

Atwood, J. 2011. Suspension, Ban or Hellban? Coding Horror, June 4, 2011. http://blog.codinghorror.com/suspension-ban-or-hellban/. Retrieved 2/5/12.

Bakker, A. B., K. I. Van Der Zee, K. A. Lewig, M. F. Dollard. 2007. The Relationship between the Big Five Personality Factors and Burnout: A Study among Volunteer Counselors. *The Journal of Social Psychology*, 146(1), 31-50.

Barack, L. 2005. Wary of Wikipedia. *School Library Journal*, 51(10), 30.

Barrick, M. R. and M. K. Mount. 2006. The Big Five Personality Dimensions and Job Performance: A Meta-Analysis. *Personnel Psychology*, 44(1), 1-26.

Barron, D. N., E. West, M. T. Hannan. 1994. A Time to Grow and a Time to Die: Growth and Mortality of Credit Unions in New York City, 1914-1990. *American Journal of Sociology*, 100(2), 381-421.

Becker, G. S. 1965. A Theory of the Allocation of Time. *The Economic Journal*, 75(299), 493-517.

Beehr, T. A. and N. A. Gupta. 1978. Note on the Structure of Employee Withdrawal. *Organizational Behavior and Human Performance*, 21(1), 73-79.

Bryk, A. S. and S. W. Raudenbush. 1992. Hierarchical Linear Models for Social and Behavioural Research: Applications and Data Analysis Methods. Newbury Park, CA: Sage Publications.

Buckels, E. E., P. D. Trapnell, D. L. Paulhus. 2014. Trolls Just Want to Have Fun. *Personality and Individual Differences*, 67, 97-102.

Burnham, K. P. and D. R. Anderson. 2004. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods Research*, 33, 261-304.

Butler, B. S. 2001. Membership Size, Communication Activity, and Sustainability: A Resource-Based Model of Online Social Structures. *Information Systems Research*, 12(4), 346-362.

Butler, B. S., L. Sproull, S. Kiesler, R. E. Kraut. 2007. Community Effort in Online Groups: Who Does the Work and Why? In *Leadership at a Distance*, S. Weisband and L. Atwater (Eds.), New York: Lawrence Erlbaum Associates.

Callison-Burch, C. 2009. Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk. In *Proc. EMNLP 2009*, 286–295.

Callison-Burch, C. and M. Dredze. 2010. Creating Speech and Language Data with Amazon's Mechanical Turk. In *Proc. NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 1-12.

Cameron, J. and D. Pierce. 1994. Reinforcement, Reward, and Intrinsic Motivation: A Meta-Analysis. *Review of Educational Research*, 64(3), 363-423.

Chandler, D. and A. Kapelner. 2012. Breaking Monotony with Meaning: Motivation in Crowdsourcing Markets. Working Paper. University of Chicago.

Chaturvedi, A. R., D. R. Dolk, P. L. Drnevich. 2011. Design Principles for Virtual Worlds. *MIS Quarterly*, 35(3), 673-684.

Chen, J., Y. Ren, J. Riedl. 2010. The Effects of Diversity on Group Productivity and Member Withdrawal in Online Volunteer Groups. In *Proc. CHI 2010*.

Cordes, C. L. and T. W. Dougherty. 1993. A Review and an Integration of Research on Job Burnout. *Academy of Management Review*, 18(4), 621-656.

Cosley, D., D. Frankowski, L. Terveen, J. Riedl. 2006. Using Intelligent Task Routing and Contribution Review to Help Communities Build Artifacts of Lasting Value. In *Proc. CHI 2006*, 1037-1046.

Cotton, J. L. and J. M. Tuttle. 1986. Employee Turnover: A Meta-Analysis and Review with Implications for Research. *Academy of Management Review*, 11(1), 55-70.

Cross, R. and J. N. Cummings. 2004. Tie and Network Correlates of Individual Performance in Knowledge-Intensive Work. *Academy of Management Journal*, 47(6), 928-937.

Cuddy, L. and J. Nordlinger. 2009. World of Warcraft and Philosophy: Wrath of the Philosopher King. *Popular Culture and Philosophy*, 45.

Daniel, S. L. and E. I. Diamant. 2008. Network Effects in OSS Development: The Impact of Users and Developers on Project Performance. In *Proc. ICIS 2008*.

Deci, E. L., R. Koestner, R. M. Ryan. 1999. A Meta-Analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation. *Psychological Bulletin*, 125(6), 627-668.

Dennis, A. R. and S. T. Kinney. 1998. Testing Media Richness Theory in New Media: The Effects of Cues, Feedback, and Task Equivocality. *Information Systems Research*, 9(3), 256-274.

Dewey, M. E. 1983. Coefficients of Agreement. *British Journal of Psychiatry*, 143, 487-489.

Dishion, T. J., K. A. Dodge, J. E. Lansford. 2008. Deviant by Design: Risks Associated with Aggregating Deviant Peers into Group Prevention and Treatment Programs. *The Prevention Researcher*, 15(1), 8-11.

Doig, W. 2008. Homophobosphere. The Advocate, Feb.-Mar. http://www.advocate.com/article.aspx?id=22197. Retrieved 12/16/11.

Downs, J. S., M. B. Holbrook, S. Sheng, L. F. Cranor. 2010. Are Your Participants Gaming the System? Screening Mechanical Turk Workers. In *Proc. 28th International Conference on Human Factors in Computing Systems*.

Evans, R. D. 2011. Examining the Informal Sanctioning of Deviance in a Chat Room Culture. *Deviant Behavior*, 22(3), 195-210.

Farmer, S. M. and D. B. Fedor. 1999. Volunteer Participation and Withdrawal: A Psychological Contract Perspective on the Role of Expectations and Organizational Support. *Nonprofit Management and Leadership*, 9, 349-368.

Feng, D., S. Besana, R. Zajac. 2009. Acquiring High Quality Non-Expert Knowledge from On-Demand Workforce. In *Proc. 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, 51-56.

Fleiss, J. L. 1981. Statistical Methods for Rates and Proportions, 2nd Ed., New York: John Wiley.

Frey, B. S. and M. Osterloh. 2001. Successful Management by Motivation: Balancing Intrinsic and Extrinsic Incentives. New York, USA: Springer.

Frick, P. J. and S. F. White. 2008. Research Review: The Importance of Callous-Unemotional Traits for Developmental Models of Aggressive and Antisocial Behavior. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, 49(4), 359-375.

Gefen, D. and E. Carmel. 2008. Is the World Really Flat? A Look at Offshoring in an Online Programming Marketplace. *MIS Quarterly*, 32(8), 367-384.

Glaser, B. G. and A. L. Strauss. 1967. Discovery of Grounded Theory. Mill Valley, CA: Sociology Press.

Grant, A. M. 2008a. Does Intrinsic Motivation Fuel the Prosocial Fire? Motivational Synergy in Predicting Persistence, Performance, and Productivity. *Journal of Applied Psychology*, 93(1), 48-58.

Grant, A. M. 2008b. The Significance of Task Significance: Job Performance Effects, Relational Mechanisms, and Boundary Conditions. *Journal of Applied Psychology*, 93(1), 108-124.

Grant, A. M., E. M. Campbell, G. Chen, K. Cottone, D. Lapedis, K. Lee. 2007. Impact and the Art of Motivation Maintenance: The Effects of Contact with Beneficiaries on Persistence Behavior. *Organizational Behavior and Human Decision Processes*, 103, 53-67.

Griffeth, R. W., P. W. Hom, S. Gaertner. 2000. A Meta-Analysis of Antecedents and Correlates of Employee Turnover: Update, Moderator Tests, and Research Implications for the Next Millennium. *Journal of Management*, 26(3), 463-488.

Hackman, J. R. and G. R. Oldham. 1976. Motivation Through the Design of Work: Test of a Theory. *Organizational Behavior and Human Performance*, 16(2), 250-279.

Hackman, J. R. and G. R. Oldham. 1980. Work Redesign. Reading, MA: Addison-Wesley Publishing.

Halfaker, A., A. Kittur, J. Riedl. 2011. Don't Bite the Newbies: How Reverts Affect the Quantity and Quality of Wikipedia Work. In *Proc. WikiSym 2011*.

Hardaker, C. 2010. Trolling in Asynchronous Computer-Mediated Communication: From User Discussions to Academic Definitions. *Journal of Politeness Research*, 6(2), 215-242.

Han, S. and B. J. Kim. 2008. Network Analysis of an Online Community. *Physica A*, 387(23), 5946-5951.

Hawk, T. F. and A. J. Shah. 2007. Using Learning Style Instruments to Enhance Student Learning. *Decision Sciences Journal of Innovative Education*, 5(1), 1-19.

Haythornthwaite, C. 2009. Crowds and Communities: Light and Heavyweight Models of Peer Production. In *Proc. HICSS 2009*.

Herring, S., K. Job-Sluder, R. Scheckler, S. Barab. 2002. Searching for Safety Online: Managing "Trolling" in a Feminist Forum. *The Information Society*, 18(5), 371-384.

Hofmann, D. A., R. Jacobs, S. J. Gerras. 1992. Mapping Individual Performance Over Time. *Journal of Applied Psychology*, 77(2), 185-195.

Howe, J. 2006. The Rise of Crowdsourcing. *Wired*, 14(6), June 2006.

Howe, J. 2008. Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business. New York: Crown Business.

Huang, E., H. Zhang, D. C. Parkes, K. Z. Gajos, Y. Chen. 2010. Toward Automatic Task Design: A Progress Report. In *Proc. ACM SIGKDD Workshop on Human Computation*.

Humphrey, S. E., J. D. Nahrgang, F. P. Morgeson. 2007. Integrating Motivational, Social, and Contextual Work Design Features: A Meta-Analytic Summary and Theoretical Extension of the Work Design Literature. *Journal of Applied Psychology*, 92(5), 1332-1356.

Ipeirotis, P. G., F. Provost, J. Wang. 2010. Quality Management on Amazon Mechanical Turk. In *Proc. ACM SIGKDD Workshop on Human Computation*.

Irani, L. 2009. Agency and Exploitation in Mechanical Turk. In *Proc. Internet as Playground and Factory Conference*.

Jarvenpaa, S. L. and A. Majchrzak. 2010. Vigilant Interaction in Knowledge Collaboration: Challenges of Online User Participation Under Ambivalence. *Information Systems Research*, 21(4), 773-784.

Jenkins, G. D., A. Mitra, N. Gupta, J. D. Shaw. 1998. Are Financial Incentives Related to Performance? A Meta-Analytic Review of Empirical Research. *Journal of Applied Psychology*, 83(5), 777-787.

Kaisser, M. and J. B. Lowe. 2008. Creating a Research Collection of Question Answer Sentence Pairs with Amazon's Mechanical Turk. In *Proc. Fifth International Conference on Language Resources and Evaluation (LREC-2008)*.

Kalyuga, S., P. Chandler, J. Sweller. 2000. Incorporating Learner Experience into the Design of Multimedia Instruction. *Journal of Educational Psychology*, 92(1), 126-136.

Kelly, J., D. Fisher, M. Smith. 2006. Friends, Foes, and Fringe: Norms and Structure in Political Discussion Networks. In *Proc. 2006 International Conference on Digital Government Research*, 412-417.

Kirman, B., C. Linehan, S. Lawson. 2012. Exploring Mischief and Mayhem in Social Computing or How We Learned to Stop Worrying and Love the Trolls. In *Proc. CHI EA 2012*, 121-130.

Kittur, A. and R. E. Kraut. 2008. Harnessing the Wisdom of Crowds in Wikipedia: Quality Through Coordination. In *Proc. CSCW 2008*.

Kittur, A., and R. E. Kraut. 2010. Beyond Wikipedia: Coordination and Conflict in Online Production Groups. In *Proc. CSCW 2010*.

Kittur, A., B. Suh, B. A. Pendleton, E. H. Chi. 2007a. He Says, She Says: Conflict and Coordination in Wikipedia. In *Proc. CHI 2007*.

Kittur, A., E. Chi, B. A. Pendleton, B. Suh, T. Mytkowicz. 2007b. Power of the Few vs. Wisdom of the Crowd: Wikipedia and the Rise of the Bourgeoisie. In *Proc. Alt.CHI at CHI 2007*.

Kittur, A., E. H. Chi, B. Suh. 2008. Crowdsourcing User Studies with Mechanical Turk. In *Proc. SIGCHI Conference on Human Factors in Computing Systems*.

Lakhani, K. R. and R. Wolf. 2005. Why Hackers Do What They Do: Understanding Motivation and Effort in Free/Open Source Software Projects. In *Perspectives on Free and Open Source Software*, J. Feller et al. (Eds.), Cambridge, MA: MIT Press.

Lakhani, K. R. and K. J. Boudreau. 2009. How to Manage Outside Innovation. *MIT Sloan Management Review*, 50(4).

Lampe, C., E. Johnston, P. Resnick. 2007. Follow the Reader: Filtering Comments on Slashdot. In *Proc. SIGCHI 2007*.

Little, G., L. B. Chilton, M. Goldman, R. C. Miller. 2010. Exploring Iterative and Parallel Human Computation Processes. In *Proc. ACM SIGKDD Workshop on Human Computation*.

Lyubomirsky, S. and S. Nolen-Hoeksema. 1995. Effects of Self-Focused Rumination on Negative Thinking and Interpersonal Problem Solving. *Journal of Personality and Social Psychology*, 69(1), 176-190.

Markus, M. L. 1987. Toward a "Critical Mass" Theory of Interactive Media. *Communication Research*, 14(5), 491-511.

McEvoy, G. M. and W. F. Cascio. 1987. Do Good or Poor Performers Leave? A Meta-Analysis of the Relationship between Performance and Turnover. *Academy of Management Journal*, 30(4), 744-762.

McPherson, J. M., P. A. Popielarz, S. Drobnic. 1992. Social Networks and Organizational Dynamics. *American Psychological Review*, 57(2), 153-170.

McPherson, M., L. Smith-Lovin, J. M. Cook. 2001. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27, 415-444.

Mitchell, T. R., B. C. Holtom, T. W. Lee, C. J. Sablynski, M. Erez. 2001. Why People Stay: Using Job Embeddedness to Predict Voluntary Turnover. *Academy of Management Journal*, 44(6), 1102-1121.

Morgeson, F. P. and M. A. Campion. 2003. Work Design. In *Handbook of Psychology: Industrial and Organizational Psychology*, W. Borman et al. (Eds.), 12, 423-452. Hoboken, NJ: Wiley.

Mullen, E. and J. Nadler. 2008. Moral Spillovers: The Effect of Moral Violations on Deviant Behavior. *Journal of Experimental Social Psychology*, 44, 1239-1245.

O'Reilly, C. A., III, D. F. Caldwell, W. P. Barnett. 1989. Work Group Demography, Social Integration, and Turnover. *Administrative Science Quarterly*, 34(1), 21-37.

Phillips, W. 2015. This is Why We Can't Have Nice Things: Mapping the Relationship Between Online Trolling and Mainstream Culture. Cambridge, MA: MIT Press.

Poor, N. 2005. Mechanisms of an Online Public Sphere: The Website Slashdot. *Journal of Computer-Mediated Communication*, 10(2), article 4.

Preece, J. 2000. Online Communities: Designing Usability and Supporting Sociability. Chichester: John Wiley & Sons.

Priedhorsky, R., J. Chen, S. K. Lam, K. Panciera, L. Terveen, J. Riedl. 2007. Creating, Destroying, and Restoring Value in Wikipedia. In *Proc. GROUP 2007*.

Randel, A. E. and K. S. Jaussi. 2003. Functional Background Identity, Diversity, and Individual Performance in Cross-Functional Teams. *Academy of Management Journal*, 46(6), 763-774.

Ransbotham, S. and G. C. Kane. 2011. Membership Turnover and Collaboration Success in Online Communities: Explaining Rises and Falls from Grace in Wikipedia. *MIS Quarterly*, 35 (3), 613-627.

Rashtchian, C., P. Young, M. Hodosh, J. Hockenmaier. 2010. Collecting Image Annotations Using Amazon's Mechanical Turk. In *Proc. NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 139-147.

Riding, R. and S. Rayner. 1998. Cognitive Styles and Learning Strategies: Understanding Style Differences in Learning and Behaviour. London: David Fulton Publishers.

Robinson, S. R. and A. M. O'Leary-Kelly. 1998. Monkey See, Monkey Do: The Influence of Work Groups on the Antisocial Behavior of Employees. *Academy of Management Journal*, 41(6), 658-672.

Robles, G., J. M. Gonzalez-Barahona, M. Michlmayr. 2005. Evolution of Volunteer Participation in Libre Software Projects: Evidence from Debian. In *Proc. First International Conference on Open Source Systems*.

Ross, J., L. Irani, M. S. Silberman, A. Zaldivar, B. Tomlinson. 2010. Who Are the Crowdworkers? Shifting Demographics in Amazon Mechanical Turk. In *CHI EA conference*, 2863-2872.

Shachaf, P. and N. Hara. 2010. Beyond Vandalism: Wikipedia Trolls. *Journal of Information Science*, 36(3), 357-370.

Snow, R., B. O'Connor, D. Jurafsky, A. Y. Ng. 2008. Cheap and Fast – But Is It Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In *Proc. EMNLP 2008*.

Spencer, D. G. and R. M. Steers. 1981. Performance as a Moderator of the Job Satisfaction-Turnover Relationship. *Journal of Applied Psychology*, 66, 511-514.

Sturman, M. C. 2003. Searching for the Inverted U-Shaped Relationship between Time and Performance: Meta-Analyses of the Experience/Performance, Tenure/ Performance, and Age/Performance Relationships. *Journal of Management*, 9(5), 609-640.

Su, Q., D. Pavlov, J. Chow, W. C. Baker. 2007. Internet-Scale Collection of Human-Reviewed Data. In *Proc. 16th International Conference on World Wide Web*.

Suh, B., G. Convertino, E. H. Chi., P. Pirolli. 2009. The Singularity is Not Near: Slowing Growth of Wikipedia. In *Proc. WikiSym2009*.

Taylor, L. C. 2009. Thousands of Editors Leaving Wikipedia. Thestar.com. http://www.thestar.com/news/sciencetech/technology/article/729552--thousands-of-editors-leaving-wikipedia. Retrieved 11/18/10.

Thompson, C. 2009. Clive Thompson on the Taming of Comment Trolls. *Wired*, 17(4), March 2009.

Wagner, W. G., J. Pfeffer, C. A. O'Reilly, III. 1984. Organizational Demography and Turnover in Top-Management Group. *Administrative Science Quarterly*, 29(1), 74-92.

Wang, L. S., J. Chen, Y. Ren, J. Riedl. 2012. Searching for the Goldilocks Zone: Trade-Offs in Managing Online Volunteer Groups. In *Proc. CSCW 2012*, 989–998.

Wiersma, U. J. 1992. The Effect of Extrinsic Rewards in Intrinsic Motivation: A Meta-Analysis. *Journal of Occupational and Organizational Psychology*, 65, 101-114.

Wilson, C. 2007. Taming Internet Flamers and Attracting Adults to Boot. *U.S. News & World Report*, 142(22), 28.

Wilson, J. 2000. Volunteering. *Annual Review of Sociology*, 26, 215-240.

Wise, K., B. Hamman, K. Thorson. 2006. Moderation, Response Rate, and Message Interactivity: Features of Online Communities and Their Effects on Intent to Participate. *Journal of Computer-Mediated Communication*, 12(1), 24-41.

APPENDIX A: INTERVIEW QUESTIONS

Background:

How long have you been a member of this site?

Why did you first join this site? What did you like or dislike?

Do you feel like a part of the community? Why/why not?

Have you ever considered leaving this site?

What caused you to consider leaving? Why did you decide to stay?

Online Community Problems:

Have you seen people behave in a way that disrupts others or the community?

Can you give a specific example when you have seen that happen?

Can you describe or define what this behavior is in your own words?

What has been the impact of this behavior on you? On the community as a whole?

What do you think are the main motivations behind this behavior?

What are some methods you would recommend for preventing this behavior?

Personal Experience:

Have you ever behaved in a similar manner on this site? Have you ever behaved in a similar manner on another site?

Can you describe what happened?

What were your reasons for doing so?

How do you think that behavior impacted others or the community?

What could have been done to prevent this type of situation?

Do you have any other comments or concerns about the community?

APPENDIX B: CODING SCHEME

Events / Constructs	Definitions	Examples	Notes
COMMENT LEVEL			
[Method of trolling]	what suspected trolls do	--	Trolls may use a combination of methods in one post.
- **lack of proof or argument for claims**	trolling by making (often) wild and outrageous claims without giving rationale or evidence; no argument for debatable opinions	"I do believe that the USA has the most freedom. I cannot say that there won't be another free country like mine, but we can only hope." "If it were constitutional, I would demand a test in order to get your voter card. Yes too many idiots vote, that's why we have Obama." "Pagans believe in false gods. When you die and see that it is G-d waiting for you and not Hercules or whatever, you'll regret being such a know it all!"	Label troll comments like this as "lack of proof."
- **inflammatory language**	swearing or vitriol	*Any troll post with offensive language usage.*	Label troll comments using curse words or slurs as "inflammatory language."
- **all caps**	an entire comment or large portion in all capital letters	"For some reason' people think that due to my behavior I don't deserve to have my questions answered. But all questions must be answered! ALL QUESTIONS MUST BE ANSWERED! There are only 2 exceptions: when the answer is obvious or unknown! Yet these people don't believe me! WHY WON'T THEY ANSWER MY QEUSTIONS?!" "GEt. OUT. OF. MY.THREATD!"	Does not include capitalizing individual word for emphasis. Label troll comments using a majority of capital letters as "all caps."
- **personal attacks**	attack directed at an individual, using hurtful comments	"Did you fail basic reading comprehension in elementary school or something?" "I suspect that your body will be damaged goods - not usable, go away." "Man, you really suck at arguing valid points […] you fucking moron."	Label troll comments that target an individual as "personal attack."

[Troll target]	who trolling is directed at	--	
- **targeting many**	open ended comments addressing a general audience, or presenting general opinion, rather than individual response	"What is your thoughts on it? Personally, I am okay with non-practicing pedophiles. The moment you touch a child without their consent, I view it as wrong." (opens debate to all) "Atheism is scepticism, agnosticism is muddleheadedness. Does everyone agree?"	Most OP troll comments fall under this category. Label troll comments not directed at specific individuals as "targeting many."
- **targeting individual**	comment obviously directed at an individual	"no your the faggot here u FAGGOT!!!" "[username] sucks [username] sucks [username] sucks" "Did you fail basic reading comprehension in elementary school or something?"	Typically uses the word "you," but not always. Label troll comments targeting identifiable individuals as "targeting individual."
[Identification of trolling]	the act of someone calling another user a troll	--	
- **firm identification of troll**	definite judgment made that someone is trolling	"oh yeah you're a troll, i forgot" "Go troll someone else, you are not every entertaining."	Label identification comments with clear judgment as "firm identification of troll."
- **tentative identification of troll**	suspected trolling but uncertain	"Is this a troll thread?" "Not sure if trolling or just really egotistical. You aren't the greatest artist in the world. Stop acting like it." "I legitimately can't tell if he's being serious or not. Three threads that carried on from his butthurt so far, so I thought he was just trolling. In his first thread, I sincerely believe it started out of pure, concentrated butthurt but then he turned into this Lord of the Lesser Lights troll." "As God is my witness, I can't tell if this is a troll..."	Includes any instance where user suspects trolling but is not firm. Label these identification comments as "tentative identification."
- **firm identification of not troll**	other user defending suspected troll by arguing they are not trolling	"trust me [username], I'm a troll an' this bubba, well he ain't a troll. he's one a them true blue pissant brainers" "This is not trolling. Real trolling is funny, real trolling takes intelligence and wit. Posting a comment that could have easily been done by a 12 year old or an ignorant zealot takes neither of these." "actually, this one is no troll and no bullshit, so I don't see the point..."	Label identification comments clearly judging non-trolling as "firm identification of not troll."

[Evidence to support the identification]	possible cues used by person making an identification to decide if user is or is not a troll	--	Includes evidence for and against trolling.
- **content of posts (or lack thereof)**	whether the suspected troll posts lack valid content and/or if there are outrageous comments	Troll: "This has got to be a troll post... claiming that other religions and belief systems lack proof but theirs does not? Please." ""You've gotta be a troll... There is no way an intelligent person would use that argument."	Label identification comments using content for evidence as "content of posts."
- **cues from community**	citing evidence provided by other users or identification made by others	Troll: "*points to the other replies in the comment thread calling him a troll and the image that brought him to his page* yeah no i'm not the only one who says your a troll" "If you are a troll, as some have suggested, you have done an expert job of being a troll. Bravo." "I realized it was a troll when I looked at other comments, which I should have done in the first place. Thank you though, and I will tell others about it."	Label identification comments using other users for evidence as "cues from community."
- **profile**	judgment based on cues from profile page or other user information (such as tenure of potential troll)	Troll: "Troll. Just look at the profile -_-" "Lol, or just look at the icon and assume that you're trolling." "They are a troll, hell, their username is pretty trollish." "It was hard, but based of the comments on the front of their page, clearly a troll." (refers to profile page)	Label identification comments using troll's user information for evidence as "profile."
- **past behavior**	recognizing user or judging from other posts user has made (not current thread)	Troll: "[username] is a troll. Don't bother with him." "Oh look, [username], is trying to troll again." "I can confirm that he is trolling, and I can back that up with a link to my battle." Non-troll: "trust me [username], I'm a troll an' this bubba, well he ain't a troll. he's one a them true blue pissant brainers"	Label identification comments using troll's past behavior for evidence as "past behavior."
- **personal experience**	judgment based on user's own personal experience, possibly off-site	Non-troll: "I'd say that she probably isn't a troll, [username]. I've been meeting people like her in real life, for my entire life, and they're deadly serious."	Label identification comments referencing user's own experience as "personal experience."

[Recipient of identification]	who the identification statement is addressed to	--	
- **direct response**	identification directed at suspected troll	"You make no sense, you are proven wrong at every turn, and overall, you are just a troll." "Fail troll." "Hey there, troll." "Please, tell me you're trolling and not serious. Please."	Label identification comments that are direct replies to troll as "direct response."
- **response to other users**	identification is part of reply to another user	"[username] is a troll. Don't bother with him." "I think this guy is just a troll that you've been feeding far too much." "She either sincerely believes the dribble she's spewing, or is a troll...perhaps the latter."	Label identification comments that reply to other users as "response to other users."
- **general response**	identification may be a reply to troll but is directed to all users on the thread	"We really have shitty trolls, don't we?" "This person is a troll. I don't know why you all are feeding it XD" "Ignore the Troll. -____-"	This generally happens only when troll is the OP. Label identification comments that are not directed at an individual as "general response."
[Tone of identification comment]	sentiment of user who makes the identification	--	For each identification, judge whether user considers trolling good or bad. If no obvious sentiment, don't code.
- **negative identification**	negative sentiment calling out a troll as something bad	"Being a troll is not an accomplishment, especially when you fail at trolling." "Trolling is a sin." "no one cares about you, troll" "Do not be a Jerk. Do not be a troll."	Label identification comments with a negative tone as "negative identification."
- **positive identification**	comment appears to enjoy or approve of trolling or offers friendly advice	"I think you're the only troll I like." "There's your homework, mate. Read up on the Lord of the Lesser Lights and his trolling and you'll understand why we love him." ""Nice trolling. […] Very well executed in comparison to other attempts. You'll get a rise out of quite a few people, I'm sure." "lol some trolls can be really entertaining after all XD"	Label identification comments with a positive tone as "positive identification."

[Troll's response]	how a suspected troll responds to identification	--	
- **deny**	denies actions are trolling	"I'm not trolling." "Troll? Me? No. You have misunderstood me." ""No I didn't admit to being a troll somebody made a fake account identical to mine and is saying that. Don't believe them!!" "nope, not trolling"	Label troll comments following an identification that denies it as "deny."
- **admit**	claims to be trolling (whether accused or not)	"I'm trolling you. In case you're confused. Now go away" "unless i'm trolling you..." "sshhhhhh! I'm trolling."	Label ANY troll comments admitting to trolling as "admit."
- **ignore**	make no reference to the accusation of trolling	*Any replies, immediately following identification, that does not reference trolling.*	Label any troll reply to an identification that does not deny or admit it as "ignore."
[Leaving]	user mentions leaving or being fed up with one of the following	--	
- **thread withdrawal**	user mentions leaving the particular thread	"Well I'll leave when you give up and I'm going to the kitchen to get a tasty snack." "Imma just hope you're trolling and go now. Don't wanna waste my time. ^_^"	Label any comment about leaving the thread as "thread withdrawal."
- **subforum withdrawal**	user mentions leaving the particular subforum (complaints, politics, or philosophy and religion) where thread takes place		Label any comment about leaving the subforum as "subforum withdrawal."
- **forum withdrawal**	user mentions leaving the entire forum		Label any comment about leaving the forum as "forum withdrawal."
- **community withdrawal**	user mentions leaving the community/site (deviantART or DA)	"I can dump this site at any time and never look back."	Label any comment about leaving the site as "community withdrawal."

[Tone of non-identifying comments]	sentiment of comments responding to troll that do not make an identification	--	Majority of comments will be neutral or not show a clear emotion. Only obvious positive or negative emotions will be coded. Only comments replying to troll or clearly being influenced by troll will be coded.
- **negative tone**	response indicates negative emotions like anger or annoyance	"What you've just said is one of the most insanely idiotic things I have ever heard. At no point in your rambling, incoherent response were you even close to anything that could be considered a rational thought. Everyone in this room is now dumber for having listened to it. I award you no points, and may God have mercy on your soul." "I think you made this post to get attention." "No. Just... no. Fuck you, fuck your prejudices and fuck your closed-mindedness."	Label any comment following a potential troll comment with a negative sentiment as "negative tone."
- **positive tone**	response indicates positive reactions like humor or camaraderie	"This thread is lolzy" (i.e., finds it laughable or entertaining) "For them to tell you your art isn't good enough for their group is low. This site is supposed to accept all forms of art (with a few exceptions, of course) and that goes for the groups too. There are a lot of artists on here at different levels of skill, and you're pretty skilled." (this user provides positive encouragement and suggestions) "Job well done [username]. you managed to make all kids really mad." (possibly encouraging trolling) "Why stop the entertainment?" (in response to another user's comment to stop feeding the troll)	Label any comment following a potential troll comment with a positive sentiment as "positive tone."

THREAD LEVEL			
[Thread-level trolling method]	patterns in a troll's actions at the thread level	--	
- **repetition**	making similar statements repeatedly, or other statements without substantial content	"[username] sucks [username] sucks [username] sucks" "For some reason' people think that due to my behavior I don't deserve to have my questions answered. But all questions must be answered! ALL QUESTIONS MUST BE ANSWERED!" (followed by) "All questions must be answered, yet nobody answers mine!"	Label thread as "repetition" if this tactic is used by a troll.
[Community consensus]	what majority of users decide in the end	--	"End" can be considered end of discussion regarding troll or end of thread, depending on where troll comments appear. Needs subjective judgment.
- **troll consensus**	majority of users suspect user is a troll	*Based on subjective judgment of overall response and consensus.*	Label thread as "troll consensus."
- **not troll consensus**	majority of users do not make a trolling identification or argue against it	*Based on subjective judgment of overall response and consensus.*	Label thread as "not troll consensus."
- **no consensus**	no majority or clear distinction either way	*Based on subjective judgment of overall response and consensus.*	Label thread as "no consensus."
[(Call for) moderator action]	user calls for moderator action or moderator takes an action	--	Actions coded at thread level and possibly also comment level.
- **locked by moderator**	thread is closed ("locked") by an administrator	(moderator response and last comment of thread) "I'll lock it for you. […] At OP's request."	Label thread as "locked by moderator."
- **other moderator action**	any other interference by someone acting in a formal administrative capacity	*Any official action.*	Label thread (and specific comment if applicable) as "other moderator action."
- **call for action**	user or users call for administrative action or discuss need for it	"if there was a way to remove this thread i would." "Why admins read people being rude, inconsiderate and just a foul part of the community and let it be is beyond me." "This thread should be locked."	Label thread (and specific comment if applicable) as "call for action."