# MODELING THE SUCCESSES AND FAILURES OF CONTENT-BASED PLATFORMS

BY

HIMEL DEV

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois Urbana-Champaign, 2021

Urbana, Illinois

Doctoral Committee:

Associate Professor Hari Sundaram, Chair
Professor Karrie Karahalios
Professor Chengxiang Zhai
Dr. Xiaolin Shi, Snap Inc.

## ABSTRACT

Online platforms, such as Quora, Reddit, and Stack Exchange, provide substantial value to society through their original content. Content from these platforms informs many spheres of life—software development, finance, and academic research, among many others. Motivated by their content's powerful applications, we refer to these platforms as content-based platforms and study their successes and failures. The most common avenue of studying online platforms' successes and failures is to examine user growth. However, growth can be misleading. While many platforms initially attract a massive user base, a large fraction later exhibit post-growth failures. For example, despite their enormous growth, content-based platforms like Stack Exchange and Reddit have struggled with retaining users and generating high-quality content. Motivated by these post-growth failures, we ask: when are content-based platforms sustainable? This thesis aims to develop explanatory models that can shed light on the long-term successes and failures of content-based platforms. To this end, we conduct a series of large-scale empirical studies by developing explanatory and causal models. In the first study, we analyze the community question answering websites in Stack Exchange through the economic lens of a "market". We discover a curious phenomenon: in many Stack Exchange sites, platform success measures, such as the percentage of the answered questions, decline with an increase in the number of users. In the second study, we identify the causal factors that contribute to this decline. Specifically, we show that impression signals such as contributing user's reputation, aggregate vote thus far, and position of content significantly affect the votes on content in Stack Exchange sites. These unintended effects are known as voter biases, which in turn affect the future participation of users. In the third study, we develop a methodology for reasoning about alternative voting norms, specifically how they impact user retention. We show that if the Stack Exchange community members had voted based upon content-based criteria, such as length, readability, objectivity, and polarity, the platform would have attained higher user retention. In the fourth study, we examine the effect of user roles on the health of content-based platforms. We reveal that the composition of Stack Exchange communities (based on user roles) varies across topical categories. Further, these communities exhibit statistically significant differences in health metrics. Altogether, this thesis offers some fresh insights into understanding the successes and failures of content-based platforms.

***To my parents, Parikhit Kumar Dev and Nilima Dev;***
***and my wife, Songjukta Datta.***

## ACKNOWLEDGMENTS

This thesis would not have been possible without the generous guidance, outstanding mentorship, and incredible support from my advisor, committee members, collaborators, mentors, family, friends, and many others. I owe an enormous debt of gratitude to these kind people.

I will begin, rightfully, with my advisor, Professor Hari Sundaram. When I first met him, I was a feeble graduate student in search of a supportive advisor. Prof. Sundaram took me under his wings and spent a great deal of time teaching me how to read, write, and think. It would be impossible to list the numerous ways he shaped my research and thinking. Yet, one aspect I can not help but mention is his remarkable ability to simplify complex ideas into intuitive concepts. Thanks to his numerous demonstrations, I am more optimistic about parsing and developing sophisticated research ideas. Beyond the academic support, he was a source of emotional support for me. I would like to extend my deepest gratitude to Prof. Sundaram for his generosity, guidance, and support.

I was very fortunate to have Prof. Karrie Karahalios, Prof. Chengxiang Zhai, and Dr. Xiaolin Shi as my committee members. Prof. Karahalios shepherded the fourth chapter of this thesis, which received a best paper honorable mention at the ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW). Without her invaluable feedback, this accolade would not have been possible. Prof. Zhai asked many thought-provoking questions and gave insightful feedback, which helped me to better position my work. Dr. Shi gave me the fantastic opportunity to intern in the data science research group at Snapchat, where I could work on exciting problems and learn from my esteemed peers. She has been an inspiring mentor who steered me in the right direction regarding data science research—to think about fundamentals. I would like to express my deepest appreciation to my committee members for their outstanding mentorship.

I have been blessed with some wonderful collaborators, Chase Geigle and Ashish Aggarwal, who played a supporting role in developing this thesis. Chase is the primary contributor of the paper that appeared at the AAAI Conference on Web and Social Media (ICWSM) and eventually became the sixth chapter of this thesis. I would like to extend my sincere thanks to Chase and Ashish for their valuable contributions. Thanks also to my mentees, Qingtao Hu and Jiahui Zheng, for their contributions. I have been fortunate to have interned in the research groups at Adobe, Visa, and Snap, where I could grow as a researcher. I am grateful to my internship mentors, Neil Shah, Hossein Hamooni, and Zhicheng Liu, for their

excellent mentorship. I would also express my gratitude to my undergraduate advisors, Prof. Mohammed Eunus Ali, Prof. Tanzima Hashem, Prof. A. B. M. Alim Al Islam, and Prof. Charisma Choudhury, for their valuable contributions to my growth.

My family stood by me, for better and for worse. I owe them everything and perhaps more. To my parents, wife, sister, brother-in-law, uncles, aunts, cousins, and in-laws: thank you for all the sacrifices. I would not have made it through if it weren't for the kind support of my friends. Many thanks to my friends who were away but still close in spirit: Madhusudan Basak, Saikat Chakraborty, Rajshakhar Paul, Tanmoy Sen, Sumit Debnath, Anup Chowdhury, Manoj Reddy. My heartfelt gratitude to my Chambana friends who have been my support system for the better part of my Ph.D. journey: Shegufta Bakht Ahsan, Syeda Persia Aziz, Sajjadur Rahman, Saraf Tarannum, Tanvir Amin, Priyanka Sarker, Anupam Aich. The Bangladeshi community members at the University of Illinois at Urbana-Champaign (UIUC) have also been a source of joy and support. It is an incomplete list, but I very much appreciate the support of Md. Iftekharul Islam Sakib, Afeefa Rahman, Rifat Mahmud, Magfura Khatun, Ahmed Khurshid, Ibtesam Nazim Ahmed, Shakil Bin Kashem, Kazi Nusrat Jahan, Reaz Mohiuddin, Tuba Yasmin, Muntasir Raihan Rahman, Farhana Afzal, Mehedi Bakht, Karishma Muntashir, among many others.

The University of Illinois system provided tremendous support throughout the journey. A big nod to the friends I have made at UIUC: Mangesh Bendre, Dipannita Dey, Amin Javari, Long Pham, Joon Park, Liqi Xu, Doris Lee. I would like to acknowledge the help of crowd dynamics lab members, especially Subham De, Suhansanu Kumar, Adit Krishnan, Aravind Sankar, Kanika Narang, Ziang Xiao, with whom I have had many enlightening conversations. I would like to acknowledge the assistance of the faculty who mentored me during my research and teaching assistantships, especially Prof. Dolores Albarracin, Prof. Aditya Parameswaran, Prof. Jiawei Han, Prof. Saurabh Sinha, Prof. Bertram Ludäscher, Prof. Dan Roth, and Prof. Abdussalam Alawini. Finally, I would like to recognize the help that I received from the officials and staff, including Viveka Perera Kudaligama, Kara Lynn MacGregor, Maggie Metzger Chappell, Holly Ann Bagwell, Kathy Runck, Kimberly Bogle, and Allison Mette.

## TABLE OF CONTENTS

# CHAPTER 1: INTRODUCTION

Since the beginning of Web 2.0—the inception of user-generated content in online platforms—online platforms have ruled over the Internet for more than two decades. A lot has happened in these two decades: the sudden demise of Myspace and Orkut, the meteoric rise of Facebook and Twitter, and the recent struggle of Stack Exchange and Reddit, to name a few. A running theme of Web 2.0 history has been the rise and fall of various online platforms, which ensued from the mass adoption and occasional abandonment of these platforms by Internet users.

While we witnessed many online platforms' rise and fall, their sustainability has largely remained a mystery to us. We do not know why some online platforms succeed in the long run while others fail. The question is increasingly important as online platforms continue to get entangled in our social fabric through a growing amount of content generation and consumption. With the availability of large-scale log data from some platforms, we now have the opportunity to investigate this question in detail.

## 1.1 SUCCESSES AND FAILURES OF CONTENT-BASED PLATFORMS

In the post Web 2.0, content is integral to studying platform sustainability as they keep the platforms alive by serving "nowness" in the user's social web experience. Most of today's online platforms enable users to generate content and share them with the masses. We can readily organize these platforms into verticals: social networks (e.g., Facebook, Twitter), media-sharing platforms (e.g., Flickr, Youtube), blogs (e.g., Blogger, Medium), forums (e.g., Reddit, Stack Exchange), wikis (e.g., Wikipedia, Wiki How), and the list goes on. Amongst these, blogs, forums, and wikis are of growing interest due to their prominence as high-quality information sources [1, 2, 3, 4, 5]. Content from these platforms informs software development [6], financial decisions [7], and academic research [8], among many others. Motivated by the numerous applications of their content, we refer to these platforms as content-based platforms.

Quintessential content-based platforms, such as Quora, Reddit, and Stack Exchange, have revolutionized how we can crowdsource the search for specific information by offering focused content from online users. The users of these platforms generate bona fide content about a wide variety of subjects: science (e.g., `physics.stackexchange`, `r/physics`), technology (e.g., `android.stackexchange`, `r/android`), and recreation (e.g., `movies.stackexchange`, `r/movies`), to name a few. Thanks to their content's ingenuity, these platforms have expe-

rienced enormous user growth in the last decade. For instance, the monthly active users in Quora, Reddit, and Stack Exchange have exceeded 300 million, 400 million, and 100 million, respectively [9, 10, 11].

However, their growth has come at a cost. For example, Stack Exchange has shown clear signs of decline in retaining users and generating content. The percentage of stable users in Stack Overflow—the largest and oldest of the Stack Exchange sites—steadily decreased from 41.05% in 2011 to 34.89% in 2014 [12]. At the same time, the percentage of deleted or unanswered questions increased from 22.45% to 39.43% [12]. Reddit has also struggled with user retention and shown evidence of voting failures. Reddit users upvoted or downvoted 73% of the posts without first viewing the content [13]. Motivated by these post-growth failures, we ask: when are content-based platforms sustainable? Here, by sustainability, we refer to a platform's ability to maintain different success metrics, such as user retention, content generation, and informed voting, at scale.

To date, the sustainability of online platforms has mainly been studied from the perspective of user growth [14, 15, 16, 17, 18]. There have been recent works on understanding content dynamics [19, 20, 21, 22, 23] and designing incentives for users [24, 25, 26, 27, 28, 29, 30]. Prior work, however, concentrates on individual aspects of platform sustainability, paying little attention to the holistic dynamics of users, content, and votes. Consequently, they fail to explain the post-growth retention failures (users leaving the platform), participation failures (decline in content contribution), and gamification failures (an increase of low-quality content) in these platforms.

In this thesis, we attempt to understand how user participation affects the long-term successes and failures of content-based platforms. Through a series of large-scale empirical studies conducted using explanatory models, we provide a holistic understanding of platform sustainability. Our models provide causal explanations to understand why some content-based platforms succeed in the long run while others fail. The models also inform mechanism design to develop sustainable content-based platforms. What follows is an overview of the four empirical studies conducted in this thesis.

- **Why Content-Based Platforms Fail:** In the first study, we investigate the factors that drive content generation in a platform and how their relationship affects platform sustainability.

- **How Biases Affect Votes On Content:** In the second study, we examine the factors that affect votes on content and show how voters may be unconsciously biased against newcomers.

- **How Voting Affects User Participation:** In the third study, we reason about different community voting norms and how these voting norms affect the participation of users, including newcomers.

- **How User Roles Affect Platform Health:** In the fourth study, we discover the behavioral roles users assume during participation and how the mixture of these roles affects platform health.

## 1.2 PRIOR WORK ON PLATFORM SUSTAINABILITY

The initial success of any online platform depends on its ability to attract new users. For this reason, much work has focused on understanding user growth in online platforms [14, 15, 16, 17, 18]. We now discuss several pioneering works on understanding user growth.

Kumar et al. [14] studied network structure evolution in two large online social networks. They classified network members into three groups: users not interacting in the network; small groups interacting with one another but not interacting with the network as a whole; large groups connected through the network paths. They also proposed a simple network growth model, which builds on the concept of these salient structures. Specifically, the model introduces a disparity between the ease of finding potential connections within the large groups and the difficulty of locating potential connections in the small groups.

Backstorm et al. [15] empirically examined the membership, growth, and evolution of large online communities through a series of prediction tasks. They first developed models to predict an individual's probability of joining a community based on its features and the user's ties to the community. They then developed models to identify the communities that will grow significantly over a given period. They also developed a methodology to measure the movement of individuals between communities.

Kairam et al. [16] studied group growth mechanisms to understand the growth and longevity of groups. They identified two different group growth mechanisms: i) diffusion growth wherein groups attract new members through ties to current members, ii) non-diffusion growth wherein individuals with no prior ties to current members become members themselves. Their analysis revealed two main findings. First, while group clustering increases diffusion growth, groups that grow more from diffusion tend to reach smaller eventual sizes. Second, while past growth features predict short-term growth more accurately, network structural features better predict long-term growth for small groups.

Ribeiro et al. [17] studied the growth of the daily number of users (DAU) in membership-based websites. They proposed a DAU prediction model, which utilizes a set of reaction-

diffusion-decay equations that describe the interactions between active members, inactive members, and not-yet-members of the website. The model can accurately classify membership-based websites as sustainable and unsustainable.

While user growth is important for a platform's initial success, we argue that the long-term success of a platform based on content production can perhaps be more meaningful [19, 20, 21, 22, 23]. We now discuss two notable works on understanding content dynamics.

Guo et al. [19] empirically examined how users generate content in three content-based platforms: a blog site, a bookmark sharing platform, and a question-answering network. Their analysis revealed that users' posting behavior in content-based platforms exhibits daily and weekly patterns. Further, the user posting behavior in these platforms follows stretched exponential distributions instead of power-law distributions. They also discovered that the distributions of user contributions on high-quality content in these platforms have smaller stretch factors for the stretched exponential distribution. This finding implies that a small number of core users can not drive content generation in the platforms.

Walk et al. [20] modeled user-level activity dynamics in content-based platforms using two factors: intrinsic activity decay and positive peer influence. Intrinsic activity decay captures the notion that users lose the motivation to contribute without incentives, and thus platforms become inactive. Positive peer influence captures the notion that the action of their peers influences users. Using the model, they discovered that a platforms' activity dynamics has a natural fixed point—the point of complete inactivity—where all users' contributions have seized. However, through external stimuli, it is possible to destabilize the fixed point, resulting in a potential increase in activity.

Another line of related work focuses on designing incentives for users, especially in the form of badges. We now discuss some recent works on steering user behavior through badges [28, 29, 30].

Anderson et al. [29, 30] studied the effects of badges on user behavior. In [30], they proposed a formal model for reasoning about user behavior in the presence of badges. They found that if a badge rewards a certain level of activity of a particular type, users will increase their activity of this type as they approach the level needed for the badge. They also discovered that different badges lead to different amounts of steering, and the degree of steering depends on how close the user is to achieve the badge. In [29], Anderson et al. conducted a large-scale deployment of badges as incentives for learner engagement in a Coursera MOOC. They conducted randomized experiments in which they varied the presentation of badges across different groups. They found that making badges more salient increased engagement in the MOOC.

Immorlica et al. [28] studied badge design mechanisms to maximize users' total contribu-

tion. They characterized badge mechanisms based on coarse partitioning (i.e., awarding the same badge to many users) vs. fine partitioning (i.e., awarding a unique badge to most users). They found that the optimal badge design mechanism utilized both fine partitioning and coarse partitioning. They found that coarse partitioning is necessary for any approximately optimal mechanism when status valuations exhibit a decreasing marginal value property. In contrast, fine partitioning is necessary for approximate optimality when status valuations exhibit an increasing marginal value property.

In the following few sections, we shall examine the research questions addressed in this thesis in more detail.

## 1.3 WHY CONTENT-BASED PLATFORMS FAIL

> Why did Yahoo Answers fail? Why is Stack Overflow declining? Will Quora survive in 2020?

When are content-based platforms like Stack Exchange sustainable? To answer this question, we interpret the community question answering websites on the Stack Exchange platform as "knowledge markets" and analyze why these markets can fail at scale. A knowledge market framing allows site operators to reason about market failures and design policies to prevent them. Our goal is to provide insights on large-scale knowledge market failures through an interpretable model. To this end, we explore a set of interpretable economic production models to analyze the dynamics of content generation in knowledge markets. Among these, the Cobb-Douglas model best explains empirical data and provides an intuitive explanation for content generation through the concepts of elasticity and diminishing returns. Content generation depends on user participation and how specific types of content (e.g., answers) depend on other types (e.g., questions). We show that these factors of content generation have *constant elasticity*—a percentage increase in any of the inputs (e.g., number of users) leads to a constant percentage increase in the output (e.g., number of answers). Furthermore, markets exhibit *diminishing returns*—the marginal output (e.g., marginal answer contribution) decreases as the input (e.g., number of users) is incrementally increased. Knowledge markets also vary on their *returns to scale*—the increase in output (e.g., number of answers) resulting from a proportionate increase in all inputs (e.g., number of questions and number of users). Notably, many knowledge markets exhibit *diseconomies of scale*—measures of market health (e.g., the percentage of questions with an accepted answer) decrease as a function of system size (i.e., the number of participants). The implications of our work are two-fold: i) site operators ought to design incentives as a function

of size; ii) the market lens should shed insight into complex dependencies among different content types and participant actions in content-based platforms.

## 1.4 HOW BIASES AFFECT VOTES ON CONTENT

The premise of this study lies in our previous empirical finding: users who join a Stack Exchange site more recently contribute fewer answers per month compared to those who joined at an early stage. Furthermore, the gap between the participation of old users and newcomers increases significantly as the community grows in terms of the number of users. This empirical phenomenon, along with qualitative evidence on biases in community feedback, raises the following question:

> Could it be that what is causing the newcomers to give up is not the poor quality of their content, but the biased social judgment of their peers?

To answer this question, we examine biases in community feedback, specifically in the form of votes. In content-based platforms like Stack Exchange, the aggregate of votes is commonly used as the "gold standard" for measuring content quality. Use of vote aggregates, however, is at odds with the existing empirical literature, which suggests that voters are susceptible to different *biases*—reputation (e.g., of the poster), social influence (e.g., votes thus far), and position (e.g., answer position). Our goal is to quantify, in an observational setting, the degree of these voter biases in online platforms. Specifically, what are the *causal effects* of different impression signals—such as the reputation of the contributing user, aggregate vote thus far, and position of content—on a participant's vote on content? Estimating causal effects from observational data is challenging: there may be unobserved confounders (e.g., content quality) that explain the association between the impression signals (e.g., the reputation of the contributing user) and the observed votes. To address this issue, we adopt the instrumental variable (IV) method, a causal inference technique that enables a researcher to quantify causal effects from observational data. IV relies on careful reasoning to identify valid instruments that co-vary with the independent variable and cannot influence the dependent variable through the unobserved confounder. To quantify voter biases, we identify a set of candidate instruments, carefully analyze their validity through argumentation, and then use the valid instruments to reveal the effects of the impression signals on observed votes. Our empirical study using log data from Stack Exchange websites shows that the bias estimates from our IV approach differ from the bias estimates from the ordinary least squares (OLS) method. In particular, OLS underestimates reputation bias (1.6–2.2x for gold

badges) and position bias (up to 1.9x for the initial position) and overestimates social influence bias (1.8–2.3x for initial votes). The implications of our work include redesigning the user interface to avoid voter biases, making changes to platforms' policy to mitigate voter biases, detecting other forms of biases in online platforms.

## 1.5 HOW VOTING AFFECTS USER PARTICIPATION

In the first study, we show that platforms fail because newcomers ease off. In the second study, we show that community voting is biased against them. Now we ask:

> Would the newcomers have stayed longer or contributed more if the community had voted objectively?

In content-based platforms like Stack Exchange, votes are the "social currency" that persuades users to create content. The deficiency of votes, therefore, is likely to impact the future participation of users. From platform operators' perspective, the interplay between votes and retention is crucial as it allows them to make strategic decisions. For instance, if platform operators could identify voting outcomes that would improve user retention, they could purposefully engineer the voting conditions to steer votes towards those outcomes. Understanding the relationship between different community voting norms and user retention is an essential first step towards achieving such altercation capabilities. In this study, we develop a methodology to reason about alternative community voting norms. We refer to the existing community voting norm as the control norm, whose voting outcomes can be observed from log data. Our goal is to reason about alternative community voting norms whose voting outcomes can not be observed in the data. For instance, how user retention in Stack Exchange would be different if the community issues vote based upon the length of the content. To enable such reasoning about alternative norms, we must perform a counterfactual analysis. To this end, we first develop a propensity model for quantifying the probabilities of different voting outcomes under each voting norm. We then define a utility model for quantifying the derived utility of users from the votes they acquire. Finally, we develop a counterfactual model for reasoning about user retention under different voting norms. We adopt an inverse propensity sampling (IPS) estimator to perform our counterfactual analysis. We conduct extensive experiments on Stack Exchange websites comparing the default voting norm in these sites with six alternative norms: random (i.e., content receive an arbitrary number of votes), uniform (i.e., all content receive the same number of votes), length (i.e., content that contain more words receive more votes), readability (i.e., content that have

higher readability receive more votes), objectivity (i.e., content that express facts rather than opinions receive more votes), and polarity (i.e., content that express positive emotion receive more votes). Our main finding is that if the community members had voted based upon the length, readability, objectivity, or polarity of the content, the platform would have observed *higher* retention. The main design implication of this study is that site operators need to promote content based on factors that are intrinsic to the content.

## 1.6 HOW USER ROLES AFFECT PLATFORM HEALTH

In a content-based platform, users assume various action-based roles, e.g., asker, answerer, moderator, etc. Prior work suggests that different user roles are crucial for sustainable content production in these communities [31]. Motivated by the importance of user roles, we ask the following research questions.

> Does a community's composition over user roles vary as a function of the topic? How does it relate to the health of the underlying community?

In this study, we propose a generative model for discovering user roles and community role compositions in Community Question Answering (CQA) platforms. While past research shows that participants play different roles in online communities, automatically discovering these roles and providing a summary of user behavior that is readily interpretable remains an important challenge. Furthermore, there has been relatively little insight into the distribution of these roles between communities. The generative model proposed in this work, the mixture of Dirichlet-multinomial mixtures (MDMM) behavior model can (1) automatically discover interpetable user roles (as probability distributions over atomic actions) directly from log data, and (2) uncover community-level role compositions to facilitate such cross–community studies. A comprehensive experiment on all 161 non-meta communities on the Stack Exchange platform demonstrates that our model can be useful for a wide variety of behavioral studies, and we highlight three empirical insights. First, we show interesting distinctions in question-asking behavior on Stack Exchange (where two distinct types of askers can be identified) and answering behavior (where two distinct roles surrounding answers emerge). Second, we find statistically significant differences in behavior compositions across topical groups of communities on Stack Exchange, and that those groups that have statistically significant differences in health metrics also have statistically significant differences in behavior compositions, suggesting a relationship between behavior composition and health. Finally, we show that the MDMM behavior model can be used to demonstrate similar but distinct evolutionary patterns between topical groups.

## 1.7 SUMMARY OF CONTRIBUTIONS

Our first study aimed to investigate the factors that drive content generation in platforms. We now briefly discuss the contributions of this work.

1. **Content Generation Model.** We adopt macroeconomic production functions to describe content generation in platforms, a novel application of economic production functions. Our best-fit model, the Cobb-Douglas function, provides an intuitive explanation for content generation in platforms.

2. **Insights into Platform Sustainability.** The Cobb-Douglas model provides three critical insights: stable core, size-dependent distribution, and diseconomies of scale. Briefly, there is a stable core of users who substantially contribute to platforms for a long time. In many platforms, the size of this stable core does not increase with the number of users. This discrepancy results in a size-dependent activity distribution, i.e., the expected user behavior changes with community size. As a result, these platforms exhibit diseconomies of scale—platform health declines with the increase in the number of users.

Our second study aimed to examine the factors that affect votes on content. We now briefly discuss the contributions of this work.

1. **Voter Bias Quantification.** We quantify the degree of voter biases—namely, reputation bias, social influence bias, and position bias—in content-based platforms. To derive these bias estimates, we measure different impression signals' effects on observed votes through a novel application of instrumental variables.

2. **Stronger Evidence of Biases.** Our analysis reveals that prior work has significantly underestimated the impact of voter biases. In particular, our empirical study using log data from Stack Exchange sites shows that the bias estimates from our IV approach differ from the bias estimates from the ordinary least squares (OLS) method. OLS underestimates reputation bias (by a factor of 1.6–2.2 for gold badges) and position bias (by a factor of 1.3-2.0 for the initial position).

Our third study aimed to develop a methodology to reason about alternative voting norms. We now briefly discuss the contributions of this work.

1. **Propensity Model for Voting Outcomes.** We adopt a Dirichlet-multinomial model for quantifying the probabilities of different voting outcomes under a community voting

norm. Our propensity model explains the community voting behavior wherein voters examine the current context to evaluate and vote on the answers.

2. **Counterfactual Model for Community Voting.** We develop an inverse propensity score (IPS) model for reasoning about user retention under different community voting norms. To the best of our knowledge, this is the first attempt to use IPS to reason about alternative community voting norms.

3. **Causal Insights into Voting Outcomes.** Our analysis provides some of the first causal insights into voting outcomes. For example, content-based voting norms can improve user retention significantly compared to the current voting norm. Even partial (20%) content-based norms significantly outperform the current norm for various retention metrics.

Our fourth study aimed to develop a methodology to discover the behavioral roles users assume during participation. We now briefly discuss the contributions of this work.

1. **Role Discovery Model.** We propose a generative model for discovering action-based user roles along with community role compositions. The proposed model, mixture of Dirichlet-multinomial mixtures (MDMM) behavior model, can automatically discover interpretable user roles directly from log data and uncover community-level role compositions to facilitate cross-community studies.

2. **Insights into Platform Health.** Our model showed statistically significant differences in behavior compositions across topical groups of communities on Stack Exchange. Further, the model also revealed that the groups with statistically significant differences in health metrics also have statistically significant differences in behavior compositions, suggesting a relationship between behavior composition and platform health.

## 1.8 ORGANIZATION OF THE THESIS

The rest of the thesis is organized as follows. Chapter 2 provides an overview of the cross-disciplinary research efforts on understanding platform sustainability. Chapter 3 describes our first empirical study: what factors drive content generation and how their relationship affects platform sustainability. Chapter 4 describes our second empirical study: what factors affect votes on content and show how some of these factors lead to biases. Chapter 5 describes our third empirical study: how different community voting norms affect the participation

of users. Chapter 6 describes our fourth empirical study: what roles users assume during participation and how the mixture of these roles affects platform health.

# CHAPTER 2: BACKGROUND

In this thesis, we adopt an economic lens to study the successes and failures of content-based platforms. Through a series of large-scale empirical studies conducted using economic and causal models, we provide insights into the basis of platform success. Perhaps the most relevant research to ours is prior work that studied content-based platforms from an economic perspective, specifically from a "market" viewpoint. This chapter provides an overview of the cross-disciplinary research landscape on understanding content-based platforms as knowledge markets, reasoning about knowledge market failures, and learning how to build successful knowledge markets. In the rest of the chapter, we will often refer to content-based platforms as knowledge markets and users as participants. Plese note that a more in-depth analysis of the prior work related to the problems studied in this thesis can be found in their associated chapters.

## 2.1 CONTENT-BASED PLATFORMS AS KNOWLEDGE MARKETS

Once flourishing content-based platforms, such as Google Answers[1] and Yahoo Answers[2], have drawn much attention from the research community. A growing body of Economics and Computer Science literature has studied these platforms as "markets" [32, 33, 34].

Notably, Chen et al. [32] conducted a field experiment at Google Answers to investigate the effects of various design features on knowledge markets' performance. They particularly examined the effects of price, tips, and reputation systems on the answerer's effort and answer quality under different pricing schemes. Their analysis revealed several interesting findings. First, a higher price leads to a longer but not better answer. Second, the level and type of tip do not affect answer length and quality. Third, an answerer with a high reputation typically provides an answer with higher quality. The implication of these findings is that reputation systems are critical for maintaining content quality in knowledge markets.

Daphne Ruth Raban [33] studied Google Answers to understand the incentive structure in information markets when both economic and social incentives are present. Her analysis confirmed the preeminence of economic incentives: the most important predictor for answerer participation in Google Answers was the anticipated tip. However, further analysis of two answerer subgroups (frequent answerers and occasional answerers) revealed the importance of social incentives: i) for frequent answerers, the crucial predictors include comments and

[1]Google Answers was a price-based Q&A platform offered by Google.
[2]Yahoo Answers is a free Q&A platform offered by Yahoo.

ratings; ii) for occasional answerers, the crucial predictors include comments. These findings imply that social incentives play a much more critical role in answerer engagement in information markets than previously understood.

Benjamin Edelman [34] analyzed the questions and answers from Google Answers to examine earnings and ratings. He found several interesting trends in answerer behavior. First, experienced answerers provide answers that askers tend to value more, thus acquiring higher ratings. Second, specialists, i.e., answeres who focus on specific question categories, provide higher quality answers than generalists. Third, on the whole, more experienced answerers tend to be more specialized. These findings highlight the importance of experienced answerers in operating knowledge markets.

Gary Hsieh conducted a series of studies highlighting the importance of applying market mechanisms to support information exchange [35, 36, 37]. These studies show how markets can improve welfare for the participants in information exchange, the operationalization of markets for Q&A platforms (knowledge markets), and how they affect interpersonal relationships.

In [35], Hsieh et al. explored market mechanisms in communication systems. They compared three communication systems modeled on questioning and answering: i) baseline system, ii) variable-price market system, iii) fixed-price market system. In the baseline system, the sender (asker) sends a request without any financial incentives, and the receiver (answerer) has to decide whether to respond. In the variable-price market system, the askers offer to pay an individually set price for an answer, and receivers accept communication if this price exceeds their individually set reservation price. In the fixed-price market system, askers pay answerers a fixed price if communication occurs. The study revealed that market systems (both fixed and variable-price markets) lead to improved communication compared to the non-market system (baseline system). This finding establishes the importance of market mechanisms in information exchange.

In [36], Hsieh et al. designed two variants of a real-time Q&A system. The first variant adopted a simple reputation system, whereas the second variant adopted an explicit market-based system using a currency. The authors compared how the two systems were used and then conducted a controlled study on question answering. Their analysis revealed that in the explicit market-based system, askers and answerers were more selective in what they ask and answer. This finding implies that the market-based system reduces low-quality questions and answers at the cost of a reduction in overall content generation and community engagement.

In [37], Hsieh et al. analyzed randomly selected questions from Mahalo Answers—a price-based Q&A service that allows its users to ask both free and for-pay questions. The authors examined the factors that impact the decision to pay for answers and how financial rewards

affect answers. They found that askers are more likely to pay for factual answers and when questions are difficult to answer. The results also confirm the prior finding that paying more leads to longer but not necessarily higher quality answers. These findings highlight the factors that affect pricing in the knowledge market.

## 2.2 KNOWLEDGE MARKET FAILURES

The recent failures of content-based platforms, such as Google Answers[3] and Yahoo Answers[4], have made researchers curious about knowledge market failures. There is a recent body of work on understanding knowledge market failures in terms of user retention, content generation, and content quality.

Dror et al. [38] studied the problem of churn prediction on Yahoo Answers, specifically for new users. They compiled a wide variety of demographic (e.g., age, gender), behavioral (e.g., answering time, answering rate), and social (e.g., interaction with other users) features to predict whether a new user is likely to churn out. Their feature analysis revealed that the strongest indicators of user churn are the number of answers created by the user and the user's degree of recognition in the form of best answers, thumbs up, and positive responses. These results indicate that social incentives are crucial for retaining participants in knowledge markets.

Harper et al. [39] conducted a comparative field study of five Q&A sites—Library Reference Services, Google Answers, AllExperts, Yahoo Answers, and Live QnA—to analyze answer quality. They focused on answering two research questions: i) How do Q&A sites differ in the quality and characteristics of answers? ii) What can askers do to receive better answers on a Q&A site? They found that price-based Q&A sites such as Google Answers provide better answers with more diverse opinions. Among the free Q&A sites, Yahoo Answers provided higher-quality answers, enabled by its large yet dedicated user community. Overall, participants get what they pay for in knowledge markets, and a knowledge market's success relies on the market size.

Srba and Bielikova [12] conducted a case study on why Stack Overflow, the largest Stack Exchange site, is failing. They found that the churn rate of Stack Overflow is increasing, especially for newcomers. Empirically, the proportion of one-time contributors, users who leave after their first contribution, is steadily increasing (from 30.8% in 2011 to 33.1% in 2014). In contrast, the proportion of stable users, users who contribute for a prolonged period, is rapidly decreasing (from 41.05% in 2011 to 34.89% in 2014).

[3]Google shut down Google Answers in 2006.

[4]Yahoo Answers has experienced a massive drop in traffic since 2011.

Dearman et al. [40] conducted a survey to understand why users do not answer questions in Yahoo Answers. They surveyed active members of Yahoo Answers, who revealed their lack of motivation for answering questions. They found that top and regular contributors do not answer questions for the same reasons. Also, questions that already received several responses are less likely to be answered; this happens because the answerers fear that the new response will not be noticed.

## 2.3 BUILDING SUCCESSFUL KNOWLEDGE MARKETS

While the failures of Google Answers and Yahoo Answers are concerning, the success of Wikipedia gives researchers design ideas to develop successful knowledge markets. Notably, Karut et al. have drawn on the literature in psychology, economics, and other social sciences, as well as their own research, to improve the design of knowledge markets [41, 42, 43, 44].

Ren et al. [41] took one of the earliest steps towards mining social science theories by arguing that online community design influences how people become attached to the community. They explored two group attachment theories with design implications for online communities, namely common identity theory and common bond theory. The former asserts the causes and consequences of people's attachment to the group (macro). In contrast, the latter asserts the causes and consequences of people's attachment to individuals in the group (micro). They explained how design decisions, notably recruiting newcomers vs. retaining existing members and limiting group size vs. allowing uncontrolled growth, can lead to different degrees and forms of community participation by those so motivated.

Ren et al. [42] also explored social psychology theory to understand how online communities develop member attachment. They implemented two sets of community features for building member attachment by strengthening either group identity or interpersonal bonds. To support identity-based attachment, they gave members information about group activities and intergroup competition. They also provided members with group-level communication tools. To support bond-based attachment, they gave members information about the activities of individual members and interpersonal similarity. They also provided members with interpersonal communication tools. They found that, under both conditions, members' attachment to groups increased. Further, community features supporting identity-based attachment had more substantial effects than features supporting bond-based attachment. The new features also had more substantial effects on newcomers than on old users.

Zhu et al. [43] explored the role of shared leadership in building successful online communities in the context of Wikipedia. They proposed a shared leadership model, which asserts that leadership behaviors may come from members at all levels. Using propensity

score matching, they investigated how different leadership behaviors and the position of the people who deliver them (say formal leadership positions or not) influence the contributions that other participants make. Their main finding is that leadership behavior performed by members at all levels significantly influenced other members' motivation. They also found that transactional and person-focused leadership effectively motivated others to contribute, whereas aversive leadership decreased other contributors' motivations.

Zhu et al. [44] also investigated how combining group identification, and direction setting can motivate volunteers in online communities. They argued that group identity triggers in-group favoritism, while direction setting steers people's group-oriented motivation towards the group's essential tasks. They tested their hypotheses in the context of Wikipedia's Collaborations of the Week (COTW), a social event within Wikiprojects. They found that publicizing important group goals via COTW can substantially motivate editors who have voluntarily identified themselves as group members. Further, the positive effects of goals spill over to non-goal-related tasks. Finally, editors exposed to group role models are more likely to perform similarly to the models on group-relevant citizenship behaviors.

## 2.4 SUMMARY

The diversity of ideas that prior literature has collectively considered motivates the present work on platform sustainability. In Section 2.1, we surveyed prior work on understanding content-based platforms as knowledge markets. The two main takeaways of the prior work on knowledge markets are: i) reputation system plays a pivotal role in maintaining quality and engagement in knowledge markets, ii) market mechanisms can improve welfare for the participants in information exchange. In Chapter 4, we show the presence of biases in the reputation systems that utilize votes. These biases affect the retention of participants in knowledge markets. In Section 2.2, we surveyed prior work on knowledge market failures. The two main takeaways of the prior work on knowledge market failures are: i) a knowledge market's success relies on its size, ii) social incentives are crucial for retaining participants in knowledge markets. In Chapter 3, we show how the increase in the size of a knowledge market can adversely affect its health. In Chapter 5, we examine the impact of social incentives by studying the impact of voting norms on the retention of participants. In Section 2.3, we surveyed prior work on building successful knowledge markets. The two main takeaways of the prior work on successful knowledge markets are: i) community design diversely affects the membership attachments of old users and newcomers, ii) leadership behaviors performed by members at all levels significantly influence other members' motivation. While these studies illuminate design decisions for designing successful knowledge markets, they primarily focus

on the role of membership attachment and shared leadership. In contrast, this thesis focuses on the impact of voting norms and user roles.

# CHAPTER 3: MODELING CONTENT GENERATION

In content-based platforms, generating content is integral to the platforms' sustainability. For instance, Q&A platforms like Stack Exchange often use content generation metrics, such as the percentage of answered questions and the percentage of questions with an accepted answer, to monitor platform health. This chapter discusses our first study on modeling content generation and exploring the relationship between content generation and platform sustainability.

## 3.1 INTRODUCTION

In this study, we analyze a large group of community question answering (CQA) websites on Stack Exchange network through the Economic lens of a market. Framing Stack Exchange sites as knowledge markets has intuitive appeal: in a hypothetical knowledge market, if no one wants to answer questions, but only ask, or conversely, there are individuals who want to only answer but not ask questions, the "market" will collapse. What, then, is the required relationship among actions (say between questions and answers) in such a knowledge market for us to deem it healthy? Are larger markets with more participants healthier since there will be more people to ask and answer questions?

Studying CQA websites through an economic lens allows site operators to reason about whether they should grow the user base. Since most of the popular CQA websites (e.g., Quora, Stack Exchange) do not charge participants, but instead depend on site advertisements for revenue, there is a natural temptation for operators of these sites to grow the user base so that there is increase in revenue. As we show in this study, for most Stack Exchange sites, growth in the user base is counter-productive in the sense that they turn unhealthy—specifically, more questions remain unanswered.

Explaining the macroscopic behavior of knowledge markets is important, yet challenging. One can regress some variable of interest (say number of questions) on variables including number of users, time spent in the website among others. However, explaining why the regression curve looks the way it does is hard. As we show in this work, using an economic lens of a market allows us to model dependencies between number of participants and the amount of content, and to predict the production of content.

Our main contribution is to model CQA websites as knowledge markets, and to provide insight on the relationship between size and health of these markets. To this end, we develop models to capture content generation dynamics in knowledge markets. We analyze a set of

basis functions (the functional form of how an input contributes to output) and interaction mechanisms (how the inputs interact with each other), and identify the optimal *power basis* function and the *interactive essential* interaction form using a prediction task on the outputs (questions, answers, and comments). This form is the well-known Cobb-Douglas form that connects production inputs with output. Using the best model fits for each Stack Exchange site, we show that the Cobb-Douglas model predicts the production of content with high accuracy.

The Cobb-Douglas function provides an intuitive explanation for content generation in Stack Exchange markets. It demonstrates that, in Stack Exchange markets, 1. factors such as user participation and content dependency have *constant elasticity*—percentage increase in any of these inputs will have constant percentage increase in output; 2. in many markets, factors exhibit *diminishing returns*—decrease in the marginal (incremental) output (e.g., answer production) as an input (e.g. number of people who answer) is incrementally increased, keeping the other inputs constant; 3. markets vary according to their *returns to scale*—the increase in output resulting from a proportionate increase in all inputs; and 4. many markets exhibit *diseconomies of scale*—measures of health decrease as a function of overall system size (number of participants).

There are two reasons why we see diminishing returns in the Stack Exchange markets. First, the total activity of participants for any Stack Exchange market unsurprisingly follows a power-law pattern. What is interesting is that the power-law exponent falls with increase in size for most markets, implying that new users do not participate in the same manner as earlier users. Second, we can identify a stable core of users who actively participate for long periods of time, contributing to the market health.

Finally, we show diseconomies of scale through experiments on system size, analysis of health metrics, and user exchangeability. For most Stack Exchange markets, we see that as system size grows, the ratio of answers to questions falls below a critical point, when some questions go unanswered. Furthermore, using health metrics of the number of questions with an accepted answer, and the number of questions with at least one answer, we observe that most Stack Exchange markets decline in health with increase in size. Finally, we compare the top contributors with the bottom contributors to see if they are "exchangeable." Most Stack Exchange markets are not exchangeable in the sense the contributions of the top and the bottom contributors are qualitatively different and differ in absolute terms. These experiments on diseconomies of scale are consistent with the insight from Cobb-Douglas model of production that predicts diminishing returns.

## 3.2 RELATED WORK

Our work draws from, and improves upon, several research threads.

**Sustainability.** Srba et al. [12] conducted a case study on why Stack Overflow, the largest and oldest of the sites in Stack Exchange network, is failing. They shed some insights into knowledge market failure such as novice and negligent users generating low quality content perpetuating the decline of the market. However, they do not provide a systematic way to understand and prevent failures in these markets. Wu et al. [45] introduced a framework for understanding the user strategies in a knowledge market—revealing the importance of diverse user strategies for sustainable markets. In this study, we present an alternative model that provides many interesting insights including knowledge market sustainability.

**Activity Dynamics.** Walk et al. [20] modeled user-level activity dynamics in Stack Exchange using two factors: intrinsic activity decay, and positive peer influence. However, the model proposed there does not reveal the collective platform dynamics, and the eventual success or failure of a platform. Abufouda et al. [46] developed two models for predicting the interaction decay of community members in online social communities. Similar to Wu et al. [20], these models accommodate user-level dynamics, whereas we concentrate on the collective platform dynamics. Wu et al. [47] proposed a discrete generalized beta distribution (DGBD) model that reveals several insights into the collective platform dynamics, notably the concept of a size-dependent distribution. In this study, we improve upon the concept of a size-dependent distribution.

**Economic Perspective.** Kumar et al. [48] proposed an economic view of CQA platforms, where they concentrated on the growth of two types of users in a market setting: users who provide questions, and users who provide answers. In this study, we concentrate on a subsequent problem—the "relation" between user growth and content generation in a knowledge market. Butler et al. [49] proposed a resource-based theory of sustainable social structures. While they treat members as resources, like we do, our model differs in that it concentrates on a market setting, instead of a network setting, and takes the complex content dependency of the platform into consideration. Furthermore, our model provides a systematic way to understand successes and failures of knowledge markets, which none of these models provide.

**Scale Study.** Lin et al. [50] examined Reddit communities to characterize the effect of user growth in voting patterns, linguistic patterns, and community network patterns. Their study reveals that these patterns do not change much after a massive growth in the size of the user community. Tausczik et al. [51] investigated the effects of crowd size on solution quality in Stack Exchange communities. Their study uncovers three distinct levels of group

size in the crowd that affect solution quality: topic audience size, question audience size, and number of contributors. In this study, we examine the consequence of scale on knowledge markets from a different perspective by using a set of health metrics.

**Stability.** Successes and failures of platforms have been studied from the perspective of user retention and stability [52, 53, 54, 55]. Notably, Patil et al. [52] studied the dynamics of group stability based on the average increase or decrease in member growth. Our study examines stability in a different manner—namely, by considering the relative exchangeability of users as a function of scale.

**User Growth.** Successes and failures of user communities have also been widely studied from the perspective of user growth [14, 15, 16, 17, 18]. Kairam et al. [16] examined diffusion and non-diffusion growth to design models that predict the longevity of social groups. Ribeiro et al. [17] proposed a daily active user prediction model which classifies membership based websites as sustainable and unsustainable. While this perspective is important, we argue that studying the successes and failures of communities based on content production can perhaps be more meaningful [21, 22, 23].

**Modeling CQA Websites.** There is a rich body of work that extensively analyzed CQA websites [1, 2, 3, 56, 57], along with user behavior [58, 59, 60, 61, 62], roles [63, 64], and content generation [65, 66, 67]. Notably, Yang et al. [66] noted the *scalability problem* of CQA—namely, the volume of questions eventually subsumes the capacity of the answerers within the community. Understanding and modeling this phenomenon is one of the goals of this study.

## 3.3 PROBLEM FORMULATION

The goal of this study is to develop a model for content generation in knowledge markets. Content is integral to the success and failure of a knowledge market. Therefore, we aim to better understand the content generation dynamics.

A model for content dynamics should have the following properties: macro-scale, explanatory, predictive, minimalistic, comprehensive.

*Macro-scale:* The model should capture content generation dynamics via aggregate measures. Aggregate measures help us understand the collective market by summarizing a complex array of information about individuals, which is especially important for policy-making.

*Explanatory:* The model should be insightful about the behavior of a knowledge market. Understanding market behavior is a crucial first step in designing policies to maintain a resilient, sustainable market.

*Predictive:* The model should allow us to make predictions about future content generation and resultant success or failure. These market predictions are integral to the prevention and mitigation of market failures.

*Minimalistic:* The model should have as few parameters as necessary, and still closely reflect the observed reality.

*Comprehensive:* The model should encompass content generation dynamics for different content types (e.g., question, answer, comment) in varieties of knowledge markets. This is important for developing a systematic way to understand the successes and failures of knowledge markets.

In remaining sections we propose models that meet the aforementioned requirements, and show that our best-fit model accurately reflects the content generation dynamics and resultant successes and failures of real-world knowledge markets.

## 3.4 MODELING KNOWLEDGE MARKETS

In this section, we introduce economic production models to capture content generation dynamics in real-world knowledge markets. We first draw an analogy between economic production and content generation, and report the content generation factors in knowledge markets (Section 3.4.1). Then, we concentrate on the knowledge markets in Stack Exchange—presenting production models for different content types (Section 3.4.2).

### 3.4.1 Preliminaries

Economic production mechanisms well describe content generation in knowledge markets. In economics, *production* is defined as the process by which human labor is applied, usually with the help of tools and other forms of capital, to produce useful goods or services—the *output* [68]. We assert that participants of a knowledge market function as labor to generate content such as questions and answers. Analogous to economic output, content contributes to participant utility.

Motivated by the production analogy, we design macroeconomic production models to capture content generation dynamics in knowledge markets. In these models, instead of directly modeling content generation as a dynamic process (function of time), we model it in terms of associated factors which are dynamic.

There are two key factors that affect content generation in knowledge markets, namely user participation and content dependency. User participation is the most important factor in deciding the quantity of generated content. The participation of more users induce more

questions, answers, and other contents in a knowledge market. Content dependency also affects the quantity of generated content for different types. Content dependency refers to the dependency of one type of content (e.g., answers) on other type of content (e.g., questions). In absence of questions, there will be no answers in a knowledge market, even in the presence of many potential participants who are willing to answer.

### 3.4.2 Modeling Stack Exchange

Stack Exchange is a network of community question answering websites where each site is based on a focused topic. Each user of the Stack Exchange network participates in one or more of these sites based on their interests. Stack Exchange sites are free knowledge markets where participants generate content for non-monetary reputation-based incentives. These markets are diverse, varying in theme (subject matter), size (number of users and amount of activity), and age (number of days in existence).

We design production models for three primary content types in Stack Exchange: questions (the root content), answers (which nest below questions), and comments (which can nest either beneath questions or answers). Based on the content dependency and user roles in content generation, we propose the following relationships for question, answer and comment generation in Stack Exchange (see Table 3.1 for notation).

Table 3.1: Notation used in the model

| Symbol | Definition |
|---|---|
| $U_q(t)$ | The number of users who asked questions at time $t$ |
| $U_a(t)$ | The number of users who answered questions at time $t$ |
| $U_c(t)$ | The number of users who made comments at time $t$ |
| $N_q(t)$ | The number of active questions at time $t$ |
| $N_a(t)$ | The number of answers to active questions at time $t$ |
| $N_c^q(t)$ | The number of comments to active questions at time $t$ |
| $N_c^a(t)$ | The number of comments to active answers at time $t$ |
| $N_c(t)$ | The number of comments to active questions/answers at time $t$ |
| $f_x$ | The functional relationship for content type $x$ |

There is a single factor in generating $N_q$ questions: the number of users $U_q$ who ask questions (askers).

$$N_q = f_q(U_q) \tag{3.1}$$

There are two key factors in generating $N_a$ answers: the number of questions $N_q$, and the number of users $U_a$ who answer questions (answerers).

$$N_a = f_a(N_q, U_a) \tag{3.2}$$

There are two types of comments: comments on questions, and comments on answers. Accordingly, there are three key factors in generating $N_c$ comments: the number of questions $N_q$, the number of answers $N_a$, and the number of users $U_c$ who make comments (commenters).

$$N_c^q = f_{c^q}(N_q, U_c) \tag{3.3}$$

$$N_c^a = f_{c^a}(N_a, U_c) \tag{3.4}$$

$$N_c = N_c^q + N_c^a \tag{3.5}$$

The aforementioned relationships imply that the amount of generated content of each type depends on the function describing its factor dependent growth, and the availability of factor(s). These relationships make three assumptions. First, different content types interact only through their use of factors. Second, the functional relationships depend on the consumption or usage of each factor. Third, the functional relationships depend on the interaction among the factors—how the factors of a particular content type interact.

Now, we transform the functional relationships into production models by first choosing a basis function to capture how a content type consumes its factor(s), and then choosing an interaction type to capture the interaction among factors.

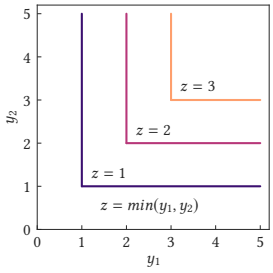
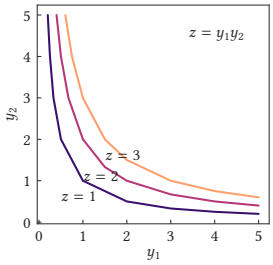
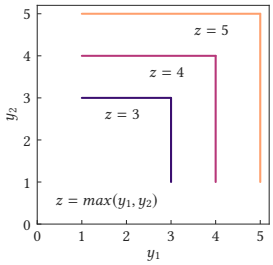
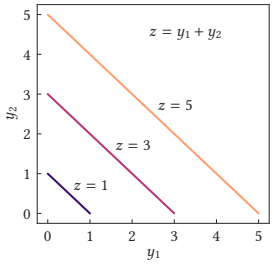
**Basis Function.** We use a basis function to capture the effect of a given factor on a particular content type. While there is a variety of basis functions available for regression, we consider three basis functions widely used in economics and growth modeling [69]: power— $g(x) = ax^{\lambda}$; exponential— $g(x) = ab^{x}$; and sigmoid— $g(x) = \frac{L}{1+e^{k(x-x_0)}}$.

**Interaction among the Factors.** We use an aggregate function to capture the interaction among multiple factors of a given content type. Specifically, we consider the pairwise interaction functions listed in Table 3.2.

We combine a basis function and an interaction type to design production models for different content types. For example, answer generation can be modeled using power basis and essential interaction as $N_a = min(a_1 N_q^{\lambda_1}, a_2 U_a^{\lambda_2})$. We consider twelve such possible models (combination of three basis and four interaction type) for answer and comment generation in Stack Exchange.

**User Role Distribution.** A fundamental assumption of our model is the awareness of user roles (e.g., asker, answerer, and commenter) and their distribution (e.g., how many

Table 3.2: Pairwise interaction between factors with contour

| | |
|---|---|
| Essential: Essential factors are both required for content generation, with zero marginal return for a single factor. For a pair of essential factors, content generation is determined by the more limiting factor: $z = min(y_1, y_2)$ [70]. This is known as Liebig's law of the minimum. |   |
| Interactive Essential: In interactive essential interaction, we get diminishing return (instead of zero return) for a single factor: $z = y_1 y_2$ [70]. If factors are consumed using power basis function, i.e., $y_i = ax_i^{\lambda}$, it captures Cobb-Douglas production function. |   |
| Antagonistic: For antagonistic factors, content generation is determined solely by the availability of the factor which yields the largest return: $z = max(y_1, y_2)$ [70]. This interaction implies that the production process has maximum possible efficiency. |   |
| Substitutable: Factors that can each support production on their own are substitutable relative to each other: $z = w_1 y_1 + w_2 y_2$ [70]. This implies that there exists some equivalence between the two factors. This is analogous to the general additive models. |   |

users are askers?). We empirically observe that all Stack Exchange markets have a stable distribution of user roles. In fact, given the number of users, we can accurately predict the number of participants for each role.

We apply linear regression to determine the number of participants $U_x$ of a particular role $x \in \{q, a, c\}$ from the number of users $U$ in a Stack Exchange market. For each market, we compute three distinct coefficients of determination, $R^2$, for predicting three roles (asker, answerer, and commenter) using linear regression. In Figure 6.2 we show the distribution of $R^2$ for regressing user roles across 156 Stack Exchange markets. We use letter value plots[1] to present these distributions—showing precise estimates of their tail behavior. We observe

[1]The letter-value plot display information about the distribution of a variable [71]. It conveys precise estimates of tail behavior using letter values; boxplots lack such precise estimation.

that, in most markets, the $R^2$ values are close to 1. Further, the tail capturing low $R^2$ values consists of markets with a relatively small number of monthly users.

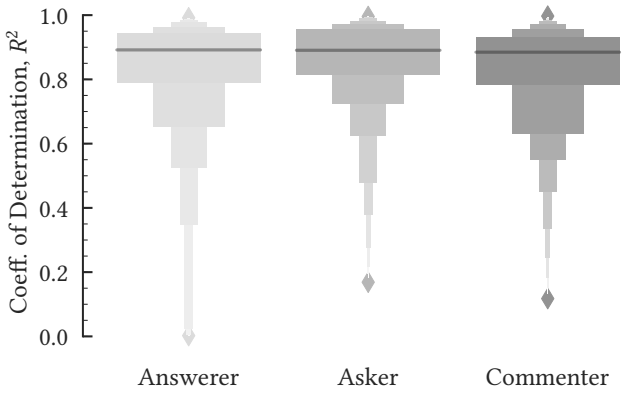



Figure 3.1: The distribution of coefficient of determination, $R^2$, for regressing user roles across 156 Stack Exchange. In most Stack Exchange sites, the role distribution is stable—as manifested by the arrangement of $R^2$ in letter-value plots.

**Number of Users.** The number of users is the only free input to our content generation models; the remaining inputs are functions of the number of users. In all these models, the growth or decline of number of users is exogenous—determined outside the model, by non-economic forces.

## 3.5 DATASET

We collected the latest release (September, 2017) of the Stack Exchange dataset. This snapshot is a complete archive of all activities in Stack Exchange sites. There are 169 sites in our collected dataset. For the purpose of empirical analysis, we only consider the sites that have been active for at least 12 months beyond the *ramp up* period (site created, but few or no activity). There are 156 such sites. The age of these sites vary from 14 months to 111 months, number of users from 1,072 to 547,175, and the number of posts (questions and answers) from 1,600 to 1,985,869. Further, the sites have small overlaps in user base; therefore, we can reasonably argue that the underlying markets are independent.

In Figure 3.2 we present letter value plots (in log-scale) to show the distribution of number of months (age), number of users, and number of posts for 156 Stack Exchange sites.

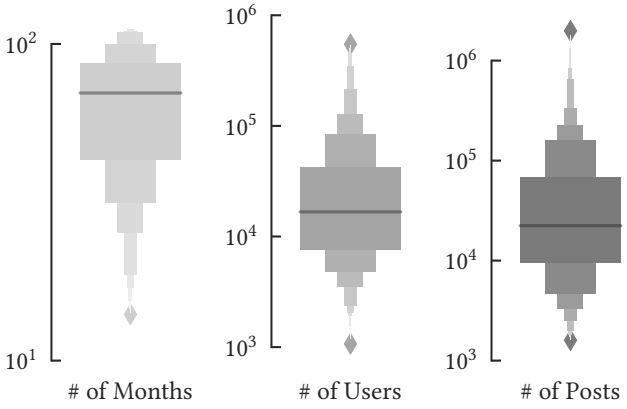



Figure 3.2: The log-scale distribution of number of months (age), number of users, and number of posts for 156 Stack Exchange sites.

## 3.6 EVALUATING OUR PROPOSED MODELS

In this section, we identify optimal models (basis and interaction) based on the accuracy of fitting content generation time series observed in our dataset (Section 3.6.1), and evaluate the performance of the optimal models in predicting content volume in long run (Section 3.6.2).

### 3.6.1 Model Fitting

We fit each variant of production model (basis and interaction), for each content type, to the observed content generation time series (monthly granularity), in each Stack Exchange site. Notice that among the different variants of production models, the models using power or exponential basis have a parsimonious set of parameters. For example, answer generation model using power basis function requires only three parameters for interactive essential interaction (See Section 3.4.2), and four parameters for remaining interaction types. In contrast, answer generation model using sigmoid basis function requires five parameters for interactive essential interaction, and six parameters for remaining interaction types.

**Parameter Estimation.** We learn the best-fit parameters for capturing the observed content generation time series. We restrict some parameters of our production models to be non-negative, e.g., non-negative exponents in power basis. These restrictions are important because the underlying factors positively affect the output. We use the trust-region reflective algorithm [72] to solve our constrained least square optimization problem. The algorithm is appropriate for solving non-linear least squares problems with constraints.

**Evaluation Method.** We evaluate fitting accuracy using four metrics: root mean square error (RMSE), normalized root mean square error (NRMSE), explained variance

score (EVS), and Akaike information criterion (AIC). Given two series for each content type, the observed series $N(t)$, and the prediction $\hat{N(t)}$ of the series by a model with $k$ parameters, we compute the four metrics as follows.

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(N(t) - \hat{N(t)})^2} \tag{3.6}$$

$$\text{NRMSE} = \frac{\text{RMSE}}{max(N(t)) - min(N(t))} \tag{3.7}$$

$$\text{EVS} = 1 - \frac{Var(N(t) - \hat{N(t)})}{Var(N(t))} \tag{3.8}$$

$$\text{AIC} = T * ln(\frac{1}{T}\sum_{t=1}^{T}(N(t) - \hat{N(t)})^2) + 2k \tag{3.9}$$

Among the four metrics, RMSE and NRMSE are error metrics (low value implies good fit), AIC is an information theoretic metric to capture the trade-off between model complexity and goodness-of-fit (low value implies good model), and EVS refers to a model's ability to capture variance in data (high value implies good model).

**Fitting Results.** We compare the fitting accuracy of production models for all Stack Exchange sites using the four metrics. Each metric is summarized via the mean, across all sites, for each content type. We use content generation time series with monthly granularity as observed data. We found that the models with the exponential and sigmoid basis functions do not fit the data for many Stack Exchange sites. Accordingly, in Table 3.3 we only present the results for production models with the power basis and different interaction types. Notice that the models with interactive essential interaction outperform the remaining models for all metrics and content types. We performed paired $t$-tests to determine if the improvements for interactive essential interaction are statistically significant; the results are positive with $p < 0.01$.

Thus we use production models with power basis and interactive essential interaction for prediction tasks.

### 3.6.2 Forecasting Content Generation

We apply production models with power basis and interactive essential interaction to forecast content volume in long run—one year ahead in the future. Specifically, we train each model using the content generation data from the first 12 months (beyond the ramp

Table 3.3: The comparison of fitting accuracy of production models (with power basis and different interaction types) for all Stack Exchange sites. The models with interactive essential interaction outperform the remaining models for all metrics and content types. The improvements for interactive essential interaction are statistically significant, validated via paired t-tests, where $p < 0.01$.

| Content | Interaction Type | Avg. RMSE | Avg. NRMSE | Avg. EVS | Avg. AIC |
|---|---|---|---|---|---|
| Question | Single Factor | 25.74 | 0.09 | 0.79 | 104.47 |
| Answer | Essential | 70.31 | 0.09 | 0.79 | 208.82 |
| | I. Essential | **64.62** | **0.08** | **0.83** | **196.39** |
| | Antagonistic | 72.77 | 0.09 | 0.78 | 210.96 |
| | Substitutable | 68.90 | 0.09 | 0.81 | 207.61 |
| Comment | Essential | 146.64 | 0.08 | 0.83 | 328.25 |
| | I. Essential | **137.23** | **0.08** | **0.85** | **318.24** |
| | Antagonistic | 155.97 | 0.09 | 0.82 | 334.12 |
| | Substitutable | 155.43 | 0.09 | 0.82 | 335.10 |

period), and then examine how well the model forecasts content dynamics in the next 12 months. We validate the forecasting capability by examining the overall prediction error (NRMSE).

We compute the prediction NRMSE across all Stack Exchange sites, and summarize the results using the mean ($\mu$) and variance ($\sigma$)— (i) question: $\mu = 0.11$, $\sigma = 0.08$; (ii) answer: $\mu = 0.12$, $\sigma = 0.09$; (iii) comments: $\mu = 0.11$, $\sigma = 0.10$. Notice that our models can forecast future content dynamics with high accuracy. We performed these experiments for different time granularity, e.g., week, month, quarter, and reached a consistent conclusion. We do not report these results for brevity.

## 3.7 CHARACTERIZING KNOWLEDGE MARKETS

In this section, we characterize the knowledge markets in Stack Exchange. We explain the best-fit models and their foundations (Section 3.7.1), reveal two key distributions that control the markets (Section 3.7.2), and uncover the stable core that maintains market equilibrium (Section 3.7.3).

### 3.7.1 Model Interpretation

First, we explain the best-fit models found in Section 3.6.1. We observe that content generation in Stack Exchange markets are best modeled through the combination of power basis and interactive essential interaction. In addition, we found that the best-fit exponents ($\lambda$ parameter in basis $g(x) = ax^{\lambda}$, where $x$ is a factor) of these models lie between 0 and 1 (inclusive), for all factors of all content types, for all Stack Exchange markets.

A model that uses the power basis (where exponents lie between 0 and 1) and interactive essential interaction is known as the Cobb-Douglas production function [73]. In its most standard form for production of a single output $z$ with two inputs $x_1$ and $x_2$, the function is as follows.

$$z = ax_1^{\lambda_1} x_2^{\lambda_2} \tag{3.10}$$

Here, the coefficient $a$ represents the *total factor productivity*—the portion of output not explained by the amount of inputs used in production [73]. As such, its level is determined by how efficiently the inputs are utilized in production. The exponents $\lambda_i$ represent the *output elasticity* of the inputs—the percentage change in output that results from the percentage change in a particular input [73].

The Cobb-Douglas function provides intuitive explanation for content generation in Stack Exchange markets. In particular, the explanation stands on three phenomena or principles: constant elasticity, diminishing returns, and returns to scale.

**Constant Elasticity.** In Stack Exchange markets, factors such as user participation and content dependency have *constant elasticity*—percentage increase in any of these inputs will have constant percentage increase in output [73], as claimed by the corresponding exponents in the model. For example, in ACADEMIA ($N_a = 6.93N_q^{0.18}U_a^{0.65}$), a 1% increase in number of answerers ($U_a$) leads to a 0.65% increase in number of answers ($N_a$).

**Diminishing Returns.** For a particular factor, when the exponent is less than 1, we observe *diminishing returns*—decrease in the marginal (incremental) output as an input is incrementally increased, while the other inputs are kept constant [73]. This "law of diminishing returns" has many interesting implications for the Stack Exchange markets, including the diminishing benefit of having a new participant in a market. For example, in ACADEMIA, if the number of answerers is 100, then the marginal contribution of a new answerer is $c(101^{0.65} - 100^{0.65}) = 0.129c$, where $c$ is a constant; in contrast, if the number of answerers is 110, then the marginal contribution of a new answerer is $c(111^{0.65} - 110^{0.65}) = 0.125c$. Thus, for answer generation in ACADEMIA, including a participant when the number of participants (system size) is 110 is likely to be less beneficial compared to including a participant when the system size is 100.

**Returns to scale.** The knowledge markets in Stack Exchange vary in terms of scale efficiency, as manifested by their *returns to scale*—the increase in output resulting from a proportionate increase in all inputs [73]. If a market has high returns to scale, then greater efficiency is obtained as the market moves from small- to large-scale operations. For example, in ACADEMIA, for answer generation, the returns to scale is $0.18 + 0.65 = 0.83 < 1$. The market becomes less efficient as answer generation is expanded, requiring more questions and answerers to increase the number of answers by same amount.

### 3.7.2 Two Key Distributions

Next, we discuss two key distributions that control content generation in knowledge markets, namely participant activity and subject POV (perspective). These two distributions induce the three phenomena reported in section 3.7.1.

**Participant Activity.** The distribution of participant activities implicitly drives a market's return in terms of user participation, as manifested by the corresponding exponent. For example, in a hypothetical knowledge market where each answerer contributes equally, the answer generation model should be $N_a = AN_q^{\lambda_1}U_a^{1.0}$. In reality, the distribution of participant activities is a size dependent distribution controlled by the number of participants (system size). As the system size increases, most participants contribute to the head of the distribution (few activities), whereas very few join the tail (many activities).

We systematically reveal the size dependent distribution for participant activities in three steps. First, we empirically a fit power-law distribution to the activities of participants in a month, for each month, for each Stack Exchange market. We follow the standard procedure to fit a power-law distribution [74]. We observe that the power-law well describes the monthly

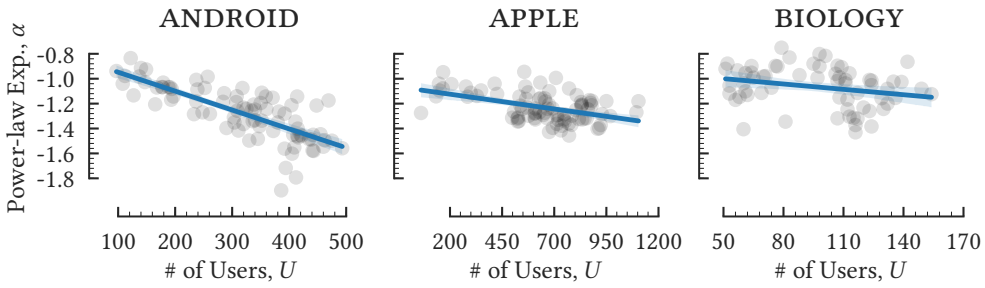



Figure 3.3: The visibility of size dependent distribution: strong—ANDROID; moderate—APPLE; and weak—BIOLOGY. In most Stack Exchange markets, the power-law exponent decreases with system size, similar to ANDROID. In other markets, there exists a non-zero correlation between system size and power-law exponent.

activity distributions. Second, we plot the exponents of the power-law against the number of participants for all observed months in a market, for each market in Stack Exchange. We observe that for most Stack Exchange markets, the power-law exponent decreases as the system size increases. Third, we apply linear regression to reveal the relationship between power-law exponent and system size. We observe that in general power-law exponent is negatively correlated with system size. This negative correlation is strongly visible in big knowledge markets that have at least 500 monthly participants in each month.

In Figure 3.3 we present empirical evidence of the size dependent distribution for answer generation in three Stack Exchange markets: ANDROID, APPLE, and BIOLOGY. We choose these examples to cover three possible visibilities of the size dependent distribution, as manifested by the correlation between the power-law exponent and system size—strong correlation ($|r^2| \geq 0.5$), moderate correlation ($0.3 \leq |r^2| < 0.5$), and weak correlation ($|r^2| < 0.3$).

**Subject POV.** The distribution of subject POV implicitly drives a market's return in terms of content dependency, as manifested by the corresponding exponent. Subject POV refers to the number of distinct perspectives on a particular content (e.g., questions) that imposes a conceptual limit to the number of dependent contents (e.g., answers). For example, an open-ended question such as 'What's your favorite book?' has many possible answers, whereas a close-ended question such as 'What's the solution for 3x+5 = 2?' has a single correct answer. In reality, most questions are neither completely open-ended nor completely closed; however, from an answerer's perspective, there's a diminishing utility in answering a question that already has an answer. This diminishing utility varies from question to question—questions asking for recommendations attract many answers, whereas questions seeking factual information attract few answers.

### 3.7.3 Uncovering the Stable Core

We uncover a stable user community in each Stack Exchange market, that maintains the *dynamic equilibria*—the increase or decrease in overall user community does not affect the Cobb-Douglas models. We assert that this stable user community generates a large fraction of high-threshold contents that require more effort, e.g, answers and comments, whereas the remaining users are unstable and contribute a small fraction of high-threshold contents.

We reveal the presence of the stable core by summarizing the answer contribution of users with different tenure levels (# of active months). First, for each Stack Exchange market, we apply equal-width binning to categorize its users into five tenure levels. Then, we plot the distribution of monthly answer contribution by the users of each category using a letter-value

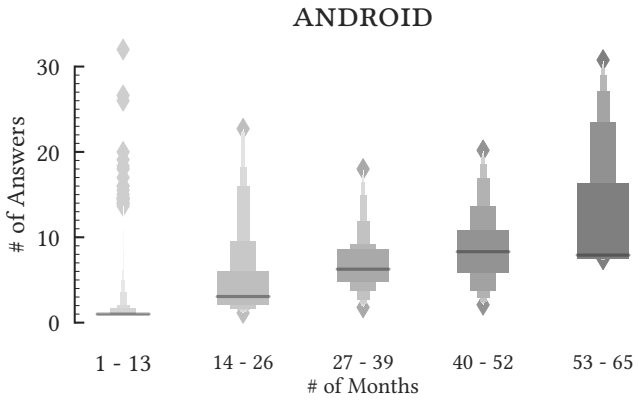



Figure 3.4: The distribution of monthly answer contribution of users with different # of active months, for ANDROID. The users who contribute for many months also contribute a large number of answers.

plot. We present the letter-value plots for ANDROID in Figure 3.4. We observe that monthly answer contribution is an increasing function of tenure level—the users who contribute for many months also contribute a large number of answers.

## 3.8 FAILURES AT SCALE

In this section, we discuss how and why knowledge markets may fail at scale. We first empirically examine diseconomies of scale (Section 3.8.1), then analyze the effects of scale on market health (Section 3.8.2), and finally study user exchangeability under scale changes (Section 3.8.3).

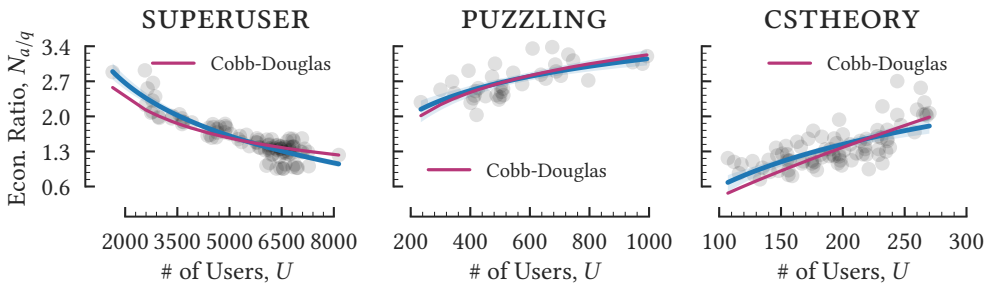



Figure 3.5: Diseconomies/economies of scale: the ratio of answers to questions decreasing/increasing with the increase in number of users. Most Stack Exchange markets exhibit diseconomies of scale. Examples: strong diseconomies—SUPERUSER; weak economies—PUZZLING; and strong economies—CSTHEORY.

### 3.8.1 Diseconomies of Scale

First, we examine disceconomies of scale—the ratio of answers to questions declining with the increase in number of users. The opposite of diseconomies is economies, when the ratio increases with the increase in number of users. The concept of diseconomies is important because a decrease in the answer to question ratio implies an increase in the gap between market supply (answer) and demand (question). In fact, if the ratio falls below 1.0, the gap becomes critical—guaranteeing there will be some questions with no answers.

In Figure 3.5 we present the economies and diseconomies of scale in three Stack Exchange markets: CSTHEORY, PUZZLING and SUPERUSER. We choose these examples to cover three cases: strong diseconomies, strong economies, and weak economies. Among the three markets, SUPERUSER shows strong diseconomies of scale: if the number of users increases by 1%, then the answer to question ratio declines by 0.95%. The other two markets show economies of scale, where CSTHEORY shows strong economies: if the number of users increases by 1%, then the answer to question ratio increases by 0.8%; and PUZZLING shows weak economies: if the the number of users increases by 1%, then the answer to question ratio increases by 0.2%. Note that most markets, especially the ones with more than 500 monthly active participants, exhibit diseconomies of scale similar to SUPERUSER. Only five markets exhibit strong economies of scale in Stack Exchange: CSTHEORY, EXPRESSIONENGINE, PUZZLING, JA_STACKOVERFLOW, and SOFTWAREENGINEERING.

The Cobb-Douglas curves well fit the empirical trends of economies and diseconomies (as shown in Figure 3.5). We derive these curves by dividing the answer models by the corresponding question models, and subsequently developing curves that capture economies and diseconomies ($N_{a/q}$) as a function of number of users (system size).

The Cobb-Douglas models well explain the economies and diseconomies of scale. As per the models, the primary cause of diseconomies is the difference between the diminishing returns of questions and answers for user participation. In other words, in most markets, for user input, the marginal question output is higher compared to the marginal answer output, i.e., an average user is likely to ask more questions and provide few answers. This causes the ratio of answers to questions to decline with an increase in the number of users.

### 3.8.2 Analyzing Health

Next, we examine the disadvantage of scale through two health metrics: $H_1$—the fraction of answered questions (questions with at least one answer); and $H_2$—the fraction of questions with an accepted answer (questions for which the asker marked an answer as accepted). $H_1$

and $H_2$ capture the true gap between market supply (answers) and demand (questions). An increase in the number of users may cause a decline in $H_1$ and $H_2$, as both metrics are related to the ratio of answers to questions. In fact, if the ratio falls below 1.0, it guarantees the decline of both metrics.

In Figure 6.4 we present the health advantage and disadvantage of scale (through $H_1$ and $H_2$) for three Stack Exchange markets: CSTHEORY, PUZZLING and SUPERUSER. We observe that the results are consistent with our analysis of economies and diseconomies—CSTHEORY exhibits health advantage at scale, PUZZLING remains stable, whereas SUPERUSER exhibits disadvantage at scale. These three examples cover the possible health effects of scale in knowledge markets.

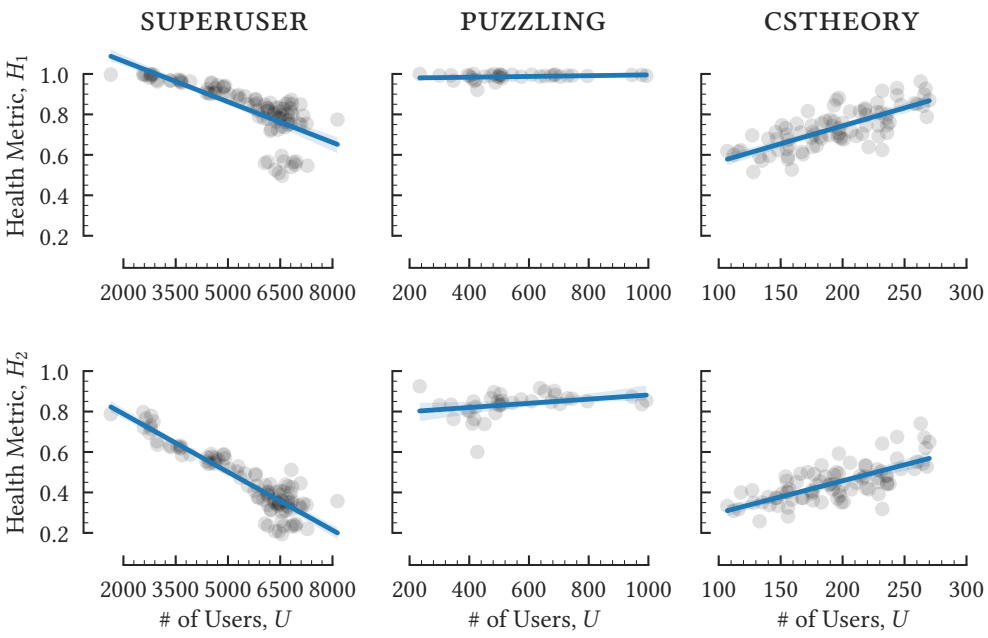



Figure 3.6: Health disadvantage/advantage of scale: $H_1$—the fraction of answered questions, and $H_2$—the fraction of questions with accepted answer, decreasing/increasing with the increase in number of users. Most Stack Exchange markets exhibit health disadvantage at scale. Examples: disadvantage—SUPERUSER; neutral—PUZZLING; and advantage—CSTHEORY.

### 3.8.3 Effects on Exchangeability

Finally, we empirically study the effects of scale on user exchangeability. By exchangeability, we specifically mean the gap between the top contributors and other participants in a knowledge market. Studying this gap is important because it can reveal if a market's success or failure depends on a small group of users.

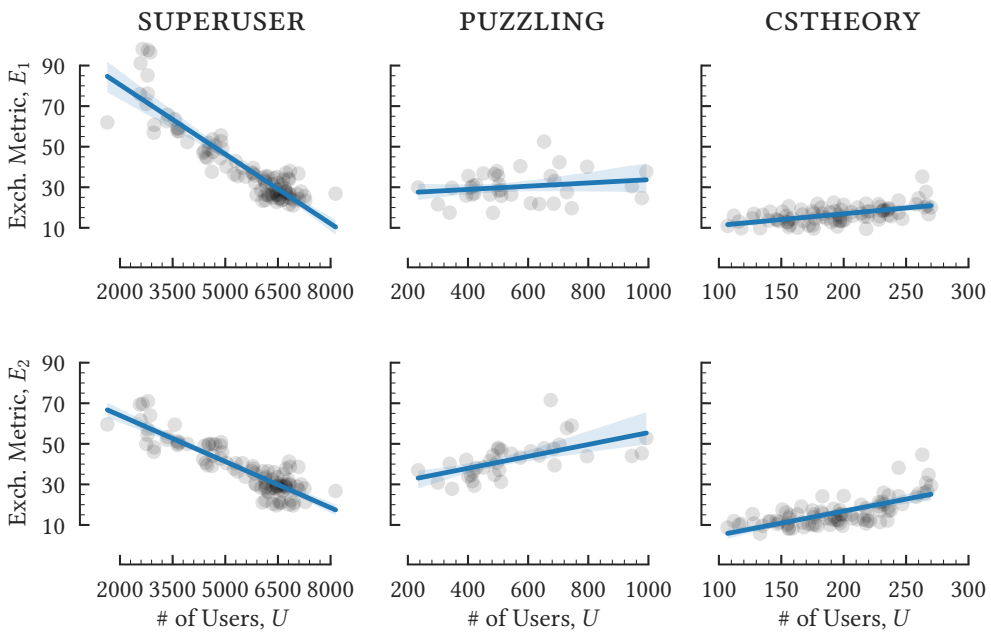



Figure 3.7: User exchangeability under scale: the gap ($E_1$ and $E_2$) between the top contributors and other participants in a knowledge market decreasing or increasing with the increase in number of users. Most markets exhibit a large gap between the top contributors and other participants. Examples: high dissimilarity—SUPERUSER; moderate dissimilarity—PUZZLING; and low dissimilarity—CSTHEORY.

To empirically study user exchangeability, we define two metrics that reflect the gap between the top contributors and other participants in a knowledge market. Note that, we only consider the active participants who contributed at least one content. The first metric $E_1$ is defined as the ratio of contribution between the top 5% and the bottom 5% of users. For computing $E_1$, we measure the contribution of a user $v$ as the ratio $N^v_{a/q}$ of the number of answers $N^v_a$ provided by the user to the number of questions $N^v_q$ asked by the user. Notice that $E_1$ is a ratio based metric and we define user contribution to be consistent with this metric. The second metric $E_2$ is defined as the sum of two distances: (i) the distance between the contribution of the top 5% of users and the median 5% of users, and (ii) the distance between the contribution of the median 5% of users and the bottom 5% of users. For computing $E_2$, we measure the contribution of a user $v$ as a tuple $< N^v_a, N^v_q >$, consisting of the number of answers $N^v_a$ provided by the user and the number of questions $N^v_q$ asked by the user. Notice that $E_2$ is an interval based metric and we define user contribution to be consistent with this metric. While both metrics have certain limitations, e.g., they are sensitive to outliers, these metrics allow us to comprehend user exchegeability to some extent.

In Figure 3.7 we present the exchangeability of users under scale changes (through $E_1$ and $E_2$) for three Stack Exchange markets: CSTHEORY, PUZZLING and SUPERUSER. Among the three markets, SUPERUSER exhibits the highest gap between the top contributors and the other participants. However, as the number of participants increases, this gap decreases, i.e., the users become more exchangeable. In contrast, CSTHEORY exhibits the lowest gap between the top contributors and the other participants. However, as the number of participants increase, this gap increases, i.e., the users become less exchangeable.

## 3.9 IMPLICATIONS

Our work promotes two new research directions—size-dependent mechanism design and content dependency in social media—while advancing several others—metrics of market health, power law of participation, and microfoundations of knowledge markets.

**Size-Dependent Mechanism Design.** We reveal that the health of a knowledge market depends on the market's size. A natural implication of this dependency is that site operators should adjust mechanisms based on the number of participants. For example, a site operator can decide between retaining existing users (via incentives) and attracting new users (via advertising) based on the number of participants and their activity distribution.

**Content Dependency in Social Media.** We observe that many social media platforms support several possible user actions with "complex dependencies". For example, in Facebook, a post is the root content (primary), comments on the post nest below the post (secondary), and replies to these comments nest beneath the original comments (tertiary). Further, a user can react to any of these content types with several possible reactions. Overall user activity in Facebook is distributed across these possible actions with complex dependencies, which drives the platform's health.

**Metrics of Market Health.** We demonstrate the presence of diseconomies of scale with several metrics that partially capture the health of a knowledge market. While we concentrate on content-generation based *production metrics*, our concepts can be extended for page-view based *consumption metrics* as well. Also, there is room for developing new health metrics that capture a more detailed picture of a knowledge market's health including *market efficiency*—the degree to which market price (amount of responses and reactions) is an unbiased estimate of the true value of the investment (user effort in content generation) [75].

**Power Law of Participation.** In Stack Exchange markets, a small fraction of the user community participate in high-engagement activities (e.g., linking similar questions), whereas the larger fraction participate in low-threshold activities (e.g., voting). This asymmetry leads to a *Power Law of Participation* [76]. We assert that both low-threshold and

high-engagement activities are required for a knowledge market's survival, and should proportionately increase with the increase in number of participants. However, in reality, for most knowledge markets, the size of the user community contributing high-engagement activities does not scale with the system size. This creates a "gap" between market supply and demand, and consequently affects market health.

**Microfoundations of Knowledge Markets.** The size-dependent distribution of user contribution implies that users who join a community later in its lifecycle exhibit different behavior than those who were present from the beginning. This very well may imply that the distribution of individual user behaviors (not just their overall production) is "also" a function of the system size. We should expect to see a stable user behavior distribution over time for markets that appear to be more scale-insensitive; preliminary results suggest that this may indeed be the case [77].

## 3.10 LIMITATIONS

We discuss several limitations of our work. First, the economic production models do not account for user growth. While there exist several user growth models for two-sided markets [48], membership based websites [17], and online social networks [15, 16, 18], it would be useful to introduce economic user growth models that complement our proposed content generation models. Specifically, there is a need to develop resource-based user growth models that account for market health. A potential direction in this research is to extend the Malthusian growth model [78]. Second, the proposed production models inherit the fundamental assumptions of macroeconomics: an aggregate is homogeneous (without looking into its internal composition), and aggregates are functionally related etc. [79]. It would be useful to empirically study these assumptions for real-world knowledge markets.

## 3.11 CONCLUSION

In this study, we examined the CQA websites on Stack Exchange platform through an economic lens by modeling them as knowledge markets. In particular, we designed a set of production models to capture the content generation dynamics in these markets. The resulting best-fit model, Cobb-Douglas, predicts the production of content in Stack Exchange markets with high accuracy. We showed that the model provides intuitive explanations for content generation. Specifically, it reveals that factors of content generation such as user participation and content dependency have *constant elasticity*; in many markets, factors

exhibit *diminishing returns*; markets vary according to their *returns to scale*; and finally many markets exhibit *diseconomies of scale*. We further investigated these prognoses by showing the presence of diseconomies of scale in terms of content production, and several measures of market health. The implications of our work are two-fold: site operators need to design incentives as a function of number of participants; there is a need to develop Economic lenses that can shed insights into the complex dependencies amongst different content types and participant actions in general social networks.

In this chapter, we showed how users generate content. As the size of a community grows, new users do not contribute as much as old users. In the next chapter, we will learn more about the reason why new users do not contribute.

# CHAPTER 4: QUANTIFYING VOTER BIASES

Our first study showed that users who join a Stack Exchange site more recently contribute fewer answers per month than those who joined at an early stage. This observation raises the question: is there any discrepancy between the motivation and perhaps the incentives of old users and newcomers. Specifically, could it be that what is causing the newcomers to give up is not the poor quality of their content but the biased social judgment of their peers? This chapter discusses our second study, where we examine biases in social judgment, specifically in the form of votes.

## 4.1 INTRODUCTION

In many online platforms, users receive up- and down- votes on content from fellow community members. An aggregate of the votes is commonly used as a proxy for content quality in a variety of applications, such as search and recommendation [80, 81, 82, 83]. The principle of the *wisdom of the crowds* underlies this quantification, where the mean of judgments on content tends to its true value. The principle rests on the assumption that individuals can make *independent* judgments, and that the crowd comprises agents with *heterogeneous* cognitive abilities [84].

However, in most online platforms, individuals are prone to using cognitive heuristics that influence their voting behavior and prevent independent judgments [85, 86]. These heuristics incorporate different impression signals adjacent to the content—such as the reputation of the contributing user [87], aggregate vote thus far [88], and position of content [89]—as input to help individuals make quick decisions about the quality of content. Prior literature suggests that the use of impression signals as shortcuts to make voting decisions results in biases [88, 89], where the aggregate of votes becomes an unreliable measure for content quality. We designate these biases as *voter biases*, which stem from the use of impression signals by voters.

There is a plethora of research on detecting and quantifying voter biases in online platforms [13, 85, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98]. Broadly, researchers have adopted one of the following two approaches: 1) conduct experiments to create different voting conditions for studying participants [13, 88, 89, 90, 91, 93, 98]; 2) develop statistical models to analyze historical voting data [85, 92, 94, 95, 96, 97]. Both approaches have limitations. First, it is hard to perform randomized experiments in actual platforms due to feasibility, ethical issue, or cost [99]. In addition, researchers not employed at a social media platform

are at a disadvantage in conducting such experiments on that platform. Second, statistical models on voter biases often lack causal validity: the derived estimates measure only the magnitude of association, rather than the magnitude and direction of causation required for quantifying voter biases. These limitations of prior research motivate the present work.

**Present Work.** In this study, we quantify the degree of voter biases in online platforms. We concentrate on three distinct biases that appear in many platforms—namely, *reputation bias*, *social influence bias*, and *position bias*. Reputation bias captures how the reputation of a user who creates content affects the aggregate vote on that content; social influence bias captures how the initial votes affect the subsequent votes on the same content; position bias captures how the position of content affects the aggregate vote. We study these biases in an observational setting, where we estimate the causal effects of their associated impression signals on the observed votes.

The key idea of our approach is to formulate voter bias quantification as a causal inference problem. Motivated by the successes of the instrumental variable (IV) framework in studying causal phenomena in Economics and other social science disciplines—e.g., how education affects earnings [100], how campaign spending affects senate elections [101], and how income inequality affects corruption [102]—we adopt the IV framework to solve our bias quantification problem. The IV framework consists of four components: outcome (dependent variable), exposure (independent variable), instrument (a variable that affects the outcome only through the exposure), and control (other covariates of interest). We operationalize these IV components using variables compiled from log data. We use impression signals as exposure, aggregate feedback as the outcome, and estimate the causal effect of exposure on the outcome by identifying proper instrument and control.

Identifying an instrument is hard [103]. A valid instrument must satisfy three conditions as follows. First, the *relevance* condition requires the instrument to be correlated with the exposure. Second, the *exclusion restriction* requires that the instrument does not affect the outcome directly. Third, the *marginal exchangeability* requires that the instrument and the outcome do not share causes. Of these three conditions, only the relevance condition is empirically verifiable; the remaining two conditions need to be justified through argumentation [104]. Using large-scale log data from Stack Exchange websites, we identify a set of nuanced instrumental variables for quantifying voter biases. We carefully analyze our proposed instruments to reason about their ability to meet the three instrumental conditions and then select a final set of instruments. We use the final instruments to estimate the causal effects of impression signals on the observed votes using two-stage least squares (2SLS) regression. These regression coefficients provide unbiased causal estimates for quantifying voter biases.

This study makes the following contributions.

**Bias quantification.** We quantify three types of voter biases by estimating the causal effects of impression signals on the aggregate of votes. Prior research has either used randomized experiments or statistical modeling for quantifying voter biases. While the former can help us identify causal effects, randomized trials are not an option for researchers who work outside the social media platform with observational data. Statistical models help us identify correlation, *not* causation. In contrast, we use an instrumental variable framework by first identifying a set of instrumental variables and then carefully analyzing their validity. The significance of our contribution lies in our framework's ability to identify from observational data, causal factors (impression signals) that affect an individual's vote.

**Findings.** We find that prior work on bias estimation with observational data has significantly underestimated the degree to which factors influence an individual's vote. Our empirical results show that OLS underestimates reputation bias (1.6–2.2x for gold badges) and position bias (up to 1.9x for the initial position), and overestimates social influence bias (1.8–2.3x for initial votes). Furthermore, we find that different impression signals vary in their effect: the badge type (gold, silver, bronze) plays a bigger role in influencing the vote than does reputation score. Also, we find the degree to which each impression signal influences vote depends on the community. This result is significant for two reasons: first, the influence of some of these factors is much more ($\sim$100% more) than previously understood from statistical models on observational data; despite statistical models estimating regression coefficients, prior work used these coefficients to impute causation, an incorrect inference. Second, had platforms attempted to de-bias with results from prior work, they would have significantly underestimated the effects of reputation and answer position.

**Significance.** Our identification of causal factors that influence votes has a significant bearing on research in voter bias in particular, as well as the broader CSCW community. First, there are practical implications. Impression signals (answer position, user reputation, prior vote) play a significant role in influencing an individual's vote, at times twice as much as previously understood. Furthermore, the effect of these signals varies by community type (with different content and social norms governing discussions). Second, our work has implications on the future interface design of these platforms. For example, these platforms may conceal impression signals prior to the vote, or delay the vote itself to address social influence bias. Future research is needed, however, to understand the effect of these suggestions.

Third, our work informs policy. By identifying causal factors, our work offers social media platforms a way to transparently de-bias votes. The de-biasing may be community dependent. Finally, by introducing the instrumental variable approach to the CSCW community, to identify causal factors from observational data, we hope that more researchers will adopt it to study other questions of interest: e.g., gender and racial bias online.

The rest of this study is organized as follows. We define our problem in Section 4.2 and discuss the related work in Section 4.3. We describe our data in Section 4.4. We then explain how our method works in Section 4.5. Section 4.6 reports the results of our study. We discuss the implications of our research in Section 4.7 and the limitations in Section 6.5. Finally, we conclude in Section 6.6.

## 4.2 VOTER BIAS

The goal of this study is to quantify the degree of voter biases in online platforms. We concentrate on three distinct biases: reputation bias, social influence bias, and position bias. To quantify these biases, we estimate the causal effects of their associated signals on the observed votes. In Figure 4.1, we present a sample page from English Stack Exchange, annotated with different signals that may induce the above-mentioned biases.

**Reputation Bias.** In content-based platforms (such as Stack Exchange and Reddit), reputation system incorporates the votes on content into the content creator's reputation [105, 106]. In Stack Exchange, for example, votes on content translate into reputation score and badges for the contributing user [30, 105]. The reputation score and badges acquired by each user are visible to all community members, who may use this information to infer the quality of the user's future contributions. Inferring content quality based on user reputation forms the basis of *reputation bias*—when the reputation of a user influences the votes he/she receives on content. We know from prior work that reputation exhibits a Matthew effect [107]: early reputation increases the chances of future reputation via upvotes. Consider a counterfactual scenario, where two users with different levels of reputation create "identical" content; then, reputation bias implies that the user with a higher reputation will receive more upvotes.

**Social Influence Bias.** The concept of *social influence* in collective action is well-known [91]: contrary to the *wisdom of the crowds* principle, individuals do not make independent decisions; instead, their decision is influenced by the prior decision of peers. Social influence affects a variety of user activities in online platforms, including voting behavior on content [85, 86, 88, 96]. Since most platforms reveal the aggregate vote thus far, the initial votes act as a social signal to influence the subsequent voters, forming the basis of *social*

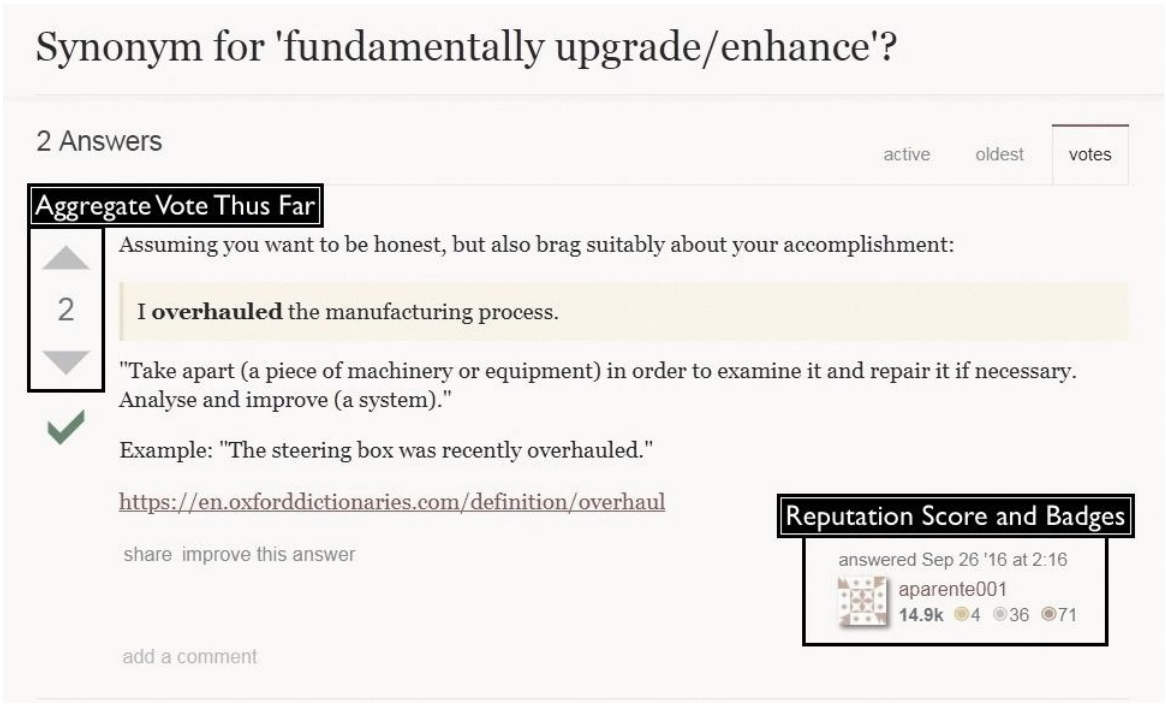



Figure 4.1: A sample page from English Stack Exchange, annotated with different signals that may induce voter biases. For all answers to a question: 1) the score at top left corner shows the aggregate vote thus far (*social signal*), which may induce social influence bias; 2) the statistics at bottom right corner shows the reputation score and badges acquired by the answerer (*reputation signal*), which may induce reputation bias; 3) the answers are presented in a sequential order (*position signal*), which may induce position bias.

*influence bias*. We know from prior work that for platforms that reveal social signal, users exhibit a herding effect [13]: the first few votes on content can unduly skew the subsequent votes. Consider a counterfactual scenario, where two "identical" content initially receive dissimilar votes; then, social influence bias implies that the content with higher aggregate vote thus far will receive more upvotes.

**Position Bias.** Many online platforms present content in some order, using a list-style format. For example, in Stack Exchange, answers are sorted based on the aggregate vote thus far. The position of content in a list-style format plays a critical role in deciding how many users will pay attention to it, and interact with it via clicks [108] or votes [89]. Users pay more attention to items at the top of a list, creating a skewed model of interaction for the items. A consequence of this skewed interaction is *position bias*—when the position of content influences the votes on it. Consider a counterfactual scenario, where two "identical" content are located in different positions within a web page; then, position bias implies that the content at the higher position will receive more upvotes.

**Relationship between Social Influence and Position.** In many platforms, the presentation order of content depends on the aggregate user feedback. In Stack Exchange sites, the default presentation order of answers is the aggregate vote thus far. Quora uses a wide variety of factors to decide the order of answers, including the upvotes and downvotes on the answers. Such vote-dependent ordering scheme imposes a critical challenge in estimating the causal effects of social influence signal and position signal, as the two signals vary together. As such, *the lack of longitudinal variation* in the relationship between the two signals makes it difficult to isolate the effects of their corresponding biases.

## 4.3 RELATED WORK

Our work draws from and improves upon, a rich literature on online voting behavior and voter biases. Since this study focuses on quantifying voter biases, we provide a taxonomy of related work on voter biases (Table 4.1). We now discuss several pioneering works on understanding voting behavior and voter biases.

**Voting Behavior.** Recent research has made significant advancements towards the understanding of rating and voting behaviors in online platforms [109, 110, 111, 112, 113]. Gilbert [109] reported the widespread *underprovisioning of votes on Reddit*: the users overlooked 52% of the most popular links the first time they were submitted. Using data from Amazon product reviews, Sipos et al. [110] showed that users do not make independent voting decisions. Instead, the decision to vote and the polarity of vote depend on the *context*: a review receives more votes if it is misranked, and the polarity of votes becomes more positive/negative with the degree of misranking. Glenski et al. [112] found that most Reddit users do not read the article that they vote on. In a later work, Glenski et al. [113] used an Internet game called GuessTheKarma to collect independent preference judgments (free from social and ranking effects) for 400 pairs of images. They found that Reddit scores are not very good predictors of the actual preferences for items as measured by GuessTheKarma. In this study, we examine three distinct cognitive biases that affect user voting behavior. We quantify these biases by estimating the causal effects of their associated signals on the observed votes.

**Reputation Bias.** Prior works on online reputation suggest that past reputation may be useful in predicting current success [87, 114, 115, 116, 117, 118] (also known as "superstar economics" [119]). Beuscart et al. [87] observed that in MySpace Music, most of the audience is focused on a few stars. These stars are established music artists who signed on major labels. Based on a user study on Twitter, Pal et al. [114] reported that the popular users get a boost in their authority rating due to the "name value". Tausczik et al. [115] found

that in MathOverflow, both offline and online reputation are correlated with the perceived quality of contributions. Paul et al. [116] found that Quora users judge the reputation of other users based on their past contributions. Liang [117] showed that in Reddit, users with higher comment karma tend to produce questions and comments with higher ratings. Budzinski et al. [118] analyzed a sample of YouTube stars to show that past success positively and significantly influences current success. While these prior studies show the evidence of reputation bias, they do not provide any bias quantification. In this study, we provide a quantification of reputation bias through causal estimates.

Table 4.1: A taxonomy of existing literature on voter biases.

| **Bias** | **Approach** | **References** | **Summary** |
|---|---|---|---|
| Reputation Bias | Correlation Study | [87], [114], [115], [116], [117], [118] | Show some evidence of correlation between a user's past reputation and current success [87, 114, 115, 116, 117, 118]. |
| Social Influence Bias | Randomized Experiment | [90], [91], [93], [98], [13] | Create different decision making (say voting) conditions for study participants by varying the availability of preceding decisions [90, 91, 98], and purposefully engineered initial decision [13, 93]. |
| | AMT Simulation | [88], [120], [86] | Simulate alternative voting conditions of platform in Amazon Mechanical Turk (AMT) by varying the availability of preceding decisions [86, 88, 120]. |
| | Statistical Model | [92], [94], [95], [96], [97], [85] | Develop statistical model for quantifying bias: Pólya Urn [92, 97], nonparametric significance test [94], additive generative model [95], Poisson regression [96], logistic regression [85]. |
| | Matching Method | [121], [122] | Contrast aggregate user feedback (say ratings) on the same object in two different platforms via matching [122]. |
| Position Bias | Randomized Experiment | [89], [88], [98] | Create different decision making (say voting) conditions for study participants by varying content ordering policies [88, 89, 98] |
| | Statistical Model | [96], [97], [123] | Develop statistical model for studying bias: Poisson regression [96], Pólya Urn [97], counterfactual inference [123]. |
| | Matching Method | [124], [108] | Contrast aggregate user feedback (say ratings) for objects occupying similar positions [108, 124]. |

**Social Influence Bias.** Since the musiclab experiment by Salganik et al. [90], a large body of work has been devoted to the social influence bias [85, 86, 88, 91, 92, 93, 94, 96, 97, 98, 121], and its resultant herding effect [13, 95, 120, 122]. A majority of the work tends to fall into one of two categories—1) Experimental Study: randomized experiment [13, 90, 91, 93, 98], simulation via Amazon Mechanical Turk (AMT) [86, 88, 120]; and 2) Observational Study: statistical model [85, 92, 94, 95, 96, 97], matching method [121, 122]. Randomized experiments provide a nuanced way to quantify the degree of social influence bias in online platforms; however, often, these experiments are infeasible due to ethical issues, or cost. AMT based simulations fall short in representing the actual voting conditions of a platform. Prior observational studies have used a wide variety of statistical models—Pólya Urn [92, 97], nonparametric significance test [94], additive generative model [95], Poisson regression [96], logistic regression [85]—for quantifying social influence bias. However, these studies lack causal validation: the estimates measure only the magnitude of association, rather than the magnitude and direction of causation. For example, in a regression-based herd model, herding behavior could be correlated with the intrinsic quality of content [97]. Therefore, it is difficult to separate the social influence bias from the inherent quality and quantify its effect. In this study, we adopt the method of instrumental variables to quantify social influence bias.

**Position Bias.** In recent years, there has been significant interest in studying position bias in online platforms [88, 89, 96, 97, 98, 108, 123, 124]. Notably, researchers performed several experimental studies in AMT, where they created different voting conditions for study participants by varying content ordering policies [88, 89, 98]. Hogg et al. [88] revealed that social signals affect item popularity about half as much as position and content do. Abeliuk et al. [98] showed that the unpredictability of voting outcome is a consequence of the ordering policy. Lerman et al. [89] found that different policies for ordering content could improve peer recommendation by steering user attention. In this study, we study position bias in an observation setup, in which it is difficult to isolate the position bias from the social influence bias. To address this problem, we develop a joint IV model that quantifies both position bias and social influence bias.

## 4.4 DATA AND VARIABLES

In this section, we first discuss the choice of our data source (Section 4.4.1), then describe the datasets that we use in this study (Section 4.4.2); and finally present the variables that we accumulate from the datasets (Section 4.4.3).

### 4.4.1 Choice of Data Source

We seek online platforms that satisfy the following criteria: content is user-generated and integral to the platform's success, the position of content and reputation of the contributing user depend upon votes, and the user interface contains various impression signals that may influence the votes. Content-based online platforms such as Quora, Reddit, and Stack Exchange satisfy these criteria. Among them, Reddit and Stack Exchange have publicly available datasets.

We selected the Stack Exchange dataset over Reddit for the following reasons: 1) the Stack Exchange dataset is a complete archive with no missing data (prior work [125] indicates that the Reddit dataset is not complete), which prevents potential selection bias; 2) the governing rules are the same for all Stack Exchange sites (in contrast, subreddits can have different governing rules), which allows us to compare the results across different Stack Exchanges; and 3) the incentives in Stack Exchange sites have been designed for getting to a "correct" answer to a question rather than invoking a discussion as is sometimes the case in Reddit, which makes the Stack Exchange content more focused.

### 4.4.2 Stack Exchange Dataset

Stack Exchange is a network of community question answering websites, where millions of users regularly ask and answer questions on a variety of topics. In addition to asking and answering questions, users can also evaluate answers by voting for them. The votes, in aggregate, reflect the community's feedback about the quality of content and are used by Stack Exchange to recognize the most helpful answers.

Table 4.2: Descriptive statistics for the selected Stack Exchange sites.

| **Site** | **Category** | **# Users** | **# Questions** | **# Answers** |
|---|---|---|---|---|
| English | Culture | 169,037 | 87,679 | 210,338 |
| Superuser | Technology | 547,175 | 356,866 | 529,214 |
| Math | Science | 356,699 | 822,059 | 1,160,697 |

**Use of Published Data.** We obtained Stack Exchange data from `https://archive.org/details/stackexchange` on September 2017 (published by Stack Exchange under the CC BY-SA 3.0 license). This snapshot is a complete archive of user-contributed content on the Stack Exchange network. In this study, we analyze three Stack Exchange sites: English, Superuser, and Math.

**Inclusion Criteria.** We select the above-mentioned sites for several reasons. First, the three sites represent the three major themes or categories in Stack Exchange: culture [English], technology [Superuser], and science [Math]. Second, apart from Superuser, the remaining two sites are the largest in their category in terms of the number of answers. Superuser is the second largest site in its category, followed by StackOverflow; we discard StackOverflow due to its massive scale difference in comparison to the remaining sites. Third, the sites vary in terms of their susceptibility to voter biases, owing to content that requires interpretation. For example, the quality of answers in English is a lot more subjective compared to the quality of answers in Math. Table 4.2 presents descriptive statistics for the three sites analyzed in this study.

### 4.4.3 Variables

In Stack Exchange sites, questions and answers are the primary content. Answer quality is especially important for these platforms as they thrive to provide answers. For this reason, we analyze the votes on answers. We compile a wide range of variables to capture the voter biases, the factors related to these biases, and the potential effects of these biases. Table 4.3 describes the variables used in this study.

## 4.5 METHOD

In this section, we first discuss our choice of method for voter bias quantification (Section 4.5.1), then explain the fundamentals of the chosen method (Section 4.5.2); and finally present our models for quantifying voter bias (Section 4.5.3 and 4.5.4)

### 4.5.1 Choice of Method

The goal of this study is to quantify the degree of voter biases in online platforms. To determine these biases, we need to estimate the *causal effects* of different impression signals on the observed votes. Estimating causal effects from observational data is exceptionally challenging [126]. The main reason is that there may exist hidden confounders that affect both independent (say impression signal) and dependent (observed votes) variables. A hidden confounder may explain the degree of association between the variables, which prevents standard regressions methods from providing causal estimates [103, 126]. We observe that our voter bias quantification problem is susceptible to several hidden confounders, such as

Table 4.3: The description of variables used in this study. The variables fall into four groups based on the following constructs: site (the Stack Exchange site), question (the question that has been addressed by the answer), answer (the answer in consideration), and answerer (the user who created the answer).

| ID | Variable | Description |
|---|---|---|
| $V_1$ | Site | The Stack Exchange site in consideration |
| $V_2$ | T | The limiting time of bias formation specific to the question |
| $V_3$ | QuestionViewCount | No. of users who viewed the question |
| $V_4$ | QuestionFavoriteCount | No. of users who favorited the question |
| $V_5$ | QuestionScore | Aggregate vote (total upvotes - total downvotes) on the question |
| $V_6$ | QuestionScoreT- | Aggregate vote on the question before time T |
| $V_7$ | QuestionScoreT+ | Aggregate vote on the question after time T |
| $V_8$ | QuestionCommentCount | No. of comments on the question |
| $V_9$ | QuestionCommentCountT- | No. of comments on the question before time T |
| $V_{10}$ | QuestionCommentCountT+ | No. of comments on the question after time T |
| $V_{11}$ | QuestionAnswerCount | No. of answers to the question |
| $V_{12}$ | QuestionAnswerCountT- | No. of answers to the question before time T |
| $V_{13}$ | QuestionAnswerCountT+ | No. of answers to the question after time T |
| $V_{14}$ | AnswerDayOfWeek | The day of answer creation |
| $V_{15}$ | AnswerTimeOfDay | The time of answer creation |
| $V_{16}$ | AnswerEpoch | Time gap between between the 1st post in site and the answer |
| $V_{17}$ | AnswerTimeliness | Time gap between the question and the answer |
| $V_{18}$ | AnswerOrder | Chronological order of the answer |
| $V_{19}$ | AnswerScore | Aggregate vote on the answer |
| $V_{20}$ | AnswerScoreT- | Aggregate vote on the answer before time T |
| $V_{21}$ | AnswerScoreT+ | Aggregate vote on the answer after time T |
| $V_{22}$ | AnswerPosition | Position of the answer based on the aggregate vote |
| $V_{23}$ | AnswerPositionT- | Position of the answer before time T |
| $V_{24}$ | AnswerPositionT+ | Position of the answer after time T |
| $V_{25}$ | AnswerCommentCount | No. of comments on the answer |
| $V_{26}$ | AnswerCommentCountT- | No. of comments on the answer before time T |
| $V_{27}$ | AnswerCommentCountT+ | No. of comments on the answer after time T |
| $V_{28}$ | AnswererPostCount | No. of posts (questions and answers) written by answerer |
| $V_{29}$ | AnswererAnswerCount | No. of answers written by answerer |
| $V_{30}$ | AnswererActiveAge | Time gap between between answerer's 1st post and the answer |
| $V_{31}$ | AnswererReputation | Total score of questions and answers written by answerer |
| $V_{32}$ | AnswererReputationViaAnswer | Total score of answers written by answerer |
| $V_{33}$ | AnswererGoldCount | No. of gold badges acquired by answerer |
| $V_{34}$ | AnswererSilverCount | No. of silver badges acquired by answerer |
| $V_{35}$ | AnswererBronzeCount | No. of bronze badges acquired by answerer |
| $V_{36}$ | AnswererBadgeDistribution | [GoldCount, SilverCount, BronzeCount] |
| $V_{37}$ | AnsweredQuestionViewTotal | No. of users who viewed past questions answered by answerer |
| $V_{38}$ | AnsweredQuestionFavoriteTotal | No. of users who favorited past questions answered by answerer |
| $V_{39}$ | AnsweredQuestionScoreTotal | Total score of past questions answered by answerer |
| $V_{40}$ | AnsweredQuestionCommentTotal | No. of comments on past questions answered by answerer |
| $V_{41}$ | AnsweredQuestionAnswerTotal | No. of answers to past questions answered by answerer |

the quality of the content (from the perspective of voters) and the ability of users (to generate high-quality content). These confounders (e.g., the ability of users) may affect both the impression signals (e.g., the reputation of the contributing user) and the observed votes. Ergo, we need to eliminate the effects of these confounders for estimating the causal effect.

The instrumental variable (IV) approach has been successfully used in the social sciences [100, 101, 102] to estimate causal effects (e.g., how education affects earning [100], how campaign spending affects senate selection [101], and how income inequality affects corruption [102]) from observational data. The IV method is especially useful for estimating effects in the presence of hidden confounders [103, 104]. The technique requires identifying candidate instruments that are correlated with the independent variable of interest. It then relies on careful argumentation (thought experiments) to eliminate the candidate instruments that may affect the hidden confounders. This process implies that the remaining instruments co-vary only with the independent variable, and cannot influence the dependent variable through a hidden confounder. As such, instrumental variables allow us to estimate causal effects, even in the presence of hidden confounders.

Prior research on voter biases regress aggregate vote on impression signals using ordinary least squares (OLS) and interpret the regression coefficients as effects. However, OLS only captures the correlation among variables; the resultant estimates are *non-causal*. For instance, a positive OLS estimate corresponding to an impression signal does not imply that the signal has a positive effect on the aggregate vote; the effect could be zero or even negative. This argument is especially applicable in the presence of hidden confounders. In fact, in such a case, the OLS estimate is biased [103].

Table 4.4: The parallels between voter bias quantification and instrumental variable method.

| **IV Terminology** | **Bias Terminology** | **Example** |
|---|---|---|
| Outcome | Aggregate Feedback | An aggregation (say sum or mean) of votes on content |
| Exposure | Impression Signal | Reputation of the contributing user in the form of scores and badges |
| Confounder | Unobserved Quality | What a voter assesses the quality of the content to be |
| Regression Coefficient | Voter Bias | How the reputation of the contributing user affects the aggregate vote |

The key conceptual difference between the IV and OLS is: IV relies on argumentation to reason about the underlying causal structure. If all we have access to is observational data,

then careful argumentation is necessary to establish the causal structure. As pointed out by Judea Pearl, *"behind every causal conclusion there must lie some causal assumption that is not testable in observational studies"* [127]. As we can not conduct randomized control trials on the actual platforms, and only have access to the observational data, IV is a reasonable approach for estimating causal effects. Further, our problem aligns well with the use case of IV: estimating causal effect in the presence of hidden confounders (In Table 4.4, we show the parallels between our problem and IV). For these reasons, we adopt the IV method to quantify voter biases.

### 4.5.2 Instrumental Variable Estimation

To motivate the use of IVs, we now explain a classic well-understood example: the causal effect of education on earnings [100]. In general, education enables individuals to earn more money, say through employment that is reserved for college graduates. One can estimate the return to education by simply regressing the earnings of individuals on their education level. However, this simplistic approach has a major limitation in the form of omitted variable—*the unobserved ability of individuals*. Unobserved ability (*confounder*) might be correlated with the level of education that an individual attains (*exposure*), and the wage he/she receives (*outcome*). Specifically, higher intellectual ability increases the probability of graduating from college, and individuals with more ability also tend to earn higher wages. This complication is popularly known as the "ability bias" [100]. The ability bias suggests that standard regression (OLS) coefficient would be a biased estimate of the causal effect of education on earnings.

Over the past decades, researchers have attempted to solve the problem of "ability bias" in a number of ways. Notably, a number of studies controlled for the effect of ability bias directly by including measures of ability such as IQ and other test scores within the regression model [128]. However, there are concerns over whether these types of variables are a good proxy for wage-earning ability. An alternative strategy which has been the focus of much of the literature is to identify one or more variables which affect education but do not affect earnings either directly or indirectly through some other aspect. If such variables can be found, they can be used as *instrumental variables* to derive a consistent estimate of the return to education. A large body of literature has been devoted to identifying proper instruments for estimating the causal effect of education on earnings. Some notable instruments include—differences in education owing to the—proximity to college, quarter-of-birth, and state variation when children have to commence compulsory schooling. A consistent finding across IV studies is that the estimated return to education is 20-40%

above the corresponding OLS estimate [100]. These IV studies motivate the question: *could we use IV for quantifying voter bias in Stack Exchange?*

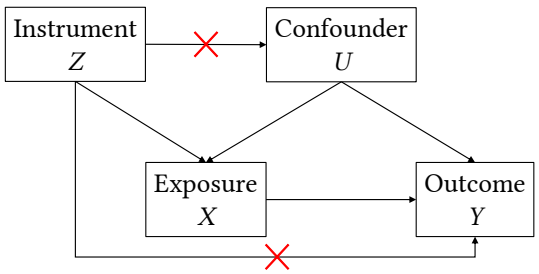



Figure 4.2: General structure of an instrumental variable model. The paths from $U$ to $X$, and $U$ to $Y$ introduces confounding in estimating the causal effect of $X$ on $Y$. For a valid instrument $Z$, the pathways from $Z$ to $X$, and $X$ to $Y$ must exist; whereas the pathways from $Z$ to $U$, and $Z$ to $Y$ must cease to exist.

Figure 4.2 depicts the general structure of an IV model. Designing an IV model requires identifying a valid *instrument* $Z$—a variable to eliminate the effects of confounders—that must satisfy the following conditions [104]:

1. *Relevance Condition:* The instrument $Z$ is correlated with the exposure $X$. For example, while estimating the causal effect of education on earnings, proximity to college ($Z$) is correlated with college education ($X$).

2. *Exclusion Restriction:* The instrument $Z$ does not affect the outcome $Y$ directly, except through its potential effect on the exposure $X$. This independence can be conditional upon other covariates. For example, proximity to college ($Z$) should not affect earnings ($Y$), except through its effect on college education ($X$). One can argue that—for people who work at college but are not college graduate themselves—the independence of proximity to college from earnings depends on the job.

3. *Marginal Exchangeability:* The instrument $Z$ and the outcome $Y$ do not share causes. For example, no common factor influences both proximity to college ($Z$) and earnings ($Y$).

Of the three instrumental conditions mentioned above, only the relevance condition is empirically verifiable [104]. Therefore, in an observational study such as ours, we can not test if a proposed instrument is a valid instrument. The best we can do is to use our subject matter knowledge to build a case for why a proposed instrument may be reasonably assumed to meet the exclusion restriction and marginal exchangeability.

In IV literature, if the correlation between the instrument $Z$ and the exposure $X$ is strong, then $Z$ is called a *strong instrument*; otherwise, it is called a *weak instrument*. A weak instrument has three major limitations. First, a weak instrument yields parameter estimates with a wide confidence interval. Second, any inconsistency from a small violation of the exclusion restriction gets magnified by the weak instrument. Third, a weak instrument may introduce bias in the estimation process and provide misleading inferences about parameter estimates and standard errors. In this study, we seek a strong instrument for quantifying each of the three voter biases.

In the remaining subsections, we develop IV models for reputation bias, social influence bias, and position bias. For each voter bias, we operationalize the IV components (outcome, exposure, instrument, and control) using our compiled variables (Table 1).

### 4.5.3 IV Model for Reputation Bias

We develop an IV model for quantifying reputation bias in Stack Exchange sites. We estimate the causal effect of the reputation of the user who contributes an answer (*exposure*) on the aggregate of votes on that answer (*outcome*). To this end, we operationalize the four IV components (outcome, exposure, instrument, and control) as follows.

**Outcome.** Our outcome of interest is the aggregate vote on the answer. We represent this outcome via variable `AnswerScore` $\langle V_{19} \rangle$ [1].

**Exposure.** Our exposure of interest is the reputation of the answerer. To represent this exposure, we compute several reputation measures for the answerer, based on the reputation and badge system in Stack Exchange. In Stack Exchange sites, the primary means to gain reputation and badges is to post good questions and useful answers. We compute the reputation measures for each answerer, *per answer*, based on the answerer's achievements prior to creating the current answer. Our reputation measures are as follows: `AnswererReputation` $\langle V_{31} \rangle$, `AnswererReputationViaAnswer` $\langle V_{32} \rangle$, `AnswererGoldCount` $\langle V_{33} \rangle$, `AnswererSilverCount` $\langle V_{34} \rangle$, and `AnswererBronzeCount` $\langle V_{35} \rangle$.

Note that, for a given answer, different voters may observe different reputation score and badges for the answerer, depending on their time of voting. The voters who participate later typically observe higher reputation score and badges, as the answerer may acquire more upvotes on other answers. Our dataset does not provide the exact state of reputation score and badges of the answerer for a particular vote. To get around this problem, we assume that all voters observe the same state of reputation: the reputation score and badges

[1] We shall use this syntax consistently throughout this study. The first term is variable name and the second term is variable id in Table 4.3. Please see Table 4.3 for the description of variables.

acquired by the answerer before creating the current answer. In general, reputation increases monotonically; therefore, our assumption is conservative.

Notice that, both our outcome (aggregate votes on the answer) and exposure (reputation of the answerer) of interest can be influenced by the *unobserved ability of the answerer*. Specifically, an answerer with high-ability is expected to generate high-quality answers that would receive many upvotes, increasing his/her reputation. The unobserved ability of the answerer and associated unobserved quality of answers prevent us from distilling the effect of the answerer's reputation on observed votes. We need instruments to eliminate the confounding effect of the answerer's ability.

**Instrument.** Now, how can we find instruments to uncover the effect of an impression signal (exposure) on the aggregate vote (outcome)? In the social science literature that employs IV's [101, 102], researchers use domain knowledge to identify variables that are likely to influence the exposure and thus satisfy the *relevance condition* (these are candidate instruments). Then for each candidate instrument, they use argumentation to determine if it meets the remaining IV conditions—exclusion restriction and marginal exchangeability.

Motivated by the social science approach to IV, we seek candidate instruments that contribute to an answerer's reputation. Based on our literature review, we identify two such factors: 1) answerer's activity level (number of posts, especially answers contributed by the answerer) [105], and 2) popularity of the answered questions (number of views, comments, and answers attracted by the questions) [129]. Note that an answerer's reputation increases with the volume of his/her activities. Also, a popular question allows contributing answerers to obtain more reputation by attracting more views (voters). To capture these two factors, we compute several measures for each answerer, *per answer*—namely, `AnswererPostCount` $\langle V_{28} \rangle$, `AnswererAnswerCount` $\langle V_{29} \rangle$, `AnswererActiveAge` $\langle V_{30} \rangle$, `AnsweredQuestionViewTotal` $\langle V_{37} \rangle$, `AnsweredQuestionFavoriteTotal` $\langle V_{38} \rangle$, `AnsweredQuestionScoreTotal` $\langle V_{39} \rangle$, `AnsweredQuestionCommentTotal` $\langle V_{40} \rangle$, and `AnsweredQuestionAnswerTotal` $\langle V_{41} \rangle$. We use these variables as are our candidate instruments.

We now scrutinize the candidate instruments to reason about their ability to meet the three instrumental conditions described in Section 4.5.2. Note that, all three conditions must be met for a candidate instrument to be valid. We divide the candidate instruments into two groups for qualitative reasoning: A) answerer's activity level [`AnswererPostCount` $\langle V_{28} \rangle$, `AnswererAnswerCount` $\langle V_{29} \rangle$, `AnswererActiveAge` $\langle V_{30} \rangle$]; and B) popularity of past questions responded to by the answerer [`AnsweredQuestionViewTotal` $\langle V_{37} \rangle$, `AnsweredQuestionFavoriteTotal` $\langle V_{38} \rangle$, `AnsweredQuestionScoreTotal` $\langle V_{39} \rangle$, `AnsweredQuestionCommentTotal` $\langle V_{40} \rangle$, `AnsweredQuestionAnswerTotal` $\langle V_{41} \rangle$]. Both groups of candidate instruments empirically satisfy the relevance condition. Therefore, we concentrate on the remaining two IV

conditions: exclusion restriction, and marginal exchangeability. In other words, we aim to identify instruments that affect the obtained reputation of the answerer (*exposure*) without affecting the votes on current answer (*outcome*), either directly or through the ability of the answerer (*confounder*).

Notice that the first group of candidate instruments—based on the answerer's activity level—may contribute to the ability of the answerer (*confounder*), which in turn may affect the quality of the answer and resultant votes on the answer (*outcome*). For example, a user who posted many answers may learn from experience to provide better quality answers in the future. Thus, the first group of candidate instruments may violate marginal exchangeability. In contrast, the second group of candidate instruments—based on the popularity of past questions responded to by the answerer—may affect the votes on the current answer (*outcome*) only through the answerer's reputation (*exposure*). These candidate instruments do not inform us about the ability of answerer (*confounder*). The second group of candidate instruments satisfies both exclusion restriction and marginal exchangeability. Ergo, we use the second group of instruments to estimate the effects of reputation signals on observed votes.

Based on the IV components mentioned above—exposure (reputation of the answerer), outcome (votes on the answer), confounder (the ability of the answerer to create high-quality answers), and instrument (popularity of the past questions)—we present the causal diagram of our model in Figure 4.3. Please note that our causal diagram follows the general structure of the instrumental variable framework (in Figure 2).

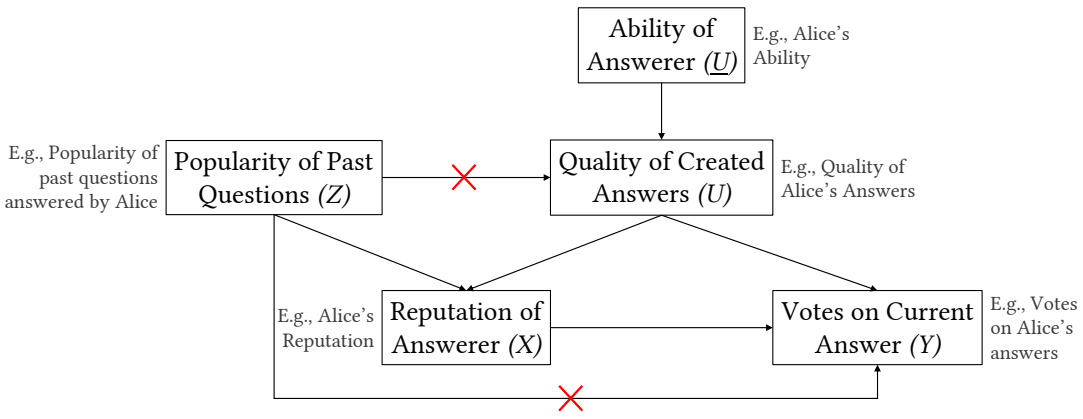



Figure 4.3: Causal diagram of our IV model for quantifying reputation bias. Here, the unobserved ability of answerer introduces confounding via the unobserved quality of created answers. To eliminate this confounding, we propose the popularity of the past questions responded to by the answerer as the instrument.

**Control.** While our claimed instruments (based on the popularity of past questions responded to by the answerer) are unlikely to affect the outcome (votes on current answer), we take further precautions in the form of controls, to establish the conditional independence of proposed instruments from the outcome. To this end, we propose the following controls in our IV specifiction: `Site` $\langle V_1 \rangle$, `QuestionViewCount` $\langle V_3 \rangle$, `QuestionFavoriteCount` $\langle V_4 \rangle$, `QuestionScore` $\langle V_5 \rangle$, `QuestionCommentCount` $\langle V_8 \rangle$, and `QuestionAnswerCount` $\langle V_{11} \rangle$.

Each Stack Exchange site accommodates a distinct audience, who may exhibit a distinct voting behavior; ergo, we control for `Site` $\langle V_1 \rangle$ via stratification. The remaining controls capture the popularity of current question, which establish the conditional independence of proposed instruments from the outcome. Specifically, given the popularity of current question, the popularity of past questions responded to by the answerer should not affect the votes on current answer. We incorporate these control variables into our model as regressors. For the outcome (`AnswerScore` $\langle V_{19} \rangle$) and exposure of interest (e.g., `AnswererReputationViaAnswer` $\langle V_{32} \rangle$), we can select one or more instrumental variables (say `AnsweredQuestionViewTotal` $\langle V_{37} \rangle$), and appropriate controls (`Site` and `QuestionViewCount` $\langle V_3 \rangle$) to estimate the causal effect of the exposure on the outcome.

#### 4.5.4 Joint IV Model for Social Influence Bias and Position Bias

In Stack Exchange sites, the default presentation order of answers is the aggregate vote thus far. This ordering scheme imposes a critical challenge in isolating the effect of position bias from the social influence bias, as the two biases vary together. To address this challenge, we develop a joint IV model to quantify social influence bias and position bias in the same model. We estimate the causal effects of initial votes and resultant position on subsequent votes by specifying the IV components as follows.

**Outcome.** Our outcome of interest is the aggregate vote on the answer after an initial *bias formation period*—the time required for social influence signal (initial votes) and position signal (answer position) to come into effect. We represent this outcome via `AnswerScoreT+` $\langle V_{21} \rangle$: a response variable that captures the aggregate vote on the answer based on the votes after time $T$, where $T$ is the limiting time of bias formation specific to the question.

**Exposure.** We have two exposures of interest corresponding to the initial votes and resultant position of the answer. To represent these exposures, we compute the aggregate vote and resultant position of answer at the limiting time of bias formation $T$. Our exposures are as follows: 1. `AnswerScoreT-` $\langle V_{20} \rangle$ captures the aggregate vote on answer based on the votes before time $T$; 2. `AnswerPositionT-` $\langle V_{23} \rangle$ captures the position of answer based on the aggregate vote before time $T$.

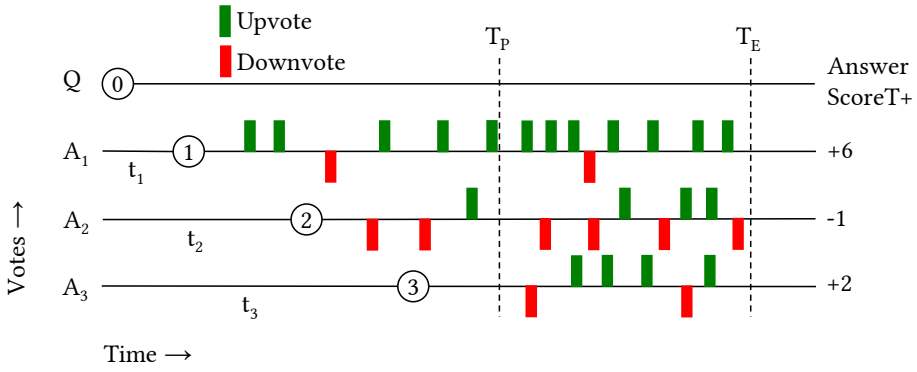



Figure 4.4: An illustration of the bias formation period to quantify our outcome (`AnswerScoreT+` $\langle V_{21} \rangle$) and exposures (`AnswerScoreT-` $\langle V_{20} \rangle$ and `AnswerPositionT-` $\langle V_{23} \rangle$). The creation of question $Q$ marks the beginning of our observation period. Then, three answers $A_1$, $A_2$, and $A_3$ that refer to $Q$ arrive after time $t_1$, $t_2$, and $t_3$ respectively. Finally, $T_E$ marks the end of our observation period (the time of data collection). Notice that, a total of 30 votes (20 upvotes, 10 downvotes) are casted on $A_1$, $A_2$, and $A_3$ by time $T_E$. We consider the time by which $P\%$ of total votes are casted on on $A_1$, $A_2$, and $A_3$ as the bias formation period; $T_P$ marks the limiting time of this bias formation period. In this example, the value of $P$ is 30.

We define a bias formation period to quantify our outcome and exposures. We define this period based on the dynamics of votes on the answers to each question. Specifically, we define the bias formation period of a question as the time by which $P\%$ of total votes on its answers are cast. Figure 4.4 shows an illustration of bias formation period, and how we use this period to quantify our outcome and exposures. The creation of question $Q$ marks the beginning of our observation period. Then, three answers $A_1$, $A_2$, and $A_3$ that refer to $Q$ arrive after time $t_1$, $t_2$, and $t_3$ respectively. Finally, $T_E$ marks the end of our observation period (the time of data collection). Notice that, a total of 30 votes (20 upvotes, 10 downvotes) are casted on $A_1$, $A_2$, and $A_3$ by time $T_E$. We consider the time by which $P\%$ of total votes are cast on $A_1$, $A_2$, and $A_3$ as the bias formation period; $T_P$ marks the limiting time of this bias formation period. In this example, the value of $P$ is 30 (in our experiments, we use different values of $P$ ranging from 5 to 30). The aggregate vote on answer before time $T_P$ is quantified as `AnswerScoreT-` $\langle V_{20} \rangle$, and the resultant position as `AnswerPositionT-` $\langle V_{23} \rangle$. The values of `AnswerScoreT-` $\langle V_{20} \rangle$ for answer $A_1$, $A_2$, $A_3$ in Figure 4.4 are +4, -1, 0 respectively. The resultant values of `AnswerPositionT-` $\langle V_{23} \rangle$ for $A_1$, $A_2$, $A_3$ are 1, 3, 2 respectively. The aggregate vote on answer from Time $T_P$ to time $T_E$ is quantified as `AnswerScoreT+` $\langle V_{21} \rangle$. The values of `AnswerScoreT+` $\langle V_{21} \rangle$ for $A_1$, $A_2$, $A_3$ are +6, -1, +2 respectively.

Notice that, both our exposures and outcome of interest can be influenced by the *unobserved quality of the answer*. We seek instruments to eliminate the confounding effect of answer quality.

**Instrument.** We seek candidate instruments that can uncover the effects of initial votes and position on subsequent votes. Same as before, we identify factors that contribute to the initial votes and position, thereby likely to satisfy the *relevance condition*. For the time being, we do not focus on the remaining IV conditions, exclusion restriction and marginal exchangeability. Prior work on voting behavior in Stack Exchange suggest several factors that contribute to initial votes, notably, activities on the question (number of views, comments, and answers attracted by the question) [129], time of answer (day of the week, hour of the day) [85], and timeliness of answer (time gap between question and answer) [96]. To capture these factors, we compute several measures—namely, `QuestionScoreT-` $\langle V_6 \rangle$, `QuestionCommentCountT-` $\langle V_9 \rangle$, `QuestionAnswerCountT-` $\langle V_{12} \rangle$, `AnswerDayOfWeek` $\langle V_{14} \rangle$, `AnswerTimeOfDay` $\langle V_{15} \rangle$, `AnswerEpoch`, `AnswerTimeliness` $\langle V_{17} \rangle$, and `AnswerOrder` $\langle V_{18} \rangle$. These variables are our candidate instruments.

We now scrutinize the candidate instruments to reason about their ability to meet the three instrumental conditions described in Section 4.5.2. Recall that, all three conditions must be met for a candidate instrument to be valid. We divide the candidate instruments into three groups for qualitative reasoning: A) activities on the question within the bias formation period [`QuestionScoreT-` $\langle V_6 \rangle$, `QuestionCommentCountT-` $\langle V_9 \rangle$, `QuestionAnswerCountT-` $\langle V_{12} \rangle$]; B) actual time of answer [`AnswerDayOfWeek` $\langle V_{14} \rangle$, `AnswerTimeOfDay` $\langle V_{15} \rangle$, `AnswerEpoch` $\langle V_{16} \rangle$]; and C) relative timeliness of answer [`AnswerTimeliness` $\langle V_{17} \rangle$, `AnswerOrder` $\langle V_{18} \rangle$]. All three groups of candidate instruments satisfy the relevance condition. The activities on a question within the bias formation period positively influence the votes on its answers within that period. The actual time of answer creation affects the initial votes due to the varying amount of voter activity across time.The timeliness of an answer affects its initial votes due to the amount time available for voting. Therefore, we concentrate on the remaining two IV conditions: exclusion restriction, and marginal exchangeability. In other words, we aim to identify the instruments that affect the initial votes or position (*exposure*) without affecting the subsequent votes (*outcome*), either directly or through the quality of the answer (*confounder*).

Notice that the first group of candidate instruments—based on the activities on the question within the bias formation period—may be influenced by the popularity of question (*confounder*), which in turn may contribute to both initial votes (*exposure*) and subsequent votes on the answer (*outcome*). For example, a popular question may induce a high amount of activity both within and beyond the bias formation period. The popularity of the question

may also explain the initial and subsequent votes on the answer. Thus, the first group of candidate instruments may violate marginal exchangeability. In contrast, the second group of candidate instruments—based on the actual time of answer—may directly influence both initial votes (*exposure*) and subsequent votes on the answer (*outcome*). Thus, the second group of candidate instruments may violate exclusion restriction. Finally, the third group of candidate instruments—based on the relative timeliness of answer—affect the subsequent votes primarily through the initial votes and position. For example, if Bob posts the 2nd answer to a particular question, then his initial votes within the bias formation period will be affected by the fact that he is the 2nd answerer. However, the subsequent votes after the bias formation period will not be affected by the same fact. Note that, the timeliness of an answer may be affected by the answerer's expertise. The answerer's expertise may also affect the outcome (subsequent votes on the answer) [130]. We address this issue by incorporating the answerer's expertise as a control variable in our IV model. Notice that, the third group of candidate instruments does not inform us about the quality of the answer (*confounder*) and help us avoid the primary confounder. These candidate instruments are reasonably assumed to satisfy both exclusion restriction and marginal exchangeability. Ergo, we use the third group of instruments to estimate the effects of initial votes and position on subsequent votes.

Based on the IV components mentioned above—exposure (initial votes and position of the answer), outcome (subsequent votes on the answer), confounder (quality of answer), and instrument (timeliness of answer)—we present the causal diagram of our model in Figure 4.5.

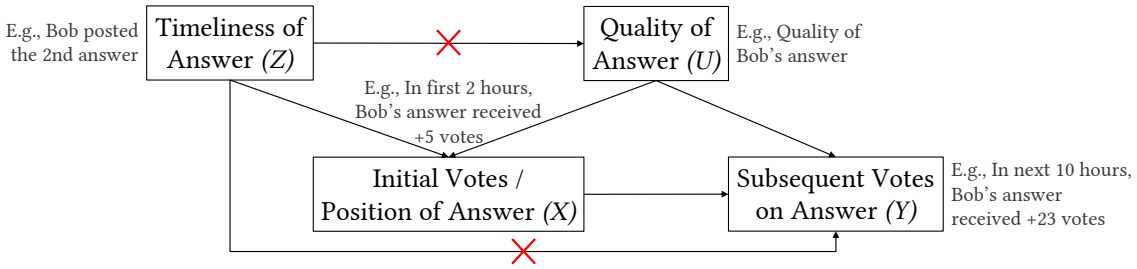



Figure 4.5: Causal diagram of our IV model for quantifying social influence bias and position bias. Here, the unobserved quality of answer act as a confounder. To eliminate this confounder, we propose the timeliness of answer as the instrument.

**Control.** While our claimed instruments (based on the relative timeliness of answer) are unlikely to affect the outcome (votes on answer after the bias formation period), we take further precautions in the form of controls, to establish the conditional independence of proposed instruments from the outcome. To this end, we propose the following controls in

our IV specifiction: `Site` $\langle V_1 \rangle$ and `AnswererReputationViaAnswer` $\langle V_{32} \rangle$.

In the joint IV model, we control for `Site` $\langle V_1 \rangle$ (via stratification) to account for the distinct audience in each Stack Exchange site. We also control for `AnswererReputationViaAnswer` $\langle V_{32} \rangle$ (via regression) as a proxy for the answerer's expertise. Recall that, our claimed instrument (the timeliness of an answer) may be affected by the answerer's expertise. The answerer's expertise may also affect our outcome (subsequent votes on the answer). While we acknowledge that `AnswererReputationViaAnswer` $\langle V_{32} \rangle$ is not a proxy for the answerer's expertise, it helps us to reduce the degree of bias in causal estimation.

In this section, we explain how to measure the effects of different impression signals on observed votes through the instrumental variable method. We identify instruments that co-vary only with the impression signals and do not influence the observed through a hidden confounder. These instruments allow us to estimate the causal effect of impression signals on votes.

## 4.6 RESULTS

In this section, we report the results of our study[2]. We begin by presenting the two-stage least squares (2SLS) method for implementing IV models. We then present our bias estimates for Stack Exchange sites—reputation bias (Section 4.6.1), social influence bias and position bias (Section 4.6.2).

**Two-Stage Least Squares (2SLS) Method.** Two-stage least squares (2SLS) is a popular method for computing IV estimates. The 2SLS method consists of two successive stages of linear regression. In the first stage, we regress each exposure variable on all instrumental and control variables in the model and obtain the predicted values from the regressions. In the second stage, we regress the outcome variable on the predicted exposures from the first stage, along with the control variables. The resultant regression coefficients corresponding to the predicted exposures in second stage yield the IV estimates. More details can be found in the supplementary material.

### 4.6.1 Quantifying Reputation Bias

We quantify reputation bias by estimating the causal effects of reputation score and badges on the aggregate vote. We have one outcome variable ($V_{19}$), five exposure variables ($V_{31}$, $V_{32}$, $V_{33}$, $V_{34}$, $V_{35}$), five instrumental variables ($V_{37}$, $V_{38}$, $V_{39}$, $V_{40}$, $V_{41}$), and six control

[2]The source code is available at `https://github.com/CrowdDynamicsLab/Quantifying_Voter_Biases`

variables ($V_1$, $V_3$, $V_4$, $V_5$, $V_8$, $V_{11}$). We use `Site` $\langle V_1 \rangle$ to stratify the data based on Stack Exhchange site. We incorporate the remaining variables into 2SLS regression framework to develop our IV models. We develop 10 IV models [5 (exposure) $\times$ 2 (with or without control)] that use all instruments, two for each exposure (with or without control). We develop another 50 IV models [5 (exposure) $\times$ 5 (instrument) $\times$ 2 (with or without control)] to analyze the performance of individual instrument. We also develop a baseline OLS model for each IV model. We perform log modulus transformation [$L(x) = sign(x) * log(|x| + 1>)$] of variables before using them in regression; this is required to linearize the relationship among variables. The use of log transformation in IV models is well-established [131].

We compare the performance of OLS and IV models by examining their estimates (regression coefficients). Table 4.5, 4.6, and 4.7 present the OLS and IV estimates for quantifying the causal effects of reputation score and badges on the aggregate vote, for English, Math, and Superuser respectively. We make the following observations from these estimates.

**Relevance Condition.** The final instruments for estimating the causal effects of reputation score and badges on the aggregate vote satisfy the *relevance condition* (stated in Section 4.5.2). For all IV estimates reported in Table 4.5–4.7, we observe low $p$-values and high $t$-statistics in the first stage of 2SLS. We do not report these numbers for brevity. Notice that the IV estimates in Table 4.5–4.7 have a small confidence interval, which is a byproduct of identifying *strong instruments*.

**Causal Effect of Reputation Score.** Prior research would interpret the regression coefficients from OLS in a causal way. In this study, we interpret the IV estimates as causal effects. For all three sites, the causal effect of reputation score on the aggregate vote is small. While OLS and IV provide similar estimates for quantifying the effect of reputation score, OLS assigns a slightly higher weight to the reputation score. Control variables rectify the estimates from both OLS and IV by increasing weights.

**Causal Effects of Badges.** For all three sites, the causal effects of badges on the aggregate vote is significant. The effects vary across the level of badges: high effect for gold badges, a moderate effect for silver badges, and low effect for bronze badges. This finding is consistent with the rarity of these badges. Stack Exchange sites grant a few gold badges, some silver badges, and lots of bronze badges to their users. OLS and IV differ a lot in quantifying the effects of badges. OLS tends to assign equal weights to all badges, whereas IV assigns more weight to gold badges (1.6–2.2`x` of OLS weights). In other words, *OLS underestimates the causal effect of gold badges significantly.* Control variables rectify the estimates from both OLS and IV by increasing weights.

Table 4.5: Causal effects (regression coefficients) of answerer's reputation score and badges on the aggregate vote in ENGLISH. All results presented in this table are statistically significant—validated via two-tailed t-tests—with $p < 0.001$. The results suggest that OLS and IV provide similar estimates for reputation score, whereas they differ a lot in estimating the effects of badges. Notably, *OLS tends to assign equal weights to all badges, whereas IV assigns more weights to gold badges.*

| | Instrument and Control | $Y$ = AnswerScore $\langle V_{19} \rangle$ | | | |
|---|---|---|---|---|---|
| | (*for estimating the effect of Exposure*) | $X$ = AnswererReputation $\langle V_{31} \rangle$ | | $X$ = AnswererReputationViaAnswer $\langle V_{32} \rangle$ | |
| Site | $Z + C$ | OLS | IV | OLS | IV |
| English | AnsweredQuestionViewTotal $\langle V_{37} \rangle$ | 0.092 (± 0.001) | 0.089 (± 0.001) | 0.090 (± 0.002) | 0.088 (± 0.002) |
| | $V_{37}$ + QuestionViewCount $\langle V_3 \rangle$ | 0.101 (± 0.002) | 0.098 (± 0.001) | 0.099 (± 0.002) | 0.097 (± 0.002) |
| | AnsweredQuestionFavoriteTotal $\langle V_{38} \rangle$ | 0.092 (± 0.001) | 0.088 (± 0.002) | 0.090 (± 0.002) | 0.086 (± 0.001) |
| | $V_{38}$ + QuestionFavoriteCount $\langle V_4 \rangle$ | 0.101 (± 0.002) | 0.093 (± 0.001) | 0.099 (± 0.001) | 0.092 (± 0.002) |
| | AnsweredQuestionScoreTotal $\langle V_{39} \rangle$ | 0.092 (± 0.001) | 0.086 (± 0.002) | 0.090 (± 0.002) | 0.084 (± 0.001) |
| | $V_{39}$ + QusestionScore $\langle V_5 \rangle$ | 0.100 (± 0.001) | 0.092 (± 0.001) | 0.099 (± 0.002) | 0.090 (± 0.001) |
| | AnsweredQuestionCommentTotal $\langle V_{40} \rangle$ | 0.092 (± 0.001) | 0.070 (± 0.002) | 0.090 (± 0.002) | 0.068 (± 0.001) |
| | $V_{40}$ + QuestionCommentCount $\langle V_8 \rangle$ | 0.093 (± 0.001) | 0.070 (± 0.001) | 0.091 (± 0.002) | 0.069 (± 0.002) |
| | AnsweredQuestionAnswerTotal $\langle V_{41} \rangle$ | 0.092 (± 0.001) | 0.076 (± 0.001) | 0.090 (± 0.002) | 0.075 (± 0.002) |
| | $V_{41}$ + QuestionAnswerCount $\langle V_{11} \rangle$ | 0.100 (± 0.001) | 0.084 (± 0.001) | 0.098 (± 0.001) | 0.083 (± 0.002) |
| | $V_{37}$, $V_{38}$, $V_{39}$, $V_{40}$, $V_{41}$ | 0.092 (± 0.001) | 0.081 (± 0.001) | 0.090 (± 0.002) | 0.079 (± 0.002) |
| | $V_{37}$, $V_{38}$, $V_{39}$, $V_{40}$, $V_{41}$ + $V_3$, $V_4$, $V_5$, $V_8$, $V_{11}$ | 0.098 (± 0.002) | 0.087 (± 0.001) | 0.096 (± 0.001) | 0.085 (± 0.001) |

| | Instrument and Control | $Y$ = AnswerScore $\langle V_{19} \rangle$ | | | | | |
|---|---|---|---|---|---|---|---|
| | (*for estimating the effect of Exposure*) | $X$ = AnswererGoldCount $\langle V_{33} \rangle$ | | $X$ = AnswererSilverCount $\langle V_{34} \rangle$ | | $X$ = AnswererBronzeCount $\langle V_{35} \rangle$ | |
| Site | $Z + C$ | OLS | IV | OLS | IV | OLS | IV |
| English | AnsweredQuestionViewTotal $\langle V_{37} \rangle$ | 0.184 (± 0.006) | 0.712 (± 0.014) | 0.138 (± 0.003) | 0.225 (± 0.004) | 0.157 (± 0.003) | 0.178 (± 0.003) |
| | $V_{37}$ + QuestionViewCount $\langle V_3 \rangle$ | 0.219 (± 0.005) | 0.794 (± 0.014) | 0.158 (± 0.003) | 0.250 (± 0.004) | 0.183 (± 0.002) | 0.198 (± 0.003) |
| | AnsweredQuestionFavoriteTotal $\langle V_{38} \rangle$ | 0.184 (± 0.006) | 0.543 (± 0.009) | 0.138 (± 0.003) | 0.187 (± 0.003) | 0.157 (± 0.003) | 0.175 (± 0.003) |
| | $V_{38}$ + QuestionFavoriteCount $\langle V_4 \rangle$ | 0.206 (± 0.006) | 0.579 (± 0.010) | 0.153 (± 0.002) | 0.200 (± 0.003) | 0.176 (± 0.003) | 0.186 (± 0.003) |
| | AnsweredQuestionScoreTotal $\langle V_{39} \rangle$ | 0.184 (± 0.006) | 0.570 (± 0.010) | 0.138 (± 0.003) | 0.192 (± 0.003) | 0.157 (± 0.003) | 0.170 (± 0.003) |
| | $V_{39}$ + QusestionScore $\langle V_5 \rangle$ | 0.199 (± 0.005) | 0.613 (± 0.010) | 0.151 (± 0.003) | 0.207 (± 0.003) | 0.177 (± 0.003) | 0.183 (± 0.003) |
| | AnsweredQuestionCommentTotal $\langle V_{40} \rangle$ | 0.184 (± 0.006) | 0.447 (± 0.010) | 0.138 (± 0.003) | 0.153 (± 0.003) | 0.157 (± 0.003) | 0.135 (± 0.003) |
| | $V_{40}$ + QuestionCommentCount $\langle V_8 \rangle$ | 0.183 (± 0.006) | 0.448 (± 0.010) | 0.138 (± 0.003) | 0.154 (± 0.004) | 0.157 (± 0.002) | 0.136 (± 0.003) |
| | AnsweredQuestionAnswerTotal $\langle V_{41} \rangle$ | 0.184 (± 0.006) | 0.500 (± 0.011) | 0.138 (± 0.003) | 0.170 (± 0.003) | 0.157 (± 0.003) | 0.149 (± 0.003) |
| | $V_{41}$ + QuestionAnswerCount $\langle V_{11} \rangle$ | 0.201 (± 0.006) | 0.551 (± 0.010) | 0.150 (± 0.003) | 0.188 (± 0.004) | 0.173 (± 0.003) | 0.165 (± 0.003) |
| | $V_{37}$, $V_{38}$, $V_{39}$, $V_{40}$, $V_{41}$ | 0.184 (± 0.006) | 0.338 (± 0.009) | 0.138 (± 0.003) | 0.143 (± 0.003) | 0.157 (± 0.003) | 0.145 (± 0.003) |
| | $V_{37}$, $V_{38}$, $V_{39}$, $V_{40}$, $V_{41}$ + $V_3$, $V_4$, $V_5$, $V_8$, $V_{11}$ | 0.195 (± 0.005) | 0.382 (± 0.008) | 0.149 (± 0.003) | 0.157 (± 0.003) | 0.176 (± 0.003) | 0.167 (± 0.003) |

Table 4.6: Causal effects (regression coefficients) of answerer's reputation score and badges on the aggregate vote in MATH. All results presented in this table are statistically significant—validated via two-tailed t-tests—with $p < 0.001$. The results suggest that OLS and IV provide similar estimates for reputation score, whereas they differ a lot in estimating the effects of badges. Notably, *OLS tends to assign equal weights to all badges, whereas IV assigns more weights to gold badges.*

| | Instrument and Control | $Y$ = AnswerScore $\langle V_{19} \rangle$ | | | |
|---|---|---|---|---|---|
| | (*for estimating the effect of Exposure*) | $X$ = AnswererReputation $\langle V_{31} \rangle$ | | $X$ = AnswererReputationViaAnswer $\langle V_{32} \rangle$ | |
| Site | $Z + C$ | OLS | IV | OLS | IV |
| Math | AnsweredQuestionViewTotal $\langle V_{37} \rangle$ | 0.056 (± 0.001) | 0.055 (± 0.001) | 0.053 (± 0.001) | 0.051 (± 0.001) |
| | $V_{37}$ + QuestionViewCount $\langle V_3 \rangle$ | 0.067 (± 0.001) | 0.061 (± 0.001) | 0.063 (± 0.001) | 0.057 (± 0.001) |
| | AnsweredQuestionFavoriteTotal $\langle V_{38} \rangle$ | 0.056 (± 0.001) | 0.057 (± 0.001) | 0.053 (± 0.001) | 0.053 (± 0.001) |
| | $V_{38}$ + QuestionFavoriteCount $\langle V_4 \rangle$ | 0.061 (± 0.001) | 0.057 (± 0.001) | 0.058 (± 0.001) | 0.053 (± 0.001) |
| | AnsweredQuestionScoreTotal $\langle V_{39} \rangle$ | 0.056 (± 0.001) | 0.055 (± 0.001) | 0.053 (± 0.001) | 0.051 (± 0.001) |
| | $V_{39}$ + QusestionScore $\langle V_5 \rangle$ | 0.058 (± 0.001) | 0.053 (± 0.001) | 0.055 (± 0.001) | 0.049 (± 0.001) |
| | AnsweredQuestionCommentTotal $\langle V_{40} \rangle$ | 0.056 (± 0.001) | 0.040 (± 0.001) | 0.053 (± 0.001) | 0.037 (± 0.001) |
| | $V_{40}$ + QuestionCommentCount $\langle V_8 \rangle$ | 0.057 (± 0.001) | 0.041 (± 0.001) | 0.054 (± 0.001) | 0.038 (± 0.001) |
| | AnsweredQuestionAnswerTotal $\langle V_{41} \rangle$ | 0.056 (± 0.001) | 0.040 (± 0.001) | 0.053 (± 0.001) | 0.037 (± 0.001) |
| | $V_{41}$ + QuestionAnswerCount $\langle V_{11} \rangle$ | 0.060 (± 0.001) | 0.043 (± 0.001) | 0.057 (± 0.001) | 0.040 (± 0.001) |
| | $V_{37}$, $V_{38}$, $V_{39}$, $V_{40}$, $V_{41}$ | 0.056 (± 0.001) | 0.048 (± 0.001) | 0.053 (± 0.001) | 0.043 (± 0.001) |
| | $V_{37}$, $V_{38}$, $V_{39}$, $V_{40}$, $V_{41}$ + $V_3$, $V_4$, $V_5$, $V_8$, $V_{11}$ | 0.062 (± 0.001) | 0.055 (± 0.001) | 0.059 (± 0.001) | 0.050 (± 0.001) |

| | Instrument and Control | $Y$ = AnswerScore $\langle V_{19} \rangle$ | | | | | |
|---|---|---|---|---|---|---|---|
| | (*for estimating the effect of Exposure*) | $X$ = AnswererGoldCount $\langle V_{33} \rangle$ | | $X$ = AnswererSilverCount $\langle V_{34} \rangle$ | | $X$ = AnswererBronzeCount $\langle V_{35} \rangle$ | |
| Site | $Z + C$ | OLS | IV | OLS | IV | OLS | IV |
| Math | AnsweredQuestionViewTotal $\langle V_{37} \rangle$ | 0.086 (± 0.001) | 0.234 (± 0.002) | 0.076 (± 0.001) | 0.104 (± 0.001) | 0.090 (± 0.001) | 0.112 (± 0.001) |
| | $V_{37}$ + QuestionViewCount $\langle V_3 \rangle$ | 0.122 (± 0.002) | 0.262 (± 0.003) | 0.094 (± 0.001) | 0.116 (± 0.001) | 0.117 (± 0.001) | 0.125 (± 0.001) |
| | AnsweredQuestionFavoriteTotal $\langle V_{38} \rangle$ | 0.086 (± 0.001) | 0.217 (± 0.002) | 0.076 (± 0.001) | 0.099 (± 0.001) | 0.090 (± 0.001) | 0.115 (± 0.001) |
| | $V_{38}$ + QuestionFavoriteCount $\langle V_4 \rangle$ | 0.105 (± 0.002) | 0.218 (± 0.002) | 0.083 (± 0.001) | 0.099 (± 0.001) | 0.103 (± 0.001) | 0.115 (± 0.001) |
| | AnsweredQuestionScoreTotal $\langle V_{39} \rangle$ | 0.086 (± 0.001) | 0.214 (± 0.002) | 0.076 (± 0.001) | 0.098 (± 0.001) | 0.090 (± 0.001) | 0.112 (± 0.001) |
| | $V_{39}$ + QusestionScore $\langle V_5 \rangle$ | 0.100 (± 0.001) | 0.206 (± 0.002) | 0.078 (± 0.001) | 0.094 (± 0.001) | 0.098 (± 0.001) | 0.107 (± 0.001) |
| | AnsweredQuestionCommentTotal $\langle V_{40} \rangle$ | 0.086 (± 0.001) | 0.154 (± 0.002) | 0.076 (± 0.001) | 0.072 (± 0.001) | 0.090 (± 0.001) | 0.081 (± 0.001) |
| | $V_{40}$ + QuestionCommentCount $\langle V_8 \rangle$ | 0.089 (± 0.002) | 0.157 (± 0.002) | 0.077 (± 0.001) | 0.073 (± 0.001) | 0.092 (± 0.001) | 0.083 (± 0.001) |
| | AnsweredQuestionAnswerTotal $\langle V_{41} \rangle$ | 0.086 (± 0.001) | 0.153 (± 0.002) | 0.076 (± 0.001) | 0.072 (± 0.001) | 0.090 (± 0.001) | 0.081 (± 0.001) |
| | $V_{41}$ + QuestionAnswerCount $\langle V_{11} \rangle$ | 0.094 (± 0.001) | 0.165 (± 0.002) | 0.081 (± 0.001) | 0.077 (± 0.001) | 0.098 (± 0.001) | 0.087 (± 0.001) |
| | $V_{37}$, $V_{38}$, $V_{39}$, $V_{40}$, $V_{41}$ | 0.086 (± 0.001) | 0.133 (± 0.002) | 0.076 (± 0.001) | 0.079 (± 0.001) | 0.090 (± 0.001) | 0.092 (± 0.001) |
| | $V_{37}$, $V_{38}$, $V_{39}$, $V_{40}$, $V_{41}$ + $V_3$, $V_4$, $V_5$, $V_8$, $V_{11}$ | 0.113 (± 0.002) | 0.179 (± 0.002) | 0.085 (± 0.001) | 0.090 (± 0.001) | 0.108 (± 0.001) | 0.110 (± 0.001) |

Table 4.7: Causal effects (regression coefficients) of answerer's reputation score and badges on the aggregate vote in SUPERUSER. All results presented in this table are statistically significant—validated via two-tailed t-tests—with $p < 0.001$. The results suggest that OLS and IV provide similar estimates for reputation score, whereas they differ a lot in estimating the effects of badges. Notably, *OLS tends to assign equal weights to all badges, whereas IV assigns more weights to gold badges.*

| | Instrument and Control | $Y$ = AnswerScore $\langle V_{19} \rangle$ | | | |
|---|---|---|---|---|---|
| | (*for estimating the effect of Exposure*) | $X$ = AnswererReputation $\langle V_{31} \rangle$ | | $X$ = AnswererReputationViaAnswer $\langle V_{32} \rangle$ | |
| Site | $Z + C$ | OLS | IV | OLS | IV |
| Superuser | AnsweredQuestionViewTotal $\langle V_{37} \rangle$ | 0.054 (± 0.001) | 0.045 (± 0.001) | 0.052 (± 0.001) | 0.043 (± 0.001) |
| | $V_{37}$ + QuestionViewCount $\langle V_3 \rangle$ | 0.067 (± 0.001) | 0.062 (± 0.001) | 0.065 (± 0.001) | 0.060 (± 0.001) |
| | AnsweredQuestionFavoriteTotal $\langle V_{38} \rangle$ | 0.054 (± 0.001) | 0.054 (± 0.001) | 0.052 (± 0.001) | 0.052 (± 0.001) |
| | $V_{38}$ + QuestionFavoriteCount $\langle V_4 \rangle$ | 0.065 (± 0.001) | 0.062 (± 0.001) | 0.063 (± 0.001) | 0.060 (± 0.001) |
| | AnsweredQuestionScoreTotal $\langle V_{39} \rangle$ | 0.054 (± 0.001) | 0.052 (± 0.001) | 0.052 (± 0.001) | 0.050 (± 0.001) |
| | $V_{39}$ + QusestionScore $\langle V_5 \rangle$ | 0.065 (± 0.001) | 0.061 (± 0.001) | 0.064 (± 0.001) | 0.059 (± 0.001) |
| | AnsweredQuestionCommentTotal $\langle V_{40} \rangle$ | 0.054 (± 0.001) | 0.038 (± 0.001) | 0.052 (± 0.001) | 0.036 (± 0.001) |
| | $V_{40}$ + QuestionCommentCount $\langle V_8 \rangle$ | 0.054 (± 0.001) | 0.038 (± 0.001) | 0.052 (± 0.001) | 0.036 (± 0.001) |
| | AnsweredQuestionAnswerTotal $\langle V_{41} \rangle$ | 0.054 (± 0.001) | 0.045 (± 0.001) | 0.052 (± 0.001) | 0.044 (± 0.001) |
| | $V_{41}$ + QuestionAnswerCount $\langle V_{11} \rangle$ | 0.062 (± 0.001) | 0.053 (± 0.001) | 0.060 (± 0.001) | 0.052 (± 0.001) |
| | $V_{37}$, $V_{38}$, $V_{39}$, $V_{40}$, $V_{41}$ | 0.054 (± 0.001) | 0.048 (± 0.001) | 0.052 (± 0.001) | 0.046 (± 0.001) |
| | $V_{37}$, $V_{38}$, $V_{39}$, $V_{40}$, $V_{41}$ + $V_3$, $V_4$, $V_5$, $V_8$, $V_{11}$ | 0.063 (± 0.001) | 0.060 (± 0.001) | 0.062 (± 0.001) | 0.057 (± 0.001) |

| | Instrument and Control | $Y$ = AnswerScore $\langle V_{19} \rangle$ | | | | | |
|---|---|---|---|---|---|---|---|
| | (*for estimating the effect of Exposure*) | $X$ = AnswererGoldCount $\langle V_{33} \rangle$ | | $X$ = AnswererSilverCount $\langle V_{34} \rangle$ | | $X$ = AnswererBronzeCount $\langle V_{35} \rangle$ | |
| Site | $Z + C$ | OLS | IV | OLS | IV | OLS | IV |
| Superuser | AnsweredQuestionViewTotal $\langle V_{37} \rangle$ | 0.106 (± 0.004) | 0.414 (± 0.009) | 0.081 (± 0.002) | 0.139 (± 0.003) | 0.082 (± 0.002) | 0.097 (± 0.002) |
| | $V_{37}$ + QuestionViewCount $\langle V_3 \rangle$ | 0.175 (± 0.004) | 0.591 (± 0.009) | 0.116 (± 0.002) | 0.196 (± 0.003) | 0.123 (± 0.001) | 0.137 (± 0.002) |
| | AnsweredQuestionFavoriteTotal $\langle V_{38} \rangle$ | 0.106 (± 0.004) | 0.399 (± 0.007) | 0.081 (± 0.002) | 0.143 (± 0.002) | 0.082 (± 0.002) | 0.123 (± 0.002) |
| | $V_{38}$ + QuestionFavoriteCount $\langle V_4 \rangle$ | 0.147 (± 0.004) | 0.459 (± 0.006) | 0.103 (± 0.001) | 0.165 (± 0.002) | 0.110 (± 0.002) | 0.142 (± 0.002) |
| | AnsweredQuestionScoreTotal $\langle V_{39} \rangle$ | 0.106 (± 0.004) | 0.406 (± 0.007) | 0.081 (± 0.002) | 0.144 (± 0.005) | 0.082 (± 0.002) | 0.117 (± 0.002) |
| | $V_{39}$ + QusestionScore $\langle V_5 \rangle$ | 0.162 (± 0.003) | 0.481 (± 0.006) | 0.109 (± 0.001) | 0.170 (± 0.002) | 0.116 (± 0.002) | 0.139 (± 0.002) |
| | AnsweredQuestionCommentTotal $\langle V_{40} \rangle$ | 0.106 (± 0.004) | 0.266 (± 0.006) | 0.081 (± 0.002) | 0.099 (± 0.003) | 0.082 (± 0.002) | 0.082 (± 0.002) |
| | $V_{40}$ + QuestionCommentCount $\langle V_8 \rangle$ | 0.106 (± 0.004) | 0.266 (± 0.007) | 0.081 (± 0.002) | 0.099 (± 0.003) | 0.081 (± 0.001) | 0.082 (± 0.002) |
| | AnsweredQuestionAnswerTotal $\langle V_{41} \rangle$ | 0.106 (± 0.004) | 0.349 (± 0.007) | 0.081 (± 0.002) | 0.124 (± 0.002) | 0.082 (± 0.002) | 0.100 (± 0.002) |
| | $V_{41}$ + QuestionAnswerCount $\langle V_{11} \rangle$ | 0.144 (± 0.003) | 0.419 (± 0.007) | 0.102 (± 0.002) | 0.148 (± 0.002) | 0.105 (± 0.002) | 0.120 (± 0.002) |
| | $V_{37}$, $V_{38}$, $V_{39}$, $V_{40}$, $V_{41}$ | 0.106 (± 0.004) | 0.244 (± 0.006) | 0.081 (± 0.002) | 0.110 (± 0.002) | 0.082 (± 0.002) | 0.093 (± 0.002) |
| | $V_{37}$, $V_{38}$, $V_{39}$, $V_{40}$, $V_{41}$ + $V_3$, $V_4$, $V_5$, $V_8$, $V_{11}$ | 0.152 (± 0.003) | 0.337 (± 0.005) | 0.105 (± 0.002) | 0.141 (± 0.002) | 0.113 (± 0.002) | 0.131 (± 0.002) |

### 4.6.2 Quantifying Social Influence Bias and Position Bias

We quantify social influence bias and position bias by jointly estimating the causal effects of initial votes and position on the subsequent votes. We have one outcome variable ($V_{21}$), two exposure variables ($V_{20}$, $V_{23}$), two instrumental variables ($V_{17}$, $V_{18}$), and two control variables ($V_1$, $V_{32}$). We use `Site` $\langle V_1 \rangle$ to stratify the data based on Stack Exchange site. We incorporate the remaining variables into 2SLS regression framework to develop one comprehensive IV model. Note that, we need all instruments and controls to develop our IV model, as there are multiple exposure variables and confounders. For this reason, we can not study the effect of an individual instrument.

The measurement of variables in this model relies on the specification of the bias formation period, $T$. We define the bias formation period of a question as the time by which $P$% of total votes on its answers are cast. We vary the value of $P$ from 5 to 30, with an increment of 5, to create six different instances of this model. We also develop a baseline OLS instance for each IV instance.

Table 4.8: The causal effects (IV estimates) of initial votes and position on subsequent votes in English, Superuser and Math. All results presented in this table are statistically significant—validated via two-tailed t-tests—with $p < 0.001$. The results suggest that OLS and IV differ a lot in quantifying the effects of initial votes and position. Notably, *OLS underestimates reputation bias and overestimates social influence bias significantly.*

| | | $Y$ = AnswerScoreT+ $\langle V_{21} \rangle$, $Z_1$ = AnswerTimeliness $\langle V_{17} \rangle$, $Z_2$ = AnswerOrder $\langle V_{18} \rangle$ | | | |
|---|---|---|---|---|---|
| | | $X_1$ = AnswerScoreT- $\langle V_{20} \rangle$ | | $X_2$ = AnswerPositionT- $\langle V_{23} \rangle$ | |
| Site | $T$ | OLS | IV | OLS | IV |
| English | $T_{0.05}$ | 0.803 (± 0.007) | 0.442 (± 0.087) | 0.215 (± 0.014) | 0.401 (± 0.037) |
| | $T_{0.10}$ | 0.821 (± 0.006) | 0.403 (± 0.080) | 0.205 (± 0.012) | 0.337 (± 0.030) |
| | $T_{0.15}$ | 0.819 (± 0.005) | 0.385 (± 0.073) | 0.184 (± 0.010) | 0.300 (± 0.025) |
| | $T_{0.20}$ | 0.791 (± 0.005) | 0.354 (± 0.067) | 0.161 (± 0.009) | 0.270 (± 0.022) |
| | $T_{0.25}$ | 0.752 (± 0.004) | 0.323 (± 0.061) | 0.126 (± 0.008) | 0.230 (± 0.018) |
| | $T_{0.30}$ | 0.699 (± 0.004) | 0.289 (± 0.057) | 0.100 (± 0.008) | 0.204 (± 0.016) |
| Math | $T_{0.05}$ | 0.802 (± 0.003) | 0.359 (± 0.037) | 0.470 (± 0.007) | 0.483 (± 0.010) |
| | $T_{0.10}$ | 0.880 (± 0.003) | 0.355 (± 0.036) | 0.446 (± 0.005) | 0.445 (± 0.009) |
| | $T_{0.15}$ | 0.920 (± 0.003) | 0.352 (± 0.035) | 0.380 (± 0.005) | 0.399 (± 0.008) |
| | $T_{0.20}$ | 0.921 (± 0.003) | 0.342 (± 0.034) | 0.339 (± 0.004) | 0.373 (± 0.007) |
| | $T_{0.25}$ | 0.885 (± 0.002) | 0.331 (± 0.034) | 0.284 (± 0.004) | 0.343 (± 0.007) |
| | $T_{0.30}$ | 0.833 (± 0.002) | 0.324 (± 0.033) | 0.240 (± 0.003) | 0.319 (± 0.006) |
| Superuser | $T_{0.05}$ | 1.814 (± 0.010) | 0.800 (± 0.122) | 0.842 (± 0.025) | 1.209 (± 0.058) |
| | $T_{0.10}$ | 1.939 (± 0.008) | 0.742 (± 0.108) | 0.784 (± 0.021) | 1.018 (± 0.045) |
| | $T_{0.15}$ | 1.983 (± 0.007) | 0.689 (± 0.097) | 0.705 (± 0.017) | 0.899 (± 0.037) |
| | $T_{0.20}$ | 1.888 (± 0.005) | 0.633 (± 0.087) | 0.594 (± 0.014) | 0.793 (± 0.030) |
| | $T_{0.25}$ | 1.633 (± 0.004) | 0.583 (± 0.076) | 0.463 (± 0.012) | 0.712 (± 0.025) |
| | $T_{0.30}$ | 1.477 (± 0.003) | 0.526 (± 0.067) | 0.363 (± 0.009) | 0.630 (± 0.021) |

We compare the performance of OLS and IV models by examining their estimates (regression coefficients). Table 4.8 presents the OLS and IV estimates for quantifying the

causal effects of initial votes and position on the subsequent votes, for ENGLISH, MATH, and SUPERUSER. We make the following observations from these estimates.

**Relevance Condition.** The final instruments for estimating the causal effects of initial votes and position on the subsequent votes satisfy the *relevance condition*. For all IV estimates reported in Table 4.8, we observe low $p$-values and high $t$-statistics in the first stage of 2SLS. We do not report these numbers for brevity. Notice that the IV estimates in Table 4.8 have a small confidence interval, which is a byproduct of identifying *strong instruments*.

**Causal Effect of Initial Votes.** For all three sites, the causal effect of initial votes on subsequent votes is significant. OLS and IV differ a lot in quantifying the effect of initial votes. OLS assigns high weights to initial votes, 1.8–2.3x of IV weights (based on initial 5% votes). In other words, *OLS overestimates the causal effect of initial votes significantly*.

**Causal Effect of Initial Position.** For all three sites, the causal effect of initial position on subsequent votes is significant. OLS and IV differ a lot in quantifying the effect of initial position. IV assigns high weights to initial position, at times 1.9x of OLS weights (based on initial 5% votes). In other words, *OLS underestimates the causal effect of initial position significantly*.

**Effect of Bias Formation Period.** For all three sites, increasing the bias formation period $T$ leads to a decrease in causal effects for both initial votes and position. This finding implies that *the first few votes significantly skew the subsequent votes*.

In addition to the above-mentioned definition of bias formation period, we also define it based on the day of question creation. Specifically, we use the votes on answers during the day of question creation for computing `AnswerScoreT-` $\langle V_{20} \rangle$ and `AnswerPositionT-` $\langle V_{23} \rangle$. We use the votes on subsequent days for computing `AnswerScoreT+` $\langle V_{21} \rangle$. The results are available in the supplementary material.

## 4.7 DISCUSSION

In the presented work, we quantify the degree of voter biases in online platforms. To derive these bias estimates, we make a methodological contribution in the study: how to measure the effects of different impression signals on observed votes through a novel application of instrumental variables. Our findings have implications for studying online voting behavior,

making changes to the platforms' interface, changes to the policy, and broader research within the CSCW community.

### 4.7.1 Implications for Online Voting Behavior

Our work has provided some of the first *causal insights* into online voting behavior.

**How Community Type Affects Voting.** Our results show that the effects of impression signals on votes widely vary across Stack Exchange sites. For example, the effect of gold badges in English is twice as high as in Math. Again, the effect of content position in Superuser is twice as high as in Math. This finding implies that what impression signals voters pay attention to and what cognitive heuristics they use to transform the signals into up- and down- votes may vary based upon the community type. For instance, English, Superuser, and Math belong to different themes—culture, technology, and science—which cater to different subsets of participants. On the one hand, different themes induce a varying degree of content interpretation, e.g., content interpretation in English is perhaps more subjective compared to content interpretation in Math [80]. On the other hand, users who are interested in different themes may be driven by different factors to contribute [85]. Overall, the communities appropriate the platforms in different ways as they deal with different themes and define their own understanding of what is good content or what signals competent users. Our finding, coupled with the above-mentioned corollaries suggest that voter bias may vary as a function of community type. We follow up on the design implications of these insights in Section 4.7.3.

**On Social Prestige of Badges.** Our results show that different reputation signals have varying effects on votes. While both badges and reputation score are indicative of user reputation, badges exhibit higher influence on votes compared to reputation score. An interpretation of this finding is that badges are perhaps deemed more "prestigious" than reputation score by voters. Recent work by Merchant et al. [132] investigated the role of reputation score and badges in characterizing social qualities. By adopting a regression approach, they found that reputation score and badges positively correlate with popularity and impact. Our finding, in contrast, provides *causal evidence* in favor of the social prestige of badges [133], over reputation score. This evidence, coupled with growing concerns about user engagement in online platforms [134] suggest that badge systems may put newcomers at a significant disadvantage. Our results also reveal the relationship between the prestige of badges and their exclusivity. Gold badges are the

rarest among the three types of badges, and their effect is *two to three times* higher compared to that of silver and bronze badges.

### 4.7.2 Implications for User Interface Design

Our research reveals how impression signals in user interface affect the votes and lead to biases. These findings have the potential to inform interface design to avoid biases.

**Conceal Impression Signals.** Our results show that impression signals, such as prior votes and badges, heavily influence voting behavior. An interface design implication of this finding is to conceal these signals from voters. Online platforms may adopt different interface design techniques to conceal impression signals from voters. For example, impression signals can be moved from the immediate vicinity of content; these signals may appear in other places, e.g., badges may still appear in the profile pages of the contributing users. Alternatively, impression signals can be concealed from voters till vote casting; a voter may access the signals only after casting his/her vote. The concept of concealing impression signals has been explored in another context: Grosser et al. [135] prescribed removing impression metrics (e.g., number of followers, likes, retweets, etc.) from social media feed to prevent users from feeling compulsive, competitive, and anxious. Note that, while concealing impression signals may eliminate the influence of these signals on voters, it is hard to anticipate how voters will react in the absence of such signals. For instance, voters may then rely on other factors, such as the offline reputation of the contributing user, to make voting decisions. Further, the interface changes may also impact the contributing users, who may adopt new strategic behaviors to maintain their online reputation.

**Delay the Votes.** Recall that, to uncover the effects of prior votes and position on subsequent votes; we use the timeliness of answers as the instrument. The main motivation of our chosen instrument is that early-arriving answers get more time to acquire votes. A design implication of this finding is to prevent the early arriving answer(s) from accumulating higher initial votes. Platforms could withhold the provision of voting for a fixed amount of time to achieve this. The withholding period could be decided based on the historical time gap between the arrival time of questions and answers.

**Randomize Presentation Order.** Our results show that the position of content also exhibits a strong influence on voters. As the position of content cannot be concealed in a webpage, the design implication is to eliminate position bias via other means.

Platforms may randomize the order of answers for each voter and thus prevent any answer from gaining a position advantage (on average). Lerman et al. [89] studied the effects of different ranking policies on votes, including the randomized ordering policy. They found that random policy is best for unbiased estimates of preferences. However, since a small fraction of user-generated content is interesting, users will mainly see uninteresting content under the random policy.

### 4.7.3 Informing Policy Design

Our research could also inform policy design to mitigate biases.

**De-biasing Votes.** What can a platform operator do to mitigate voting biases? A natural remedy is to de-bias the feedback scores *post-hoc*. Our research provides a major step in this regard by providing accurate bias estimates using the IV approach. Apart from such a remedial approach, platform operators could also use a preventive approach, including adopting more evolved aggregation mechanisms to combine individual feedback from voters. Such complex aggregation already occurs on some websites. For example, Amazon no longer displays the voter average for each product but instead uses a proprietary Machine Learning algorithm to compute the aggregate ratings [136]. The aggregation policy for votes may account for potential biases, say by weighting the votes based on their arrival time (later votes are more susceptible to herding behavior), history of the voter (differentiating novice voters from the more experienced voters), and content type. While prior work has considered weighted voting—to identify the answer that received most of the votes when most of the answers were already posted [137]—the weighting mechanism for bias mitigation merits further investigation. It's especially important to understand the effects of weighted voting on participation bias, as different weighting mechanisms may attract different subsets of the voter population to participate. For instance, any weighted voting policy where all votes are not equal is likely to dissuade the disadvantaged voters from participation.

**Community Dependent Policy Design.** Our research revealed how community type could affect the degree of voter biases. A policy design implication of this finding is to design policies based on the type of community. Instead of using the same vote aggregation and content ranking function for all Stack Exchange sites, platform operators could use variants of the same function for different sites, accounting for the behavior of the underlying voter base. How variation in policy (across sites) may affect the users who participate in multiple communities is an interesting direction for future research.

### 4.7.4 Impact on CSCW Research

We show how to estimate the degree to which a factor bias votes through an application of instrumental variables (IV) method. We believe that IV is a valuable tool for use in CSCW research, in particular, for researchers studying biases and online behavior.

**IV for Studying Biases.** The presented research concentrates on quantifying voter biases in the light of impression signals. However, online platforms also accommodate other more serious forms of biases, such as race and gender biases [138, 139, 140]. Jay Hanlon—the vice president of community growth at Stack Overflow—acknowledged the presence of race and gender biases in Stack Exchange: "Too many people experience Stack Overflow as a hostile or elitist place, especially newer coders, women, people of color, and others in marginalized groups." [140]. Vasilescu et al. [138] revealed the gender representation in Drupal, WordPress, and StackOverflow: only 7-10% of the participants in these communities are women. Through semi-structured interviews and surveys, Ford et al. [139] identified some of the barriers for female participation in Stack Overflow, such as lack of awareness about site features and self-doubt about qualification. Estimating the causal effects of race and gender on the perceived community feedback could reveal the degree of race and gender biases in online platforms. We believe IV could be a valuable tool in this regard. The argumentation based underpinning of IV is well-suited for studying biases in observational setup; it prompts researchers to reason about the underlying causal process.

## 4.8 LIMITATIONS

The observational nature of our study imposes several constraints on our analysis, which requires us to make a number of assumptions. First, we assume that all voters observe the same state of reputation and badges for the answerer. In reality, voters arrive at different times, and the reputation score and badges of the answerer may change between the voter arrivals. Second, we assume that the voters who arrive after the bias formation period observe the same state of initial votes. However, due to the sequential nature of voting, the observed votes may change from one voter to the next. We also assume that the positions of answers do not change after the bias formation period. Third, we ignore the effects of external influence. For example, a voter may be influenced by Google search results or Twitter promotion to upvote an answer. Fourth, while the default presentation order of answers in Stack Exchange is to sort them by votes, we can not track the views that individuals used to make voting decisions. We assume that the default presentation order is the one that influences voter

judgment. Finally, we inherit the key limitation of the instrumental variables method, relying on two untestifiable assumptions: exclusion restriction and marginal exchangeability.

## 4.9 CONCLUSION

In content-based platforms, an aggregate of votes is commonly used as a proxy for content quality. However, empirical literature suggests that voters are susceptible to different biases. In this study, we quantify the degree of voter biases in online platforms. We concentrate on three distinct biases: reputation bias, social influence bias, and position bias. The key idea of our approach is to formulate voter bias quantification using the instrumental variable (IV) framework. The IV framework consists of four components: outcome, exposure, instrument, and control. Using large-scale log data from Stack Exchange sites, we operationalize the IV components by employing impression signals as exposure and aggregate feedback as outcome. Then, we estimate the causal effect of exposure on outcome by using a set of carefully chosen instruments and controls. The resultant estimates quantify the voter biases. Our empirical study shows that the bias estimates from our IV approach differ from the bias estimates from the ordinary least squares (OLS) approach. The implications of our work include: redesigning user interface to avoid voter biases; making changes to platforms' policy to mitigate voter biases; detecting other forms of biases in online platforms.

In this chapter, we showed that voters are biased. We found that voters use reputation as a proxy for content quality, which hurts new users. In the next chapter, we will investigate if voters had voted impartially, how it would have affected the retention of users.

# CHAPTER 5: EVALUATING VOTING NORMS

Our first study showed that platforms fail because newcomers are not as devoted to platforms as old users. Our second study showed that community voting is biased against newcomers. These findings raise the question: would the newcomers have stayed longer or contributed more if the community had voted impartially? This chapter discusses our third study on evaluating voting norms, specifically how they affect user retention.

## 5.1 INTRODUCTION

Creating content and receiving feedback on the content are at the core of today's social web experience. Feedback on content—which comes in the form of votes (e.g., Stack Exchange, Reddit), likes (e.g., Facebook, Twitter), etc.—captures how much a community values the content. More so, it captures how much the community appreciates the user who created the content. For example, in Stack Exchange and Reddit, votes on content translate into social rewards such as reputation [105], badges [30], and karma [106]. These vote-based social rewards incentivize users to create more content [141, 142].

Owing to the underlying social value, community feedback such as votes can alter user behavior. Indeed, prior research suggests that votes on content may significantly affect the future participation of users [143]. The degree to which votes affect the participation of users is important. Votes not only does it form the basis of most platform's policies for rewarding behavior but also because we know that many online platforms (e.g., Stack Exchange) are not welcoming to newcomers [144] and whose membership is falling [12]. Could it be that what is causing these newcomers to give up is *not* the lack of interest in the platform, but the "poor" social judgment of their peers? That is, had their peers voted differently by focusing on the objective qualities of content (e.g., length, readability, objectivity, polarity)—instead of using impression signals (e.g., the reputation of the content creator)—fewer individuals would have left the platform. In this study, we ask a counterfactual question: *what would have happened if the members of a community had voted based on some objective criteria?*

Much of the work thinking about understanding behavior on online social platforms has centered around the question of mechanism design, especially the role of badges [26, 30]. Another line of related work examines voter biases in online platforms [88, 89], suggesting that individuals are prone to using cognitive heuristics that influence their voting behavior and prevent independent judgments. Specifically, voters tend to use different impression signals adjacent to the content—such as the reputation of the contributing user, aggregate

vote thus far, and position of content—as an input to make quick decisions about the quality of content. Impression signals as shortcuts to make voting decisions result in biases, which makes aggregate votes an unreliable measure for content quality. Also, voter biases may adversely affect user retention, as they put certain users, especially newcomers, at a significant disadvantage.

**Present Work.** In this study, we develop a methodology for quantifying the effects of different counterfactual community voting norms on user retention. Our methodology provides a formalism to reason about the retention effects of never-experienced voting norms, e.g., how user retention in Stack Exchange would be different if the community issues votes based upon the length of content. By developing a methodology to answer such counterfactual questions, we identify voting norms that improve user retention.

To perform our counterfactual analysis, we must understand the probability of different outcomes under a given voting condition. To this end, we develop a model for quantifying the propensity of voting outcomes. Our propensity model explains the voting behavior of a community from a contextual perspective. In this context, voters examine all answers to a given question (the context) to evaluate and subsequently vote on the answers. We adopt a Dirichlet-multinomial model that captures the joint distribution of votes for all answers to a question. The posterior distribution from the model quantifies the propensity of voting outcomes.

To study the effect of votes on users, we need to transform the propensities of voting outcomes into propensities of user utility. To this end, we first aggregate the observed (empirical) voting results from the content level to compute utility at the user level. Then, using the propensity of votes, we compute the propensity of user utility.

Using the propensities computed above, we develop a counterfactual model to reason about alternative voting norms. The main idea of our approach is to interpret community voting norms as functions that determine the votes on content. The current (unaltered) community voting condition acts as the *control* norm, whose outcomes can be observed from the log data. We then adopt a counterfactual setup to reason about the outcomes of the alternative voting norms. We adopt an inverse propensity sampling (IPS) estimator to perform our counterfactual analysis. An IPS estimator has three components: context, action, and reward. Context, in our case, refers to the content created by a user; action refers to the allocation of votes on content, determined by the norm in consideration; reward refers to the retention of the user.

We conduct extensive experiments on Stack Exchange websites comparing the default voting norm in these sites with six alternative norms: random (i.e., content receive an arbitrary number of votes), uniform (i.e., all content receive the same number of votes),

length (i.e., content that contain more words receive more votes), readability (i.e., content that have higher readability receive more votes), objectivity (i.e., content that express facts rather than opinions receive more votes), and polarity (i.e., content that express positive emotion receive more votes). Our main findings are that had the community members voted based upon content-based criteria, such as length, readability, or objectivity or polarity, the platform operator would have observed *higher* retention.

This study's main design implication is that platform operators need to promote content based on factors that are intrinsic to the content (e.g., length). The payoff: a higher retention rate amongst the community members.

## 5.2 PROBLEM FORMULATION

In this section, we provide an informal problem description, introduce the terminology we use throughout the study, and discuss our problem statement in detail.

**Informal Description.** Consider an online social platform where users post content, and they receive up and down votes on the content from peers. Broadly, the users derive utility from the votes they receive on content. Some of these users would continue to participate in the platform at a given time, while others would depart. Our goal is to understand that if the peers as a community had voted differently, would many of these departed users have stayed.

**Terminology.** In this study, we use the following terms.

1. Voting Criteria: Voting criteria refers to the factors that an individual takes into account while casting votes on content, say the quality of content and the reputation of content creator.

2. Voting Norm: Voting norm refers to the normative voting criteria around which a community votes, say the members of a community may cast votes on content based upon the reputation of content creator.

3. Voting Outcome: Voting outcome refers to the distribution of votes over content (resulting from a voting norm), say the distribution of votes over the answers to a question in Stack Exchange.

**Problem Statement.** *The goal of this study is to develop a methodology for evaluating the effects of alternative voting norms on user retention.*

In content-based social platforms (such as Stack Exchange and Reddit), users receive up- and down- votes on content from fellow community members. An aggregate of the votes

is displayed alongside content, which allows the community to recognize the most helpful content. The votes also contribute to an online reputation system that measures how much the community values any content and the user who contributed the content. The vote-based reputation is the main form of social reward that incentivizes users to create content on these platforms.

However, prior research on online communities suggests that voters tend to use different impression signals adjacent to the content—such as the reputation of the content creator, aggregate vote thus far, and position of content—as their primary voting criteria, which leads to systematic biases. Indeed, impression signals' use as voting criteria puts certain users (say, newcomers) at a significant disadvantage. This evidence, coupled with growing concerns about user retention in content-based platforms, suggests that voting norm may impact users' retention in these platforms.

The goal of this study is to infer, in an *observational* setting, the outcomes of counterfactual scenarios where the voting norm is different from the one in use. Specifically, given the historical logs of user content creation and associated community voting, is it possible to investigate scenarios where the community adopts a voting norm different from the one observed in log data? More importantly, what happens when the community votes differently? In this study, we develop a methodology to reason about the consequences of counterfactual voting norms using log data, specifically in terms of user retention. Using our methodology, we examine several alternative voting norms and quantify their effects on user retention.

Without loss of generality, consider the following scenario in a Stack Exchange website. A user Alice has provided an answer to a particular question. The question, overall, has three answers, one each from Alice, Bob, and Carol. Now, imagine that the voting norm in Stack Exchange is to cast votes based upon the content creator's reputation, which causes Alice to receive fewer votes than Bob and Charles. How is this voting norm going to affect Alice's tenure on the platform? Further, how would Alice's tenure be different if the voting norm were different, say the readability of answers? In the rest of the study, we develop a methodology that allows answering such counterfactual questions from log data.

## 5.3 OPERATOR'S REWARD: USER RETENTION

User retention is perhaps the most important factor that determines the success or failure of any online social platform. With more and more platforms failing to maintain their user-base [12, 145, 146, 147], retention remains a major concern for the platform operators. In this study, we examine user retention in content-based social platforms such as Stack Exchange and Reddit. We specifically study retention from a content generation standpoint—whether

a user participates in content generation or not.

**Retention Metrics.** There is a wide range of metrics for measuring user retention in online platforms, such as duration of membership (also known as "lifetime) [148, 149] and average time gap between visits. We are particularly interested in metrics that capture the tenure of a user based on content generation. To this end, we adopt the following metrics for measuring the retention of a user: the number of content created by the user, the number of active months (an active month refers to a month during which the user created at least one content), and the range of active months (time gap between the user's first active moth and last active month). Note that, within the range of active months, the user may be active in some months and inactive in others; therefore, the number of active months and the range of active months may be different.

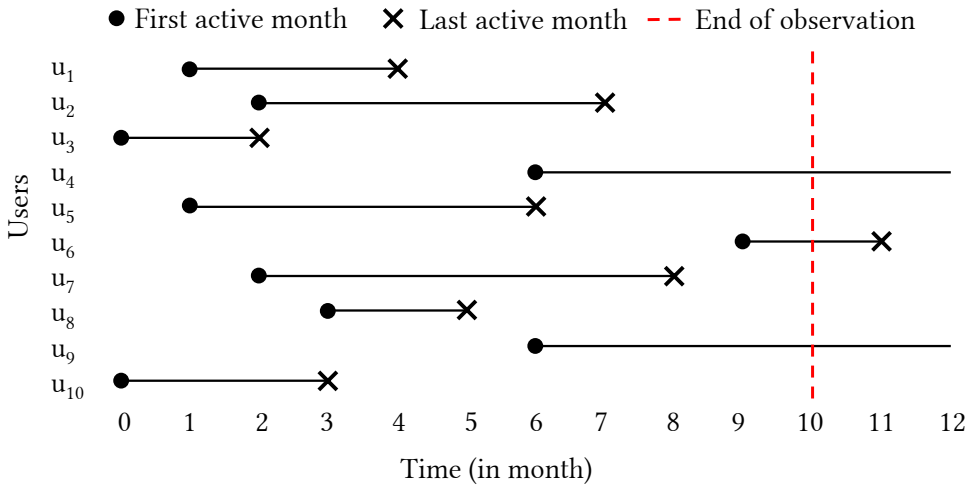



Figure 5.1: Censorship in measuring retention metrics. If we collect data at the end of month 10, we do not observe the last active month for three users ($u_4$, $u_6$, and $u_9$).

**Data Censorship.** While our retention metrics are practical, they are susceptible to data censorship. Specifically, at the end of our observation period (or the time of data collection), we encounter two types of users: users who are already inactive (labeled as "dead"), and users who are still active (labeled as "censored"). In Figure 5.1, for instance, if we collect data at the end of month 10, three users ($u_4$, $u_6$, and $u_9$) are censored. These censored users may continue to be active in future. If we use the observed final value of a retention metric as its limiting value, we may underestimate the retention of censored users. This faulty estimation, in turn, will affect the overall retention statistics, such as the average retention of users.

**Survival Curves.** To overcome the data censorship issue, we examine a detailed view of

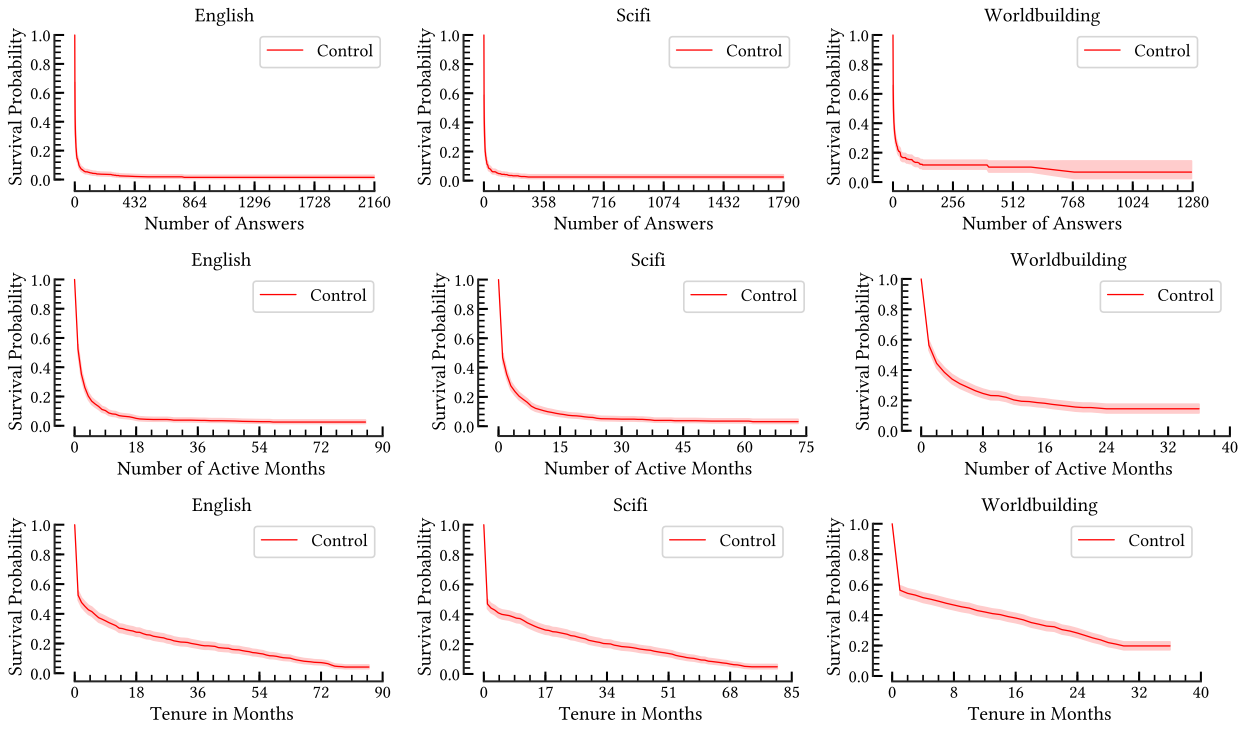



Figure 5.2: Survival curves for user retention in three Stack Exchange sites: `English`, `Scifi`, and `Worldbulding`.

the retention metrics through survival curves. The survival curves in Figure 5.2 show the probability that a user will be active after a specified period say a certain number of active months or a certain number of answer contributions. Notice that retention is particularly low for newcomers, as exhibited by the sharp decline in survival curves.

In this study, we ask the following question: is it possible that if the members of a community (say, English) vote differently, the platform operators would see higher retention? The goal of the rest of this study is to develop a methodology to evaluate alternative voting norms to investigate how they can improve user retention.

## 5.4 COUNTERFACTUALS OF VOTING

In this section, we develop a counterfactual formalism to reason about alternative voting norms using log data. We first describe the alternative voting norms that we will examine in this study (Section 5.4.1). We then explain our counterfactual setup to evaluate these norms (Section 5.4.2). Finally, we present statistical estimators for our counterfactual evaluation (Section 5.4.3).

### 5.4.1 Alternative Voting Norms

In this study, we develop a methodology to reason about alternative voting norms based on the principles of counterfactual evaluation. We refer to the existing voting norm as the *control norm*, whose voting outcomes can be observed from log data. Our goal is to evaluate the outcomes of alternative voting norms that were never observed. To this end, we examine several voting norms that express how the members of a community can vote based upon different voting criteria. Specifically, we examine six alternative voting norms defined as follows.

**Random:** Under the random voting norm, the members of a community cast votes on content without following any common voting criteria. For the purpose of our analysis on Stack Exchange, we assume the following distribution of votes to instantiate the random voting norm: the total number of votes on all answers to a question are distributed randomly among the answers.

**Uniform:** Under the uniform voting norm, the members of a community cast votes on content based on "the principle of equality", where all content receive the same number of votes. We instantiate the uniform voting norm as follows: the total number of votes on all answers to a question are distributed equally among the answers.

**Length:** Under the length voting norm, the members of a community cast votes on content based on the number of words in content, where content that contain more words receive more votes compared to content that contain fewer words. We instantiate the length voting norm as follows: the total number of votes on all answers to a question are distributed in such a way that the number of votes on each answer is proportional to the number of words it contains.

**Readability:** Under the readability voting norm, the members of a community cast votes on content based on the degree of ease in understanding content, where content with higher readability receive more votes compared to content with lower readability. We instantiate the readability voting norm as follows: the total number of votes on all answers to a question are distributed in such a way that the number of votes on each answer is proportional to its Flesch reading score [150]. Flesch reading score is a real number within the range [0.0, 1.0], where 0.0 implies confusing to read and 1.0 implies easy to read content.

**Objectivity:** Under the objectivity voting norm, the members of a community cast votes on content based on the degree to which content express facts rather than opinions,

where content with higher objectivity receive more votes compared to content with lower objectivity. We instantiate the objectivity voting norm as follows: the total number of votes on all answers to a question are distributed in such a way that the number of votes on each answer is proportional to its objectivity score. We compute objectivity score using a lexicon based approach [151], which returns a number within the range [0.0, 1.0], where 0.0 implies highly subjective and 1.0 implies highly objective content.

**Polarity:** Under the polarity voting norm, the members of a community cast votes on content based on the type of emotion expressed in content, where content with positive emotion receive more votes compared to content with neutral or negative emotion. We instantiate the polarity voting norm as follows: the total number of votes on all answers to a question are distributed in such a way that the number of votes on each answer is proportional to its polarity score. We compute polarity score using a lexicon based approach [151], which returns a number within the range [-1.0, 1.0], where -1.0 implies negative emotion, 0.0 implies neutral emotion and 1.0 implies positive emotion in content.

In the next subsection, we develop a counterfactual formalism to evaluate the retention effects of alternative voting norms. What would have happened (in terms of user retention) if the members of a community had cast votes on content following one of these norms?

### 5.4.2 Counterfactual Evaluation of Voting Norm

Given the control voting norm and an alternative voting norm, we now explain our counterfactual setup to evaluate the outcomes of the alternative voting norm in terms of user retention. Recall that, the outcomes of the control norm are experiential (recorded in log data), whereas the outcomes of the alternative norm are counterfactual (never-experienced). The key idea of our counterfactual evaluation is to use the outcomes of the control norm as references for evaluating the outcomes of the alternative norm in terms of user retention. In the remaining of this section, we introduce the key constructs of our counterfactual setup, describe the characteristics of our log data which embody these constructs, and explain the mechanics of our counterfactual evaluation using the log data.

**Constructs.** The key elements of our counterfactual setup are three co-varying constructs : user ($u$), votes ($\boldsymbol{v}$), and retention ($r$). Below we explain each of these constructs.

**User ($u$):** Our first construct refers to a user $u$ who created content. For instance, consider

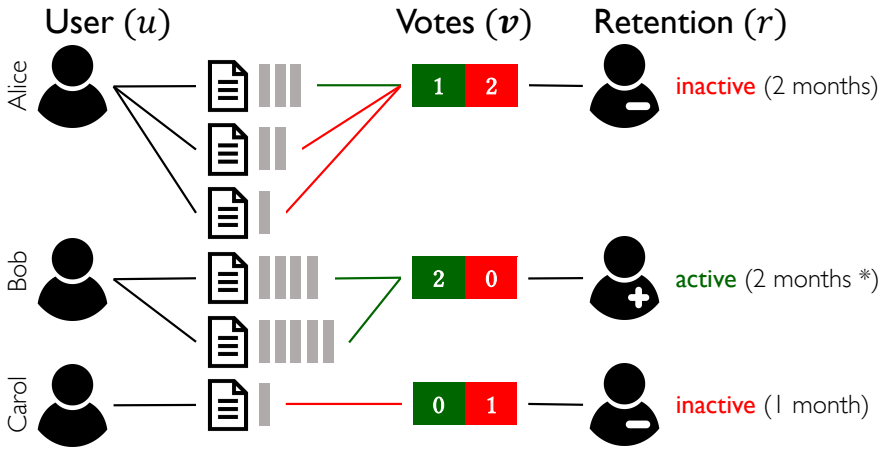



Figure 5.3: An example scenario of community voting in Stack Exchange. A user Alice created three answers. The answers received 3, 2, and 1 votes respectively. Alice left Stack Exchange after two months. Could a change in voting norm make her stay?

the scenario illustrated in Figure 5.3, where a user Alice created three answers in a Stack Exchange website.

**Votes ($\boldsymbol{v}$):** Our second construct refers to the collection of votes $\boldsymbol{v}$ that user $u$ received on his/her content. For instance, in the scenario illustrated in Figure 5.3, Alice received 3, 2, and 1 votes on her three succeeding answers, summarized using a utility vector.

**Retention ($r$):** Our third construct refers to the retention $r$ of user $u$. For instance, in the scenario illustrated in Figure 5.3, Alice left Stack Exchange website after two months.

**Log Data.** Our log data $\mathcal{L}$ can be seen as a set of $n$ historical observations of form $(u, \boldsymbol{v}, r)$ that embody the three constructs: $\mathcal{L} = \{(u_1, \boldsymbol{v_1}, r_1), \ldots, (u_n, \boldsymbol{v_n}, r_n)\}$. Note that, the observations in our data are based upon the control voting norm. To reason about an alternative voting norm in a counterfactual setup, we must adjust the weights of these observations as per the alternative norm. The concept of propensity score weighting underlies this adjustment, where the probability of a particular observation under the control and alternative norm may be different.

**Mechanics.** We develop a counterfactual formalism to evaluate the outcomes of an alternative voting norm in terms of user retention. Our formalism relies on two functions that capture voting norm and user retention respectively.

**Voting Norm $\pi(\boldsymbol{v}|u)$:** We consider platforms where voting norm decides the votes on content. The key idea of our counterfactual evaluation is that under different voting norms,

the probability of a user $u$ receiving a particular collection of votes $\boldsymbol{v}$ may be different. Accordingly, we conceptualize each voting norm (both control and alternative) $\pi$ as a function of form $\pi(\boldsymbol{v}|u)$ that takes user $u$ as input and generates the conditional probability of collection of votes $\boldsymbol{v}$ as output. For control norm, we estimate $\pi(\boldsymbol{v}|u)$ from observational data; whereas for alternative norm we estimate $\pi(\boldsymbol{v}|u)$ from synthetic data generated using the functions described in Section 5.4.1. We defer the discussion of this estimation to the next section. For the remaining of this section, we assume that, for a given voting norm, there's a way for us to estimate $\pi(\boldsymbol{v}|u)$.

**User Retention** $r(u, \boldsymbol{v})$**:** Our evaluation metric of interest is user retention, which can be captured via any of the retention metrics described in Section 5.3. To evaluate voting norms, we conceptualize user retention as a function $r(u, \boldsymbol{v})$ that takes user $u$ and the collection of votes $\boldsymbol{v}$ that he/she received as input and generates the retention of the user as output.

We refer to our control and alternative norm as $\pi_c$ and $\pi_a$ respectively. The goal of our counterfactual evaluation is to evaluate norm $\pi_a$ using a log $\mathcal{L}$ collected under norm $\pi_c$, where our evaluation metric is user retention. To perform this evaluation, we need to compute the expected retention $R(\pi_a)$ for norm $\pi_a$.

Now, the expected retention $R(\pi)$ of a norm $\pi$ is a mathematical expectation of the retention function $r(u, \boldsymbol{v})$ under the distribution of users $p(u)$ and norm $\pi(\boldsymbol{v}|u)$.

$$R(\pi) = \sum_{u} \sum_{\boldsymbol{v}} p(u)\pi(\boldsymbol{v}|u)r(u, \boldsymbol{v}) \tag{5.1}$$

For control norm $\pi_c$, we can estimate this expected retention $\hat{R}(\pi_c)$ directly from log $\mathcal{L} = \{(u_1, \boldsymbol{v_1}, r_1), \ldots, (u_n, \boldsymbol{v_n}, r_n)\}$ as follows.

$$\hat{R}(\pi_c) = \frac{1}{n} \sum_{i} r_i \tag{5.2}$$

Recall that we do not have any log data collected under the alternative voting norm in $a$. Therefore, we can only use the data collected under control norm $c$ for evaluating the alternative norm. Intuitively, since we can not change the observed votes, what we are interested in asking is that what if the alternative norm generated the observed votes. We can achieve this by using the Inverse Propensity Score Estimator or IPS estimator, which simply adjusts the weights of observations from log data based on the ratios of their occurrence probabilities under the control and alternative norms. The average retention in the adjusted data provides us the counterfactual retention for the alternative norm.

### 5.4.3 Inverse Propensity Score Estimators

As log $\mathcal{L}$ is collected under $\pi_c$, we can not use its observations directly for computing expected retention $\hat{R}(\pi_a)$ for $\pi_a$. Rather, we must adjust the weights of the observations in $\mathcal{L}$ to generate pseudo-observations for n$\pi_a$. In this subsection, we present several counterfactual estimators for performing this task. The basic concept of counterfactual estimators lies in importance sampling or inverse propensity score (IPS) [152]. We present several IPS based estimators to determine the expected retention under an alternative norm.

**Basic IPS Estimator (IPS).** The basic IPS estimator adjusts the weights of observations from log $\mathcal{L} = \{(u_1, \boldsymbol{v_1}, r_1), \ldots, (u_n, \boldsymbol{v_n}, r_n)\}$ based on the ratios of their occurrence probabilities under the control and alternative norm. It is a Monte-Carlo estimator defined as follows

$$\hat{R}_{IPS}(\pi_a) = \frac{1}{n} \sum_i r_i \frac{\pi_a(\boldsymbol{v_i}|u_i)}{\pi_c(\boldsymbol{v_i}|u_i)} \tag{5.3}$$

IPS is an unbiased estimator that is guaranteed to converge to the expected retention with sufficient samples. The key limitation of this estimator is its high variance, which may make it unsuitable for norm comparison. There are two sources of variance in IPS: the variability of retention function $r(u, \boldsymbol{v})$, and the mismatch in probabilities between the norm $\pi_c$ and $\pi_a$. To address the second source of variance, control variates are a popular choice. Control variates exploit information about known quantities to reduce the variance of an estimate of an unknown quantity.

**Normalized IPS Estimator (NIPS).** The normalized IPS estimator uses expected sample size $\hat{W}(\pi_a)$ as a multiplicative control variate to reduce the variance of the basic IPS estimator. It can be described as follows.

$$\hat{R}_{NIPS}(\pi_a) = \frac{1}{\hat{W}(\pi_a)} \sum_i r_i \frac{\pi_a(\boldsymbol{v_i}|u_i)}{\pi_c(\boldsymbol{v_i}|u_i)} \tag{5.4}$$

$$\hat{W}(\pi_a) = \frac{1}{n} \sum_i \frac{\pi_a(\boldsymbol{v_i}|u_i)}{\pi_c(\boldsymbol{v_i}|u_i)} \tag{5.5}$$

Unlike basic IPS, NIPS provides a biased estimate of the expected retention, however, with a lower variance compared to the basic IPS. Further, the bias itself decreases in proportion to the sample size.

Using the IPS estimators described above, we can compute the expected retention $\hat{R}(\pi_a)$ for norm $\pi_a$. The only required items are $\pi_c(\boldsymbol{v}|u)$ and $\pi_a(\boldsymbol{v}|u)$, which we need to estimate from data. In the next section, we show how we estimate these functions.

## 5.5 ESTIMATING VOTING NORMS

In this section, we first quantify how a user derives utility from votes (Section 5.5.1), then describe our models for estimating the control norm $\pi_0$ (Section 5.5.2) and alternative norm $\pi_a$ (Section 5.5.3).

### 5.5.1 Quantifying User's Utility

To evaluate the effects of voting norms on user retention, we need to represent the collection of votes that a user received on his/her content in terms of derived utility. Our goal is to keep this quantification as simple as possible while still capturing the acquisition of votes on content as incentive. Motivated by the prior work on community voting [143], we adopt a proportion based approach to quantify how an answerer derives utility from the votes he/she receives on answers in Stack Exchange.

Without loss of generality, let's assume that in Stack Exchange an answerer's utility can be discretized into two utility bins (see Table 5.1 for definition): $B_{high}$, $B_{low}$. The utility bins capture the percentage of votes an answer receives with respect to the total votes across all answers to its parent question. We quantify the utility of an answerer as the distribution of his/her answers over these two utility bins. For instance, consider the following scenario in Stack Exchange. For instance, consider the scenario illustrated in Figure 5.4. A user, Alice, created three answers on a Stack Exchange site. She received 3, 2, and 1 vote on her three succeeding answers. As per the observed votes, the first answer one falls into the high utility bin, whereas the next two answers fall into the low utility bin. We represent this distribution as a utility vector, as shown in the Figure.

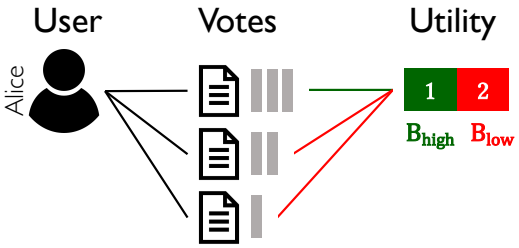



Figure 5.4: Example of constructing utility vector.

For each user user $u$, we represent the collection of received votes $\boldsymbol{v}$ using the two utility bins ($B_{low}$, $B_{high}$) as follows: $\boldsymbol{v} = [\lambda_{low}, \lambda_{high}]$, where $\lambda_{low}$ represents the number of answers that falls into the bin $B_{low}$, and $\lambda_{high}$ represents the number of answers that falls into the bin

Table 5.1: The two utility bins ($B_{high}$, $B_{low}$) capture the percentage of votes an answer received with respect to the total votes across all answers to its parent question.

| **Bin** | **Definition** |
|---|---|
| $B_{low}$ | Answer received less than 50% of total votes. |
| $B_{high}$ | Answer received more than or equal to 50% of total votes. |

$B_{high}$. We use this representation to estimate $\pi(\boldsymbol{v}|u)$ for the control norm $\pi_0$ and alternative norm $\pi_a$. Our estimation relies on two main hypotheses. First, for any voting norm, voting outcomes are probabilistic. Second, the probabilities of voting outcomes may vary from one voting norm to another. On the basis of these two hypotheses, we estimate the voting norms in successive steps as follows. First, for a given community voting norm, we either observe (control norm) or compute (alternative norm) the votes on each answer. We then derive a probability distribution per answer that explains the uncertainity around the votes on that answer. Next, using these probability distributions, we compute the probability of each answer to belong to each of the two utility bins defined in Table 5.1. Finally, for each answerer, based upon the probabilities associated with each of his/her answers (to belong to the two utility bins), we compute the probability of the answerer to receive the utility observed under the control norm. In the next subsections, we show how we estimate the control and alternative voting norms as discussed above.

### 5.5.2 Estimating Control Norm $\pi_0$

To estimate the control norm $\pi_0$, we begin by examining our log data. The log data reports the voting outcomes—the distribution of votes over the answers to a question—under the control norm. However, there is uncertainty associated with the votes. We develop a propensity model for quantifying the uncertainties in voting outcomes. Propensity, in our setup, refers to the probability that a certain voting outcome will be observed. For instance, given a question with three answers, what is the probability that the answers will receive 5, 5, and 2 votes respectively?

Our propensity model explains the voting behavior of a community from a contextual perspective, in which voters examine the current context to evaluate and subsequently vote on the answers. We define the *context* of an answer as all answers to its parent question. There is a natural tendency among voters to cast vote on an answer in the light of other answers to its parent question. We jointly model the votes on the answers to a given question

using a Dirichlet-multinomial model, which offers high flexibility and expressive power in quantifying uncertainity.

In our Dirichlet-multinomial model, we represent the distribution of votes over all answers to a given question as a multinomial distribution. In addition, we use a Dirichlet distribution with symmetric parameters as a conjugate prior for the parameters of the multonomial.

To generate a voting outcome for all $k$ answers to a given question, we first draw a $k$-dimensional probability vector $\boldsymbol{\theta}$ from a Dirichlet distribution with parameter vector $\boldsymbol{\alpha}$. We then draw a $k$-dimensional observation $\boldsymbol{x}$ from the multinomial distribution with the probability vector $\boldsymbol{\theta}$ and the total number of votes $n$.

$$\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k) \sim \text{Dir}\,(\boldsymbol{\alpha}) \tag{5.6}$$

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_k) \sim \text{Multi}\,(n, \boldsymbol{\theta}) \tag{5.7}$$

To learn the propensities of voting outcomes from the observed votes, we need to estimate the posterior distribution $p(\boldsymbol{\theta}|D)$, where $D$ refers to the observed data. In our Bayesian setup, $p(\boldsymbol{\theta}|D) \propto p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})$. If the prior of a multinomial distribution is a Dirichlet distribution with parameter vector $\boldsymbol{\alpha}$, then its posterior reduces to a Dirichlet distribution with parameter vector $\boldsymbol{N} + \boldsymbol{\alpha}$, where $\boldsymbol{N}$ is the the observed distribution of votes over the answers to a question.

$$p(\boldsymbol{\theta}|D) \sim \text{Dir}\,(\boldsymbol{N} + \boldsymbol{\alpha}) \tag{5.8}$$

We use this posterior distribution to quantify the uncertainty in the distribution of votes over the answers to a given question. We can further quantify the uncertainly for individual answers by computing the marginal distributions of our posterior Dirichlet, which are beta distributions.

Using the marginal beta distributions described above, we can compute the probability of an answer to belong to each of the two utility bins defined in Table 5.1. For each answerer, we can further combine these probabilities to compute the probability of the answerer to receive the utility derived under the control norm. The underlying probability model is a generalization of the multiple biased coin model—given $z$ biased coins, each with a different Bernoulli parameter $p_i$, determine the probability that $k$ of these coins will land on heads. In our case, given $z$ content generated by a user, each with a different probability distribution over the utility bins, determine the probability that $k$ of these content will belong to the high utility bin.

### 5.5.3 Estimating Alternative Norm $\pi_a$

To estimate the alternative norm $\pi_a$, we begin by generating synthetic data using the functions described in Section 5.4.1. The synthetic data captures counterfactual voting outcomes—the distribution of votes over the answers to a question—under the alternative norm. However, same as the control norm, there is uncertainty associated with the votes under the alternative norm. Accordingly, we develop a Dirichlet-multinomial model to capture the uncertainty around votes on the answers, as per the alternative norm. Further, using the steps described in Section 5.5.2, we compute the probability of the answerer to receive the same utility under the alternative norm.

In Figure 5.5, we show how we estimate the alternative norm. Consider our hypothetical user Alice, who created three answers and received 3, 2, and 1 vote, summarized using a utility vector. In the control norm, the first answer's probability of belonging to the high utility bin is high (0.7). For the remaining two answers, the probabilities are low (0.2 and 0.1). The overall probability of the utility vector (2 answers in low and 1 answer in high utility bin), can be computed as a joint. For the control norm, the probability is high (0.576). In contrast, for the alternative norm, the first answer's probability of belonging to the high utility bin is high (0.8). For the remaining two answers, the probabilities are also high (0.4 and 0.7). The overall probability of the utility vector under the alternative norm is low. In other words, Alice is less likely to receive low votes under the alternative norm.

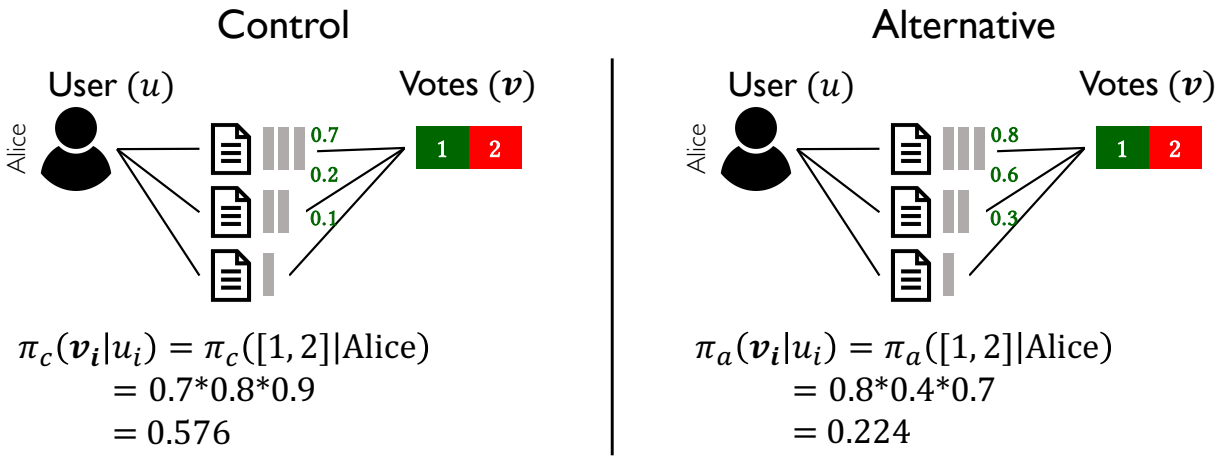



Figure 5.5: Example of estimating alternative norm.

Based on the estimation of control norm $\pi_0$ (0.576) and alternative norm $\pi_a$ (0.224), we conclude that Alice has a higher probability of receiving the aforementioned utility under

the control norm compared to the alternative norm.

## 5.6 EXPERIMENTAL EVALUATION

In this section, we identify the best performing community voting norm on the basis of user retention. We report the experimental setup (Section 5.6.1), compare different voting norms using our IPS framework (Section 5.6.2), and analyze the performance of the best performing norm through the lens of survival curve (Section 5.6.3).

### 5.6.1 Experimental Setup

We first describe our experimental setup: the dataset, the evaluation metrics, and the candidate policies.

**Dataset.** We collected the latest release (September, 2017) of the Stack Exchange dataset. This snapshot is a complete archive of all activities in Stack Exchange sites. There are 169 sites in our collected dataset. For the purpose of analysis, we show the results for three Stack Exchange sites: English (English linguistics), Scifi (science fiction), and Worldbuilding (building imaginary worlds). We select these three sites for following reasons. First, the sites contain many open-ended questions that accommodate more than one correct answer. The average number of answers per question for the three sites are as follows: 2.4 for English, 1.96 for Scifi, and 4.82 for Worldbuilding. Second, the topics of the sites promote descriptive content with rich text, which make them suitable for our text based quality metrics. The average number of words per answer for the three sites are as follows: 110 for English, 203 for Scifi, and 271 for Worldbuilding. Third, the sites cover diverse categories: life (Worldbuilding), culture (English), and recreation (Scifi).

Table 5.2: Summary statistics for Stack Exchange sites analyzed in this section.

| Site | # of Users | # of Questions | # of Answers | Answer Length |
|---|---|---|---|---|
| English | 31381 | 87679 | 210338 | 110 |
| Scifi | 19144 | 41943 | 82130 | 203 |
| Worldbuilding | 8546 | 12853 | 61966 | 271 |

**Evaluation Metrics.** We use the following retention metrics for comparing the performance of our community voting norms.

Table 5.3: Comparison of community voting norms for English.

| Voting Norm | Answer Contributions | Active Months | Site Tenure |
|---|---|---|---|
| Control | 23.27 (± 7.73) | 4.42 (± 0.57) | 16.60 (± 1.53) |
| Uniform | 20.59 (± 16.68) | 4.59 (± 1.75) | 16.99 (± 3.79) |
| Random | 21.10 (± 16.73) | 4.74 (± 1.80) | 16.65 (± 3.68) |
| Length | 87.98 (± 75.57) | 11.07 (± 5.37) | 29.22 (± 10.16) |
| Readability | 34.03 (± 38.55) | 6.81 (± 3.04) | 23.34 (± 7.83) |
| Objectivity | 24.59 (± 12.94) | 6.93 (± 2.96) | 24.44 (± 8.64) |
| Polarity | 70.10 (± 65.95) | 9.89 (± 4.67) | 28.46 (± 9.47) |

Table 5.4: Comparison of community voting norms for Scifi.

| Voting Norm | Answer Contributions | Active Months | Site Tenure |
|---|---|---|---|
| Control | 15.42 (± 5.32) | 4.81 (± 0.60) | 15.96 (± 1.43) |
| Uniform | 10.99 (± 5.46) | 4.63 (± 1.57) | 15.59 (± 3.45) |
| Random | 14.00 (± 6.47) | 5.88 (± 2.10) | 17.36 (± 3.97) |
| Length | 37.80 (± 34.38) | 7.33 (± 3.00) | 19.74 (± 5.53) |
| Readability | 16.68 (± 8.19) | 6.63 (± 2.50) | 19.17 (± 4.90) |
| Objectivity | 13.06 (± 6.65) | 5.73 (± 2.25) | 18.28 (± 4.74) |
| Polarity | 14.24 (± 6.74) | 6.18 (± 2.25) | 19.04 (± 4.68) |

1. **Answer Contributions:** The average number of answers contributed by the users.
2. **Active Months:** The average number of active months (when a user contributes) for the users.
3. **Site Tenure:** The average tenure (in months) for the users.

We compute the expected value of these metrics for different voting norms using the IPS estimator (as described in Section 5).

**Community Voting Norms.** We estimate user retention for the community voting norms described at Section 5.4.1. For each of these seven norms—control, random, uniform, length, readability, objectivity, and polarity—we compute the expected value of the retention metrics.

Table 5.5: Comparison of community voting norms for WORLDBUILDING.

| Voting Norm | Answer Contributions | Active Months | Site Tenure |
|---|---|---|---|
| Control | 17.72 (± 4.63) | 4.79 (± 0.41) | 11.83 (± 0.79) |
| Uniform | 27.26 (± 26.57) | 4.99 (± 1.28) | 11.62 (± 1.72) |
| Random | 31.67 (± 27.23) | 5.59 (± 1.52) | 12.36 (± 2.13) |
| Length | 67.12 (± 46.24) | 9.99 (± 3.52) | 18.36 (± 5.00) |
| Readability | 28.19 (± 13.82) | 7.69 (± 2.52) | 15.73 (± 3.90) |
| Objectivity | 60.13 (± 37.42) | 9.52 (± 3.14) | 17.29 (± 4.39) |
| Polarity | 66.92 (± 44.39) | 9.79 (± 3.22) | 17.63 (± 4.51) |

### 5.6.2 Voting Norm Comparison

We compare the performance of the community voting norms by estimating the expected value of retention metrics through IPS estimator. We specifically examine the retention of first 1000 users in the platforms. Table 5.3 shows the comparison of community voting norms for ENGLISH; Table 5.4 for SCIFI; Table 5.5 for WORLDBUIDLING. We make the following observations from these results.

1. **Content Matters.** For all three sites, the four content-based voting norms, namely, length, readability, polarity, and objectivity, typically outperform the control norm. In some cases, the content-based voting norms achieve order of magnitude improvement in user retention over the control norm.

2. **Effort Matters.** For all three sites, length based voting norm outperforms the remaining voting norms across all retention metrics. Since a longer answer typically requires more effort by the answerer, it is intuitive that voting based on the effort of the answerer significantly improves retention.

3. **Equality is Not Equity.** For all three sites, the uniform voting norm typically underperform the content-based norms. This is consistent with our expectation that a uniform voting norm may demotivate a user from contributing as the utility is low.

4. **Not So Random.** Across all sites, the random voting norm typically underperform the content-based norms. Community voting norms are stochastic, in which randomness is a necessary element. However, a completely random voting norm may demotivate users.

In summary, we observe that content-based voting norms can improve user retention in

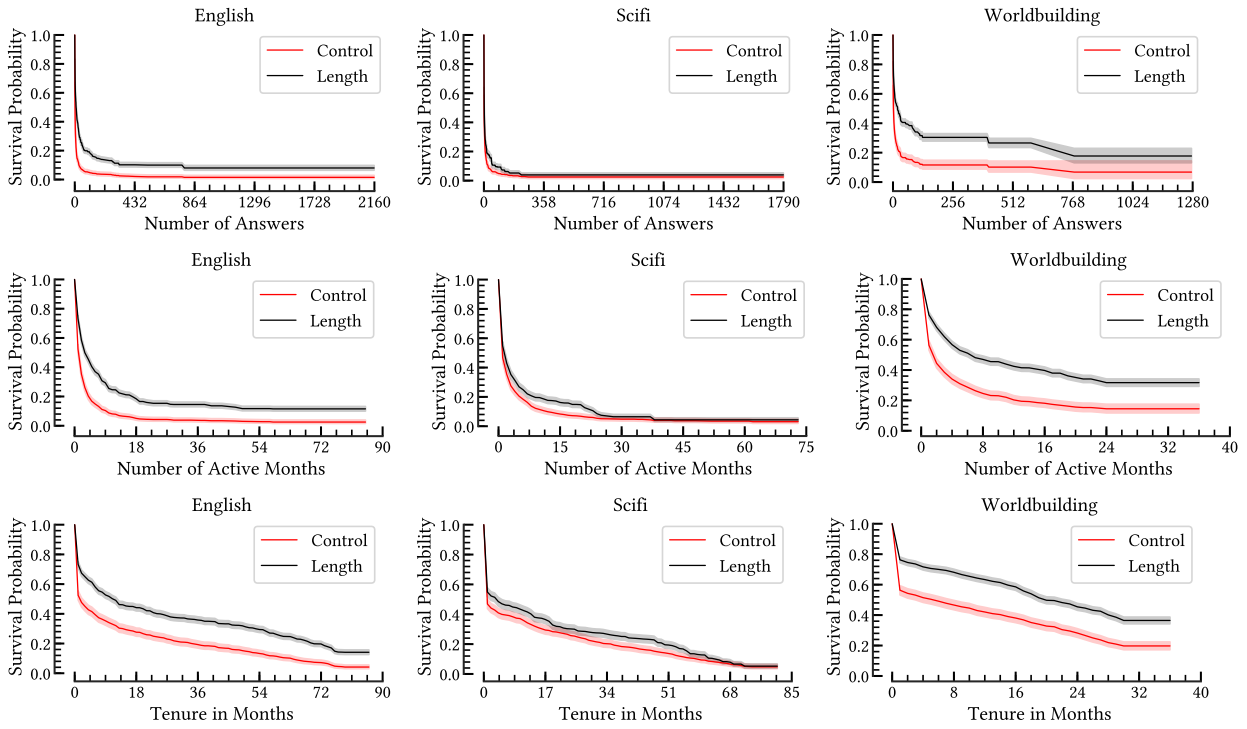



Figure 5.6: Survival curves for control and length voting norm in three Stack Exchange sites: `English`, `Scifi`, and `Worldbulding`.

Stack Exchange wsites. In the next subsection, we examine the best content-based voting norm (length based voting) in the light of survival curves.

### 5.6.3 Survival Curve

We observe that the length based voting norm improves the average retention of users for all sites. However, this could be achieved by increasing the probability of tail users (users with high life expectancy), while ignoring other users. A more evolved norm should improve the overall survival probability for entire population. We investigate whether the high retention in length based voting norm is due to the tail users or not. To this end, we compare the survival functions for the control norm and the length based voting norm. Specifically, we use KM estimator to compute survival functions for the control norm and the length based norm. Figure 5.6 presents these survival functions using survival curves. Notice that, the length based voting norm improves survival probability of users at all stages. The norm's improvement therefore comes from the entire population.

**Survival Function.** The survival function $S(t)$ captures the probability that a user will survive past time $t$, and can be expressed as $S(t) = p(T > t)$. It has the following properties:

$0 \leq S(t) \leq 1$, $S(t + \delta t) \leq S(t)$, and $\lim_{t \to \infty} S(t) = 0$.

**Kaplan-Meier Estimator (KM).** Kaplan-Meier estimator is a non-parametric statistic used to estimate the survival function $S(t)$ from data. The KM estimator can be expressed as:

$$\hat{S}(t) = \prod_{t_i \leq t} \frac{n_i - d_i}{n_i}, \tag{5.9}$$

where $d_i$ is the number of death events at time $t$ and $n_i$ is the number of subjects at risk of death just prior to time $t$.

## 5.7 RELATED WORK

Our work draws from, and improves upon, a wealth of research on social media—notably, voting behavior, policy design, user retention, content evaluation, and counterfactual estimation.

**Voting Behavior.** Recent research has made great advancements towards the understanding of rating and voting behaviors in online platforms. Much of the work has focused on studying biases, including social influence bias (especially herding behavior) [13, 88, 90, 93, 95, 97, 98, 120, 122], and reputation bias [114, 153]. The most relevant work to ours is research that has investigated community voting behavior [96, 143], and its consequences. Notably, Cheng et al. [143] proposed a propensity matching framework for examining how votes on a content affect the future behavior of the content creator. It is important to note the contrast between our proposed framework and the framework by Cheng et al.—that our framework allows examining the effects of *future (i.e., never experienced) voting behavior* on user retention.

**Policy Design.** Designing policies, in particular badges, for maximizing user participation has attracted a lot of attention in recent years [24, 25, 26, 27, 28, 29, 30]. Notably, Anderson et al. [29, 30] developed both theoretical and empirical methods for examining the effects of badges on user behavior, and studied badge design in MOOCs [29] and question-answering platforms [30]. Immorlica et al. [28] studied badge design mechanisms for maximizing the total contribution of users. In this study, we examine policy design from a behavioral perspective, in which we conceptualize community behavior as a "policy". Therefore, our policy design problem is *in fact a behavior adoption problem*. We are not looking to deploy better incentives (say badges), rather identify the desired voting behavior of a community, which is expected to improve user retention.

**User Retention.** Another line of related work studied user retention in online platforms [149, 154, 155, 156, 157, 158, 159]. Yang et al. developed ClusChurn [154]—a

framework for interpretable user clustering and churn prediction, based on an analysis of Snapchat's userbase. Kapoor et al. [149] applied Cox's proportional hazard model to predict the return time of users in an online music service. In this study, we examine user retention in the light of *content generation and subsequent community feedback*; specifically, how community feedback on content affects the retention of users who generated these content.

**Content Evaluation.** Evaluating social media content is crucial for recommendation and search, and has been extensively studied by researchers [80, 81, 82, 83, 160, 161, 162]. The general theme of these works is to predict content quality based on various textual (e.g., length) and non-textual (e.g., user reputation) features. Most of these works determine the quality of a content by aggregating the raw opinion scores (e.g., votes) from users. However, raw opinion scores from users are often unreliable, due to various biases resulting from the absence of independent assessment [110]. In this study, we investigate if alternative voting schemes that do not accommodate these biases could improve user retention.

**Counterfactual Estimation.** There is a rich body of work on counterfactual estimation, and its applications to the offline evaluation of recommendation systems and search engines [163, 164, 165]. Similar to these works, we adopt an inverse propensity sampling (IPS) framework to reason about new policies. However, unlike any previous work on counterfactual evaluation, we evaluate *community voting policies*. Our goal is to identify an evolved voting policy for the purpose of improving user retention.

## 5.8 CONCLUSION

In this study, we developed a counterfactual framework for examining the effects of community voting on user retention. We conducted extensive experiments on 169 Stack Exchange sites comparing the default voting policy in these CQA forums with six alternative policies: random, uniform, length, readability, objectivity, and polarity. Our main finding is that had the community members voted on the basis of content properties, we would have observed increased user retention in Stack Exchange sites.

In this chapter, we showed how voting norms affect the retention of users. In the next chapter, we will study how user roles affect platform health.

# CHAPTER 6: DISCOVERING USER ROLES

During the early years of this thesis, we investigated if the lack of platform sustainability is due to the roles users play in platforms. Specifically, could it be that the abundance or shortage of certain types of users causes platforms to decline? This chapter discusses our fourth study, where we discover the roles users assume in content-based platforms and examine how the mixture of these roles affects platform health.

## 6.1 INTRODUCTION

Discovering user roles and community role compositions on Community Question Answering (CQA) platforms is an important challenge. CQA platforms such as the Stack Exchange platform[1] play an incredibly important role in today's society, and recent years have seen an increase in both the number of such CQA communities and the user populations within each community. For example, in 2017, StackOverflow[2] added over 200,000 new questions and over 130,000 new users every month; many software developers regularly depend StackOverflow to be effective at work. An understanding of behavior within communities can help to inform the decisions made by platform providers to steer the communities to be maximally effective.

It is well established that users in these communities play important, distinct roles [31, 166, 167, 168, 169], but it remains an important scalability challenge to automatically uncover these distinct user roles across a large number of communities. Stack Exchange as a platform, for example, facilitates 161 distinct websites. Manual investigation of user behavior compositions within and across these communities is prohibitively expensive to do without some level of automation, and with these communities continuing to grow over time, the need for automated role discovery intensifies.

Existing approaches fall short of our needs in a number of ways. Many existing models for role discovery do not consider the case of modeling many communities at once, yet such a cross-community understanding of behavior is important to enable comparative studies across communities. Previous work often defines roles based on a graph-centric approach [170, 171, 172], which fails to uncover many distinct roles beyond "answer people" and "discussion people." Other approaches require a manual definition of individual features to describe roles [173, 174], which can fail to cover all of the empirically present role patterns

[1] `https://stackexchange.com`

[2] `https://stackoverflow.com`, the largest community on the Stack Exchange platform

in the data.

In this paper, we propose a generative model for discovering action-based user roles and community role compositions in CQA platforms directly from log data. We formally define an action-based user behavior role as a probability distribution over atomic actions a user may take with respect to the CQA community within one browsing session. We also directly model the role compositions across all communities within the platform to facilitate comparative analysis of communities. This is achieved via the use of a mixture of Dirichlet-multinomial Mixtures (MDMM), which allows us to use statistical inference to uncover the latent user roles and community role compositions from log data directly, which can facilitate studies into user behavior both within and across communities on a CQA platform at scale. We envision that with the assistance of our model, human analysts can "see" more patterns than what they could see otherwise. Such a tool provides a useful "lens" through which to view behavior data, and opens up many directions for future studies that would not otherwise be possible.

To demonstrate that such a model is indeed useful as a tool to assist human discovery of behavior patterns within and between CQA communities, we perform a comprehensive experiment on all 161 non-meta communities on the Stack Exchange CQA platform that delivers three empirical insights. First, we show interesting distinctions in question-asking behavior on Stack Exchange (where two distinct types of askers can be identified) and answering behavior (where two distinct roles surrounding answers emerge). Second, we find statistically significant differences in behavior compositions across topical groups of communities on Stack Exchange, and that those groups that have statistically significant differences in health metrics also have statistically significant differences in behavior compositions, suggesting a relationship between behavior composition and health. Finally, we show that the MDMM behavior model can be used to demonstrate similar but distinct evolutionary patterns between topical groups.

## 6.2 RELATED WORK

The presence of roles on CQA platforms has been argued by many. For example, Adamic et al.[166] demonstrate that, on the Yahoo! Answers platform, there are at least three distinct user types—answerers, askers, and *discussion persons*. Mamykina et al.[167] argue for the presence of at least four distinct user roles on StackOverflow: *community activists*, *shooting stars*, *low-profile users*, and *lurkers and visitors*. Other studies have explored whether roles characterized by a single action are separate or overlapping [168, 169]. Developing tools to automatically uncover distinct user behavior types is a major thrust of this paper.

Many approaches for discovering these distinct user roles in the CQA setting require practitioners to define individual features used to describe the discovered roles [173, 174], and early work in the domain of user role modeling could only easily identify two critical roles ("answer people" and "discussion people") through the use of a graph-centric modeling approach [170, 171, 172]. More recent work explores a mixed-membership approach to user behavior modeling [175] in order to identify more user roles, but still takes a graph-centric modeling approach. In this work, we explore the newer direction of action-focused probabilistic modeling for user behavior in order to automatically discover roles in a way that requires less hands-on effort to define features and is flexible enough to be able to capture more nuance within the roles of "answer people" and "discussion people".

The application of probabilistic modeling for user behavior understanding has been explored before [176, 177, 178]. We extend this body of research by modeling the behavior composition at a community level, rather than just at a user level. This allows us to understand the behavior at the level of an entire community as it relates to others.

Perhaps the most relevant probabilistic behavior model to ours is the one proposed by Han et al.[179], where they attempt to jointly model three phenomena: social network link formation, community discovery, and behavior prediction. Their definition of user behavior differs from ours, however, as it considers only posting and reposting as the two possible actions a user can take. We attempt to define a much more comprehensive behavioral action set in this work. Furthermore, their discovered role distributions model real-valued user attributes, rather than behavior directly, which makes interpretation challenging. Our work, in comparison, assumes a different generative process over user action lists that leads to a set of readily interpretable probability distributions that define our roles.

CQA data, and in particular the Stack Exchange CQA platform, have been analyzed in many ways in previous literature [166, 167, 168, 169, 174, 180], but many do not discuss user roles in depth. Furtado et al.[174], however, do explore user roles and their dynamics using five of the communities on the Stack Exchange platform, but their definition of user roles arises from manual construction of user attributes and an agglomerative clustering approach. Our model, in comparison, is more general in that it should be applicable to any CQA community (or any social network) where articulating the set of actions users can take within the community is the only manual supervision required.

Our session-focused approach is closely related to the notion of clickstream mining [181, 182, 183, 184, 185, 186, 187], where a variety of clustering techniques is applied to find users that share similar clickstream traces. Many of these techniques utilize Markov models and focus on the task of predicting a user's next action. In this paper, we instead focus on characterizing the behavior of users in an interpretable way that also facilitates cross-

community comparisons.

The model we propose in this paper is essentially similar to topic models such as PLSA [188] or LDA [189], but the key difference is that the data modeled by our model are the user actions whereas topic models generally model text data where the input tokens are individual words within topics. The Dirichlet-multinomial mixture (DMM) [190, 191] is the closest related model to ours in this space. A DMM assumes that individual documents exhibit only one topic—our generative framework also assumes that one user session exhibits only one role.

Other approaches for user behavior modeling on CQA communities consider both actions and textual content to generate topic-specific action distributions [178, 192]. These distributions are similar to what we call roles, but the meaning they capture is very different—in their work these capture how users interact with a specific topic, whereas in our work they describe how to characterize an individual user's entire browsing session.

## 6.3 MODEL

The design of our model is motivated by our goal of discovering interpretable descriptions of functional roles played by users on CQA platforms, as well as a representation for each community as a mixture over these user roles. We explore a definition of user roles that considers the co-occurrence behavior of actions users take within individual browsing sessions. To accomplish this, we represent the roles as probability distributions that describe the likelihood of taking individual actions when a user is assuming a particular functional role in one session. This definition is advantageous: first, it is general, and thus should be applicable to any CQA platform (or even any social network); second, roles represented in this way can be readily interpreted by inspection; and third, it is able to capture the uncertainty associated with assigning users to roles.

### 6.3.1 Generative Process and Inference

The first step in the use of our action-based role discovery model is to define the set $A$ of actions users may take within a community. Defining the actions in this action set is very important in order to capture meaningful roles under our model, so careful attention should be paid to the construction of a set of disjoint actions whose proportions can meaningfully reflect a type of domain-relevant behavior.

Next, one must identify the collection of *observed* communities $C_{1:N}$ to analyze that all share the same action set $A$. We do not address the problem of community discovery in this

paper; rather, these communities are treated as input to the model. Each community must share the same types of allowed actions. In our case, we use individual websites that are all part of the same CQA platform (but focus on different topical domains) to ensure that by defining $A$ with respect to the CQA platform itself we can represent behavior across all of these communities.

To automatically discover distinctive user behavior types, which we will call our roles, we appeal to the general technique of probabilistic graphical models [193] and model user behavior using a mixed membership approach. The model assumes that there are $K$ distinct user roles, each of which is characterized with a categorical distribution $\phi_k$ over actions from some $A$; each of the roles $\phi_k$ is assumed to be drawn from a Dirichlet distribution with parameter $\beta$. With these user roles defined, we further assume that each community $C_i$ is associated with a mixing distribution $\theta_i$ (drawn from another Dirichlet distribution with parameter $\alpha$) that governs the distribution over the user roles for each *user session* that occurs *within that community*. If a user makes actions in multiple communities within one browsing session, we subdivide their browsing session into a collection of sessions, one for each community they participated in.

More concretely, we represent each community $C_i$ with a list of the user sessions $\langle \mathbf{s_{i,1}}, \mathbf{s_{i,2}}, \ldots, \mathbf{s_{i,M}} \rangle$ associated with it. Each session is itself a list of actions $\mathbf{s_{i,j}} = \langle a_{i,j,1}, a_{i,j,2}, \ldots, a_{i,j,T} \rangle$, with each $a_{i,j,t} \in A$. Each individual session $\mathbf{s_{i,j}}$ is associated with one particular user role $z_{i,j}$ that indicates the role distribution $\phi_{z_{i,j}}$ from which each of the actions within the session is drawn (note that an individual user is free to exhibit a different roles between different sessions). The full generative process is thus

1. For $k = 1$ to $K$ (number of roles), draw an action distribution $\phi_k \sim \text{Dirichlet}(\beta)$
2. For each community $C_i$:
   (a) Draw a role mixing distribution $\theta_i \sim \text{Dirichlet}(\alpha)$
   (b) For each $\mathbf{s_{i,j}}$ in community $C_i$:
      i. Draw a role for the session $z_{i,j} \sim \text{Categorical}(\theta_i)$
      ii. For $t = 1$ to $|\mathbf{s_{i,j}}|$ (the length of the session), draw a single action within the session $a_{i,j,t} \sim \text{Categorical}(\phi_{z_{i,j}})$

and is depicted using plate notation in Figure 6.1.

The resulting model is quite similar to a Dirichlet-multinomial mixture (DMM), which has seen use in the text mining community for clustering [191] and classification [190]. A major difference from our model, however, is that in a DMM one learns a *single* distribution $\theta$ that

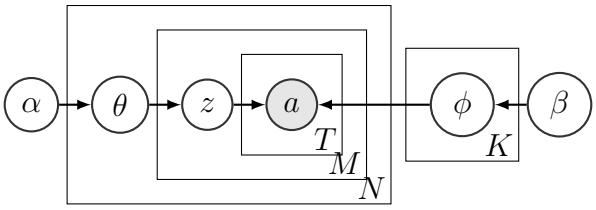



Figure 6.1: Plate notation for the role discovery model we propose. $\alpha$ parameterizes a Dirichlet distribution from which each community's role proportions, $\theta_i$, are drawn. $z$ represents the role assignment for a specific user session, and $a$ represents the actions taken within that user session. $\beta$ parameterizes a Dirichlet distribution from which each of the user roles $\phi_k$ are drawn, each of which is a categorical distribution over the possible action types.

governs the mixing proportions over the components $\phi_k$ that is shared for each element $C_i$, whereas our model instead learns a *separate* distribution $\theta_i$ for each individual community, but shares the description of the components $\phi_k$ between each. This allows us to compare two communities by their role proportions in a meaningful way since each $\theta_i$ will be a distribution over the same set of roles $\phi_k$. If one were instead to fit multiple DMMs, one for each community, comparison of the $\theta$ distributions would not necessarily be immediately obvious due to the fact that each model would learn a separate set of roles $\phi_k$. Thus, we view our model as a principled mixture of DMMs (MDMM) where we have made a deliberate decision to share a global set of role components between all communities $C_i$.

There are several approaches to inference in a DMM. Nigam et al.[190] use maximum a posteriori (MAP) estimation to obtain a point estimate. We instead choose to follow a more fully Bayesian approach similar to Yin et al.[191] and instead appeal to Markov-chain Monte Carlo methods to approximate the desired posterior distribution. Specifically, we integrate out $\theta$ and $\phi$ in order to then derive a collapsed Gibbs sampler that iteratively updates the latent role assignments $z_j$ by sampling new values from the full conditional distribution. When this chain has converged, we extract a MAP estimate for each $\theta_i$ and $\phi_k$ from the current state of the Markov chain.

Formally, we can define the full conditional distribution

$$p(z_{m,n} = z \mid \mathbf{Z}_{\neg m,n}, \mathbf{S}, \alpha, \beta) = \frac{p(\mathbf{Z}, \mathbf{S} \mid \alpha, \beta)}{p(\mathbf{Z}_{\neg m,n}, \mathbf{S} \mid \alpha, \beta)} \propto \frac{p(\mathbf{Z}, \mathbf{S} \mid \alpha, \beta)}{p(\mathbf{Z}_{\neg m,n}, \mathbf{S}_{\neg m,n} \mid \alpha, \beta)}, \quad (6.1)$$

where $\mathbf{Z}_{\neg m,n}$ indicates the set of all the assignments of $z_{i,j}$ with only $z_{m,n}$ excluded, and similarly $\mathbf{S}_{\neg m,n}$ indicate the set of all user sessions with only the specific session $\mathbf{s_{n,m}}$ absent. We begin by noting $p(\mathbf{Z}, \mathbf{S} \mid \alpha, \beta) = p(\mathbf{S} \mid \mathbf{Z}, \beta)P(\mathbf{Z} \mid \alpha)$, and focus on each term separately.

Following a similar argument to Yin et al.[191], we have $p(\mathbf{Z} \mid \alpha) = \prod_{i=1}^{N} \frac{B(\eta_i+\alpha)}{B(\alpha)}$, where $B(\alpha)$ is the multivariate beta function and $\eta_i$ is a vector where $\eta_{i,k}$ indicates the number of times role $k$ is chosen as the role assignment for a session in community $C_i$. Similarly, $p(\mathbf{S} \mid \mathbf{Z}, \beta) = \prod_{k=1}^{K} \frac{B(\tau_k+\beta)}{B(\beta)}$ where $\tau_k$ is a vector with $\tau_{k,a}$ indicating the number of times action type $a$ was assigned to role $k$ through its session's role assignment. From here, we can derive the sampling probability through cancellation of terms and exploiting the property of the gamma function that $\Gamma(1+x) = x\Gamma(x)$ and arrive at

$$\begin{aligned} & p(z_{m,n} = z \mid \mathbf{Z}_{\neg m,n}, \mathbf{S}, \alpha, \beta) \\ & \propto \frac{\alpha_z + \eta_{i,z}^{\neg m,n}}{\sum_{k=1}^{K} \alpha_k + \eta_{i,k}^{\neg m,n}} \\ & \quad \times \frac{\prod_{a \in \mathbf{s_{m,n}}} \prod_{j=1}^{c(a,\mathbf{s_{m,n}})} (\beta_a + \tau_{z,a}^{\neg m,n} + j - 1)}{\prod_{j=1}^{|\mathbf{s_{m,n}}|} \left( \left( \sum_{a=1}^{|A|} \beta_a + \tau_{z,a}^{\neg m,n} \right) + j - 1 \right)}, \end{aligned} \tag{6.2}$$

where $c(a, \mathbf{s_{m,n}})$ indicates the number of occurrences of action type $a$ in session $\mathbf{s_{m,n}}$.

As a practical matter, computing this probability is susceptible to underflow issues due to the products occurring in the second term. To prevent this issue, we use the Gumbel-max trick [194] to sample from this discrete distribution. This trick works by first computing the sampling proportions in log-space $\gamma_k = \log \tilde{p}(z_{m,n} = k \mid \mathbf{Z}_{\neg m,n}, \mathbf{S}, \alpha, \beta)$, where $\tilde{p}$ represents the un-normalized probability in equation 6.2, which effectively prevents the underflow issues. We then can sample from the original discrete distribution by sampling $k$ values $g_k \sim$ Gumbel(0), and taking the sample $z_{m,n} = \arg\max_k \gamma_k + g_k$. We have open-sourced the implementation of our inference algorithm under a liberal license[3].

### 6.3.2 Choosing the Number of Roles

The number of roles, $K$, remains a hyperparameter of the MDMM behavior model. How should one choose the "optimal" value for $K$? This is a similar question that is asked for nearly any mixed-membership or clustering model. We note, first, that the choice of $K$ can be an empirical parameter that is sometimes beneficial as it can give users *control* over the granularity of the model, much like a user can adjust the zoom level of a microscope. If the user does not know how to set $K$ a priori, we describe a procedure that can help choose a particular value of $K$ that may be optimal.

In our specific case, not only do we wish to discover distributions over actions that can

[3] `https://github.com/CrowdDynamicsLab/stackoverflow-stream`

adequately describe a user's behavior within a single session, but we wish for these distributions to be *meaningfully different* from one another. An ad-hoc approach, then, is to simply run the model for different values of $K$ in some range, and then investigate the roles $\phi_{1:K}$ that are produced. When moving from $k$ to $k+1$ roles, if a new role arises that is not meaningfully different from all of the $k$ roles found previously, this suggests that $k$ was the optimal number of roles for the data being modeled.

One can define a simple quantitative heuristic to capture this intuition. Formally, let $\phi_{1:k}$ be the $k$ roles proposed by the model previously, and let $\hat{\phi}_{1:k+1}$ be the $k+1$ roles proposed by the model when incrementing $K$. Consider a single new role $\hat{\phi}_i$. We can compute how different it is from each of the previously proposed roles $\phi_{1:k}$ by using the KL-divergence metric [195]. By taking the minimum divergence from the newly proposed role $\hat{\phi}_i$ to each of the $k$ previous roles, we have a measure for how "surprising" this new role is compared to the previous roles. If it is very similar to one of the existing roles, it will have a very low minimum KL-divergence; on the other hand, should it be very different from all of the previous roles, it would have a very large minimum KL-divergence.

If we then take the maximum value of this measure over all of the $k+1$ newly proposed roles $\hat{\phi}_{1:k+1}$, we obtain a number that reflects the largest minimum divergence between the set of $k$ old roles and the set of $k+1$ new roles. The smaller this value, the more redundant the set of $k+1$ new roles is compared to the set of $k$ previous roles. Formally, we can define this measure $\text{MaxMinKL}_{k \to k+1}$

$$\text{MaxMinKL}_{k \to k+1} = \max_{\hat{\phi}_i} \left( \min_{\phi_j} KL(\phi_j \mid\mid \hat{\phi}_i) \right). \tag{6.3}$$

To find the optimal value of $K$, one can run the model for $K$ in a range of values to be considered, computing $\text{MaxMinKL}_{k \to k+1}$ for each transition. When this value drops substantially, this is a sign that the new set of roles is not meaningfully different from the previous set of roles, and we should stop increasing $K$.

### 6.3.3 Applications of the Model

The MDMM behavior model is a tool to enable humans to discover new knowledge, explore new hypotheses, and test those hypotheses about user behavior in ways that they were unable to before. There are a number of different applications of the model beyond just the discovery of user behavior roles. We outline a few of them below, but note that this list is not exhaustive—exploring those opportunities are interesting future directions.

**Community Profiling.** A secondary output of the model are the mixing proportions $\theta_i$

over the roles for each community. These distributions provide a profile of the behavior of users within the community, which can be used as a representation for that community in downstream tasks. To explore this in more detail, in Section 6.4.3 and 6.4.4 we explore how we can use this representation to uncover communities with different behavior profiles, and show how these groups are correlated with many metrics of community success.

**User Profiling.** The model can also be used to infer the roles of a user by averaging over the roles they assume in their sessions. This output can then be used in downstream tasks that relate to understanding user behavior on an individual level and can be used as a representation of a user for other machine learning algorithms.

**Behavior Dynamics of Communities.** We can also uncover temporal community representations by further segmenting the user browsing sessions into buckets relating to different points in time. This allows us to study how behavior proportions evolve over time as community age. We explore this in more depth in Section 6.4.5.

**Behavior Dynamics of Users.** In much the same way we can uncover community representations over time, we can also uncover user representations over time. This output could be used to understand how individual users, or groups of users, change their behavior over time.

## 6.4 EXPERIMENTS

The goal of our experiments is to demonstrate the usefulness of the MDMM user behavior model as a tool for investigating user behavior in different ways. Our goal is *not* to be completely comprehensive or conclusive in our study of user behavior, but rather to lay a framework for future studies in a variety of different directions that could not otherwise be studied.

Our MDMM user behavior model provides two important outputs to characterize user behavior in CQA communities: (1) the latent role representations, and (2) the degree to which each latent role is present within each of the CQA communities. We apply our model to communities from the Stack Exchange CQA platform[4] in order to better understand its utility for role discovery and CQA community behavior analysis tasks. We take the entire Stack Exchange dataset consisting of a total of 322 websites and discard all "meta" websites (websites discussing one of the other Stack Exchange websites), leaving us with 161 non-meta websites (communities) for our analysis.

[4]The dataset is available here: `https://archive.org/details/stackexchange`. We used a dataset from 2016-12-12, which covers from 2008-07-31 through 2016-12-11.

Table 6.1: Action names and their definitions for our application of the MDMM behavior model on StackExcahnge. (m: "my", o: "other", q: "question", a: "answer")

| Action Name | Action Definition |
|---|---|
| question | Posting a new question |
| answer-mq | Answering your own question |
| answer-oq | Answering someone else's question |
| comment-mq | Commenting on your own question |
| comment-oq | Commenting on someone else's question |
| comment-ma-mq | Commenting on your own answer to your own question |
| comment-ma-oq | Commenting on your own answer to someone else's question |
| comment-oa-mq | Commenting on someone else's answer to your own question |
| comment-oa-oq | Commenting on someone else's answer to someone else's question |
| edit-mq | Editing your own question |
| edit-oq | Editing someone else's question |
| edit-ma | Editing your own answer |
| edit-oa | Editing someone else's answer |
| mod-vote | Voting for moderation action |
| mod-action | Moderating a post |

### 6.4.1 Dataset Construction

A critical component of the use of the MDMM in our setting is properly defining the action space to be considered, as the roles discovered are to be distributions over that action space. The flexibility of defining actions outside of the MDMM model makes it easy to accommodate analysis of action patterns at different levels of granularity by adjusting the granularity of the action space to be analyzed itself. However, in any specific application, carefully choosing the exact action set used is naturally very important. If the space of actions is defined too narrowly, this prevents discovering subtle differences between user roles.

To analyze the Stack Exchange dataset, we defined an action space based on the inherent content hierarchy present on the Stack Exchange platform (see Table 6.1 for a list of the action set we consider). Content on the Stack Exchange platform comes in three main types: questions (the root content), answers (which nest below questions), and comments (which can nest either beneath questions or answers), so it is natural to consider an action set consisting of the creation action for each of these three types of content. However, limiting the action space to just these three actions will fail to uncover meaningful differences in

commenting behavior, the most frequently generated type of content. We subdivide the commenting action by first distinguishing between comments that occur on questions from comments that occur on answers, and then further dividing these based on the original poster of the parent content further up in the content tree. Concretely, we arrive at six separate commenting action types: commenting on my own question (comment-mq), commenting on others' questions (comment-oq), commenting on my answer to my question (comment-ma-mq), commenting on my answer to others' questions (comment-ma-oq), commenting on others' answers to my question (comment-oa-mq), and finally commenting on others' answers to others' questions (comment-oa-oq). Similarly, we can subdivide the answering action into answering my own question (answer-mq) and answering others' questions (answer-oq).

While creation actions are arguably the most important actions to consider for modeling user behavior with respect to the generation of content, it is also important to consider the role that editors play within the communities. We define four types of edit actions: editing my question (edit-mq), editing others' questions (edit-oq), editing my answer (edit-ma), and editing others' answers (edit-oa). We also include two actions related to moderation (the closing, locking, deleting, moving, etc. of posts) on Stack Exchange with two actions: voting for moderation activity (mod-vote) and the actual application of moderation (mod-action).

Once we have defined our action space, we can then begin the session segmentation process. We start with a chronologically ordered list of all of the actions from the action space taken within a community, and then partition this list into separate action lists associated with each individual user. Then, we define a session as a contiguous chunk of a user's action list such that the gap between consecutive actions is less than six hours to roughly capture a day's worth of activity per session. The collection of all of these sessions, grouped by community, serves as the MDMM's input.

We further decompose the community session lists by segmenting them into month-long chunks to enable temporal analysis of the behavior compositions over time for our communities. We define the "birth" of a community as the timestamp of the very first action taken in any user session associated with it, and then use that as the reference point for constructing the monthly session lists. This gives us 49,768,660 user sessions across 9117 community-month pairs.

### 6.4.2 Analysis of the Discovered Roles

We start our analysis by examining the usefulness of the discovered roles $\phi_{1:K}$. Because the number of roles, $K$, is a hyperparameter of our model, it must be chosen in advance of our investigation into the roles. Our MaxMinKL heuristic suggests a value of $K = 5$ for our

Table 6.2: The MaxMinKL heuristic for the MDMM behavior model applied to the Stack Exchange dataset. Notice the substantial drop when moving from $K = 5$ to $K = 6$, indicating redundancy obtained in the set of new roles. This matches our own visual inspection of the role distributions; hence, we choose $K = 5$ for the remaining experiments.

| Transition | MaxMinKL |
|---|---|
| $2 \to 3$ | 2.95 |
| $3 \to 4$ | 3.35 |
| $4 \to 5$ | 3.30 |
| $5 \to 6$ | 1.73 |
| $6 \to 7$ | 1.75 |

dataset (see Table 6.2 for the scores for each transition), and manual inspection also indicated role redundancies found at $K > 5$. We ran our model on an Intel(R) Core(TM) i7-5820K CPU, and each iteration takes approximately 20 seconds. We ran the model for 100 total iterations, as we found the output stopped changing appreciably after about 40 iterations. Each role we discovered at $K = 5$ is depicted in Figure 6.2, along with labels constructed from our own interpretation of the roles. These results directly help us understand what the "typical" roles assumed by users are in CQA communities.

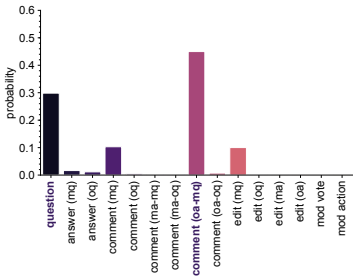
**"Eager asker"** (Figure 6.2a): Users exhibiting this role tend to ask questions, and comment on others' answers to their questions.

**"Careful asker"** (Figure 6.2d): While both this role and the previous role tend to ask questions in the same proportion within a session, a "careful asker" tends to comment a lot in discussions on their own question rather than on answers to their question, and they also have a much higher chance of updating their question when compared to the "eager asker" role. This subtle difference in asking behavior types would be lost if we had not carefully subdivided the commenting action by considering both the type and originator of the parent content of the comment.

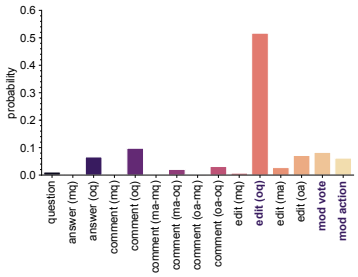
**"Answerer"** (Figure 6.2c): For the most part, this reflects a user that is concerned about their own answers. They provide their answers, they comment on their answers, and they update their answers. They may also seek clarification on a question by engaging in the discussion on that question, but not nearly as much as the next role.

**"Clarifier"** (Figure 6.2e): Users exhibiting this role tend to engage in the discussion on a question (by far their most frequent action) before answering; they also tend to comment on others' answers to others' questions more than any other role.

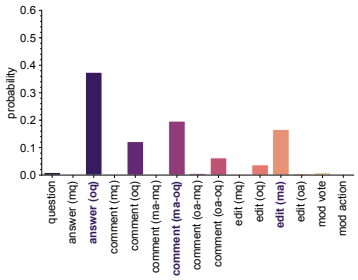
**"Editor/moderator"** (Figure 6.2b): This role captures nearly all of the observed moderation activity, and the most common action is to update someone else's question.



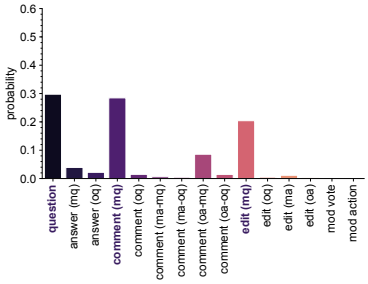

(a) An asker role we call "eager asker." In comparison to Figure 6.2d, we see that when a user exhibiting this role chooses to comment, they tend to comment on others' answers to their own question.



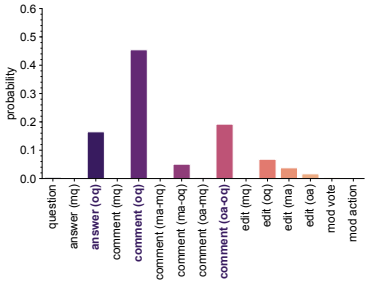

(b) An "editor/moderator" role. This role is the only role that exhibits moderation behavior, and we see that the vast majority of actions a user takes when exhibiting this role are to update others' questions.




(c) An "answerer" role. The majority of the time, users exhibiting this role answer others' questions, engage in discussion on their provided answers, and update their answers accordingly.




(d) An asker role we call "careful asker." In comparison to Figure 6.2a, we see that when a user exhibiting this role chooses to comment, they tend to do so on their own question. This may indicate engagement with users exhibiting the "clarifier" role (see Figure 6.2e) to improve the question's quality before obtaining an answer.




(e) A "clarifier" role. The majority of this user's activity is centered on commenting behavior, and is predominately engaging in discussion on others' questions. This is likely a result of this type of user engaging with others exhibiting the "careful asker" role (see Figure 6.2d) in order to clarify the question before providing an answer.

Figure 6.2: The role distributions discovered by our MDMM behavior model fit to 161 Stack Exchange communities. The labels given to these roles reflect our own interpretation of the role and are given here to make disambiguating the roles easier in the text. The MDMM behavior model can uncover subtle distinctions in asking behavior (see Figure 6.2a vs Figure 6.2d) and answering behavior (see Figure 6.2c vs Figure 6.2e).

While it might not be very surprising to see two distinct roles corresponding to primarily asking questions and primarily answering questions, the model goes beyond discovering such "obvious" roles to provide further fine distinction of interesting variations of roles for both question askers and question answerers, which may not be easy to discover otherwise by simply manually examining their behaviors. Our MDMM behavior model is able to

uncover these meaningful user behavior roles, including those with subtle differences, in a completely unsupervised way directly from log data once given an appropriate action space. Note that due to the generality of the MDMM model, we can easily refine action categories to potentially discover even finer-grained variations of user roles than what we have seen here—in this way, our model naturally supports multi-resolution analysis of user behavior.

### 6.4.3 Analysis of Behavior Compositions

The MDMM behavior model also outputs role proportions $\theta_i$ for each community in the dataset. These proportions provide an informative summary of the composition of behaviors in a community, i.e., a behavior profile. This profile provides a representation of a community that can be further analyzed, as we will discuss in this section.

We start with the following question: are there systematic differences in role proportions between groups of communities in our dataset? To answer this question, we grouped each community in the Stack Exchange dataset using the taxonomy provided by Stack Exchange itself[5]: (1) Technology, (2) Culture/Recreation, (3) Life/Arts, (4) Science, (5) Professional, and (6) Business. To allow for a "warm-up" period for the community and to eliminate the issue of noisy proportion vectors arising due to data sparsity during community launch, we discard the first 12 months of role proportion data for each community. We then only consider communities that have at least 12 months of data beyond that warm-up period to allow for computing an average proportion vector to represent the community over at least one year. After filtering, the "Professional" and "Business" groups have only five and four communities, respectively, so we consider only the four larger groups. "Technology" had 52 communities, "Culture/Recreation" had 36, "Life/Arts" had 20, and "Science" had 17. We show the group memberships in Table 6.3.

These four groups' role proportions are visualized in Figure 6.3. Visually, we can see a number of differences. First, the "eager asker" role is more prominent in the "Technology" group than all three others. Both the "Technology" and "Science" groups have higher prominence of the "careful asker" role when compared against "Culture/Recreation" and "Life/Arts". We can also see that the "clarifier" role is diminished in the "Technology" compared to the others.

There are also notable commonalities between groups. The "Culture/Recreation" and "Life/Arts" groups are quite similar across nearly all of the roles. The "editor/moderator" role prevalence is similar across all of the groups (with only a slight increase present for the "Culture/Recreation" group). "Answerer" prevalence is similar across all of the groups

[5] `https://stackexchange.com/sites`

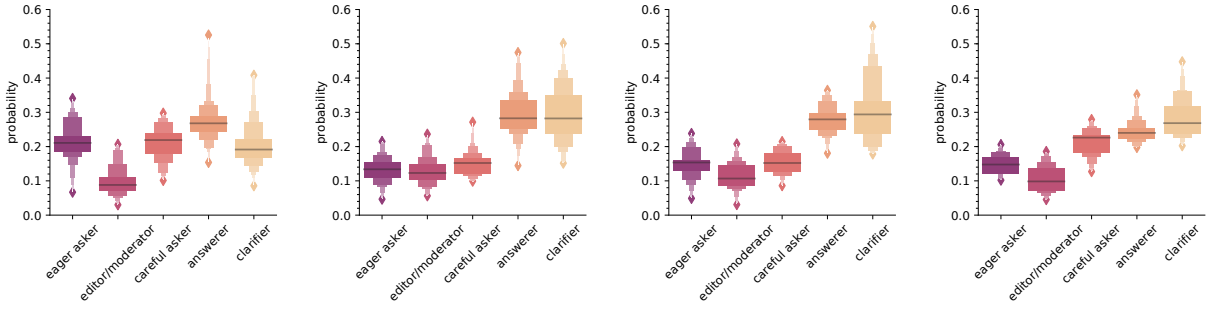
Table 6.3: Communities belonging to each of the four groups we consider from Stack Exchange's own taxonomy.

| Group | Members |
|---|---|
| Technology | android, apple, arduino, askubuntu, bitcoin, blender, codegolf, codereview, craftcms, crypto, datascience, dba, drupal, dsp, ebooks, electronics, emacs, expressionengine, gamedev, gis, ja.stackoverflow, joomla, magento, mathematica, networkengineering, opendata, programmers, pt.stackoverflow, raspberrypi, reverseengineering, robotics, ru.stackoverflow, salesforce, security, serverfault, sharepoint, softwarerecs, sound, space, sqa, stackapps, stackoverflow, superuser, tex, tor, tridion, unix, ux, webapps, webmasters, windowsphone, wordpress |
| Culture/Recreation | anime, beer, bicycles, boardgames, bricks, buddhism, chess, chinese, christianity, ell, english, french, gaming, german, ham, hermeneutics, hinduism, history, homebrew, islam, italian, japanese, judaism, martialarts, mechanics, outdoors, poker, politics, puzzling, rpg, rus, russian, skeptics, spanish, sports, travel |
| Life/Arts | academia, avp, cooking, diy, expatriates, fitness, gardening, genealogy, graphicdesign, lifehacks, money, movies, music, parenting, pets, photo, productivity, scifi, sustainability, worldbuilding |
| Science | astronomy, biology, chemistry, cogsci, cs, cstheory, earthscience, economics, hsm, linguistics, math, matheducators, mathoverflow.net, philosophy, physics, scicomp, stats |

(where the reduction in variance in "Life/Arts" and "Science" likely attributable to there being fewer communities in those groups).

To quantify the statistical significance of the above observations, we use a Kruskal-Wallis $H$ test [196] to perform a one-way ANOVA test to determine the existence of a difference between a single role proportion across all four groups, for each role proportion. Then, if a statistically significant difference between the groups is reported, we use a post-hoc Conover-Iman test [197] to determine which of the groups exhibit statistically significant differences in that role proportion. To correct for multiple testing in both cases, we use the Holm-Bonferroni method [198] to correct the $p$-values. We report our findings in Table 6.4. On the whole, we see that the "Technology" group differs strongly from the other three groups in terms of its proportion of "eager asker" (where it is higher) and "clarifier" roles (where it is lower). We also see that the "careful asker" role is more prominent in the communities from the "Technology" and "Science" groups and less prominent in the "Culture/Recreation" and "Life/Arts" groups. This suggests that the more technical communities in "Technology" and "Science" require more discussion around questions than the less technical communities of "Culture/Recreation" and "Life/Arts".

Thus, we have demonstrated the utility of using the MDMM behavior model for understanding differences in user behavior across communities. This is easily facilitated because it learns a role proportion vector $\theta_{1:N}$ that, by design, can be readily interpreted in the context




(a) Technology (b) Culture/Recreation (c) Life/Arts (d) Science

Figure 6.3: Letter-value plots of the role proportion vectors for the four largest Stack Exchange groups after filtering communities with less than 12 months of data after filtering a start-up period of 12 months.

of the discovered roles $\phi_{1:K}$.

### 6.4.4 Behavior Compositions and their Relationship to Community Success

As another example of what one can learn by studying the community role compositions that can be discovered by the MDMM user behavior model, we now ask the following question: how does the proportion of roles within a community relate to its success? In order to explore this, we first need to be able to define what we mean by "success" in a CQA community. We have taken a content-focused approach to understanding behavior, so we also choose to define the success of a community in terms of its content generation. Borrowing from Dev et al.[199], we have the following metrics: (1) the ratio of the number of answers $N_a$ to the number of questions $N_q$, which is a reflection of the ability of a community to cope with question load; (2) the percentage of questions that receive an answer; (3) the percentage of questions that receive an "accepted" answer[6], which reflects the community's ability to provide high-quality answers to new questions; and finally (4) the average time before the arrival of the first answer[7], which measures the timeliness of the community's answering capabilities.

Each of these metrics can be computed for each monthly snapshot of a community (by considering the questions that are asked within that time period). Then, we can average the value for a metric across all of the months of a community to obtain an overall score for that metric for that community. We again only consider the communities that, after dropping 12

[6]On Stack Exchange, the original poster of a question can designate one of the answers provided as being "correct" by "accepting" that answer.

[7]We compute this only for questions that did receive an answer.

Table 6.4: Statistical significance tests for differences in role proportions across the four groups. All $p$-values are adjusted using the Holm-Bonferroni method. Shown are only those tests that are statistically significant at a threshold of 0.05. We notice strongly significant differences ($p < 1 \times 10^{-5}$) in role proportions for the "eager asker", "careful asker", and "clarifier" roles.

| Role | $p$-value | Group Pair | $p$-value |
|---|---|---|---|
| eag. ask. | $3.87 \times 10^{-11}$ | cult. tech. | $1.49 \times 10^{-14}$ |
| | | life vs. tech. | $5.41 \times 10^{-7}$ |
| | | sci. vs. tech. | $6.63 \times 10^{-7}$ |
| edit/mod. | $1.10 \times 10^{-2}$ | cult. vs. tech. | $2.40 \times 10^{-3}$ |
| care. ask. | $7.53 \times 10^{-9}$ | cult. vs. sci. | $3.00 \times 10^{-6}$ |
| | | cult. vs. tech. | $3.41 \times 10^{-9}$ |
| | | life vs. sci. | $5.80 \times 10^{-5}$ |
| | | life vs. tech. | $3.07 \times 10^{-6}$ |
| answerer | $1.10 \times 10^{-2}$ | cult. vs. sci. | $4.44 \times 10^{-3}$ |
| clarifier | $4.41 \times 10^{-8}$ | cult. vs. tech. | $5.22 \times 10^{-8}$ |
| | | life vs. tech | $1.69 \times 10^{-6}$ |
| | | sci. vs. tech | $2.30 \times 10^{-5}$ |

months of "warm-up" period data, have at least 12 months of data.

The results are visualized in Figure 6.4. While differences in these metrics are small, they are statistically significant (see Table 6.5). In particular, we notice that the "Culture/Recreation" and "Life/Arts" groups have a higher ratio of answers to questions (Figure 6.4a) and a higher fraction of answered questions (Figure 6.4b) when compared to the "Science" and "Technology" groups. These same pairs exhibit statistically significantly different proportions of the "careful asker" role.

This provides an interesting insight: groups of communities that have a higher propensity for the "careful asker" role exhibited *lower health metrics* across multiple measures. In fact, every pair of groups that exhibited a statistically significant difference in this role proportion also had statistically significant differences present in at least two metrics (with one pair with three and another with four). While we cannot say whether this correlation is causal, this opens the door for more studies into impact of the "careful asker" profile on community health—a question we could not have raised without first having a tool like the MDMM behavior model to aid our efforts to understand user behavior.

Furthermore, notice that groups that do *not* exhibit differences in their behavior profiles (namely "Culture/Recreation" and "Life/Arts") also do not exhibit differences in any of our

Table 6.5: Statistical significance tests for differences in health metrics across the four groups. All $p$-values are adjusted using the Holm-Bonferroni method. Shown are only those tests that are statistically significant at a threshold of 0.05. We note that, with a single exception ("Science" vs "Technology"), when there is a statistically significant difference in role proportions, there is a statistically significant difference in at least one of the four health metrics we explore. Similarly, groups that do not have different role proportions ("Culture/Recreation" and "Life/Arts") do not have significant differences in health metrics.

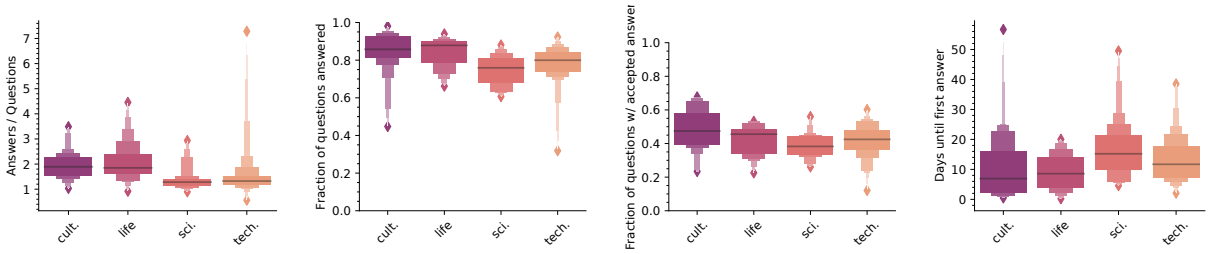
| Metric | $p$-value | Group Pair | $p$-value |
|---|---|---|---|
| $N_a/N_q$ | $7.08 \times 10^{-7}$ | cult. vs. sci. | $4.26 \times 10^{-5}$ |
| | | cult. vs. tech. | $3.51 \times 10^{-6}$ |
| | | life vs. sci. | $1.20 \times 10^{-4}$ |
| | | life vs. tech. | $6.50 \times 10^{-5}$ |
| % ans. | $6.34 \times 10^{-5}$ | cult. vs. sci. | $7.44 \times 10^{-5}$ |
| | | cult. vs. tech. | $4.12 \times 10^{-4}$ |
| | | life. vs. sci. | $3.16 \times 10^{-3}$ |
| | | life. vs. tech. | $3.36 \times 10^{-2}$ |
| % acc. ans. | $1.08 \times 10^{-2}$ | cult. vs. sci. | $6.68 \times 10^{-3}$ |
| Resp. time | $1.08 \times 10^{-2}$ | cult. vs. sci. | $2.31 \times 10^{-2}$ |
| | | cult. vs. tech. | $3.34 \times 10^{-2}$ |

four health metrics.

### 6.4.5 Evolution of Behavior Composition

The questions we have explored so far have focused mainly on static snapshots of the CQA communities in our dataset. However, these communities do not exist in a vacuum—they continually evolve over time as they acquire new users and address new topics. How can we understand how community behavior changes over time as these communities grow and evolve? Here, we explore one potential solution using the MDMM behavior model as yet another example application.

Because we segmented the user sessions by month for each community, we have a role proportion associated with each (community, month) pair. With this information in hand, we can then plot a collection of time-series for each community by considering the role proportions for each individual role over the life of the community. This plot can allow us to understand how role proportions fluctuate as the community evolves. In Figure 6.5, we show the evolution of the top three oldest communities belonging to the "Technology" and "Culture/Recreation" groups, respectively. We start plotting the time series at the month



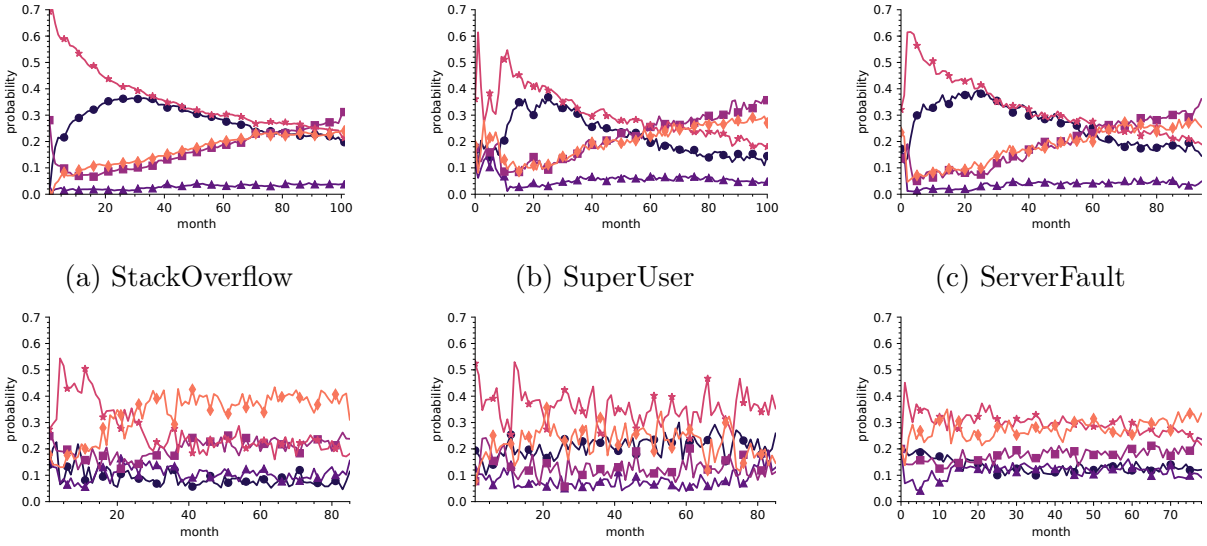

(a) $N_a/N_q$ (b) Frac. of $N_q$ with ans. (c) Frac. of $N_q$ with accepted ans. (d) Avg. days until first answer

Figure 6.4: Health metrics for each of the four groups of Stack Exchanges considered in Section 6.4.3. Differences are small but statistically significant (see Table 6.5). $N_a/N_q$ is higher for "Culture/Recreation" and "Life/Arts" than for "Science" and "Technology". Similarly, "Culture/Recreation" and "Life/Arts" enjoy a higher fraction of answered questions compared to "Science" and "Technology", and also have faster average response times (though only "Culture/Recreation" statistically significantly so). "Culture/Recreation" also has a higher fraction of questions with an accepted answer compared to the other three groups.

when the community first has at least 100 browsing sessions.

We can see a few trends occurring. First, we can see a common trend in Figure 6.5a–6.5c, where the proportions for the "eager asker" role grow, reach a peak within the first quarter or so of the community's life, and then begin a steady decline over time. We also notice that the "careful asker" and "clarifier" roles tend to increase steadily over time, nearly in tandem. Second, we can see in Figure 6.5d–6.5f that the role proportions tend to be more consistent over time for members of the "Culture/Recreation" group than for "Technology". Note, however, that the exact composition that is remaining stable varies between the communities. That is to say, communities in "Culture/Recreation" appear to be more stable relative to themselves over time, but exhibit variation in what that stability looks like.

Why does this behavior shift happen in "Technology" while "Culture/Recreation" communities remain more stable? While we cannot yet provide an answer to this question, we note that without first being able to see that this kind of behavior evolution is even taking place (which requires a model like our MDMM behavior model), we could not even begin to ask such a question. This shows that the MDMM behavior model opens new interesting research directions in understanding user behavior in ways we were not able to before.




(a) StackOverflow (b) SuperUser (c) ServerFault

(d) Judaism (e) Homebrew (f) Gaming

Figure 6.5: Role proportions over time for the three oldest communities belonging to the "Technology" and "Culture/Recreation" groups. (a)–(c) belong to the "Technology" group, and (d)–(f) belong to the "Culture/Recreation" group. We can see a common trend in (a)–(c) where the proportion of the "eager asker" role grows until it peaks, and then declines as the community ages. The "clarifier" and "careful asker" roles increase over time, almost in tandem in this group. However, in (d)–(f) we see that that communities belonging to "Culture/Recreation" tend to have role proportions that remain more consistent over time (in that they do not demonstrate long-term trends.)

## 6.5 DISCUSSION AND LIMITATIONS

The goal of this work is to contribute a new and general tool for role discovery and analysis of community role compositions. There are two key ideas in the design of the proposed model. The first is the formalization of a *shared* set of user roles, distributions over user actions, across communities. This is an expressive representation of a user role as the distribution can vary to capture subtle differences between user roles while also allowing us to discover user roles empirically from the data using sound statistical principles. The second is the direct modeling of the composition of user roles in a CQA community with another distribution over the user roles. This second distribution provides a general and flexible way to model variations in the composition of user roles that may exist in different communities, and again allows us to use statistical inference to discover each community's role composition.

The use of a generative model over user actions to discover user roles and community role compositions is advantageous in that it allows the model to be very general and applied in a variety of different analysis scenarios without requiring hand-crafted features to be defined in

order to describe user roles. On the other hand, the use of a generative model is not without some cost. Because statistical inference of such a model is intractable, we must resort to approximate posterior inference methods. In this paper, we have used Gibbs sampling to approximate the posterior, but this comes with some risk—it is difficult to determine whether the sampler has actually converged to the true posterior, despite there being a theoretical guarantee that it will do so given enough time. Had we instead opted for a different inference method like variational inference which instead optimizes a variational lower bound, we trade the convergence question for a question about the quality of the solution found by the optimization because the variational lower bound is highly non-convex. In practice, we can attempt to mitigate these concerns via multiple runs of the sampler (or multiple randomly initialized optimizations for variational inference)—we found multiple runs of the model all converged to nearly identical solutions.

Because the model does not impose an action set upon the user, they are free to specify a different action set for different analysis purposes. This again makes the model quite flexible, but also requires some up front work to define an appropriate action set for the model. Feeding the model with less meaningful actions can lead to the output of less meaningful role patterns. Fortunately, in the case of CQA communities, defining an action set based on the content hierarchy and content ownership semantics is a reasonable choice that should lead to interpretable roles as demonstrated here for Stack Exchange. However, a user *does* need to manually interpret the role distributions $\phi_{1:K}$ discovered by the model.

Finally, our model makes a strong assumption that a user only performs one role in a given session. While this assumption is valid in most cases, there are situations where users potentially perform more than one role in a given browsing session. In these cases, the model will incorrectly conflate these two roles and this will contribute some "noise actions" to that role.

## 6.6 CONCLUSION AND FUTURE WORK

Computational analysis of user roles on CQA platforms is important not only for the understanding of users in such a new social network environment, but also for improving their efficiency and utility. To this end, we proposed a general probabilistic model for discovering and analyzing action-based roles on CQA platforms. The generative model assumes that the observed user actions in a single session are samples drawn from the same, but unknown, action distribution (the role). Individual communities are modeled as mixtures over these role distributions, allowing for cross-community analysis. Through a comprehensive experiment on all 161 non-meta communities on the Stack Exchange CQA

platform, we demonstrated that our model is indeed useful for understanding user behavior on these platforms. We were able to show interesting distinctions in asking and answering behavior on the platform are captured through our roles, that different groups of communities exhibit statistically significant differences in role composition, and those communities also exhibit statistically significant differences in a variety of health measures. Finally, we were also able to uncover two clear and distinct trends of role compositions over time between the "Technology" and "Culture/Recreation" groups on Stack Exchange.

The proposed model is very general and does not require labeled data for training. It can thus be applied to analyze any CQA platform immediately. Since the definition of actions is outside the model, analysts can vary the granularity of actions as needed; this flexibility allows for multi-resolution analysis of user actions, behavior, and roles. An interesting future work is to fully exploit this flexibility to further analyze roles with even more refined actions on CQA platforms as well as to apply the model to other social networks. Another interesting future direction is to develop tools based on this model for monitoring the "well-being" of those CQA platforms and helping the community managers to improve the utility and efficiency of a community so as to maximize the utility of all the CQA communities.

# CHAPTER 7: CONCLUSION

This chapter provides a summary of the four empirical studies conducted in this thesis. It then discusses avenues of future work for studying platform sustainability.

## 7.1 THESIS SUMMARY

In Chapter 3, we interpret community question answering websites in Stack Exchange as knowledge markets. Our objective is to model content generation in these markets and reason about their sustainability. In any knowledge market, users generate different types of content such as questions, answers, and comments. Content has dependency; for example, comments depend on questions and answers. We want to capture these two factors—user participation and content dependency—in our content generation model. To this end, we model content generation in knowledge markets using production functions, which is a natural choice to model output in a market. These functions involve two components: a basis function and an interaction type. A basis function captures the relationship between individual input and output. We consider three possible relationships: exponential, power, and sigmoid. Interaction types capture how different inputs interact to produce an output. We consider four possible interactions: essential, interactive essential, antagonistic, and substitutable. Considering these choices of basis function and interaction types, we found the power basis and interactive essential interaction provided the best fit to our data. In the broader field of economics, this is known as the Cobb-Douglas model. The Cobb-Douglas model gives us three critical insights: stable core, size-dependent distribution, and diseconomies of scale. Briefly, in Stack Exchange websites, there is a stable core of users who substantially contribute to the websites for a long period. In most websites, the size of this stable core does not increase with the number of users. This discrepancy results in a size-dependent activity distribution—the expected user behavior changes with community size. As a result, these websites exhibit diseconomies of scale; for example, the fraction of answered questions declines with the increase in number of users.

In Chapter 4, we examine biases in social judgment, specifically in the form of votes. We concentrate on votes because they are the primary social feedback in content-based platforms, such as Stack Exchange, Reddit, and Quora. Votes are crucial as they form the basis of most platform's ranking and recommendation services and their rewarding policies. Motivated by such powerful implications of votes, we ask the following questions. What are the factors that affect votes on content? Which of these factors may lead to biases?

What is the degree to which biasing factors affect the votes on content? Our goal is to answer these causal questions in an observational setup. Prior research suggests several factors that may affect votes on content. Some usual suspects include: 1) content topic, 2) content quality, 3) presentation order, 4) social influence, and 5) author reputation. Among these five factors, the first two factors (content topic and content quality) are the content's properties. The remaining factors are impression signals that appear in the user interface and act as sources of biases. Our goal is to quantify, in an observational setting, the degree of voter biases in online platforms. Specifically, what are the causal effects of different impression signals—such as the reputation of the contributing user, aggregate vote thus far, and position of content—on the observed votes on content? Our observational setup's main challenge arises in the form of unobserved confounders: unobserved factors that may explain the association between the impression signals and observed votes. In this work, we adopt the instrumental variable (IV) method to eliminate unobserved confounders' impact. Using the popularity of the past questions responded to by the answerer as the instrument, we show that badges significantly influence votes, at times twice as much as the effect suggested by the existing research. Further, using the timeliness of the answer as the instrument, we show that the answer position has a significant influence on votes, almost twice as much as the effect suggested by the existing research.

In Chapter 5, we develop a counterfactual framework for community voting. The framework allows us to study the effects of alternative voting behaviors on user retention. We refer to the existing voting norm as the control norm, whose voting outcomes can be observed from log data. Our goal is to evaluate the outcomes of alternative voting norms that were never observed. To perform this counterfactual analysis, we must understand the probability of different voting outcomes under each norm. To this end, we examine several voting norms that express how the members of a community can vote based upon different voting criteria, such as length and readability. Given the control voting norm and an alternative voting norm, we propose a counterfactual framework to evaluate the alternative voting norm's outcomes in terms of user retention. We first develop a propensity model for quantifying the probabilities of different voting outcomes under each voting norm. We then define a utility model for quantifying the derived utility of users from the votes they acquire. Finally, we develop a counterfactual model for reasoning about user retention under different voting norms. We use this framework to analyze voting behavior in Stack Exchange sites. Our main findings are that had the community members voted based upon the length, readability, or objectivity norm, the platform operator would have observed *higher* retention.

In Chapter 6, we examine user roles and the composition of communities based on these roles. Motivated by the importance of user roles in platform sustainability, we ask the fol-

lowing research questions. How do user roles affect the health of content-based platforms? How can we discover user roles in a content-based platform? How do we extract the composition of communities in terms of roles? To answer these questions, we develop a general framework for discovering user roles and the composition of communities in terms of these roles. The design of our model is motivated by our goal of discovering interpretable descriptions of functional roles played by users, as well as a representation for each community as a mixture of these user roles. Specifically, we represent the roles as probability distributions that describe the likelihood of taking individual actions when a user assumes a particular functional role in one session. To automatically discover user roles from data, we appeal to the general technique of probabilistic graphical models. A comprehensive experiment on all 161 non-meta communities on the Stack Exchange network demonstrates that our model can shed light on user roles' effect on platform sustainability. First, we show interesting distinctions in question-asking behavior on Stack Exchange and answering behavior. Second, we find statistically significant differences in behavior compositions across topical groups of communities on Stack Exchange. Those groups that have statistically significant differences in health metrics suggest a relationship between behavior composition and health.

## 7.2 FUTURE WORK

How do platforms fail? Why do they fail? When do they fail?

Drawing from the literature on community behavior modeling, causal inference, and to a less extent, survival analysis, we developed models that answer these questions. The intersection of these three areas is a fertile ground for future research on platform sustainability. Some promising directions include modeling community composition, learning user representations for explaining survival, and unbiased learning to rank for content-based platforms.

**Studying the Composition of User Community.** We recognized an interesting fact about the content-based online platforms, that many of these platforms have existed for more than a decade: Stack Exchange (11 years), Reddit (14 years), and Quora (10 years). Since the inception of the platforms to the present day, individuals have joined and abandoned the platforms at different times, forming *stage-structured populations*—in which individuals vary regarding their age (time spent in the platform), experience (active participation in content creation), and expertise (quality of created content) in the platform. These stage-structured populations provide us a unique opportunity to study the composition and evolution of the underlying user community. Studying the age, experience, and expertise composition of the user community is essential. Notably, the distribution of age in a user community portrays a holistic picture of the retention of individuals in the accommodating platform. Similarly,

the distribution of experience in the community depicts a detailed view of the engagement of individuals. Further, the distribution of expertise in the community provides a circumstantial view of the quality of content in the platform. Overall, the composition of a community has serious implications for retention, engagement, and content quality in the accommodating platform. Therefore, studying the composition of crowds is crucial for understanding their evolution, predicting the successes and failures of platforms, and identifying platforms that will sustain for a long time.

**Learning Representations for Survival.** Owing to the recent success of neural representation learning methods such as word2vec [200] and doc2vec [201], representation learning has become a popular research area. A plethora of recent works has developed methods to learn representations for various machine learning tasks, such as item recommendation [202] and churn prediction [154]. The successful adoption of representation learning in various domains motivates us to ask: can we learn representations of users for survival analysis? Prior work has typically relied on manual features or embeddings learned for other tasks to perform survival analysis. In contrast, we propose to construct user embeddings specifically for predicting the survival (i.e., time to leave platform) of users. By constructing such embeddings, one can identify the vulnerable user groups. Also, by analyzing these embeddings, it may be possible to identify some common root causes behind users leaving the platform and design interventions to retain users.

**Unbiased Learning to Rank for Content-Based Platforms.** In Chapter 4, we showed that in content-based platforms, community feedback in the form of votes suffers from different voter biases: reputation bias, social influence bias, and position bias. These voter biases impose a significant challenge in using votes for downstream applications, such as search ranking and content recommendation in content-based platforms. For instance, position bias in search rankings strongly influences how many views a result receives. For this reason, directly using votes as a training signal in traditional learning to rank (LTR) methods yield sub-optimal results. The problem of unbiased learning to rank from biased data is well-studied in the context of position bias in web search [123, 124, 165]. However, the problem has not been explored in the context of content-based platforms, where other forms of biases such as reputation bias and social influence bias also play a role in biasing the training data. Developing unbiased learning to rank mechanism for content-based platforms can prevent subsequent reinforcement of these biases. Preventing such reinforcement of biases has the potential to improve platform sustainability by providing new users a fair shot at obtaining a higher reputation.

There are other new threads of research to understand the sustainability of content-based platforms better.

**Reasoning about Information Markets.** Content-based platforms and the underlying information markets are core institutions that underpin today's attention economy, as envisioned by Herbert A. Simon (1971) — "In an information-rich world, the wealth of information means a dearth of something else: a scarcity of whatever it is that information consumes. What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention and a need to allocate that attention efficiently among the overabundance of information sources that might consume it." Content production and consumption are core processes that drive these markets and the encompassing attention economy. Yet, our understanding of these processes and their relationship is limited. For instance, since users in content-based platforms have limited time and they are distributing their time across production and consumption activities (e.g., In Facebook, a user distributes daily time between writing own posts and comments, and reading posts and comments from friends), what can be inferred about the relationship between production and consumption? Further, if users invest most of their time in consumption activities (say reading posts and comments), or conversely, in production activities (say generating posts and comments), how does it affect the platform's sustainability, specifically in terms of future production and consumption? Answering these questions is crucial for developing a deeper understanding of the economic views of platforms and reason about their sustainability.

**Studying the Impact of Anti-Social Behaviors.** Anti-social behaviors, such as trolling, harassment, bullying, abuse, and hate speech, have become a significant issue in content-based platforms. Users who face these issues are more likely to abandon the platform. There has been considerable interest in studying anti-social behaviors in platforms like Reddit [203, 204] and the role of moderators in preventing these behaviors [205]. Yet, the relationship between anti-social behavior and platform sustainability is still under-explored. For instance, while platforms continue to dedicate more resources to preventing anti-social behaviors, the amount of abusive content on the Internet is still rising. We do not fully understand why despite the increased moderation, anti-social behaviors continue to rise. We also do not understand the role of social norms in mitigating the effects of anti-social behaviors. For instance, how does community support affect the users who face anti-social behaviors to stay on the platform? Revealing the impact of anti-social behaviors is crucial for engineering norms to retain users in the platform.

## REFERENCES


[1] L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman, "Knowledge sharing and yahoo answers: Everyone knows something," in *Proceedings of the 17th International Conference on World Wide Web*, ser. WWW '08. New York, NY, USA: ACM, 2008. [Online]. Available: http://doi.acm.org/10.1145/1367497.1367587 pp. 665–674.

[2] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec, "Discovering value from community activity on focused question answering sites: A case study of stack overflow," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '12. New York, NY, USA: ACM, 2012. [Online]. Available: http://doi.acm.org/10.1145/2339530.2339665 pp. 850–858.

[3] G. Wang, K. Gill, M. Mohanlal, H. Zheng, and B. Y. Zhao, "Wisdom in the social crowd: An analysis of quora," in *Proceedings of the 22Nd International Conference on World Wide Web*, ser. WWW '13. ACM, 2013. [Online]. Available: http://doi.acm.org/10.1145/2488388.2488506 pp. 1341–1352.

[4] R. S. Geiger and A. Halfaker, "Using edit sessions to measure participation in wikipedia," in *Proceedings of the 2013 conference on Computer supported cooperative work*, 2013, pp. 861–870.

[5] A. Halfaker, R. S. Geiger, J. T. Morgan, and J. Riedl, "The rise and decline of an open collaboration system: How wikipedia's reaction to popularity is causing its decline," *American Behavioral Scientist*, vol. 57, no. 5, pp. 664–688, 2013.

[6] S. M. Nasehi, J. Sillito, F. Maurer, and C. Burns, "What makes a good code example?: A study of programming q&a in stackoverflow," in *2012 28th IEEE International Conference on Software Maintenance (ICSM)*. IEEE, 2012, pp. 25–34.

[7] N. Schrading, C. O. Alm, R. Ptucha, and C. Homan, "An analysis of domestic abuse discourse on reddit," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 2577–2583.

[8] A. N. Medvedev, R. Lambiotte, and J.-C. Delvenne, "The anatomy of reddit: An overview of academic research," in *Dynamics on and of Complex Networks*. Springer, 2017, pp. 183–204.

[9] G. Marvin, "Quora introduces broad targeting, says audience hits 300 million monthly users," https://marketingland.com/quora-introduces-broad-targeting-says-audience-hits-300-million-monthly-users-248261, 2018, online; accessed 5 April 2021.

[10] S. Perez, "Reddit's monthly active user base grew 30% to reach 430m in 2019," https://techcrunch.com/2019/12/04/reddits-monthly-active-user-base-grew-30-to-reach-430m-in-2019/, 2019, online; accessed 5 April 2021.

[11] D. Fullerton, "State of the stack 2019: A year in review," https://stackoverflow.blog/2019/01/18/state-of-the-stack-2019-a-year-in-review/, 2019, online; accessed 5 April 2021.

[12] I. Srba and M. Bielikova, "Why is stack overflow failing? preserving sustainability in community question answering," *IEEE Software*, vol. 33, no. 4, pp. 80–89, July 2016.

[13] M. Glenski and T. Weninger, "Rating effects on social news posts and comments," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 8, no. 6, p. 78, 2017.

[14] R. Kumar, J. Novak, and A. Tomkins, "Structure and evolution of online social networks," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '06. New York, NY, USA: ACM, 2006. [Online]. Available: http://doi.acm.org/10.1145/1150402.1150476 pp. 611–617.

[15] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan, "Group formation in large social networks: Membership, growth, and evolution," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '06. New York, NY, USA: ACM, 2006. [Online]. Available: http://doi.acm.org/10.1145/1150402.1150412 pp. 44–54.

[16] S. R. Kairam, D. J. Wang, and J. Leskovec, "The life and death of online groups: Predicting group growth and longevity," in *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, ser. WSDM '12. New York, NY, USA: ACM, 2012. [Online]. Available: http://doi.acm.org/10.1145/2124295.2124374 pp. 673–682.

[17] B. Ribeiro, "Modeling and predicting the growth and death of membership-based websites," in *Proceedings of the 23rd International Conference on World Wide Web*, ser. WWW '14. New York, NY, USA: ACM, 2014. [Online]. Available: http://doi.acm.org/10.1145/2566486.2567984 pp. 653–664.

[18] C. Zang, P. Cui, and C. Faloutsos, "Beyond sigmoids: The nettide model for social network growth, and its applications," in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: ACM, 2016. [Online]. Available: http://doi.acm.org/10.1145/2939672.2939825 pp. 2015–2024.

[19] L. Guo, E. Tan, S. Chen, X. Zhang, and Y. Zhao, "Analyzing patterns of user content generation in online social networks," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 369–378.

[20] S. Walk, D. Helic, F. Geigl, and M. Strohmaier, "Activity dynamics in collaboration networks," *ACM Trans. Web*, vol. 10, no. 2, pp. 11:1–11:32, May 2016. [Online]. Available: http://doi.acm.org/10.1145/2873060

[21] R. E. Kraut and A. T. Fiore, "The role of founders in building online groups," in *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, ser. CSCW '14. New York, NY, USA: ACM, 2014. [Online]. Available: http://doi.acm.org/10.1145/2531602.2531648 pp. 722–732.

[22] H. Zhu, R. E. Kraut, and A. Kittur, "The impact of membership overlap on the survival of online communities," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '14. New York, NY, USA: ACM, 2014. [Online]. Available: http://doi.acm.org/10.1145/2556288.2557213 pp. 281–290.

[23] H. Zhu, J. Chen, T. Matthews, A. Pal, H. Badenes, and R. E. Kraut, "Selecting an effective niche: An ecological view of the success of online communities," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '14. New York, NY, USA: ACM, 2014. [Online]. Available: http://doi.acm.org/10.1145/2556288.2557348 pp. 301–310.

[24] T. Kuśmierczyk and M. Gomez-Rodriguez, "On the causal effect of badges," 2018.

[25] A. Segal, K. Gal, E. Kamar, E. Horvitz, and G. Miller, "Optimizing interventions via offline policy evaluation: Studies in citizen science," 2018.

[26] D. Easley and A. Ghosh, "Incentives, gamification, and game theory: an economic approach to badge design," *ACM Transactions on Economics and Computation (TEAC)*, vol. 4, no. 3, p. 16, 2016.

[27] Y. Lv and T. Moscibroda, "Fair and resilient incentive tree mechanisms," *Distributed Computing*, vol. 29, no. 1, pp. 1–16, 2016.

[28] N. Immorlica, G. Stoddard, and V. Syrgkanis, "Social status and badge design," in *Proceedings of the 24th International Conference on World Wide Web*. ACM, 2015, pp. 473–483.

[29] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec, "Engaging with massive online courses," in *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 2014, pp. 687–698.

[30] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec, "Steering user behavior with badges," in *Proceedings of the 22nd International Conference on World Wide Web (WWW)*. ACM, 2013, pp. 95–106.

[31] L. Wu, J. A. Baggio, and M. A. Janssen, "The role of diverse strategies in sustainable knowledge production," *PLOS ONE*, vol. 11, no. 3, pp. 1–13, 03 2016.

[32] Y. Chen, T.-H. Ho, and Y.-m. Kim, "Knowledge market design: A field experiment at google answers," *Journal of Public Economic Theory*, vol. 12, no. 4, pp. 641–664, 2010.

[33] D. R. Raban, "The incentive structure in an online information market," *Journal of the American Society for Information Science and Technology*, vol. 59, no. 14, pp. 2284–2295, 2008.

[34] B. Edelman, "Earnings and ratings at google answers," *Economic Inquiry*, vol. 50, no. 2, pp. 309–320, 2012.

[35] G. Hsieh, R. Kraut, S. E. Hudson, and R. Weber, "Can markets help? applying market mechanisms to improve synchronous communication," in *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, 2008, pp. 535–544.

[36] G. Hsieh and S. Counts, "mimir: A market-based real-time question and answer service," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2009, pp. 769–778.

[37] G. Hsieh, R. E. Kraut, and S. E. Hudson, "Why pay? exploring how financial incentives are used for question & answer," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2010, pp. 305–314.

[38] G. Dror, D. Pelleg, O. Rokhlenko, and I. Szpektor, "Churn prediction in new users of yahoo! answers," in *Proceedings of the 21st International Conference on World Wide Web*, 2012, pp. 829–834.

[39] F. M. Harper, D. Raban, S. Rafaeli, and J. A. Konstan, "Predictors of answer quality in online q&a sites," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2008, pp. 865–874.

[40] D. Dearman and K. N. Truong, "Why users of yahoo! answers do not answer questions," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2010, pp. 329–332.

[41] Y. Ren, R. Kraut, and S. Kiesler, "Applying common identity and bond theory to design of online communities," *Organization studies*, vol. 28, no. 3, pp. 377–408, 2007.

[42] Y. Ren, F. M. Harper, S. Drenner, L. Terveen, S. Kiesler, J. Riedl, and R. E. Kraut, "Building member attachment in online communities: Applying theories of group identity and interpersonal bonds," *Mis Quarterly*, pp. 841–864, 2012.

[43] H. Zhu, R. Kraut, and A. Kittur, "Effectiveness of shared leadership in online communities," in *Proceedings of the ACM 2012 conference on computer supported cooperative work*, 2012, pp. 407–416.

[44] H. Zhu, R. Kraut, and A. Kittur, "Organizing without formal organization: group identification, goal setting and social modeling in directing online production," in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, 2012, pp. 935–944.

[45] L. Wu, J. A. Baggio, and M. A. Janssen, "The role of diverse strategies in sustainable knowledge production," *PLoS ONE*, 2016.

[46] M. Abufouda, "Community aliveness: Discovering interaction decay patterns in online social communities," *CoRR*, vol. abs/1707.04477, 2017. [Online]. Available: http://arxiv.org/abs/1707.04477

[47] L. Wu and J. Zhang, "Accelerating growth and size-dependent distribution of human online activities," *Phys. Rev. E*, vol. 84, p. 026113, Aug 2011. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.84.026113

[48] R. Kumar, Y. Lifshits, and A. Tomkins, "Evolution of two-sided markets," in *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, ser. WSDM '10. New York, NY, USA: ACM, 2010. [Online]. Available: http://doi.acm.org/10.1145/1718487.1718526 pp. 311–320.

[49] B. S. Butler, "Membership size, communication activity, and sustainability: A resource-based model of online social structures," *Information systems research*, vol. 12, no. 4, pp. 346–362, 2001.

[50] Z. Lin, N. Salehi, B. Yao, Y. Chen, and M. Bernstein, "Better when it was smaller? community content and behavior after massive growth," in *Proceedings of the 11th International AAAI Conference on Web and Social Media*, 2017.

[51] T. Yla, W. Ping, and J. Choi, "Which size matters? effects of crowd size on solution quality in big data q&a communities," in *Proceedings of the 11th International AAAI Conference on Web and Social Media*, 2017.

[52] A. Patil, J. Liu, and J. Gao, "Predicting group stability in online social networks," in *Proceedings of the 22Nd International Conference on World Wide Web*, ser. WWW '13. ACM, 2013. [Online]. Available: http://doi.acm.org/10.1145/2488388.2488477 pp. 1021–1030.

[53] D. Garcia, P. Mavrodiev, and F. Schweitzer, "Social resilience in online communities: The autopsy of friendster," in *Proceedings of the First ACM Conference on Online Social Networks*, ser. COSN '13. New York, NY, USA: ACM, 2013. [Online]. Available: http://doi.acm.org/10.1145/2512938.2512946 pp. 39–50.

[54] K. Kapoor, M. Sun, J. Srivastava, and T. Ye, "A hazard based approach to user return time prediction," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '14. New York, NY, USA: ACM, 2014. [Online]. Available: http://doi.acm.org/10.1145/2623330.2623348 pp. 1719–1728.

[55] K. Ellis, M. Goldszmidt, G. Lanckriet, N. Mishra, and O. Reingold, "Equality and social mobility in twitter discussion groups," in *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, ser. WSDM '16. New York, NY, USA: ACM, 2016. [Online]. Available: http://doi.acm.org/10.1145/2835776.2835814 pp. 523–532.

[56] Y. Chen, T.-H. HO, and Y.-M. KIM, "Knowledge market design: A field experiment at google answers," *Journal of Public Economic Theory*, vol. 12, no. 4, pp. 641–664, 2010.

[57] I. Srba and M. Bielikova, "A comprehensive survey and classification of approaches for community question answering," *ACM Trans. Web*, vol. 10, no. 3, pp. 18:1–18:63, Aug. 2016. [Online]. Available: http://doi.acm.org/10.1145/2934687

[58] J. Zhang, M. S. Ackerman, and L. Adamic, "Expertise networks in online communities: Structure and algorithms," in *Proceedings of the 16th International Conference on World Wide Web*, ser. WWW '07. New York, NY, USA: ACM, 2007. [Online]. Available: http://doi.acm.org/10.1145/1242572.1242603 pp. 221–230.

[59] L. Qiaoling and A. Eugene, "Modeling answerer behavior in collaborative question answering systems," in *Proceedings of the 33rd European conference on Advances in information retrieval*, 2011.

[60] P. Aditya, C. Shuo, and K. Joseph, A, "Evolution of experts in question answering communities," in *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*, 2012.

[61] B. V. Hanrahan, G. Convertino, and L. Nelson, "Modeling problem difficulty and expertise in stackoverflow," in *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work Companion*, ser. CSCW '12. New York, NY, USA: ACM, 2012. [Online]. Available: http://doi.acm.org/10.1145/2141512.2141550 pp. 91–94.

[62] U. Upadhyay, I. Valera, and M. Gomez-Rodriguez, "Uncovering the dynamics of crowdlearning and the value of knowledge," in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, ser. WSDM '17. ACM, 2017. [Online]. Available: http://doi.acm.org/10.1145/3018661.3018685 pp. 61–70.

[63] A. Furtado, N. Andrade, N. Oliveira, and F. Brasileiro, "Contributor profiles, their dynamics, and their importance in five q&a sites," in *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, ser. CSCW '13. ACM, 2013. [Online]. Available: http://doi.acm.org/10.1145/2441776.2441916 pp. 1237–1252.

[64] V. Kumar and N. Pedanekar, "Mining shapes of expertise in online social q&a communities," in *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*, ser. CSCW '16 Companion. New York, NY, USA: ACM, 2016. [Online]. Available: http://doi.acm.org/10.1145/2818052.2869096 pp. 317–320.

[65] R. Baeza-Yates and D. Saez-Trumper, "Wisdom of the crowd or wisdom of a few?: An analysis of users' content generation," in *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, ser. HT '15. ACM, 2015. [Online]. Available: http://doi.acm.org/10.1145/2700171.2791056 pp. 69–74.

[66] J. Yang, A. Bozzon, and G.-J. Houben, ""harnessing engagement for knowledge creation acceleration in collaborative q&a systems"," in *Proceedings of the 23rd International Conference on user Modeling, Adaptation, and Personalization*, ser. UMAP, F. Ricci, K. Bontcheva, O. Conlan, and S. Lawless, Eds., 2015, pp. 315–327.

[67] E. Ferrara, N. Alipourfard, K. Burghardt, C. Gopal, and K. Lerman, "Dynamics of content quality in collaborative knowledge production," in *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017.*, 2017, pp. 520–523.

[68] J. Stanford, "Economics for everyone: On-line glossary of terms & concepts," 2008.

[69] D. Fekedulegn, S. Mártín Pádraig Mac, and J. J. Colbert, "Parameter estimation of nonlinear models in forestry," *Finnish Society of Forest Science and the Finnish Forest Research Institute*, 1999.

[70] D. Tilman, "Resources: A graphical-mechanistic approach to competition and predation," *The American Naturalist*, vol. 116, no. 3, pp. 362–393, 1980. [Online]. Available: https://doi.org/10.1086/283633

[71] H. Hofmann, H. Wickham, and K. Kafadar, "Letter-value plots: Boxplots for large data," *Journal of Computational and Graphical Statistics*, vol. 26, no. 3, pp. 469–477, 2017.

[72] M. A. Branch, T. F. Coleman, and Y. Li, "A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems," *SIAM Journal on Scientific Computing*, vol. 21, no. 1, pp. 1–23, 1999.

[73] Glossary of economics, "Glossary of economics — Wikipedia, the free encyclopedia," 2017, [Online; accessed 30-October-2017]. [Online]. Available: https://en.wikipedia.org/wiki/Glossary_of_economics

[74] L. A. Adamic, "Zipf, power-laws, and pareto-a ranking tutorial," *Xerox Palo Alto Research Center, Palo Alto, CA, http://ginger. hpl. hp. com/shl/papers/ranking/ranking. html*, 2000.

[75] A. Damodaran, *Investment Valuation: Tools and Techniques for Determining the Value of Any Asset*, ser. Wiley finance series. Wiley, 2002.

[76] R. Mayfield, "Power law of participation," 2006, [Online; accessed 30-October-2017]. [Online]. Available: http://ross.typepad.com/blog/2006/04/power_law_of_pa.html

[77] C. Geigle, H. Dev, H. Sundaram, and C. Zhai, "Discovering and analyzing action-based roles in community question answering networks," Tech. Rep., 2018.

[78] T. R. Malthus, *An essay on the principle of population, as it affects the future improvement of society*, 1809, vol. 2.

[79] C. A. Sims, "Macroeconomics and reality," *Econometrica: Journal of the Econometric Society*, pp. 1–48, 1980.

[80] G. Gkotsis, K. Stepanyan, C. Pedrinaci, J. Domingue, and M. Liakata, "It's all in the content: state of the art best answer prediction based on discretisation of shallow linguistic features," in *Proceedings of the 2014 ACM Conference on Web Science (WebSci)*. ACM, 2014, pp. 202–210.

[81] C. Shah and J. Pomerantz, "Evaluating and predicting answer quality in community qa," in *Proceedings of the 33rd International ACM SIGIR International Conference of Research and Development in Information Retrieval (SIGIR)*. ACM, 2010, pp. 411–418.

[82] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, "Finding high-quality content in social media," in *Proceedings of the First International Conference on Web Search and Data Mining (WSDM)*. ACM, 2008, pp. 183–194.

[83] J. Jeon, W. B. Croft, J. H. Lee, and S. Park, "A framework to predict the quality of answers with non-textual features," in *Proceedings of the 29th International ACM SIGIR International Conference of Research and Development in Information Retrieval (SIGIR)*. ACM, 2006, pp. 228–235.

[84] L. Hong and S. E. Page, "Groups of diverse problem solvers can outperform groups of high-ability problem solvers," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 101, no. 46, pp. 16 385–16 389, 2004.

[85] K. Burghardt, E. F. Alsina, M. Girvan, W. Rand, and K. Lerman, "The myopia of crowds: Cognitive load and collective evaluation of answers on stack exchange," *PLOS One*, vol. 12, no. 3, p. e0173610, 2017.

[86] K. Burghardt, T. Hogg, and K. Lerman, "Quantifying the impact of cognitive biases in question-answering systems," in *Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM)*, 2018.

[87] J.-S. Beuscart and T. Couronné, "The distribution of online reputation: Audience and influence of musicians on myspace," in *Proceedings of the Third International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2009.

[88] T. Hogg and K. Lerman, "Disentangling the effects of social signals," *Human Computation*, vol. 2, no. 2, 2015.

[89] K. Lerman and T. Hogg, "Leveraging position bias to improve peer recommendation," *PLOS One*, vol. 9, no. 6, p. e98914, 2014.

[90] M. J. Salganik, P. S. Dodds, and D. J. Watts, "Experimental study of inequality and unpredictability in an artificial cultural market," *Science*, vol. 311, no. 5762, pp. 854–856, 2006.

[91] J. Lorenz, H. Rauhut, F. Schweitzer, and D. Helbing, "How social influence can undermine the wisdom of crowd effect," *Proceedings of the National Academy of Sciences (PNAS)*, vol. 108, no. 22, pp. 9020–9025, 2011.

[92] C. Krumme, M. Cebrian, G. Pickard, and S. Pentland, "Quantifying social influence in an online cultural market," *PLOS One*, vol. 7, no. 5, p. e33785, 2012.

[93] L. Muchnik, S. Aral, and S. J. Taylor, "Social influence bias: A randomized experiment," *Science*, vol. 341, no. 6146, pp. 647–651, 2013.

[94] S. Krishnan, J. Patel, M. J. Franklin, and K. Goldberg, "A methodology for learning, analyzing, and mitigating social influence bias in recommender systems," in *Proceedings of the 8th ACM Conference on Recommender systems (RecSys)*. ACM, 2014, pp. 137–144.

[95] T. Wang, D. Wang, and F. Wang, "Quantifying herding effects in crowd wisdom," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 2014, pp. 1087–1096.

[96] G. Stoddard, "Popularity dynamics and intrinsic quality in reddit and hacker news," in *Proceedings of the Ninth International AAAI Conference on Web and Social Media (ICWSM)*, 2015, pp. 416–425.

[97] P. Van Hentenryck, A. Abeliuk, F. Berbeglia, F. Maldonado, and G. Berbeglia, "Aligning popularity and quality in online cultural markets," in *Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM)*, 2016, pp. 398–407.

[98] A. Abeliuk, G. Berbeglia, P. Van Hentenryck, T. Hogg, and K. Lerman, "Taming the unpredictability of cultural markets with social influence," in *Proceedings of the 26th International Conference on World Wide Web (WWW)*. ACM, 2017, pp. 745–754.

[99] H. Oktay, B. J. Taylor, and D. D. Jensen, "Causal discovery in social media using quasi-experimental designs," in *Proceedings of the First Workshop on Social Media Analytics*. ACM, 2010, pp. 1–9.

[100] D. Card, "The causal effect of education on earnings," in *Handbook of Labor Economics*. Elsevier, 1999, vol. 3, pp. 1801–1863.

[101] A. Gerber, "Estimating the effect of campaign spending on senate election outcomes using instrumental variables," *American Political Science Review*, vol. 92, no. 2, pp. 401–411, 1998.

[102] Y. Jong-Sung and S. Khagram, "A comparative study of inequality and corruption," *American Sociological Review*, vol. 70, no. 1, pp. 136–157, 2005.

[103] J. D. Angrist and J.-S. Pischke, *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

[104] M. Hernán and J. Robins, "Causal inference." Chapman & Hall/CRC, 2019, ch. 16.

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

[106] A. Richterich, "'karma, precious karma!' karmawhoring on reddit and the front page's econometrisation," *Journal of Peer Production*, vol. 4, no. 1, 2014.

[107] R. K. Merton, "The matthew effect in science: The reward and communication systems of science are considered," *Science*, vol. 159, no. 3810, pp. 56–63, 1968.

[108] Y. Yue, R. Patel, and H. Roehrig, "Beyond position bias: Examining result attractiveness as a source of presentation bias in clickthrough data," in *Proceedings of the 19th International Conference on World Wide Web (WWW)*. ACM, 2010, pp. 1011–1018.

[109] E. Gilbert, "Widespread underprovision on reddit," in *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW)*. ACM, 2013, pp. 803–808.

[110] R. Sipos, A. Ghosh, and T. Joachims, "Was this review helpful to you?: it depends! context and voting patterns in online content," in *Proceedings of the 23rd International Conference on World Wide Web (WWW)*. ACM, 2014, pp. 337–348.

[111] J. Thebault-Spieker, D. Kluver, M. A. Klein, A. Halfaker, B. Hecht, L. Terveen, and J. A. Konstan, "Simulation experiments on (the absence of) ratings bias in reputation systems," *Proceedings of the ACM on Human-Computer Interaction*, vol. 1, no. CSCW, p. 101, 2017.

[112] M. Glenski, C. Pennycuff, and T. Weninger, "Consumers and curators: Browsing and voting patterns on reddit," *IEEE Transactions on Computational Social Systems*, vol. 4, no. 4, pp. 196–206, 2017.

[113] M. Glenski, G. Stoddard, P. Resnick, and T. Weninger, "Guessthekarma: A game to assess social rating systems," *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, p. 59, 2018.

[114] A. Pal and S. Counts, "Identifying topical authorities in microblogs," in *Proceedings of the Fourth International Conference on Web Search and Data Mining (WSDM)*. ACM, 2011, pp. 45–54.

[115] Y. R. Tausczik and J. W. Pennebaker, "Predicting the perceived quality of online mathematics contributions from users' reputations," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 2011, pp. 1885–1888.

[116] S. A. Paul, L. Hong, and E. H. Chi, "Who is authoritative? understanding reputation mechanisms in quora," in *Collective Intelligence*, 2012.

[117] Y. Liang, "Knowledge sharing in online discussion threads: What predicts the ratings?" in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*. ACM, 2017, pp. 146–154.

[118] O. Budzinski and S. Gaenssle, "The economics of social media stars: An empirical investigation of stardom, popularity, and success on youtube," 2018.

[119] G. M. MacDonald, "The economics of rising stars," *The American Economic Review*, pp. 155–166, 1988.

[120] L. E. Celis, P. M. Krafft, and N. Kobe, "Sequential voting promotes collective discovery in social recommendation systems," in *Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM)*, 2016, pp. 42–51.

[121] F. Wu and B. A. Huberman, "How public opinion forms," in *International Workshop on Internet and Network Economics (WINE)*. Springer, 2008, pp. 334–341.

[122] G. Lederrey and R. West, "When sheep shop: Measuring herding effects in product ratings with natural experiments," in *Proceedings of the 2018 World Wide Web Conference (WWW)*. IW3C2, 2018, pp. 793–802.

[123] T. Joachims, A. Swaminathan, and T. Schnabel, "Unbiased learning-to-rank with biased feedback," in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM)*. ACM, 2017, pp. 781–789.

[124] F. Radlinski and T. Joachims, "Minimally invasive randomization for collecting unbiased preferences from clickthrough logs," in *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, vol. 21, no. 2. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006, p. 1406.

[125] D. Gaffney and J. N. Matias, "Caveat emptor, computational social science: Large-scale missing data in a widely-published reddit corpus," *PLOS One*, vol. 13, no. 7, p. e0200162, 2018.

[126] C. Winship and S. L. Morgan, "The estimation of causal effects from observational data," *Annual Review of Sociology*, vol. 25, no. 1, pp. 659–706, 1999.

[127] J. Pearl, "Bayesianism and causality, or, why i am only a half-bayesian," in *Foundations of Bayesianism*. Springer, 2001, pp. 19–36.

[128] M. Dickson, "The causal effect of education on wages revisited," *Oxford Bulletin of Economics and Statistics*, vol. 75, no. 4, pp. 477–498, 2013.

[129] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec, "Discovering value from community activity on focused question answering sites: a case study of stack overflow," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 2012, pp. 850–858.

[130] L. Wu, J. A. Baggio, and M. A. Janssen, "The role of diverse strategies in sustainable knowledge production," *PLOS One*, vol. 11, no. 3, p. e0149151, 2016.

[131] J. D. Angrist and J.-S. Pischke, "The credibility revolution in empirical economics: How better research design is taking the con out of econometrics," *Journal of Economic Perspectives*, vol. 24, no. 2, pp. 3–30, 2010.

[132] A. Merchant, D. Shah, G. S. Bhatia, A. Ghosh, and P. Kumaraguru, "Signals matter: Understanding popularity and impact of users on stack overflow," in *The World Wide Web Conference (WWW)*. ACM, 2019, pp. 3086–3092.

[133] A. Halavais, K. H. Kwon, S. Havener, and J. Striker, "Badges of friendship: Social influence and badge acquisition on stack overflow," in *2014 47th Hawaii International Conference on System Sciences (HICSS)*. IEEE, 2014, pp. 1607–1615.

[134] H. Dev, C. Geigle, Q. Hu, J. Zheng, and H. Sundaram, "The size conundrum: Why online knowledge markets can fail at scale," in *Proceedings of the 2018 World Wide Web Conference (WWW)*. IW3C2, 2018, pp. 65–75.

[135] B. Grosser, "What do metrics want? how quantification prescribes social interaction on facebook," *Computational Culture*, 2014.

[136] M. Eslami, S. R. Krishna Kumaran, C. Sandvig, and K. Karahalios, "Communicating algorithmic process in online behavioral advertising," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 2018, p. 432.

[137] D. Romano and M. Pinzger, "Towards a weighted voting system for q&a sites," in *2013 IEEE International Conference on Software Maintenance*. IEEE, 2013, pp. 368–371.

[138] B. Vasilescu, A. Capiluppi, and A. Serebrenik, "Gender, representation and online participation: A quantitative study of stackoverflow," in *2012 International Conference on Social Informatics*. IEEE, 2012, pp. 332–338.

[139] D. Ford, J. Smith, P. J. Guo, and C. Parnin, "Paradise unplugged: Identifying barriers for female participation on stack overflow," in *Proceedings of the 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering (FSE)*. ACM, 2016, pp. 846–857.

[140] J. Hanlon, "Stack overflow isn't very welcoming. it's time for that to change." https://stackoverflow.blog/2018/04/26/stack-overflow-isnt-very-welcoming-its-time-for-that-to-change/, 2019.

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

[142] Y. R. Tausczik and J. W. Pennebaker, "Participation in an online mathematics community: differentiating motivations to add," in *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW)*. ACM, 2012, pp. 207–216.

[143] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec, "How community feedback shapes user behavior," in *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[144] J. Hanlon, "Stack overflow isn't very welcoming. it's time for that to change." https://stackoverflow.blog/2018/04/26/stack-overflow-isnt-very-welcoming-its-time-for-that-to-change/, 2018.

[145] M. Torkjazi, R. Rejaie, and W. Willinger, "Hot today, gone tomorrow: on the migration of myspace users," in *Proceedings of the 2nd ACM workshop on Online social networks*. ACM, 2009, pp. 43–48.

[146] D. Garcia, P. Mavrodiev, and F. Schweitzer, "Social resilience in online communities: The autopsy of friendster," in *Proceedings of the first ACM conference on Online social networks*. ACM, 2013, pp. 39–50.

[147] B. Ribeiro, "Modeling and predicting the growth and death of membership-based websites," in *Proceedings of the 23rd international conference on World wide web*. ACM, 2014, pp. 653–664.

[148] C. Danescu-Niculescu-Mizil, R. West, D. Jurafsky, J. Leskovec, and C. Potts, "No country for old members: User lifecycle and linguistic change in online communities," in *Proceedings of the 22nd international conference on World Wide Web*. ACM, 2013, pp. 307–318.

[149] K. Kapoor, M. Sun, J. Srivastava, and T. Ye, "A hazard based approach to user return time prediction," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 1719–1728.

[150] R. Flesch, "A new readability yardstick." *Journal of applied psychology*, vol. 32, no. 3, p. 221, 1948.

[151] S. Loria, "Textblob: Simplified text processing," https://textblob.readthedocs.io/en/dev/, 2018.

[152] P. R. Rosenbaum and D. B. Rubin, "The central role of the propensity score in observational studies for causal effects," *Biometrika*, vol. 70, no. 1, pp. 41–55, 1983.

[153] M. Rokicki, E. Herder, and C. Trattner, "How editorial, temporal and social biases affect online food popularity and appreciation." in *ICWSM*, 2017, pp. 192–200.

[154] C. Yang, X. Shi, L. Jie, and J. Han, "I know you'll be back: Interpretable new user clustering and churn prediction on a mobile social application," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 914–922.

[155] Z. Lin, T. Althoff, and J. Leskovec, "I'll be back: On the multiple lives of users of a mobile activity tracking application," in *Proceedings of the 27th International Conference on World Wide Web*, vol. 2018. ACM, 2018, p. 1501.

[156] H. Jing and A. J. Smola, "Neural survival recommender," in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, 2017, pp. 515–524.

[157] Q. Wu, H. Wang, L. Hong, and Y. Shi, "Returning is believing: Optimizing long-term user engagement in recommender systems," in *Proceedings of the 2017 ACM Conference on Information and Knowledge Management*. ACM, 2017, pp. 1927–1936.

[158] K. Kapoor, K. Subbian, J. Srivastava, and P. Schrater, "Just in time recommendations: Modeling the dynamics of boredom in activity streams," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. ACM, 2015, pp. 233–242.

[159] A. Patil, J. Liu, and J. Gao, "Predicting group stability in online social networks," in *Proceedings of the 22nd international conference on World Wide Web*. ACM, 2013, pp. 1021–1030.

[160] F. M. Harper, D. Moy, and J. A. Konstan, "Facts or friends?: distinguishing informational and conversational questions in social q&a sites," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2009, pp. 759–768.

[161] Y. Liu, J. Bian, and E. Agichtein, "Predicting information seeker satisfaction in community question answering," in *Proceedings of the 31st International ACM SIGIR International Conference of Research and Development in Information Retrieval*. ACM, 2008, pp. 483–490.

[162] J. Bian, Y. Liu, E. Agichtein, and H. Zha, "Finding the right facts in the crowd: factoid question answering over social media," in *Proceedings of the 17th International Conference on World Wide Web*. ACM, 2008, pp. 467–476.

[163] B. Carterette and P. Chandar, "Offline comparative evaluation with incremental, minimally-invasive online feedback," in *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2018, pp. 705–714.

[164] A. Gilotte, C. Calauzènes, T. Nedelec, A. Abraham, and S. Dollé, "Offline a/b testing for recommender systems," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 2018, pp. 198–206.

[165] L. Li, W. Chu, J. Langford, and X. Wang, "Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms," in *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*. ACM, 2011, pp. 297–306.

[166] L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman, "Knowledge sharing and yahoo answers: Everyone knows something," in *Proc. WWW*, ser. WWW '08, 2008, pp. 665–674.

[167] L. Mamykina, B. Manoim, M. Mittal, G. Hripcsak, and B. Hartmann, "Design lessons from the fastest q&a site in the west," in *Proc. CHI*, ser. CHI '11, 2011, pp. 2857–2866.

[168] K. K. Nam, M. S. Ackerman, and L. A. Adamic, "Questions in, knowledge in?: A study of naver's question answering community," in *Proc. CHI*, ser. CHI '09, 2009, pp. 779–788.

[169] S. Wang, D. Lo, and L. Jiang, "An empirical study on developer interactions in stackoverflow," in *Proc. SAC*, ser. SAC '13, 2013, pp. 1019–1024.

[170] D. Fisher, M. Smith, and H. T. Welser, "You are who you talk to: Detecting roles in usenet newsgroups," in *Proceedings of the 39th Annual Hawaii International Conference on System Sciences - Volume 03*, ser. HICSS '06. Washington, DC, USA: IEEE Computer Society, 2006, pp. 59.2–.

[171] H. T. Welser, E. Gleave, D. Fisher, and M. Smith, "Visualizing the signatures of social roles in online discussion groups," *Journal of Social Structure*, vol. 8, pp. 1–31, 2007.

[172] V. D. Barash, M. Smith, L. Getoor, and H. T. Welser, "Distinguishing knowledge vs social capital in social media with roles and context," in *Proceedings of the Third International AAAI Conference on Weblogs and Social Media*, 2009.

[173] J. Chan, C. Hayes, and E. M. Daly, "Decomposing discussion forums and boards using user roles," in *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 2010, pp. 215–218.

[174] A. Furtado, N. Andrade, N. Oliveira, and F. Brasileiro, "Contributor profiles, their dynamics, and their importance in five q&a sites," in *Proc. CSCW*, ser. CSCW '13, 2013, pp. 1237–1252.

[175] A. White, J. Chan, C. Hayes, and T. B. Murphy, "Mixed membership models for exploring user roles in online fora," in *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, 2012, pp. 599–602.

[176] E. Manavoglu, D. Pavlov, and C. L. Giles, "Probabilistic user behavior models," in *ICDM*, Nov 2003, pp. 203–210.

[177] Z. Xu, Y. Zhang, Y. Wu, and Q. Yang, "Modeling user posting behavior on social media," in *Proc. SIGIR*, ser. SIGIR '12, 2012, pp. 545–554.

[178] M. Qiu, F. Zhu, and J. Jiang, "It is not just what we say, but how we say them: Lda-based behavior-topic model," in *Proc. SDM*, ser. SDM '13, May 2013, pp. 794–802.

[179] Y. Han and J. Tang, "Probabilistic community and role model for social networks," in *Proc. KDD*, ser. KDD '15, 2015, pp. 407–416.

[180] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec, "Discovering value from community activity on focused question answering sites: A case study of stack overflow," in *Proc. KDD*, ser. KDD '12, 2012, pp. 850–858.

[181] G. Wang, X. Zhang, S. Tang, H. Zheng, and B. Y. Zhao, "Unsupervised clickstream clustering for user behavior analysis," in *Proc. CHI*, ser. CHI '16, 2016, pp. 225–236.

[182] c. Gündüz and M. T. Özsu, "A web page prediction model based on click-stream tree representation of user behavior," in *Proc. KDD*, ser. KDD '03, 2003, pp. 535–540.

[183] C. Geigle and C. Zhai, "Modeling mooc student behavior with two-layer hidden markov models," *Journal of Educational Data Mining*, vol. 9, no. 1, pp. 1–24, 2017.

[184] Q. Su and L. Chen, "A method for discovering clusters of e-commerce interest patterns using click-stream data," *Electron. Commer. Rec. Appl.*, vol. 14, no. 1, pp. 1–13, Jan. 2015.

[185] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida, "Characterizing user behavior in online social networks," in *Proc. IMC*, ser. IMC '09, 2009, pp. 49–62.

[186] L. Lu, M. Dunham, and Y. Meng, "Mining significant usage patterns from clickstream data," in *Proc. WebKDD*, ser. WebKDD'05, 2006, pp. 1–17.

[187] N. Sadagopan and J. Li, "Characterizing typical and atypical user sessions in click-streams," in *Proc. WWW*, ser. WWW '08, 2008, pp. 885–894.

[188] T. Hofmann, "Probabilistic latent semantic analysis," in *Proc. UAI*, ser. UAI'99, 1999, pp. 289–296.

[189] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[190] K. Nigam, A. K. Mccallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using em," *Machine Learning*, vol. 39, no. 2, pp. 103–134, May 2000.

[191] J. Yin and J. Wang, "A dirichlet multinomial mixture model-based approach for short text clustering," in *Proc. KDD*, ser. KDD '14, 2014, pp. 233–242.

[192] A. McCallum, X. Wang, and A. Corrada-Emmanuel, "Topic and role discovery in social networks with experiments on enron and academic email," *J. Artif. Int. Res.*, vol. 30, no. 1, pp. 249–272, Oct. 2007.

[193] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009.

[194] C. J. Maddison, D. Tarlow, and T. Minka, "A* sampling," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 3086–3094.

[195] S. Kullback, *Information Theory and Statistics*. John Wiley & Sons, 1959.

[196] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.

[197] W. J. Conover and R. L. Iman, "On multiple-comparisons procedures," Los Alamos Scientific Laboratory, Tech. Rep., 1979.

[198] S. Holm, "A simple sequentially rejective multiple test proceedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979.

[199] H. Dev, C. Geigle, Q. Hu, J. Zheng, and H. Sundaram, "The size conundrum: Why online knowledge markets can fail at scale," in *Proceedings of WWW 2018: The Web Conference*. New York, NY, USA: ACM, 4 2018, pp. 65–75.

[200] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *arXiv preprint arXiv:1310.4546*, 2013.

[201] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International conference on machine learning*. PMLR, 2014, pp. 1188–1196.

[202] O. Barkan and N. Koenigstein, "Item2vec: neural item embedding for collaborative filtering," in *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2016, pp. 1–6.

[203] E. Chandrasekharan, M. Samory, S. Jhaver, H. Charvat, A. Bruckman, C. Lampe, J. Eisenstein, and E. Gilbert, "The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales," *Proceedings of the ACM on Human-Computer Interaction*, vol. 2, no. CSCW, pp. 1–25, 2018.

[204] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, and E. Gilbert, "You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech," *Proceedings of the ACM on Human-Computer Interaction*, vol. 1, no. CSCW, pp. 1–22, 2017.

[205] E. Chandrasekharan, C. Gandhi, M. W. Mustelier, and E. Gilbert, "Crossmod: A cross-community learning-based system to assist reddit moderators," *Proceedings of the ACM on human-computer interaction*, vol. 3, no. CSCW, pp. 1–30, 2019.