Quantifying Voter Biases in Online Platforms: An Instrumental Variable Approach

HIMEL DEV, University of Illinois at Urbana-Champaign, USA
KARRIE KARAHALIOS, University of Illinois at Urbana-Champaign, USA
HARI SUNDARAM, University of Illinois at Urbana-Champaign, USA

In content-based online platforms, use of aggregate user feedback (say, the sum of votes) is commonplace 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 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? We adopt an instrumental variable (IV) framework to answer this question. We identify a set of candidate instruments, carefully analyze their validity, and then use the valid instruments to reveal the effects of the impression signals on 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 user interface to avoid voter biases; making changes to platforms' policy to mitigate voter biases; detecting other forms of biases in online platforms.

Additional Key Words and Phrases: reputation bias, social influence bias, position bias, instrumental variables

ACM Reference Format:
Himel Dev, Karrie Karahalios, and Hari Sundaram. 2019. Quantifying Voter Biases in Online Platforms: An Instrumental Variable Approach. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 120 (November 2019), 27 pages. https://doi.org/10.1145/3359222

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 [20, 52, 2, 30]. 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 [29].

However, in most online platforms, individuals are prone to using cognitive heuristics that influence their voting behavior and prevent independent judgments [9, 10]. These heuristics incorporate different impression signals adjacent to the content—such as the reputation of the

Authors' addresses: Himel Dev, hdev3@illinois.edu, University of Illinois at Urbana-Champaign, USA; Karrie Karahalios, kkarahal@illinois.edu, University of Illinois at Urbana-Champaign, USA; Hari Sundaram, hs1@illinois.edu, University of Illinois at Urbana-Champaign, USA.

2573-0142/2019/11-ART120 $15.00
https://doi.org/10.1145/3359222

contributing user [7], aggregate vote thus far [28], and position of content [36]—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 [36, 28], 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 [51, 38, 34, 43, 36, 33, 59, 28, 54, 57, 9, 1, 23]. Broadly, researchers have adopted one of the following two approaches: 1) conduct experiments to create different voting conditions for studying participants [51, 38, 43, 36, 28, 1, 23]; 2) develop statistical models to analyze historical voting data [34, 33, 59, 54, 57, 9]. Both approaches have limitations. First, it is hard to perform randomized experiments in actual platforms due to feasibility, ethical issue, or cost [44]. 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 paper, 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 [11], how campaign spending affects senate elections [18], and how income inequality affects corruption [32]—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 [5]. 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 [27]. 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 paper 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 (~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.

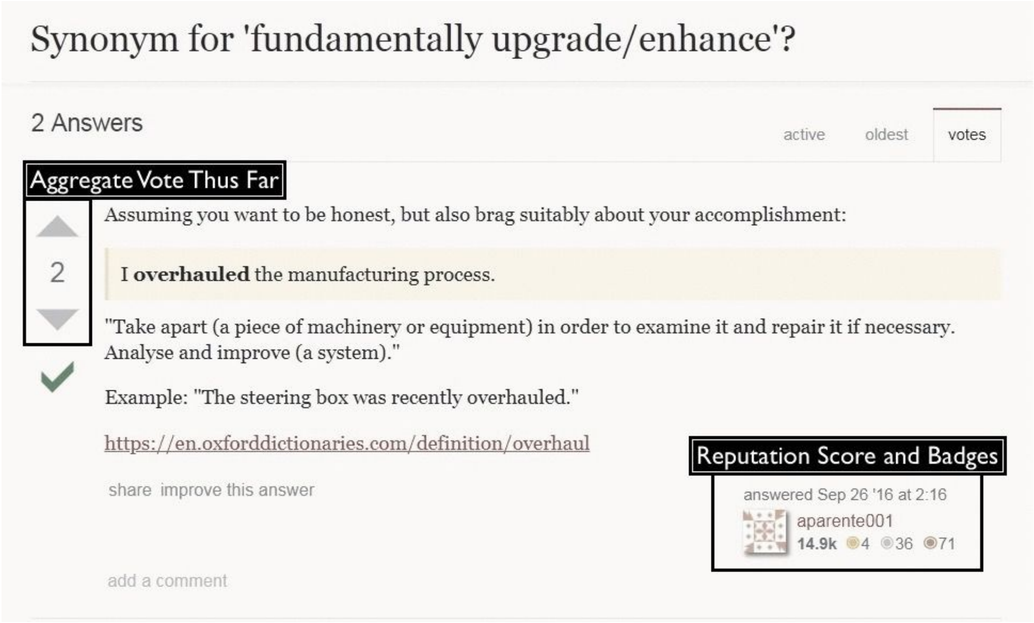
Significance. Our identification of causal factors that influence votes has a significant bearing on research in voter bias in particular, as well as the broader CSCW community. First, there are practical implications. Impression signals (answer position, user reputation, prior vote) play a significant 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 paper is organized as follows. We define our problem in Section 2 and discuss the related work in Section 3. We describe our data in Section 4. We then explain how our method works in Section 5. Section 6 reports the results of our study. We discuss the implications of our research in Section 7 and the limitations in Section 8. Finally, we conclude in Section 9.

2 VOTER BIAS

The goal of this paper 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 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 [42, 49]. In Stack



Synonym for 'fundamentally upgrade/enhance'?

2 Answers

Assuming you want to be honest, but also brag suitably about your accomplishment:

> I **overhauled** the manufacturing process.

"Take apart (a piece of machinery or equipment) in order to examine it and repair it if necessary. Analyse and improve (a system)."

Example: "The steering box was recently overhauled."

https://en.oxforddictionaries.com/definition/overhaul

answered Sep 26 '16 at 2:16
aparente001
14.9k ●4 ●36 ●71

Fig. 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.

Exchange, for example, votes on content translate into reputation score and badges for the contributing user [42, 4]. 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 [41]: 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 [38]: 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 [54, 28, 9, 10]. 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 influence bias*. We know from prior work that for platforms that reveal social signal, users exhibit a herding effect [23]: 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 [63] or votes [36]. 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.

3 RELATED WORK

Our work draws from and improves upon, a rich literature on online voting behavior and voter biases. Since this paper focuses on quantifying voter biases, we provide a taxonomy of related work on voter biases (Table 1).

Table 1. A taxonomy of existing literature on voter biases.

Bias	Approach	References	Summary
Reputation Bias	Correlation Study	[7], [45], [55], [46], [37], [8]	Show some evidence of correlation between past reputation and current success [7, 45, 55, 46, 37, 8].
Social Influence Bias	Randomized Experiment	[51], [38], [43], [1], [23]	Create different decision making (say voting) conditions for study participants by varying the availability of preceding decisions [51, 38, 1], and purposefully engineered initial decision [43, 23].
	AMT Simulation	[28], [12], [10]	Simulate alternative voting conditions of platform in Amazon Mechanical Turk (AMT) by varying the availability of preceding decisions [28, 12, 10].
	Statistical Model	[34], [33], [59], [54], [57], [9]	Develop statistical model for quantifying bias: Pólya Urn [34, 57], nonparametric significance test [33], additive generative model [59], Poisson regression [54], logistic regression [9].
	Matching Method	[61], [35]	Contrast aggregate user feedback (say ratings) on the same object in two different platforms via matching [35].
Position Bias	Randomized Experiment	[36], [28], [1]	Create different decision making (say voting) conditions for study participants by varying content ordering policies [36, 28, 1]
	Statistical Model	[54], [57], [31]	Develop statistical model for studying bias: Poisson regression [54], Pólya Urn [57], counterfactual inference [31].
	Matching Method	[48], [63]	Contrast aggregate user feedback (say ratings) for objects occupying similar positions [48, 63].

Voting Behavior. Recent research has made significant advancements towards the understanding of rating and voting behaviors in online platforms [19, 53, 56, 21, 22]. Gilbert [19] 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. [53] 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. [21] found that most Reddit users do not read the article that they vote on. In a later work, Glenski et al. [22] 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 paper, we study 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 [7, 45, 55, 46, 37, 8] (also known as "superstar economics" [39]). Beuscart et al. [7] 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. [45] reported that the popular users get a boost in their authority rating due to the "name value". Tausczik et al. [55] found that in MathOverflow, both offline and online reputation are correlated with the perceived quality of contributions. Paul et al. [46] found that Quora users judge the reputation of other users based on their past contributions. Liang [37] showed that in Reddit, users with higher comment karma tend to produce questions and comments with higher ratings. Budzinski et al. [8] 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 paper, we provide a quantification of reputation bias through causal estimates.

Social Influence Bias. Since the musiclab experiment by Salganik et al. [51], a large body of work has been devoted to the social influence bias [61, 38, 34, 43, 33, 54, 28, 57, 1, 9, 10], and its resultant herding effect [59, 12, 23, 35]. A majority of the work tends to fall into one of two categories—1) Experimental Study: randomized experiment [51, 38, 43, 1, 23], simulation via Amazon Mechanical Turk (AMT) [28, 12, 10]; and 2) Observational Study: statistical model [34, 33, 59, 54, 57, 9], matching method [61, 35]. 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 [34, 57], nonparametric significance test [33], additive generative model [59], Poisson regression [54], logistic regression [9]—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 [57]. Therefore, it is difficult to separate the social influence bias from the inherent quality and quantify its effect. In this paper, 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 [48, 63, 36, 28, 54, 57, 1, 31]. Notably, researchers performed several experimental studies in AMT, where they created different voting conditions for study participants by varying content ordering policies [36, 28, 1]. Hogg et al. [28] revealed that social signals affect item popularity about half as much as position and content do. Abeliuk et al. [1] showed that the unpredictability of voting outcome is a consequence of the ordering policy. Lerman et al. [36] found that different policies for ordering content could improve peer recommendation by steering user attention. In this paper, 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 DATA AND VARIABLES

In this section, we first discuss the choice of our data source (Section 4.1), then describe the datasets that we use in this study (Section 4.2); and finally present the variables that we accumulate from the datasets (Section 4.3).

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 [17] 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.

4.2 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 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 paper, 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 2 presents descriptive statistics for the three sites analyzed in this paper.

Table 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	Number of users who viewed the question
V_4	QuestionFavoriteCount	Number 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	Number of comments on the question
V_9	QuestionCommentCountT-	Number of comments on the question before time T
V_{10}	QuestionCommentCountT+	Number of comments on the question after time T
V_{11}	QuestionAnswerCount	Number of answers to the question
V_{12}	QuestionAnswerCountT-	Number of answers to the question before time T
V_{13}	QuestionAnswerCountT+	Number 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 based on the aggregate vote before time T
V_{24}	AnswerPositionT+	Position of the answer based on the aggregate vote after time T
V_{25}	AnswerCommentCount	Number of comments on the answer
V_{26}	AnswerCommentCountT-	Number of comments on the answer before time T
V_{27}	AnswerCommentCountT+	Number of comments on the answer after time T
V_{28}	AnswererPostCount	Number of posts (questions and answers) written by the answerer
V_{29}	AnswererAnswerCount	Number of answers written by the answerer
V_{30}	AnswererActiveAge	Time gap between between the answerer's 1st post and the answer
V_{31}	AnswererReputation	Total score of questions and answers written by the answerer
V_{32}	AnswererReputationViaAnswer	Total score of answers written by the answerer
V_{33}	AnswererGoldCount	Number of gold badges acquired by the answerer
V_{34}	AnswererSilverCount	Number of silver badges acquired by the answerer
V_{35}	AnswererBronzeCount	Number of bronze badges acquired by the answerer
V_{36}	AnswererBadgeDistrribution	[AnswererGoldCount, AnswererSilverCount, AnswererBronzeCount]
V_{37}	AnsweredQuestionViewTotal	Total number of users who viewed past questions answered by the answerer
V_{38}	AnsweredQuestionFavoriteTotal	Total number of users who favorited past questions answered by the answerer
V_{39}	AnsweredQuestionScoreTotal	Total score of past questions answered by the answerer
V_{40}	AnsweredQuestionCommentTotal	Total number of comments on past questions answered by the answerer
V_{41}	AnsweredQuestionAnswerTotal	Total number of answers to past questions answered by the answerer

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 3 describes the variables used in this study.

5 METHOD

In this section, we first discuss our choice of method for voter bias quantification (Section 5.1), then explain the fundamentals of the chosen method (Section 5.2); and finally present our models for quantifying voter bias (Section 5.3 and 5.4)

5.1 Choice of Method

The goal of this paper 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 [60]. 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 [60, 5]. We observe that our voter bias quantification problem is susceptible to several hidden confounders, such as 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 [11, 18, 32] to estimate causal effects (e.g., how education affects earning [11], how campaign spending affects senate selection [18], and how income inequality affects corruption [32]) from observational data. The IV method is especially useful for estimating effects in the presence of hidden confounders [5, 27]. 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 [5].

Table 4. The parallels between voter bias quantification and instrumental variable method.

IV Terminology	Bias Terminology	Example
Outcome	Aggregate Feedback	Mean of votes on content
Exposure	Impression Signal	Reputation of the contributing user
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 mean 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"* [47]. 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, we show the parallels between our problem and IV). For these reasons, we adopt the IV method to quantify voter biases.

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 [11]. 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" [11]. 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 [14]. 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 [11]. These IV studies motivate the question: *could we use IV for quantifying voter bias in Stack Exchange?*

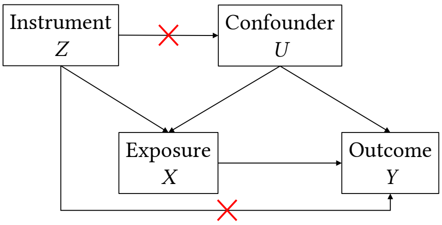


Fig. 2. General structure of an instrumental variable model. The paths from U to X, and U to Y introduces confounding in estimating the causal effect of X on Y. For a valid instrument Z, the pathways from Z to X, and X to Y must exist; whereas the pathways from Z to U, and Z to Y must cease to exist.

Figure 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 [27]:

(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 [27]. 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 paper, 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).

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

[1] We shall use this syntax consistently throughout this paper. The first term is variable name and the second term is variable id in Table 3. Please see Table 3 for the description of variables.

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 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 [18, 32], 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) [42], and 2) popularity of the answered questions (number of views, comments, and answers attracted by the questions) [3]. 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 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 3. Please note that our causal diagram follows the general structure of the instrumental variable framework (in Figure 2).

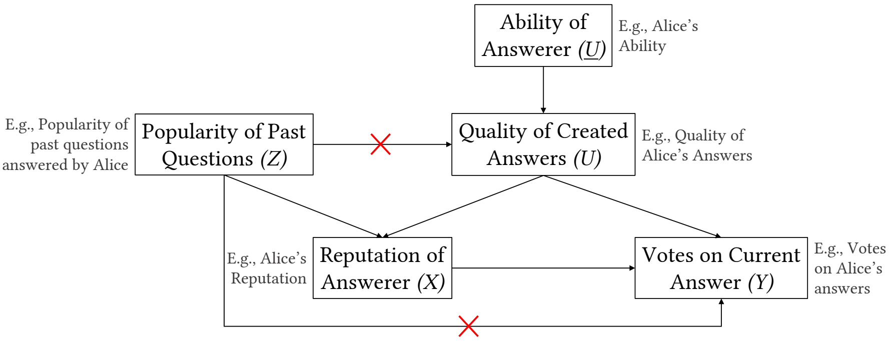

Ability of Answerer (U)
E.g., Alice's Ability
E.g., Popularity of past questions answered by Alice
Popularity of Past Questions (Z)
Quality of Created Answers (U)
E.g., Quality of Alice's Answers
E.g., Alice's Reputation
Reputation of Answerer (X)
Votes on Current Answer (Y)
E.g., Votes on Alice's answers

Fig. 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 specification: `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.

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.

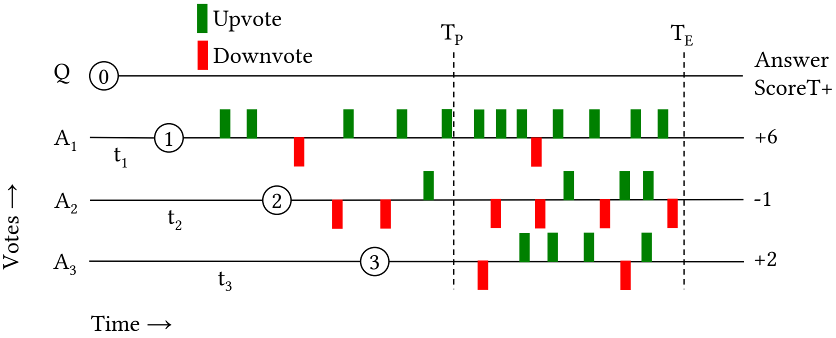


Fig. 4. An illustration of the bias formation period to quantify our outcome (`AnswerScoreT+` $\langle V_{21} \rangle$) and exposures (`AnswerScoreT-` $\langle V_{20} \rangle$ and `AnswerPositionT-` $\langle V_{23} \rangle$). The creation of question Q marks the beginning of our observation period. Then, 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 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 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) [3], time of answer (day of the week, hour of the day) [9], and timeliness of answer (time gap between question and answer) [54]. 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 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) [62]. 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 5.

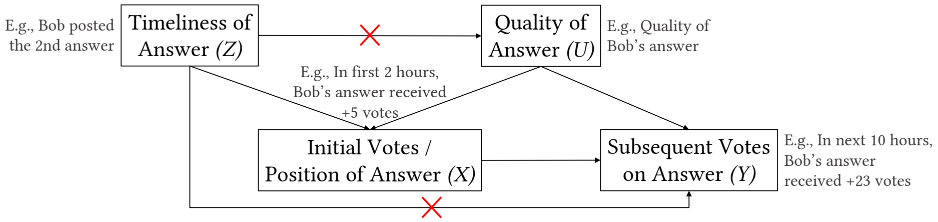

E.g., Bob posted the 2nd answer
Timeliness of Answer *(Z)*
Quality of Answer *(U)*
E.g., Quality of Bob's answer
E.g., In first 2 hours, Bob's answer received +5 votes
Initial Votes / Position of Answer *(X)*
Subsequent Votes on Answer *(Y)*
E.g., In next 10 hours, Bob's answer received +23 votes

Fig. 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 specification: `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.

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 6.1), social influence bias and position bias (Section 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.

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

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 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) × 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) × 5 (instrument) × 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 [6].

We compare the performance of OLS and IV models by examining their estimates (regression coefficients). Table 5, 6, and 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 5.2). For all IV estimates reported in Table 5–7, 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 5–7 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 paper, 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 (1.6–2.2x 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.

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.

Table 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 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 ⟨V$_{19}$⟩			
	(*for estimating the effect of Exposure*)	X = AnswererReputation ⟨V$_{31}$⟩		X = AnswererReputationViaAnswer ⟨V$_{32}$⟩	
Site	$Z + C$	OLS	IV	OLS	IV
Math	AnsweredQuestionViewTotal ⟨V$_{37}$⟩	0.056 (± 0.001)	0.055 (± 0.001)	0.053 (± 0.001)	0.051 (± 0.001)
	V$_{37}$ + QuestionViewCount ⟨V$_{3}$⟩	0.067 (± 0.001)	0.061 (± 0.001)	0.063 (± 0.001)	0.057 (± 0.001)
	AnsweredQuestionFavoriteTotal ⟨V$_{38}$⟩	0.056 (± 0.001)	0.057 (± 0.001)	0.053 (± 0.001)	0.053 (± 0.001)
	V$_{38}$ + QuestionFavoriteCount ⟨V$_{4}$⟩	0.061 (± 0.001)	0.057 (± 0.001)	0.058 (± 0.001)	0.053 (± 0.001)
	AnsweredQuestionScoreTotal ⟨V$_{39}$⟩	0.056 (± 0.001)	0.055 (± 0.001)	0.053 (± 0.001)	0.051 (± 0.001)
	V$_{39}$ + QusestionScore ⟨V$_{5}$⟩	0.058 (± 0.001)	0.053 (± 0.001)	0.055 (± 0.001)	0.049 (± 0.001)
	AnsweredQuestionCommentTotal ⟨V$_{40}$⟩	0.056 (± 0.001)	0.040 (± 0.001)	0.053 (± 0.001)	0.037 (± 0.001)
	V$_{40}$ + QuestionCommentCount ⟨V$_{8}$⟩	0.057 (± 0.001)	0.041 (± 0.001)	0.054 (± 0.001)	0.038 (± 0.001)
	AnsweredQuestionAnswerTotal ⟨V$_{41}$⟩	0.056 (± 0.001)	0.040 (± 0.001)	0.053 (± 0.001)	0.037 (± 0.001)
	V$_{41}$ + QuestionAnswerCount ⟨V$_{11}$⟩	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 ⟨V$_{19}$⟩					
	(*for estimating the effect of Exposure*)	X = AnswererGoldCount ⟨V$_{33}$⟩		X = AnswererSilverCount ⟨V$_{34}$⟩		X = AnswererBronzeCount ⟨V$_{35}$⟩	
Site	$Z + C$	OLS	IV	OLS	IV	OLS	IV
Math	AnsweredQuestionViewTotal ⟨V$_{37}$⟩	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 ⟨V$_{3}$⟩	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 ⟨V$_{38}$⟩	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 ⟨V$_{4}$⟩	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 ⟨V$_{39}$⟩	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 ⟨V$_{5}$⟩	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 ⟨V$_{40}$⟩	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 ⟨V$_{8}$⟩	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 ⟨V$_{41}$⟩	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 ⟨V$_{11}$⟩	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 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)

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 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+ ⟨V_{21}⟩, Z_1 = AnswerTimeliness ⟨V_{17}⟩, Z_2 = AnswerOrder ⟨V_{18}⟩			
		X_1 = AnswerScoreT- ⟨V_{20}⟩		X_2 = AnswerPositionT- ⟨V_{23}⟩	
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 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 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 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.

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 paper: 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.

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 [20]. On the other hand, users who are interested in different themes may be driven by different factors to contribute [9]. 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 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. [40] 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 [25], over reputation score. This evidence, coupled with growing concerns about user engagement in online platforms [13] 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.

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. [24] 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. [36] 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.

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 [15]. 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 [50]—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.

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 [58, 16, 26]. 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." [26]. Vasilescu et al. [58] 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. [16] 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.

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.

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 paper, 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.

REFERENCES

[1] Andrés Abeliuk, Gerardo Berbeglia, Pascal Van Hentenryck, Tad Hogg, and Kristina Lerman. 2017. Taming the unpredictability of cultural markets with social influence. In *Proceedings of the 26th International Conference on World Wide Web (WWW)*. ACM, 745–754.

[2] Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. 2008. Finding high-quality content in social media. In *Proceedings of the First International Conference on Web Search and Data Mining (WSDM)*. ACM, 183–194.

[3] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2012. Discovering value from community activity on focused question answering sites: a case study of stack overflow. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 850–858.

[4] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2013. Steering user behavior with badges. In *Proceedings of the 22nd International Conference on World Wide Web (WWW)*. ACM, 95–106.

[5] Joshua D Angrist and Jörn-Steffen Pischke. 2008. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press. Part 4.

[6] Joshua D Angrist and Jörn-Steffen Pischke. 2010. The credibility revolution in empirical economics: how better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, 24, 2, 3–30.

[7] Jean-Samuel Beuscart and Thomas Couronné. 2009. 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)*.

[8] Oliver Budzinski and Sophia Gaenssle. 2018. The economics of social media stars: an empirical investigation of stardom, popularity, and success on youtube.

[9] Keith Burghardt, Emanuel F Alsina, Michelle Girvan, William Rand, and Kristina Lerman. 2017. The myopia of crowds: cognitive load and collective evaluation of answers on stack exchange. *PLOS One*, 12, 3, e0173610.

[10] Keith Burghardt, Tad Hogg, and Kristina Lerman. 2018. Quantifying the impact of cognitive biases in question-answering systems. In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM)*.

[11] David Card. 1999. The causal effect of education on earnings. In *Handbook of Labor Economics*. Vol. 3. Elsevier, 1801–1863.

[12] L Elisa Celis, Peter M Krafft, and Nathan Kobe. 2016. Sequential voting promotes collective discovery in social recommendation systems. In *Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM)*, 42–51.

[13] Himel Dev, Chase Geigle, Qingtao Hu, Jiahui Zheng, and Hari Sundaram. 2018. The size conundrum: why online knowledge markets can fail at scale. In *Proceedings of the 2018 World Wide Web Conference (WWW)*. IW3C2, 65–75.

[14] Matt Dickson. 2013. The causal effect of education on wages revisited. *Oxford Bulletin of Economics and Statistics*, 75, 4, 477–498.

[15] Motahhare Eslami, Sneha R Krishna Kumaran, Christian Sandvig, and Karrie Karahalios. 2018. Communicating algorithmic process in online behavioral advertising. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 432.

[16] Denae Ford, Justin Smith, Philip J Guo, and Chris Parnin. 2016. 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, 846–857.

[17] Devin Gaffney and J Nathan Matias. 2018. Caveat emptor, computational social science: large-scale missing data in a widely-published reddit corpus. *PLOS One*, 13, 7, e0200162.

[18] Alan Gerber. 1998. Estimating the effect of campaign spending on senate election outcomes using instrumental variables. *American Political Science Review*, 92, 2, 401–411.

[19] Eric Gilbert. 2013. Widespread underprovision on reddit. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW)*. ACM, 803–808.

[20] George Gkotsis, Karen Stepanyan, Carlos Pedrinaci, John Domingue, and Maria Liakata. 2014. 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, 202–210.

[21] Maria Glenski, Corey Pennycuff, and Tim Weninger. 2017. Consumers and curators: browsing and voting patterns on reddit. *IEEE Transactions on Computational Social Systems*, 4, 4, 196–206.

[22] Maria Glenski, Greg Stoddard, Paul Resnick, and Tim Weninger. 2018. Guessthekarma: a game to assess social rating systems. *Proceedings of the ACM on Human-Computer Interaction*, 2, CSCW, 59.

[23] Maria Glenski and Tim Weninger. 2017. Rating effects on social news posts and comments. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 8, 6, 78.

[24] Benjamin Grosser. 2014. What do metrics want? how quantification prescribes social interaction on facebook. *Computational Culture*.

[25] Alexander Halavais, K Hazel Kwon, Shannon Havener, and Jason Striker. 2014. Badges of friendship: social influence and badge acquisition on stack overflow. In *2014 47th Hawaii International Conference on System Sciences (HICSS)*. IEEE, 1607–1615.

[26] Jay Hanlon. 2019. 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).

[27] MA Hernán and JM Robins. 2019. Causal inference. In Chapman & Hall/CRC. Part 16.

[28] Tad Hogg and Kristina Lerman. 2015. Disentangling the effects of social signals. *Human Computation*, 2, 2.

[29] Lu Hong and Scott E Page. 2004. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences (PNAS)*, 101, 46, 16385–16389.

[30] Jiwoon Jeon, W Bruce Croft, Joon Ho Lee, and Soyeon Park. 2006. 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, 228–235.

[31] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017. Unbiased learning-to-rank with biased feedback. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM)*. ACM, 781–789.

[32] You Jong-Sung and Sanjeev Khagram. 2005. A comparative study of inequality and corruption. *American Sociological Review*, 70, 1, 136–157.

[33] Sanjay Krishnan, Jay Patel, Michael J Franklin, and Ken Goldberg. 2014. 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, 137–144.

[34] Coco Krumme, Manuel Cebrian, Galen Pickard, and Sandy Pentland. 2012. Quantifying social influence in an online cultural market. *PLOS One*, 7, 5, e33785.

[35] Gael Lederrey and Robert West. 2018. When sheep shop: measuring herding effects in product ratings with natural experiments. In *Proceedings of the 2018 World Wide Web Conference (WWW)*. IW3C2, 793–802.

[36] Kristina Lerman and Tad Hogg. 2014. Leveraging position bias to improve peer recommendation. *PLOS One*, 9, 6, e98914.

[37] Yuyang Liang. 2017. 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, 146–154.

[38] Jan Lorenz, Heiko Rauhut, Frank Schweitzer, and Dirk Helbing. 2011. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences (PNAS)*, 108, 22, 9020–9025.

[39] Glenn M MacDonald. 1988. The economics of rising stars. *The American Economic Review*, 155–166.

[40] Arpit Merchant, Daksh Shah, Gurpreet Singh Bhatia, Anurag Ghosh, and Ponnurangam Kumaraguru. 2019. Signals matter: understanding popularity and impact of users on stack overflow. In *The World Wide Web Conference (WWW)*. ACM, 3086–3092.

[41] Robert K Merton. 1968. The matthew effect in science: the reward and communication systems of science are considered. *Science*, 159, 3810, 56–63.

[42] Dana Movshovitz-Attias, Yair Movshovitz-Attias, Peter Steenkiste, and Christos Faloutsos. 2013. 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, 886–893.

[43] Lev Muchnik, Sinan Aral, and Sean J Taylor. 2013. Social influence bias: a randomized experiment. *Science*, 341, 6146, 647–651.

[44] Hüseyin Oktay, Brian J Taylor, and David D Jensen. 2010. Causal discovery in social media using quasi-experimental designs. In *Proceedings of the First Workshop on Social Media Analytics*. ACM, 1–9.

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

[46] Sharoda A Paul, Lichan Hong, and Ed H Chi. 2012. Who is authoritative? understanding reputation mechanisms in quora. In *Collective Intelligence*.

[47] Judea Pearl. 2001. Bayesianism and causality, or, why i am only a half-bayesian. In *Foundations of Bayesianism*. Springer, 19–36.

[48] Filip Radlinski and Thorsten Joachims. 2006. Minimally invasive randomization for collecting unbiased preferences from clickthrough logs. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)* number 2. Vol. 21. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 1406.

[49] Annika Richterich. 2014. 'karma, precious karma!' karmawhoring on reddit and the front page's econometrisation. *Journal of Peer Production*, 4, 1.

[50] Daniele Romano and Martin Pinzger. 2013. Towards a weighted voting system for q&a sites. In *2013 IEEE International Conference on Software Maintenance*. IEEE, 368–371.

[51] Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *Science*, 311, 5762, 854–856.

[52] Chirag Shah and Jefferey Pomerantz. 2010. 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, 411–418.

[53] Ruben Sipos, Arpita Ghosh, and Thorsten Joachims. 2014. 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, 337–348.

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

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

[56] Jacob Thebault-Spieker, Daniel Kluver, Maximilian A Klein, Aaron Halfaker, Brent Hecht, Loren Terveen, and Joseph A Konstan. 2017. Simulation experiments on (the absence of) ratings bias in reputation systems. *Proceedings of the ACM on Human-Computer Interaction*, 1, CSCW, 101.

[57] Pascal Van Hentenryck, Andrés Abeliuk, Franco Berbeglia, Felipe Maldonado, and Gerardo Berbeglia. 2016. Aligning popularity and quality in online cultural markets. In *Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM)*, 398–407.

[58] Bogdan Vasilescu, Andrea Capiluppi, and Alexander Serebrenik. 2012. Gender, representation and online participation: a quantitative study of stackoverflow. In *2012 International Conference on Social Informatics*. IEEE, 332–338.

[59] Ting Wang, Dashun Wang, and Fei Wang. 2014. Quantifying herding effects in crowd wisdom. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 1087–1096.

[60] Christopher Winship and Stephen L Morgan. 1999. The estimation of causal effects from observational data. *Annual Review of Sociology*, 25, 1, 659–706.

[61] Fang Wu and Bernardo A Huberman. 2008. How public opinion forms. In *International Workshop on Internet and Network Economics (WINE)*. Springer, 334–341.

[62] Lingfei Wu, Jacopo A Baggio, and Marco A Janssen. 2016. The role of diverse strategies in sustainable knowledge production. *PLOS One*, 11, 3, e0149151.

[63] Yisong Yue, Rajan Patel, and Hein Roehrig. 2010. 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, 1011–1018.

Received April 2019; revised June 2019; accepted August 2019