Proceedings of the Fourteenth International AAAI Conference on Web and Social Media (ICWSM 2020)

“Trust Me, I Have a Ph.D.”: A Propensity Score Analysis on the Halo Effect of Disclosing One’s Offline Social Status in Online Communities

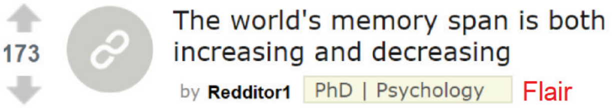
Kunwoo Park,[†] **Haewoon Kwak,**[‡] **Hyunho Song,**[§,¶] **Meeyoung Cha**[¶,§]
[†]University of California, Los Angeles, [‡]Qatar Computing Research Institute,
[§]Korea Advanced Institute of Science and Technology, [¶]Institute for Basic Science
kunwpark@ucla.edu, haewoon@acm.org, hyun78@kaist.ac.kr, mcha@ibs.re.kr

Abstract

Online communities adopt various reputation schemes to measure content quality. This study analyzes the effect of a new reputation scheme that exposes one’s offline social status, such as an education degree, within an online community. We study two Reddit communities that adopted this scheme, whereby posts include tags identifying education status referred to as *flairs*, and we examine how the “transferred” social status affects the interactions among the users. We computed propensity scores to test whether flairs give ad-hoc authority to the adopters while minimizing the effects of confounding variables such as topics of content. The results show that exposing academic degrees is likely to lead to higher audience votes as well as larger discussion size, compared to the users without the disclosed identities, in a community that covers peer-reviewed scientific articles. In another community with a focus on casual science topics, exposing mere academic degrees did not obtain such benefits. Still, the users with the highest degree (e.g., Ph.D. or M.D.) were likely to receive more feedback from the audience. These findings suggest that reputation schemes that link the offline and online worlds could induce halo effects on feedback behaviors differently depending upon the community culture. We discuss the implications of this research for the design of future reputation mechanisms.

Introduction

Online communities strive to encourage high-quality content from their members; however, judging the quality of myriads of content has been a great challenge (Brandtzæg and Heim 2008). A popular solution to this problem is reputation tracking whereby community members evaluate the quality of the content generated by other members by giving votes, and the aggregated votes constitute each member’s *reputation*. Studies have found that crowdsourced votes from peers positively correlate to content quality (Stoddard 2015). However, this reputation mechanism has a critical limitation. That limitation is known as the cold start problem (Lampe and Resnick 2004) whereby there is no prior record for newly joining members or freshly uploaded content. Social influence and pre-existing records can also bias individuals’ behavior when judging content quality (Muchnik, Aral, and Taylor 2013; Aral 2014; Berry and Taylor 2017).



The world's memory span is both increasing and decreasing

by Redditor1 PhD | Psychology

Figure 1: An example of flair information

An alternative to building a reputation online from scratch is to “bring” an established *social status* from the real world (e.g., academic degree or job affiliation). For example, on the question-and-answer site Quora, people can choose to reveal their domain expertise by listing job affiliations. Such “transfer” of offline to online status gives its members an ad hoc social status. This transfer provides additional information about the writer’s knowledge, which helps the other members to better judge the credibility of the content. This mechanism may be promising because domain experts can promote their content better than others and, at the same time, overcome the cold start problem. However, this belief has yet to be tested in a data-driven manner. *How does a community respond to the act of revealing offline social status such as academic degrees?*

This research brings attention to two online communities on Reddit that utilize information about members’ academic degrees and domain of expertise to promote quality content: *r/Science*[1] and *r/EverythingScience*[2].

The members in these communities use tags on posts and comments, called “flairs,” as demonstrated in Figure 1. To acquire a flair, the members must send proof of certification, such as a diploma, to the community moderators for verification. According to the board announcement, the purpose of the flair mechanism was “to enable the general public to distinguish between an educated opinion and a random comment without a background related to the topic.”[3] As of March 2020, the number of subscribers in these two communities was 23.6 M and 218 K, respectively. Gathering data from these two large communities enables us to investigate how the reputation mechanism that links offline and

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[1] https://www.reddit.com/r/science

[2] https://www.reddit.com/r/everythingscience

[3] Do you have a college degree or higher in science? Get flair indicating your expertise in /r/science! http://tiny.cc/2wqlqy

online worlds works in the wild. We investigate the transfer of offline social status to the online world with a data-driven approach and compare the findings between the two communities.

The analysis of million-scale data over several years can identify how flairs affect the feedback behavior of community members. However, confounding factors, such as topics of content and user reputation, present challenges in measuring the effects of exposing academic status. Therefore, we conducted a propensity score analysis to infer the effect of flairs while controlling for the influences of the confounding variables on the community feedback. We found that r/Science members who expose their academic degrees receive a more substantial amount of feedback from the other members against those who have used the community over a similar period and posted content of the same quality without any flairs. Among the users with flairs, the average effect size of exposing the highest academic degrees (such as a Ph.D. or M.D.) was three times greater than that of showing lower academic degrees such as an M.S. or indicating graduate student status. In r/EverythingScience, which covers more casual topics on science, the benefits of revealing the offline social status disappeared but remained only for the highest degree group.

Our research sheds light on the design choices for reputation mechanisms for future online communities. Exposing one's offline social status may induce cognitive bias on the perceived level of content quality, such as the halo effect (Kahneman and Egan 2011), which in turn can draw a disproportionate amount of community feedback to only certain content produced by individuals with high status. While this skewed attention might have been intended, it could make it difficult for online communities to promote high-quality content on the merit of the content and therefore calls for more careful designs in introducing offline status online. We hope that our findings will benefit the practitioners and designers of online reputation systems.

Related Works

Community Feedback as a Key for Participation

Motivating users to actively participate and contribute high-quality content is critical for a thriving online community (Malinen 2015). Among the various factors that facilitate member participation, usage motivation (Arguello et al. 2006), personality traits (Nov et al. 2013), and social networks have been found to play a role. For instance, social support boosts future engagement, as demonstrated in a study in which people committed to long-term health behaviors when they form connections (Park et al. 2016). However, social networks can discourage user participation when members receive negative feedback from the community, as shown in a study related to toxic behavior in online games (Shores et al. 2014). Negative feedback can be fostered throughout a community, as negatively evaluated users are likely to rate others more negatively (Cheng, Danescu-Niculescu-Mizil, and Leskovec 2014; 2015). These studies suggest that community feedback and particularly a positive perspective are crucial for user participation in online communities, which is broadly the topic of this research.

Level of Anonymity

Significant research has investigated the anonymity level of individuals within a community. One extreme is full anonymity where one member cannot be distinguished from another by any identifier. Complete anonymity is detrimental for the credibility of a system (Rains 2007), and it provokes a negative culture involving toxic behaviors (Kilner and Hoadley 2005; Suler 2004; Kwak, Blackburn, and Han 2015). Some studies find that full anonymity can promote open conversations in public discourse (Bernstein et al. 2011). The other extreme is complete openness in which one's offline identity is exposed in the online world. Complete transparency may increase the trust and accountability of the system (Kusumasondjaja, Shanka, and Marchegiani 2012) and promote a polite communication culture (Millen and Patterson 2003). A study investigating Amazon reviews reported that users more positively rate reviews containing identity-related information (Forman, Ghose, and Wiesenfeld 2008). Another study showed that fake reviews could be better detected when the author's identity information is disclosed (Munzel 2016). However, full openness may restrain individuals from freely sharing their views and acting naturally, as several studies have reported that privacy concerns might hinder member participation (Frost, Vermeulen, and Beekers 2014; Liao et al. 2012).

Most online communities advocate pseudonymity, representing a middle ground between complete anonymity and full openness, in which members create pseudonyms to build their identity. However, pseudonymity has a downside because members may develop more than one character by creating multiple user accounts, which are called *sockpuppets* (Kumar et al. 2017). Not all sockpuppets are harmful as multiple accounts are sometimes useful for people seeking social support (Andalibi et al. 2016; De Choudhury and De 2014) by allowing them to build temporary identities (Leavitt 2015). Nonetheless, malicious attackers can exploit such multiple accounts with fake identities to harm other community members (Wang et al. 2013).

Reputation and Social Status in Online Communities

Reputation systems promote quality content in online communities. For example, the StackOverflow website aggregates votes on the historical answers of each member as a measure of reputation (Bosu et al. 2013). Similarly, Reddit employs a reputation system called "Karma" based on the voted scores of its members' historical activities. Communication studies report that online reputation creates social status—i.e., 'an actor's relative standing in a group by prestige, honor, or deference' (Sauder, Lynn, and Podolny 2012)—within an online community, which further drives active participation and altruistic behaviors by motivating members to achieve higher status (Lampel and Bhalla 2007; Bateman, Gray, and Butler 2011). Online reputation systems can also help members to more easily identify experts on a given topic. Studies concerning Yahoo! Answers (Shah and

Pomerantz 2010) and StackOverflow (Movshovitz-Attias et al. 2013) have shown that online reputation scores are directly predictive of answer quality in question answering communities.

The main challenge in implementing any online reputation mechanism is the cold start problem. One method to resolve this weakness is to transfer the offline social status, which has a hierarchy that correlates with expertise, such as academic profiles and job affiliations. MathOverflow.net uses real names instead of pseudonyms, which allows the members to match each other's offline identity to the online profile. A study involving 3,470 users of MathOverflow found a correlation between voting scores and offline social status (Tausczik and Pennebaker 2011), suggesting that transferring offline status to online communities might be useful in promoting content with excellent quality.

Nonetheless, unexpected consequences might occur that could hurt the health of online communities. Social status can be divided into hierarchical levels. For example, a Ph.D. is higher than an MS, and a senior engineer is higher than an entry-level engineer. Introducing this relative difference to an online community provides a new social status structure among the members. Hence, community members can evaluate quality more generously when a user with high status posts content as observed in an interview-based study that found that Quora users perceive answers written by experts who have first-hand information on a topic to be more authoritative (Paul, Hong, and Chi 2012). If users with high-status receive more feedback than general users when they post content of equal quality, this may not be a desirable outcome, particularly in systems that address many niche areas of content. Despite its importance, to the best of our knowledge, no prior study has conducted an in-depth examination of the effect of introducing offline social status online on the feedback behaviors of online community members. This study aims to fill this gap using data-driven approaches.

Problem and Data

Problem Definition

We pose the following research question:

> *RQ. Does exposing one's offline social status to an online community lead to more feedback?*

To answer this question, we utilized logs from Reddit, which is a link-sharing and discussion website that has over a million sub-communities dedicated to specific subjects called *subreddits*. Subreddit names are prefixed with 'r/,' such as r/Sports or r/Gaming. Reddit members communicate by sharing web links and commenting on shared links. Each subreddit has a small group of moderators who oversee the shared content. The moderators decide the rules (e.g., terms of use violation) and pinned content (e.g., member notices). The moderators can also define the tags and flairs — what members can show next to their profiles or posts. For example, in the r/loseit subreddit, where members share the common goal of losing weight, members adopt a flair that displays their weight loss progress (e.g., "-50 lb"). Such flairs and tags have been shown to exert a positive peer effect (Cunha, Weber, and Pappa 2017).

In this study, we refer to the two Reddit communities, r/Science and r/EverythingScience, as Sci and Eve, respectively. The topics of these communities are either scientific manuscripts or news articles related to scientific research. The community members judge the quality of the shared content by voting up or down and, if they wish, they can participate in the comment threads associated with each post. The code of conduct is similar to that in other Reddit communities; however, their topics are limited to science. While a post may link to anything in Eve as long as its focus is on science, Sci posts are limited to peer-reviewed scientific articles that have been published in the last six months. Interestingly, the moderators of these subreddits have adopted a flair mechanism that lets members expose their education degrees and domains of expertise, such as 'PhD | Psychology', as shown in Figure 1.

The logs containing profiles of users with different degree types and the kinds of feedback their content received lend these subreddits to a natural experiment. Analyzing the flair dataset is advantageous for several reasons. First, the behavioral traces span nearly four years. Therefore, we are able to cover various types of scientific content over time, which is not feasible in randomized trials or interview studies. Second, their topics are limited to science, and hence, the effects of the content topics are better controlled. Third, the two similar-yet-different communities provide an opportunity to investigate the common or differing impacts of exposing offline status online. Notably, this paper is based on data-driven approaches and, hence, allows us to objectively measure the changes in community feedback without the risk of possible biases such as the social desirability bias (Nederhof 1985) that can arise in surveys or interview-based studies.

Data

The data analyzed were obtained from a well-known database that operates on the Reddit API (Baumgartner et al. 2020). We downloaded the primary action logs, posts, and comments that appeared on the two subreddits since the launch of Eve in January 2014 until December 2017. The dataset contains information regarding (1) the authors' information (e.g., name and flair); (2) the content information (e.g., identifier, timestamp, text content, and net score); (3) the crawl information (e.g., crawled time), etc.

We carefully cleaned and sanitized the data. One challenge was to estimate the exact time when a user adopted a flair because the Reddit API does not provide this information. However, the crawled dataset contained an individual's flair status at the time of data collection instead of the time of posting. We downloaded multiple snapshots of the Reddit crawls and repeatedly checked each user to determine the times when she was last seen without a flair. Finally, for each user, we could identify the earliest data collection time when her post appeared with a flair, and we utilized all logs only after that time point. Our decision to exclude the data points is a conservative choice that guarantees reliable flair information at the time of posting.[4] This step removed 6,224 and 2,977 posts and 86,078 and 7,104 comments from Sci and

[4] None of the posts without a flair were removed from the data.

Subreddit	Type	Data entry (count)	Users (count)	Average per user (count)	Mean (score)	Median (score)	Std. Error (score)
r/Science	Posts	193,441	73,233	2.641	144.04	1.0	3.302
	Comments	2,487,480	543,524	4.576	9.446	1.0	0.434
r/EverythingScience	Posts	55,966	12,724	4.398	20.842	3.0	0.059
	Comments	130,542	31,237	4.179	5.757	2.0	0.091

Table 1: Descriptive statistics of posts and comments in r/Science and r/EverythingScience (Period: 2014/01-2017/12)

Eve, respectively. The Reddit snapshots had been crawled regularly over several years, with a median time difference of 27.2 days. Hence, the flair adoption time could be estimated with this margin of error.

Table 1 summarizes the data statistics after the above filtering step and reveals that we have ample data regarding the posts and comments for the analysis. The score distributions were heavily skewed, and only a small proportion of the posts were popular, as has been observed in other Reddit communities (Gilbert 2013). The median scores of the posts and comments in Sci are both 1.0.[5] The median scores are slightly higher in Eve, i.e., 3.0 for posts and 2.0 for comments. These statistics do not diverge much across the whole period, suggesting that community feedback regarding content is comparable over time.

Type	User		Post		Comment	
	Sci	Eve	Sci	Eve	Sci	Eve
DR (Doctoral)	940	129	895	203	20,871	467
MS (Master)	657	69	689	893	5956	405
GS (Grad Student)	1,104	112	567	270	10,370	362
BS (Bachelor)	1,185	72	233	23	6,136	196

Table 2: Distribution of degree types in Sci and Eve

The academic degree information was extracted from the collected flair tags via regular expression. We grouped similar degree types; for instance, all variants of doctoral degrees (e.g., Ph.D. and PharmD) and positions requiring an equivalent degree (e.g., Professor) were grouped. After iterative processes, we obtained the following four main degree types based on the natural hierarchy: DR (doctoral degree), MS (master's degree), GS (currently a graduate student), and BS (bachelor's degree). We further validated this automated process with the correct labels of the sampled users, which were provided by the moderators of those subreddits. Table 2 describes the final counts of 3,886 flair users, whose degree information was successfully retrieved. A user can only achieve a flair in Eve through the process of the Sci subreddit by the community design such that flair users in Eve are automatically a subset of Sci flair users. In the analysis, we exclude one flair type, i.e., the AMA (Ask Me Anything) flair, because it is given to science celebrities who draw a considerable amount of attention from the community.

This work is an observational study based on data gathered through the public Reddit API; hence, obtaining IRB approval is not necessary. The researchers did not intervene with the Reddit users nor process identifiable private information in this study.

Direct Comparison Across Flairs

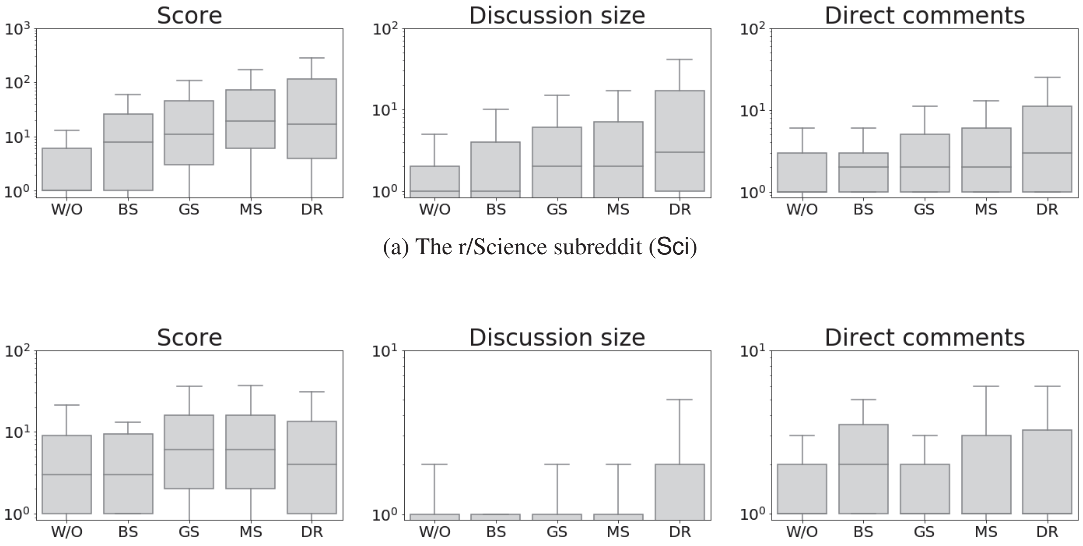
To answer the research question of whether flairs lead to any changes in member response, we first compared the amount of community feedback against the levels of academic degrees exposed via flairs. The following features can quantify the amount of community feedback:

- *Score* (integer value): The net scores of the target post are calculated as #upvotes − #downvotes. This variable captures how positively or negatively the community members evaluated the target post. While Reddit adds fuzziness to the number of upvotes and downvotes to prevent spam bots, the score, which is the difference between these votes, remains unchanged.
- *Discussion Size* (integer value): The number of comments on the target post represents the community members' participation level in the discussion.
- *Direct Comments* (integer value): This variable reflects the number of root comments (i.e., depth=1) on the discussion tree of the target post. Compared to the discussion size, this variable estimates the amount of direct feedback on the post. Posts with zero comments (45.4% in Sci, 71.2% and Eve) were excluded to capture a distinct pattern against the discussion size.

Figure 2 shows the log-scaled boxplot of the amount of community feedback that each degree group received in the two subreddits. For comparison, we show the distribution of the members without any flairs, denoted as "W/O" in the figure. The three feedback measures are skewed, such that they do not meet the normality assumption. Therefore, instead of using a one-way ANOVA, the Kruskal-Wallis test was employed, followed by Dunn's test as a post hoc analysis to identify the pairs of degree groups with a significant difference.

Figure 2(a) suggests that exposing academic degrees through flairs is correlated with the amount of feedback received by members on their posts in Sci. All three measures showed a significant difference (p<0.001 in each degree group combination). Compared with those without flairs, the flair posts received a larger amount of community feedback except for the bachelor's degree group (p<0.001 in all comparisons). Eve, however, exhibits a distinct pattern in Figure 2(b). Most of the users received a similar amount of feedback, and we only discovered substantial differences between the MS and GS groups. The master's degree group was likely to enjoy a higher score than those

[5] The initial score of the comments in Reddit was 1.



(a) The r/Science subreddit (Sci)

(b) The r/EverythingScience subreddit (Eve)

Figure 2: Distributions of the community feedback measures across academic degree groups. "W/O" represents no flair.

without a flair (p<0.001); however, it drew a smaller discussion size (p<0.001). Additionally, the posts written by the current graduate students were likely to receive a higher score (p<0.001) and engage a larger audience in the discussion (p<0.001) compared to those that did not expose their status.

The above results demonstrate that a type of academic degree is likely to yield different effects. Sci is a clear example where the amount of community feedback correlates with the levels of academic degrees. At a glance, the group's members seem to judge the posts by high degree holders to be more attractive, as denoted by the enhanced feedback. However, users with ordinary degree holders, such as a bachelor's degree, receive even less feedback than those without any flairs.

However, the more general-science-oriented Eve community did not show the same pattern. We also note that there might exist confounding factors in these observations as has been reported by the previous studies on the effect of content quality (Cheng, Danescu-Niculescu-Mizil, and Leskovec 2014; Keneshloo et al. 2016; Singer et al. 2016) and user reputation (Bosu et al. 2013; Shah and Pomerantz 2010; Movshovitz-Attias et al. 2013). What if those with a doctoral degree received more responses not because of their flairs but because their content was significantly more interesting? These degree holders may have posted better content that discusses the latest and most significant advances in science, which the Sci community values. In contrast, academic degrees may not be a determinant of a successful post on the Eve community because anyone can engage in general posts related to science.

To test the possibility of confounding effects by content quality and user reputation, we measured the (1) lexicon count and (2) estimated user karma[6] of every post, and we investigated their correlations with the three measures on community feedback. Table 3 presents the Spearman's correlation of post quality and online reputation with our measures on community feedback. In Sci, the amount of community feedback exhibits a moderate and high level of correlations with content quality and online reputation, which make it possible to confound the effects of academic degrees on community feedback that was found in Figure 2. Eve only exhibits a negligible level of correlations, and, in combination with the previous findings in Figure 2, these findings show that there are no clear signals that affect the size of community feedback in the subreddit.

Community	Content Quality		User Reputation	
Feedback	Sci	Eve	Sci	Eve
Score	0.394	0.133	0.449	0.110
Discussion Size	0.309	0.113	0.232	-0.075
Direct Comments	0.322	0.140	0.296	0.041

Table 3: The Spearman's correlation of content quality (measured by lexicon count) and user reputation (measured by aggregate score) on community feedback.

The results in this section showed that there is a correlation between exposing academic degree and community feedback; however, the findings also suggest that content quality and user reputation could confound the relationship. A Reddit post may receive great feedback because of the flairs or because of the confounding variables. In the following section, we apply a method to control this unwanted signal in our observational data.

[6] As the Reddit API provides only user-level karma scores at the time of API call, we estimated it as the cumulative voted scores of past posts uploaded by the same user until the time of post upload in each community.

Estimating the Effects of Flairs

To infer the effects of exposing one's academic degree on community feedback while controlling for the confounding influences of the covariates such as content quality features, we utilized the propensity score matching framework (Rosenbaum and Rubin 1983), which is widely used in observational studies due to its ability to mitigate selection bias (Guo and Fraser 2015). The framework consists of three steps: (1) propensity score modeling, (2) propensity score matching, and (3) estimating a treatment effect after a successful balance check. First, for each scenario that aims at testing whether exposing one of the degree types, defined in Table 2, affects the amount of community feedback, a propensity model is trained to estimate the likelihood of having the treatment condition (exposing a degree type) from covariate features. Second, for each post with the treatment condition, one or more appropriate instances are matched among the control groups, which are posts without a flair, by utilizing the treatment probability estimated via the propensity model. Third, we check whether the covariate distribution of the treatment group is statistically identical to that of the control group. If it passes the balance test on the covariate distribution, the propensity score matching framework enables the estimation of the effects of the treatment condition on community feedback. For the case in which a treatment group and its matched control group have different distribution on any of covariates, the analysis framework does not allow for the estimation of the effect of a treatment condition.

The statistical analysis framework separates out the effects of the treatment conditions (i.e., exposing one's academic degree) on community feedback from the confounding influences of the covariates (e.g., content quality features), by matching appropriate instances within the control groups to each treatment unit (i.e., degree group). After this step, the covariate distribution of the treatment group should become statistically identical to that of the control group. This process approximates randomized controlled trials in which the treatment and control groups are randomly distributed with regard to covariates. Hence, the risks of confounding effects due to covariates are minimized.

Propensity Score Modeling

The matching process is based on the estimated propensity of each post to expose each degree type through a flair (i.e., DR, MS, GS, and BS) compared to posts without flairs as the control group. The propensity score can be modeled by any function that produces a likelihood of receiving treatment from covariates, ranging from 0 to 1. We utilized a logistic regression with Lasso regularization ($\lambda = 0.001$) because this approach is known to identify essential features among a large pool of variables in the online community research (Cunha, Weber, and Pappa 2017; Park et al. 2017). This step models the propensity scores after discarding less important covariates that can vary against experiment settings.

Covariate Features Propensity score analysis makes the conditional independence assumption of causal inference (Cunningham 2018), suggesting that the likelihood of having a treatment condition must be almost the same as that of being random. That is, to estimate a treatment effect accurately, one should model the propensity of receiving treatment using as many confounding variables as possible. Therefore, in addition to lexicon count and estimated user karma, which correlates with community feedback in the previous section, we considered the following variables as covariates, which quantify content quality (Cheng, Danescu-Niculescu-Mizil, and Leskovec 2014; Keneshloo et al. 2016; Singer et al. 2016) and user reputation (Bosu et al. 2013; Shah and Pomerantz 2010; Movshovitz-Attias et al. 2013) from diverse perspectives.

- *Catchiness* (numeric value): Catchy titles entice a larger audience and thus are likely to incur feedback regarding the target post. We measured catchiness by applying a pre-trained machine learning classifier (Chakraborty et al. 2016) that detects clickbait news headlines to post titles. Since the model was trained based on news articles, we manually tested its adaptability to posts in Sci and Eve by sampling 50 post articles from each subreddit. The results showed a moderate agreement rate of 0.46 and 0.32 as measured by Cohen's Kappa, suggesting that the pre-trained model can estimate the catchiness of the post titles in the subreddits with a low margin of error.
- *Readability* (numeric value): The Gunning-Fog score (Gunning 1952) estimates the years of formal education a person needs to understand the text in the first reading; thus, this score is widely used to measure online text quality (e.g., news articles (Keneshloo et al. 2016) and Reddit (Singer et al. 2016)). A higher value indicates that a given text is written with more complex lexicons and has longer sentences. We also considered lexicon count as a covariate because it affects the readability of each post. We applied the Python textstat library to the post titles to measure the two variables.[7]
- *Sentiment* (numeric value): Sentiments conveyed through post titles are known to affect the extent to which the user is likely to click the link (Tatar et al. 2014; Reis et al. 2015; Ferrara and Yang 2015). We utilized positive and negative sentiments as measured via VADER sentiment lexicons (Hutto and Gilbert 2014).
- *Topic Distribution* (numeric value): Certain topics might attract more member attention. We measured the topic distribution of a post by applying the latent semantic indexing (Papadimitriou et al. 2000) topic modeling method to the title with the parameter of the number of topics set to 50. The results were almost the same across several variations on a larger number of topics (i.e., 100 and 150).
- *User Reputation* (numeric value): A user's reputation can influence the evaluation of the future content uploaded by that user (Shah and Pomerantz 2010; Movshovitz-Attias et al. 2013). Reddit reveals the karma score that quantifies user reputation in the user profile, which might lead to a different amount of attention. We estimated the user-

[7] https://pypi.org/project/textstat/

level karma score at the time of post upload to be the cumulative scores of the previous posts in Reddit uploaded by the user. Additionally, a community-specific reputation was also similarly estimated by relying only on the posts within the community.

Propensity Score Matching

We matched the control units (i.e., posts without flairs) to each treatment unit (i.e., posts with flairs) based on the propensity score. The primary goal of matching is to obtain a balanced set, allowing for the ruling out of the effects of covariates on the outcome variables. While there are various options for matching, such as exact matching and caliper matching, we applied the k-nearest neighbor algorithm ($k = 5$) to each treatment unit. The similarity is measured by the *Mahalanobis distance*, which is a normalized distance measure between two targets in multivariate space.

A successful matching process should yield a balanced set in terms of confounding factors. To ensure the success of this process, we measured the standardized mean difference of the propensity scores d_c for each covariate c. As a rule of thumb, two groups are considered "balanced" if the absolute value of the standardized mean difference is below 0.1 (Austin 2011). We repeatedly adjusted the hyperparameter values (e.g., λ in Lasso, and k in the nearest neighbor algorithm) until we achieved balanced matching.

Estimating the Average Treatment Effects

We estimate the effect of a treatment condition (e.g., exposing doctoral degrees) on each outcome variable (e.g., score and other member feedback). The estimated average treatment effect (EATE) of an outcome variable y was measured by the following equation:

$$\sum_{t}^{T}\sum_{m}^{M_t}\Big(\frac{y_t - y_m}{N_{M_t}}\Big)/N_T \qquad (1)$$

where T is a set of treatment units and M_t is a set of control units matched to treatment unit t. y_t and y_m are the outcomes measured for t and m, respectively. N_T and N_{M_t} are the numbers of treatment units and matched control sets, respectively. As we isolate the effects of flairs from the covariates through propensity score matching, the EATE is interpreted as the number of benefits or disadvantages that are gained by disclosing academic degrees, compared to a post with the same quality that is uploaded by a user with a similar reputation yet with no flair. As the size of the effect may vary by treatment units (posts), the standard error of the treatment effect is also reported to carefully interpret the results.

Matched Results

How much advantage (or disadvantage) does a user gain upon adopting a flair compared to another user with a similar reputation and who shares content of similar quality without any flair? Table 4 shows the answer to this question; the average treatment effect of having a flair was computed by the *Equation (1)*. This analysis requires finding a balanced control group for each treatment group. The balance analysis revealed that the BS group did not meet this condition, leading to bias in the match analysis. Therefore, we show only the results for the DR, MS, and GS groups. The matching analysis obtained several key findings.

Degree	Score	Discussion Size	Direct Comments
DR	**188.16** (43.46)	**41.82** (8.19)	**15.77** (3.5)
MS	64.65 (54.62)	13.75 (6.64)	3.28 (3.09)
GS	67.96 (41.04)	11.19 (5.14)	2.34 (2.21)

(a) Sci community

Degree	Score	Discussion Size	Direct Comments
DR	**13.79** (6.02)	**2.06** (0.78)	0.24 (0.33)
MS	-0.83 (2.14)	-0.23 (0.26)	**0.32** (0.22)
GS	3.3 (3.44)	0.57 (0.43)	-1.98 (1.94)

(b) Eve community

Table 4: The mean effect of exposing the academic degree with the standard error in parenthesis. The largest value appears in bold text.

The first set of observations is based on the Sci subreddit. The individuals with the highest education level of doctoral degree were likely to gain better feedback from the community members even when they offer content of the same quality. The average treatment effects on all three variables related to community feedback were positive with a significant magnitude, and the EATE on the post score was 188.16 with a standard error of 43.46, indicating that doctorate flair could trigger more community feedback on the post. The other education levels of master's degree and graduate also show the desired effect across the three variables but to approximately one-third EATE of the former.

The next set of observations is on the Eve subreddit. We confirm a similar positive outcome in the doctoral degree group. Still, the average treatment effect was 1 to 2 orders of magnitude smaller than that observed in Sci, which makes the effects on direct comments negligible. The reduced magnitude is possibly due to the difference in the popularity of these two communities (notably, Eve has two orders of magnitude fewer subscribers.) In the master's degree and graduate student groups, no significant effect on community feedback was observed from exposing academic degrees online. This notable difference may be due to the group's community cupture, which covers more casual science topics, unlike Sci, which allows only peer-reviewed articles to be uploaded. Due to such disparity, its members may no longer be biased to hold positive perceptions with regard to the doctorate group, the members of which are expected to hold expertise in their respective domains.

In summary, the propensity score matching reveals a consistent trend across the two communities: exposing a higher education degree was likely to incur more member feedback on uploaded posts; however, the mere existence of an academic degree did not guarantee the same effect. Even though the flair mechanism that was intended to invigorate community members and help identify high-quality content produced unexpected biases concerning community feedback, its results may vary against the code of conduct in each community.

Discussion and Conclusion

The reputation mechanism is commonly used in many online communities. To overcome the cold-start problem of online reputation and further promote high-quality content, some communities borrow users' offline status from the real world, as shown in the two Reddit communities studied in this work. In r/Science, exposing academic degrees through flairs corresponds to a more substantial amount of community feedback compared to the matched posts of similar quality that are uploaded by the users with almost same reputation. Moreover, the most significant effects were observed for the highest academic degrees (e.g., Ph.D.). The *halo effect* (Kahneman and Egan 2011), which is a type of cognitive bias where one trait contributes to the overall judgment of a person, may partially explain these underlying dynamics. While promoting the high-quality content of educated users might have been intended by the community moderators, the disproportionate amount of benefits toward the education status might cause feelings deprivation in the overall population because the other users could not obtain sufficient feedback despite sharing content of the same quality. Feeling under-served and receiving steadily less feedback can further cause users to resort to lurking within the community (rather than participating) or leaving (Malinen 2015).

The studied reputation mechanism had different effects depending on community types. The Eve subreddit has a subtle difference in the types of information its members are allowed to share compared with Sci. Whereas the latter accepts only discussions regarding peer-reviewed scientific articles, the former allows casual topics. As a result, in Eve, academic degrees may not achieve authority except for the highest degree group, which also obtained the largest effect size in Sci. The varying results across the subreddits imply that showing any offline status would not always give power to the members in any community. For example, it appears unlikely that the educational status used in those scientific communities could be useful in r/Gaming. To properly supplement the online reputation mechanism, community moderators would have to select an appropriate offline status that is well-aligned with their community culture to introduce the desired effects.

While the findings and the supporting theory in social psychology provide a plausible explanation for the effects of exposing an offline social status online, the results should be carefully interpreted based on their own merits. The use of propensity score matching enables researchers to obtain a balanced distribution of each of the covariates between a treatment group and its corresponding control; however, it cannot rule out the effects of *unobserved* covariates (Guo and Fraser 2015). Had there been another variable that significantly affected the amount of community feedback, the findings of this paper would not show the true effect of exposing academic degrees online. For example, we found that excluding the variables of user reputation from the covariates even causes the direction of the effects of exposing a master's degree to be opposite to that of the reported findings. However, based on the literature, we identified the possible factors that could affect the amount of community feedback as the covariates of the analysis framework. Therefore, we believe that this study can accurately approximate the effects of exposing academic degrees online by minimizing the impact of the significant confounders.

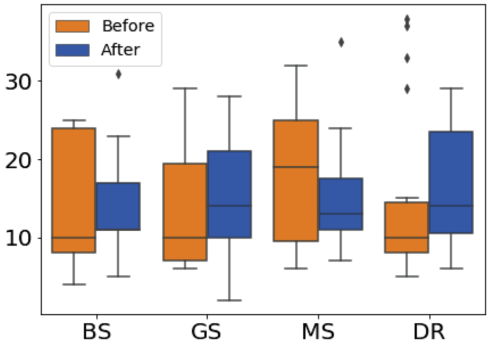


Figure 3: Change of lexicon count after flair adoption. BS means users with a bachelor's degree, GS means users who are currently a graduate student, MS means users with a master's degree, and DR means users who have a Ph.D. or an equivalent degree.

Changes in Post Quality After Flair Adoption

This paper examined the research question of how other members react to posts written by flair adopters and identified a meaningful stance change upon observing a flair. What about the flair adopters? Do flair adopters also behave differently once their academic degrees are revealed to the public? Based on the relevant studies that consider social status and text writing style (Sexton and Helmreich 2000; Hymes 2005), we hypothesize that flairs render adopters aware of their social status and, consequently, lead them to change their writing style. To test this hypothesis while minimizing the risks of temporal effects, we searched for any signal of change in post quality before and after adopting the flair (i.e., between the last post without a flair and the first post with a flair).

We compared the quality features that were used as covariates in this study of the before- and after- posts of a total of 50 users in Sci who posted at least once before and after flair adoption.[8] The nonparametric Mann-Whitney U test with continuity correction was utilized for the comparison. While based on a small data sample, the writing style of the users was found to have changed once they adopted a flair. As shown in Figure 3, the doctoral degree group began to write longer post titles once their name appeared next to

[8] The median post count per user is 1 in both communities.

'doctoral degree.' Exposing their degrees online may result in the highest social status and, in turn, cause them to feel a commitment to write post titles that are well suited to their status. We were not able to observe meaningful changes in the other groups, possibly due to the small number of users per each group.

Limitations and Future Work

This study examined the instantaneous authority enjoyed by members due to exposing their offline social status information to the online world. Two Reddit communities were reviewed to measure the impact of academic accomplishments: B.S. degree, M.S. degree, Ph.D. degree, and currently in graduate school. Applying propensity score measurements on this data allowed us to compare patterns across these academic achievements. However, the findings of this paper may not generalize to other types of offline social status such as occupation or affiliation. This study also bears a risk of Simpson's paradox as noted by recent studies (Alipourfard, Fennell, and Lerman 2018; Lerman 2018) in which a trend appears or disappears when aggregated. Unfortunately, we were unable to repeat the same analysis based on a more granular form of academic degree or other status type due to the small number of flair users. Future studies could test the generalizability of our findings by collecting a more extensive dataset on similar reputation mechanisms.

The chosen content quality measures bind the findings in this paper. Other methods could be applied to estimate the content quality. One could measure the coverage of news articles through the Altmetric API,[9] as a recent study employed this measure to identify the characteristics of popular scientific articles (MacLaughlin, Wihbey, and Smith 2018). Other than propensity score analysis, synthetic controls or a difference-in-difference framework could help in the discovery of causal effects (Cunningham 2018). In-depth interviews could also be employed to facilitate a better understanding of the psychological impact of observing flairs on Reddit. Future research may seek to answer whether and how the ad hoc authority enjoyed by users transfers across areas (e.g., a psychology scholar posting content on physics).

Investigating *who* discloses their offline status online could provide an exciting direction. Are high-status users more likely to share their offline identity? How does such disclosure affect future behaviors? Would users begin to censor their posts? While some studies report that adopting a membership badge decreases the likelihood of user churn (Anderson et al. 2013; Hamari 2017), we expect that introducing an offline status online could have distinct effects because it connects offline identity to online identity. Exposing a high level status might cause users to feel committed to posting more high-quality content. Members who cannot attain a high level of badges could also feel isolated and hence leave the community. Based on a sizable dataset, future studies could better explore such psychological and social effects of disclosing an offline status online.

Another research direction is exploring how online reputation scores and offline social status *interact*. In sociological theory, one's reputation enhances one's social status in the real world, and the same holds for one's online reputation (Lampel and Bhalla 2007). Does an online reputation have similar effects on the biases leveraged by offline social status? How does disclosed offline status affect online reputation? Is offline social status more useful in discerning high-quality content than online reputation? A recent study found that small manipulations of Reddit post scores ultimately obtained significant changes in the end (Glenski and Weninger 2017), which suggests that previously attained scores may also induce biases toward positive evaluations. In the future, we plan to answer these questions to provide extensive insights into the designs of reputation mechanisms in online communities.

We are also interested in exploring the other biases that could be introduced by the studied reputation mechanism across different platforms. For instance, revealing the author's affiliations in a single-blind review policy of peer-reviewing systems has been shown to bias outcomes: reviewers perceived papers written by authors from well-known institutions to be of higher quality (Tomkins, Zhang, and Heavlin 2017). Social media and microblogging platforms, such as Twitter, are used for a wide range of conversations, from personal thoughts and scientific discussions to public discourse. Given that individuals often expose their job affiliations in their profiles, we aim to determine how viewing and/or sharing offline social status affects community members' perceptions of the content they see or wish to share further in the network.

Acknowledgement

Majority of this work was done while K. Park was at QCRI. We thank to the moderators of r/Science who generously provided the sampled ground-truth dataset for evaluating our approach on estimating flair information. We also thank to anonymous reviewers whose comments have greatly improved this work. H. Song and M. Cha were supported by the Institute for Basic Science (IBS-R029-C2) and the Basic Science Research Program through the National Research Foundation of Korea (No. NRF-2017R1E1A1A01076400).

References

Alipourfard, N.; Fennell, P. G.; and Lerman, K. 2018. Can you Trust the Trend?: Discovering Simpson's Paradoxes in Social Data. In *Proc. of the WSDM*, 19–27. ACM.

Andalibi, N.; Haimson, O. L.; De Choudhury, M.; and Forte, A. 2016. Understanding social media disclosures of sexual abuse through the lenses of support seeking and anonymity. In *Proc. of the CHI*, 3906–3918. ACM.

Anderson, A.; Huttenlocher, D.; Kleinberg, J.; and Leskovec, J. 2013. Steering user behavior with badges. In *Proc. of the WWW*, 95–106. ACM.

Aral, S. 2014. The problem with online ratings. *MIT Sloan Management Review* 55(2):47.

Arguello, J.; Butler, B. S.; Joyce, E.; Kraut, R.; Ling, K. S.; Rosé, C.; and Wang, X. 2006. Talk to me: foundations for

[9] https://api.altmetric.com/

successful individual-group interactions in online communities. In *Proc. of the CHI*, 959–968. ACM.

Austin, P. C. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research* 46(3):399–424.

Bateman, P. J.; Gray, P. H.; and Butler, B. S. 2011. Research note—the impact of community commitment on participation in online communities. *Information Systems Research* 22(4):841–854.

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The pushshift reddit dataset. *arXiv preprint arXiv:2001.08435*.

Bernstein, M. S.; Monroy-Hernández, A.; Harry, D.; André, P.; Panovich, K.; and Vargas, G. G. 2011. 4chan and/b: An Analysis of Anonymity and Ephemerality in a Large Online Community. In *ICWSM*.

Berry, G., and Taylor, S. J. 2017. Discussion quality diffuses in the digital public square. In *Proc. of the 26th International Conference on World Wide Web*, 1371–1380.

Bosu, A.; Corley, C. S.; Heaton, D.; Chatterji, D.; Carver, J. C.; and Kraft, N. A. 2013. Building reputation in stackoverflow: an empirical investigation. In *Proc. of the MSR*, 89–92. IEEE Press.

Brandtzæg, P. B., and Heim, J. 2008. User loyalty and online communities: why members of online communities are not faithful. In *Proc. of the INTETAIN*, 11. ICST.

Chakraborty, A.; Paranjape, B.; Kakarla, S.; and Ganguly, N. 2016. Stop clickbait: Detecting and preventing clickbaits in online news media. In *Proc. of the ASONAM*, 9–16. IEEE Press.

Cheng, J.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2014. How Community Feedback Shapes User Behavior. In *ICWSM*.

Cheng, J.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2015. Antisocial Behavior in Online Discussion Communities. In *ICWSM*.

Cunha, T.; Weber, I.; and Pappa, G. 2017. A Warm Welcome Matters!: The Link Between Social Feedback and Weight Loss in/r/loseit. In *Proc. of the WWW Companion*, 1063–1072.

Cunningham, S. 2018. Causal inference: The mixtape.

De Choudhury, M., and De, S. 2014. Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity. In *ICWSM*.

Ferrara, E., and Yang, Z. 2015. Quantifying the effect of sentiment on information diffusion in social media. *PeerJ Computer Science* 1:e26.

Forman, C.; Ghose, A.; and Wiesenfeld, B. 2008. Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research* 19(3):291–313.

Frost, J.; Vermeulen, I. E.; and Beekers, N. 2014. Anonymity versus privacy: selective information sharing in online cancer communities. *Journal of medical Internet research* 16(5).

Gilbert, E. 2013. Widespread underprovision on Reddit. In *Proc. of the CSCW*, 803–808. ACM.

Glenski, M., and Weninger, T. 2017. Rating effects on social news posts and comments. *ACM Transactions on Intelligent Systems and Technology* 8(6):78.

Gunning, R. 1952. The technique of clear writing.

Guo, S., and Fraser, M. W. 2015. *Propensity score analysis*. Sage.

Hamari, J. 2017. Do badges increase user activity? a field experiment on the effects of gamification. *Computers in human behavior* 71:469–478.

Hutto, C. J., and Gilbert, E. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM*.

Hymes, D. 2005. Models of the interaction of language and social life: toward a descriptive theory. *Intercultural discourse and communication: The essential readings* 4–16.

Kahneman, D., and Egan, P. 2011. *Thinking, fast and slow*, volume 1. Farrar, Straus and Giroux New York.

Keneshloo, Y.; Wang, S.; Han, E.-H.; and Ramakrishnan, N. 2016. Predicting the popularity of news articles. In *Proc. of the SDM*, 441–449. SIAM.

Kilner, P. G., and Hoadley, C. M. 2005. Anonymity options and professional participation in an online community of practice. In *Proc. of the CSCL*, 272–280. International Society of the Learning Sciences.

Kumar, S.; Cheng, J.; Leskovec, J.; and Subrahmanian, V. 2017. An army of me: Sockpuppets in online discussion communities. In *Proc. of the WWW*, 857–866.

Kusumasondjaja, S.; Shanka, T.; and Marchegiani, C. 2012. Credibility of online reviews and initial trust: The roles of reviewer's identity and review valence. *Journal of Vacation Marketing* 18(3):185–195.

Kwak, H.; Blackburn, J.; and Han, S. 2015. Exploring cyberbullying and other toxic behavior in team competition online games. In *Proc. of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, 3739–3748. New York, NY, USA: ACM.

Lampe, C., and Resnick, P. 2004. Slash (dot) and burn: distributed moderation in a large online conversation space. In *Proc. of the CHI*, 543–550. ACM.

Lampel, J., and Bhalla, A. 2007. The role of status seeking in online communities: Giving the gift of experience. *Journal of computer-mediated communication* 12(2):434–455.

Leavitt, A. 2015. This is a throwaway account: Temporary technical identities and perceptions of anonymity in a massive online community. In *Proc. of the CSCW*, 317–327. ACM.

Lerman, K. 2018. Computational social scientist beware: Simpson's paradox in behavioral data. *Journal of Computational Social Science* 1(1):49–58.

Liao, Q.; Pan, Y.; Zhou, M. X.; and Gan, T. 2012. Your space or mine?: Community management and user participation in a chinese corporate blogging community. In *Proc. of the CSCW*, 315–324. ACM.

MacLaughlin, A.; Wihbey, J.; and Smith, D. A. 2018. Predicting News Coverage of Scientific Articles. In *ICWSM*.

Malinen, S. 2015. Understanding user participation in online communities: A systematic literature review of empirical studies. *Computers in human behavior* 46:228–238.

Millen, D. R., and Patterson, J. F. 2003. Identity disclosure and the creation of social capital. In *Proc. of the CHI EA*, 720–721. ACM.

Movshovitz-Attias, D.; Movshovitz-Attias, Y.; Steenkiste, P.; and Faloutsos, C. 2013. Analysis of the reputation system and user contributions on a question answering website: Stackoverflow. In *Proc. of the ASONAM*, 886–893. ACM.

Muchnik, L.; Aral, S.; and Taylor, S. J. 2013. Social influence bias: A randomized experiment. *Science* 341(6146):647–651.

Munzel, A. 2016. Assisting consumers in detecting fake reviews: The Role of Identity Information Disclosure and Consensus. *Journal of Retailing and Consumer Services* 32:96–108.

Nederhof, A. J. 1985. Methods of coping with social desirability bias: A review. *European journal of social psychology* 15(3):263–280.

Nov, O.; Arazy, O.; López, C.; and Brusilovsky, P. 2013. Exploring personality-targeted ui design in online social participation systems. In *Proc. of the CHI*, 361–370. ACM.

Papadimitriou, C. H.; Raghavan, P.; Tamaki, H.; and Vempala, S. 2000. Latent semantic indexing: A probabilistic analysis. *Journal of Computer and System Sciences* 61(2):217–235.

Park, K.; Weber, I.; Cha, M.; and Lee, C. 2016. Persistent sharing of fitness app status on twitter. In *Proc. of the CSCW*, 184–194. ACM.

Park, K.; Cha, M.; Kwak, H.; and Chen, K.-T. 2017. Achievement and Friends: Key Factors of Player Retention Vary Across Player Levels in Online Multiplayer Games. In *Proc. of the WWW Companion*, 445–453.

Paul, S. A.; Hong, L.; and Chi, E. H. 2012. Who is authoritative? understanding reputation mechanisms in quora. *arXiv preprint arXiv:1204.3724*.

Rains, S. A. 2007. The impact of anonymity on perceptions of source credibility and influence in computer-mediated group communication: A test of two competing hypotheses. *Communication Research* 34(1):100–125.

Reis, J.; Benevenuto, F.; de Melo, P. V.; Prates, R.; Kwak, H.; and An, J. 2015. Breaking the news: First impressions matter on online news. In *ICWSM*.

Rosenbaum, P. R., and Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55.

Sauder, M.; Lynn, F.; and Podolny, J. M. 2012. Status: Insights from organizational sociology. *Annual Review of Sociology* 38:267–283.

Sexton, J. B., and Helmreich, R. L. 2000. Analyzing cockpit communications: the links between language, performance, error, and workload. *Human Performance in Extreme Environments* 5(1):63–68.

Shah, C., and Pomerantz, J. 2010. Evaluating and predicting answer quality in community qa. In *Proc. of the SIGIR*, 411–418. ACM.

Shores, K. B.; He, Y.; Swanenburg, K. L.; Kraut, R.; and Riedl, J. 2014. The identification of deviance and its impact on retention in a multiplayer game. In *Proc. of the CSCW*, 1356–1365. ACM.

Singer, P.; Ferrara, E.; Kooti, F.; Strohmaier, M.; and Lerman, K. 2016. Evidence of online performance deterioration in user sessions on Reddit. *PloS one* 11(8):e0161636.

Stoddard, G. 2015. Popularity Dynamics and Intrinsic Quality in Reddit and Hacker News. In *ICWSM*.

Suler, J. 2004. The online disinhibition effect. *Cyberpsychology & behavior* 7(3):321–326.

Tatar, A.; de Amorim, M. D.; Fdida, S.; and Antoniadis, P. 2014. A survey on predicting the popularity of web content. *Journal of Internet Services and Applications* 5(1):8.

Tausczik, Y. R., and Pennebaker, J. W. 2011. Predicting the perceived quality of online mathematics contributions from users' reputations. In *Proc. of the CHI*, 1885–1888. ACM.

Tomkins, A.; Zhang, M.; and Heavlin, W. D. 2017. Reviewer bias in single-versus double-blind peer review. *PNAS* 114(48):12708–12713.

Wang, G.; Konolige, T.; Wilson, C.; Wang, X.; Zheng, H.; and Zhao, B. Y. 2013. You Are How You Click: Clickstream Analysis for Sybil Detection. In *USENIX Security Symposium*, volume 9, 1–008.