Identity and Opinion: A Randomized Experiment

Sean J. Taylor
New York University – Stern School of Business, staylor@stern.nyu.edu

Lev Muchnik
Hebrew University of Jerusalem, lev@muchnik.net

Sinan Aral
Massachusetts Institute of Technology, sinan@mit.edu

Identity and content are inextricably linked in social media. Content items are almost always displayed alongside the identity of the user who shares them. This social context enables social advertising but thwarts marketers' efforts to separate causal engagement effects of content from the identity of the user who shares it. To identify the role of identity in driving engagement, we conducted a large-scale randomized experiment on a social news website. For any comment on the site, 5% of random viewers could not see the commenter's identity, allowing us to measure how users interact with anonymous content. We conducted the experiment over two years, facilitating within-commenter measurements that characterize heterogeneity in identity effects. Our results establish three conclusions. First, identity cues affect rating behavior and discourse, and these effects depend on the specific identity of the content producer. Identity cues improve some users' content ratings and responses, while reducing ratings and replies for others. Second, both selective turnout and opinion change drive the results, meaning identity cues actually change people's opinions. Third, we find an association between users' past scores and identity effects, implying that users bias ratings toward past identity-rating associations. The work improves our understanding of the persuasive impact of identity and helps marketers create more effective social advertising.

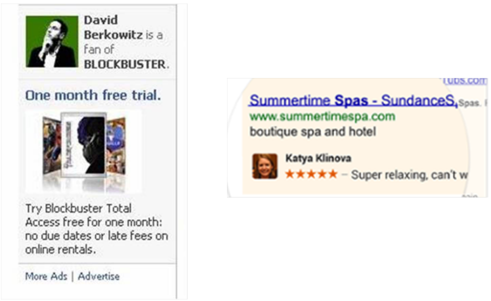
Key words: Social Networks, Field Experiments, Heterogeneity

> "Reputation is an idle and most false imposition; oft got without merit, and lost without deserving."
>
> [William Shakespeare, *Othello*]

1. Introduction

Social media usage has become pervasive over the last decade. The most recent surveys report that 72% of online adults use social networking sites, 18% use Twitter, and 6% use the online news discussion forum Reddit (Brenner and Smith 2013, Duggan and Smith 2013). These sites are important platforms for the origination, dissemination, and discussion of news for at least three reasons. First, they inspire broad participation because they are free to use and require only Internet access for participation. Second, they blur the line between producers and consumers of media content – almost all users have an audience when posting social media content. Finally, these platforms are able to match consumers to content which interests them. Social and algorithmic recommendations (such as Twitter's "retweet" feature and Reddit's ranking algorithm) increase the reach of content and ensure that it is consumed by people who find it compelling or important.

One feature that is common to all social media sites is that content is explicitly attributed to the user who produced or shared it. Social networking platforms like Facebook and microblogging sites like Twitter rely on identity as a cue to aid link formation decisions with friends and educated choices about who to follow for the best content curation. But the link between identity and content is even more pervasive. For instance, on the news discussion site Reddit, all posts include a hyperlinked username of the poster beneath them and all comments on posts include a hyperlinked username of the commenter above them. Readers of news on Reddit consume content which is always closely tied to a social identity.



David Berkowitz is a fan of BLOCKBUSTER.

One month free trial.

Try Blockbuster Total Access free for one month: no due dates or late fees on online rentals.

More Ads | Advertise

tubs.com

Summertime Spas - SundanceS, Spas.

www.summertimespa.com

boutique spa and hotel

Katya Klinova

★★★★★ – Super relaxing, can't w

Figure 1 Examples of ads augmented with social identity cues.

1.1. Social Advertising

The close link between identity and content in social media not only affects users' link formation decisions, but has also recently enabled a new form of advertising that relies on the power of consumer relationships to improve the targeting and effectiveness of advertisements. "Social advertising" is rapidly becoming one of the most novel and potentially effective means of reaching and persuading consumers about products and services. Social networking platforms like Facebook use social cues in ads to associate specific brands or products with a user's friends' decisions to purchase or like those brands or products. Search engines like Google and Bing use social identity to persuade consumers by displaying the endorsements of friends, or 'friendorsements,' in the form of friends' ratings or reviews, next to products returned in search results (see Figure 1). These new forms of advertising use social cues to nudge consumers to hold a more favorable view of the products their friends have purchased or endorsed.

The nearly universal availability of identity cues in social media and their explicit use in social advertising raises an interesting social scientific question: To what extent are content and product interactions, such as ratings, replies, purchases and reviews, caused by features of the content itself (the user generated content or advertising copy) as opposed to the identity of the associated user? The attributed identity can provide a meaningful cue that helps consumers determine which content is worth their attention and decide whether it is credible, interesting, or valuable. For discourse behavior such as comments, the identity of the producer may be just as influential as the content itself because the value from a discussion should be affected by prior relationships. On the other hand, rating behavior ideally would not change due to identity if the rating system is intended to reward high quality content in an unbiased way.

The use of social cues has been shown to increase click-through-rates and conversion rates in a variety of domains including product adoption (Aral and Walker 2011, 2012), charitable giving (Tucker 2013), information sharing (Bakshy et al. 2012b), and the adoption of premium services (Bapna and Umyarov 2011). But, the most recent research focuses almost exclusively on the average treatment effects of social cues, investigating whether the presence of social cues, like peer endorsements, increases engagement (clicks, purchases etc) on average. We extend this literature in three ways.

First, we examine the heterogeneity of social cue effects across different users. This brings the notion of identity back into the conceptualization of social cues. The presence or absence of social information is really just the beginning of the story. If marketers are to take full advantage of personalized social advertising, they will need to understand how the identity of a specific user affects the persuasive power of a social cue. For example, social cues in advertisements for clothing may on average increase engagement with social ads and sales. However, the identity cues from specific people may be much more persuasive than from others. In fact, identity cues from some unfashionable consumers may even have negative effects on engagement and clothing sales. Furthermore, specific consumers may be persuaded by identity cues from specific peers. For instance, one consumer may be persuaded by the presence of a social cue from a peer with similar tastes while another may not. This type of identity effect implies fully dyadic identity cue dependencies. Understanding this type of individual heterogeneity in identity effects is the key to personalizing social advertising.

Second, we examine the behavioral mechanisms underlying identity effects. In our setting, users give their opinions of the content they engage with by voting the content up or down. By estimating models of both turnout and the likelihood of up- and down- voting, we are able to robustly distinguish selective turnout (changes in turnout by typically negative or positive voters) from opinion change (changes in the likelihood of up and down voting). This allows us to identify the effects of identity on opinions themselves, rather than simply whether social cues attract the attention of or increase engagement from different sub-populations of consumers without moving their priors about the content they are viewing.

Third, we examine reputation effects built time. By examining the content ratings of a user's prior contributions we examine whether good prior ratings improve identity effects – in other words whether knowing the identity of a user whose content has been highly rated in the past improves the positive impact of an identity cue. We find strong evidence that it does. This is consistent with the interpretation that users learn the reputation of other users over time and bias their ratings toward past identity-rating associations. As far as we are aware, this is the first empirical evidence identifying the causal impact of accumulating reputations on others' opinions, holding constant the quality of the content produced by a given identity.

It is easy to casually theorize about the role of identity in these social systems, but much harder to empirically separate the effect of identity from the effects of content characteristics such as quality or timeliness. Latent characteristics of content producers can plausibly affect both the strength and direction of the identity cue effect as well as the effect of the content itself. For instance, if a user is a high quality commenter then she will presumably have also earned a good reputation associated with her identity that could change perceptions of her content. To disentangle the endogenous effects of quality and identity on content interactions, we conducted a large scale field experiment with a novel anonymization manipulation. Our design cleanly identifies the causal effect of identity cues presented alongside content by examining the counterfactual where users *cannot* be influenced by explicit identity cues.

1.2. From Big Data to Large Scale Experiments

The phrase “Big Data” suggests the main power in modern empirical social science and marketing research is in the size and scale of the observational data we now collect. Our view however is that the real power lies in a newfound ability to design large scale randomized experiments in complex social settings. In the 1960s, the esteemed American social psychologist Stanley Milgram pioneered the social sciences of the twentieth century by conducting several seminal experiments that documented our obedience to authority and the regularity of ‘Six Degrees of Separation’ in human social networks. It took Milgram years to design, execute and analyze these studies, which ultimately collected data on a small handful of people. Today, we can design, conduct and analyze such experiments much more rapidly; and not on samples of hundreds of people, but rather on hundreds of millions of people at a time.

Smaller scale randomized experiments are typically only sufficiently powered to estimate average treatment effects – the average effect of a policy in a population. But large scale experiments, conducted on hundreds of thousands or millions of people, allow researchers to unpack the heterogeneity of treatment effects across different sub-populations in society and to explore the behavioral mechanisms that underlie and explain the treatment effects. For example, prior work has estimated heterogeneity in the impact of influence mediating messages on different types of people among millions of Facebook users (Aral and Walker 2012) and whether opinion change or selective turnout creates social influence bias in online ratings that were seen tens of millions of times (Muchnik et al. 2013). These insights enable policies tailored for particular individuals. By understanding the causal behavioral mechanisms underlying the outcomes of specific policies, how and why those outcomes vary across different people and how they change over time, we can develop more contextual, personalized, adaptive (and therefore more effective) policies and business strategies. Large scale experimentation is therefore a critical part of the role of big data in marketing science and practice.

2. Related Work

2.1. Causal Effects of Online Social Content

As web and mobile services have added more social features, there has been a rise in the frequency and breadth of online social interactions exchanged between users. Researchers have gravitated toward these platforms in order to understand how these digitally-mediated interactions over articulated and implicit social networks influence our decision-making (Aral and Walker 2011) and create greater interdependence in our preferences (Wang et al. 2013). Social content, particularly the articulation of preferences, has also formed the basis for new form of advertising where traditional ads are augmented with or targeted using previously generated social signals from peers (Tucker 2012).

Just as with earlier work on peer effects in economics, causal inference for effects of online social content is difficult due to endogeneity in which users see signals from whom (Shalizi and Thomas 2011) as well as which users decide to generate social content (Toubia and Stephen 2013, Taylor et al. 2013). Simply being eligible to see social content from a peer selects a user into group which is more likely to interact with that content – meaning that observational studies are likely to confound influence from social cues with a process that effectively targets the most interested people.

As a result, researchers have gravitated toward using randomized field experiments in order to identify online peer effects. For instance, Aral and Walker (2011) and Aral and Walker (2012) randomly assign users of a Facebook application to conditions which vary whether their usage generates active or passive notifications to peers, allowing inferences about the relative influence of different types of viral messages. Muchnik et al. (2013) take a different approach, randomly assigning content items to conditions which vary the valence of prior social ratings.

The frontier beyond establishing that online social content affects peer behavior is measuring when and under what conditions social content is more or less influential on attitudes and behaviors. Two recent works, Bakshy et al. (2012a) and (Aral and Walker 2013) establish that features of *relationships* between the sender and recipient of social content mediate peer effects.

Our current study seeks to extend our understanding of the causal effects of social content by focusing on the role of content producers' identity in changing attitudes toward and interactions with their content. Identity cues, such as usernames and user photos, carry meaning that the viewer may associate with credibility, relevance, and trust. Specifically, we use a longer-term experimental design with repeated measures of the effects of the same identities combined with a randomization that separates the effect of identity from the quality and characteristics of content. We are therefore able to contribute to this line of work by modeling and identifying heterogeneity in the effects of identity cues and proposing a mechanism through which these effects change over time.

2.2. Economic Value of Online Reputation

In conjunction with reputation systems which afford feedback, online identities can also affect economic outcomes such as the probability of economic transactions or the price at which that transaction occurs. For instance on the online auction website eBay, having a seller account with a higher quantity of buyer feedback is associated with higher transaction prices for identical items (Resnick et al. 2006) and in online freelance labor markets, positive feedback associated with a worker identity is associated with higher probability of contracts and higher wages (Yoganarasimhan 2012).

However, in the context of an economic transaction, reputation-transmitting features of identity information is likely to be resolving information asymmetries rather than influencing a buyer's perception of quality. In the realm of social media where users can consume a product – typically a content item such as a comment, photo, or video – costlessly, there is no asymmetry about quality, reliability, or fit to resolve. Therefore identity has more of a relational component in these contexts and is developed over time, potentially affecting the opinion or attention toward content: whether users will read, believe, like, or agree with it. Forman et al. (2008), as an example, find that when online reviewers disclose more identifying information that their reviews are rated as more helpful and more likely to lead to sales.

Our study sheds some light on how online reputation is valuable in shaping the opinions and attention of content consumers, helping to answer whether the inclusion of identity cues can cause better ratings and greater probability of replies – interactions which we believe elicit content preferences.

2.3. Positive Feedback Loops in Social Systems

There is recent experimental evidence that songs and software with a higher number of downloads receive more downloads in the future (Salganik and Watts 2008, Hanson and

Putler 1996), and comments with higher ratings are more likely to receive future positive ratings (Muchnik et al. 2013). These findings of herd behavior and virtuous cycles may just as easily hold for *people* and their identities as well as the content items they produce, as first articulated by Robert Merton (Merton 1968). Positive feedback loops which produce rich-get-richer dynamics in social and economic systems are important, yet defy identification because unobserved content quality can cause both current success and future success of everything from

We are not only able to demonstrate the results of such a process – inequality or bias in how users' content is rated – but we can also identify the dynamics. The long-term nature of our experiment provides a natural setting for testing how identity effects accumulate, allowing us to contribute to the study of the dynamics of inequality formation in online ratings systems.

2.4. Measuring Content Quality and Sharer Reputation

With the dramatic increase in online content items and in users creating content, exploiting log data to predict the quality of content and users is now an active area of research. There are valuable signals present in content features, explicit quality ratings, content interactions by users which reveal user preferences, and links between users and content. Exploiting these signals using machine learning techniques can help predict future content interactions, or the editorially determined quality of user-generated content (Agichtein et al. 2008).

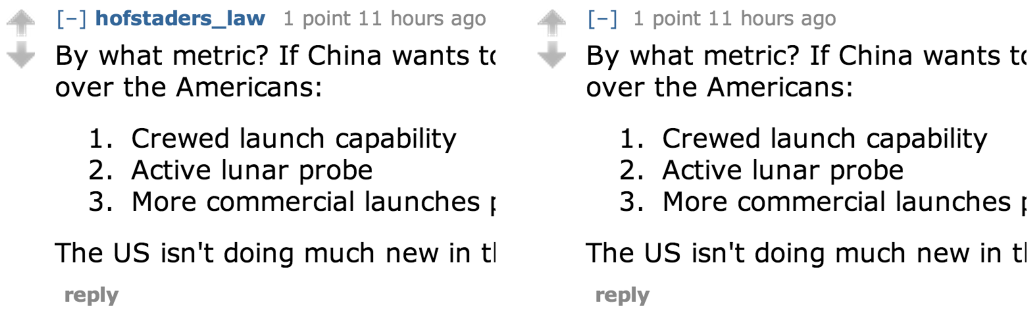
Yang et al. (2013) propose to mitigate biases in content ratings due to the presence of user identities by combining data sampled in a relatively unbiased way, measuring interactions with content exposures from an algorithmic recommendation system (LinkedIn's "Today" module) which are not displayed with a social identity. They show that including interactions from this less biased data can improve inference about content producer quality (and inversely, their identity effects). But while their model evaluation suggests that selection and response biases affect inferences about user reputation, they do not provide any rigorous measurements of biases nor interpret response bias as persuasive in nature – the subject of our work.

We show that identity bias in content ratings varies significantly (in sign and magnitude) across users and that it can be attributed to a mix of social processes. Due to the fact that we elicit both positive and negative feedback on content, we are able to identify that attentional explanations are not sufficient to explain rating outcomes, and that users' *opinions* about content are affected by associated identity information.

2.5. Blinded Academic Peer Review

The goals of peer review in the academic world are broadly similar to those of online content rating systems: providing incentives for better research and a quality filter to readers. Given the economic and scientific significance of peer review, researchers have explored the impact of blinding manuscripts for reviewers. The evidence for bias reviews due to the presence of identifying information appears to be mixed. Ross et al. (2006) find significant improvements from identity in *abstract* acceptance rates for researchers from prestigious universities and those located in the United States, indicating a potential bias to be mitigated by blinding. However, an earlier study on review of full length articles (Fisher et al. 1994) showed no effect of blinding on reviewer recommendations or editor decisions, while calling into question whether blinding actually achieves the intended purpose due to the presence of alternative cues.

While there are clear similarities between peer review in academia and online content rating systems, it is difficult to theorize about how results in one domain might generalize



[–] hofstaders_law 1 point 11 hours ago

By what metric? If China wants t
over the Americans:

1. Crewed launch capability
2. Active lunar probe
3. More commercial launches

The US isn't doing much new in t

reply

(a)

[–] 1 point 11 hours ago

By what metric? If China wants t
over the Americans:

1. Crewed launch capability
2. Active lunar probe
3. More commercial launches

The US isn't doing much new in t

reply

(b)

Figure 2 **The experiment was conducted on a social news discussion website similar to Reddit.com (displayed here). User comments are displayed either with the commenter's identity (2(a), identified condition) or, in 5% of viewer-comment exposures, without the commenter's identity (2(b), anonymous condition). Viewers of comments were able to up-vote or down-vote the comment (up and down arrows) to increment or decrement the score or reply to the comment by clicking "reply" and creating their own comments.**

to the other. One could view our study as lower cognitive effort rating task (much shorter, less complex content items) where people are more likely to rely on source cues to determine quality instead of costly evaluation of merit, and therefore a more sensitive test for identity-based rating biases.

3. Experiment

3.1. Experimental Context

We conducted our experiment on a social news aggregation website similar to Digg.com and Reddit.com. The website provides a web-based architecture for a community of people to submit synopses of news and entertainment-related posts published elsewhere on the web with links pointing to the outside content. User comments on these posts develop into long discussion threads containing dozens and even hundreds of comments. User-generated comments are also rated by community members who up- or down-vote them, providing an incentive mechanism to improve the quality of comments.

3.2. Experimental Design

For about 100 weeks in 2011-2013, for each comment created on the site we randomly assigned 5% of users exposed[1] to it to the *anonymous* treatment. A user assigned to this treatment for a comment was unable to see the username of the commenter whenever she saw the comment on the website. The anonymous condition was persistent; no matter how many times the viewer saw the comment, she was unable to see the commenter's username. The remaining 95% of comment-user exposures were assigned to the *identified* treatment – status quo for the website, listing the commenter's username with a clickable link directly above their comment.

Overall we observe 346,965 comments over the course of the experiment, yielding 17.6m user-comment exposures by 6,874 distinct users. Of these exposures, 16.7m were in the identified source condition, while about 870,000 were in the anonymous source condition. Figure 2 shows an example comment under identified and anonymous treatments, as well as a representative interface for rating and replying to the comment.[2]

Due to its obvious and slightly disruptive nature, the website administration announced the experiment to all users prior to starting the manipulation. Users of the site were informed that this change was made to the site in order to study user commenting and rating behavior. The experiment was run over a long period of time, giving us a large longitudinal sample of observations which we tested for novelty and habituation effects.

The experiment was designed to allow us to separate the effect of identity from the effect of content quality. Due to our randomization, the presence of identity is independent of

[1] We count a user as exposed if they loaded a page with the comment at least once. Subsequent views of the same comment by a user are treated as the same "exposure" – they are assigned to the same treatment group and any resulting interactions were logged as responses.

[2] Due to a confidentiality agreement with the website, we are unable to post screenshots of the exact interface used during the experiment.

content quality. This allows us to estimate the effect of identity by contrasting how viewers interact with content which has the same quality in expectation but differs only in whether an identity cue is present or not.

4. Models and Estimation

We are interested in the effect of identity on rating behavior – a trichotomous outcome that can be either an up-vote, a down-vote, or no vote after a user is exposed to a comment. We characterize rating using two sets of outcome variables. Let $i \in [1, I]$, $j \in [1, J]$, $k \in [1, K]$ index commenters, viewers, and comments respectively. We define A_{ijk} (chosen because it resembles an up arrow) denote that up-vote action: it is 1 if user j up-votes comment k by commenter i and 0 if she does not up-vote it. We also define V_{ijk} (resembling a down arrow) as 1 if user j down-votes comment k by commenter i and 0 if she does not down-vote it. The indicator variable R_{ijk} represents that viewer j replied to comment k. For convenience, we define $T_{ijk} = A_{ijk} + V_{ijk}$ (*turnout*) to be an indicator variable representing comment k was rated by user j.

Let D_{jk} be a binary indicator variable which is 1 if viewer j is shown an explicit identity cue for comment k and 0 otherwise. We would like to model three conditional probability distributions. First, $\Pr(A_{ijk} = 1 | D_{jk})$ is the probability that the viewer up-votes the comment under either the identified or anonymous conditions. Second, $\Pr(V_{ijk} = 1 | D_{jk})$ is the probability that she chooses to down-vote the comment. Third, $\Pr(R_{ijk} = 1 | D_{jk})$ is the probability that she chooses to reply to the comment.

In the next two Subsections, we describe two types of probability models we estimate for our experiment. The first type of model assumes that commenters are heterogeneous with respect to both content quality and the effect of their identity on these three outcomes. It is extremely flexible in that it contains two parameters for every commenter in our sample.

Parameter	Distribution	Description
m_i	Normal$(\bar{q}, \sigma_q^2)$	commenter i's mean content effect on up-vote rate
n_i	Normal$(\bar{r}, \sigma_r^2)$	commenter i's identity effect on up-vote rate
q_i	Normal$(\bar{q}, \sigma_q^2)$	commenter i's mean content effect on down-vote rate
r_i	Normal$(\bar{r}, \sigma_r^2)$	commenter i's identity effect on down-vote rate
s_i	Normal$(\bar{s}, \sigma_s^2)$	commenter i's mean content effect on reply rate
t_i	Normal$(\bar{t}, \sigma_t^2)$	commenter i's identity effect on reply rate
u_j	Normal$(0, \sigma_u^2)$	viewer j's propensity to up-vote
v_j	Normal$(0, \sigma_v^2)$	viewer j's propensity to down-vote
w_j	Normal$(0, \sigma_w^2)$	viewer j's propensity to reply
c_k	Normal$(0, \sigma_c^2)$	comment k's up-vote propensity
d_k	Normal$(0, \sigma_d^2)$	comment k's down-vote propensity
e_k	Normal$(0, \sigma_d^2)$	comment k's down-vote propensity

Table 1 Reference table for parameter names

The second type of model assumes that rating and discourse behavior are affected by cumulative scores of previous comments, and estimates probabilities using smooth functions of the mean cumulative score of previous comments by the commenter.

4.1. Commenter-level Effects Model

A standard approach with this many levels of heterogeneity is to assume there exist latent parameters which can rationalize the decision-making of the individuals in the study (Rossi and Allenby 2003).

There are three natural levels for parameters in our model: commenters, their comments, and the viewers who see and possibly rate the comments. We describe commenters with six parameters each: $(m_i, n_i, q_i, r_i, s_i, t_i)$, corresponding to propensity for up-vote, down-vote, and reply interactions for their comments under identity and anonymity.

We assume that commenter i writes *content* with mean up-vote propensity m_i, mean down-vote propensity q_i, and mean reply propensity s_i. Each of her comments' up-vote, down-vote, and reply parameters, c_k, d_k, and e_k respectively, are then drawn from a normal distribution with common variance parameters across users:

$$\begin{aligned} m_i &\sim \text{Normal}(\bar{m}, \sigma_s^2); & c_k &\sim \text{Normal}(0, \sigma_c^2); \\ q_i &\sim \text{Normal}(\bar{q}, \sigma_q^2); & d_k &\sim \text{Normal}(0, \sigma_d^2); \\ s_i &\sim \text{Normal}(\bar{s}, \sigma_s^2); & e_k &\sim \text{Normal}(0, \sigma_e^2). \end{aligned}$$

In addition to comment-level parameters characterizing the comment's quality, comments can have an associated identity – the effect of the commenter's username listed alongside the content. Unlike comment-level parameters, commenter-level parameters model persistent characteristics of the commenter which are the same across all of her content production. We let n_i, r_i, and t_i represent the effect of i's identity on up-vote, down-vote, and reply probabilities respectively:

$$n_i \sim \text{Normal}(\bar{n}, \sigma_n^2); \qquad r_i \sim \text{Normal}(\bar{r}, \sigma_r^2); \qquad t_i \sim \text{Normal}(\bar{t}, \sigma_t^2).$$

Viewers can vary with respect to how frequently they rate comments and their average positivity, which we represent by parameters u_j and v_j which have Normal prior distributions with means of zero:

$$u_j \sim \text{Normal}(0, \sigma_u^2); \qquad v_j \sim \text{Normal}(0, \sigma_v^2); \qquad w_j \sim \text{Normal}(0, \sigma_w^2).$$

We now linearly aggregate comment-, commenter-, and viewer-level parameters into indexes which are proportional to the probabilities of up-vote, down-vote, and reply. We use a logistic function, $F(x) = \frac{1}{1+e^{-x}}$, as a link function:

$$\Pr(A_{ijk}=1|D_{jk})=F(m_i+D_{jk}n_i+c_k+u_j+\epsilon_{ijk}); \tag{1}$$

$$\Pr(V_{ijk}=1|D_{jk})=F(q_i+D_{jk}r_i+d_k+v_j+\nu_{ijk}); \tag{2}$$

$$\Pr(R_{ijk}=1|D_{jk})=F(s_i+D_{jk}t_i+e_k+w_j+\upsilon_{ijk}), \tag{3}$$

where ϵ_{ijk}, ν_{ijk}, and υ_{ijk} represent unobserved factors that affect rating and reply behavior (e.g. interactions between content and viewers such as interest or fit). Table 1 summarizes the parameters in the model and the distributions we assume for them. We use non-informative prior distributions for all hyperparameters – the means and standard deviations of the parameter distributions.

4.2. Accumulated Cue Model

Instead of introducing a parameter for each individual commenter in our sample (a full heterogeneous treatment effects specification), we could hypothesize that there is some observed variable or variables which explain the variances in the relative risks of identity. Motivated by prior work showing accumulating returns to early success in social systems (e.g. Salganik and Watts (2008)), we hypothesize that prior ratings of a commenter's content affect the value of her identity. Specifically, if her username has been associated with high (low) quality content in the past, then viewers of her content will be biased toward rating her future content higher (lower).

We are proposing a model where at least some of the relative risks of identity on rating behavior are mediated by a commenter's past ratings. We operationalize past ratings as a scalar value $X_{ik}[t]$ – mean of the cumulative score of a commenter's i's t most recent comments before comment k. On our experimental website, cumulative scores are equal to the number of up-votes minus the number of down-votes, a common rating aggregation procedure.

Let (g_A, g_V, g_R) and (h_A, h_V, h_R) be smooth functions[3] of this mean cumulative score over commenter's the past t comments. Our mean cumulative score cue model is then specified as:

$$\Pr(A_{ijk} = 1 | D_{jk}) = F(g_A(X_{ik}[t]) + D_{jk}\, h_A(X_{ik}[t]) + c_k + u_j + \epsilon_{ijk}); \tag{4}$$

$$\Pr(V_{ijk} = 1 | D_{jk}) = F(g_V(X_{ik}[t]) + D_{jk}\, h_V(X_{ik}[t]) + d_k + v_j + \nu_{ijk}) \tag{5}$$

$$\Pr(R_{ijk} = 1 | D_{jk}) = F(g_R(X_{ik}[t]) + D_{jk}\, h_R(X_{ik}[t]) + e_k + w_j + \upsilon_{ijk}) \tag{6}$$

The functions (g_A, g_V, g_R) map the heterogeneity in quality accounted for by the mean cumulative rating of i into the increased probabilities of up-vote, down-vote, and reply which are are not caused by the identity cue. We retain random-effects parameters accounting for heterogeneity in comment quality and comment viewer rating behavior.

We are primarily interested in functions (h_A, h_V, h_R), which map the mean cumulative rating for i into the impact of her identity on probabilities of up-vote, down-vote, and reply respectively. If these functions are significantly different from 0 for some range of $X_{ik}[t]$, then we can conclude that for users with certain values of mean cumulative rating that their identity effect is associated with the past rating their comments have received.

In this model, t is a free parameter that we have no prior knowledge about. Smaller values of t may be interpreted as viewers having shorter memories about a commenter's past associated ratings, while larger values would imply a longer a memory. In the Section 5, we provide plots of relative risks of identity computed for $t \in \{1, 10, 100\}$ in order to explore how the choice of this parameter impacts the results.

[3] We use natural splines with two degrees of freedom in our estimation. The results were qualitatively the same with more complex spline fits.

4.3. Relative Risks of Identity

In the previous two sections we have presented two types of models for how identity affects rating and discourse behavior. We now discuss interpretation of these models.

In the commenter-level heterogeneous treatment effect models, the parameters of interest are the commenter-level paramters: t_i and r_i which describe the average effect of commenter i's identity on turnout and positivity respectively. However, these measurements are made on an arbitrary scale so we prefer to use our model to estimate relative risks of identity for each individual commenter:

$$\begin{aligned}
\text{RR of Up-vote}(i) &= \frac{P(A_{ijk}|q_i, r_i, D_{jk}=1)}{P(A_{ijk}|q_i, r_i, D_{jk}=0)} \\
\text{RR of Down-vote}(i) &= \frac{P(V_{ijk}|s_i, t_i, D_{jk}=1)}{P(V_{ijk}|s_i, t_i, D_{jk}=0)} \\
\text{RR of Reply}(i) &= \frac{P(R_{ijk}|q_i, r_i, D_{jk}=1)}{P(R_{ijk}|q_i, r_i, D_{jk}=0)}
\end{aligned}$$

In the accumulated cue models, we first estimate identity (g_A, g_V, g_R) and quality (h_A, h_V, h_R) functions and then use these functions to predict relative risks:

$$\begin{aligned}
\text{RR of Up-vote}(X_{ik}[t]) &= \frac{P(A_{ijk}|g_A, h_A, X_{ik}[t], D_{jk}=1)}{P(A_{ijk}|g_A, h_A, X_{ik}[t], D_{jk}=0)} \\
\text{RR of Down-vote}(X_{ik}[t]) &= \frac{P(V_{ijk}|g_V, h_V, X_{ik}[t], D_{jk}=1)}{P(V_{ijk}|g_V, h_V, X_{ik}[t], D_{jk}=0)} \\
\text{RR of Reply}(X_{ik}[t]) &= \frac{P(R_{ijk}|g_R, h_R, X_{ik}[t], D_{jk}=1)}{P(R_{ijk}|g_R, h_R, X_{ik}[t], D_{jk}=0)}
\end{aligned}$$

4.4. Parameter Inference

The scale of our data combined with the large number of target parameters present a challenge for inference. Though MCMC estimation of Bayesian models is possible at fairly large scale (e.g. over a million observations in (Taylor et al. 2013) using a state-of-the-art Hamiltonian Monte Carlo procedure implemented in highly efficient C++ code), we found it unfeasible to simulate full posteriors over all parameters for the full experiment in a reasonable amount of time.

The inherently serial nature of MCMC algorithms make it difficult to increase their speed even with the availability of a large number of machines, motivating us to use a different approach. Since we assumed that comment- and viewer-level parameters $(c_k, d_k, e_k$ and (u_j, v_j, w_j) have mean zero, our estimates of commenter-level up-vote, down-vote, and reply parameters, as well as the natural spline functions in the score cue models, can be consistently and quickly estimated by first marginalizing the observations (either by commenter or by a fine-grained score cue bin) and then using maximum likelihood or Bayesian methods.

However, marginalizing over commenters or mean cumulative score ignores dependencies due to repeated re-sampling of comments and viewers, which can vary in their quality and rating behavior. Thus the if we estimate probabilities over margnalized data, we will tend to underestimate our uncertainty about our parameter estimates. This problem is common in field experiments where users and content-items can be sampled multiple times.

In order to produce credible inferences about our uncertainty for identity effects, we use utilize a recently developed multiway bootstrap procedure (Owen and Eckles 2012), randomly resampling comments and viewers jointly to produce a sequence of model parameter estimates which can be aggregated to produce conservative confidence intervals. The procedure is fast, parallelizable, and was previously used in Bakshy et al. (2012a) to efficiently analyze an experiment with a similar dependency structure between observations.

The following is a brief explanation of how we combine multiway bootstrap with fast marginal model estimation procedures to produce unbiased probability and relative risk estimates with conservative confidence intervals:

1. Generate a series of 500 bootstrap replicates of our experiment. Each comment and viewer individually has a 50% probability of being included in each replicate.
2. For each replicate, estimate a marginal model for probabilities of up-vote, down-vote, and reply behaviors. For the hierarchical models with commenter-level parameters, we use an empirical Bayes estimator (Morris 1983) which first estimates hyperparameters for the commenter-level distributions using maximum likelihood and then estimates the commenter-level parameters conditional on those estimates. For the score cue models we use a maximum likelihood procedure.
3. For each model-replicate, predict probabilities of up-vote, down-vote, and reply for the domain of the model (treatment status interacted with commenters or mean cumulative scores). Use these predicted probabilities to additionally compute relative risk of identity for each value in the domain.
4. For each model, compute the mean and bootstrap 95% confidence intervals for each of the probability predictions and relative risks.

4.5. Isolating Opinion Change

An interesting question when studying rating behavior is under what conditions we can conclude that a cue changed the valence of a user's rating – i.e. *caused* an up-vote to become a down-vote or vice versa. This is a difficult question to answer unless we force

Proportion	Interpretation	$T_{ijk}(0)$	$T_{ijk}(1)$	$A_{ijk}(0)$	$A_{ijk}(1)$
g_1	always up-voter	1	1	1	1
g_2	always down-voter	1	1	0	0
g_3	never voter	0	0	–	–
g_4	negative opinion change	1	1	1	0
g_5	positive opinion change	1	1	0	1
g_6	suppressed up-voter	1	0	1	–
g_7	suppressed down-voter	1	0	0	–
g_8	up-vote from turnout	0	1	–	1
g_9	down-vote from turnout	0	1	–	0

Table 2 The nine observationally equivalent potential outcomes for turnout and positivity. Dashes denote states which are observationally equivalent given either value. These states are mutual exclusive and exhaustive and any given observation in our data corresponds can belong to exactly three of them.

raters to rate every comment because any randomly assigned treatments will tend to affect both whether the user decides to rate at all as well as her opinion about the quality of the content. To be more precise, an experiment with optional, binary rating can only identify two treatment effects: changes the probabilities of an up-vote and a down-vote. Any changes rating probabilities can be rationalized through increased or decreased probability of attending to the rating task by some user-comment sub-population or by changes in valence of ratings conditional on users deciding to rate.

In this section we comprehensively list the possible effects that the addition of identity can have on rating behaviors using a potential outcomes model (Rubin 2005) of rating with conditional observability. Any individual comment exposure must belong to one of class of observationally equivalent potential outcomes. Once the outcome of an individual trial is observed, it must belong to one of three of the nine possible states. We can then derive inferences about the proportion of the population belonging to each potential outcome class using the results of our experiment.

We summarize the different mechanisms using the potential outcomes presented in Table 2 alongside a parameter representing the (unobserved) proportion of our sample for which each potential outcome class applies.

There are four main groups of possible counterfactual states. The first group (g_1, g_2, g_3) consists of observations where the manipulation changes neither turnout (the choice to rate) nor positivity (the choice to up-vote) conditional on turnout. In the second group (g_4, g_5), the manipulation affects positivity but not turnout. We interpret these states as unambiguous opinion change because the users were swayed to change the valence of their ratings.

The third (g_6, g_7) and fourth (g_8, g_9) groups of states are the most subtle. In these groups, the manipulation affects turnout in a way that prevents us from ever observing counterfactual states of positivity. In principle we could force subjects to rate and elicit positivity, but we must proceed here (as in many other similar applications) as if these cannot be observed. This unobservability is in addition to the more typical problem of causal inference that we never simultaneously observe treatment and control states for any individual unit.

Though we would like to identify or place lower bounds on the values of g_4 and g_5, we have 9 parameters to identify, one linear constraint on those parameters, and only 4 quantities provided by our experiment, leading to the following system of equations which describe which counterfactual states correspond to which observations in our data:

$$1 = \sum_{n=1}^{9} g_n$$

$$P(A_{ijk}=1|D_{jk}=0)=g_1+g_4+g_6$$

$$P(V_{ijk}=1|D_{jk}=0)=g_2+g_5+g_7$$

$$P(A_{ijk}=1|D_{jk}=1)=g_1+g_5+g_8$$

$$P(V_{ijk}=1|D_{jk}=1)=g_2+g_4+g_9$$

It is clear from this system, which features four known quantities and eight degrees of freedom, that we cannot identify the opinion change parameters of interest without additional assumptions.

Computing average treatment effects on up-vote and down-vote probabilities yields two equations which have share a common term, (g_5-g_4), the net proportion of opinion change in our sample.[4]

$$P(A_{ijk}=1|D_{jk}=1)-P(A_{ijk}=1|D_{jk}=0)=(g_8-g_6)+(g_5-g_4)$$
$$P(V_{ijk}=1|D_{jk}=1)-P(V_{ijk}=1|D_{jk}=0)=(g_9-g_7)-(g_5-g_4)$$

This treatment effect system of equations succinctly shows what assumptions we need in order to identify net opinion change $(g_5-g_4\neq 0)$. If either treatment effect is non-zero and we are willing to assume no net selective turnout (that the identity does not increase the relative number of up-voters or down-voters) – $g_8-g_6=0$ or $g_9-g_7=0$ – then opinion change must explain the observed treatment effects.

[4] It is equivalent to use relative risks here instead of risk differences if we take logs of both sides of the equations. The ensuing analysis is the same except with log probability differences on the right-hand sides.

Unfortunately, selective turnout is difficult to rule out *a priori* because there are many natural situations where we would expect it. For instance, if a (favorable) identity cue causes viewers to rate when the comment is insightful and abstain when it is of poor quality, then we would expect $g_8 - g_6 > 0$ and $g_9 - g_7 < 0$. A positive net opinion change would then not be necessary to rationalize either increased up-vote rate or a decreased down-vote rate.

If we assume that net turnout change for up-voting and down-voting are the same, $g_8 - g_6 = g_9 - g_7$, then net opinion change can be identified as:

$$\begin{aligned}&= \frac{1}{2}[(P(A_{ijk} = 1|D_{jk} = 1) - P(A_{ijk} = 1|D_{jk} = 0))\\&\quad -(P(V_{ijk} = 1|D_{jk} = 1) - P(V_{ijk} = 1|D_{jk} = 0))],\end{aligned}$$

an assumption that would be justified if we believe the presence of identity uniformly increases (or decreases) probability of rating but not in a way that which is selective of up- or down-voting.

The most modest assumption we can make that allows us draw meaningful conclusions about net opinion change is that net turnout effect is weakly positive: $g_8 - g_6 \geq 0$ and $g_9 - g_7 \geq 0$. Given this assumption, if we can derive the following bounds on net opinion change:

$$\begin{aligned}g_5 - g_4 &\geq P(A_{ijk} = 1|D_{jk} = 0) - P(A_{ijk} = 1|D_{jk} = 1)\\g_5 - g_4 &\geq P(V_{ijk} = 1|D_{jk} = 1) - P(V_{ijk} = 1|D_{jk} = 0).\end{aligned}$$

We can interpret these two bounds as follows. First, if we see a significant *decrease* in up-vote rate from identity, then there has been significantly negative net opinion change. Second, if we see a significant *decrease* in down-vote rate from identity, then there has been a significantly positive net opinion change. The intuition is that if we assume the identity treatment can only increase rating behavior, then any decrease in either up-vote or down-vote probabilities must be due solely to net opinion change.

4.6. Interference

Although the assignment of our treatment to comment exposures is random, it is possible that the outcome for any individual comment exposure could be affected by the treatment status of adjacent comment exposures. These *spillover* effects would then violate the stable unit treatment value assumption (SUTVA) required for unbiased causal inferences.

As an example, consider a comment thread where two users reply to each other's comments. An anonymous exposure in such a discussion is unlikely to seem anonymous given the likelihood that most of the surrounding comments would be in the identified condition. Thus we expect leakage of identity information in our experiment due to interference. This leakage of identity is in addition to identity information carried in non-explicit identity information conveyed within the comment itself such as writing style and characteristic opinions or topics.

The positive outcome of this SUTVA violation is that we expect a downward bias in effect estimates. Because we are only able to mask identity in a limited way, any differences in behaviors between identified and anonymous conditions are likely to be conservative estimates of the true difference if we were able to more reliably prevent spillovers of identity information.

Quantity	Anonymous	Identified	Combined
Comment Exposures	16,412,070	852,271	17,264,341
Unique Comments	346,917	259,979	346,933
Unique Commenters	3,731	3,093	3,732
Unique Viewers	6,638	4,312	6,660
Unique Pairs	867,349	223,051	885,098
Up-vote Rate	0.051	0.053	0.053
Down-vote Rate	0.015	0.017	0.017
Turnout Rate	0.066	0.070	0.070
Positivity	0.768	0.758	0.758
Reply	0.011	0.012	0.012

Table 3 Experiment summary statistics

5. Results

We present our experimental results in the following three subsections. Subsection 5.1 discusses unconditional treatment effects of identity, which are estimates that are pooled over all commenters within our sample. Subsection 5.2 considers treatment effects estimated using a hierarchical model where the effect of identity is allowed to be heterogeneous within commenters – effectively exploiting within-commenter variation in rating behavior between identified and anonymous conditions to identify individual-level relative risks of identity. Finally, Subsection 5.3 explores one plausible mechanism by which users associate quality with specific identities by looking at relative risks of identity conditional on the scores associated with identities.

For reference, Table 3 describes the scale of the experiment – number of exposures, comments, commenters, and viewers – as well as summary statistics for rating and discourse behaviors under treatment and control conditions.

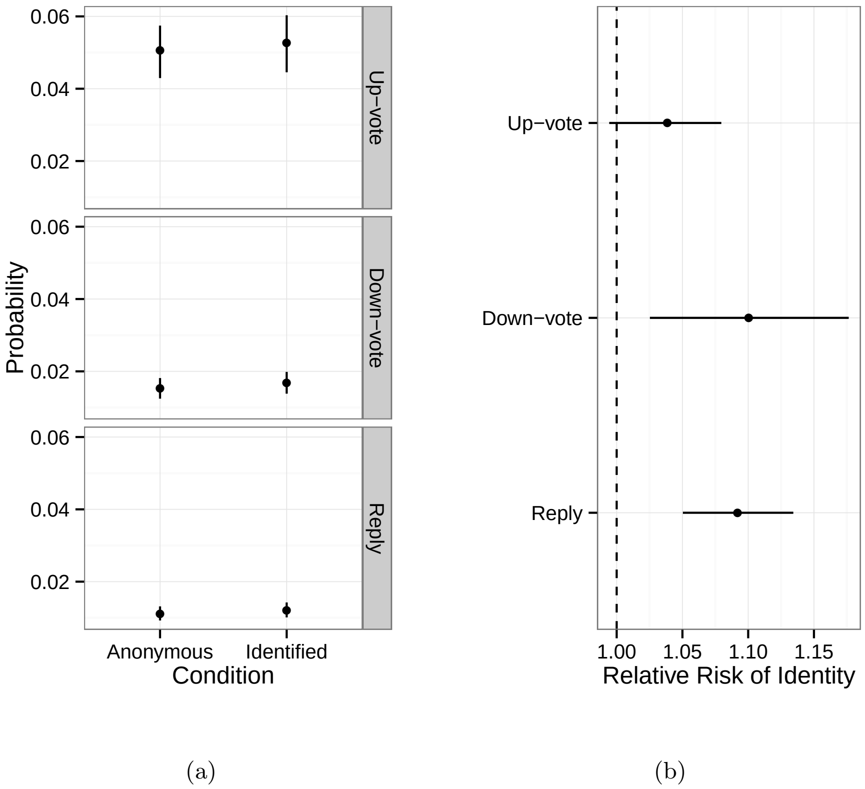


Figure 3 (a) Mean interaction outcomes for identified and anonymous conditions and (b) relative risks of identity. The 95% confidence intervals are estimated using a multiway bootstrap on comments and viewers with 500 bootstrap replicates.

5.1. Unconditional Effects

We first estimate relative risk of identity on rating and discourse behaviors pooled across all exposures in our experiment. Figure 3(a) shows rates of up-vote, down-vote and reply for identified and anonymous exposures. Rating is relative rare, with only 6-7% of exposures resulting in either and up-vote or a down-vote. Up-votes are more common, representing about three-quarters of all ratings in our data. Perhaps due to the cost of creating content, replies are even more uncommon than ratings – only about 1% of comment exposures lead to a reply.

On average, the inclusion of identity increases down-voting and reply behaviors over anonymous exposures. This suggests that at least some of a user's motivation to rate comments is driven by *whom* they are rating rather than solely by the text of the comment itself. Replies are more likely when identity cues are included, an intuitive result given that part of the utility from discourse is likely derived from knowing which specific user is involved.

These unconditional effects are averaged over many users identities. Some of these users could be receiving more positive ratings while others could be receiving more negative ratings due to their identities. In a sense, the unconditional treatment effects can obscure true countervailing effects within our experiment by averaging them to form aggregate quantities. For instance, if we had been conducting a clinical trial for a new drug and found that on average patient outcomes were no better under treatment we could not rule out the possibility that the drug worked favorably for some patients and unfavorably for others. The potential for heterogeneous treatment effects motivates an analysis which identifies subgroups within which units are likely to respond similarly to treatment.

In the following two subsections, we conduct two sets of such heterogeneous treatment effect analyses. The first subsection models the presence of each individual commenter's identity can be considered as a distinct treatment, dramatically increasing the number of parameters we need to estimate but also increasing the richness of our analysis. The second subsection describes estimates from our accumulated score cue models which posit that heterogeneity in identity effects can be attributed to the association between an identity and the ratings it has previously received.

5.2. Commenter-level Effects

A natural unit to measure the effect of identity within is the commenter. Commenters have persistent identities which are linked to a pseudonym they create when they register on the website. As they create content items over time, other users are likely to associate some

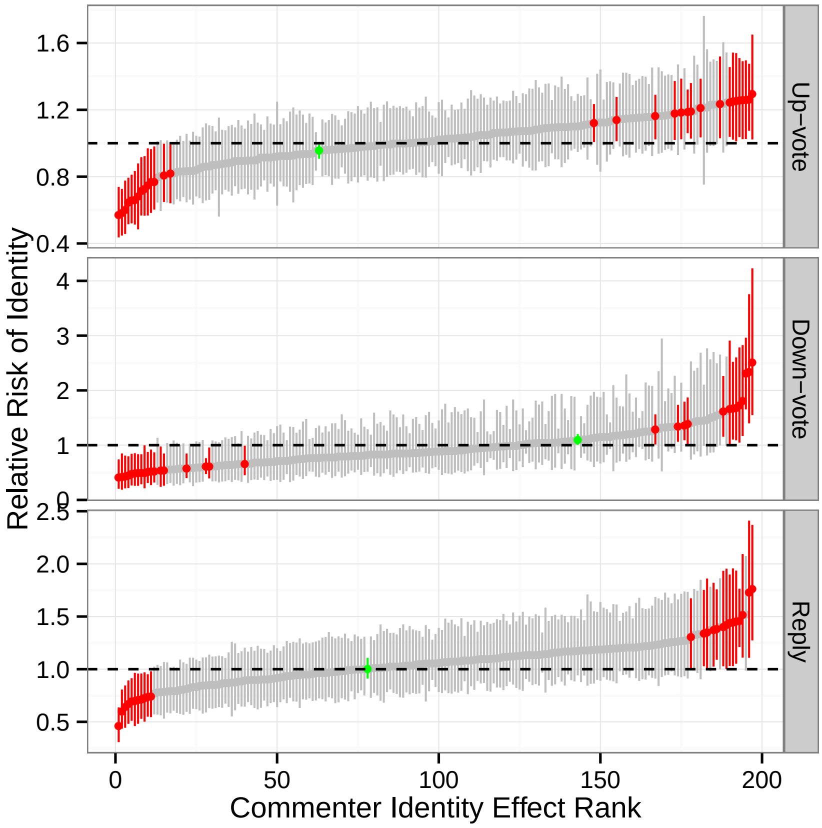


Figure 4 **Empirical Bayes estimates of relative risk of identity on up-vote, down-vote, and reply behaviors for 196 active commenters (those whose comments received at least 15,000 exposures). The green points are estimates pooled across the remaining less active commenters. The 95% confidence intervals are estimated using model estimates for each draw of a multiway bootstrap on comments and viewers, using 500 bootstrap replicates. Red points indicate estimates are significantly different from no change in relative risk at the 95% confidence level.**

quality, style, or interest information with this identity. Thus within-commenter identity effects can be interpreted naturally as the effect of a user's identity on the ratings and replies their comments receive.

Figure 4 displays estimated relative risks of identity on turnout and positivity for the 196 users whose comments generated at least 15,000 exposures in our experiment.[5] These

[5] We chose this threshold after doing a power analysis to see what sample size we would need to detect a 10% change in relative risks of identity.

users comprise a sample for which we can compute precise relative risks of identity. To compare these results to less active users where we lack statistical power, we pool together their observations into a single representative user for whom we can provide very precise estimates (displayed in green in the plots). Though we are averaging the identities of many users, these estimates provide interpretable results for identity effects pooled over new or less established users.

Under the null hypothesis of no effect of identity on any units, we would expect about 5% of these relative risks to be significant at the 95% confidence level. On the contrary, we find that 16% ($p < 0.001$), 16% ($p < 0.001$), and 13% ($p < 0.001$) of the commenters exhibit significant effects of identity on up-votes, down-votes, and replies, respectively.[6]

While the majority of users do not exhibit statistically significant relative risks of identity for rating or discourse interactions, we find evidence for every possible effect of identity on at least some of the active users. Among active commenters, some have both significantly higher and lower probabilities of receiving up-votes, down-votes, and replies because of their identity. Within this social system, identity cues appear to affect every type of interaction behavior.

We find that, on average, less active commenters have a significantly higher probability of receiving down-votes (95% confidence interval (CI) for relative risk: $[1.005, 1.203]$) and close to significantly lower probability of receiving up-votes (95% CI for relative risk: $[.907, 1.008]$). The presence of identity cues for these less active commenters produced no significant difference in reply rate (95% CI for relative risk: $[0.911, 1.109]$).

[6] Recall that we have used a hierarchical model to estimate these probabilities and relative risks, meaning that the we should consider whether the measured effects significantly differ from the mean of the empirical prior distribution on treatment effects we estimate (0.99, 1.05, and 1.04 for up-vote, down-vote, and reply relative risks). This does not substantively affect the results of any of our tests.

Our result for less active commenters – that their content is rated more negatively – is consistent with an interpretation that they users are biased against less active commenters or would like to initially provide them with stronger incentives to produce good quality content.

In Section 4.5 we derived conditions that allow us to interpret *decreases* in up-vote and down-vote rates as stemming from causal effects in viewers' opinions about comment quality. The assumption we need is that the net turnout effect of identity is weakly positive: that each individual commenter's identity will cause more users to rate than they would cause users to abstain from rating. Under these assumptions, one can interpret decreases in up-vote and down-vote rates (for which we find substantial evidence) as opinion change.

5.3. Cumulative Score Cue Effects

While the vast heterogeneity in the relative risks of identity is interesting on its own, it is worth considering how some commenters are able to achieve these effects. We have established that less active commenters, which includes a relatively large number of new users who recently created accounts, receive systematically worse ratings due to the presence of their identities. How then do the active commenters go from this state to identities which cause an increase in the probability of up-votes and a decrease in the probability of down-votes?

The explanation we explore in this section is that commenters establish a favorable or unfavorable identity through the aggregate ratings[7] their past comments have received. By creating content which in the past received high (or low) cumulative scores (possibly by creating high or low quality content), commenters may influence users into associating their identities with good (or bad) ratings.

[7] On the website for our field experiment, votes are aggregated by taking the count of up-votes and subtracting the count of down-votes. Comments begin with a rating of 0.

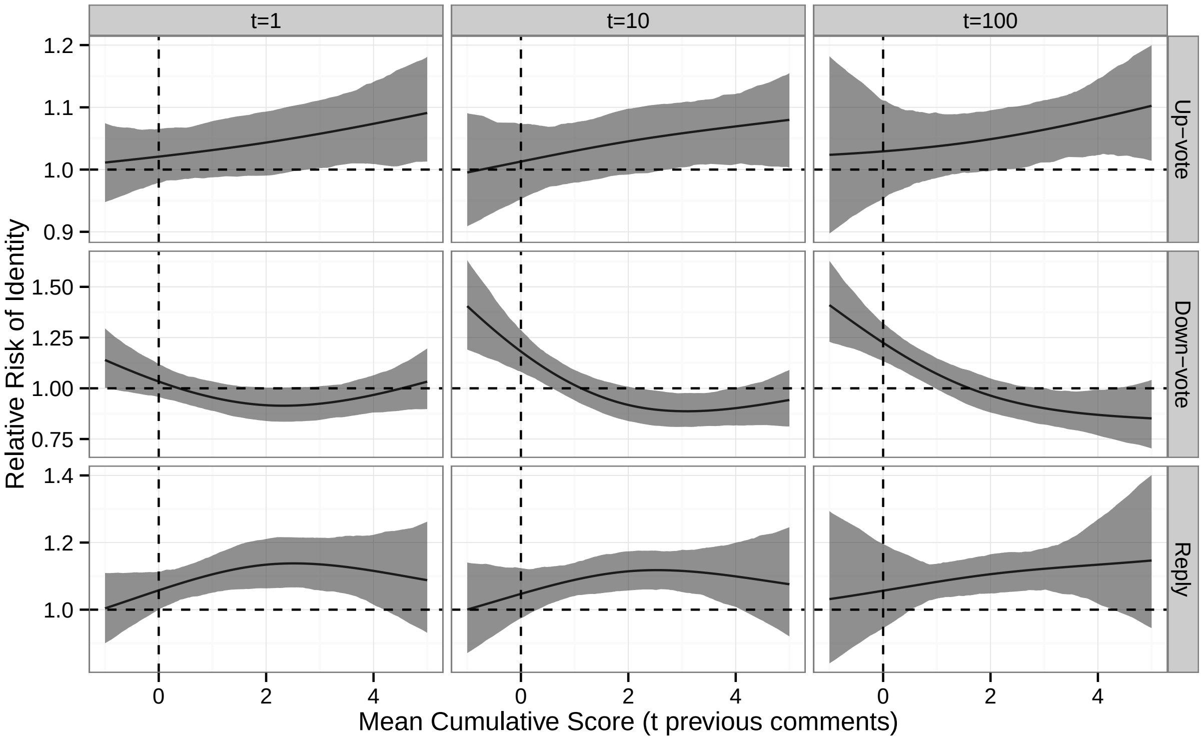


Figure 5 Relative risks of identity conditional on the mean score for commenter's previous t comments. We predict relative risks using generalized linear models with non-linear functions of mean cumulative score (using natural splines with 2 degrees of freedom). One set of relative risk predictions is made for each of 500 multiway bootstrap replicates, which we aggregate into 95% confidence intervals for each value of mean cumulative score.

We compute a quantity for each commenter's comment k which is the mean cumulative score (MCS) of her previous $t \in \{1, 10, 100\}$ comments. MCS represents a measure of quality that a viewer of the site is likely to have associated with that commenter's identity at the point that the viewer sees comment k. We compute MCS *per comment*, not per commenter, and therefore it can vary for a commenter over the course of her tenure on the site depending on the quality of her recent content.

Figure 5.3 presents relative risks of identity for comments where the commenter's identity has a MCS in range $[-1, 5]$, a range which accounts for between 80% and 85% of the comments in our experiment. Each column in the table represents a different value for t, and the roughly similar results demonstrate that the effects of MCS are robust to choice of this parameter.

It is important to note that although content quality should tend to vary by MCS, our randomization accounts for this. Along any given vertical slice of the plots in Figure 5.3, we are comparing the interactions of users with the same quality content in expectation (produced by users with the same MCS), the only difference being whether the identity cue is present or not. So though we find that MCS is indeed correlated with up-vote, down-vote, and reply behaviors as one might expect, our results are driven by exogenous variation in the presence of identity information. However, comparisons which are made left-to-right within the plots are not valid causal inferences since we did not exogenously manipulate MCS.

For discourse behavior, we see that identity causes an increase in the probability of a reply for commenters with positive MCS. Users seem to prefer to reply to commenters who consistently receive favorable ratings. The drop-off in the significance of the effect for high values of MCS is likely due to less precise estimates because we observe fewer comments in those bins.

We also see statistically significant changes in up-vote and down-vote rates caused by the presence of identities with higher and lower values of MCS. Having an identity associated with very low or negative mean scores causes commenters to have a significantly higher probability of receiving a down-vote. The effects of identity on up-vote rates are smaller yet still statistically significant – commenters with an MCS greater than 3 are significantly more likely to receive an up-vote due to the presence of their identity.

The association between MCS and relative risks of identity on ratings suggests that commenters can *accumulate* a favorable identity over time by initially producing higher quality content and then allowing that content to simultaneously increase both their current ratings as well as their future values of MCS.

6. Discussion

6.1. Large Scale Field Experiments

This study reports results from a very large scale experiment with millions of individual observations. Several themes emerge while analyzing the results of randomized field experiments on web and mobile platforms at this scale.

First, large scale experimentation can detect much fairly subtle effects for even rare events. In our study, replying to comments was relatively rare, but we were able to measure a significant increase in this rate due to identity – from about 1.1% to about 1.2%. When baseline response rates are low, a frequent occurrence in online advertising, this additional statistical power becomes dramatically more important.

Second, power from these experiments does not scale as quickly as one would expect due to dependence between observations. Unlike in lab studies, it is rare to run online experiments which do not allow repeated participation of users and content items. Ignoring these dependencies can result in anti-conservative estimates of effect sizes and an increased probability of type I errors. Our efforts to model these dependencies with the full set of experimental data introduced scalability problems and we were left unable to estimate full conditional models. While we and other marketing researchers prefer Bayesian estimation which leverages hierarchy in parameters and flexibly accounts for dependence structures, it remains a research frontier to apply these techniques with big data.

Third, larger experiments allow researchers to potentially identify and precisely estimate dramatically more parameters, creating opportunities to model heterogeneous treatment effects. Traditional field experiments focus on measuring average treatment effects, quantities which are averaged over the entire population. By incorporating information about users and content items, we were able to create richer models which showed that identity effects can vary substantially in sign and magnitude. The potential for policy improvements from this type of causal knowledge is substantial. We can make efficient use of costly resources if we know under what conditions manipulations work best and we can not apply them when they are likely to do more harm than good.

Finally, field experiments at scale create both challenges and opportunities for uncovering behavioral mechanisms underlying the raw results. The challenges are created by the fact that subjects' responses are usually voluntary – we can rarely directly elicit the revealed preferences or information we want from everyone. This introduces the possibility that the differences in behaviors we do observe at least partially result from different sub-populations participating depending on the treatment group (which we call *selective turnout*). The upshot is that despite running a potentially costly randomized experiment, we may not always be able to conclude that a treatment changed any individual's behavior.

On the other hand, more experimental data allows researchers to estimate more complex and flexible models such as our accumulated score cue models. Instead of interpreting point estimates and standard errors, we found it to be useful to present our results graphically – an approach facilitated by scale that we feel yielded deeper insights into underlying behavioral processes.

6.2. Identity Effects

Using a unique experimental design, we were able to broadly characterize how identity cues affect how users process and interact with organic content created by peers. Perhaps the most striking result of our analysis is the robust heterogeneity in the effect of identity cues on rating behaviors, both in magnitude and sign. With the group of more active users for whom we study commenter-level effects we find some who benefit strongly from their identities, receiving significantly more up-votes and fewer down-votes controlling for the quality of the content they produce. Other active users are less fortunate and receive systematically fewer up-votes or more down-votes when their identities are present than when they are anonymous.

The important implications of identity effect heterogeneity are that some identities are more valuable than others and some actually harm the perception of the content with which they are associated. Therefore any strategy which seeks to use social context cues in order to improve the probability of favorable online interactions, such a clicks or ratings, should be careful to understand for which identities it is likely to work.

Our cumulative score cue models suggest a mechanism which explains how users develop and maintain favorable or unfavorable identities. Using a fairly crude measure of average scores associated with identity cues, we find evidence for a feedback loop that creates accumulating or deteriorating identity value. Users appear to earn the good or bad biases caused by their identities, but perhaps they earn them in an unfair system that penalizes new users and rewards the results of previous upward biases. Our results show that the contributions of less active users are met with a negative ratings bias that presumably must be overcome in order to achieve fairer content ratings. Is is maybe unsurprising how closely our experimental context resembles the scientific reward systems discussed by Robert Merton almost half a century ago.

7. Conclusion

We have argued that identity plays an important role in many popular and economically important social systems. Services such as Facebook, Twitter, and Pinterest are characterized by democratic content creation, user interactions to rate (e.g. *like* or *favorite*) or disseminate (*share* or *retweet* content), ubiquitous identity cues, and repeated interactions between users over time. These social systems also employ advertising which leverages social context – implicit or explicit endorsements of branded content – in order to increase the attentional or persuasive effects of ad units. Thus we believe studying of of the role of identity cues is crucial to our understanding of user interactions with both organic content and also social advertising. Our experiment represents a microcosm of user content exposures and interaction which may be altered by the presence or absence of identity cues.

Though we believe our results are novel and will apply more broadly, we are left with substantial unanswered questions about why and under what conditions identity cues will be meaningful and persuasive for users. In particular, we look forward to seeing future work which explores the relational aspects of identity cues: how they develop *between* pairs of users, the role of prior offline interactions in determining trust and persuasion, and how users' relationships interact with content characteristics to affect behavioral outcomes beyond ratings.

References

Agichtein, Eugene, Carlos Castillo, Debora Donato, Aristides Gionis, Gilad Mishne. 2008. Finding high-quality content in social media. *Proceedings of the international conference on Web search and web data mining*. ACM, 183–194.

Aral, Sinan, Dylan Walker. 2011. Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management Science* **57**(9) 1623–1639.

Aral, Sinan, Dylan Walker. 2012. Identifying influential and susceptible members of social networks. *Science* **337**(6092) 337–341.

Aral, Sinan, Dylan Walker. 2013. Tie strength, embeddedness & social influence: Evidence from a large scale networked experiment. *Embeddedness & Social Influence: Evidence from a Large Scale Networked Experiment (January 8, 2013)* .

Bakshy, Eytan, Dean Eckles, Rong Yan, Itamar Rosenn. 2012a. Social influence in social advertising: evidence from field experiments. *Proceedings of the 13th ACM Conference on Electronic Commerce*. ACM, 146–161.

Bakshy, Eytan, Itamar Rosenn, Cameron Marlow, Lada Adamic. 2012b. The role of social networks in information diffusion. *Proceedings of the 21st international conference on World Wide Web*. ACM, 519–528.

Bapna, Ravi, Akhmed Umyarov. 2011. Are paid subscriptions on music social networks contagious? a randomized field experiment. *22nd Workshop on Information Systems Economics, Shanghai, China*.

Brenner, Joanna, Aaron Smith. 2013. *72% of Online Adults are Social Networking Site Users*. Pew Research Center's Internet & American Life Project. URL `http://pewinternet.org/Reports/2013/social-networking-sites.aspx`.

Duggan, Maeve, Aaron Smith. 2013. *6% of Online Adults are reddit Users*. Pew Research Center's Internet & American Life Project. URL `http://pewinternet.org/Reports/2013/reddit.aspx`.

Fisher, M., S.B. Friedman, B. Strauss. 1994. The effects of blinding on acceptance of research papers by peer review. *JAMA* **272**(2) 143–146. doi:10.1001/jama.1994.03520020069019. URL `+http://dx.doi.org/10.1001/jama.1994.03520020069019`.

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

Hanson, Ward A, Daniel S Putler. 1996. Hits and misses: Herd behavior and online product popularity. *Marketing letters* **7**(4) 297–305.

Merton, Robert K. 1968. The matthew effect in science. *Science* **159**(3810) 56–63.

Morris, Carl N. 1983. Parametric empirical bayes inference: theory and applications. *Journal of the American Statistical Association* **78**(381) 47–55.

Muchnik, Lev, Sinan Aral, Sean J Taylor. 2013. Social influence bias: A randomized experiment. *Science* **647** 651.

Owen, Art B, Dean Eckles. 2012. Bootstrapping data arrays of arbitrary order. *The Annals of Applied Statistics* **6**(3) 895–927.

Resnick, Paul, Richard Zeckhauser, John Swanson, Kate Lockwood. 2006. The value of reputation on ebay: A controlled experiment. *Experimental Economics* **9**(2) 79–101.

Ross, Joseph S, Cary P Gross, Mayur M Desai, Yuling Hong, Augustus O Grant, Stephen R Daniels, Vladimir C Hachinski, Raymond J Gibbons, Timothy J Gardner, Harlan M Krumholz. 2006. Effect of blinded peer review on abstract acceptance. *JAMA: the journal of the American Medical Association* **295**(14) 1675–1680.

Rossi, Peter E, Greg M Allenby. 2003. Bayesian statistics and marketing. *Marketing Science* **22**(3) 304–328.

Rubin, Donald B. 2005. Causal inference using potential outcomes. *Journal of the American Statistical Association* **100**(469).

Salganik, Matthew J, Duncan J Watts. 2008. Leading the herd astray: An experimental study of self-fulfilling prophecies in an artificial cultural market. *Social Psychology Quarterly* **71**(4) 338–355.

Shalizi, Cosma Rohilla, Andrew C Thomas. 2011. Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research* **40**(2) 211–239.

Taylor, Sean J, Eytan Bakshy, Sinan Aral. 2013. Selection effects in online sharing: consequences for peer adoption. *ACM Conference on Electronic Commerce*. 821–836.

Toubia, Olivier, Andrew T Stephen. 2013. Intrinsic vs. image-related utility in social media: Why do people contribute content to twitter? *Marketing Science* **32**(3) 368–392.

Tucker, Catherine. 2012. Social advertising. *Available at SSRN 1975897* .

Wang, Jing, Anocha Aribarg, Yves F. Atchadé. 2013. Modeling choice interdependence in a social network. *Marketing Science* **32**(6) 977–997. doi:10.1287/mksc.2013.0811.

Yang, Jaewon, Bee-Chung Chen, Deepak Agarwal. 2013. Estimating sharer reputation via social data calibration. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 59–67.

Yoganarasimhan, Hema. 2012. The value of reputation in an online freelance marketplace. *Available at SSRN 2178208* .