Supplementary Information

Social Influence Bias: A Randomized Experiment

Lev Muchnik,[1] Sinan Aral,[1, 2] & Sean J. Taylor[1]

[1]Information, Operations and Management Sciences Department, Stern School of Business, New York University.

[2]Center for Digital Business, Sloan School of Management, Massachusetts Institute of Technology.

Prior Work

Our experiment identifies the causal effect of social influence on rating behavior, as well as the mechanisms driving influence, by randomly manipulating user-generated ratings. A large body of research(*1-7*) has explored the dynamics of online reviews and ratings for books, restaurants, hotels, and more, as well as their relationship with economic phenomena such as firm strategy and consumer decision-making.(*8, 9*) Recent work in text-mining considers textual features of user reviews in addition to numerical ratings(*10, 11*). These papers use observational or quasi-experimental designs to assess the relationship between ratings and economic outcomes.

There is also substantial interest in how past ratings affect future ratings and public opinion in general, although causal estimates have been elusive. For example, Wu and Huberman(*12*) study how public opinion evolves over time. They measure rating dynamics in two different online ratings systems -- one where raters can see prior ratings before they rate and another where they cannot -- in order to understand how exposure to prior ratings affects rating behavior. As in Salganik and Watts (2008)(*13*) (discussed below), the analyses are conducted at the system level (on two separate websites) but the study is not experimental -- there are substantial differences between the two websites besides the ratings systems themselves that could confound estimates of the effect of social influence on ratings (the websites' content, purpose, user-base, etc.).

Wu and Huberman's results agree with ours on some dimensions but not others. They find that on the website on which no prior rating information is available people tend to exhibit more uniform rating patterns and that on the website where social information is made available, ratings tend to follow the group. This is broadly consistent with the positive herding we observe,

but is contrary to the correction effect we observe in the case of negative social influence. The dynamics of the correction effect may not be measurable in observational, system level outcomes. The group agreement they observe likely confounds correlated group preferences with social influence bias. The experimental method we use may therefore more accurately identify the asymmetry in social influence effects across negative and positive prior judgment.

Our experiment is related to a line of work that (quasi-)experimentally manipulates signals of the quality of items and measures the accompanying change in behavior. Friedkin and Johnson (2011)(*14*) report results of social influence experiments similar to Asch (1951)(*15*), which vary social structure in order to validate their network model of social influence. Sorenson (2007)(*16*) exploits mistaken omissions of books in the NY Times bestseller's list to identify the boost in sales from appearing on the list. Since Friedkin and Johnson's experiments are performed in a lab and Sorenson relies on the assumption that mistakes are exogenous to identify a causal effect, we will focus the comparison of our work to two prior studies that use randomized field experiments.

Hanson and Putler (1996)(*17*) randomly increase the download counters for software and observe that for those with the largest increase (100% more downloads), users are significantly more likely to download them than the baseline. Salganik and Watts (2008)(*13*) invert the ordering of songs that are ostensibly ranked by download count and track the change in the popularity of songs. In agreement with our work, both studies find herding effects in quality signals on user behavior, but there are four significant differences between our study and the experiments of Hanson and Putler (hereafter HP) and Salganik and Watts (SW).

First, in contrast to HP and SW, our study separates the effect of ratings bias from the effect of search costs. Both HP's and SW's perturbation of rankings have two possible mechanisms of effect: they change the quality signal for the items but they also change search costs in a correlated way. In both studies, the rank order of items was manipulated. As a result, users saw the highest rated song or most downloaded software at the top and had to scroll down a list to get to the bottom. The resultant downloads may in either case be due to the effect of a higher rating on the likelihood of download but may also be explained by the effect of lower search costs on the likelihood of clicking on the song/software in the first position. Result position is known to have a very strong effect on click through rates in web search results and other settings.

Separating the effect of ratings and the effects of search costs helps pin down social influence bias holding this very plausible alternative explanation constant.

Second, our study is an in vivo field experiment conducted on a live, public website while Music Lab was an online lab experiment in which a convenience sample of subjects were recruited to participate. The rating system used in our field experiment is also widely used on the web (c.f. Digg, Reddit, Hacker News). Furthermore, the manipulation used in SW, an inversion of all rankings, is not a change that is likely to be seen on a real website. HP's statistical significance result required them to increase the number of downloads by 50% or more, which one could argue is similarly unrealistic. In contrast, we attempted to design a manipulation that was realistic both in the context in which the experiment took place as well as the implemented manipulation that could represent a real change in what users' could experience, in other words to robustly identify the marginal effect of a discrete, precisely defined social signal.

Third, we report individual level effect estimates that reveal an asymmetry in voting behavior that would have been impossible to observe with the aggregated outcome measures in HP or SW. This difference is quite important because it enables our examination of heterogeneous treatment effects in positive and negative voting and for different types of users (e.g. friends or enemies). The asymmetric effects of negative and positive manipulation are theoretically and empirically important not only for our understanding of the nature of the biases created by exposure to prior ratings and opinions but also for the design of systems that attempt to aggregate collective intelligence. Our experimental design is different in that our unit of analysis is an item of user-generated content rather than systems of users acting in multiple simultaneous 'universes.' Both approaches have benefits and limitations, but we are able to estimate conditional models that identify subject- and item-specific effects. Thus, our experiment is capable of providing a micro-level explanation for what SW observe at the macro-level, providing parameters for the individual-level decision process that causes highly rated items to continue to become more highly rated. Because of our level of analysis, we are also able to conduct analyses of how much of the effect is driven by selection of subject types or drawing attention to items rather than changing rating valence (see Subgroup Analysis below).

Fourth, prior work confounds opinion and consumption. Whereas HP and SW measure the popularity of items by the number of downloads, we study users' subsequent ratings and

discourse as well as ratings and discourse mediated by social relationships and topics. In HP and SW, it is possible that the users did not always like the songs/software they downloaded. In contrast, our outcome measure is a less ambiguous elicitation of users' opinions about the items they are evaluating rather than a proxy for consumption without evaluation.

To summarize, our work is intended to extend the seminal contributions of HP and SW, et al. by a) measuring the direct effects of ratings in isolation without any other changes such as reordering content or reductions in search costs, b) quantifying the magnitude of social influence bias at the user and item level, c) estimating social influence bias across different topical domains, d) estimating the relative effect of social influence bias on friends and enemies e) identify the mechanisms driving social influence, and f) analyzing these effects in a real in vivo online environment. These differences enable novel insights that isolate the effect of past ratings on future ratings, capturing asymmetries between negative and positive herding, differences across topics and social relationships, and the relative importance of different behavioral mechanisms driving herd behavior.

Platform and Experimental Design

The experiment was conducted on a social news aggregation web site 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. These posts compete for a visible spot on the web site's front page as they are ranked by the community members. User comments on these posts develop into long discussion threads containing dozens and even hundreds of comments. User-generated comments are also ranked by community members who up- or down-vote them. The comment score observed by the users is the number of upvotes minus the number of down-votes observed until the time the impression of the comment is served. Users cannot learn the temporal evolution of the score or the identity of voters. Examples of the commenting and comment scoring features on Reddit.com are shown as an example in Figures S1a and S1b.

We worked in partnership with the web site administration to implement a randomized controlled experiment to assess the effect of comment scores on subsequent users' ratings and discourse. Every new comment submitted to the web site was assigned to either a control group or one of two treatment groups: 4% of comments were randomly set to present a score of +1 at the time the comment was created (hereafter referred to as the 'positive treatment', the 'upvoted' treatment group or the 'up-treated'), while another 2% were randomly set to present a score of -1 at the time the comment was created (hereafter referred to as the 'negative treatment,' the 'down-voted' treatment group or the 'up-treated'). The treatment group sizes were chosen to reflect the natural proportions of positive and negative votes that occur on the website. The treatment was applied at the comment level so that all users viewing that comment would be subject to the same treatment. Users do not observe the comment scores before clicking through to comments – each impression of a comment is always accompanied by that comment's current score, tying the comment to the score during users' evaluation, and thus mitigating selection bias on high (or low) rated comments. Users of the site are only allowed to vote for a comment once. They cannot change their vote once it is cast and cannot vote for a comment they themselves authored.

The experiment was conducted over months (163 days) between December 2010 and May 2011. During that time, users generated 101,281 comments of which 4,049 were positively treated and 1,942 were treated negatively. These comments were viewed over 10 million times by 3,600 users who cast 308,515 votes (Table S1). Random assignment guarantees that comments assigned to the control and treatment groups have identical quality, authorship characteristics, on-page positioning and other observable and unobservable properties in expectation. Any differences in the response of users to these comments can therefore be attributed exclusively to the experimentally manipulated scores.

We limited the comment score manipulations to +/- 1 for several reasons. First, for operational reasons we tried to avoid interfering with the normal operation of the web site. Scores highly inconsistent with comment quality may have altered the dynamics of user behavior in exaggerated ways that the website wanted to avoid. Thus, the results from our experiment are created by very subtle manipulations. Second, small perturbations are typical of attempts to defraud online ratings and are therefore of particular interest. Finally, we focused on the range of

scores the literature predicts will be the most psychologically significant, which typically occurs near zero as ratings shift from being positive to being negative (*18, 19*).

Two additional aspects of the experimental design are worth noting. First, in our analysis of the impact of ratings on the likelihood of up-voting and down-voting, we focus on the response of the first user exposed to the comment. Analysis of the response of subsequent users may potentially depend on the voting behavior of the users who saw and voted on the comment after the manipulation took place, making causal inference more difficult. We do however also separately analyze differences in the final ratings distributions of positively and negatively treated comments, which takes into account all accumulated ratings. Second, it is practically impossible for users to discover who upvoted or downvoted a comment, reducing the likelihood that experimental comments were evaluated with an eye toward reciprocity or retaliation and were instead likely to be evaluated on the merits of their content and their current scores.

Materials and Methods

In order to simplify the analysis and avoid confounding effects associated with user responses conditioned on the behavior of earlier users, we first analyzed the impact of prior ratings on rating and discourse behavior for the first viewer exposed to the comment. The first users exposed to a post experience a clean, controlled randomized treatment. The following comment viewers are affected by more complex treatment effects that confound treatment with the response (or abstention) of preceding non-experimental viewers. Reliable reconstruction of conditional response chains would require a significant increase in the number of observations and would not contribute to a meaningfully deeper understanding of the phenomena.

Model Specification

Our experiment uses randomization to guarantee that the treatment is exogenously assigned to comments, which allows us to interpret the differences in vote probabilities as average treatment effects. However, because the randomization is performed at the content level, users are recruited to become subjects non-randomly. This causes a dependency structure in our data that creates the potential for an anti-conservative bias in naive binomial proportion confidence intervals

computed across comments. To better account for the interdependence of observations in our sample, we formally model the variance components and employ Bayesian estimation of the resulting hierarchical models. We motivate the primary statistical models with a description of how we conceptualize the data generating process that describes user upvote behavior, noting that down-vote behavior is modeled equivalently except with a different dependent variable.

User *i*'s probability to upvote content c produced by user *j* can be represented by the following hierarchical generalized linear model:

$$P(UV_{ijc} \mid i, j, c, u_c, d_c) = \frac{1}{e^{-[\alpha + \beta_u u_c + \beta_d d_c + \varepsilon_{ijc}]} + 1}$$

We assume α, β_{u}, and β_{d} are random variables drawn from normal distributions with non-informative priors, and u_c and d_c are indicator variables of the up- and down-treatment respectively. The intercept α represents mean quality of content in the system under the control condition. Average treatment effects for the upvoted and down-voted treatments are measured (on a log-odds scale) by β_{u} and β_{d} respectively.

We now turn to the specification for the variance component, ϵ_{ijc}. If we assume that each potential rating event is independent, we need put no further structure on this unobserved disturbance term. However, due to the structure of our experiment, our model specification uses three random effects. First, we consider that some raters have different baseline probabilities of up-voting content, for example if they are more generous or strict raters. Second, it is possible that some users consistently create higher quality comments that are more likely to be upvoted by other users. Finally, since we repeatedly observe rater-commenter pairs, it is possible that the rater-commenter relationship affects the upvote probability, independently of the commenter's quality or the rater's generousness.[1] To account for these different sources of variance, we decompose our unobserved component into:

$$\epsilon_{ijc} = \rho_i + \kappa_j + \delta_{ij} + \nu_c.$$

Here ρ_i is the rater's random effect, $\kappa_j$is the commenter's random effect, δ_{ij} is a random effect accounting for the rating behavior between pairs of comment producers and raters. All three of

[1] Consider the case where I am a harsh rater and my friend produces poor quality content, but that I will always upvote his content because of our relationship.

these variance components are normally distributed with mean zero and variances τ_ρ, τ_κ and τ_δ respectively. These random effect variance hyperparameters are distributed as $InverseGamma(0.001, 0.001)$.

The remaining ν_c term represents the unobserved quality of the content holding variation across commenters, raters, and commenter-rater pairs constant.

Estimation

We estimate this hierarchical generalized linear model using Markov Chain Monte Carlo (specifically Gibbs sampling using the JAGS package in R, see Plummer, (2003)(*20*)). The Markov chain proceeds by generating draws from the set of conditional posterior distributions of our random parameters. For each model we ran three Markov chains for 2,000 iterations, discarding the first 1,000 iterations for "burn-in." We then used the last 1,000 draws to estimate the mean and 95% credible intervals for the posterior distributions of our parameters. We evaluated convergence by computing a potential scale reduction factor for each estimated parameter in the model (Gelman & Rubin, 1992)(*21*).

Marginal Effects

The baseline and treated comment effects are represented by the model parameters α, β_u, and β_d. The intercept α is the baseline log-odds of up-voting (or down-voting) and β_u, and β_d represent the increase in log-odds due to our two treatments. To convert these estimates to a probability scale and aid interpretation of the average treatment effects, we compute marginal effects of our treatments using simulation.

We use the fitted model and sample 1,000 draws from the joint posterior distribution of all model parameters. Each of these draws can be considered a distinct model that fits our data, and together they encompass our uncertainty about all parameters, including all random effects components. For each of these models, we compute the empirical probability of up-voting $P(UV = 1 | u_c = 0, d_c = 0)$, $P(UV = 1 | u_c = 1, d_c = 0)$, and $(UV = 1 | u_c = 0, d_c = 1)$. The 95% credible interval is then computed by taking the 2.5% and 97.5% percentiles across models. The procedure is the same for down-voting with a different dependent variable.

Integrity of Randomization

The validity of our results depends on the integrity of the random assignment of user comments to control and treatment groups. We test the integrity of the randomization procedure and confirm that no significant differences exist across any observable characteristics between either of the treatment groups and the control group. The tests of these differences are shown in Table S2.

We use one-way ANOVA to test for differences in distribution averages across the treatment groups and find no significant differences. We further confirm these results by computing the asymptotic p-values of the two-sample Kolmogorov-Smirnov test (*22*). Both ANOVA and the k-s test are applied to test differences between each of the four groups and the control group.

All posts are assigned to one of thirteen predefined categories (Figure S2 and Table S3). We find that while certain categories are prone to more commenting, voting and viewing activity, the randomization procedure ensures that comments in treatment and control groups do not display statistically significantly different distributions of post topics. We use a χ^2 test(*23*) instead of the KS-test to check whether the categories of posts hosting the comments from different groups are drawn from the same distribution because KS-tests are only valid for continuous distributions and not appropriate for categorical data.(*24*) The χ^2 test results are consistent with ANOVA and k-s tests on other observables and confirm the integrity of randomization procedure (Table S2). Table S4 summarizes the various forms of user response to each of the comment groups, showing that voting is effected by the perceived score, while the discourse is only affected by quality.

Social Network Data

Users are able to tag other users as users they ‘like’ or ‘dislike’ creating two mutually exclusive friendship graphs. Table S5 summarizes the networks formed by like and dislike relationships, which we refer to as ‘friends’ and ‘enemies.’ The average degree and the size of the network measurements exclude the isolates. Reciprocity is the fraction of bi-directional(*25*) relationships.

Average clustering coefficient(*26*) is computed for nodes with degree two or greater. In each of the networks the clustering coefficient of a single node is defined as:

$$C_i = \frac{1}{k_i(k_i - 1)} \sum_{j,k \in N_i} e_{jk} \ ,$$

Where N_i is the set of all friends (enemies) of the user *i*, and $e_{jk} \in \{0,1\}$ represents presence or absence of a directed edge between the two peers of *i*. Fig S3 displays a portion of the network of friends and enemies where each directed tie, if it exists, is mutually exclusively displayed as either a friend (blue) or enemy (enemy) relationship.

Model Fit

The models we fit in the paper are logistic regressions with random effects components to account for the dependencies between observations in our data. As necessitated by our design, we resample both comment viewers and comment authors throughout the study. Therefore, we prefer to estimate a conditional model that accounts for repeated sampling of subjects in our study in both roles. The random effects structure we specify is fairly general, modeling heterogeneity in raters' tendencies to both upvote and downvote, authors' tendencies to produce comments which are up-treated or down-treated, as well as unobservable rating effects which are idiosyncratic to specific viewer-author pairs.

The basic model we estimate makes three main assumptions that we feel are reasonable. First, the treatment effect parameters in our model are drawn from normal distributions. Second, all three levels of random effects in our model are drawn from normal distributions with mean zero. Third, we assume a logistic link function for converting our linear model into probabilities.

Note that we estimate separate models of upvote and downvote behavior to avoid making an assumption about how users' preferences for voting are related. For instance, an ordinal logit or probit would have assumed a single latent quality dimension could account for upvoting and downvoting behavior.

All models we present use only binary treatment and control variables, and include all possible interaction terms when controls are present. The models are saturated in the sense that, by combining upvote and downvote models, we have as many degrees of freedom as the full

conditional distribution of treatment effects. Therefore, the models are capable of perfectly replicating the mean voting responses by treatment group without any functional form assumptions.

To test that the model is not affecting treatment effect estimates, we now compare the marginal model with our conditional model that accounts for heterogeneity. The marginal model estimates treatment effects over the population distribution (our study contains a comprehensive set of voting behavior), obtaining treatment effect estimates by integrating out individuals' heterogeneities. This is precisely how we compute marginal effects using our random effects models (averaging over the sample), and the resulting vote distributions are centered approximately on the conditional probabilities we calculate. This can be clearly seen in Figure S4a, which compares our random effects (conditional) model estimates for the upvote and downvote models compared to the marginal model (binomial proportions). Confidence intervals in the plot are 95% simulated credible intervals for the random effects models and taken from the binomial exact test confidence intervals for the marginal model. Similar results (and more complicated figures) are obtained for the other models used in the paper for the by-category and by-relationship results. From viewing the 95% confidence intervals in the figure, our motivation for using a random effects formulation is clear. The confidence intervals produced when assuming that the observations are independent are anti-conservative. The point estimates remain unchanged by our estimation procedure.

We also consider an additional measure of model fit, the distribution of AUC (area under the ROC curve) for predicting upvotes and downvotes in-sample, conditional on the estimated parameters. This is a measure of the discrepancy between the observed voting data and the fitted model.

Since Bayesian estimation produces a posterior distribution over these parameters, we produce a distribution of AUC for the models using the following procedure. First we draw a full set of parameters from our posterior distribution, including all random effects components. We then calculate predicted probabilities of the voting behavior that is the dependent variable in the model. We compute a distribution of in-sample AUC as the area under the ROC curve produced by this procedure. The results for our basic models are presented in Figure S4b, which shows that the models fit the data fairly well (mean AUCs of 0.82 and 0.95, respectively). The downvote

model is more accurate at describing voting behavior because there are fewer downvotes in the sample and the additional degrees of freedom used in the random effects are capable of explaining more variance in downvoting.

Model Convergence

Since we make use of imbalanced and crossed random effects in our models, we use Markov chain Monte Carlo (MCMC) techniques to find posterior distributions for all parameters as well as the marginal effects of treatments on various outcomes of interest. An important assumption of MCMC estimation is that with a long enough Markov chain, we will be sampling from a stationary distribution representing the true posterior distribution of the parameters conditional on the observed data. In practice, there is no guarantee that the Markov chain has converged after any number of samples. Instead, we must rely on diagnostic tests to give us confidence that our posterior distributions have converged.

The Gelman-Rubin diagnostic test (*21*) is widely used to assess the convergence of models estimated with MCMC procedures. The diagnostic employs multiple Markov chains with different starting parameters to create pooled estimates of parameter variances. These pooled variances can be compared to the variance computed within each chain, resulting in a statistic called the potential scale reduction factor (PSRF). Typically, modelers will check if the upper confidence limit for the PSRF is > 1.2 for any particular parameter, and also compute a multivariate PSRF to test that all parameters in the model have converged.

Table S8 displays our computed PSRFs for every parameter in every model we estimate. With few exceptions, the diagnostic tests meet the heuristic of being less than 1.2. We do not see complete convergence of the Markov chain for only 3 of 64 parameters.

First, the PSRF for the dyadic random effect scale parameter in the model predicting downvote probability is slightly higher than 1.2. While there are a large number of random effects drawn with this variance parameter, they have a highly skewed distribution of appearance in the sample (i.e. some dyads appear many times while others appear only once). We believe that this parameter is slightly less well identified due to this skew in combination with the low prior probability of downvoting under either treatment.

Second, the scale parameters for the random effects in the upvote models within the enemy subsample seem to have low convergence. This is likely because these parameters are poorly identified in the data due to 1) limited data from sub-sampling and 2) a small number of random effects relative to the size of the sample, since many of enemy dyads are repeated many times.

Third, the treatment effect parameters in models with dependent variables related to discourse (number of responses and mean tree height) converged poorly. These dependent variables have unusual distributions and extremely high variance, making it more difficult for the models to fit the data well and achieve convergence.

In the first two cases, the lack of convergence occurred in variance component scale parameters in models which produced confidence intervals which led to positive results. Due to our approach of simulating over parameter estimates as well as data to compute confidence intervals, any excessive variance (which should be mild) in these hyperparameters will lead to a conservative bias in our hypothesis tests. This is because we incorporate model uncertainty into the confidence intervals and any upward bias in model uncertainty should tend to inflate them. Note also that 1) in the downvote model the PSRF is not particularly large and 2) in the enemies subsample we do not reject any of the null hypotheses so the slow convergence cannot cause a type I error. In these ways, any poor convergence we observe can only serve to make our results more conservative.

In the third case, the models were again used to correct the anti-conservative bias in the confidence intervals from marginal models and we could not reject the null hypothesis of no treatment effect. Since even the anti-conservatively biased estimates were too poorly estimated to show any treatment effect, this lack of convergence also plays little role in changing our results or conclusions.

Interference

Our experiment was conducted in the field and some subjects were repeatedly exposed to control and treated comments over the course of the experiment. Therefore, it is worth exploring whether the manipulations we employed had effects which extended beyond the observation they were randomly assigned to. Manipulating the scores of comments on the site could plausibly affect

rating behavior on comments that were shown nearby, a violation of the stable unit treatment value (SUTVA) assumption which allows us to claim an unbiased measure of causal effects.

To test for this type of interference, we conducted analyses designed to test for treatment effect spillovers between comments. We use the unique identifier of the comments (which is incremented as they are added on the site) to order them temporally. We consider a comment to be potentially subject to interference if it was created within 10 comments of a treated observation (either in the past or the future). We used different window sizes (1, 5, 15, and 20) and the results were broadly similar. With a window of +/- 10 comments, we find 55% and 32% of observations could potentially be subject to interference from an upvoted comment and downvote comment respectively. We use regression models with dummy variables representing proximity to treated comments to test whether the treatments led to interference both in first rater behavior and in the final score analysis.

Table S6 shows estimates from a logistic regression with parameters for each of our treatments before and after adding dummy variables for potential interference. We find no evidence of statistically significant effects from potential interference resulting from the proximity of either treatment. The treatment effect estimates also do not change substantially after including interference dummy variables in the regression. We conclude that interference is not likely to be playing a role in our analysis of the first viewer's rating decision.

In Table S7, we present estimates from linear regression models of 1) total number of votes and 2) ratio of positive votes for the comment final scores. We include the same four parameters as in Table S6. The results provide some weak evidence for interference between comments in our experiment as measured by the effect on accumulated rating behavior over the course of the experiment. Proximity to the comments which received the downvote treatment received a greater number of ratings and these ratings were more positive.

Despite this weak evidence for interference, we believe that the accumulated rating results we present in the experiment are still valid for three reasons. First and perhaps most importantly, the vast majority of comments that are subject to potential interference are control comments. If the interference is biasing our results, it will predominantly cause control comments to have larger turnout and positivity -- an indirect effect that would bias our direct treatment contrasts

conservatively. Second, proximity to a down-treated comment yields a very small absolute effect -- about an order of magnitude less than the direct treatment effect. Finally, because of randomization there is no systematic correlation in exposure to the direct treatments and the indirect treatments. Therefore we expect there to be no systematic bias added to our results by these small indirect effects, and this is reflected in the fact that the treatment effect estimates are unchanged by the addition of the interference dummy variables.

Estimating Behavioral Mechanisms

Summary of Findings

In our aggregate analysis, the positive manipulation created a positive social influence bias that persisted over time, generating accumulating herding effects that increased comments' final mean ratings by 25%. The negative manipulation on the other hand created no average herding in either direction due to a 'correction effect.' While these results describe the outcomes of the experiment, they do not reveal the behavioral mechanisms driving our results. We therefore analyzed changes in *turnout* (the likelihood of rating) and changes in *positivity* (the proportion of positive ratings) across subgroups in our study population to identify variance in our results that can be explained by *attention effects* and *opinion change* respectively. Here, we summarize the main findings of our analysis of behavioral mechanisms and then describe the analysis in more detail.

Our analysis of the mechanisms driving social influence bias leads to several broad summary findings that together explain the results of our experiment:

First, both treatments increase turnout within most subgroups (which are defined below), but neither creates differential turnout across any subgroup dimensions. This suggests that differential turnout by voter type (e.g. selecting different proportions of positive or negative raters, or dyads with frequent or infrequent rating interactions) cannot explain our results (at least for the theoretically motivated subgroups we considered).

Second, we find evidence for statistically significant opinion change in two of four theoretically-motivated subgroup dimensions in our data.

Up-treatment creates a systematic increase in the proportion of positive ratings for voters with little prior experience rating the particular commenter whose comment was manipulated and no decrease in positivity in any subgroup. This implies that positive opinion change explains part of the variation in the positive herding we observe.

The down-treatment on the other hand creates countervailing opinion change for positive and negative raters, canceling out any evidence of opinion change measured in the aggregate. Negative raters (users with the most negative ratings on control comments) used a higher proportion of positive ratings for down-treated comments, while positive raters (users with the most positive ratings on control comments) used a lower proportion of positive ratings under the down-treatment. These countervailing treatment effects on positivity across negative and positive raters explain why the ratio of positive ratings is similar for the control and the down-treated comments in the pooled data. This in turn helps explain why we find no aggregate trend in either direction for the negative treatment (i.e. no change in the final mean score for negatively treated comments).

Third, both treatments create a uniform increase in turnout compared to the control group, drawing attention to treated comments uniformly across voter types. This overall increase in turnout combined with a general trend toward positivity on this site creates a tendency toward positive ratings under either treatment.

Taken together, these results suggest that a mixture of a) opinion change and b) the natural tendency to upvote combined with greater turnout under both manipulations combine to create the herding effects we see. We also note that our analysis provides a conservative estimate of opinion change. This is because opinion changes in opposite directions across pooled treated observations will be reflected in higher turnout but not necessarily in changes in the proportion of positive ratings. These cases of countervailing opinion change mask the true level of aggregate opinion change created by our treatments because they cancel one another out in the aggregate.

Theoretical Basis for the Estimation of Behavioral Mechanisms

We can derive hypotheses about rating behavior by assuming a model of the rater's decision. Unfortunately framing the rating act as an economic decision yields few predictions about what

we should expect because the costs and benefits of rating are difficult to observe or quantify. Previous work on user reviews and ratings, such as Li and Hitt (2008)(*4*), tends to make the assumption that raters truthfully reveal their opinion of the item they are rating.

The economics literature on information cascades and herd behavior has provided key insights into how individuals use observable information from the past decisions of others, but the decision problems solved by individuals in these theories and experiments are precisely defined. For instance, in the seminal theories of Banerjee (1992)(*27*) and Bikhchandani et al. (1992)(*28*), agents must make a binary choice where it is assumed one of the choices yields a higher utility than the other. The subsequent lab experiments to test these theories, such as Anderson and Holt (1997)(*29*) and Çelen and Kariv (2004)(*30*), ask subjects to make similar choices between two alternatives which are easily ordered by expected utility. The rating decision in our context is inherently different because the individual does not receive an obvious benefit from being "right" in any objective sense, so we must make additional assumptions about preferences.

What costs and benefits can we assume raters receive from rating? It is natural to assume that it is costly to provide ratings, whether negative or positive, since the action occurs quite rarely. It is also relatively uncontroversial to assume a preference for honesty: i.e. that the likelihood of rating positively (negatively) should be correlated (anti-correlated) with the individual's private evaluation of the item's quality. However, these two assumptions do not tell us how the presence of prior rating information should alter rating behavior.

There are two theoretical pieces necessary to complete a theory of rating behavior in the presence of social information. First, we need a model of opinion change that describes how individuals' beliefs are changed by observing prior rating behavior. A compelling null hypothesis here might be that individuals do not change their beliefs in response to prior ratings. Second, we need to assume an objective function for rating that describes the goal of the potential rater when presented with a rating decision given any social cues and her (potentially changed) beliefs about the merits of the item.

Unfortunately, with two moving parts, it is difficult to pin down a model that makes precise theoretical predictions. While the literature on herd behavior and information cascades relies on a compelling objective function for individuals to test hypotheses about opinion change, studies

conducted in in-vivo settings are left with too many degrees of freedom. For example, it is possible to show that opinion change is observationally equivalent to a model with no opinion change but where individuals display a preference for conformity with the current social cue.

The story becomes more complex if one considers models with heterogeneity in opinion change or objective functions. For example, if some individuals change their beliefs while others don't or they experience differential response to social cues. The subgroup analyses we present below provide evidence that a more faithful model should account for this possibility.

Given that crisp, mutually exclusive theoretical hypotheses are not identifiable in our setting, we present the results of empirical analyses of behavioral mechanisms below and describe the theoretical interpretations of our findings in the following few paragraphs.

Opinion change is a satisfying explanation because it is closely related to the learning mechanisms in models of herd behavior. In our experimental setting, a theory of learning would also be similar to the theory of anchoring and adjustment (Tversky and Kahneman 1974)(*31*). However under modest assumptions, an unconditional opinion change model combined with rating behavior unaffected by social cues would predict a lower probability of positive ratings for negatively rated items – a prediction which conflicts with our finding of correction of down-treated items. Pure opinion change would also predict that a current positive rating would decrease the probability of a negative rating, but we do not find evidence for this either.

If we assume no opinion change takes place, there exist plausible objectives for raters that can explain our results. For instance, a preference for conformity with the current rating coupled with a preference for higher scores could rationalize our results without opinion change.

We can, however, rule out some possible objective functions for rating given our data. A preference for conformity with the current rating explains our positive and negative herding results but cannot rationalize the correction effect we observe. A preference for higher scores would explain the herding on positively rated items and the correction effect, but not the herding on negatively rated items.

If attention effects were solely driving our results -- which would be modeled as a preference for rating previously rated comments – then our aggregate results would have two implications.

First, negatively rated comments draw more attention than positively rated comments. Second, there is a rather particular distribution of opinion among raters. There must be a negligible number of raters who are on the margin between rating negatively and not rating (which is necessary for our null result on correction of positive manipulations), but enough raters just past this margin to account for our negative herding result.

However, if only attention effects are considered, then this would conflict with the findings of our subgroup analyses that the ratio of positive to negative ratings changes within certain subgroups due to both treatments. There must be some opinion change or differential preference for conformity in order to explain the changes in these ratios. This evidence against a pure attention effect explanation is presented in more detail in the next subsection.

After considering a number of models, our main conclusion is that our findings can be rationalized by a number of theories about rating behavior, but they can neither rule out an opinion change process nor can they be rationalized completely by an opinion change process such as learning. The main limitation in distinguishing between models using our data is that we cannot observe the rater's private opinion about the item. This limitation is inherent in any study that seeks to measure social influence in decisions where there is no past information about the subject's beliefs or preferences about something idiosyncratic. In order to better explore potential mechanisms for the effects we see, we need some additional information to serve as a proxy for the rater's opinion. In this next section, we consider treatment effects across subgroup dimensions that provide their information about users' prior opinions, their strength, and how they may interact with our treatments.

Distinguishing Attention (or Turnout) Effects from Opinion Change

Since subjects in our experiment must make a trichotomous choice, we can characterize the outcomes in our experiment along two dimensions. We define *Turnout* as the proportion of total comments which the subject rated, regardless of whether it was an upvote or downvote. We use this as a measure of attention given to comments across different subgroups and treatments. Conditional on turnout, we define *Positivity* as the proportion of votes which are upvotes. Positivity is our measure of the average opinion of users within a specific subgroup or treatment and it allows us to compute a measure of opinion change that we will argue is conservative.

A valid alternative characterization of our outcomes is to compute upvote and downvote proportions as we do in the main analyses in the paper. This calculation, given a well specified model, provides a good sense of the average treatment effect on the treated, but does not conceptually distinguish or empirically decompose changes in rating behavior due to increased attention and/or selective turnout on one hand and opinion change on the other.

The turnout-positivity characterization is meant to distinguish between effects due to increased rating without any opinion change (either uniformly or differentially across subgroups) and effects arising from treatments that change the opinions of subjects without changing their likelihood of rating.

With an appropriate outcome characterization in hand, we can look at the average treatment effect on a single outcome, say upvoting, as a measure of the change in two separate variables and their counterfactuals, specifically the change in *Turnout* (the number of votes) and the change in *Positivity* (the proportion of votes that are positive).

$$ATE_u = \sum_i (T_{i1}P_{i1}) - (T_{i0}P_{i0})$$

Here, $T_{it} \in \{0,1\}$ is comment *i*'s turnout under a particular treatment *t* (for example, here we take $t = 1$ for up-treated and $t = 0$ for control). Similarly, $P_{it} \in \{0,1\}$ is 1 if comment i is rated positively under treatment *t* and 0 otherwise. We can then decompose this expression into something more useful:

$$ATE_u = \frac{1}{N}\sum_{i=1}^{N}(T_{i1} - T_{i0})P_{i0} + (P_{i1} - P_{i0})T_{i1}$$

$$ATE_u = Attention\ Effect + Opinion\ Change$$

$$Attention\ Effect = \frac{1}{N}\sum_{i=1}^{N}(T_{i1} - T_{i0})P_{i0}$$

$$Opinion\ Change = \frac{1}{N}\sum_{i=1}^{N}(P_{i1} - P_{i0})T_{i1}$$

As is standard for randomized experiments, we only observe either T_{i0} or T_{i1} for any particular unit. We also only observe P_{it} under the condition that users turnout to vote, $T_{it} = 1$. Thus to estimate the attention effects and opinion change components we will need to derive four estimators which can be measured given what we observe.

$$E[Attention\ Effect] = E[(T_{i1} - T_{i0})P_{i0}] = E[(T_{i1} - T_{i0})]E[P_{i0}]$$

$$E[Opinion\ Change] = E[(P_{i1} - P_{i0})T_{i1}] = E[(P_{i1} - P_{i0})]E[T_{i1}]$$

P_{it} is actually a conditional random variable – we are only concerned with its distribution when $T_{it} = 1$ and therefore the two are uncorrelated. We use P_{it} and $E[P_{it}]$ to represent $P_{it}|(T_{it} = 1)$ and $E[P_{it}|T_{it} = 1]$. This gives us the necessary property to decompose the expectations, $E[P_{it}T_{it}] = E[P_{it}]\,E[T_{it}]\ \forall t = 0,1$ because as defined, P_{it} is independent of T_{it}.

We can now show simple estimators for each of our quantities of interest:

$$Change\ in\ Turnout = E[(T_{i1} - T_{i0})] = \frac{1}{N_1}\sum_{i=1}^{N_1} T_{i1} - \frac{1}{N_0}\sum_{i=1}^{N_0} T_{i0}$$

$$Turnout\ under\ Treatment = E[T_{i1}] = \frac{1}{N_1}\sum_{i=1}^{N_1} T_{i1}$$

$$Change\ in\ Positivity = E[(P_{i1} - P_{i0})] = \frac{\sum_{i=1}^{N_1} P_{i1}}{\sum_{i=1}^{N_1} T_{i1}} - \frac{\sum_{i=1}^{N_0} P_{i0}}{\sum_{i=1}^{N_0} T_{i0}}$$

$$Positivity\ under\ Control = E[P_{i0}] = \frac{\sum_{i=1}^{N_0} P_{i0}}{\sum_{i=1}^{N_0} T_{i0}}$$

For any observation *i* in our data set, we cannot measure expected turnout and positivity, or the treatment effects on these quantities because we cannot observe the necessary counterfactuals for any observation. However, these measurements are possible for any subgroup in our population (that displays at least some turnout) because there are observations in each treatment group and the treatment effect components can be estimated by calculating differences compared to the control group outcomes, which serve as counterfactuals.

If we could observe a comment-level counterfactual, this decomposition would provide an unbiased measure of the amount of opinion change taking place. But within a subgroup, the opinion change component of the ATE must be a lower bound estimate of opinion change. This is because some part of any opinion change that is inconsistent among members of the subgroup will offset and serve to change turnout but not positivity.

Our decomposition of outcomes is likely to be biased against finding evidence for opinion change, since subjects' opinions can be changed in a balanced way that preserves the aggregate proportion of positive ratings while simultaneously changing our measure of attention/turnout. For example, some people could change their opinion from positive to negative while others change their opinion from negative to positive in such a way that maintains the proportion of positive votes but that increases total turnout so as to increase the overall rating.

As an example of how this conservative bias works, consider a group of eight subjects in which four are in the control condition and four are in the treated condition. In the control condition, we observe that one subject rates positively and one rates negatively. In the treated condition two subjects rate positively and two rate negatively solely due to opinion change (meaning the two that rated positively would have not voted or rated negatively in the control and the two that rated negatively would have not rated or rated positively in the control). In the aggregate, we would measure the treatment effect in this case to be a) a doubling of turnout but b) no change in the aggregated level of positivity attributed to the treatment. This is a clear example in which opinion change is masked because of the lack of a counterfactual. If however we could separate the population of eight into two subgroups that were likely to change opinions similarly in response to the treatment, then we could detect opinion change.

Thus we expect our change in positivity to be a conservative measure of opinion change in the aggregate when there is countervailing opinion change among subgroups in the population. We also expect this downward bias to be minimized when we evaluate effects on subgroups that more unanimously change their opinions in response to treatments, rather than when we evaluate subgroups that have countervailing changes in opinions that 'cancel out.'

We also note that we can only measure in vivo opinion change effectively by studying a decision that has at least two possible outcomes (e.g. up voting or down voting). In lab experiments,

opinion change can be measured using pretreatment opinion surveys, which themselves have limitations. But, in in vivo settings in which behaviors provide proxies for opinions pre- and post-treatment, one can only measure opinion change when at least two different expressions of opinion exist (e.g. up vote/down vote; democrat/republican), above and beyond the decision to express an opinion at all (in this case the decision to vote). This is because an observed outcome can be attributed to either 1) the decision to express an opinion (or to attend to a decision at all) (attention/turnout) or 2) a change in the subject's opinion (opinion change).

For instance, in Salganik and Watts (2008)(*13*) and Hanson and Putler (1996)(*17*), the outcomes of interest are downloads of songs or software products. Though experimental, these studies are incapable of attributing the observed social influence to opinion change because the changes in observed behavior could plausibly be attributed to changes in attention. Even the seminal work on social influence by Asch (1951)(*15*) can only measure opinion change due to the experimental control of a lab experiment. Asch could force subjects to attend to the decision and he could reasonably assume that their prior beliefs (about the length of the lines) were accurate.

As we show in the next section, when two or more options are available to experimental participants, we are able to provide a conservative estimate of opinion change. We do this by looking for changes in the distribution of choices for the subjects who choose to attend to the decision.

Subgroup Analysis

There are two main motivations for performing a subgroup analysis on our data:

First, we want to test the degree to which turnout alone can explain our results. If certain subgroups differ in their positivity and also experience differential turnout in response to treatment, our results could be explained purely by selection (e.g. more positive voters turn out more in response to a treatment, there is no opinion change, but the ratio of positivity changes due to differential turnout across positive or negative voters).

Second, the previously discussed conservative bias in the opinion change component of our average treatment effect can be mitigated if we analyze subgroups which are homogenous with respect to opinion change due to treatment.

Leveraging the repeated appearance of raters and comment authors in our sample, we construct subgroups based on user rating behavior on our control comments. Under the assumption of no interference, which we validate with robustness tests described in the SOM section on Interference above, these data serve as a valid source of "pretest" variables similar to those typically gathered beforehand in a traditional pretest-posttest control group design (Campbell and Stanley, 1963)(*32*). We compute measures of activity and positivity for all raters, commenters, and rater-author pairs that we observe on the control comments. We then use these as pretest measures to divide our sample of comments into theoretically motivated subgroup dimensions that may vary in both baseline rating behavior and response to treatment.

Let $\Delta T_i^g = T_{i1}^g - T_{i0}^g$ represent the treatment effect on turnout for all subgroups contained in the set of sugroups g. Let $\Delta P_i^g = P_{i1}^g - P_{i0}^g$ represent the treatment effect on positivity for all subgroups contained in the set of sugroups g. Let n_g be the population size of g. Then, the total average treatment effect across the entire population of g is the mean ATE across all the groups in g:

$$ATE = \frac{1}{\sum_g n_g} \sum_{g=1}^{G} \sum_{i=1}^{n_g} \Delta T_i^g P_{i0}^g + \Delta P_i^g T_{i1}^g$$

When looking at two subgroups k and l, the total average treatment effect in the subpopulation made up of those two subgroups is the weighted mean ATE of the two groups:

$$ATE = \frac{1}{n_k + n_l} \left[\sum_{i=1}^{n_k} \Delta T_i^k P_{i0}^k + \Delta P_i^k T_{i1}^k + \sum_{i=1}^{n_l} \Delta T_i^l P_{i0}^l + \Delta P_i^l T_{i1}^l \right]$$

Assume for simplicity of exposition that the two subgroups are equally sized (we relax this assumption in the analysis itself), such that $n_k = n_l = n$. We can then rearrange this into the following equation:

$$2n * ATE = \sum_{i=1}^{n} \left(\Delta T_i^k P_{i0}^k + \Delta T_i^l P_{i0}^l \right) + \left(\Delta P_i^k T_{i1}^k + \Delta P_i^l T_{i1}^l \right),$$

such that:

$$2n * ATE = \sum_{i=1}^{n} \begin{matrix} \left(\Delta T_i^k - \Delta T_i^l\right)\left(P_{i0}^k - P_{i0}^l\right) + \left(\Delta T_i^k P_{i0}^l + \Delta T_i^l P_{i0}^k\right) + \\ (\Delta P_i^k - \Delta P_i^l)(T_{i1}^k - T_{i1}^l) + (\Delta P_i^k T_{i1}^l + \Delta P_i^l T_{i1}^k) \end{matrix}$$

where:

$$2n * ATE = Attention\ Effect + Opinion\ Change$$

$$Attention\ Effect = Selective\ Turnout + \left(\Delta T_i^k P_{i0}^l + \Delta T_i^l P_{i0}^k\right)$$

$$Opinion\ Change = Selective\ Opinion\ Change + (\Delta P_i^k T_{i1}^l + \Delta P_i^l T_{i1}^k)$$

and:

$$Selective\ Turnout = \left(\Delta T_i^k - \Delta T_i^l\right)\left(P_{i0}^k - P_{i0}^l\right)$$

$$Selective\ Opinion\ Change = (\Delta P_i^k - \Delta P_i^l)(T_{i1}^k - T_{i1}^l)$$

The first line of the ATE expression is the Attention Effect discussed above, decomposed into effects from two groups, while the second line is the Opinion Change effect similarly decomposed. We discuss each one in turn.

The first component of the attention effect, $\left(\Delta T_i^k - \Delta T_i^l\right)\left(P_{i0}^k - P_{i0}^l\right)$, is the differential effect of treatment on turnout between the two groups multiplied by their baseline positivity. If this term is non-zero, then the two groups differ with respect to their baseline positivity and also turnout differently in response to the treatment. Thus, this component represents the contribution of selective turnout across groups to our overall average treatment effect.

To test for selective turnout as an explanation for our treatment effect, we create sets of subgroups for which $\left(P_{i0}^k - P_{i0}^l\right)$ is non-zero, and then test the hypothesis that $\left(\Delta T_i^k - \Delta T_i^l\right)$ is non-zero.

The second component of the attention effect is the non-selective turnout effect (a uniform increase in turnout across groups), which is simply measured as the remaining variation in the attention effect.

The first component of the opinion change effect is selective opinion change -- the case where there is differential turnout between the two groups and where they have different amounts of opinion change in response to the treatment. To test for selective opinion change, we test the two hypotheses of differential turnout and differential opinion change between our subgroups.

The second component of the opinion change effect is $(\Delta P_i^k - \Delta P_i^l)$. If ΔP_i^k and ΔP_i^l have opposite signs, then despite each subgroup having non-zero opinion change, this term can be close to zero even in the presence of significant (though countervailing) opinion change. This illustrates mathematically how countervailing treatment effects between two subgroups can cancel out and diminish the measured contribution of opinion change to the treatment effect.

We suggest that the amount of (potentially opposing) opinion change should be measured by the following quantity:

$$Absolute\ Opinion\ Change = |\Delta P_i^k\ |T_{i1}^k + |\Delta P_i^l|T_{i1}^l$$

Absolute Opinion Change will be equal to Opinion Change only under the condition that either $\Delta P_i^k > 0$ & $\Delta P_i^l > 0$ or $\Delta P_i^k < 0$ & $\Delta P_i^l < 0$ i.e. that the treatment has the same direction effect on positivity for both subgroups. If this condition is not met, then the absolute value of the Opinion Change component of the ATE will be an underestimate of Absolute Opinion Change.

We now describe the theoretically motivated dimensions on which we create subgroups and test variation in response to treatment.

Subgroup Dimension 1: Rater Positivity

Raters may vary with respect to the degree to which they rate comments positively. More positive raters may be less discerning about quality or less likely to read and evaluate comments carefully and therefore rely more on social cues. Positive raters are more likely to rate positively, so some treatment effect could be explained by a selective turnout effect across this dimension.

We evaluate active raters because we need sufficient data with which to characterize the positivity of a raters' voting behavior. We define an "active rater" as one who rated at least 100 control comments (not necessarily as the first viewer), which accounts for 20% of the first viewers and 86% of the comments in our data. We call a user a "positive rater" if her positivity,

measured by the proportion of her ratings which are positive, is greater than the median positivity in the sample (0.84).

Subgroup Dimension 2: Commenter Quality

Raters may be more or less likely to rely on social cues when rating a comment as a function of the commenters' quality. Comments from high quality commenters may create more or less turnout under our treatments. Therefore, as in the rater positivity dimension, our findings could in part be a result of selective turnout along this dimension.

Within the group of active commenters (100 or more comments created, accounting for 83% of observations), the median commenter quality, measured as the proportion of positive ratings on their comments, is 0.87. We therefore divide active commenters into high and low quality subgroups whose average quality is greater than or less than this median respectively.

Subgroup Dimension 3: Rating Interaction Frequency

Frequency of rating interactions is a measure of the rater's experience evaluating the commenter. With more evaluation experience with a particular commenter, the raters may have less reliance on social cues as a substitute for forming their own opinions about comments. On the other hand, with less experience rating the commenter, raters with fewer interactions may evaluate the comment more thoroughly before deciding to rate, which could diminish the effect of the social cue.

When a user rates a comment, they form a rater-commenter pair. We observe almost 72,678 of these pairs from almost 287,750 ratings of control comments. We call a rating relationship active if the rater rated the commenter's control comments at least 20 times. These active rating relationships account for about 27% of our first viewer observations. Since 20 rating observations is not sufficient to precisely estimate relationship positivity and the median positivity among this group is quite high at 0.97, we only compare pairs with frequent versus infrequent rating interactions (rather than positive or negative relationships, given they are active). This provides an exhaustive split of our data.

Subgroup Dimension 4: Articulated Relationship

As described in the text of the paper, users on the site can articulate that they are friends or enemies with other users. Users may be more generous when rating their friends and more likely to rate comments made by their enemies poorly.

In about 37% of the observations in our experiment involve a pair where the rater had articulated a relationship with the commenter. In 84% of these observations where a relationship was articulated, the users were friends. Due to the small number of observations where the rater and commenter are enemies, our estimates are relatively imprecise for this subgroup.

Subgroup Analysis Procedure

For each subgroup we compute turnout and positivity under the control, up-treated and down-treated conditions (quantities discussed above as estimates $E[T_{i1}], E[T_{i0}], E[P_{i1}], E[P_{i0}]$). We then estimate the effects of each treatment on turnout and positivity, which we denoted $E[\Delta T_i]$ and $E[\Delta P_i]$. When comparing probabilities, we examine odds ratios instead of differences, e.g. $E[T_{i1}/T_{i0}]$ instead of $E[T_{i1} - T_{i0}]$ and $E[\Delta T_i^k/\Delta T_i^l]$ instead of $E[\Delta T_i^k - \Delta T_i^l]$, and test the null hypothesis that the ratio is not equal to 1.

We test significance in odds ratios using Fisher's exact test. For turnout, we construct a contingency table of turnout versus non-turnout for treatment versus control comments. For positivity, we construct a contingency table of positive versus negative ratings for treatment versus control comments, using only the cases where turnout occurred. Thus the contingency tables for treatment effects on positivity have much lower counts and the odds ratio estimates for positivity changes are less precisely estimated than the odds ratio estimates for turnout.

We then estimate the differences in these quantities between subgroups along our four dimensions (Rater Positivity, Commenter Quality, Rating Interaction Frequency, and Articulated Relationship). We provide contrasts in levels of turnout and positivity for comments in the control and treated conditions: $E[T_{i1}/T_{i0}]$ and $E[P_{i1}/P_{i0}]$. We then estimate differences in the treatment effect on turnout $E[\Delta T_i^k/\Delta T_i^l]$ and the treatment effect on positivity $E[\Delta P_i^k/\Delta P_i^l]$.

To compare turnout and positivity across groups, we again construct contingency tables. For turnout this is again turnout versus non-turnout for each of the two subgroups in the subgroup dimension. The contingency tables for positivity are constructed similarly and we again use Fisher's exact test to look for the significance of the odds ratio.

Finally, to test for differences in treatment effects across groups, we compare treatment effect odds ratios for each group. The log of each odds ratio is approximately normally distributed with standard errors which are straightforward to compute. We therefore test the null that the difference in these log odds ratios is equal to zero.

We summarize the results of our subgroup analyses in Tables S9 and S10 which present results by subgroup and contrasts of subgroups along the previously outlined theoretical dimensions respectively. The full results, organized by subgroup dimension and outcome variable (turnout or positivity) are shown later for reference (see Tables S11-S18).

In the next four subsections, we review the results from subgroup analyses conducted along the four subgroup dimensions we motivated earlier. Note that these are different dimensions, not groups, and represent a dichotomous split of the bulk of the data along one of the "pre-treatment" measures we describe. The groups are therefore not mutually exclusive between dimensions and comparisons will only be made within one dimension at a time, but not across dimensions. Finding evidence for selective turnout or opinion change along any one of these dimensions should be sufficient to conclude that these mechanisms are part of the story in explaining our treatment effects. Inversely, a lack of evidence for these phenomena along any dimension should not constitute evidence that they are not in play, because there may exist a subgroup dimension in which they could be found.

Results for Rater Positivity Dimension

Results for the rater positivity subgroup dimension are listed in Tables S11-12 and Figures S6-7. Both positive and negative raters are more likely to turnout than the sample mean because they are more active users (recall that we omitted raters with fewer than 100 total ratings). Partly by construction, positive raters use significantly more positive ratings on first viewing than negative raters on control comments. Under up-treatment this difference in positivity persists, but under

down-treatment the raters converge toward using the same ratio of positive and negative ratings (see Figure S6).

Both positive and negative raters are significantly more likely to turnout under either treatment than under control, but the differences between these effects are not significant across the subgroup dimension. This indicates that despite the large difference in positivity, differential turnout along this dimension cannot explain our main treatment effects.

On the other hand, we do observe statistically significant changes in positivity for both subgroups under down-treatment, which are significantly different from one another and in offsetting directions (see Figure S7). Negative raters become more positive in response to down-treatment while positive raters become more negative. Herding on down-treatment seems to be due to the behavior of positive raters, while the correction effect we observe is attributable to negative raters. One explanation is that negative raters are more discerning in general and better able to recognize when a rating conflicts with comment quality.

These subgroup treatment effects on positivity would be obscured by aggregation. If we were to compute the effect of the down-treatment on positivity for these subgroups in the aggregate, the odds ratio would be 0.995 ($p < 0.999$), showing exactly no opinion change effects from the treatment. This illustrates how this measure of opinion change is biased conservatively as we aggregate groups of observations where the treatment effect has a different sign.

The effect of up-treatment on positivity is not significant for either subgroup, nor is there a difference between the two. Negative raters remain negative, while positive raters remain positive.

Results for Commenter Quality Dimension

Results for the comment quality subgroup dimension are listed in Tables S13-14 and Figures S8-9. As with the positive rater subgroup, comments by high quality commenters receive more positive ratings under control. There are also some slight differences in turnout under control but they are not significant under either treatment (see Figure S8).

Comments by high and low quality commenters are both more likely to create turnout under down-treatment than control; comments by low quality commenters are more likely to create

turnout under up-treatment. However, neither treatment creates differential treatment effects on turnout across this subgroup dimension – again negating the possibility that differential treatment effects are driving our results (see Figure S9, bottom row).

The effects of either treatment on positivity also do not seem to significantly vary across this subgroup dimension (see Figure S9, top row).

Results for Rating Interaction Frequency Dimension

Results for the rating interaction frequency subgroup dimension are listed in Tables S15-16 and Figures S10-11. Rater-commenter pairs with frequent rating interactions (hereafter the "frequent" subgroup) are more likely to see turnout and greater positivity under control. These differences seem to be diminished by up-treatment, where positivity is essentially equal, and exacerbated by down-treatment, where the frequent subgroup is significantly more positive than the infrequent subgroup (see Figure S10).

Treatment effects on turnout are significant and positive across this subgroup dimension for down-treatment, with no difference between the subgroups. The effect of up-treatment on turnout is significant and positive for the infrequent subgroup, but the difference between the two subgroups is not significant (see Figure S11, bottom row).

There is evidence of a significant treatment effect on positivity under up-treatment for the infrequent subgroup ($p < 0.028$). While the infrequent subgroup is less positive on control comments than the frequent subgroup, under up-treatment they display approximately the same level of positivity. Analyzed along this subgroup dimension, the increase in positive ratings from up-treatment seems to be at least partly explained by changing the proportion of positive ratings used by the infrequent subgroup (see Figure S11, top-right). This fits our explanation that little rating experience leads to a greater reliance on social cues and more opinion change in response to treatment among infrequent raters.

Results for Articulated Relationship Dimension

Results for the articulated relationship subgroup dimension are listed in Tables S17-18 and Figures S12-13. It should be noted from the outset that the enemies subgroup is very small and

the measurements and treatment effects are imprecisely estimated. This also affects the precision of contrasts in treatment effects across the two subgroups.

Friends and enemies display similar probabilities of turnout under control and up-treatment, but for down-treatment friends increase their turnout while enemies' turnout is unchanged. Friends are always significantly more positive than enemies, regardless of treatment status (see Figure S12).

Friends display positive and significant effects on turnout from both up- and down-treatment, whereas enemies do not appear to change turnout behavior in response to treatment. Friends are so positive on control comments (P = 0.98) that there is little room for improvement in their positivity. On the other hand, enemies are incredibly negative on control comments (P = 0.32) and yet there is a decrease in their positivity toward up-treated comments that is mildly significant ($p < 0.097$). It appears that on the whole, friends uniformly rate positively and are more likely to turnout under treatment, while enemies are slightly more negative under up-treatment.

Subgroup Analysis Summary

We find strong support that increased turnout is a fairly universal response to our treatments. In all but three subgroups both treatments' effects on turnout are significantly positive for both the up- and down-treatments. The high quality commenters and frequent rating interaction groups still have positive effects of up-treatment on turnout, but they are not significant, though they are close ($p < .154$ and $p < .108$, respectively). The Enemies subgroup is small and imprecisely estimated, but we see no evidence of decreased turnout. Since positivity is always non-zero, we can conclude that the attention effect component of the ATE is always positive.

Recall that the selective turnout component of the attention effect is only non-zero if 1) positivity on control comments differs across subgroups ($E[P_{i0}^{k}/P_{i0}^{l}] \neq 1$) and 2) the subgroups turnout differentially ($E[\Delta T_i^k/\Delta T_i^l] \neq 1$). Through construction of our subgroup dimensions, we observe significantly different positivity on control comments across all four dimensions (see Table S10, column 4), which means condition (1) is met. However, we find no evidence at the 5% significance level or lower that turnout differs along any subgroup dimension.

We also find evidence for significant absolute opinion change for both treatments along multiple subgroup dimensions. First, we find that $E[P_{i1}^k / P_{i0}^k] \neq 1$ for positive and negative raters under down-treatment. Since these effects were different in direction, the opinion change component aggregated over the rater positivity subgroup dimension provided an underestimate of the absolute opinion change taking place under down-treatment.

Second, we find significantly increased positivity for the infrequent rating interaction subgroup under up-treatment and the difference in treatment effects on positivity across this subgroup dimension, $E[\Delta P_i^k / \Delta P_i^l] \neq 1$, is significant at the 10% level ($p < 0.078$). Turnout under up-treatment is only half as probable ($p < 0.001$) for the infrequent group. Together, these two differential effects contribute to a negative and marginally significant estimate of the selective opinion change term, $E[(\Delta P_i^k - \Delta P_i^l)(T_{i1}^k - T_{i1}^l)]$. The subgroup with the greater opinion change effect is turning out with lower probability, diminishing the total ATE.

We also tested the aggregate influence of attention effects in our findings at the level of final mean ratings. Final mean ratings could have been driven by an accumulating attention affect, whereby our manipulation drew attention to comments and increased the rate of voting rather than the relative proportion of positive and negative votes. As the final score is the number of positive votes minus the number of negative votes, a comment with nine hundred positive votes and one hundred negative votes would have a final mean score far exceeding a comment with nine positive votes and one negative vote, despite having identical proportions of positive and negative votes. However, when we compared differences in the ratio of positive to negative votes and the raw number of votes across treatment groups, we found that our manipulation a) changed the ratio of positive to negative votes in the expected directions, indicating a significant bias in the scores of manipulated comments, b) did not produce differential turnout across the treatment groups but c) did increase turnout compared to the control group.

These analyses again corroborate the explanation that a mixture of a) changing opinion and b) the natural tendency to upvote combined with greater turnout under both manipulations combine to create the herding effects we see.

A Note about Data Access

There are legal obstacles to making the data available and revealing the name of the website. The decision to not release the data was not ours but that of the website administrators. They are concerned about the re-identification of individual users and the associated risk to their privacy. We understand this concern. Although there is a scientific need for the validation of results in these types of studies, there are competing concerns about the welfare of users. We note that this is not any issue about the replication of results in other settings, but rather the validation of results using the same data. In essence, there is a tricky balancing act in the decision to release such data between access to the scientific advances that can be made with access to this kind of data and experimental setting on one hand and the associated restrictions that website administrators place on release of the data in an effort to protect the privacy and welfare of their users on the other. In the end, they are making decisions about valid risks to their users' private data and represent the entities responsible for making policy on these data. We are not in favor of a draconian policy in either direction but support appeals for more openness about the subject.

References

1. P.-Y. Chen, S. Dhanasobhon, M. D. Smith, All reviews are not created equal: The disaggregate impact of reviews and reviewers at amazon. com. *working paper, Carnegie Mellon University*, (2008).
2. C. Dellarocas, X. M. Zhang, N. F. Awad, Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive Marketing* **21**, 23 (2007).
3. F. Zhu, X. Zhang, Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of Marketing* **74**, 133 (2010).
4. X. Li, L. M. Hitt, Self-selection and information role of online product reviews. *Information Systems Research* **19**, 456 (2008).
5. N. Hu, P. A. Pavlou, J. Zhang, Can online reviews reveal a product's true quality?: empirical findings and analytical modeling of Online word-of-mouth communication. *Proceedings of the 7th ACM conference on Electronic commerce. ACM*, (2006).
6. W. Duan, B. Gu, A. B. Whinston, Do online reviews matter?—An empirical investigation of panel data. *Decision Support Systems* **45**, 1007 (2008).
7. B. Gu, M. Lin, The dynamics of online consumer reviews. *Workshop on Information Systems and Economics (WISE).* (2006).
8. M. Luca, Reviews, Reputation, and Revenue: The Case of Yelp.com. *Harvard Business School* NOM Unit Working Paper No. 12 (2011).
9. J. A. Chevalier, D. Mayzlin, The Effect of Word of Mouth on Sales: Online Book Reviews. *Journal of Marketing Research* **43**, 345 (2006).
10. A. Ghose, P. G. Ipeirotis, A. Sundararajan, The dimensions of reputation in online markets. *CeDER Working Papers* (2006).
11. N. Archak, A. Ghose, P. G. Ipeirotis, Deriving the Pricing Power of Product Features by Mining Consumer Reviews. *Management Science* **57**, 1485 (2011).
12. F. Wu, B. A. Huberman, How Public Opinion Forms. *Internet and Network Economics* **5385**, 334 (2008).
13. M. J. Salganik, D. J. Watts, Leading the Herd Astray: An Experimental Study of Self-fulfilling Prophecies in an Artificial Cultural Market. *Social Psychology Quarterly* **71**, 338 (2008).
14. N. E. Friedkin, E. C. Johnsen, *Social Influence Network Theory: A Sociological Examination of Small Group Dynamics*. (Cambridge University Press, 2011), vol. 33, pp. 390.
15. S. E. Asch, in *Groups, leadership and men,* H. Guetzkow, Ed. (Carnegie Press, Pittsburgh, PA, 1951), pp. 177-190.
16. A. T. Sorenson, Bestseller Lists and Product Variety. *The Journal of Industrial Economics* **55**, 715 (2007).
17. W. A. Hanson, D. S. Putler, Hits and Misses: Herd Behavior and Online Product Popularity. *Marketing Letters* **74**, 297 (1996).
18. P. Rozin, E. B. Royzman, Negativity Bias, Negativity Dominance, and Contagion. *Personality and Social Psychology Review* **5**, 296 (2001).
19. G. Peeters, J. Czapinski, Positive-Negative Asymmetry in Evaluations: The Distinction Between Affective and Informational Negativity Effects. *European Review of Social Psychology* **1**, 33 (1990).
20. M. Plummer, in *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*. (Vienna, 2003).
21. A. Gelman, D. B. Rubin, Inference from iterative simulation using multiple sequences. *Statistical science* **7**, 457 (1992).

22. F. J. Massey, The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association* **46**, 68 (1951).
23. K. Pearson, On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine* **5**, 157 (1900).
24. M. Jeng, Error in statistical tests of error in statistical tests. *BMC Medical Research Methodology* **6**, (2006).
25. J. Balthrop, S. Forrest, M. Newman, M. Williamson, Technological Networks and the Spread of Computer Viruses *Science* **304**, 527 (2004).
26. D. J. Watts, S. Strogatz, Collective dynamics of 'small-world' networks. *Nature* **393**, 440 (1998).
27. A. V. Banerjee, A Simple Model of Herd Behavior. *Quarterly Journal of Economics* **107**, 797 (1992).
28. S. Bikhchandani, D. Hirshleifer, I. Welch, A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades. *Journal of Political Economy* **100**, 992 (1992).
29. L. R. Anderson, C. A. Holt, Information Cascades in the Laboratory. *The American Economic Review* **87**, 847 (1997).
30. B. Çelen, S. Kariv, Distinguishing Informational Cascades from Herd Behavior in the Laboratory. *The American Economic Review* **94**, 484 (2004).
31. A. Tversky, D. Kahneman, Judgment under Uncertainty: Heuristics and Biases. *Science* **185**, 1124 (1974).
32. D. T. Campbell, J. Stanley, *Experimental and Quasi-Experimental Designs for Research* (Cengage Learning, ed. 1 edition 1963).

Supplementary Tables

Table S1. Web Site and Experiment Descriptive Statistics	
Experiment	
Experiment Start Date	07-Dec-2010
Experiment End Date	19-May-2011
Posts	
Number of Posts	10,755
Number of Upvotes	483,002
Number of Down-votes	20,608
Number of Post Categories	49
Comments	
Number of Comments	101,281
Mean Comment Size	378 bytes
Mean (Max) Number of Replies	0.57 (12)
Mean (Max) Tree Size	1.93 (38)
Mean (Max) Number of Upvotes per Comment	2.50 (89)
Mean (Max) Number of Downvotes per Comment	0.55 (69)
Page Views	
Total Number of Comment Impressions	10,361,257
Mean (Max) Impressions per Comment	102.30 (8138)
Mean (Max) Impressions per User	2,878.13 (292,901)
Users	
Total Registered Users	116,340
Number of Active Users During the Experiment	3,600
Published a Post	1,184
Voted for Post	3,443
Commented	2,125
Voted for a Comment	1,099

Table S2. Comment Descriptive Statistics by Treatment				
	Control	Positively Treated	Negatively Treated	All
	(n=95,290) 94.08%	(n=4,049) 4.00%	(n=1,942) 1.92%	(n=101,281) 100%
Number of Page Views (% of all Page Views)	9743770 (94.0%)	413634 (3.99%)	203586 (1.96%)	10361257 (100%)
Topic frequency, χ^2		0.17	0.08	1.00
Mean text size in bytes (ANOVA vs. Control)	378	360 (0.12)	392 (0.41)	378 (0.88)
K-S Test		0.71	0.99	1.00
Mean unique viewers per comment (ANOVA vs. Control)	55.2	55.1 (0.96)	57.9 (0.02)	55.2 (0.89)
K-S Test		0.96	0.23	1.00

Table S3 Frequency of posts and comments by category.

	Posts	Comments
	Number(%)	Number(%)
Culture and Society	1924 (17.9%)	20947 (20.7%)
General News	1894 (17.6%)	15637 (15.4%)
Fun	1751 (16.3%)	14212 (14.0%)
Politics	1412 (13.1%)	14001 (13.8%)
IT	804 (7.5%)	7312 (7.2%)
Business	672 (6.2%)	6028 (6.0%)
Economics	648 (6.0%)	5313 (6.2%)
Science	588 (5.5%)	4916 (4.9%)
Automobile	299 (2.8%)	2786 (2.8%)
Entertainment	279 (2.6%)	1688 (1.7%)
Sports	219 (2.0%)	1217 (1.2%)
Humor	200 (1.9%)	1552(1.5%)
Portal	55 (0.5%)	5624 (5.6%)

Note: Comments are assigned to the category of the post on which they comment.

Table S4. Discourse Descriptive Statistics by Treatment				
	Control	Positive Treatment	Negative Treatment	All
	(n=95,290) 94.08%	(n=4,049) 4.00%	(n=1,942) 1.92%	(n=101,281) 100%
Mean comment score (ANOVA vs. Control)	1.93	2.42 (2e-11)	2.00 0.54	1.95 0.33
K-S test		1e-23	1e-63	0.02
Mean number of votes (ANOVA vs. Control)	3.02	3.45 (2e-8)	3.49 (2e-5)	3.05 (0.24)
K-S test		7e-29	4e-78	0.007
Mean upvotes per comment (ANOVA vs. Control)	2.48	2.94 (3e-11)	2.75 (6e-3)	2.50 (0.24)
K-S test		1e-17	1e-70	0.03
Mean downvotes per comment (ANOVA vs. Control)	0.54	0.52 (0.40)	0.75 (8e-7)	0.55 (0.73)
K-S test		0.03	8e-7	0.99
Mean number of child comments (ANOVA vs. Control)	0.58	0.57 (0.80)	0.57 (0.55)	0.57 (0.88)
K-S test		1.00	1.00	1.00
Mean tree size (ANOVA vs. Control)	1.34	1.35 (0.76)	1.31 (0.57)	1.34 (0.96)
K-S test		1.00	0.73	1.00

Table S5. Social Network Descriptive Statistics.				
Network Type	**Number of users**	**Average degree**	**Reciprocity**	**Average Clustering Coefficient**
Friends	9068	6.04	0.41	0.11
Enemies	38961	4.96	0.11	0.00

Table S6: Interference Effect of Proximate Treatment on First Viewer's Turnout and Voting Behavior

	Turnout	Turnout	Upvote	Upvote	Downvote	Downvote
Upvoted	0.263*** (0.059)	0.263*** (0.059)	0.290*** (0.062)	0.289*** (0.062)	0.022 (0.184)	0.025 (0.184)
Downvoted	0.657*** (0.072)	0.657*** (0.072)	0.648*** (0.077)	0.647*** (0.077)	0.623 (0.197)	0.627** (0.197)
Upvoted Interference		-0.009 (0.026)		0.007 (0.027)		-0.116 (0.072)
Downvoted Interference		0.025 (0.027)		0.039 (0.029)		-0.077 (0.079)
N	103019	102999	103019	102999	103019	102999

Notes: Logistic regression estimated with maximum likelihood. *, **, and *** denote significance at the 10%, 5%, and 1% levels respectively.

Table S7: Interference Effect of Proximate Treatment on Comment Total Votes and Proportion of Positive Votes

	Total Ratings	Total Ratings	Proportion Positive	Proportion Positive
Upvoted	0.435*** (0.077)	0.435*** (0.077)	0.058*** (0.005)	0.058*** (0.005)
Downvoted	0.484*** (0.110)	0.483*** (0.110)	0.106*** (0.008)	0.106*** (0.008)
Upvoted Interference		-0.012 (0.030)		0.0039 (0.002)
Downvoted Interference		0.073* (0.032)		0.0052* (0.002)
N	103019	102999	103019	102999

Notes: Linear regression estimated with OLS. *, **, and *** denote significance at the 10%, 5%, and 1% levels respectively.

Table S8. Gelman-Rubin Statistics

Model		Treatment		Random Effects Hyperparameters			Multivariate Potential Scale Reduction Factor
	β_0	$\beta_{up\ voted}$	$\beta_{down\ voted}$	τ_{author}	τ_{viewer}	τ_{pair}	
Figure 1a							
Upvote	1.0105(1.0246)	1.0005(1.0017)	1.0002(1.0006)	1.0236(1.057)	1.0067(1.0169)	1.0286(1.0755)	1.045527
Downvote	1.0403(1.1051)	1.0005(1.0016)	1.0003(1.0009)	1.0143(1.037)	1.0046(1.011)	1.1155(1.2914)	1.11115
Figure 1b							
Comment Score	1.0198(1.0477)	1.0002(1.001)	0.9999(1.0001)	1.0059(1.0153)			1.01702
Figure 2a							
Business	1.0003(1.0009)	1(1.0001)	1(1.0001)	1.0033(1.0073)			1.002379
Culture & Society	1.0009(1.0023)	1.0001(1.0004)	1(1)	1.0006(1.0014)			1.000751
Politics	1.0007(1.0018)	1.0002(1.0006)	1.0001(1.0005)	1.0022(1.0058)			1.00226
IT	1.0007(1.0017)	1.0002(1.0005)	1(1.0001)	1.0038(1.0097)			1.003487
Fun	1.001(1.0025)	1.0001(1.0003)	1(1.0002)	1.0027(1.0041)			1.001517
Economics	1.0004(1.0008)	1(1.0002)	1.0004(1.0011)	1.0017(1.0032)			1.000845
General News	1.0012(1.0024)	0.9999(1)	1(1.0002)	1.0014(1.0029)			1.000868
Figure 2b							
Upvote by friends	1.0143(1.0385)	1.0001(1.0001)	1(1.0001)	1.0251(1.0523)	1.0041(1.0103)	1.0118(1.0257)	1.024868
Figure 2c							
Upvote by “enemies”	1.0196(1.0517)	1.0006(1.0017)	1.0008(1.002)	1.337(2.4716)	1.0038(1.008)	1.7783(4.4137)	1.609671
Figure 3a							
Number of responses	1.2031(1.2688)	1.2504(1.3977)	1.1899(1.4059)	1.0021(1.0053)			1.032432
Figure 3b							
Comment Tree Height	1.1699(1.3274)	1.3259(2.4905)	1.3543(3.1237)	1.0023(1.0063)			1.097036

Table S9: Fisher's Exact Test of Turnout and Positivity Effects of Upvote & Downvote Treatments

Subgroup	N	Upvote Effect on Turnout	Downvote Effect on Turnout	Upvote Effect on Positivity	Downvote Effect on Positivity
Negative Raters	38,930	↑***	↑***	—	↑**
Positive Raters	50,070	↑***	↑***	—	↓**
Low Quality Commenter	43,186	↑***	↑***	—	—
High Quality Commenter	42,686	—	↑**	—	—
Infrequent Rating Interactions	75,628	↑***	↑***	↑**	—
Frequent Rating Interactions	27,391	—	↑***	—	—
Enemies	5,939	—	—	—	↓*
Friends	31,665	↑***	↑***	—	—

Notes: ↑ and ↑ denote odds ratios greater than and less than 1, respectively. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table S10: Contrasts of Subgroup Responses to Upvote and Down Treatments

Contrast	Control Turnout	Upvote on Turnout	Downvote on Turnout	Control Positivity	Upvote on Positivity	Downvote on Positivity
Negative vs Positive Raters	↓***	—	—	↓***	—	↑**
Low vs High Quality Commenters	↓***	—	—	↓***	—	—
Infrequent vs Frequent Rating Interactions	↓***	↑*	—	↓***	↑*	↓*
Enemies vs Friends	↓**	—	—	↓***	—	—

Notes: ↑ and ↑ denote odds ratios greater than and less than 1, respectively. *, **, and *** denote significance at the 10%, 5%, and 1% levels, respectively.

Table S11: Turnout by Rater Positivity				
	Negative	Positive	odds ratio	p(odds ratio = 1)
control	0.065	0.072	0.893	0.000
up-treated	0.081	0.094	0.856	0.211
down-treated	0.129	0.123	1.057	0.715
odds up-treated / control	1.273	1.329	0.958	0.731
p(u/c = 1)	0.013	0.000		
odds down-treated / control	2.136	1.804	1.184	0.257
p(d/c = 1)	0.000	0.000		

Table S12: Positivity by Rater Positivity				
	Negative	Positive	odds ratio	p(odds ratio = 1)
control	0.790	0.950	0.196	0.000
up-treated	0.808	0.974	0.115	0.000
down-treated	0.880	0.887	0.935	1.000
odds up-treated / control	1.121	1.931	0.580	0.291
p(u/c = 1)	0.735	0.166		
odds down-treated / control	1.953	0.410	4.765	0.000
p(d/c = 1)	0.031	0.008		

Table S13: Turnout by Commenter Quality				
	Low Quality	High Quality	odds ratio	p(odds ratio = 1)
control	0.057	0.070	0.800	0.000
up-treated	0.077	0.079	0.964	0.799
down-treated	0.108	0.123	0.863	0.352
odds up-treated / control	1.372	1.140	1.204	0.155
p(u/c = 1)	0.001	0.154		
odds down-treated / control	1.999	1.854	1.078	0.634
p(d/c = 1)	0.000	0.000		

Table S14: Positivity by Commenter Quality				
	Low Quality	High Quality	odds ratio	p(odds ratio = 1)
control	0.845	0.958	0.238	0.000
up-treated	0.883	0.979	0.166	0.002
down-treated	0.874	0.970	0.215	0.023
odds up-treated / control	1.378	1.985	0.694	0.577
p(u/c = 1)	0.312	0.376		
odds down-treated / control	1.264	1.406	0.899	0.876
p(d/c = 1)	0.547	0.798		

Table S15: Turnout by Rating Interaction Frequency				
	Infrequent	Frequent	odds ratio	p(odds ratio = 1)
control	0.044	0.112	0.364	0.000
up-treated	0.063	0.128	0.456	0.000
down-treated	0.081	0.200	0.355	0.000
odds up-treated / control	1.461	1.166	1.253	0.065
p(u/c = 1)	0.000	0.108		
odds down-treated / control	1.935	1.987	0.974	0.858
p(d/c = 1)	0.000	0.000		

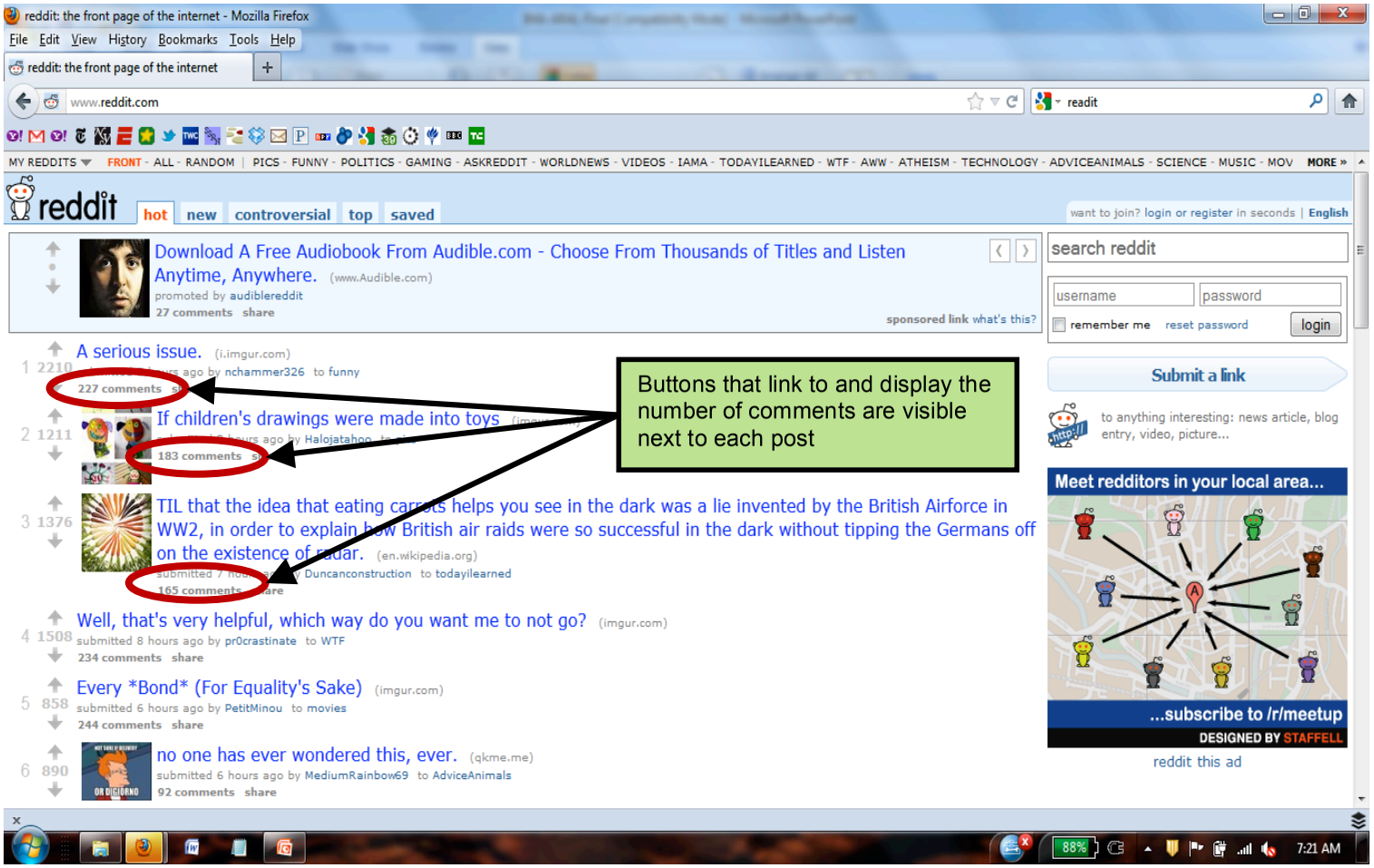
Table S16: Positivity by Rating Interaction Frequency				
	Infrequent	Frequent	odds ratio	p(odds ratio = 1)
control	0.850	0.914	0.532	0.000
up-treated	0.907	0.902	1.062	1.000
down-treated	0.814	0.952	0.220	0.002
odds up-treated / control	1.716	0.860	1.995	0.078
p(u/c = 1)	0.028	0.633		
odds down-treated / control	0.771	1.879	0.410	0.088
p(d/c = 1)	0.294	0.210		

Table S17: Turnout by Articulated Relationship				
	Enemies	Friends	odds ratio	p(odds ratio = 1)
control	0.084	0.094	0.890	0.025
up-treated	0.086	0.113	0.739	0.261
down-treated	0.103	0.181	0.520	0.049
odds up-treated / control	1.025	1.234	0.831	0.459
p(u/c = 1)	0.906	0.024		
odds down-treated / control	1.248	2.134	0.585	0.115
p(d/c = 1)	0.481	0.000		

Table S18: Positivity by Articulated Relationship				
	Enemies	Friends	odds ratio	p(odds ratio = 1)
control	0.319	0.980	0.010	0.000
up-treated	0.143	0.986	0.003	0.000
down-treated	0.364	0.971	0.019	0.000
odds up-treated / control	0.356	1.454	0.245	0.143
p(u/c = 1)	0.097	1.000		
odds down-treated / control	1.219	0.690	1.767	0.515
p(d/c = 1)	0.751	0.469		

Supplementary Figures

Figures S1. a. and **b.** show screenshots from Reddit.com, a English-language social news aggregation website similar to ours. Users post URLs to content, which are then browsed by other users on the main page (Panel a.). Clicking on the comment link (circled) takes users to a discussion page (Panel b.), where they can read or contribute comments about the URL. Users may use up or down arrows (circled) to upvote or downvote the comment once. The current score is displayed next to each comment (also circled).



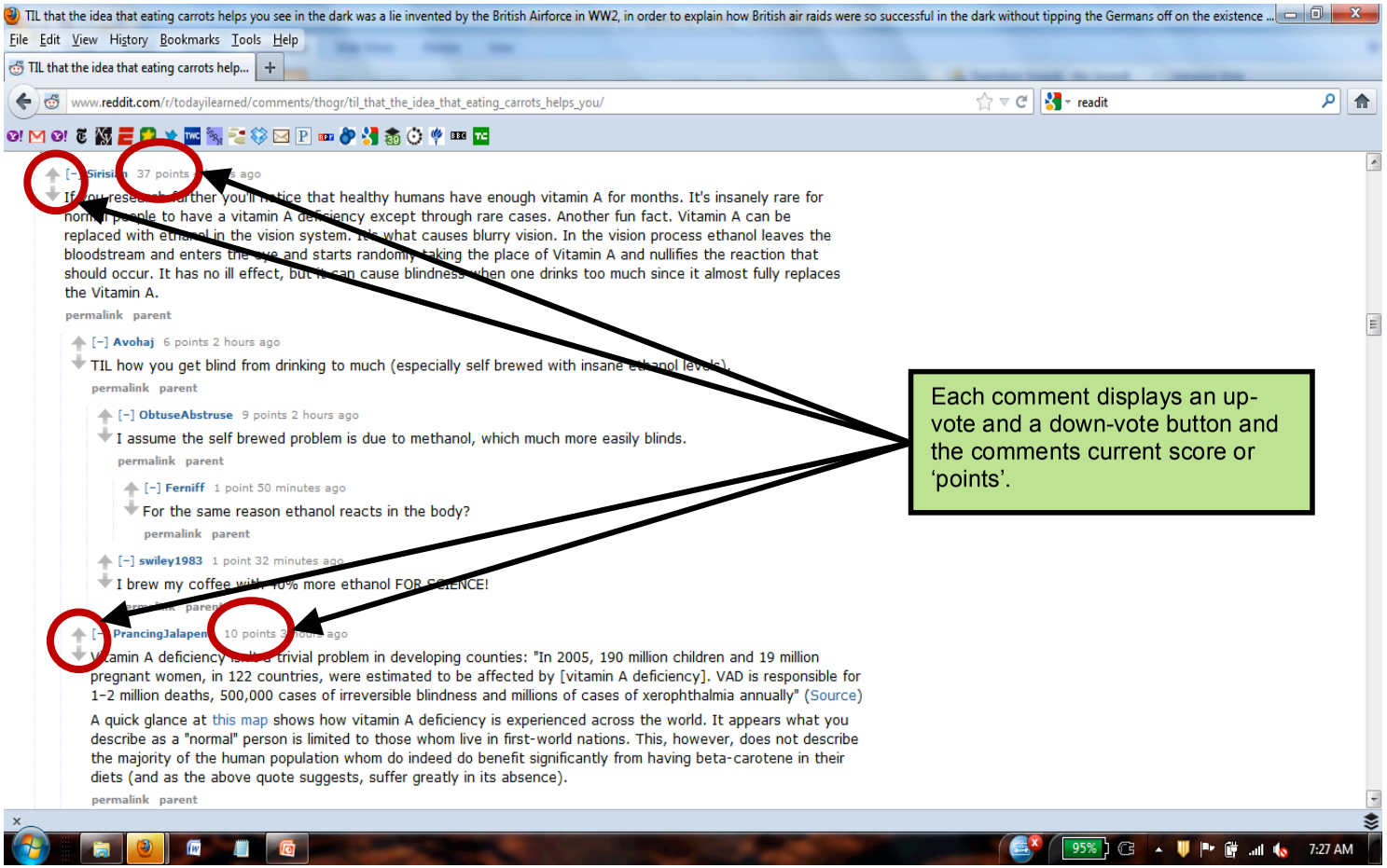

TIL that the idea that eating carrots helps you see in the dark was a lie invented by the British Airforce in WW2, in order to explain how British air raids were so successful in the dark without tipping the Germans off on the existence ...
File Edit View History Bookmarks Tools Help
TIL that the idea that eating carrots help...
www.reddit.com/r/todayilearned/comments/thogr/til_that_the_idea_that_eating_carrots_helps_you/
readit
[–] Sirisian 37 points ago
If you research further you'll notice that healthy humans have enough vitamin A for months. It's insanely rare for normal people to have a vitamin A deficiency except through rare cases. Another fun fact. Vitamin A can be replaced with ethanol in the vision system. It's what causes blurry vision. In the vision process ethanol leaves the bloodstream and enters the eye and starts randomly taking the place of Vitamin A and nullifies the reaction that should occur. It has no ill effect, but it can cause blindness when one drinks too much since it almost fully replaces the Vitamin A.
permalink parent
[–] Avohaj 6 points 2 hours ago
TIL how you get blind from drinking to much (especially self brewed with insane ethanol levels).
permalink parent
[–] ObtuseAbstruse 9 points 2 hours ago
I assume the self brewed problem is due to methanol, which much more easily blinds.
permalink parent
[–] Ferniff 1 point 50 minutes ago
For the same reason ethanol reacts in the body?
permalink parent
[–] swiley1983 1 point 32 minutes ago
I brew my coffee with 10% more ethanol FOR SCIENCE!
[–] PrancingJalapeno 10 points ago
Vitamin A deficiency is not a trivial problem in developing counties: "In 2005, 190 million children and 19 million pregnant women, in 122 countries, were estimated to be affected by [vitamin A deficiency]. VAD is responsible for 1–2 million deaths, 500,000 cases of irreversible blindness and millions of cases of xerophthalmia annually" (Source)
A quick glance at this map shows how vitamin A deficiency is experienced across the world. It appears what you describe as a "normal" person is limited to those whom live in first-world nations. This, however, does not describe the majority of the human population whom do indeed do benefit significantly from having beta-carotene in their diets (and as the above quote suggests, suffer greatly in its absence).
permalink parent
95%
7:27 AM
Each comment displays an up-vote and a down-vote button and the comments current score or 'points'.

Figure S2 displays the frequency of posts (green) and comments (blue) by topic during the period over which the experiment was conducted.

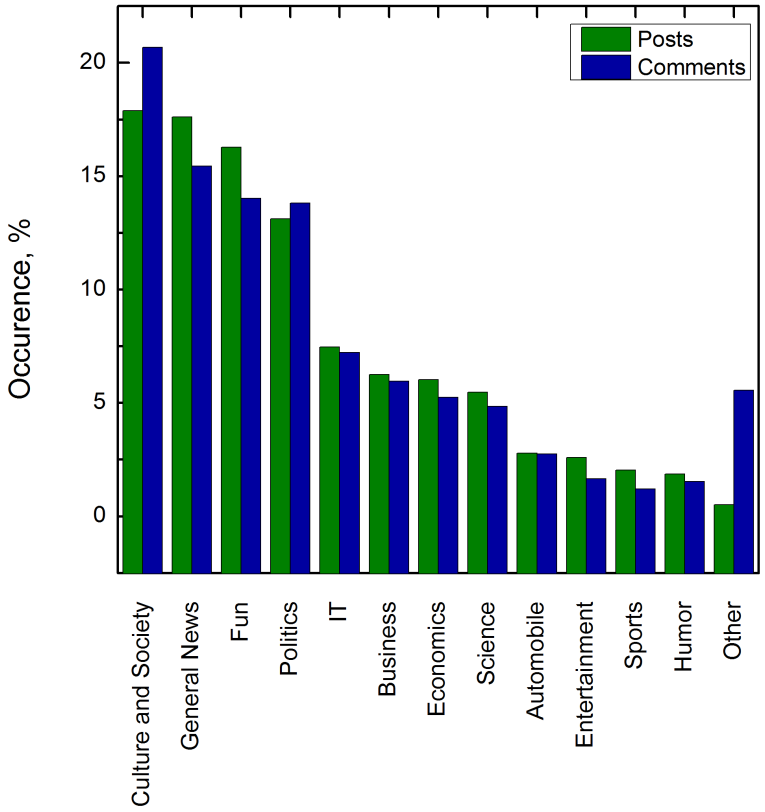


Figure S3 displays the largest strongly connected component of the social network composed of the 8000 most reputable users. The component contains 762 users linked with 5374 directed edges in which each tie is mutually exclusively displayed as either a friend (blue) or enemy (enemy) relationship. There are 3373 friend relationships and 2001 enemy relationships in the graph. The nodes are sized in proportion to their in-degree across both types of relationships.

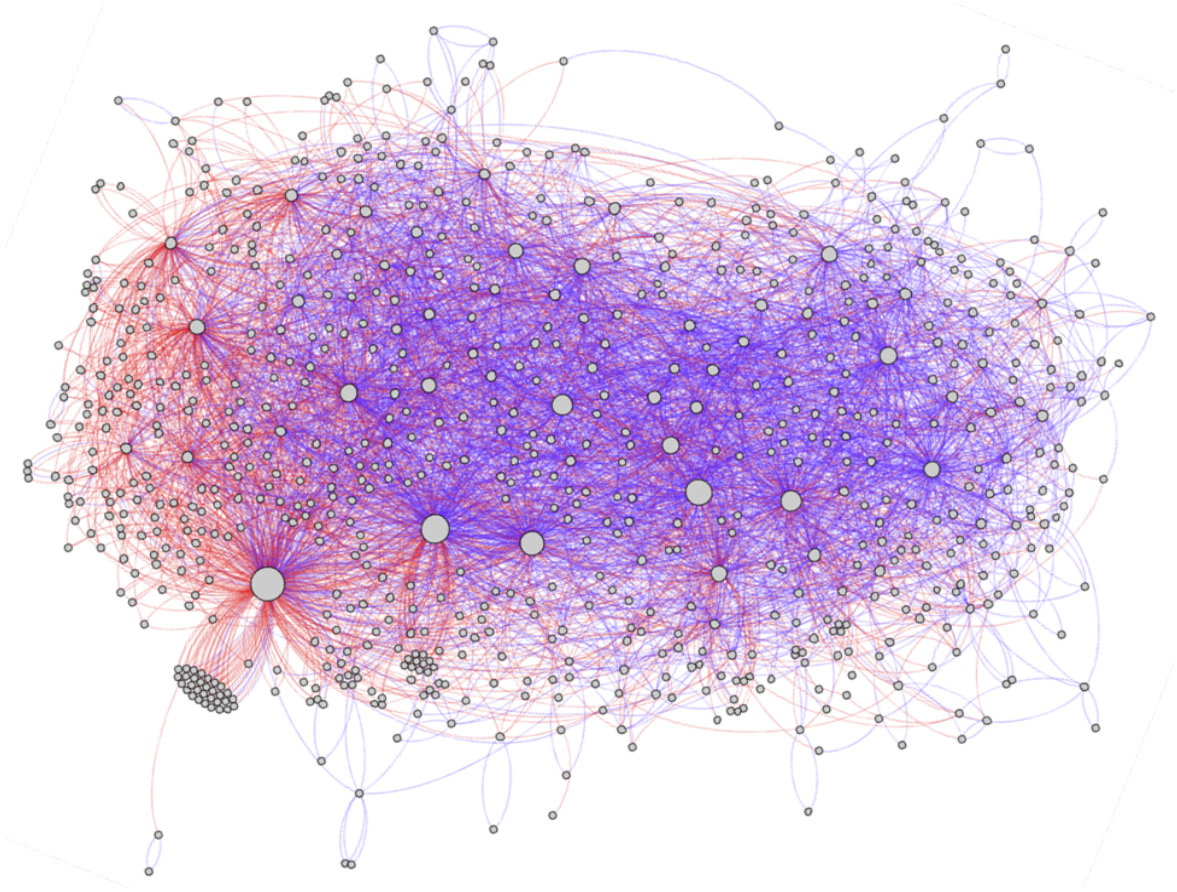


Figure S4 a. displays the treatment effect estimates and 95% confidence intervals for conditional and marginal models. Conditional models include random effects components and provide more conservative intervals. The estimates of treatment effects are identical. **S4 b.** displays a histogram of the area under the ROC curve (AUC) for simulated conditional models of upvoting (left panel) and downvoting (right panel) draw from posterior parameter distributions.

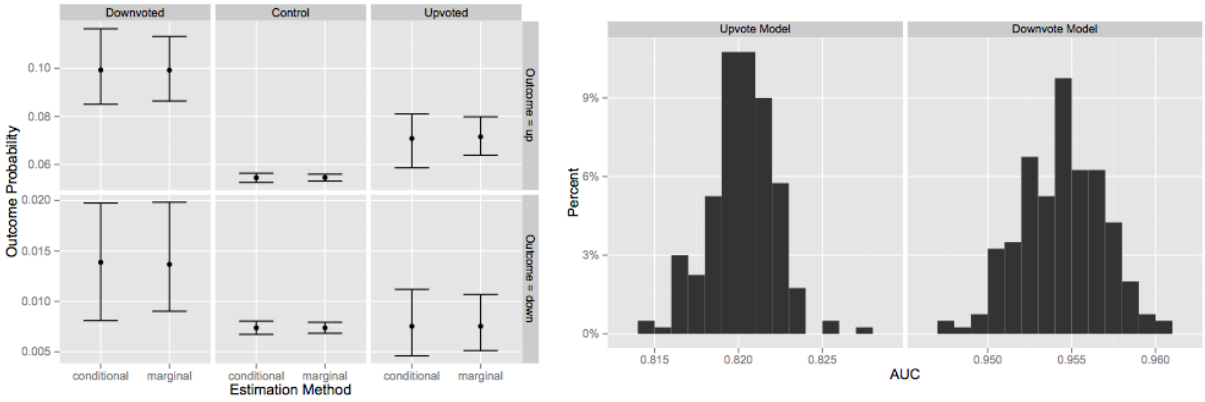


Figure S5 displays a) the difference of positive to negative votes (this simply reorders the results from Fig 1.b. in the main text for ease of comparison) and b) the ratio of positive to negative votes under each treatment.

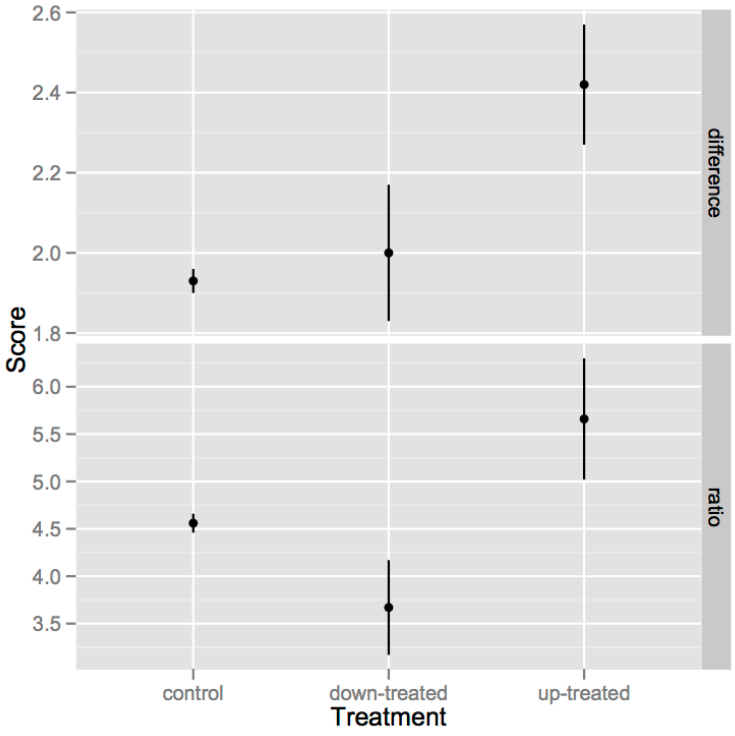


Figure S6 displays positivity and turnout estimates by rater positivity. Dashed lines are overall sample means. Positive raters are more positive as first viewers of control and up-treated comments (top-middle and top-right), but their positivity is indistinguishable from negative raters for down-treated comments (top-left). Turnout is roughly the same across this subgroup dimension except on control comments.

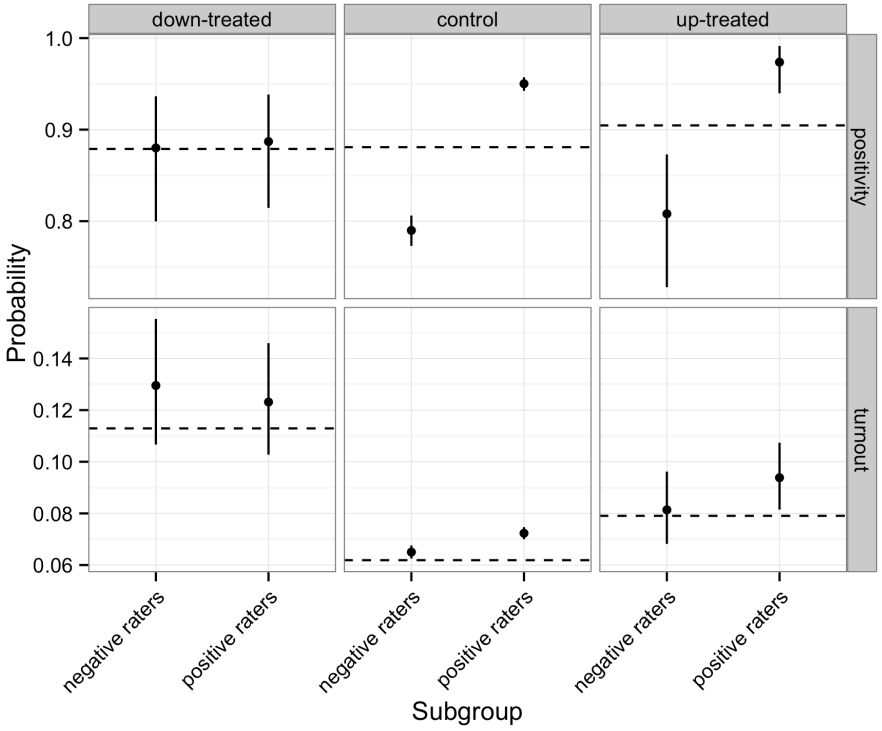


Figure S7 displays treatment effects on positivity and turnout by rater positivity (left and middle dot-plots in each pane) and the ratio of treatment effects (rightmost dot-plot in each pane). Dashed lines are the null hypotheses of odds ratios equal to one and red dot-plots indicate significance at the 95% confidence level. The positivity odds ratio scale is truncated at 5 to aid comparison. Both subgroups respond to both treatments with increased turnout, but the difference in treatment effects is not significant (bottom row). Negative (positive) raters use a significantly higher proportion of positive (negative) ratings on down-treated comments (top-left). Changes in positivity from the up-treatment are not statistically significant for either subgroup.

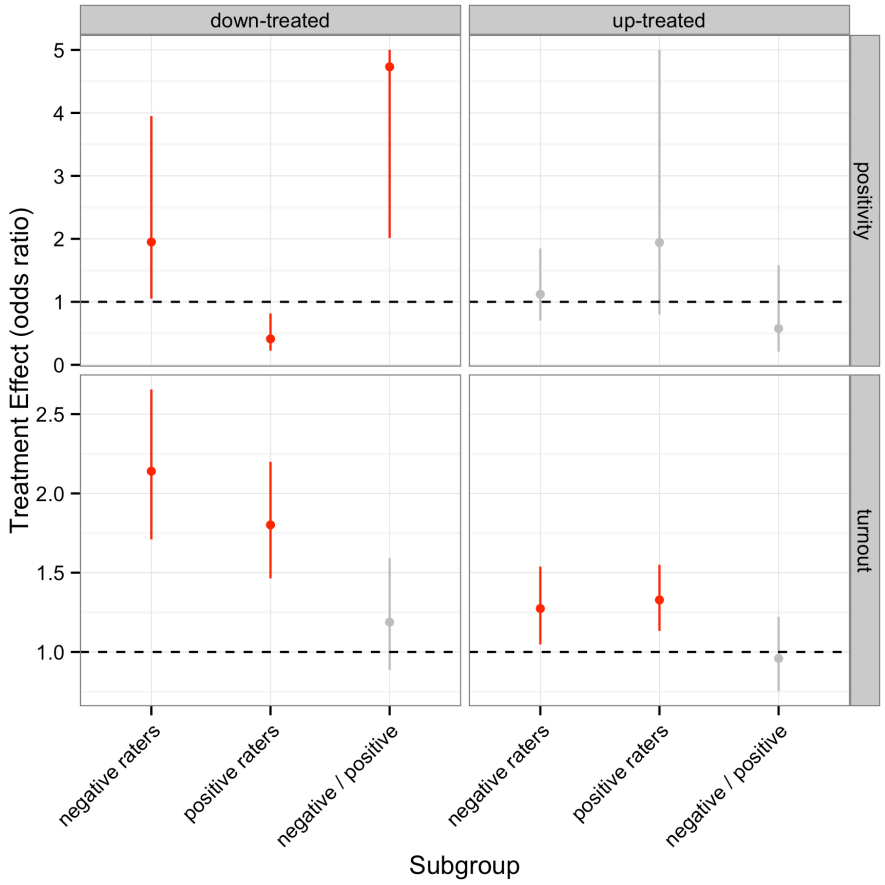


Figure S8 displays positivity and turnout estimates by commenter quality. Dashed lines are overall sample means. High quality commenters receive more a significantly greater proportion of positive ratings across all three treatment groups (top row). Turnout is equally likely for both subgroups except for on control comments (bottom-middle).

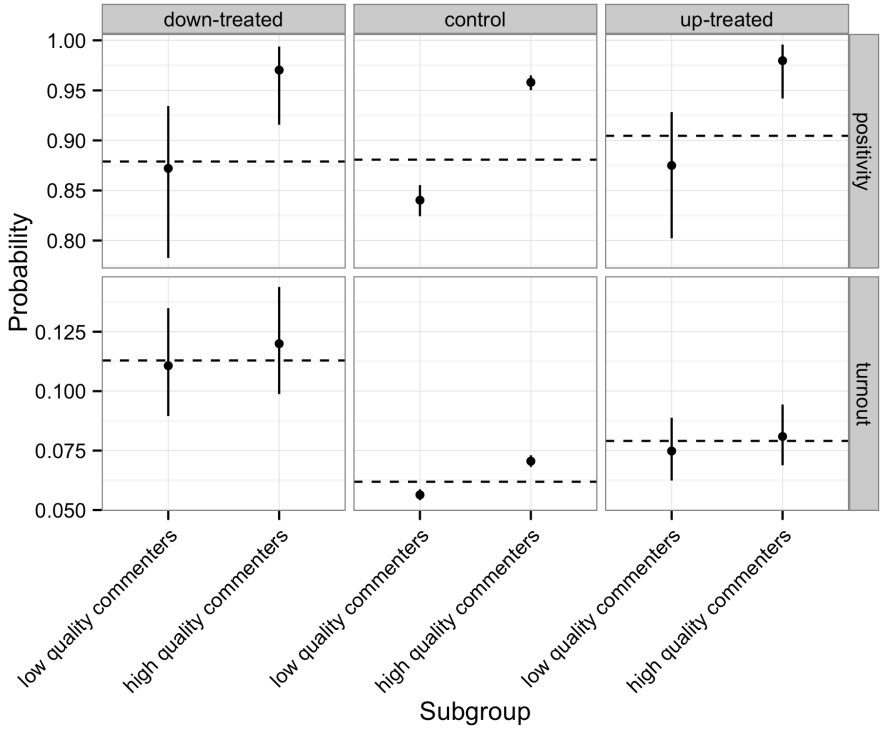


Figure S9 displays treatment effects on positivity and turnout by commenter quality (left and middle dot-plots in each pane) and the ratio of treatment effects (rightmost dot-plot in each pane). Dashed lines are the null hypotheses of odds ratios equal to one and red dot-plots indicate significance at the 95% confidence level. Neither subgroup experiences significant treatment effects on positivity (top row). Turnout is increased for both subgroups under down-treatment and under up-treatment for comments from low quality commenters (bottom row). Differences in treatment effects on turnout are not significant.

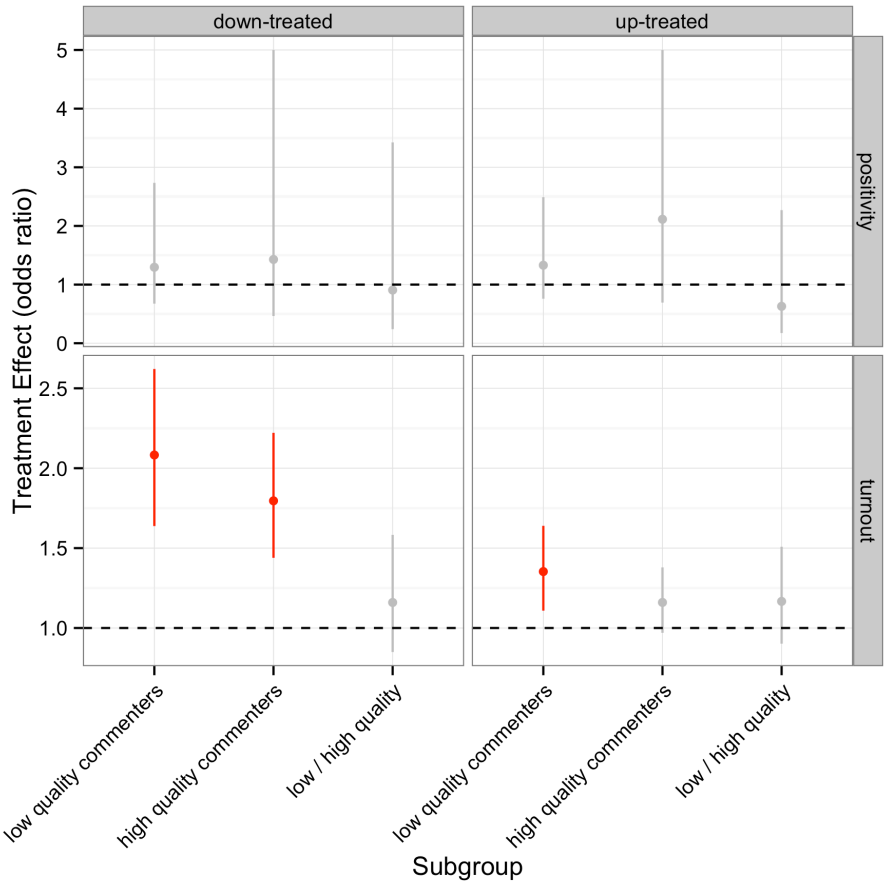


Figure S10 displays positivity and turnout estimates by rating interaction frequency. Dashed lines are overall sample means. Pairs with frequent interactions are more likely to turnout, regardless of treatment status (bottom row). Pairs with frequent interactions have a higher proportion of positive ratings under control and down-treatment (top-left and top-middle), but have identical proportions under up-treatment (top-right).

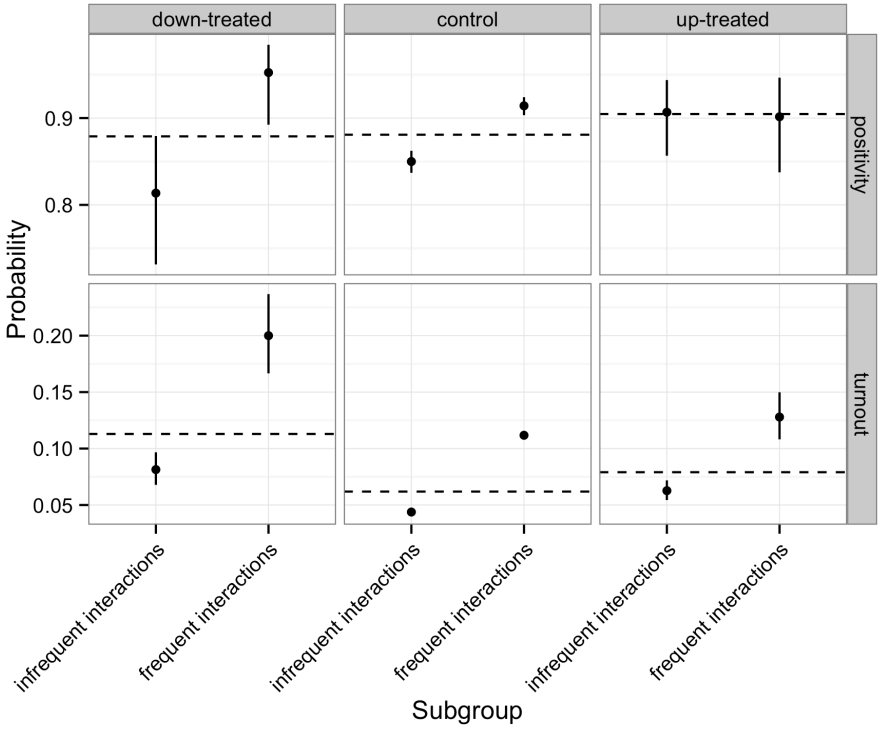


Figure S11 displays treatment effects on positivity and turnout by rating interaction frequency (left and middle dot-plots in each pane) and the ratio of treatment effects (rightmost dot-plot in each pane). Dashed lines are the null hypotheses of odds ratios equal to one and red dot-plots indicate significance at the 95% confidence level. The infrequent interactions subgroup receives a greater proportion of positive ratings under up-treatment (top-right), indicating significant opinion change for this group. The different in treatment effects across subgroups is significant at the 7% level but not the 5% level. Turnout is increased for both subgroups under down-treatment and under up-treatment for pairs with infrequent rating interactions (bottom row). Differences in treatment effects on turnout are not significant.

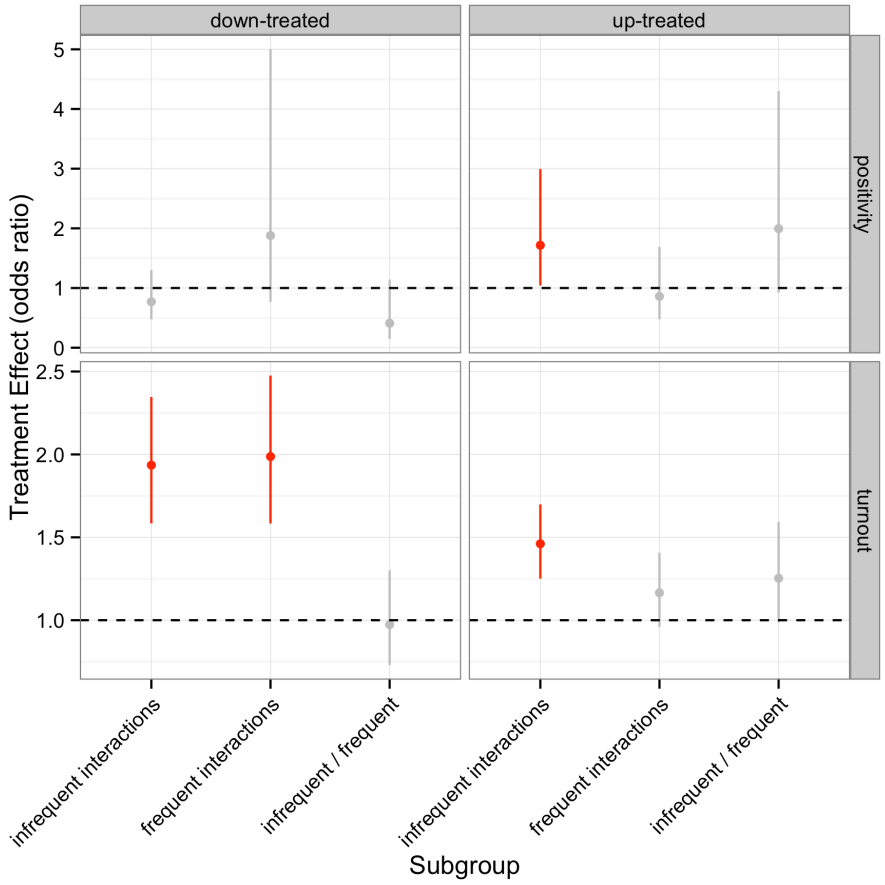


Figure S12 displays positivity and turnout estimates by articulated relationship frequency. Dashed lines are overall sample means. Friends are more positive conditional on turnout than the sample mean and enemies for all three treatment groups (top row). Friends also have a higher probability of turnout on down-treated and control comments ($p < 0.05$), but not on up-treated comments.

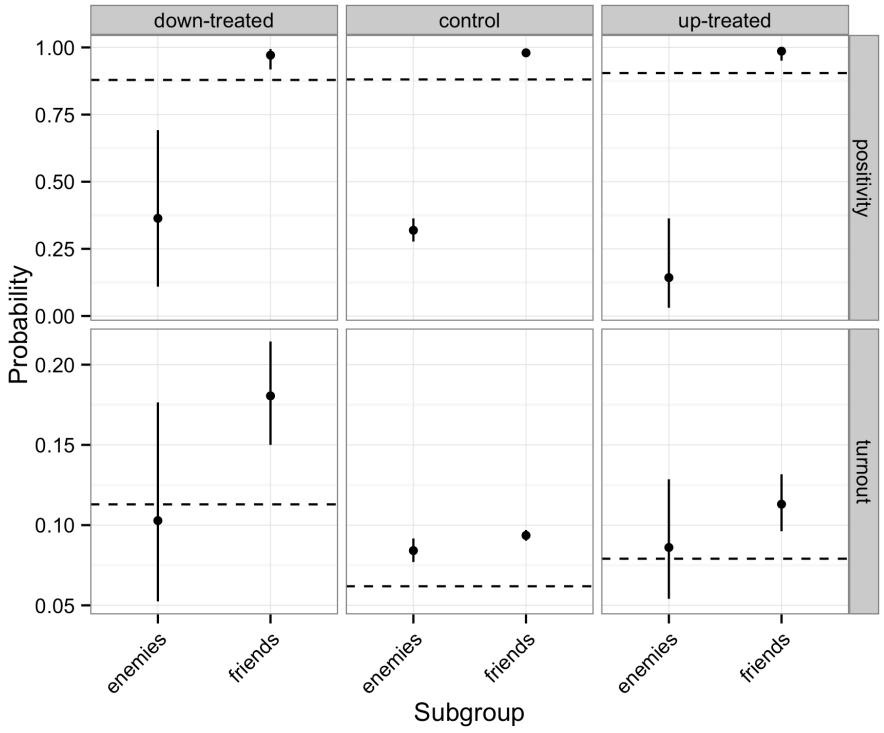


Figure S13 displays treatment effects on positivity and turnout by articulated relationship (left and middle dot-plots in each pane) and the ratio of treatment effects (rightmost dot-plot in each pane). Dashed lines are the null hypotheses of odds ratios equal to one and red dot-plots indicate significance at the 95% confidence level. Friends are significantly more likely to turnout under either treatment, but not significantly differently from enemies (bottom row). We cannot reject the null hypothesis that neither subgroup displays significant treatment effects on positivity, nor that their treatment effects differ (top row). The estimates in these plots are less precise than the others because of both subgroups are quite small.

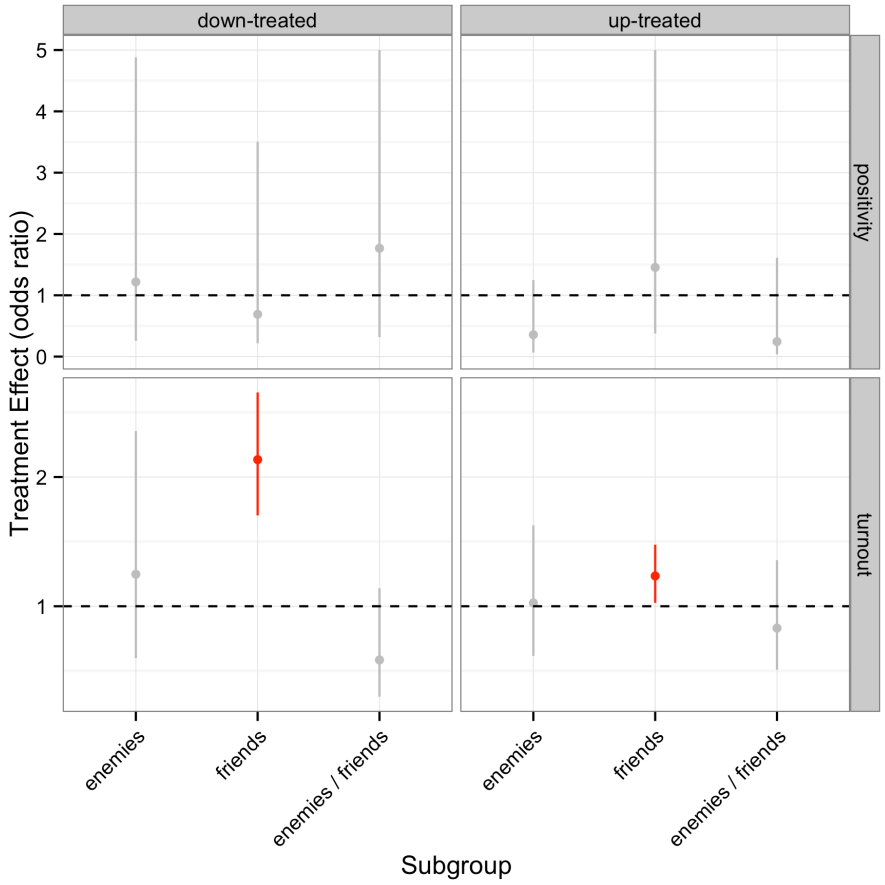


Figure S14 displays positivity and turnout estimates by topic. Dashed lines are overall sample means. Topics display some heterogeneity with respect to turnout and positivity, but few of the differences are statistically significant.

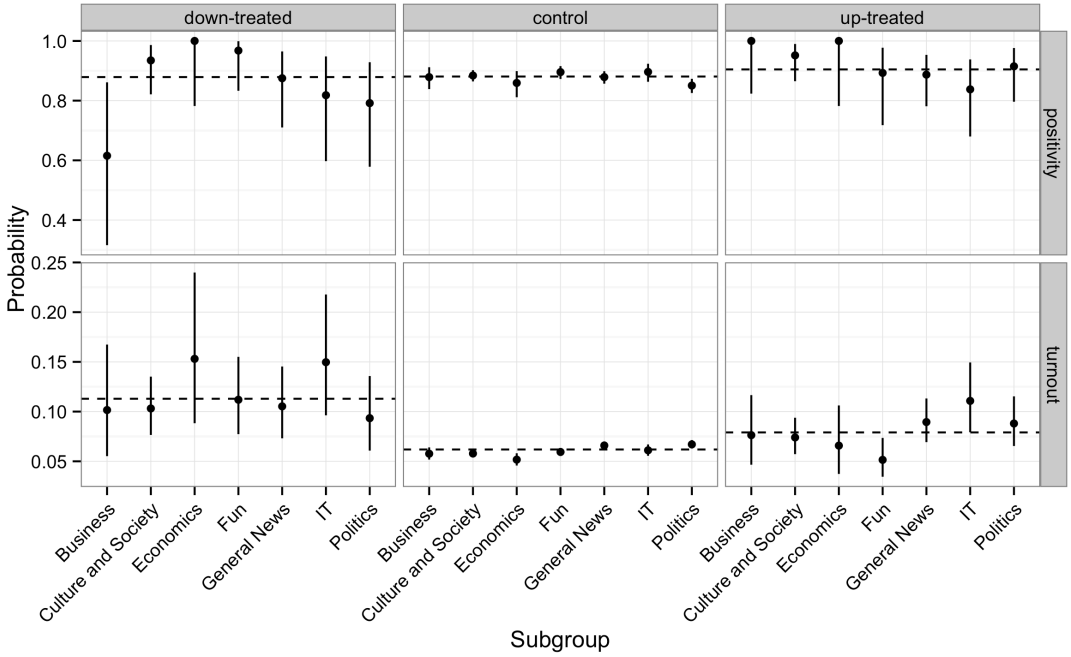


Figure S15 displays treatment effects on positivity and turnout by topic (left and middle dot-plots in each pane) and the ratio of treatment effects (rightmost dot-plot in each pane). Dashed lines are the null hypotheses of odds ratios equal to one and red dot-plots indicate significance at the 95% confidence level. Five out of seven topics display statistically significant increases in turnout for down-treated comments (bottom-left) and likewise for two of the topics in the up-treated condition (bottom-right). Treatment effects on positivity are imprecisely estimated due to small subsamples, with only Business displaying statistically significant evidence of a negative in positivity for down-treated comments (top-right).

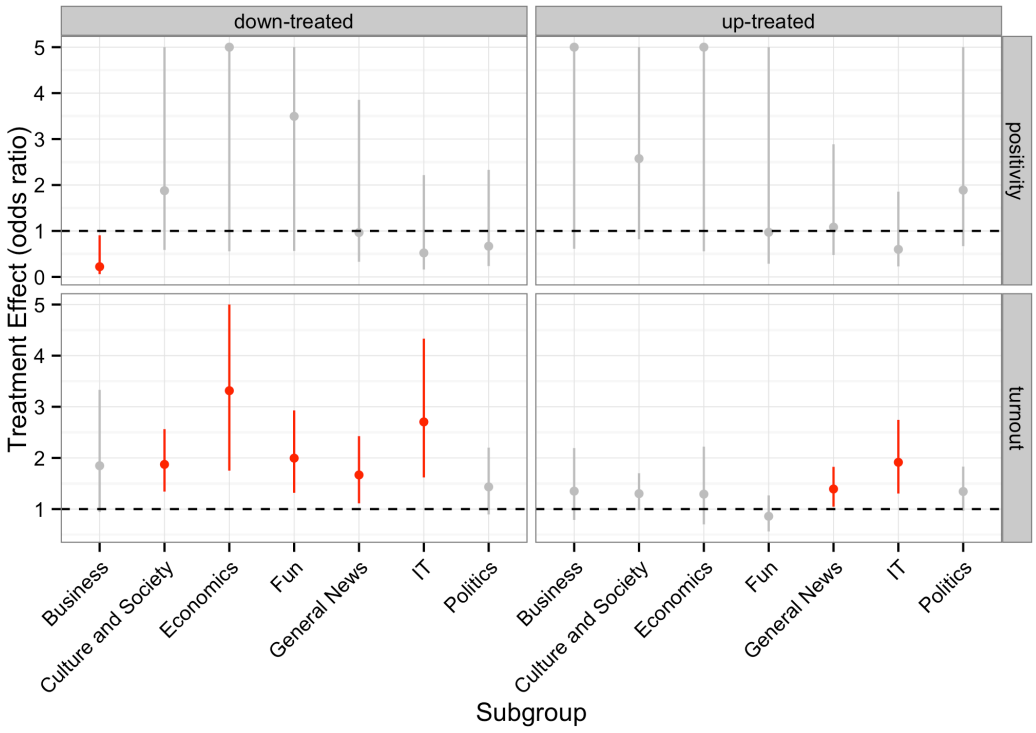


Figure S16 displays the mean final scores of negatively manipulated and control group comments with 95% confidence intervals inferred from Bayesian linear regression of the final comment score with commenter random effects across the seven most active topic categories on the site, ordered by the magnitude of the difference between the mean final score of positively manipulated comments and the mean final score of control comments in each category: Business, Culture and Society, Politics, IT, Fun, Economics, and General News. We did not include these results in the main text of the paper as none of them are statistically significant.