Threading is Sticky: How Threaded Conversations Promote Comment System User Retention

CEREN BUDAK, University of Michigan
R. KELLY GARRETT, Ohio State University
PAUL RESNICK, University of Michigan
JULIA KAMIN, University of Michigan

The Guardian—the fifth most widely read online newspaper in the world as of 2014—changed conversations on its commenting platform by altering its design from non-threaded to single-level threaded in 2012. We studied this naturally occurring experiment to investigate the impact of conversation threading on user retention as mediated by several potential changes in conversation structure and style. Our analysis shows that the design change made new users significantly more likely to comment a second time, and that this increased stickiness is due in part to a higher fraction of comments receiving responses after the design change. In mediation analysis, other anticipated mechanisms such as reciprocal exchanges and comment civility did not help to explain users' decision to return to the commenting system; indeed, civility did not increase after the design change and reciprocity declined. These analyses show that even simple design choices can have a significant impact on news forums' stickiness. Further, they suggest that this influence is more powerfully shaped by affordances—the new system made responding easier—than by changes in users' attention to social norms of reciprocity or civility. This has an array of implications for designers.

CCS Concepts: • **Human-centered computing** → *Empirical studies in HCI*;

Additional Key Words and Phrases: commenting systems; interrupted time series design; mediation analysis; design principles; stickiness

ACM Reference format:
Ceren Budak, R. Kelly Garrett, Paul Resnick, and Julia Kamin. 2017. Threading is Sticky: How Threaded Conversations Promote Comment System User Retention. *Proc. ACM Hum.-Comput. Interact.* 1, 2, Article 27 (November 2017), 20 pages.
https://doi.org/10.1145/3134662

1 INTRODUCTION

News organizations have long struggled with how to create online commenting systems that promote thoughtful engagement with news content while avoiding the vitriol for which online comments are best known [18]. They have financial motivations: using human editors to hunt down and remove hate speech and other unacceptable posts is costly, while failing to do so leads

This work is supported by the National Science Foundation, under grant IIS-1717688 and grant IIS-1617820.
Authors' addresses: Ceren Budak, University of Michigan School of Information, 105 S. State Street Ann Arbor, MI 48109; R. Kelly Garrett, Ohio State University, 154 N. Ovall Mall, 3016 Derby Hall, Columbus, OH 43210; Paul Resnick, University of Michigan School of Information, 105 S. State St. Ann Arbor, MI 48109; Julia Kamin, Department of Political Science, University of Michigan, 505 South State Street Ann Arbor, MI 48109.

2573-0142/2017/11-ART27 $15.00
https://doi.org/10.1145/3134662

to an environment that can drive news consumers away [37, 51]. News organizations also have motivations that are more closely related to their core mission: comments can influence how readers interpret the associated news article [1] and how much they trust the news site [36]. Furthermore, issues related to comment quality affect how journalists perform their work [21]. Solutions for improving news commenting systems are not trivial. Some attempts include abusive-content flagging by community members, moderation, and newsroom engagement [21, 44]. Faced with the pitfalls and costs of comment systems, some news organizations have simply shut them down [23], but others continue to explore whether design can help.

The Guardian is among those that have sought to improve online comments. In 2012 the news organization introduced single-level threading to its commenting system [33], providing a unique opportunity to examine the influence of that design feature. This study treats the design change as a natural experiment. Analyzing millions of comments posted over the course of a year, and comparing posting behavior before and after the change, we found compelling evidence of a variety of substantively important effects. We begin by demonstrating that threading was associated with an increase in the probability that newcomers—individuals participating in *The Guardian* commenting system for the first time—would return to post again. In other words, threading made the commenting system stickier.

Next, we examine changes in commenting behaviors that we would expect to be associated with threading, and that could help explain the increased stickiness. We focus on three potential behavior changes. First, threading could make it more likely that comments receive a response. Second, threading could increase reciprocity, encouraging users to reply to those who respond to them. And third, threading could promote civility. The change to threaded presentation of comments could also, however, have an adverse effect on any of these three. For example, users might game the system, only responding to others when it helps them achieve higher visibility for their own messages. Such behavior would likely reduce reciprocal exchanges.

In this study, we first analyzed the change in each of these three measures after the design change. Then, we conducted a mediation analysis to assess the extent to which each of these factors can contribute to explaining whether first-time commenters return to comment again. Results suggest that the increased probability of receiving a response following the design change plays a uniquely important role in promoting site stickiness. Before we turn to the analysis, we provide an overview of the motivation for—and implementation of—the design change, and we discuss a body of related work that motivates our predictions.

Background. News site commenting systems were originally seen as a way of drawing readers in and keeping them engaged. Over time, however, they developed a reputation for their vitriol and their lack of substance. Contrary to their mission, news sites often play host to a range of harmful communication practices, including incivility [18], harassment [41], and hate speech [52]. Indeed, several news organizations, including the *Chicago Sun-Times*, *Popular Science*, *Reuters*, and *USA Today*, eliminated their commenting systems altogether in order to avoid the problems they introduce [23]. As a consequence, many people have come to see discussion spaces as something to avoid.

By altering its commenting platform *The Guardian* hoped to create a space that was more attractive to users. In the words of the newspaper's staff, the addition of threading was intended to "make conversations more coherent and help users to engage with each other" [37]. This reasoning appears justified because related work demonstrates that threading can increase conversation coherence [53]. Yet, the positive effect of threading is not universal [20, 26, 47, 63]. It can also result in decreased *perceived* conversation quality and lower participant satisfaction [53]. In addition, even though threading makes it easy for the discussion to scale well, it can make it harder to find the

newest contributions to the conversation [31], and the overall suitability of threading as a design choice depends on the application [20]. *The Guardian*, after implementing threading, reported that although a small number of participants complained, the response from the community was largely positive [33], a claim supported by the fact that threading is still in use today. Empirical evidence for these conclusions, however, has not been released publicly.

UI Change. Prior to the introduction of threading at *The Guardian* in late 2012, comments were displayed chronologically, in the order in which they were posted. Prior to this design change, individuals could reply to one another either by clicking on a "reply" button or by including *@username* at the start of a message, but those responses were always appended to the end of the list of comments. The new design changed the way replies were displayed. In the old design, comments made in reply to a top-level message were indented and listed in chronological order below that post. The new design introduced only one level of threading, meaning that replies to messages below the top level are presented chronologically on the same level as the messages to which they are responding.[1] All but the first few of these replies are collapsed by default[2]. No other aspect of the commenting system was changed—including the notifications system and the mechanisms by which users post comments and replies. *The Guardian* community members use pseudonyms and have to be logged in in order to contribute to a conversation. Such characteristics were also unchanged.

All design changes, including the threading change studied in this paper, were announced on the *Inside the Guardian Blog*, which lists changes and updates to theguardian.com [60]. To the best of our knowledge, no other major change that would affect the commenting system was implemented during the period studied in this paper, though *The Guardian* has changed other aspects of the interface since then. In June 2013, the site enabled commenting on its mobile website [25]. In January 2015, it fully launched a new design for the site that it had been testing in a number of sections [9]. In addition to changing the navigation by introducing a number of personalization options, the new design also affected the commenting section by providing an option to highlight some comments on the first screen of an article. We plan to study the effect of such changes in future work.

Although the change to single-level threading was a fairly modest interface change, a small subset of the news site's users responded vigorously. A pair of posts to the *Inside the Guardian Blog* explaining that the newspaper was testing a threading system in early 2012 generated almost 700 comments, while announcements about the site-wide rollout elicited more than 2,5000 responses [6, 33, 37, 61]. Some users wrote to express their support, but most comments were critical of the change.

2 RELATED WORK

This is not the first time scholars have examined the influence of threading on commenting behavior. Previous work has examined the impact of threading in experimental settings and has focused on domains such as email [63], user groups [54], chat [26, 53] and social media [47]. For instance, [53] constructed an experimental chat platform and found that a threaded view increased coherence but, surprisingly, that users preferred the linear view and perceived conversations in this mode to be of higher quality. Given the mixed findings, a number of studies aimed to improve on purely

[1] We have not located a contemporaneous archive showing the commenting system before and after the change. The current interface, when showing historical articles, shows the difference between unthreaded and one-level threading, though it may reflect some other interface changes that have occurred since then. For example articles with and without comment threading, please see the auxiliary materials.

[2] Readers have an option to expand all or to expand/collapse individual threads.

threaded solutions by proposing new designs [20, 26, 47, 63]. For instance, [26] complemented the threaded view by providing *BackTalk*, which displays new replies both in the appropriate thread and at the bottom of the history pane.

Our study differs from previous work both in the domain of interest—we focus on a news site forum—and methodology—we examine the impact of threading in a real online community as opposed to an in-lab setting. The closest to our study, in this sense, is a recent study of Menéame, a Digg-like social media service based in Spain [2]. This study focused on the impact of threading on conversation reciprocity and found that the shift from a linear conversation view to one that displayed comments hierarchically was associated with sharp increase in two-way conversations among users. We found contrary results on reciprocity, and we consider potential explanations and implications of these contrasting results in the discussion section.

Threading is only one of the many design elements that can influence commenting behavior. Other changes that were studied include ending anonymity, introducing moderation, and cuing desired behaviors. Here we provide a short review of such studies. For a broader review of deliberation online, we refer the reader to [29].

Ending anonymity is the most extensively researched change to commenting systems. Many early systems allowed users to post comments anonymously, and there is compelling evidence that this often contributed to the high levels of incivility observed in these spaces. One large-scale observational study analyzed 42 million comments posted on the *Huffington Post* collected both before and after it ended anonymous posting [27]. Analyses show that users were less likely to post comments after the change, and that the magnitude of this effect varied by topic (e.g., comments in the world section dropped by 88%, while those in parenting dropped by about 50%). This may be due in part to a reduction in spam campaigns: stories with tags that were commonly associated with spam campaigns before the change (e.g., South Korea, Kim Jong-un, NSA surveillance) saw larger-than-average drops in commenting. Furthermore, the comments posted after the change appeared more civil. For example, there were fewer typos (a 35% drop), fewer words in all caps (10%), and a reduction in words classified as offensive (18%).

These effects are not limited to the U.S. In 2005 South Korea enacted the Real Name Verification Law, which forbade anonymous online posting on election-related discussion forums. Although the policy was ultimately overturned by courts, researchers collected data in 2012 during a period when some news sites required users to identify themselves, while others made identification voluntary [16]. Interestingly, voluntary identification was more effective at stemming the use of profanity than mandatory identification, an effect that was most pronounced among the heaviest commenters.

Moderation of online comments has also been credited with improving the quality of discourse, yet studies show that censorship of comments can make readers reticent to post, which can reduce participation [49, 62]. Direct engagement with commenters has likewise been shown to improve deliberation, but only when that engagement is with an identifiable journalist as opposed to an unknown staff worker [57]. While influencing behavior through moderation may come with costs in both time and traffic, research also suggests that deliberation can be improved simply by cuing pro-social norms [40].

Finally, given that one of the main focuses of our paper is on repeat participation, we note the rich literature identifying factors that contribute to this desired outcome [3, 4, 10, 11, 13, 34, 35]. For instance, [11] found that both receiving a response, and linguistic matching between a newcomer and the group are predictive of repeated participation in Twitter chats. [3] found receiving a response to be important in Usenet groups, and they identified characteristics that are predictive of receiving a response. [10] showed that the decreasing participation on Norwegian online groups was due to a number of complex characteristics including lack of interesting attendees, low-quality

content, low usability, and harassment. Unlike these studies, which are based on correlational analysis and surveys, we utilize a natural experiment to determine which factors affected site stickiness. The present study is also unique in that we worked to understand the mechanisms by which threading influences site stickiness.

3 HYPOTHESES

Repeat participation. A primary goal of *The Guardian's* design change was to improve user satisfaction, creating a system that would encourage readers to post comments more regularly in response to news content. Staff members hoped that the new system would promote repeated participation, and so we make this our key outcome. Although providing evidence that threading made users more likely to return is valuable, identifying the mechanisms by which threading promotes stickiness is just as important. Among other things, it might suggest further refinements to how threading is implemented that could maximize its effects. It also allows designers to consider whether interface changes unrelated to threading might have similar effects. We identify three processes by which threading might encourage users to return: (1) increasing the probability that comments receive responses, (2) increasing reciprocal exchanges among commenters, and (3) increasing the civility of comments. We consider both how threading could promote each change in commenting behavior, and how the change would in turn encourage users to return to the commenting system at a later date.

Receiving a response. The affordances of a threaded conversation system could promote responses. Although the mechanism for creating a reply was unchanged, the new interface provided a potentially powerful visual clue about how the system is meant to be used [43]. In contrast to a chronological display, which tends to focus users' attention on the most recent comments [7], nesting responses below the originating comment subtly calls attention to prior instances of this desired behavior. Although users still have the option of commenting at top level about the original article, the threaded presentation accentuates the fact that the system is intended to be a platform for conversation among readers.

The threading change might have also created attentional incentives for responding rather than posting at top level. The new system rewards users for responding to comments that have few prior responses. This is an outgrowth of the way that comments are displayed. Top-level comments are displayed chronologically, and initially only a few responses to each top-level comment are shown. Thus, a strategy for increasing attention to one's comment is to respond to the first top-level comment that lacks other responses. If users were to do this consistently, the number of users who received at least one response should increase with threading.

Further, we anticipate that individuals who received a response to a comment they posted would be more likely to contribute again. Posting a comment is a means of social expression, and without a response commenters have little indication that their voice has been heard. Replying closes the loop, rewarding the original poster for his or her contribution. Research has confirmed that individuals' continued contribution to online discussion spaces is influenced by the level of engagement experienced, or in other words, the probability of receiving a response [11, 34].

In sum, we anticipate that one important way that threading encourages repeated use of the commenting system is by promoting responses from other users.

Reciprocity. Reciprocity goes beyond simply receiving a response; an exchange is reciprocal when each party responds to the other. There can be no reciprocity without response, which means that as responses increase, so do the opportunities for reciprocal exchanges. To the extent that the visual layout signals that the system is intended to facilitate conversation with other users, and that it calls attention to instances where this is happening, we would expect users' behavior to

change accordingly. In other words, because reciprocal interaction among commenters is more easily noticeable in the threaded interface, users should be more likely to recognize this behavior as encouraged, or even expected and thus do it more [17]. Indeed, the introduction of (multi-level) threading at a popular Spanish social networking site was followed by a sharp increase in reciprocity [2].

The attentional incentives in *The Guardian's* implementation of single-level threading, however, might act to reduce reciprocal exchanges. If a user wants his or her comment to be seen, responding to a top-level comment will almost always be more effective than replying to a response, especially because only the first few replies to a comment are shown by default. Thus, the author of a top-level comment who is faced with the choice of responding to a responder (creating a reciprocal exchange), or responding to some other top-level comment that has few responses (not creating a reciprocal exchange), can expect greater visibility for the latter. If responses to top-level comments increase, while replies to those responses remained unchanged, reciprocity levels would decline. In the face of these contradictory potential outcomes we do not have a strong hypothesis about whether, and in what direction, threading will influence reciprocity.

If threading does alter reciprocity, in either direction, we would expect that change in reciprocity to influence repeat participation levels. Reciprocal exchange is a foundation of social attachment and community building, and repeated interactions between people can contribute to a sense of belonging, which is a powerful incentive for continued participation [46]. In other words, if threading promotes reciprocity, repeat participation should increase; if it reduces reciprocity, repeat participation should fall.

Civility. Threading has the potential to reduce the depersonalization so prevalent in online communication spaces and consequently to promote civility. Online communication has long been plagued by uncivil behavior [55]. One of the core explanations for this behavior pattern is that many aspects of computer-mediated communication, notably anonymity, promote a shift of attention from individual identity to social identity [48]. The increasing salience of social identity often leads individuals to endorse their ingroup identity, to express hostility toward outgroups, and to employ negative outgroup stereotypes. The visual interface of the threaded comment system could, however, help shift attention back toward individual identity. The more users see themselves as engaged in an evolving conversation with other individuals, the less likely they should be to exhibit hostility toward those with whom they disagree.

Many users express trepidation about using commenting systems based on the negativity they have observed [12], and there is ample evidence that people do have an aversion to discussion conflict [42]. Given this, we would anticipate that threading-induced civility could promote repeat participation.

4 DATA

Section	Article Count	Participant Count	Avg. Comments/Article
Business	2274	24980	114.81
Environment	1836	21484	90.05
Tech	2356	22997	117.05
U.S.	2288	30597	90.12
World	2672	42643	129.28

Table 1. Guardian comment section descriptives for the period between 180 days before and after the design change.

Our dataset spans the period between May 3, 2012 and May 21, 2013. We examine commenting behavior observed over five sections in *The Guardian*: business, environment, tech, U.S., and world

news. We chose these sections to provide broad coverage of political and non-political topics that had significant participation. Threading was implemented on October 29, 2012, for the environment section and on November 22, 2012, for all other sections.

This dataset includes 11,425 unique articles, 84,135 unique forum participants and 1,253,811 comments generated by these commenters. Table 1 provides high-level descriptive statistics at the section level for the 360-day period surrounding the design change. For most analysis, we focus on the 100-day period surrounding the threading change (50 days before and 50 days after). We expand the period to 360 days surrounding the design change when looking at site-level stickiness because this characteristic is better measured for a longer period to determine more accurately whether a community member came back for further contributions.

5 MEASUREMENT

Repeated participation. We measure repeated participation in two ways. First, we measure the change in repeated participation at the *article level* through the reduction of one-off commenting. In other words, we measure, for each article, the percentage of commenters who posted at least two comments on it. Then, for each day, repeated participation at the article level is defined as the median of this measure, across all articles.

Second, we measure the change in repeated participation at the *site level*. In other words, we measure how sticky *Guardian* attendance is over time. To compute this measure, we estimate the probability of a newcomer who made a comment for the first time on day i to come back to make at least one more comment in the next k days as $|N_i^+|/|N_i|$, where $|N_i|$ is the number of newcomers on day i and $|N_i^+|$ is the number of newcomers on day i who ended coming back for further participation.

Receiving a response. We measure the response rate of conversations on *The Guardian* on a given day i as the median response rate of all articles on day i. The response rate of an article a is defined as the fraction of all comments on article a written as a response to an earlier comment on the same article.

Reciprocity. To compute reciprocity on a daily basis, we first generate daily graphs $G_i = \{N_i, E_i\}$ for each day i. N_i is the set of participants who have at least one comment on day i. For each participant u who responded to participant v on day i, there is an edge $e_{i,u,v} \in E_i$ with weight $w_{i,u,v}$ where $w_{i,u,v}$ is the number of times u responded to v on day i. Having constructed graphs on a daily basis, we measure reciprocity in two ways.[3]

(1) *Unweighted reciprocity:* For a directed network G_i, a simple measure of reciprocity can be defined as: $r_i = \frac{|E_i'|}{|E_i|}$, where E_i' is the set of bidirectional (reciprocated) edges. However, this measure is strongly influenced by network size. To ensure that network size is not driving any observed effects, we use the corrected reciprocity proposed in [30], which measures the relative level of reciprocity in comparison to a random network with the same number of nodes and edges as:

$$r_{unweighted,i} = \frac{r_i - d_i}{1 - d_i} \tag{1}$$

where $d_i = |E_i|/(|N_i| * (|N_i| - 1))$ is the density of graph G_i.

[3]We also measured reciprocity at the article (as opposed to day) level and defined reciprocity of a day i as the reciprocity of the average article on day i. The results were consistent with the daily graphs and are omitted for brevity.

(2) *Weighted reciprocity:* Given a graph G_i, we also measure the weighted reciprocity as proposed by [56]:

$$r_{weighted,i} = \frac{W_i'}{W_i} = \frac{\sum_u \sum_{v \neq u} w'_{i,u,v}}{\sum_u \sum_{v \neq u} w_{i,u,v}} \quad (2)$$

where $w'_{i,u,v}$ is the minimum of the edge weight from u to v and the edge weight from v to u on day i (computed as $min(w_{i,u,v}, w_{i,v,u})$). This definition, unlike the previous, takes the weight of edges (frequency or interactions) into account. Both $r_{unweighted,i}$ and $r_{weighted,i}$ have also been used in [2], so our results are comparable to theirs.

Civility. We examine civility (or incivility) observed on a given day i through three measures:

(1) *Comments deleted:* Moderators at the *Guardian* remove comments "...if they go against the Community Standards & Participation Guidelines..." [59]. An inspection of the guidelines reveals that the standards are strongly tied to civility measures [58]. Thus, in order to measure incivility on a given day i, we measure the fraction of all comments posted on a given day i that are deemed unacceptable by *Guardian* moderators and deleted. We note that this is a fairly conservative measure because moderators only remove unambiguous violations of the community standards. Many forms of incivility, including insults and stereotyping, are often allowed to remain.

(2) *Use of all caps:* Consistent with conventions of online etiquette, we treat words typed in all capital letters as the equivalent of shouting. The frequency with which users type in all caps has been used as a measure of incivility in other work [27]. We measure this form of incivility on a given day i as the fraction of comments that included at least one word in all capitals. In order to identify all caps words, we only consider words with at least four characters and exclude common abbreviations (e.g., NASA).

(3) *Use of swear words:*Use of offensive language is a strong indicator of incivility [27]. To measure incivility through this lens on a given day i, we measure the fraction of swear word usage across all comments on day i. Swear words are identified using the LIWC dictionary [45].

6 METHODOLOGY

Interrupted time series design. In order to test the hypotheses listed above, we rely on a quasi-experimental approach called interrupted time series analysis. The interrupted time series design comprises several waves of observations collected over a specified time interval, Δt, before and after the introduction of a treatment. Given such data, the interrupted time series analyses use segmented regression analysis specified as follows: Let T define time, X_t a dummy variable indicating the period (0 for pre-intervention and 1 for post-intervention), and Y_t the outcome at time t. We estimate the impact of the treatment using segmented regression model as:

$$Y_t = \beta_{level} + \beta_{slope} T + \beta_{levelChange} X_t + \beta_{slopeChange} T X_t \quad (3)$$

We use this method to study the change in *repeated participation, civility, response rate*, and *reciprocity* 50 days before and 50 days after the treatment (design change).[4] Here, β_{level} defines the baseline behavior (e.g. response rate) just before threading was introduced, β_{slope} represents how that behavior changed as a function of time pre-threading, $\beta_{levelChange}$ is the change in levels after threading, and $\beta_{slopeChange}$ indicates the slope change after threading was introduced.

Threats to internal validity: There is no comparison group in this interrupted time series design, meaning that we cannot account for unobserved confounds. Therefore, events happening at the same time as the treatment might be responsible for any effects we observe. However, studying the

[4]We repeat our analyses setting the time interval Δt to various values between 50 and 100 to ensure the robustness of our results. Findings are qualitatively consistent.

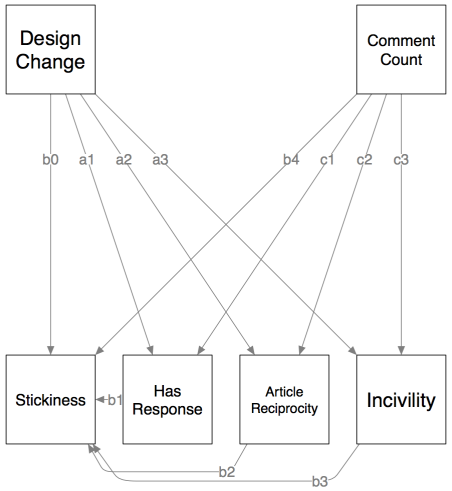

Fig. 1. Mediation Analysis Diagram

behavior across different sections alleviates some of the concerns, since they did not all change to threading on the same date. Patterns that are consistent across sections are less likely to be due to factors external to the news site.

A number of recent studies have relied on regression discontinuity design [2, 39] to study the effect of similar design changes. We do not use a regression discontinuity design (RDD), despite the stronger causal interpretation it would provide, because time series data are generally incompatible with its assumptions. RDD assumes that observations on either side of the cutoff between comparison groups (in our case, the time of design change) are identical on average except for treatment status. When applied to time series data, RDD can only estimate effects for small time intervals around the cutoff date. In our case, the farther an observation is from the introduction of threading, the more likely it is that users being compared will differ in ways that are unrelated to the platform change. RDD is ill-suited to studying dynamic impacts that cascade over time, such as a sustained increase in response rate after the introduction of threading.

Mediation analysis. One of the main goals of this paper is to identify the mechanisms by which threading influences site stickiness. To that end, we use mediation analysis. Mediation analysis is a statistical method that qualifies the underlying mechanism or process by which one variable influences another variable through one or more mediator variables [38]. Figure 1 shows the mediation model, which predicts the final outcome variable, stickiness, measured as the probability that a user u who first commented on day i also commented on any article for a future day $(i, i+k]$. Design change is a binary variable indicating whether day i was before or after the design change. The parameter $b0$ captures the direct impact of the design change on stickiness. The parameter $a1$ captures the impact of the design change on whether u received at least one response on day i from any other commenter (has response) and $b1$ captures the impact of receiving a response on the final outcome. Together, $a1 * b1$ captures the indirect effect of the design change on stickiness that is mediated by the change in the probability of getting a response. Similarly, $a2 * b2$ captures the indirect effect that is mediated by the reciprocity observed on the article u commented on[5].

[5]If there are no responses, reciprocity is zero.

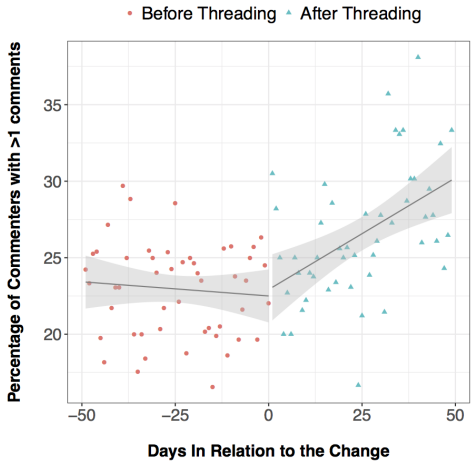

Fig. 2. Article-level repeated participation over time, all sections.

	$Model_{repeat,article}$
β_{level}	$22.54\ (1.01)^{***}$
β_{slope}	$-0.02\ (0.04)$
$\beta_{levelChange}$	$0.31\ (1.39)$
$\beta_{slopeChange}$	$0.17\ (0.05)^{**}$
R^2	0.32

$^{***}p < 0.001, ^{**}p < 0.01, ^{*}p < 0.05$

Table 2. Segmented Regression Analysis for article-level repeated participation, all sections.

Likewise, $a3$ and $b3$ capture the indirect effect of incivility, which is measured as the fraction of comments deleted from the conversation on article a that u commented on. The model includes article popularity (comment count) as a control variable, measured as the number of comments on the article that u commented on [6].

We use structural equation modeling (SEM) to estimate our mediation models given its various advantages over other methods [32], and individual indirect effects are estimated using bootstrap confidence intervals based on 10,000 samples. All SEMs are estimated using lavaan [50].

7 RESULTS

Repeated participation. Article level: Figure 2 shows time-series data for repeated article-level participation, overlaid with linear estimates based on segmented regression analysis. Although there is no immediate effect of threading (the change in level is not significant), the figure shows that the fraction of individuals who posted multiple comments on an article increased steadily after the platform change (a significant increase in slope; see Table 2 for coefficients). These findings support the first hypothesis.

Site level (stickiness): Next, we estimate the probability that an individual who contributed their first comment at time t would make another contribution between $(t, t + t_K]$. We do so by simply

[6]If u commented on multiple articles on the same day, we take the average for reciprocity, incivility and comment count. Such users are rare.

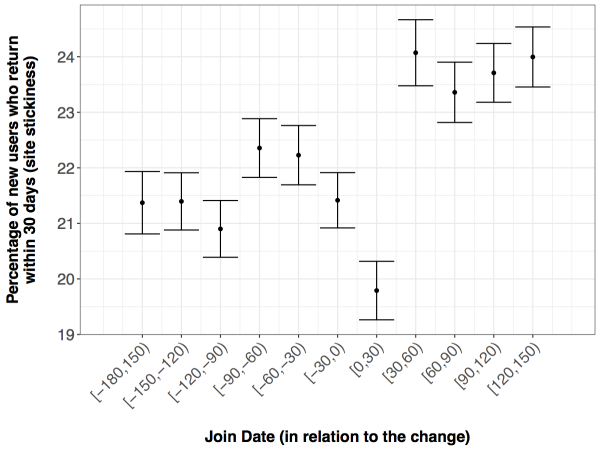

Fig. 3. Site-level repeated participation over time. Whiskers give standard errors.

computing the fraction of newcomers in a given period who commented again within K days of their first comment. Results for $t_K = 30$ are presented visually in Figure 3.[7] The horizontal axis indicates when individuals made their first comment. For instance, the x values $[-180, -150)$ denote the group that made its first contributions between 180 and 150 days before the threading change. The y value for this group gives the estimated stickiness during this 30-day time window. Within each bin, the stickiness is computed as the fraction of newcomers in that bin that came back within $K = 30$ days.

An interesting pattern emerges in Figure 3: the user retention rate hovers around 22% before the change, drops to 20% around the time of the change, and increases to 24% after this transition period. In other words, site stickiness was at its lowest among individuals who joined within the first month following the introduction of threading, but quickly exceeded pre-threading rates in the months that followed. An inspection of comments around the time of the design change reveals that the community reacted strongly to the change, with comments that were atypically aggressive. Such conditions can be suboptimal for newcomers. However, after this transition phase, the platform change seems to have increased *The Guardian's* ability to retain new members of the community.

Receiving a response. We measure the response rate of *Guardian* conversations through segmented regression analysis and present the findings in Figure 4. As expected, use of the reply feature increased after threading was introduced. As is evident from Figure 4, the fraction of comments posted in response to other messages increased more rapidly following the platform change. Reviewing the model coefficients confirms that this change is statistically significant (see Table 3). After threading was introduced, the response rate increased by $\approx 1\%$ every 4 days over all sections (last column). Reviewing the sections separately, we find that those with lower response rates initially (U.S., business, and tech) experienced a steady increase in responsiveness ($\beta_{slopeChange}$ is positive and significant). This effect was not, however, statistically significant in sections that already had the highest rates of response (world, and environment).[8]

[7] We varied t_K between 10, 20, and 30 days and found the same overall qualitative findings.
[8] Note that there is less variability in outcome—the daily percentages of comments that are responses—when pooling all sections versus individual sections. Thus, the model's R-squared value is much higher when pooling all sections.

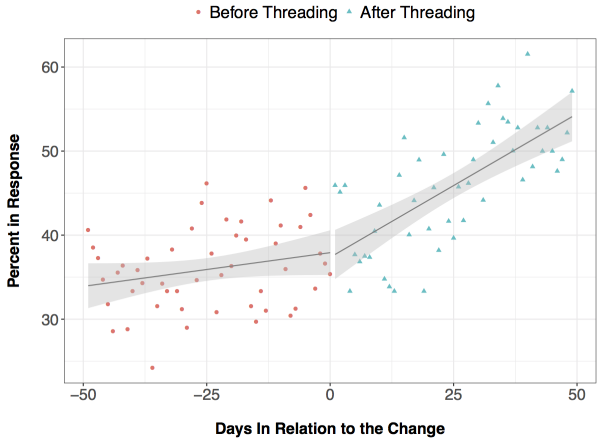

Fig. 4. Response rate over time, all sections.

	U.S.	World	Business	Environment	Tech	Overall
β_{level}	33.97 (3.08)***	43.63 (3.05)***	29.59 (3.92)***	42.49 (4.69)***	25.93 (3.48)***	38.13 (1.44)***
β_{slope}	−0.01 (0.11)	0.10 (0.11)	−0.07 (0.14)	0.05 (0.16)	0.03 (0.12)	0.09 (0.05)
$\beta_{levelChange}$	0.67 (4.27)	1.29 (4.23)	−1.07 (5.47)	−0.48 (6.43)	0.30 (4.90)	−0.92 (1.99)
$\beta_{slopeChange}$	0.38 (0.15)*	0.02 (0.15)	0.48 (0.19)*	0.07 (0.22)	0.53 (0.17)**	0.26 (0.07)***
R^2	0.26	0.11	0.14	0.02	0.39	0.61

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 3. Segmented regression analysis for response rate, by section and overall.

Reciprocity. We assess threading's influence on reciprocity using segmented regression analysis (see Table 4), and by plotting the results for one of these measures ($r_{unweighted}$ calculated for daily networks; see Figure 5). Surprisingly, there was a significant drop in the level of reciprocity immediately after threading was introduced, though it held steady for the remainder of the study ($\beta_{slopeChange}$ is non-significant). This finding holds across all four measures of reciprocity.

Civility. The results of the segmented regression analysis of civility are illustrated in Figure 6. Although visual inspection might appear to suggest that threading influenced civility, neither the level nor the slope changed significantly for any of the three measures after the platform change (see Table 5). This lack of finding is consistent across all sections of the newspaper (results not shown).

Mediation. As noted at the outset, our primary goal is to understand whether, and how, the introduction of threading made the commenting system stickier. Analyses thus far have shown that the platform change did promote repeated use, and that it also influenced two behaviors expected to

We also investigated whether the increase in response rate was driven by topic-specific changes. Using newspaper-assigned keywords (e.g. Apple, Microsoft, Barack Obama, climate change), we identified topics that were covered in at least 50 articles before and after the system change, comparing response rate changes associated with each topic. The results suggest that most topics had an increase in responding behavior. Although there is variation, no one topic appears uniquely influential.

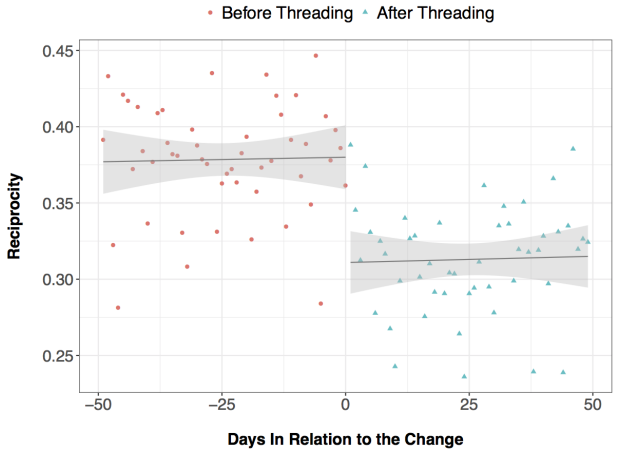

Fig. 5. Reciprocity over time, all sections.

	$Model_{r_{weighted},day}$	$Model_{r_{unweight},day}$
β_{level}	0.43 (0.01)***	0.38 (0.01)***
β_{slope}	0.00 (0.00)	0.00 (0.00)
$\beta_{levelChange}$	−0.09 (0.02)***	−0.07 (0.01)***
$\beta_{slopeChange}$	−0.00 (0.00)	−0.00 (0.00)
R^2	0.45	0.44

$^{***}p < 0.001, ^{**}p < 0.01, ^{*}p < 0.05$

Table 4. Segmented regression analysis for reciprocity, all sections.

help explain the relationship. Taken together, this suggests that threading may indirectly influence site stickiness.

Figure 7 summarizes the decomposition of the direct and indirect effects from the mediation analysis. It shows estimates of direct and indirect effects obtained through bootstrapped samples. Results confirm that the introduction of threading is associated with an increase in site stickiness, and that this effect is mediated by the increasing fraction of comments that receive responses ($(a1 * b1$ in Figure 1; 95% CI 0.009, 0.012). Reciprocity worked in the opposite direction. The increase in stickiness following threading was constrained by the fact that threading reduces reciprocity and reciprocity is positively associated with stickiness ($a2 * b2$ in Figure 1; 95% CI -0.006, -0.004). There was no evidence that incivility mediated the relationship between threading and repeated use of the commenting system. The total indirect effect, which accounts for both the positive influence via receiving a response and the negative influence via reciprocity, was positive (95% CI 0.004, 0.007).

8 DISCUSSION

The introduction of single-layer hierarchical threading to the comment section of *The Guardian* newspaper was followed by an increase in stickiness, the fraction of individuals posting a comment who returned to post again, both on any given article, and via the commenting service as a whole. That such a straightforward design change could have this effect should be encouraging to online services seeking to promote user contributions. In our view, however, insights about the *mechanisms*

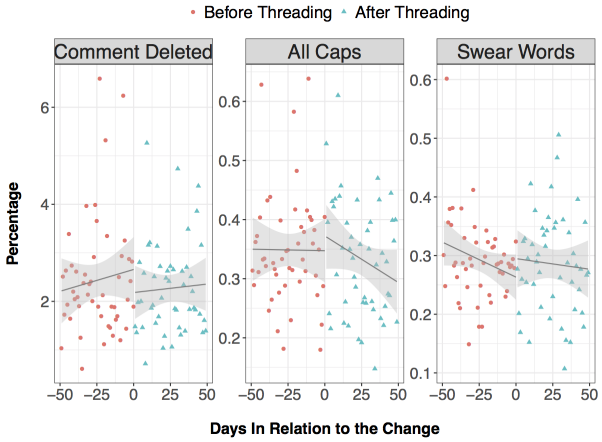

Fig. 6. Civility measures over time, all sections.

	Comment Deleted	All Caps	Swear Words
β_{level}	2.7213 (0.3143)***	0.3425 (0.0277)***	0.2581 (0.0225)***
β_{slope}	0.0109 (0.0109)	−0.0002 (0.0010)	−0.0013 (0.0008)
$\beta_{levelChange}$	−0.5622 (0.4358)	0.0324 (0.0383)	0.0390 (0.0312)
$\beta_{slopeChange}$	−0.0068 (0.0152)	−0.0015 (0.0013)	0.0009 (0.0011)
R^2	0.0195	0.0369	0.0345

$^{***}p < 0.001, ^{**}p < 0.01, ^{*}p < 0.05$

Table 5. Segmented regression analysis for civility, all section.

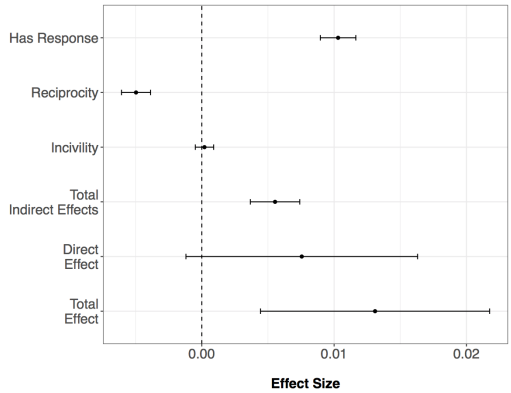

Fig. 7. Direct effect and indirect effect of threading on site stickiness via three mechanisms (95% confidence intervals shown)

by which threading promoted repeat contributions are considerably more important. Understanding these processes can help designers to refine their approach to threading, accentuating the paths that do the most to promote participation, and altering any that constrain it. This might also suggest design changes unrelated to threading that could further boost engagement.

The most important way that threading promoted repeat participation in this study was by making it more likely that someone who posted a comment would receive a response. There are a variety of ways that designers might capitalize on this knowledge. One straightforward way to reduce unanswered comments is to make them easier to find via optional filters or by highlighting unanswered comments of high quality, including those that introduce new information, share novel perspectives, or raise questions. Furthermore, a commenting system might actively solicit responses to content still awaiting a reply. For instance, inspired by systems that aim to increase participation in collaborative crowdsourcing platforms [14, 19], commenting systems may be able to reduce the cost, and increase the personal value, of responding to others by directing community members' attention to comments that best fit their knowledge or expertise. Indeed, research shows that individuals are more likely to contribute when they are reminded of their uniqueness, and when they are given specific goals [5]. Effectively matching commenters with comments can be automated by building accurate profiles of commenters through analysis of their past comments and interactions with other members.

More subtly, it may be possible to adjust the threading interface to provide hints that the system is meant to be used to respond to other users. Displaying statistics describing the proportion of comments receiving replies in the current thread, and highlighting this information when the proportion is high, could serve as an influential signal to users. Other interface changes could also help. Research shows that social comparison and competition through the use of feedback mechanisms can promote desirable behavior [15, 22]. For example, ranking individuals on a leader board based on the frequency with which they respond to others might encourage a higher response rate. Similarly, inspired by research that shows team-focused leader boards increase engagement more than individually focused leader boards [24], designers might consider sorting users into "teams", encouraging different sections of *The Guardian* to compete with one another on various measures of discussion quality. Of course, caution must be taken. If external incentives lead to less meaningful responses, or if recipients view responses as externally motivated, then the link between receiving a response and returning might fade.

The more troubling pattern revealed by our analysis is that threading reduced reciprocity, attenuating the positive effect that threading had on repeat participation. A number of factors may help account for this counter-intuitive finding. First, it is plausible that at least some fraction of the increase in the response rate was due to individuals choosing to reply in order to promote the visibility of their message, not to engage in genuine conversation. Such replies are unlikely to elicit reciprocation. Second, the *particular* threading design used by *The Guardian*—allowing only a single level of responses and hiding all replies beyond the first few by default—might have inadvertently disincentivized reciprocation. If visibility is the primary motivation, then replying to the first unanswered comment, regardless of its content, is the most obvious strategy; a reply to a reply, which creates a reciprocal exchange, is seen by far fewer users if there are already several comments in that thread. Indeed, a recent study that analyzed the introduction of multi-level hierarchical threading to a popular Spanish news site [2] on which all comments are visible by default found the change led to an increase in reciprocity. While it is possible that other factors (e.g., cultural context, types of news services, etc.) also played a role, we believe that the divergent effects are due, at least in part, to the different ways in which threading was implemented. We cannot, however, say conclusively which mechanism led to the drop of reciprocation, and we encourage future work to explore this phenomenon.

Although the effect of the threading design change on reciprocity was not in the desired direction, our results suggest that reciprocity is positively associated with site stickiness. There may be other ways that designers can promote reciprocity, and doing so should make users more likely to return. For example, a commenting system might aim to make reciprocal exchanges easier, notifying

users when they receive a reply via a channel outside the comment system (e.g., email or text). Efforts to promote reciprocity could also extend beyond pairs of communicators. Designers might aim to increase visibility of reciprocal conversations, thereby creating visibility incentives for engaging. Designers could also explicitly draw attention to contributions that exemplify high reciprocity. Changes in descriptive norms—in this case, community members' perceptions about how other users typically behave in the communication space—can have a powerful influence on individuals' behaviors [17]. Furthermore, when high-quality comments are shown first, the quality of subsequent posts tends to be higher [8]. Following this idea, the system could present threads exhibiting high reciprocity at the top of the comment list and encouraging other users to join in. As we explain later, however, this approach would likely need to be introduced in conjunction with other more direct mechanisms promoting reply and reciprocity.

Intriguingly, we found no evidence that civility is substantively influenced by threading, or that civility mediates the influence of threading on stickiness. There are a number of possible explanations for this non-finding. First, civility on *The Guardian* was substantially higher than for many of the other online commenting systems that have been studied. For instance, [28] found that more than a third of tweets and of newspaper site comments included insults. A study of a different newspaper's website found that about one in five comments (22%) exhibited incivility, most often name calling (14% of all comments) [18]. In comparison, the fraction of comments *The Guardian* deleted for being uncivil was as low as 2% before the threading design change. Of the remaining comments, only three in a thousand included a swear word. Thus, it is possible that our failure to detect a change in civility reflected the fact that civility was already very high, creating a ceiling effect. Second, we had predicted that incivility would fall because the introduction of threading would encourage users to view their contributions as part of a discussion, not as anonymous voices in a crowd. Shifting attention from group to individual identities should help to reduce incivility [48]. Such an outcome, however, may be contingent on increased reciprocity, which we did not observe here. Finally, it could be that our hypothesis was simply wrong: perhaps civility is unrelated to the level of interactivity. Our data did not allow us to assess which of these explanations is correct. To adjudicate among them, researchers need to identify discussion spaces exhibiting more incivility and then observe what happens when changes effectively promoting replies and/or reciprocity are introduced.

Lessons learned here may have other implications as well. For example, our results suggest that the success of Facebook's social ranking algorithm ("Top comments") in promoting high-quality contributions [8] is likely due in part to the alignment between descriptive norms and other contribution incentives. On Facebook, the comments that users see are of higher quality, which may help set a higher bar for their own contributions (descriptive norms). This approach to ranking may also remind users that they can be rewarded for their effort via higher visibility to other users (a complementary visibility incentive). In contrast, when descriptive norms and visibility incentives are in tension with each other, as they were at *The Guardian*, results are less encouraging. We observed that threading could make the practice of reciprocity more salient—a descriptive norm—but that individuals whose primary motivation was for their messages to be seen would have an incentive to reply to new comments instead of replying to a reply. The fact that threading was associated with a drop in reciprocity suggests users were more concerned with the second behavior. Thus, one important risk that designers must guard against is of creating competing incentives: if users have to choose between contributing to a richer conversation and having their own message seen by more people, they will often choose the latter.

9 LIMITATIONS

Several limitations are worth noting. First, our study uses observational data, which are vulnerable to a variety of internal validity threats, including the risk of unobserved disturbances in the environment. Randomized experiments would be an invaluable tool for addressing these threats, thereby helping to establish causality. Second, we study one particular implementation of threading on one particular news sites. Commenting systems on other news site, and alternative threading implementation could lead to different results. Furthermore, an online community found at a news organization could function very differently from online communities that prioritize social relationships over news discussions (e.g., Facebook). Therefore we urge caution when generalizing from these findings.

10 CONCLUSION

The Guardian newspaper's introduction of single-layer hierarchical threading to its comment section creates a natural experiment that we exploit in order to better understand the consequences of this design change. Consistent with the publisher's aims, we show that the new design was followed by an increase in the rate of individuals returning to post again, both on any given article, and via the commenting service as a whole. For roughly two months leading up to the design change, the proportion of users who made more than one comment on an article hovered between 20% and 25%. After threading was introduced, the proportion of repeat commenters on a single article began to climb at the rate of more than a percent each week. After 50 days, the proportion of users leaving multiple comments had grown to 30%. Similarly, though less dramatically, the percentage of first-time commenters who returned to make a comment on a subsequent day was a couple of percentage points higher 30 days after the design change.

Mediation analysis indicates that increased stickiness on *The Guardian* was in large part due to the increase in the proportion of users receiving responses on the site. Further, we find that some of the potential benefits of threading were reduced by the drop in reciprocity associated with the interface change. By identifying the mechanisms through which threading promotes stickiness, this study provides insights that can help designers both improve how threading is implemented and make other design changes that might encourage participation by promoting responsiveness and reciprocity.

REFERENCES

[1] Ashley A. Anderson, Dominique Brossard, Dietram A. Scheufele, Michael A. Xenos, and Peter Ladwig. 2014. The Nasty Effect: Online Incivility and Risk Perceptions of Emerging Technologies. *Journal of Computer-Mediated Communication* 19, 3 (2014), 373–387.

[2] Pablo Aragón, Vicenç Gómez, and Andreas Kaltenbrunner. 2017. To Thread or Not to Thread: The Impact of Conversation Threading on Online Discussion. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017*. AAAI Press, 12–21. https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15609

[3] Jaime Arguello, Brian S. Butler, Elisabeth Joyce, Robert Kraut, Kimberly S. Ling, Carolyn Rosé, and Xiaoqing Wang. 2006. Talk to me: foundations for successful individual-group interactions in online communities. In *CHI '06*. ACM, 959–968. https://doi.org/10.1145/1124772.1124916

[4] Patrick J. Bateman, Peter H. Gray, and Brian S. Butler. 2006. Community Commitment: How Affect, Obligation, and Necessity Drive Online Behaviors. In *Proceedings of the International Conference on Information Systems, ICIS 2006, Milwaukee, Wisconsin, USA, December 10-13, 2006*. Association for Information Systems, 63. http://aisel.aisnet.org/icis2006/63

[5] Gerard Beenen, Kimberly S. Ling, Xiaoqing Wang, Klarissa Chang, Dan Frankowski, Paul Resnick, and Robert E. Kraut. 2004. Using social psychology to motivate contributions to online communities. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work, CSCW 2004, Chicago, Illinois, USA, November 6-10, 2004*. ACM, 212–221. https://doi.org/10.1145/1031607.1031642

[6] Martin Belam. 2012. An experiment with threaded comments. (29 February 2012). https://www.theguardian.com/help/insideguardian/2012/feb/29/threaded-comments

[7] Roy Bendor, Susanna Haas Lyons, and John Robinson. 2012. What's there not to 'like'? sustainability deliberations on facebook. *JeDEM-eJournal of eDemocracy and Open Government* 4, 1 (2012), 67–88.

[8] George Berry and Sean J. Taylor. 2017. Discussion Quality Diffuses in the Digital Public Square. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*. ACM, 1371–1380. https://doi.org/10.1145/3038912.3052666

[9] Wolfgang Blau. 2015. Welcome to the new Guardian website. https://www.theguardian.com/help/insideguardian/2015/jan/28/welcome-to-the-new-guardian-website. (2015).

[10] Petter Bae Brandtzæg and Jan Heim. 2008. User loyalty and online communities: why members of online communities are not faithful. In *2nd International Conference on INtelligent TEchnologies for interactive enterTAINment, INTETAIN 2008, Cancun, Mexico, January 8-10, 2008*. ICST/ACM, 11. https://doi.org/10.4108/ICST.INTETAIN2008.2481

[11] Ceren Budak and Rakesh Agrawal. 2013. On Participation in Group Chats on Twitter. In *Proceedings of the 22Nd International Conference on World Wide Web (WWW '13)*. ACM, New York, NY, USA, 165–176. https://doi.org/10.1145/2488388.2488404

[12] Pew Research Center. 2016. *The Political Environment on Social Media*. Technical Report. Pew Research Center.

[13] Calvin M. L. Chan, Mamata Bhandar, Lih-Bin Oh, and Hock-Chuan Chan. 2004. Recognition and Participation in a Virtual Community. In *HICSS '04*. IEEE Computer Society, 70194.2–. http://dl.acm.org/citation.cfm?id=962755.963100

[14] Yan Chen, Rosta Farzan, Robert Kraut, Iman YeckehZaare, and Ark Fangzhou Zhang. 2017. ExpertIdeas: Incentivizing Domain Experts to Contribute to Wikipedia. (2017).

[15] Coye Cheshire and Judd Antin. 2008. The social psychological effects of feedback on the production of Internet information pools. *Journal of Computer-Mediated Communication* 13, 3 (2008), 705–727.

[16] Daegon Cho and K. Hazel Kwon. 2015. The impacts of identity verification and disclosure of social cues on flaming in online user comments. *Computers in Human Behavior* 51, Part A (2015), 363–372.

[17] Robert B. Cialdini and Noah J. Goldstein. 2004. Social Influence: Compliance and Conformity. *Annual Review of Psychology* 55, 1 (2004), 591–621.

[18] Kevin Coe, Kate Kenski, and Stephen A. Rains. 2014. Online and Uncivil? Patterns and Determinants of Incivility in Newspaper Website Comments. *Journal of Communication* 64, 4 (2014), 658–679.

[19] Dan Cosley, Dan Frankowski, Loren G. Terveen, and John Riedl. 2007. SuggestBot: using intelligent task routing to help people find work in wikipedia. In *Proceedings of the 12th International Conference on Intelligent User Interfaces, IUI 2007, Honolulu, Hawaii, USA, January 28-31, 2007*. ACM, 32–41. https://doi.org/10.1145/1216295.1216309

[20] Kushal Dave, Martin Wattenberg, and Michael Muller. 2004. Flash Forums and forumReader: Navigating a New Kind of Large-scale Online Discussion. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work (CSCW '04)*. ACM, New York, NY, USA, 232–241. https://doi.org/10.1145/1031607.1031644

[21] Nicholas Diakopoulos and Mor Naaman. 2011. Towards Quality Discourse in Online News Comments. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work (CSCW '11)*. ACM, New York, NY, USA, 133–142. https://doi.org/10.1145/1958824.1958844

[22] Tawanna R. Dillahunt and Jennifer Mankoff. 2014. Understanding Factors of Successful Engagement Around Energy Consumption Between and Among Households. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)*. ACM, New York, NY, USA, 1246–1257. https://doi.org/10.1145/2531602.2531626

[23] Justin Ellis. 2015. What happened after 7 news sites got rid of reader comments. (November 7 2015). http://www.niemanlab.org/2015/09/what-happened-after-7-news-sites-got-rid-of\-reader-comments/

[24] Rosta Farzan, Laura A. Dabbish, Robert E. Kraut, and Tom Postmes. 2011. Increasing Commitment to Online Communities by Designing for Social Presence. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work (CSCW '11)*. ACM, New York, NY, USA, 321–330. https://doi.org/10.1145/1958824.1958874

[25] Julian Fitzell. 2013. Commenting now available on our mobile website. https://www.theguardian.com/help/2013/nov/18/commenting-now-available-on-our-mobile-website. (2013).

[26] David Fono and Ron Baecker. 2006. Structuring and Supporting Persistent Chat Conversations. In *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work (CSCW '06)*. ACM, New York, NY, USA, 455–458. https://doi.org/10.1145/1180875.1180944

[27] Rolf Fredheim, Alfred Moore, and John Naughton. 2015. Anonymity and Online Commenting: The Broken Windows Effect and the End of Drive-by Commenting. In *Proceedings of the ACM Web Science Conference, WebSci 2015, Oxford, United Kingdom, June 28 - July 1, 2015*. ACM, 11:1–11:8. https://doi.org/10.1145/2786451.2786459

[28] Deen Freelon. 2015. Discourse architecture, ideology, and democratic norms in online political discussion. *New Media & Society* 17, 5 (2015), 772–791.

[29] Dennis Friess and Christiane Eilders. 2015. A systematic review of online deliberation research. *Policy & Internet* 7, 3 (2015), 319–339.
[30] Diego Garlaschelli and Maria I Loffredo. 2004. Patterns of link reciprocity in directed networks. *Physical review letters* 93, 26 (2004), 268–701.
[31] Werner Geyer, Andrew J. Witt, Eric Wilcox, Michael Muller, Bernard Kerr, Beth Brownholtz, and David R. Millen. 2004. Chat Spaces. In *Proceedings of the 5th Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques (DIS '04)*. ACM, New York, NY, USA, 333–336. https://doi.org/10.1145/1013115.1013173
[32] Douglas Gunzler, Tian Chen, Pan Wu, and Hui Zhang. 2013. Introduction to mediation analysis with structural equation modeling. *Shanghai archives of psychiatry* 25, 6 (2013), 390.
[33] Natalie Hanman. 2012. Threading arrives on Comment is free. (3 December 2012). https://www.theguardian.com/commentisfree/2012/dec/03/threading-arrives-on-comment-is-free
[34] Cliff Lampe and Erik Johnston. 2005. Follow the (slash) dot: effects of feedback on new members in an online community. In *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*. ACM Press, Sanibel Island, Florida, USA, 11–20.
[35] Cliff Lampe, Rick Wash, Alcides Velasquez, and Elif Ozkaya. 2010. Motivations to Participate in Online Communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 1927–1936. https://doi.org/10.1145/1753326.1753616
[36] Eun-Ju Lee. 2012. That's Not the Way It Is: How User-Generated Comments on the News Affect Perceived Media Bias. *Journal of Computer-Mediated Communication* 18, 1 (2012), 32–45.
[37] Bella Mackie. 2012. A threading experiment on Comment is free. (March 12 2012). https://www.theguardian.com/commentisfree/2012/mar/12/threading-experiment-comment-is-free
[38] David P MacKinnon, Amanda J Fairchild, and Matthew S Fritz. 2007. Mediation Analysis. *Annual Review of Psychology* 58 (2007), 593.
[39] Momin M. Malik and Jürgen Pfeffer. 2016. Identifying Platform Effects in Social Media Data. In *Proceedings of the Tenth International Conference on Web and Social Media, Cologne, Germany, May 17-20, 2016*. AAAI Press, 241–249. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13163
[40] Edith Manosevitch, Nili Steinfeld, and Azi Lev-On. 2014. Promoting online deliberation quality: cognitive cues matter. *Information, Communication & Society* 17, 10 (2014), 1177–1195.
[41] Adrienne Massanari. 2017. #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society* 19, 3 (2017), 329–346. https://doi.org/10.1177/1461444815608807
[42] Diana C. Mutz. 2006. *Hearing the other side: deliberative versus participatory democracy*. Cambridge University Press, New York.
[43] Donald A. Norman. 1990. *The design of everyday things*. Doubleday, New York, NY.
[44] Deok Gun Park, Simranjit Singh Sachar, Nicholas Diakopoulos, and Niklas Elmqvist. 2016. Supporting Comment Moderators in Identifying High Quality Online News Comments. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, May 7-12, 2016*. ACM, 1114–1125. https://doi.org/10.1145/2858036.2858389
[45] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71 (2001), 2001.
[46] Robert D. Putnam. 1995. Bowling alone: America's declining social capital. *Journal of Democracy* 6, 1 (1995), 65–78.
[47] Stuart Reeves and Barry Brown. 2016. Embeddedness and Sequentiality in Social Media. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. ACM, New York, NY, USA, 1052–1064. https://doi.org/10.1145/2818048.2820008
[48] S. D. Reicher, R. Spears, and T. Postmes. 1995. A Social Identity Model of Deindividuation Phenomena. *European Review of Social Psychology* 6, 1 (1995), 161–198.
[49] June W Rhee and Eun-mee Kim. 2009. Deliberation on the net: Lessons from a field experiment. In *Online deliberation: Design, research, and practice*, Todd Davies and Seeta PeÃČÂśa Gangadharan (Eds.). CSLI Publications, Standford, CA, 223–232.
[50] Yves Rosseel. 2012. Lavaan: an R package for structural equation modeling. *JOURNAL OF STATISTICAL SOFTWARE* 48, 2 (2012), 1–36.
[51] Marie K. Shanahan. 2013. More news organizations try civilizing online comments with the help of social media. (2013). http://www.poynter.org/news/mediawire/218284/more-news-organizations-try-civilizing-online-comments\-with-the-help-of-social-media/
[52] Leandro Araújo Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the Targets of Hate in Online Social Media. In *Proceedings of the Tenth International Conference on Web and Social Media, Cologne, Germany, May 17-20, 2016*. AAAI Press, 687–690. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/view/13147

[53] Marc Smith, Jonathan J. Cadiz, and Byron Burkhalter. 2000. Conversation trees and threaded chats. In *CSCW 2000, Proceeding on the ACM 2000 Conference on Computer Supported Cooperative Work, Philadelphia, PA, USA, December 2-6, 2000.* ACM, 97–105. https://doi.org/10.1145/358916.358980

[54] Marc A. Smith and Andrew T. Fiore. 2001. Visualization Components for Persistent Conversations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '01).* ACM, New York, NY, USA, 136–143. https://doi.org/10.1145/365024.365073

[55] Lee Sproull and Sara Kiesler. 1998. *Connections: new ways of working in the networked organization.* The MIT Press, Cambridge, MA.

[56] Tiziano Squartini, Francesco Picciolo, Franco Ruzzenenti, and Diego Garlaschelli. 2013. Reciprocity of weighted networks. *Scientific Reports* 3 (2013), 2729.

[57] Natalie Jomini Stroud, Joshua M. Scacco, Ashley Muddiman, and Alexander L. Curry. 2015. Changing Deliberative Norms on News Organizations' Facebook Sites. *Journal of Computer-Mediated Communication* 20, 2 (2015), 188–203.

[58] The Guardian. 2009. Community standards and participation guidelines. (2009). https://www.theguardian.com/community-standards

[59] The Guardian. 2009. Frequently asked questions about community on the Guardian website. (2009). https://www.theguardian.com/community-faqs

[60] The Guardian. 2017. Inside the Guardian blog. https://www.theguardian.com/help/insideguardian. (2017).

[61] Hannah Waldram. 2012. An update to news comments - what do you think? (22 November 2012). https://www.theguardian.com/news/blog/2012/nov/22/news-community-comments-threading

[62] Kevin Wise, Brian Hamman, and Kjerstin Thorson. 2006. Moderation, Response Rate, and Message Interactivity: Features of Online Communities and Their Effects on Intent to Participate. *Journal of Computer-Mediated Communication* 12, 1 (2006), 24–41.

[63] Ka-Ping Yee. 2002. Zest: Discussion mapping for mailing lists. (2002). http://zesty.ca/pubs/cscw-2002-zest.pdf

Received April 2017; revised July 2017; accepted November 2017