Self-Censorship on Facebook

Sauvik Das[1] and Adam Kramer[2]

[1]sauvik@cmu.edu
Carnegie Mellon University

[2]akramer@fb.com
Facebook, Inc.

Abstract

We report results from an exploratory analysis examining "last-minute" self-censorship, or content that is filtered after being written, on Facebook. We collected data from 3.9 million users over 17 days and associate self-censorship behavior with features describing users, their social graph, and the interactions between them. Our results indicate that 71% of users exhibited some level of last-minute self-censorship in the time period, and provide specific evidence supporting the theory that a user's "perceived audience" lies at the heart of the issue: posts are censored more frequently than comments, with status updates and posts directed at groups censored most frequently of all sharing use cases investigated. Furthermore, we find that: people with more boundaries to regulate censor more; males censor more posts than females and censor even more posts with mostly male friends than do females, but censor no more comments than females; people who exercise more control over their audience censor more content; and, users with more politically and age diverse friends censor less, in general.

Introduction

Self-censorship is the act of preventing oneself from speaking. Important in face-to-face communication, it is unsurprising that it manifests in communications mediated through social networking sites (SNS). On these venues, self-censorship may be caused by artifacts unique to, or exacerbated by, social media. For example, users may seek to maintain presentation of their self-images across multiple social contexts simultaneously, may be unwilling to diverge from the community's perceived social norms (such as avoiding negative expressions), or may fear "spamming" friends with uninteresting or unnecessary content (Frederic & Woodrow 2012; Sleeper et al., 2013; Tufekci 2007; Wisniewski, Lipford & Wilson 2012).

Social media also affords users the ability to type out and review their thoughts prior to sharing them. This feature adds an additional phase of filtering that is not available in face-to-face communication: filtering after a thought has been formed and expressed, but before it has been shared. Filtering at this phase is what we call *last-minute self-censorship* because it occurs at the last minute, whereas other lower-level forms of self-censorship might prevent a user from thinking or articulating thoughts at all. Hereafter, we may refer to last-minute self-censorship simply as self-censorship, but one should keep the distinction in mind. Last-minute self-censorship is of particular interest to SNSs as this filtering can be both helpful and hurtful. Users and their audience could fail to achieve potential social value from not sharing certain content, and the SNS loses value from the lack of content generation. Consider, for example, the college student who wants to promote a social event for a special interest group, but does not for fear of spamming his other friends—some of who may, in fact, appreciate his efforts. Conversely, other self-censorship is fortunate: Opting not to post a politically charged comment or pictures of certain recreational activities may save much social capital.

Understanding the conditions under which censorship occurs presents an opportunity to gain further insight into both how users *use* social media and how to *improve* SNSs to better minimize use-cases where present solutions might unknowingly promote value diminishing self-censorship. In this paper, we shed some light on which Facebook users self-censor under what conditions, by reporting results from a large-scale exploratory analysis: the behavior of 3.9 million Facebook users. As the first study of its kind at this scale, we were motivated to (1) understand the magnitude of self-censorship on Facebook, (2) identify patterns of self-censorship that are exhibited across different products (groups vs. status updates vs. comments, etc.), and (3) construct a data-driven model of self-censorship based on behavioral, demographic, and social graph features of users.

We found that 71% of the 3.9 million users in our sample self-censored at least one post or comment over the course of 17 days, confirming that self-censorship is common. Posts are censored more than comments (33% vs. 13%). Also, we found that decisions to self-censor content strongly affected by a user's perception of audience: Users who target specific audiences self-censor more than users who do not. We also found that males censor more posts, but, surprisingly, also that males censor more than females when more of their friends are male. Additionally, we found that people with more boundaries to regulate censor more posts; older users censor fewer posts but more comments; and, people with more politically and age diverse friends censor fewer posts.

Related Work

Self-censorship on social media has scarcely been studied in its own right, but many fields have alluded to the phenomenon. It is known that people present themselves differently in distinct social circle (Goffman 1959). Researchers have also found that the dynamic "presentation-of-self phenomenon" extends into the virtual world, noting that users manage multiple identities when sharing online and behave differently based on the (virtual) social group with which they are sharing (Farnham & Churchill 2011). In face-to-face communication, totally distinct social circles are rarely collocated, so people have a reasonable expectation of their audience and therefore can easily portray themselves appropriately. Conversely, social networking sites collapse many distinct social contexts into one. Consequently, several researchers have observed that SNS users often have trouble with "boundary regulation," or maintaining consistency of presentation across the "boundaries" of multiple social contexts (Acquisiti & Gross 2006; Frederic & Woodrow 2012; Kairam et al. 2012; Marwick & boyd 2010; Wisniewski, Lipford & Wilson 2012).

Self-censorship in social media, then, can be construed as a boundary regulation strategy. Users who experience episodes of "regret" for sharing content that is inappropriate for parts of their audience might resort to self-censorship to avoid repeating a similar episode (Wang et al. 2011). Users may also self-censor as a strategy for managing group copresence in social media (Marwick & boyd 2010; Wisniewski, Lipford & Wilson 2012), perhaps only sharing content that would be reasonable for the "lowest common denominator"—content that would be appropriate for *any* of the user's distinct social circles. Others have observed that to deal with boundary regulation, users "imagine" an audience upon sharing content, and that this imagined audience modulates user-level self-censorship: If some of the "imagined" audience is not "appropriate," users are likely to censor themselves (Marwick & boyd 2010).

Beyond self-presentation, others have noted that "privacy" concerns (defined by having *unknown* or potentially inappropriate audiences gain access to a user's content) may lead to censorship as well (Acquisiti & Gross 2006; Tufecki 2007). Though users generally tend to underestimate their actual audience (Bernstein et al. 2013), those aware of potential privacy breaches may utilize SNS tools such as Facebook privacy settings to avoid content withdrawal or to allow for less vigilance, thus reducing self-censorship. Recent work on the sharing behaviors of Google+ users seems to corroborate this claim, noting that Google+ users construct and utilize "circles" to selectively share or exclude content from individuals (Kairam et al. 2012). Taken together, one might reasonably expect that users who utilize such selective content sharing tools might self-censor less.

Perhaps the most directly related work is a qualitative examination of the motivations users stated for their self-censorship on Facebook (Sleeper et al. 2013). Through a diary study of 18 Facebook users (who reported on content that they "thought about" posting but ultimately did not post), the authors found that users' reasons for self-censoring content could be classified into one of five categories: (1) users did not want to instigate or continue an argument; (2) users did not want to offend others; (3) users did not want to bore others; (4) users did not want to post content that they believed might be inconsistent with their self-representations; and (5) users neglected to post due to technological constraints (e.g., inconvenience of using a mobile app). The first four of these reasons align with our concept of censorship as an audience control strategy.

The first two categories, concerns about offending ties or instigating arguments, map well to the findings of Hayes, Scheufele & Huge (2006), which notes that people in a polarized opinion climate may refrain from participating in discourse out of fear of criticism or disagreement. Indeed, the first two reasons are reported to be especially prevalent for political content (Sleeper et al. 2013). The third category is related to the "imagined audience" described by Marwick & boyd (2010). Surely, the self-censoring user had an audience in mind when she decided that the content was too boring. Others have noted a similar effect, mentioning that the users in their study "indicated a concern for not overwhelming audience information streams" (Frederick & Woodrow 2012). The fourth reason, and perhaps the most prevalent (Sleeper et al. 2013), is consistent with the findings of the boundary regulation literature and once again suggests that maintaining consistency across social contexts is important to users (Marwick and boyd 2013; Wisniewski, Lipford & Wilson 2012). Finally, as most of their participants either did not utilize or were unaware of audience management tools on Facebook, participants would have censored as much as 50% less often if they could specifically and easily share and/or prevent certain ties in their network from viewing the content (Sleeper et al. 2013). These findings stress the importance of "audience" control in content sharing decisions.

However, while self-censorship in social media is noted, we have not yet quantified its prevalence. Furthermore, it is unknown how self-censorship manifests across different use cases of social media: Do users self-censor undirected status updates more than posts directed at a specific friend's timeline? Finally, while there are many hypotheses regarding which users self-censor more or less, there has yet to be any empirical validation of these findings at a large scale. We offer insight into these gaps in the literature.

Methodology

Broadly, our analysis was guided by three questions: (1) How many people censor content, and how often? (2) What are the patterns of self-censorship exhibited across Facebook and how do they differ? And, (3) What factors are associated with being a more frequent self-censor?

For our purposes, we operationalize "self-censorship" as any non-trivial content that users *began* to write on Facebook but ultimately did not *post*. This method is fast and

Demographic	Behavioral	Social Graph
Gender [GEN]	Messages sent [AUD]	Number of friends [DIV]
Age [AGE]	Photos added [CTRL]	Connected components [DIV]
Political affiliation [CTRL]	Friendships initiated [CTRL]	Biconnected components [DIV]
Media (pic/video) privacy [CTRL]	Deleted posts [CTRL]	Average age of friends [AGE]
Wall privacy [CTRL]	Deleted comments [CTRL]	Friend age entropy [DIV]
Group member count [AUD]	Buddylists created [CTRL]	Mostly (conservative/liberal/moderate) Friends [CTRL]
Days since joining Facebook [CTRL]	Checkins [CTRL]	Percent male friends [GEN]
	Checkins deleted [CTRL]	Percent friends (conservative/liberal/moderate) [DIV]
	Created posts [CTRL]	Friend political entropy [DIV]
	Created comments [CTRL]	Density of social graph [DIV]
		Average number of friends of friends [DIV]

Table 1. Main effect features in our model. Columns represents what aspect of the user the feature describes, codes in square brackets represent high level categories to which the features belong: GEN are features related to gender, AGE are features related to age, AUD are features related to audience selection, DIV are features related to social graph diversity, EXP are features related to a user's length of experience with Facebook, and CTRL are other control features that we account for in our models.

lightweight enough to not affect users' experience of Facebook. We also believe that this approach captures the essence of self-censorship behavior: The users produced content, indicating intent to share, but ultimately decided against sharing. Note that we do not claim to have captured all self-censoring behavior with this method. The user authored content, but censored it at the last minute.

Computing a Measure of Self-Censorship

This research was conducted at Facebook by Facebook researchers. We collected self-censorship data from a random sample of approximately 5 million English-speaking Facebook users who lived in the U.S. or U.K. over the course of 17 days (July 6-22, 2012). This time period was selected for pragmatic purposes, as a summer intern conducted this work. Censorship was measured as the number of times a user censored content over the 17-day period. We define *posts* as threads initiated by the user (e.g., status updates or posts on a friend's timeline) and *comments* as responses to extant threads (e.g., replies to a status update or timeline post). We aggregated censorship data separately for posts and comments, under the intuition that the two forms of content represent sufficiently distinct use cases. We further collected data about where the censorship occurred (e.g., on a user's own timeline or on a friend's), hypothesizing that censorship may vary across these domains.

To measure censorship, we instrumented two user interface elements of the Facebook website, shown in Figure 1: the "composer"—the HTML form element through which users can post standalone content such as status updates—and the "comment box"—the element through which users can respond to existing content such as status updates and photos. To mitigate noise in our data, content was tracked only if at least five characters were entered into the composer or comment box. Content was then marked as "censored" if it was not shared within the subsequent ten minutes; using this threshold allowed us to record only the *presence or absence* of text entered, not the keystrokes or content. If content entered were to go unposted for ten minutes and then be posted, we argue that it was indeed censored (albeit temporarily). These analyses were conducted in an anonymous manner, so researchers were not privy to any specific user's activity. Furthermore, all instrumentation was done on the client side. In other words, the content of self-censored posts and comments was not sent back to Facebook's servers: Only a binary value that content was entered at all.

Figure 1. Screenshot of the "composer" (top) and comment box (bottom) HTML elements on Facebook.

Model Features

For each user, we gathered or calculated features that were hypothesized to affect self-censorship. These features can broadly be categorized into three distinct buckets: (1) demographic and (2) behavioral features of the user, and (3) aggregate demographic and differential features of the user's social graph (e.g., average number of friends of friends, the user's own political ideology compared to that of his or her friends). The main effect features, all included as controls, are listed in Table 1. As most of the features are self-explanatory, for brevity, we will talk about the meaning of non-obvious features as they become more pertinent to the discussion.

In addition to these main effect features, we explored three interactions: (1) *User's political affiliation given the modal affiliation of the user's friends* (e.g., the effect of being a conservative when most of your friends are liberal), (2) *political affiliation given the political entropy of friends* (e.g., the effect of being a liberal given a politically homogenous or heterogeneous social graph), and (3) *gender given the gender diversity of the social graph*.

We selected variables by applying the following considerations: (1) *accessibility*—we selected features that were easily accessible; (2) *computational feasibility & reliability*—we selected features that could be computed reliably given the available data (some users freely provide their

political affiliation, as part of their profile, but there is no way to specify one's race); (3) *interpretation & model parsimony*—we selected features and interactions which were open to straightforward interpretation and which we believed would have a straightforward impact on self-censorship based on our own intuitions and the background literature. Interactions were selected after carefully examining the main effects of the model. Together, this analytic approach is designed to be primarily exploratory, as this construct has not yet been examined. However, we do form hypotheses in order to give this measure of self-censorship context within the current literature.

Hypotheses

As posts require substantially more effort to construct and users who are concerned with posting "uninteresting" content should also be more likely to censor posts than comments, we expected **H1: posts will be censored more than comments.**

The literature on the impact of gender on self-disclosure suggests that women are more comfortable with disclosing than men (Cho 2007; Dindia & Allen 1992; Seamon 2003). In turn, greater propensity for self-disclosure is likely inversely related to propensity to self-censor, because disclosing more information online is an indicator of greater comfort with sharing. Taken together, we predicted **H2: men will self-censor more than women**. Furthermore, we know from (Dindia & Allen 1992) that both men and women self-disclose more to same-sex partners, while (Walther 2007) noted that users of CMC technologies work harder and edit their writing more when it is directed at the opposite sex than at the same sex. These findings suggest **H3: users with more opposite-sex friends will self-censor more**.

Younger SNS users have been suggested to be "more motivated for publicity and more willing to give up their privacy" (Tufecki 2007) and older users, even within a youth-skewed sample, have been shown to find social media more cognitive demanding suggesting **H4: younger users will self-censor less.** We also suspected **H5: users with older friends will censor more**, based on the intuition that older audiences are likely to be more judgmental of what constitutes "appropriate" shared content.

It is also often suggested that imagined audience and presentation of self across boundaries is important to users (Frederic & Woodrow 2012; Marwick & boyd 2010; Sleeper et al. 2013; Wisniewski, Lipford & Wilson 2012) and that the perceived lack of boundary control is a primary cause for "self-censorship" on Facebook (Sleeper et al. 2013). Balanced with the observation that many users desire larger audiences than they perceive they have for the content they share (Bernstein et al. 2013), we expected **H6: users who more frequently used audience selection tools self-censor less**. Audience selection tools include usage of Facebook groups, buddy lists and private messages.

Relatedly, because maintaining presentation of self across social contexts might get harder with more social contexts and because more diverse social graphs likely come with tumultuous opinion climates that can cause users to self-censor more for fear of criticism (Hayes, Scheufele & Huge 2006), we expected **H7: users with more diverse friends will self-censor more**.

Results

User Demographics and Descriptive Statistics

We were able to derive all but the politics metrics listed in Table 1 for 3,941,161 users, out of the initial set of five million. We were able to derive the politics metrics for a smaller subset of 322,270 users within the United States, and use this subset of users in analyses where those mertics become pertinent. Users were young, with an average age of 30.9 years (sd 14.1). Furthermore, our population was comprised of 57% females. Experience with Facebook was roughly normally distributed with a mean of 1,386 days.

Magnitude and Patterns of Self-Censorship

Over the 17-days, 71% of all users censored content at least once, with 51% of users censoring at least one post and 44% of users censoring at least one comment. The 51% of users who censored posts censored 4.52 posts on average, while the 44% of users who censored comments censored 3.20 comments on average.

Figure 2 shows the distribution of comments and posts censoring rates. The two distributions are similar, with notable spikes at both 0% and 100%; this spike can be explained by the short, finite time interval over which we collected data. In other words, some users had a very small denominator: A user who only had one potential post would escore 100% or 0% if they censored it or not, respectively.

Figure 3 shows a breakdown of how post and comment censorship rates vary across popular Facebook products and use cases. For posts, users tended to censor their own status updates (34%) and group posts (38%) more than posts on their friends timelines (25%) or event timelines (25%). Most of these numbers agree with **H6**, that posts with vaguer audiences—the audience agnostic posts such as status updates—will be censored more. However, there was an exception: Group posts were censored most frequently. For comments, users tended to primarily censor comments on photos (15%) and group messages (14%) more than they censored comments on timeline posts (11%) and status updates (12%). That comments on photos are censored most of all seems to make sense—it may well be that photos elicit comments that users feel more apprehensive about sharing.

Modeling Self-Censorship

We modeled censorship for posts and comments separately because of the large differences in the distribution of censorship between them. Due to the skew in the distributions (see Figure 2), we elected to model censorship rates for posts and comments using a zero-inflated negative binomial regression, or ZINB (Lambert 1992; Mwalili, Lessaffre, & Declerck 2008). Zero-inflation attempts to model the over-abundance of users who censored zero times in the time interval. We used an intercept-only model for zero-

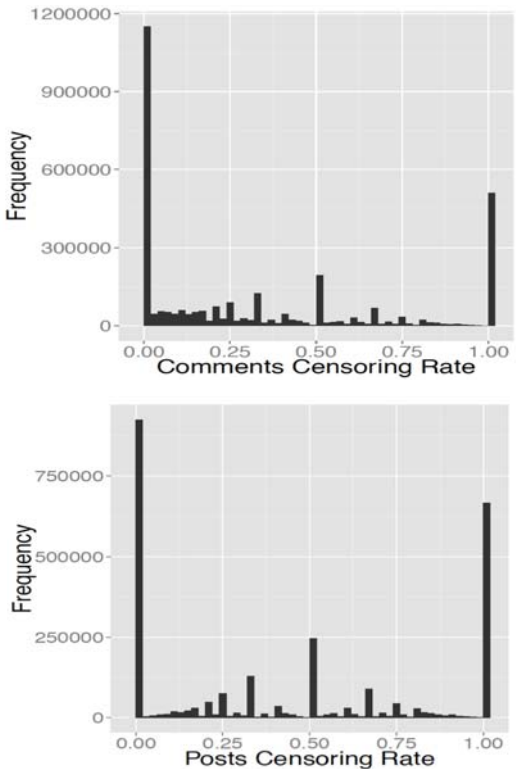

Figure 2. Distribution of comments (top) and posts (bottom) self-censoring rates.

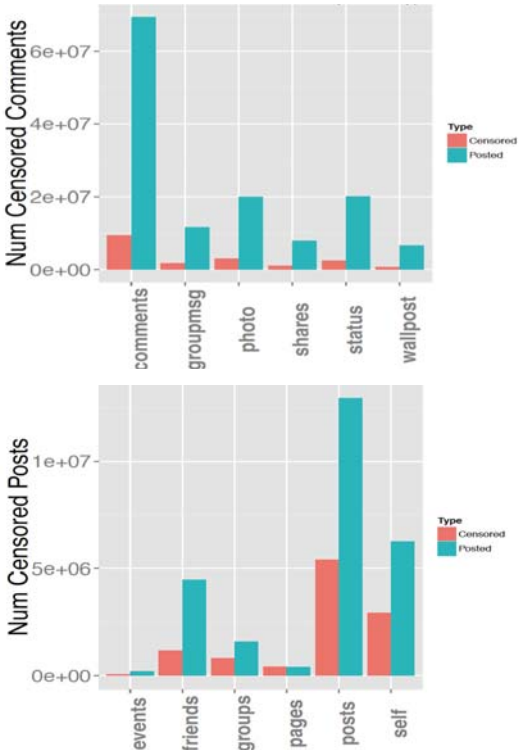

Figure 3. Number of censored (red) and shared (blue) comments (top) and posts (bottom) grouped by location. Location represents where the content was shared (e.g., "groupmsg" means comments on group posts). The "posts" and "comments" locations are the values aggregated across all locations.

inflation, but found that it was insignificant at predicting whether a user who censored zero times was a "true" zero (truly never censors) or a happenstance zero (happened to not censor in the time period we measured). The negative binomial distribution was favored over Poisson because of the presence of overdispersion in the distributions of the response variables (Lambert 1992)—total censored posts for posts, and total censored comments for comments. We also offset the response variables for each model by the total number posts or comments actually posted by the user, as users who post more also have more opportunity to censor, so the frequency of censorship should be proportional to the frequency of posts. The predictor variables we used are described in the "Model Features" section above. This model does not generate an R-squared value, so we report the regression coefficients as our effect size measure.

In the subsections to follow, we will make reference to coefficients in the ZINB model. These coefficients can be read as follows: For numeric values, a 1% increase in the feature value yields a difference in the response equivalent to multiplying the estimated response by the coefficient of the feature, holding all other features at their mean. For categorical variables, all but one of the discrete values of the variable are compared to a baseline value (the value that is left out). All reported coefficients are significant at the p=.05 level; nonsignificant results are reported as such.

For example, consider the *Age* feature, with a coefficient of 0.85 in the posts model. If x is the baseline estimate for censored posts when all predictors are held at their mean, a 1% increase in the feature results in an estimate of 0.85x. In other words, a 1% increase from the mean of the number of days a user has been on Facebook indicates that the user censors 85% as many (i.e., 15% fewer) posts. Similarly, the categorical feature, *Gender: Male*, has a coefficient of 1.26. Thus, compared to the baseline users who have the feature *Gender: Female*, users with the feature *Gender: Male* censor 1.26 times as much (i.e., 26% more).

The regression results we share below are abbreviated for brevity, but the full regression table can be found at: http://sauvik.me/icwsm_selfcensor.

Hypotheses Revisited

Pertinent model coefficients for the ZINB are presented in Table 2 for the posts model and Table 3 for the comments model. Posts were censored at a higher rate than comments (33% vs 13%, χ^2=4.3e6, p=2e-16), lending credence to **H1**. From Table 2, we also see that males censored more posts (coeff. 1.26—26% more) than females, but that males censored even *more* posts than females as the proportion of their male friends increased (coeff. 1.11). However, we do not observe either of these effects for comments—gender does not affect censorship of comments. Thus, **H2** is supported for posts but unsupported for comments, and **H3** is unsupported for both.

For posts, older users seemed to censor substantially fewer posts than younger users (coeff. 0.85—15% less per 1% increase in age from the mean). The opposite is true for comments, however (coeff. 1.11). Thus, **H4** is refuted for posts but supported for comments. Also, while the average age of one's friends was unrelated to posts censorship, we see that users with older friends censored fewer comments (coeff. 0.87) refuting **H5** for posts and comments.

Curiously, users who were part of more groups censored more posts (coeff. 1.29) and comments (coeff. 1.14). Similarly, users who used another audience selection tool, buddy lists, censored more posts (coeff. 1.13) and comments (coeff. 1.07). One notable exception to the trend was the use of private messaging, which was associated with lower censorship for comments (coeff. 0.92). One explanation for this finding is that comments are flat, coercing many branching discussions into a single thread; so, any one comment might be perceived as irrelevant to other discussions that are occurring within the thread. To more specifically target an audience, then, users might utilize private messaging. Nevertheless, combined with our previous finding that posts and comments directed at specific groups were censored substantially more than the mean, it seems that **H6** is unsupported: users who are more aware of audience censor more, not less.

Diversity (**H7**) had a mixed effect. A greater number of distinct friend communities—which we took as a measure of diversity—predicted for increased censorship. Users with a higher average number of friends-of-friends, a correlate of distinct communities in one's extended friend network, censored more posts (coeff. 1.29). Similarly, users with more biconnected components, the number of 2-connected subgraphs in one's social graph that we used as rough measure of distinct friend groups (Ugander et al. 2012), censored more posts (coeff. 1.12). Furthermore, users with a higher friendship density, or users with more friends who were friends with one another, censored less (coeff. 0.97) as is consistent with prior findings because higher density social networks should have fewer separate communities. The effect was more slight and tumultuous for comments, however. While users with more biconnected components did predict for more comments censorship (coeff. 1.03), users with more friends of friends had the opposite, though similarly small (coeff. 0.95), effect. This finding might suggest that users are apprehensive towards generating new content towards large, impersonal audiences but are willing to join ongoing discussions within such communities.

Alternatively, users with more diverse friends generally censored less. The presence of more politically identified friends, which we took as a measure of diversity because most users do not have many friends who share their political identification on Facebook, predicted for dramatically reduced self-censorship for both posts (coeffs. 0.77, 0.77, 0.95 for liberal, conservative, moderate) and comments (coeffs. 0.91, 0.87, 0.95 for liberal, conservative, moderate). Users with more politically diverse friends also censored fewer posts (coeff. 0.92) and comments (coeff. 0.94). Finally, users with more age diverse friends censored fewer posts (coeff. 0.96), but more comments (coeff. 1.17).

Taken together, it seems that "diversity" has two effects. Users with more distinct communities represented in their social graph self-censored more, whereas users with more diverse friends—of varied ages and political affiliations—self-censored less. The one exception to this rule is that users with more age diverse friends censor more comments. One possible explanation for the exception is that users may not know how to appropriately respond to discussions started by friends from different age generations (Sleeper et al. 2013). Thus, support for **H7** is mixed, suggesting a need for future research to tease apart these dual aspects of friendship diversity: diversity of community membership and diversity of individual friend features.

Discussion

While 71% of our users did last-minute self-censor at least once, we suspect, in fact, that all users employ last-minute self-censorship on Facebook at some point. The remaining 29% of users in our sample likely didn't have a chance to self-censor over the short duration of the study. Surprisingly, however, we found that relative rates of self-censorship were quite high: 33% of all potential posts written by our sample users were censored, and 13% of all comments. These numbers were higher than anticipated and are likely inflated with false positives because of the imprecise nature of our metric. Nevertheless, our metric should be strongly correlated with self-censorship, so while the exact numbers we report might be rough, self-censorship on Facebook does seem to be a common practice. Furthermore, the frequency of self-censorship seems to vary by the nature of the content (e.g., post or a comment?) and the context surrounding it (e.g., status update or event post?).

The decision to self-censor also seems to be driven by two simple principles: People censor more when their audience is hard to define, and people censor more when the relevance or topicality of a CMC "space" is narrower. For example, posts are unsurprisingly censored more than comments. After all, posts create new discussion threads over which the user claims ownership, are more content-rich, tend to require more energy and thought to craft, and require more effort to share, as users have to explicitly hit a "submit" button. However, posts also make it hard to conceptualize an "audience" relative to comments, because many posts (e.g., status updates) are undirected projections of content that might be read by anyone in one's friend list. Conversely, comments are specifically targeted, succinct replies to a known audience. Even groups of users who are known to be comfortable with more self-disclosure are often only comfortable with such disclosure to a well-specified audience, such as close relationships (Dindia & Allen 1992; Bowman, Westerman & Claus 2012), so it makes sense that users pay special attention to posts to ensure that the content is appropriate for the "lowest common denominator" (Wisniewski, Lipford & Wilson 2012).

Type	Feature	Coefficient
GEN	Gender: Male (#)	1.26 ***
GEN	Gender: Male X Percentage male friends (##)	1.11 ***
AGE	Age	0.85 ***
AUD	Group member count	1.29 ***
AUD	Buddylists created	1.13 ***
DIV	Average number of friends of friends	1.32 ***
DIV	Biconnected components	1.12 ***
DIV	Percentage friends liberal	0.77 ***
DIV	Percentage friends conservative	0.77 ***
DIV	Percentage friends moderate	0.95 ***
DIV	Friend political entropy	0.92 ***
DIV	Friendship density	0.97 ***
DIV	Friend age entropy	0.96 **

(#) baseline: female, (##) baseline: female x percentage male friends
*** p < 0.01, *** p < 2e-16*

Table 2. Coefficients for the ZINB model of Posts Censorship. For categorical variables, baselines are explained at the bottom. Interactions are specified with an "X".

Clarity of audience is not the only factor that influences sharing decisions, however. For example, posts directed only towards members of a specific group were censored substantially more than posts on events and on friends' timelines. This finding was surprising because groups provide users with a quick and easy way to target a specific audience with known interests and/or expertise—a strategy that is often considered an alternative to self-censorship in prior work (Acquisiti & Gross 2006; Marwick & boyd 2010; Wisniewski, Lipford & Wilson 2012). However, knowing one's audience is only one part of the battle—a known audience is a double-edged sword: Topics relevant to the group may be easier to post due to the established audience, but far fewer thoughts, statements, or photos are relevant to group. This finding is further corroborated by the observation that users who used more audience selection features (e.g., were members or more groups and created buddylists to specifically share or exclude sharing content with) actually censored more posts and comments than other users.

User-specific factors also seemed to impact the frequency of self-censorship. The background literature was able to correctly predict the impact of some of these factors, but was just as often wrong. For example, we found that males censor more posts than females—an outcome we expected given the knowledge that males tend to be less comfortable with self-disclosure (Dindia & Allen 1992; Seamon 2003). However, males with more male friends censored more posts than females with a comparable proportion of male friends—a finding that we did not expect given prior findings that users of CMC technologies work harder and edit their writing more when it is directed at the opposite sex than at the same sex (Walther 2007). Other unexpected results include older users censoring fewer posts but more comments than younger users, and users with older friends censor no more and no fewer posts but fewer comments than other users. Further research will be required to better understand these effects.

Type	Feature	Coefficient
AGE	Age	1.11 ***
AGE	Friends average age	0.87 ***
AUD	Group member count	1.14 ***
AUD	Buddylists created	1.07 ***
AUD	Messages sent	0.92 ***
DIV	Friends age entropy	1.17 ***
DIV	Biconnected components	1.03 ***
DIV	Average number of friends-of-friends	0.95 ***
DIV	Percentage friends moderate	0.95 ***
DIV	Friend political entropy	0.94 *
DIV	Percentage friends liberal	0.91 ***
DIV	Percentage friends conservative	0.87 ***

*** p < 0.05, *** p<2e-16*

Table 3. Coefficients for the ZINB model of Comments Censorship.

We also found support for prior work suggesting that self-censorship might be a boundary regulation strategy (Sleeper et al. 2013; Wisniewski, Lipford & Wilson 2012), even after controlling for privacy settings and usage of audience selection tools. Users part of a larger number of distinct friend communities and who had a larger extended friend network censored substantially more. This finding suggests that present audience selection tools and privacy settings on Facebook are not very effective at mitigating self-censorship that results from boundary regulation problems (i.e., refraining from posting content in fear that an inappropriate community of friends might it). On the other hand, users with a diverse set of friends in fewer distinct communities actually self-censor less, suggesting that Facebook users more often initiate or engage in ongoing discussions with a diverse audience over a homogenous one. Thus, users with a small set of diversely populated groups of friends seem to self-censor least of all. Future work will be necessary to understand why we observe this pattern.

Recall that one of our motivations in understanding the phenomenon of self-censorship in social media is to understand when it is adaptive. This question is still open, though through this work, we have developed a better understanding of self-censorship behaviors on Facebook. However, it would be inappropriate to optimize against the metric we present in this paper because it is too general. Rather, it would be pertinent to optimize against *undesirable* instances of self-censorship, as in the case of the college student who avoids posting a status update directed at a special interest group because she is afraid of spamming friends outside of that group. At this point, we have evidence that self-censorship is motivated in part by concerns regarding an audience, suggesting that cleaner and easier-to-use audience selection tools are desirable. Future work is necessary to further understand when self-censorship is adaptive.

There are also some biases in our sample that prevent us from over-generalizing. The short time period over which we collected data suggests that our sample comprises relatively active Facebook users. Furthermore, by only modeling those users for whom we could gather the demographic information in Table 1, we may have biased our sample

towards users who are comfortable with personal disclosure. We also do not make claims of causality. The findings we present are strictly observational correlations. Experimental approaches are necessary to determine that diversity is a cause of self-censorship, rather than a correlate. For example, if people who censor more come off as more likeable and less offensive, they might be seen as more broadly attractive, resulting in a broader group of friends. Indeed, the role of censorship in relationship building also provides an interesting avenue of study.

Nevertheless, we now know that current solutions on Facebook do not effectively prevent self-censorship caused by boundary regulation problems. Users with more boundaries to regulate self-censor more, even controlling for their use of audience selection and privacy tools. One reason for this finding is that users might distrust the available tools to truly restrict the audience of a post; another possibility is that present audience selection tools are too static and content agnostic, rendering them ineffective in allowing users to selectively target groups on the fly (Sleeper et al. 2013). Future work will be necessary to unpack the nature of this effect and strategies to improve the tools available to users.

Conclusion

We studied the last-minute self-censorship habits of 3.9 million English speaking Facebook users, and found that a large majority (71%) self-censored content at least once. Decisions to self-censor appeared to be driven by two principles: people censor more when their audience is harder to define, and people censor more when the relevance of the communication "space" is narrower. In other words, while posts directed at vague audiences (e.g., status updates) are censored more, so are posts directed at specifically defined targets (e.g., group posts), because it is easier to doubt the relevance of content directed at these focused audiences.

We also uncovered the relationships between various user-specific factors and the frequency of self-censorship on Facebook. Indeed, as the first study examining self-censorship on Facebook at this scale, we directly supported some prior findings—for example, that males and people with more boundaries to regulate "self-censor" more—while showing findings from other work to be less consistent—for example, that people who exercise more control over their audience with audience selection tools self-censor less. Through this work, we have arrived at a better understanding of how and where self-censorship manifests on social media; next, we will need to better understand what and why.

Acknowledgements

Special thanks go to Michael Bernstein, Moira Burke, A.T. Fiore, Bill Fumerola, Robert Kieffer and Solomon Messing for their help in the ideation, execution and articulation of this research.

References

Acquisti, A. & Gross, R. Imagined Communities: Awareness, Information Sharing, & Privacy on the Facebook. Priv. Enhancing Tech. 4258, (2006), 36–58.

Bernstein, M., Bakshy, E., Burke, M., and Karrer, B. Quantifying the Invisible Audience in Social Networks. Proc. CHI 2013.

Bowman, N.D., Westerman, D.K., & Claus, C.J. How demanding is social media: Understanding social media diets as a function of perceived costs and benefits – A rational actor perspective. Computers in Human Behavior, 28 (2012), 2298.

Cho, S.H. Effects of motivations and gender on adolescents' self-disclosure in online chatting. Cyberpsychology & Behavior 10, 3 (2007), 339–45.

Dindia, K. & Allen, M. Sex differences in self-disclosure: A meta-analysis. Psych. Bulletin 112 (1992), 106.

Farnham, S.D. & Churchill, E.F. Faceted identity, faceted lives. Proc. CSCW 2011, 359.

Frederic, S. & Woodrow, H. Boundary regulation in social media. Proc. CSCW 2012, 769.

Goffman, E. The presentation of self in everyday life. Doubleday Anchor, Garden City, NY, 1959.

Hayes, A.F., Scheufele, D.A., & Huge, M.E. Nonparticipation as Self-Censorship: Publicly Observable Political Activity in a Polarized Opinion Climate. Pol. Behavior 28 (2006), 259.

Kairam, S., Brzozowski, M., Huffaker, D., & Chi, E. Talking in circles. Proc. CHI 2012, 1065.

Lambert, D. Zero-Inflated Poisson With an Application to Defects in Manufacturing. Technometrics 34, (1992), 1.

Lampinen, A., Lehtinen, V., Lehmuskallio, A., & Tamminen, S. We're in it together. Proc. CHI 2011, 3217.

Marwick, A.E. & boyd, d. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. New Media & Society, 13 (2010), 114.

Mwalili, S.M., Lesaffre, E., & Declerck, D. The zero-inflated negative binomial regression model with correction for misclassification: an example in caries research. Stat. meth. med rsrc 17 (2008), 123.

Seamon, C.M. Self-Esteem, Sex Differences, and Self-Disclosure: A Study of the Closeness of Relationships. Osprey J. Ideas & Inquiry, (2003).

Sleeper, M., Balebako, R., Das, S., McConohy, A., Wiese, J., & Cranor, L.F. The Post that Wasn't: Exploring Self-Censorship on Facebook. Proc. CSCW 2013.

Tufekci, Z. Can You See Me Now? Audience and Disclosure Regulation in Online Social Network Sites. Bull. Sci., Tech, & Soc, 28 (2007), 20.

Ugander, J., Backstrom, L., Marlow, C., Kleinberg, J. Structural diversity in social contagion. PNAS. 109 (2012), 5962.

Walther, J.B. Selective self-presentation in computer-mediated communication: Hyperpersonal dimensions of technology, language, and cognition. Comp. in Hum. Behav. 23 (2007), 2538.

Wang, Y., Norcie, G., Komanduri, S., Acquisti, A., Leon, P.G., & Cranor, L.F. "I regretted the minute I pressed share." Proc. SOUPS 2011, 1.

Wisniewski, P., Lipford, H., & Wilson, D. Fighting for my space. Proc. CHI 2012, 609.