The More Social Cues, The Less Trolling? An Empirical Study of Online Commenting Behavior*

Daegon Cho† Alessandro Acquisti‡

H. John Heinz III College, Carnegie Mellon University

June 3, 2013

[PRELIMINARY AND INCOMPLETE DRAFT– PLEASE CONTACT THE AUTHORS FOR THE MOST UPDATED VERSION OF THIS MANUSCRIPT.]

Abstract

We examine how online commenting is affected by different degrees of commenters' identifiability: 1) real name accounts on social networking sites (or "real name-SNS accounts"; e.g., Facebook); 2) pseudonym accounts on social networking sites (or "pseudonym-SNS account"; e.g., Twitter); 3) pseudonymous accounts outside social networking sites (or "non-SNS accounts"; e.g., an online newspaper website's account). We first construct a conceptual model of the relationship between the degree of identifiability and commenting behavior. When users can freely choose the account type between a non-SNS account and an SNS account to write a comment, the decision determines degree of identifiablity. This decision will be correlated to the propensity of using 'offensive' words (classified from a comprehensive list of swear terms in the Korean language) in their comments. To take endogeneity into consideration, we estimate a bivariate probit model between the choice of account type and the choice of using offensive words. We apply our model to a unique set of data consisting in over 75,000 comments attached to news stories collected from a variety of online news media websites. Our analysis highlights interesting dynamics between the degree of identifiability and commenters' behavior. When commenters use an SNS account (which consists in a more identifiable condition) rather than a non-SNS account, they are less likely to use offensive words. We also find that the use of real name-SNS accounts (which provide an even more identifiable condition due to the disclosure of one's real name), is associated with lower occurrence of offensive words than the case in which commenters use a pseudonym-SNS account for commenting. While the disclosure of true identity is likely to reduce the probability of using offensive words, the greater number of users seems to prefer participating in the commenting activity by using their pseudonym accounts.

Keywords: Online anonymity, pseudonym, true identity, online comment, social networking site, social commenting system.

---

* An author, Daegon Cho, thanks Jae Yeon Kim for helpful comments and discussions as well as generous support for this study. The author is also thankful to Beomjin Kim, Mikyun Kim and staffs in CIZION for sharing and refining the data for this study and to Byeong Jo Kim and Seokho Lee for sharing their expertise in the news media industry.

† Corresponding author. Email: daegonc@andrew.cmu.edu

‡ Email: acquisti@andrew.cmu.edu

1 Introduction

Most online news providers nowadays have established commenting services. As reported by the American Society of News Editors (ASNE)'s survey in 2009, 87.6% of newsrooms enabled users to post online comments regarding specific stories. Adding a commenting system can yield higher advertising revenues by increasing the number of page views.[1] Through commenting platforms, users can post their views, read and discuss other users' comments, or even vote "like" or "dislike" to those comments. Users appear to appreciate these features. According to a 2010 survey report from the Pew Internet & American Life Project,[2] 25% of Internet users in the United States have commented on a news article[3]; in addition, over three million comments per month are posted at HuffingtonPost.com as of 2011 (Diakopoulos and Naaman 2011).

Managing commenting systems, therefore, has become more important over time. According to the 2009 ASNE survey, 38.9% of respondents reported they have closed at least one comment thread for a specific story due to undesirable trolls and cyberbullies within the past year.[4] Not surprisingly, offensive comments (e.g., flames, swear words and provocative languages) can negatively affect other users' experience and consequently cause damage to the news outlets (Raskauskas and Stolz 2007).

Online news comments can be moderated in a variety of ways. One involves using automated filtering systems to block comments including swear words or fowl language. This approach, however, is not always adequate – for instance, comments on sensitive topics (such as politics or religion) may be offensive without actually containing offensive terms. An alternative solution relies on crowdsourcing (Mishra and Rastogi 2012), letting the set of commenters' self-discipline by – for instance – upvoting or downvoting a comment. Another approach consists

[1] For example, according to an April 2010 column by Washington Post's ombudsman, Andrew Alexander, "the growth [in online comments] is critical to The Post's financial survival in the inevitable shift from print to online."

[2] Source: http://pewinternet.org/Reports/2010/Online-News.aspx

[3] According to the same survey, 37% of online news users (and 51% of 18-29 year olds) think that commenting on news stories is an important feature.

[4] According to survey, primary reasons for shutdowns of comments are: (1) discriminatory comments involving race, ethnicity, gender or sexual orientation, (2) hurtful comments and (3) obscenities, profanities, foul language. (Source: http://tae.asne.org/Default.aspx?tabid=65&id=458)

in forcing commenters to publicly and personally identify themselves – under the expectation that public identification may lead to more civil discourse.

Online interactions can be indeed different from offline communications in many aspects; one of those is the ability to remain anonymous. In this respect, the degree of identity disclosure may well play a pivotal role in influencing online commenting participation and behaviors. The issue is how – and the literature in this area provides contrasting evidence.

For instance, strict identity verification policies (i.e., the absence of anonymity) could deter users' online participation.[5] In contrast, some studies paradoxically highlighted that highly anonymous conditions can discourage voluntary contributions (because individuals are less motivated in the absence of social interactions and recognitions by others: see Andreoni and Petrie 2004). In addition, elements of anonymity may or may not produce a high likelihood of antinormative behaviors[6] (Postmes and Spears 1998; Suler 2005). These noticeable nuances arising from both academic studies and anecdotal evidence[7] suggest, at the very least, that online news organizations' choice between anonymous, pseudonymous, or fully identified commenting systems may have significant effects on readers' choice to participate in them and on their subsequent commenting behavior.

Several online news media, blog-publishing services, and online forum services have recently moved away from anonymous commenting systems.[8] In fact, an increasing number of news organizations have adopted "social commenting systems" through which users' comments are linked to their accounts on social network sites (SNS). This transition raises interesting issues: (1) users may or may not be concerned over how their comments on a news side will reflect on their social image associated to their SNS accounts; (2) other readers interested

[5]Cho and Kim (2012) studied the impact of real name verification policy in South Korea. Their finding suggests that the policy significantly reduced user participations compared to a period in which the law was not in place.

[6]According to Postmes and Spears (1998), antinormative behaviors are defined in relation to general norms of conduct rather than specific situational norms. In this broad respect, the occurrence of deindividuated behaviors can be regarded as antinormative behaviors.

[7]Anecdotal evidence suggests that the quality of pseudonymous comments is higher than comments by completely anonymous users *or* users with real names, implying that a certain level of identity protection may provide positive outcomes by fostering more intimate and open conversations (see : http://www.poynter.org/latest-news/mediawire/159078/people-using-pseudonyms-post-the-most-highest-quality-comments-disqus-says/)

[8]See for more information: http://www.nytimes.com/2010/04/12/technology/12comments.html?ref=media

in who wrote a particular comment may visit the commenter's SNS profile and further communicate with the commenter; and (3) a commenter's offline true identity is more likely to be disclosed to other readers through her real name and her activities presented in her SNS. In short, connecting an SNS account to a reader's online commenting significantly alters her expectation of anonymity and in turn affects her commenting behavior.

Table 1 presents current features of the major news organizations' commenting systems, suggesting significant heterogeneity in terms of functions and policies. While both the Wall Street Journal (WSJ) and the New York Times (NYT) run proprietary platforms, the WSJ holds a strict real name policy, which is not the case of the NYT. CNN and TechCrunch instead have adopted a third-party platform and Facebook commenting system.

News Medium	Platform Type	Real Name Policy	Selected Functions Available
The Wall Street Journal	Proprietary	Yes	Recommend, Subscriber badge
The New York Times	Proprietary	No	Recommend, Flag, Share with SNS
Huffington Post	Porprietary	No	Badge, Fans, Permanent Link, Share
CNN	Disqus	N/A*	Vote Up (or Down)
NPR	Disqus	N/A*	Vote Up (or Down)
TechCrunch	Facebook	N/A*	Like, Mark as spam, Top commenter
Los Angeles Times	Facebook	N/A*	Top commenter, Like, Follow post
Slashdot	Proprietary	No	Scoring by peer rating
BBC	Proprietary	No	Editor's Picks, Vote

*User may use either pseudonym or real name according to their preference. Users of this commenting platform may not hold multiple pseudonyms, because it would be costly to change it frequently.

Table 1: Commenting Platform Examples of Major Global Online News Websites

In this manuscript, we empirically examine how different degree of commenters' identifiability affects their commenting behavior on news sites. We use a unique data set of social commenting system attached to news stories by which users can freely choose either non-SNS account or SNS account for commenting. Specifically, we focus on the relationship between the user's account choice and their commenting behavior. Answering that question may not only help better understand anonymity-related user behavior, but also contribute to an untested and novel debate: how can online news organizations facilitate user participations and lead them to behaving more discreetly? Throughout this paper, we define antisocial behaviors as comments

that include designated offensive expressions such as swear words.[9]

We begin our analysis by proposing a conceptual model of the relationship between the degree of identifiability and commenting behavior. We apply the model to a large data set of over 75,000 comments written by over 23,000 commenters on a number of online news media websites. The data was collected from the largest third-party online commenting platform provider in South Korea. Online news websites equipped with the commenting system equally allow users to choose one of three types of accounts for posting comments.[10] In other words, to write a comment users have to sign-in by choosing one of the following account types: (1) a non-SNS account (e.g., a news website's account); (2) a pseudonym-SNS account (e.g., Twitter); and (3) a real name-SNS account (e.g., Facebook). Hence, this data set includes comments (and commenters) that may be substantially less identifiable (when a commenter uses a non-SNS account without the display of real name), identified either under pseudonyms or the user's real name (when a commenter uses an SNS account). The users' account choice is likely to affect their amount of disclosure (of personal identifiable information) and self-disclosure in posted comments; as a result, their commenting behavior may be related to their choice of account type, which determines the degree of identifiability.

To take this important aspect into consideration, we employ a bivariate probit model that allows us to estimate parameters in the consideration of interdependent decisions by the same actor. By using this empirical approach, not only does this approach account for correlation between the account choice and the commenting behavior, but we can also compute conditional and marginal effects of parameters of interest.

Our main results are as follows. We show that, when a commenter uses an SNS account (which provides a higher degree of identifiability), they are less likely to use offensive words and expressions such as swear words. On the other hand, we find that the use of a real name-

[9]To define antisocial behaviors, we conduct content analysis to check whether a comment includes offensive words or not. More details will be described in following sections.

[10]We consider online news websites that adopted the third-party commenting system in our analysis. Since some other news websites in South Korea operate their own proprietary commenting systems such as the WSJ and the NYT, our sample represents a fraction of the entire domestic news websites available in South Korea. This fact may cause a selection bias in empirical analysis. News websites in our sample however show sufficient variations in several aspects, and we will explain a greater detail in Section 4.

SNS account, which provides an even higher identifiable condition, is positively and significantly correlated to the lower occurrence of using offensive terms. Regardless of the account choice, when one's real name is visually represented on the screen with a comment, commenters are less likely to use offensive words. Our results also demonstrate that offensive comments tend to receive a larger number of positive votes (as well as negative votes), providing an important implication for news outlets in designing their ranking mechanism.

A key conclusion of these findings is that commenters are more likely to use offensive words under the less identifiable conditions. Prior work documented that Internet anonymity indeed implies apparent pros and cons. Instead of either using excessive identification policy instruments or maintaining a state of high anonymity, our findings suggest that the use of an SNS account might naturally lead to self-disclosure of identity. Commenters using their SNS account, therefore, are (consciously or unconsciously) less likely to be online flamers or trolls.

To the best of our knowledge, empirical investigations of online commenting behaviors under different settings of anonymity have not been common in prior empirical work, although many studies in a similar context have been conducted by using data from other types of online communications and transactions. Furthermore, while empirical research using real world data is recently increasing, a majority of studies still relies on either laboratory experiments or surveys.

The rest of this paper is organized as follows. In section 2, we present a literature review and in section 3, we propose a conceptual model of how users are likely to behave in commenting when their account choice is related to the degree of identifiability and social image concerns. In section 4, we describe our data in detail. We present our estimation model in section 5, and we document the results of our model in section 6. We conclude in section 7.

2 Related Literature

In this paper, we focus on analyzing how the degree of identifiability and social image concern would affect commenters' behaviors on the Internet. The first important strand of literature in this regard consists in studies of Internet anonymity. In the field of social psychology, the

effect of anonymity on user behavior has been initially examined based on "deindividuation theory" (Zimbarbo 1969), in which an unidentifiable deindividuated state in a crowd is seen as a path to greater uninhibited expression. A majority of studies suggest that reduced awareness without contexts associated with social cues and social evaluations increase a likelihood of antinormative behaviors (see for more literature survey, Chistopherson 2007).

Lea et al. (1992) and Postmes et al. (2001) also highlight that online flaming may decrease when users pay more attention to their social contexts under the setting where their social identity is more salient.[11] Suler (2005) explored how anonymity affects "online disinhibition", and why people may behave differently on the Internet from face-to-face communications. He highlighted that the behavior could be either positive (benign disinhibition) or negative (toxic disinhibition).

Some have argued that anonymity is a key factor motivating deviation from a person's real identity by falsifying, exaggerating or omitting information about oneself (Noonan 1998; Cinnirella and Green 2007). As a consequence, an environment of identifiability may promote self-presentation that corresponds to normative expectations and accountable actions. In line with this perspective, others have suggested that real names and pseudonyms can help promote trust, cooperation and accountability (Millen and Patterson 2003) and that anonymity may make communication impersonal and undermine credibility (Hiltz et al. 1986; Rains 2007).

In contrast, some scholars suggest that high level of anonymity could be beneficial in a certain context (Grudin 2002; Lampe and Resnick 2004; Ren and Kraut 2011). Researchers (particularly, in the field of Human Computer Interaction) have extensively explored the idea that computer-mediated communication (CMC) may provide a more equal place for communicators without revealing their social identity (Sproull and Kiesler 1991), and that anonymous speech helps construct free discussion environment through the autonomous disclosure of personal identity (Zarsky 2004). This positive effect of anonymity in CMC was termed the *equalization hypothesis* by Dubrovsky et al. (1991). According to the so-called Social Identity Model of Deindividuation Effects (SIDE) (Reicher et al. 1995; Spears and Lea 1994), an update on

[11] Some law scholars also have argued that an anonymous environment is more likely to lead to defamation, threats, and lander by users (Cohen 1995).

the previous deindividuation theory – anonymity can accentuate the desire to follow a socially dominant normative response when social identity is strong and personal identity is low. In sum, theories predict a variety of manners in which anonymity can indeed influence individual behavior.

Similarly nuanced are the results of numerous empirical studies of anonymity in the fields of psychology, organizational behavior, and information systems. Jessup et al. (1990) suggested that anonymity would lead to a reduction in behavioral constraints and enable individuals to engage in discussions that they would not engage in when they are identifiable. Yet, findings of a greater number of empirical research tends to be associated with the dark side of anonymity. Some experimental studies challenged the equalization hypothesis by finding that CMC would not be substantially helpful in increasing equality in communication between individuals of different status (Hollinghead 1993; Strauss 1996). Sia et al. (2002)'s finding suggests a tendency of group polarization under the condition of anonymity, and Liu et al. (2007) found that low level of anonymity is linked to a higher quality of comments by using natural language processing. Coffey and Woolworth (2004) compared local citizens' behaviors on an anonymous discussion board provided by a local newspaper website, to those in the town meeting provided by the city authority, and found that when discussing a violent crime, threats and slanders were more frequent in the comments on the online anonymous board than the identifiable town meeting.

It is worth noting that anonymity may vary in degrees and is not dichotomous (Nissenbaum 1999), as emphasized by Qian and Scott (2007). For instance, pseudonym would contain various degree of anonymity (Froomkin 1995; Froomkin 1996). People may use either one or more persistent pseudonyms that are not connected to their true identity but sometimes others can partially recognize one's real identity from revealed personal information (Goldberg 2000). For example, when pseudonyms can be easily disposable and be cheaply obtained, this condition would facilitate anonymity. Friedman and Resnick (2001) propose a mechanism in which a central authority offers the use of free but unreplaceable pseudonym to avoid high social costs and individual misbehaviors that are likely to happen in the use of cheaply replaceable pseudonyms.

As for the second strand of literature, our study is related to the private provision of public goods. Economists attempted to model incentives of these contributions and to empirically test associated hypotheses. According to Benabou and Tirole (2006), people may contribute to public goods due to intrinsic incentives, extrinsic incentives, and social image motivations. While intrinsic and extrinsic motivations refer to altruism (or other forms of prosocial preferences) and monetary incentives, respectively (for surveys, see Meier 2007), image motivation captures people's desire to be perceived as "good" by others. An ample body of research has been conducted to examine motivations for prosocial behavior, particularly in the domain of public economics, and a majority of this work was done through surveys and controlled experiments on a variety of offline settings such as charitable giving, voluntary participations in public services, unpaid supports for local communities, etc (see, Ariely et al. 2009).[12] Findings from a number of prior studies suggest that people will act more prosocially when their social image is more concerned (Ostrom 2000; Andreoni and Petrie 2004; Dana et al. 2006).

A growing number of studies have examined how voluntary activities can be motivated in the context of advanced information technology. Lerner and Tirole (2003) investigate developers' contribution of open source software, suggesting that their primary incentives are social image and career concern. Similar to findings in economic literature, intrinsic motivation (e.g., altruism, individual attributes and self-expression) and social concerns (e.g., reputation, social affiliation and social capitals) are also highlighted as key motivations to contribute in online communities (Wasko and Faraj 2005; Jeppesen and Frederiksen 2006; Chiu et al. 2006). Zhang and Zhu (2010) also examine the causal relationship between group size and incentive to contribute public goods by using Chinese Wikipedia data, and they found that collective provision on Wikipedia are positively correlated to the participating group size.

Based on findings of motivations to contribute, researchers are naturally interested in how to design moderating mechanism in which participation is encouraged and antinormative behaviors are discouraged (see, Kiesler et al. 2010). As noted above, reputation systems are widely used and analyzed in this respect. A reputation is an expectation about an agent's be-

[12]In addition to offline settings in most studies, Sproull et al. (2005) explanatory emphasized the importance of motivational signals and trust indicators in incentivize online social behaviors.

havior based on information about or observations of, its past behavior (Resnick et al. 2000). Screening by reputation may enforce social norms, such as honesty and co-operation, in large communities. Reputation mechanism in online electronic markets (e.g., eBay) facilitates economic transactions, thereby promoting efficiency (Dellarocas 2005).[13] Analyzing reputation scheme, users may incentivize through social interaction, and Wang (2010) found that an online restaurant review website equipped with functions of social engagement and identity presentation showed significantly higher rates of participation and productivity than those in other competing websites without those social network functionalities.

This crowd-based moderation seems to be positioned as an effective mechanism to enhance the quality of content and to reduce deviation from social norms. Underlying phenomena and structures of these studies correspond to the core feature of online commenting system in our context.[14] However, despite the fact that a considerable number of studies have been conducted in the context of e-commerce, online communities and open source software, there seem to be few studies on online news media.

On the other hand, with regards to online anonymity, debates on compulsory real name policy have recently been heated:[15] proponents of real name policy argue the negative effect of anonymity on the quality of discourse. This argument is supported by experimental studies (Joinson et al. 2009) and by content analyses of online forums (Gerstenfeld et al. 2003). This group of researchers and practitioners highlights the importance of identifiable profiles to be able to hold Internet users legally accountable. Opponents of real name policy, however, state several problems such as implementation difficulties, costs, and declines in participation. Cho

[13]Resnick et al. (2000) identify three challenges that any reputation system must meet. Firstly, it must provide information that allows buyers to distinguish between trustworthy and non-trustworthy sellers. Secondly, it must encourage sellers to be trustworthy; and thirdly, it must discourage participation from those who are not trustworthy.

[14]A commenting system platform accompanies with moderating mechanism. In this respect, this is different from peer to peer platforms where users can share files and opinions without a mediator.

[15]Ruesch and Marker (2012) identify three major rationales for real name policy: (1) the possibility to restrict access to citizens; (2) the prevention of offensive communication; and (3) the strengthening of a transparent democracy. He also accounts for several major objections of the real name policy: (1) the violation of privacy rights; (2) administrative problems causing high expenditure of time and costs; (3) negative media and public attention; and (4) usability problems that may result in a low rate of participation. See Ruesch and Marker (2012), for more information.

(2013)'s empirical findings on real name policy in South Korea indicates that the policy significantly decreases user participation, whereas there is not significant impact on the decrease in antinormative behaviors in the long run.

In this context, there are various 'compromises' between complete anonymity and real name policy.[16] Assuming that people are likely to behave in a less inhibited fashion online as a consequence of anonymity and invisibility (Suler 2005), using SNS accounts in other online communicative activities may partially increase a likelihood of self-disclosure. Note that SNS affords users the opportunity to create their own profile pages and to communicate with their offline acquaintance and online friends. Gross and Acquisti (2005)'s finding suggests that a majority of users revealed pictures, date of birth, and other personally identifiable information.

In sum, we know of no prior work on online anonymity where users have a choice of account utilized by SNSs, despite the facts that a considerable number of the Internet users are currently using Facebook and Twitter[17] and a growing number of websites has allowed users to sign in their websites by using SNS accounts.

3 Conceptual Model and Hypotheses

In order to construct a testable model corresponding to our data feature, we fundamentally assume that, to write a comment, a user should sign-in by choosing one of the three given account types as shown in Figure 1: (1) a non-SNS account (an online newspaper website's account), (2) pseudonym-SNS account (e.g., Twitter) and (3) real name-SNS account (e.g., Facebook).

The premises of the degree of identifiability in choosing an account type can be specified as follows: All else given, a commenter may choose a particular account type that is associated with the willingness to disclose their real identity. The use of an SNS account is associated with

[16] According to Ruesch and Marker (2012), the level of anonymity can be ranged from no registration at all, registration with pseudonyms, registration with a real but unverified name, or registration with a hidden real name and pseudonyms, to registration with a verified name and possibly also personal data.

[17] Based on statistics from Alexa (http://www.alexa.com), the sum of daily reach of Facebook and Twitter was 50% of daily Internet consumption in the United States as of February 2011.

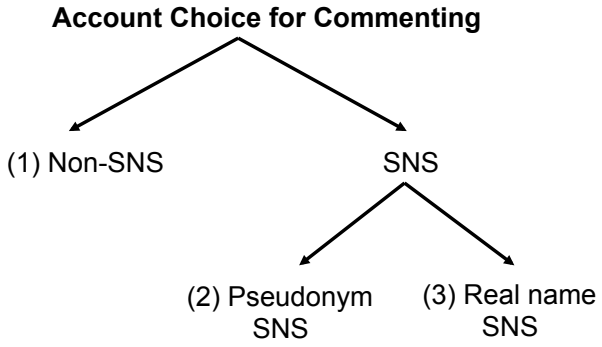

Figure 1: Structure of Commenter's Account Choice for Commenting

higher willingness to disclose her true identity. For example, a commenter using an SNS account rather than a less identifiable and pseudonymous newspaper website account (a non-SNS account) is more likely to be involved in communications and interactions through comments and her online social network. Furthermore, with regards to the degree of identifiability, there would be a significant difference between a real name-SNS account and a pseudonym-SNS account. Choosing real name-SNS account would provide even higher degree of identifiability, because her real name is displayed on screen with her comment.[18]

We also assume that users prefer a self-image as a socially decent and neat person. A commenter may suffer a loss of self-image if she deviates from social norms, e.g., her comment includes swear words. Considering a commenter does regard using offensive words as a morally inferior activity, she would experience a higher degree of self-image loss if she were circumstanced as being more identifiable. For instance, when a user writes a comment including offensive terms by signing-in her real name-SNS account, her image loss would be higher than the case in which she uses either non-SNS account or pseudonym-SNS account. This is because her personal information is more identifiable with the disclosure of real name.

In short, the use of real name-SNS account would be associated with the highest degree of identifiability, whereas the use of a non-SNS account provides the least information about

[18]In the data, a small fraction of commenters who chose a non-SNS account allow their real name to be displayed even under the non-compulsory website policy. A small fraction of commenters who chose a real name-based (or pseudonym-based) SNS-account also chooses pseudonym (or real name). We explain this in greater detail in Section 4.

the commenter's true identity. The use of pseudonym-SNS account is located between the two. As the true identity is more identifiable, the individual's self-image loss would become greater when she write a comment using offensive words.[19] For better understanding, the relationships between the account type and the degree of identifiability (or the amount of self-image loss) are shown in Figure 2.

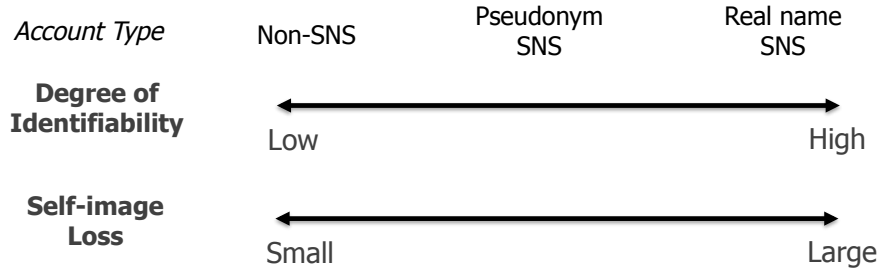

Figure 2: Relationships between Account Choice and Degree of Identifiability (or Self-image Loss)

A challenge in our framework is that two related decisions are made almost simultaneously by the same actor: a decision to choose an account type (which determines the degree of identifiability) and a decision of whether or not she writes an offensive comment.[20] One can easily anticipate that a commenter who intends to use swear words is likely to prefer a less identifiable condition (e.g., using a non-SNS account rather than a real name-SNS account).

To represent our arguments formally, we suppose that the individual's choice of whether or not to post offensive comments is discrete: either her comment does not include offensive words ($NEAT_i = 1$), or the comment includes offensive languages ($NEAT_i = 0$).[21] Note that we

[19] One might argue that a commenter may enhance self-image by using offensive words in her comments, because she may receive higher attention from a particular group of audience in the newspaper website. Nonetheless, using offensive words could lead to negative feelings for the general audience that would be associated with self-image loss.

[20] Precisely speaking, making a comment is chronologically followed by choosing an account. Signing-in an account, however, takes only a few seconds, whereas writing a comment typically takes a substantially longer time. We thus regard both activities as happening (almost) simutaneously.

[21] Note that we borrow modeling framework and empirical approaches developed in Brekke et al. (2010). They examined the impact of social influence on responsibility ascription and glass recycling behaviors. Their finding indicates that responsibility ascription is affected by social interactions and recycling intention may increase when moral responsibility is a burden.

define a term, $NEAT$ (a case in which a commenter does not use offensive words), as opposed to a case in which a comment uses offensive words. As for a user's account choice, let SNS_i is equal to one if an individual i uses an SNS account, zero if she uses a non-SNS account. Thus, we assume that a commenter i's utility, U_i, from selecting an account type can be written as

$$E[U_i] = \begin{cases} -A \text{ if } NEAT_i = 0 \text{ and } SNS_i = 1 \\ 0 \quad \text{otherwise} \end{cases} . \tag{1}$$

A is a positive arbitrary value specifying a certain level of self-image loss when a user i chooses an SNS account and uses swear words in her comment. Assuming that there is no self-image gain when she does not use offensive words, the expected utility from the other two combinations ($NEAT_i = 1$ and $SNS_i = 1$, $NEAT_i = 1$ and $SNS_i = 0$) is zero. If she chooses a non-SNS account and uses swear words in her comment ($NEAT_i = 0$ and $SNS_i = 0$), she might suffer from a certain amount of self-image loss but for the sake of simplicity we assume the loss is negligible (i.e., the self-image loss in this case is also zero), due to the fact that her personal identity is less likely to be identifiable through the pseudonym she used.[22]

Similar to this argument, we further compare the use of real name-SNS account to the use of pseudonym-SNS account. Let $REAL_i$ is one if i uses a real name-SNS account, zero if she uses a pseudonym-SNS account. This commenter i's utility from selecting an account type can be written as,

$$E[U_i|SNS = 1] = \begin{cases} -B \text{ if } NEAT_i = 0 \text{ and } REAL_i = 0 \\ -C \text{ if } NEAT_i = 0 \text{ and } REAL_i = 1 \\ 0 \quad \text{otherwise} \end{cases} . \tag{2}$$

B and C are positive arbitrary values, and we assume the value of C is greater than that of B. This assumption is in accordance with our argument in which her self-image loss would be further augmented as her true identity is more likely to be identifiable. In other words, when

[22]For instance, other readers are difficult to identify the commenter's true identity when the commenter's account is not linked to her SNS.

using offensive words the self-image loss would be greater in the case of using a real name-SNS account than the case of using a pseudonym-SNS account. For example, suppose other users in an online news site attempt to trace a commenter's real identity due to the commenter's flames. They could find the commenter's true identity more effortlessly when they recognize the commenter's real name rather than the pseudonym.[23]

In sum, our main hypotheses, taking into account that our data on the account choice and the use of offensive words are binary, can be documented as follows:[24]

HYPOTHESIS 1: The probability to use offensive words in commenting is negatively correlated to the degree of identifiability. In other words, the use of non-SNS account (associated with the lower degree of identifiability) increases the probability of using offensive words, all else equal.

HYPOTHESIS 2: The use of a real name-SNS account (associated with higher degree of identifiability) decreases the probability to use offensive words, compared to the case that a commenter uses a pseudonym-SNS account, all else equal.

4 Data

4.1 Sample Construction

We collected our unique data from the largest online commenting system provider in South Korea. The firm launched their service in July 2010, and over 100 various online websites in South Korea, including several major domestic news media, have adopted the system as of 2012. Our data covers all comments to news articles from 35 news media websites[25] during a 6-week period, March 8 – March 28, 2012 and April 12 – May 2, 2012. Resulting sample includes 75,314 comments by 23,151 commenters.[26]

[23] It is evident that a real name-SNS (Facebook) provides more abundant personal identifiable information than a pseudonym-SNS (Twitter).

[24] We revisit these main hypotheses by connecting to empirical specifications in Section 5.

[25] There are additional news media websites serviced by the third party commenting platform provider, but we did not include these websites due to the trivial number of comments therein.

We exclude the period, March 29 – April 11, 2012, because this period was the official election campaign period within which only verified users with true identities were able to comment on the website. This heavy-handed policy would indeed discourage user participation, and hence, their communication behavior may significantly change. In spite of this fact, we however chose each three-week interval before and after the official election campaign period for our study, because we expect that users are likely to express their opinions (or sentiments) on the election and politics around the election period. This provides a more desirable natural experimental setting in examining users' conscious (or unconscious) behaviors.

An additional advantage of data from such a diverse set of news media websites is that each news medium may hold idiosyncratic characteristics and perspectives in terms of politics and the economy, and hence, users may have a preference to visit a particular website to read news articles and to participate in discussion. Our data incorporates comments from Internet news media operated by some of the major nationwide daily newspapers, the largest nationwide business newspaper, a dozens of local newspapers, and numerous category-specialized online-specific news websites. This variety may alleviate possible selection concerns that would be encountered in a typical field study.

Figure 3 shows features of the commenting system by which we can obtain information of news website source, commenting date, identifiable information about commenters, connected social network sites (e.g., Facebook, Twitter), contents of comment, feedback from other users (i.e., the number of likes or dislikes). Based on the contents of the comment, we calculate the length of each comment and identify whether or not a comment turns out not to be neat enough by including designated offensive words. It is worth noting that we assume that commentators did not change their identifier during the period of our study, since it is costly and cumbersome to do so.[27]

[26]The frequency of the number of comments per commenter indeed shows highly skewed distribution: 1 comment- 13,382 commenters (56%), 2 comments- 1,693 (16%) and 3 comments- 1,025 commenters (7%) account for 80% out of total commenters. In terms of proportion of total number of comments, however, comments by these less-frequent users only account for about 30% out of total comments. We will take into account this aspect in our empirical specification.

[27]To verify our assumption, we check commenters who used multiple accounts during our study period. It appears that only fewer than 0.5% of total commenters in our sample used multiple accounts, which is negligible.

Figure 3: Commenting System Feature

Note:

(a) Sign-in identifier (newspaper or SNS accounts by thumbnail pictures)

(b) Commenter identifier (pseudonym or real name)

(c) Contents

(d) Date and time of posting

(e) Review (the numbers of positive and negative votes from other users)

What makes our data set interesting is that a commenter can choose one of three types of accounts to comment: (1) non-SNS account, (2) pseudonym-SNS account, and (3) real name-SNS accounts.[28] Since the commenting system provider offers a common platform to all clients, online news websites in our sample equally have the same feature in terms of the account choice set encountered by users. In other words, a user has to sign-in by choosing one of the given

[28]Besides Twitter, the commenting system provides two additional pseudonym SNS account options. These domestic SNSs have almost identical features to Twitter and are run by two of largest Internet portals in South Korea. Besides Facebook, the commenting system provides an additional real name SNS account option. This domestic SNS has very close functions and features to Facebook and are run by the third largest Internet portals in South Korea.

alternatives to write a comment.

As shown in Figure 3, if a commenter logs in with the newspaper account, the comment comes with the newspaper's logo, which has no personal identifiable information, and a pseudonym which is typically a user ID in the website. In contrast, if a user chooses one of their SNS accounts, her comment comes with a user's current profile picture that often contains a person's face image or other identifiable information connected to a user's real identity. Followed by a user's SNS account choice, a sign-in identifier with a small image logo of the selected SNS is displayed on screen, as seen in Figure 3. If other readers in the news website are interested in the commenter's profile, they are able to visit the commenter's SNS webpage by simply clicking the displayed image. These salient features truly make clear distinctions between non-SNS account and SNS account in terms of the degree of identifiability in accordance with what we noted in the previous section.

Another interesting feature of our data is that a comment may or may not present with a person's real name, which would also affect the degree of identifiability. One might have a question about potentials regarding the exposure of real name: a commenter using a pseudonym-SNS account uses a real name and a commenter using a real name-SNS account uses a pseudonym. To verify this likelihood, we first see the distribution of commenters by account type and the use of real name as shown in Table 2.

Account / ID Type	Pseudonym	Real name	Sum
Pseudonym SNS	11,498 (98.37%)	190 (1.63%)	11,688 (50.49%)
Real name SNS	71 (1.12%)	6,277 (98.88%)	6,348 (27.42%)
Non-SNS	3,257 (63.68%)	1,858 (36.32%)	5,115 (22.09%)
Sum	14,826 (64.04%)	8,325 (35.96%)	23,151 (100%)

Table 2: Distribution of Commenters by Account Type and the Use of Real Name

It appears that very small fractions of users contravened the rules, so our classification of pseudonym- and real name-SNSs seems to be valid. On the other hand, in a case in which she chooses a non-SNS account, her real name can be displayed or not with her comment,

according to the website's policy.[29]

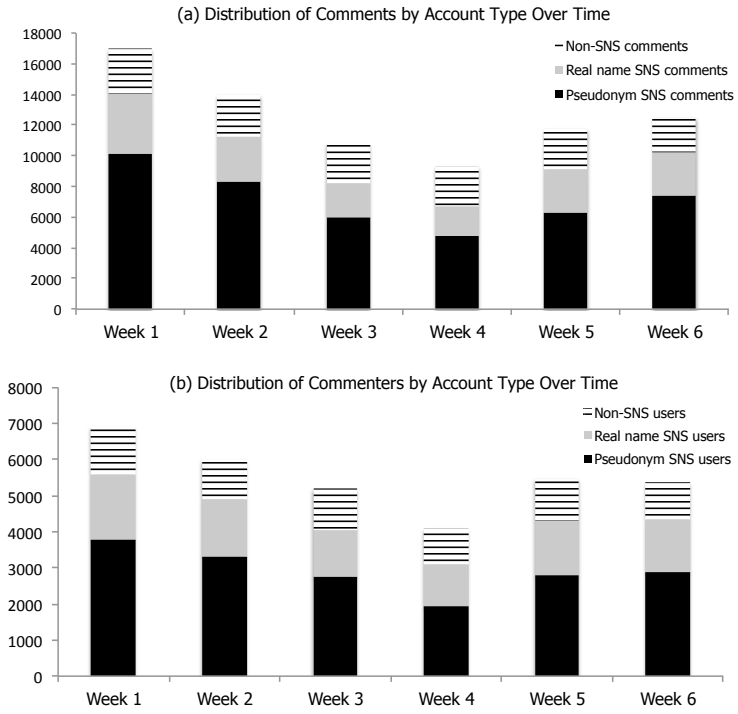

Figure 4: Compositions of Comments and Commenters by Account Type Each Week

We present compositions of comments and commenters by account type over time in Figure 4. Out of 75,314 comments, comments with pseudonym-SNS and real name-SNS account for 57% and 22%, respectively, whereas comments with non-SNS accounts for 21%. As for the composition of commenters by account type, comments with pseudonym SNS and real name SNS account for 50% and 28% respectively, whereas comments with non-SNS account for 22%. That is, a majority of comments are written with SNS accounts, implying that users may prefer to use their SNS accounts when commenting. This observation corresponds to the anecdotal evidence, which suggests that introducing a social commenting system may increase a user's participation. In other words, the convenience of using an SNS account may positively

[29] A third-party commenting system's service can be customized to each news organization's requests. Accordingly, most news sites allow commenters' pseudonymous sign-in names to be displayed on screen, whereas some of news sites in our sample required commenters to provide their real names, and this real name is disclosed with a user's comment.

contribute to collective provision in commenting.

4.2 Measure Operationalization: Content Analysis

To evaluate whether or not a comment is antisocial, it is important to appropriately classify aggressive and offensive comments from other normative comments. We identify offensive comments by conducting content analysis: we select 651 offensive words including 319 abusive words designated by *Nielsen KoreanClick*, one of the largest the Internet market research firm in South Korea. The selected words contain swear words (or commonly-used pseudo swear words to avoid automatic filtering procedures), the use of vulgar nicknames to belittle famous politicians and political parties, and other frequently-used offensive words at online communities and comments in South Korea.[30]

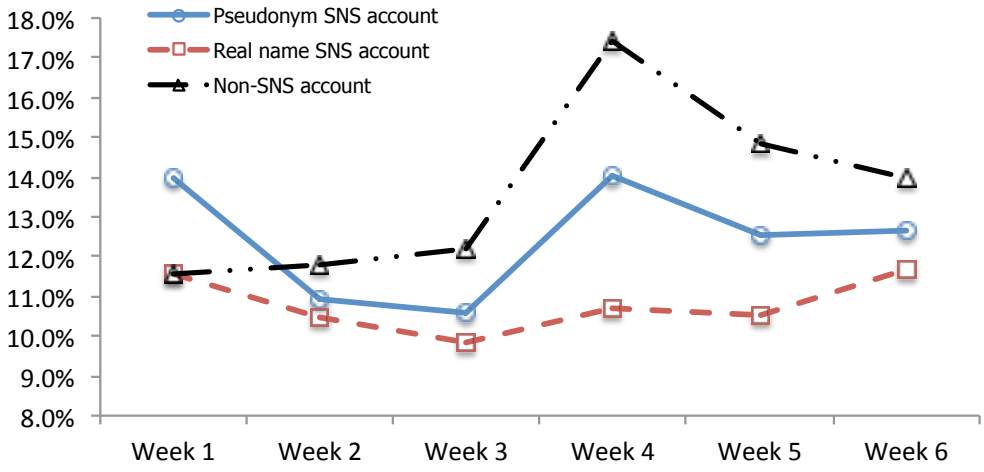

Figure 5: Proportion of the Use of Offensive Words by Account Type Each Week

Figure 5 depicts the proportion of comments including offensive words by account type. In accordance with our hypotheses in Section 3, a real name-SNS account (associated with a

[30]To verify the validity of our content analysis, we consulted two Korean Ph.D students (used to work as journalists of newspaper firms in South Korea and now studying Journalism and Organization Behaviors, respectively, in the United States) to examine selected offensive words used in this study. A few terms were additionally included in the final set of offensive words according to their suggestions and they generally agreed that a set of offensive terms in our study quite exhastively captured currently using offensive and provative experessions.

higher degree of identifiability) shows a smaller fraction of using offensive words than other conditions under the lower degree of identifiability. Similarly, in Figure 6, a comment with the displayed one's real name is less likely to include designated offensive words. This observation corresponds to our conceptual model, because the disclosure of one's true name with her comment should be associated with high degree of identifiability, no matter which type of accounts a commenter choose. While these outcomes noticeably show the fractional differences across account types in line with our hypotheses, this approach is not sufficient because the user's account choice is indeed endogenous to the propensity of the user's behavioral choice; namely, statistical analyses are required. We thus illustrate identifications of variables and then introduce our empirical approach using a bivariate probit model. By doing so, we can take the endogenous problem into consideration to some extent.

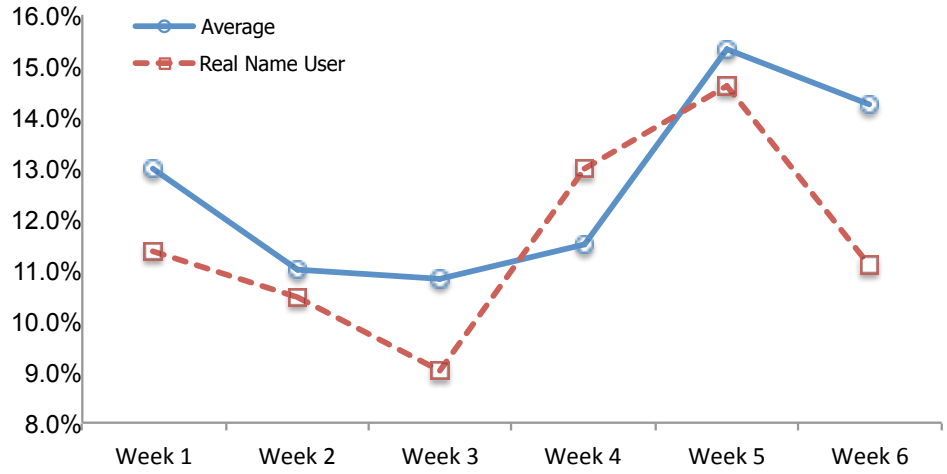

Figure 6: Proportion of the Use of Offensive Words by the Use of Real Name

4.3 Identifications

First, one of our primary interests is each commenter's account choice. Following notations in Section 3, SNS take the value 1 if the commenter chooses any types of SNS accounts and the value 0 if she chooses a non-SNS account. Similarly, the variable, $REAL$, takes the value 1 if the commenter chooses a real name-SNS account and the value 0 if she chooses a pseudonym-SNS

account. Second, the other dependent variable is whether or not a user's comment includes designated offensive words. The behavior is measured by the variable $NEAT$ (specifying a commenting behavior). As described earlier, we conducted a content analysis to distinguish comments including offensive words from others. In other words, $NEAT$ takes the value 1 if the comment does not contain the designated offensive words, and 0 otherwise.

Variable	Description	Obs.	Mean	Std. Dev.	Min	Max
	Outcome Variables:					
NEAT	Use of offensive words or not	75314	0.8765	0.3290	0	1
SNS	Use of SNS account of not	75314	0.7892	0.4079	0	1
REAL	Use of Real name-based SNS account or not	75314	0.2203	0.4144	0	1
	Comment-specific Variables:					
NAME	Comment with disclosed true name or not	75314	0.2742	0.4461	0	1
LENGTH	the length of comments (1-250 characters)	75314	85.8349	66.2763	1	250
# of LIKES	The number of positive votes from other users	75314	17.6616	38.9013	0	2856
# of DISLIKES	The number of negative votes from other users	75314	6.8748	20.8986	0	1458
	Commenter-specific Variables:					
ALL COMMENTS	The number of comments by commenter	23151	26.0081	57.9708	1	517
AVG LENGTH	The average length of comments by commenter	23151	85.1565	49.8412	1	249
AVG GOOD	The average positive votes received by commenter	23151	19.1889	28.0657	0	2856
AVG BAD	The average negative votes received by commenter	23151	7.6746	13.8848	0	783

Table 3: Descriptive Statistics

	NEAT	SNS	REAL	NAME	LENGTH	LIKES	DISLIKES	ALLCOMMENT	AVGLENGTH	AVGGOOD	AVGBAD
NEAT	1										
SNS	0.019	1									
REAL	0.024	0.275	1								
NAME	0.032	0.005	0.820	1							
LENGTH	-0.124	-0.086	-0.014	0.002	1						
LIKES	-0.079	0.010	-0.011	-0.032	0.058	1					
DISLIKES	-0.074	-0.012	-0.033	-0.050	0.093	0.342	1				
ALLCOMMENT	-0.009	-0.010	-0.114	-0.135	0.040	-0.061	-0.022	1			
AVGLENGTH	-0.079	-0.123	-0.025	0.000	0.740	0.030	0.061	0.047	1		
AVGGOOD	-0.089	0.022	-0.010	-0.043	0.033	0.664	0.236	-0.088	0.046	1	
AVGBAD	-0.094	-0.015	-0.046	-0.074	0.071	0.261	0.600	-0.029	0.095	0.383	1

Table 4: Correlation Matrix

Additional variables for explaining account choice and commenting behavior can be categorized as comment-specific measures and commenter-specific measures. The comment-specific measures are $NAME$ and $LENGTH$. The variable, $NAME$, takes one if a commenter's real name is displayed on screen, and zero if not. The variable, $LENGTH$, indicates how long a

comment is, measured between 1 and 250 characters.

The commenter-specific variables include $ALLCOMMENTS$, $AVGLENGTH$, $AVGGOOD$ and $AVGBAD$. We assume that a commenter's two choices, 1) account type and 2) the use of offensive words, may be related to individual-specific attributes. We thus consider a commenter's features in terms of her frequency of comments ($ALLCOMMENTS$), her efforts in each commenting ($AVGLENGTH$), and quality of a comment measured by feedback received from others ($AVGGOOD$ and $AVGBAD$) during our study period. All these variables are quantified by aggregating the data during the entire period of our study. In order to control user involvement, we also include group dummies according to the user's frequency of comments: Group 1 (1–3 comments), Group 2 (4–9 comments) and Group 3 (+10 comments).

Descriptive statistics including definition of each variable and correlation matrix are presented in Tables 3 and 4, respectively. Apart from the variables we explained above, in Table 3, the mean numbers of "likes" and "dislikes" on each comment are 17.6 and 6.8, respectively, suggesting that users cast more positive votes than negative votes. The mean of total comments ($ALLCOMMENT$) is 26, but there seems to be a considerable variation from 1 to 517.

5 Empirical Strategy

We present a full information maximum likelihood (FIML) joint model based on the suggested conceptual framework, which explains the role of the degree of identifiability in commenting behavior. This FIML model is formally described as a bivariate probit model as Brekke et al. (2010) did. A commenter chooses an account and then writes a comment. Thus, as emphasized earlier, two decisions (the account choice and the commenting behavior) are endogenous. To take this aspect into account, we let outcomes from both the account choice and the commenting behavior be linked through a joint error structure. By doing this, we can measure the effect of the account choice on the likelihood of using offensive words. That is, the central idea here is that the account choice alters the payoff from commenting behavior as explained in Section 3.[31]

5.1 The Joint FIML Model

We first consider an account choice (SNS versus Non-SNS accounts). Let $Z_{1i} + \varepsilon_{1i}$ represent the data generating process for the account choice such that a person i chooses SNS account ($SNS_i = 1$) if and only if $Z_{1i} + \varepsilon_{1i} > 0$, where Z_{1i} is an observable deterministic component and ε_{1i} is an stochastic component. Similarly, let $Z_{2i} + \varepsilon_{2i}$ represent the data generating process for commenting behavior (i.e., the comment does not include any designated offensive words) with deterministic component Z_{2i} and unobservable component ε_{2i}. The error terms are assumed to have zero mean, and standard deviations are σ_1 and σ_2, respectively.

We can test how a commenter's account choice is related to her commenting behavior by examining the error terms (ε_{1i} and ε_{2i}) and their correlation coefficients. ε_{1i} and ε_{2i} are assumed to be jointly and normally distributed as follows:

$$\begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho_1 \\ \rho_1 & \sigma_2^2 \end{pmatrix}\right), \tag{3}$$

where ρ_1 is the correlation coefficient that captures the extent to which the error terms are correlated. We let both σ_1 and σ_2 be normalized to one. Next, the joint probability that a commenter uses an SNS account and posts a comment that does not contain offensive terms (denoted p_{1i}) is:

$$p_{1i} = \Pr(Z_{1i}^* > -\varepsilon_{1i}^*, Z_{2i}^* > -\varepsilon_{2i}^*) = \Phi_2(Z_{1i}^*, Z_{2i}^*, \rho_1), \tag{4}$$

where $Z_{1i}^* = Z_{1i}/\sigma_1, Z_{2i}^* = Z_{2i}/\sigma_2, \varepsilon_{1i}^* = \varepsilon_{1i}/\sigma_1, \varepsilon_{2i}^* = \varepsilon_{2i}/\sigma_2$ and Φ_2 is the bivariate standard normal cumulative density function. Similarly, the joint probability of the choice of SNS account and the use of offensive words in her comment (denoted p_{2i}) can be written as

[31] It is worth noting that this process allows us to empirically measure the association between the two and the conditional probability of commenting behavior given a commenter's account choice.

$$p_{2i} = \Pr(Z^*_{1i} > -\varepsilon^*_{1i}, Z^*_{2i} \leq -\varepsilon^*_{2i}) = \Phi(Z^*_{1i}) - \Phi_2(Z^*_{1i}, Z^*_{2i}, \rho_1) \tag{5}$$

where Φ is the univariate standard normal cumulative distribution function.

To complete, the joint probability that the a commenter chooses a non-SNS account while she does not use offensive words in her comment (p_{3i}), and the joint probability of choosing a non-SNS account with using offensive words (p_{4i}) can be expressed, respectively, as follows:

$$p_{3i} = \Pr(Z^*_{1i} \leq -\varepsilon^*_{1i}, Z^*_{2i} > -\varepsilon^*_{2i}) = \Phi(Z^*_{2i}) - \Phi_2(Z^*_{1i}, Z^*_{2i}, \rho_1), \tag{6}$$

$$p_{4i} = \Pr(Z^*_{1i} \leq -\varepsilon^*_{1i}, Z^*_{2i} \leq -\varepsilon^*_{2i}) = \Phi_2(-Z^*_{1i}, -Z^*_{2i}, \rho_1). \tag{7}$$

The joint likelihood function $L(\varphi, \gamma, \lambda, \beta, \rho_1)$[32] used to estimate our model (specifying the relationship between the degree of identifiability and commenting behavior) can be written as

$$L(\varphi, \gamma, \lambda, \beta, \rho_1) = \prod\nolimits_{\forall i} (p_1^{NEAT \cdot SNS} \times p_2^{NEAT \cdot (1-SNS)} \times p_3^{(1-NEAT) \cdot SNS} \times p_4^{(1-NEAT) \cdot (1-SNS)}). \tag{8}$$

Replicating identical procedures to consider an account choice between pseudonym- and real name-SNSs, the joint likelihood function $L(\eta, \kappa, \psi, \zeta, \rho_2)$ can be written as

$$L(\eta, \kappa, \psi, \zeta, \rho_2) = \prod\nolimits_{\forall i} (p_5^{NEAT \cdot REAL} \times p_6^{NEAT \cdot (1-REAL)} \times p_7^{(1-NEAT) \cdot REAL} \times p_8^{(1-NEAT) \cdot (1-REAL)}). \tag{9}$$

where ρ_2 is the correlation coefficient.[33]

5.2 Specifications for Account Choice and Commenting Behavior

An commenter's account choice may be influenced by individual characteristics and other covariates. For comment j written by an individual i is therefore assumed to choose an SNS

[32] Parameters in this term will be presented in a following sub-section.

[33] For the sake of brevity, we do not provide details of derivation of this joint likelihood function in the text. Explanations about proposed parameters are provided in the following sub-section.

account ($SNS_i = 1$) when

$$Z_{1ij} = \varphi_0 + \varphi_1 NAME_{ij} + \varphi_2 LENGTH_{ij} + \lambda' X_{ij} > -\varepsilon_{1i}, \tag{10}$$

where X is a vector of other covariates that represent commenter-specific characteristics. We include these commenter-level characteristics because these variables would capture a commenter's general behaviors. In this context, the unconditional probability that individual i will choose an SNS account for writing comments without using offensive words equals the probability that the previous Equation (10) holds $\Pr(SNS_i = 1) = \Pr(Z_{1i} > -\varepsilon_{1i})$.

Similarly, commenting behavior is specified in terms of the individual's utility from not using offensive words in her comment. Individual i does not use offensive words in her comment ($NEAT_i = 1$) if and only if

$$Z_{2ij} = \gamma_0 + \gamma_1 NAME_{ij} + \gamma_2 LENGTH_{ij} + \beta' X_{ij} > -\varepsilon_{2i}, \tag{11}$$

where Z_{2ij} can be seen as the difference in the deterministic components of a random utility model with alternative choices (whether a commenter uses offensive words or not), and ε_{2i} denotes the difference in the stochastic random error of these alternatives. Note that account choice is not explicitly included in Equation (11). Our approach in which commenting behavior is affected by the account choice is captured through the error structure given in Equation (8) of the joint estimation model above.[34]

On the other hand, γ_1 is greater than zero if a commenter's intention not to use offensive words are related to the displayed real name. In other words, γ could capture how the disclosure of real name on screen affects the likelihood of using offensive words apart from the use of SNS account. Further, although we do not have a theoretical background for the parameter on the length of comment, but we use this variable as a proxy for efforts on commenting behavior. Finally, all other control variables including commenter-specific characteristic variables are captured in the vector X with parameter vector β. The unconditional probability that individual i

[34]We follow the suggested estimation methods in Brekke et al. (2010).

will not use offensive words is given by $\Pr(NEAT_i = 1) = \Pr(Z_{2i} > -\varepsilon_{2i})$.

This specification can be replicated for the additional comparison in line with Equation (2) when a commenter chooses either real name- or pseudonym-SNS account:

$$Z_{3ij} = \eta_0 + \eta_1 NAME_{ij} + \eta_2 LENGTH_{ij} + \psi' X_{ij} > -\varepsilon_{3i}, \tag{12}$$

Corresponding probability that an individual i will choose a pseudonym-SNS account for writing comments equals to the probability that the previous Equation (12) holds $\Pr(REAL_i = 1) = \Pr(Z_{3i} > -\varepsilon_{3i})$.

Similarly, for the sub-sample of comments with only SNS account users, an empirical specification can be presented:

$$Z_{4ij} = \kappa_0 + \kappa_1 NAME_{ij} + \kappa_2 LENGTH_{ij} + \zeta' X_{ij} > -\varepsilon_{4i}, \tag{13}$$

Probability that individual i will not use offensive words in this case is given by $\Pr(NEAT_i = 1) = \Pr(Z_{4i} > -\varepsilon_{4i})$.

The above probability expressions can be used to extract conditional mean functions for the commenting behavior outcome given the account choice outcome (Greene 2002). The mean of the expected value in commenting behavior when an SNS account is used $E[NEAT|SNS = 1]$ is:

$$E[NEAT|SNS = 1] = \frac{\Pr(NEAT = 1, SNS = 1)}{\Pr(SNS = 1)} = \frac{\Phi(Z_1^*, Z_2^*, \rho_1)}{\Phi(Z_1^*)}. \tag{14}$$

We can interprete Equation (14) as the expected share of comments which do not include offensive words among comments posted by SNS account. The expected mean value in commenting behavior when an SNS account is not used $E[NEAT|SNS = 0]$ is:

$$E[NEAT|SNS = 0] = \frac{\Pr(NEAT = 1, SNS = 0)}{\Pr(SNS = 0)} = \frac{\Phi(Z_2^*) - \Phi(Z_1^*, Z_2^*, \rho_1)}{1 - \Phi(Z_1^*)}. \tag{15}$$

Equation (15) shows that the expected share of comments which do not include offensive

words among comments posted by non-SNS account. The difference between two equations, $E[NEAT|SNS = 1] - E[NEAT|SNS = 0]$ would be a marginal effect of using SNS account on the probability of not using offensive words in her comment, all else being equal. If ρ_1 is greater than zero, the marginal effect will be positive. Our interest is to see the sign and the statistical significance of ρ_1 to test how the SNS account usage affects commenting behavior. The identical procedure can be repeated for ρ_2.

5.3 Hypothesis Tests

Combining empirical specifications described above which are in accordance with statements in Section 3, we summarize our testable hypotheses quantitatively here. The impact of the use of SNS versus non-SNS account (or the use real name-SNS versus pseudonym-SNS account) on commenting behavior can be presented, respectively,

$$\begin{pmatrix} H_0 : \rho_1 = 0 \\ H_A : \rho_1 > 0 \end{pmatrix} \text{ and } \begin{pmatrix} H_0 : \rho_2 = 0 \\ H_A : \rho_2 > 0 \end{pmatrix}. \tag{16}$$

In addition, the effect of the displayed true name (which is associated with the high degree of identifiability) on commenting behavior can be shown, respectively,

$$\begin{pmatrix} H_0 : \gamma_1 = 0 \\ H_A : \gamma_1 > 0 \end{pmatrix} \text{ and } \begin{pmatrix} H_0 : \kappa_1 = 0 \\ H_A : \kappa_1 > 0 \end{pmatrix}. \tag{17}$$

6 Results

6.1 The Relationship between Account Choice and the Use of Offensive Words

Table 5 presents estimation results for the joint choice model and constitutes the main results of this study.

The model's joint log-likelihood of -64623.51 can be compared to the sum of the two log-

	Account choice (SNS) (Z_1)			**Commenting behavior ($NEAT$) (Z_2)**		
Variables	Est. Parameter	Std Error	z-stat	Est. Parameter	Std Error	z-stat
NAME	0.0504	0.0118	4.28	0.0322	0.0143	2.24
LENGTH	-0.0005	0.0001	-4.79	-0.0027	0.0001	-23.77
ln(ALL COMMENTS)	-0.0423	0.0076	-5.53	-0.0257	0.0091	-2.82
ln(AVG LENGTH)	-0.2023	0.0114	-17.78	0.0197	0.0123	1.59
ln(AVG GOOD)	0.0731	0.0046	16.01	-0.1126	0.0058	-19.57
ln(AVG BAD)	0.0017	0.0055	0.30	-0.0781	0.0063	-12.31
GROUP2	0.2359	0.0179	13.17	-0.0379	0.0208	-1.82
GROUP3	0.2224	0.0262	8.49	-0.0692	0.0308	-2.25
Constant	1.4948	0.0433	34.53	1.7838	0.0469	38.05

Notes:
Estimated ρ_1 = 0.0425 (z-stat: 4.96, p-value: 0.000)
Joint log-likelihood = -64623.51
Sum of Independent log-likelihood= -64635.88 (LR statistic=24.58)
E(NEAT|SNS=1)=0.880
E(NEAT|SNS=0)=0.864
95% CI for E(NEAT|SNS=1)-E(NEAT|SNS=0): (0.006, 0.024)
N=75314.

Table 5: Results: Joint FIML Estimation of SNS *versus* non-SNS accout use

likelihoods of -64635.88 from separate estimations with a likelihood ratio test. Hence, estimated ρ_1 , which means joint estimation is statistically more efficient. The estimated correlation coefficient ρ_1 is 0.042, and this is statistically significant. That is, there is a positive relationship between the two outcomes. This result can be interpreted that SNS account users are less likely to use offensive words than non-SNS account users. Conditional means of $E[NEAT|SNS = 1]$ and $E[NEAT|SNS = 0]$ are 0.880 and 0.864, respectively. This indicates that using an SNS account increases the probability of not using offensive words by about 1.6%.[35]

The estimated coefficient for the use of real name ($NAME$) on commenting behavior is statistically significant and positive at 0.05 level. This result indicates that, *ceteris peribus*, a user's propensity to using offensive words is smaller when commenters' real name is displayed with a comment, consistent to our prediction model in Section 3. The coefficient estimate for the length shows a negative sign and is statistically significant at 0.01 level, suggesting that a

[35]The 95% confidence interval for this quantitative effect of using an SNS account is (0.6%, 2.4%), based on confidence bounds for the estimated correlation parameter.

longer comment is more likely to contains offensive words. It is also interesting that readers may vote more "likes" for comments including offensive words. This finding gives an important policy implication to a website operator in ranking and ordering comments, and we will explore this salient aspect in detail in the sub-section, 6.3. In addition, coefficient estimates of group dummies suggest that heavy users are more likely to use offensive words.

Some interesting findings are also observed in the account choice equation. On the one hand, the estimated coefficient of $ln(AVGGOOD)$ is statistically significant at 0.01 level with positive sign, whereas that of $ln(AVGBAD)$ is not statistically significant. This implies that comments by SNS account users are positively associated with positive feedback from other users, implying that allowing the use of SNS accounts for commenting might be beneficial to the online forum. In addition, coefficient estimates of group dummies are positive and statistically significant at 0.01 level, suggesting that SNS account users more heavily participate in the commenting activities.

We then shift our attention to the comparison between the use of a real name-SNS account and the use of pseudonym-SNS account on commenting behavior. Table 6 presents results.

	Account choice (*REAL*) (Z_3)			**Commenting behavior (*NEAT*) (Z_4)**		
Variables	Est. Parameter	Std Error	z-stat	Est. Parameter	Std Error	z-stat
NAME	4.5447	0.0375	121.31	0.0469	0.0160	2.93
LENGTH	0.0012	0.0003	4.40	-0.0028	0.0001	-21.28
ln(ALL COMMENTS)	0.2198	0.0177	12.39	0.0209	0.0107	1.95
ln(AVG LENGTH)	0.0697	0.0244	2.86	0.0127	0.0141	0.90
ln(AVG GOOD)	0.0536	0.0128	4.18	-0.1032	0.0064	-16.04
ln(AVG BAD)	0.0571	0.0139	4.12	-0.0755	0.0073	-10.37
GROUP2	-0.3607	0.0457	-7.89	-0.0933	0.0238	-3.92
GROUP3	-0.2738	0.0676	-4.05	-0.1715	0.0352	-4.87
Constant	-3.2420	0.0977	-33.18	1.7714	0.0535	33.09

Notes:
Estimated ρ_2 = 0.2273 (z-stat: 13.40, p-value: 0.000)
Joint log-likelihood = -24992.96
Sum of Independent log-likelihood=-25032.90 (LR statistic=179.46)
E(NEAT|REAL=1, SNS=1)=0.942
E(NEAT|REAL=0, SNS=1)=0.846
95% CI for E(NEAT|REAL=1, SNS=1)-E(NEAT|REAL=0, SNS=1): (0.035, 0.175)
N=59439.

Table 6: Results: Joint FIML Estimation of real name-SNS *versus* pseudonym-SNS use

First of all, the model's joint log-likelihood of -24992.96 can be compared to the sum of the two log-likelihoods of -25032.90 from separate estimations with a likelihood ratio test. The smaller value from joint estimation suggests that the joint estimation is statistically more efficient. The estimated correlation coefficient ρ_2 is 0.227, and this is positive and statistically significant. That is, this result supports our hypothesis in which comments written by real name-SNS account users are less likely to include offensive words than comments written by pseudonym SNS account users, on average. Computed conditional means of $E[NEAT|REAL = 1, SNS = 1]$ and $E[NEAT|REAL = 1, SNS = 1]$ are 0.942 and 0.846, respectively, and the difference is approximately 10%, which is positive. This finding suggests that using a real-name SNS account increases the probability of not using offensive words by 10%, all else being equal, compared the case in which a pseudonym-SNS account is used.[36] The magnitude of discrepancy is very marked, suggesting that real name-SNS users are less likely to use offensive words in their comments than pseudonym-SNS users do, in accordance with our hypothesis.

On the other hand, additional interesting findings are also observed from other parameters estimated in Table 6. The estimated coefficient for the disclosure of real name on commenting behavior is statistically significant and positive at 0.01 level. This is consistent to an analogous result from Table 5, indicating that, the disclosure of true identity is indeed negatively correlated to the likelihood of the use of offensive words. The signs of coefficient estimates of comment length, the numbers of "likes" and "dislikes", group dummies in commenting behavior equation are consistent to also results in Table 5.[37]

6.2 Robustness Checks

In our main model, we used a bivariate probit model in which equations for the probability of using offensive words and the probability of choosing an SNS account are simultaneously estimated. Despite the fact that the smaller value of joint log-likelihood than the sum of the

[36]The 95% confidence interval for this quantitative effect of using an SNS account is (3.5%, 17.5%)

[37]We also run regressions: non-SNS account versus pseudonym-SNS account and non-SNS account versus real name-SNS account. The results are consistent to our results in this text. The results will be provided by readers' requests.

two independent log-likelihoods validated that our approach is more efficient, an alternative specification can be considered to verify our finding. We simply model that the user's account choice, real name usage and other covariates are correlated to the unobservable latent variable that determines the use of offensive words. In other words, we add two indicator variables ($d.REALNAMESNS$ and $d.PSEUDONYMSNS$) denoting a user's account choice in Equation (11).[38] Results are presented in Table 7. In Columns (1) and (2), signs of all estimated coefficients are positive and statistically significant, which corresponds to our main results presented in Section 6.1. In other words, the disclosure of a person's real name with comments (i.e., identified comments) and the use of SNS accounts (i.e., higher degree of identifiability) would be positively correlated to a lower likelihood to use offensive words. When we include other covariates in Columns (3) and (4), the signs of two indicator variables for SNS account use still remained positive and statistically significant, whereas the sign of real name disclosure turns to be negative but not significant. Our results from an alternative specification correspond to our main findings.

Propensity Score Matching: the second part of our robustness check consists in documenting relationships between commenting behavior and SNS account choice (or the use of real name) by groups. To do this, we use the Propensity Score Matching method (PSM) suggested by Rosenbaum and Rubin (1983). The idea of PSM is to use a set of control variables to select some samples that are most similar to the samples in the treatment group. The matched samples are used to form a control group. If the dependent variable only correlates with those control variables, then this method produces results as good as a randomized experiment for excluding the impacts of unobserved heterogeneity. In other words, we show that

$$E(NEAT_{ij} = 1|Treatment = 1, \mathbf{X}) > E(NEAT_{ij} = 1|Treatment = 0, \mathbf{X}), \tag{18}$$

where we set treatment groups as a real name-SNS use, a pseudonym-SNS use and real name use. X is a vector covariates including comment- and commenter-specific attributes. By

[38] We do not explicitly show derivations for brevity due to the reason that it would be a similar replication to what we showed in Section 5.

DV: *NEAT*	(1)	(2)	(3)	(4)
d.NAME	0.1180*** (0.0135)		-0.0169 (0.0265)	-0.0159 (0.0265)
d.REAL NAME SNS		0.1331*** (0.0179)	0.1181*** (0.0265)	0.1153*** (0.0265)
d.PSEUDONYM SNS		0.0514*** (0.0146)	0.0556*** (0.0163)	0.0544*** (0.0164)
# of LIKES				-0.0002 (0.0001)
# of DISLIKES				-0.0005* (0.0002)
LENGTH			-0.0027*** (0.0001)	-0.0027*** (0.0001)
ln(ALL COMMENTS)			-0.0263*** (0.0009)	-0.0264*** (0.0090)
ln(AVG LENGTH)			0.0241* (0.0124)	0.0221* (0.0124)
ln(AVG LIKES)			-0.1150*** (0.0057)	-0.1123*** (0.0062)
ln (AVG DISLIKES)			-0.0784*** (0.0063)	-0.0739*** (0.0068)
d.GROUP2			-0.0409** (0.0208)	-0.0443** (0.0209)
d.GROUP3			-0.0711** (0.0030)	-0.0747** (0.0308)
Constant			1.7291*** (0.0497)	1.7356*** (0.0498)
Log likelihood	-28116.25	-28126.75	-26627.72	-26623.60
Wald chi-sq (q)	76.29	55.94	2874.66	2897.61
Prob>chi-sq	0.0000	0.0000	0.0000	0.0000
Pseudo R-sq	0.0014	0.0010	0.0542	0.0544
Number of observations	75314	75314	75314	75314

Note: Standard errors are in parentheses. *** p<0.01; ** p<0.05; * p<0.1.

Table 7: Results: Simple Probit Estimation

dividing into groups according to the frequency of commenting, results may indicate which types of commenters show particularly different behaviors in commenting, which was not explicitly captured in our main specification.

We choose PSM over regression as the baseline analysis for the following reasons. First, PSM can alleviate the multicollinearity among independent variables, which can be a serious problem in our dataset. Second, samples in our treatment group may be different from the population. By using the PSM, we can construct a control group of users, similar to comments in our treatment group. Following Brynjolfsson et al. (2011), we use the 10-nearest neighbor matching algorithm with replacement. In other words, for each sample in the treatment group, we identify 10 comments with the most similar number of comments.

Estimated results are reported in Table 8. First, signs of all cases are positive in line with our main results, indicating that a high level of self-disclosure might decrease the probability of using offensive words. The differences in the heavy user groups are most marked, whereas the average treatment effect in the middle user groups are not statistically significant. For all users, however, the average treatment effect in every row shows a positive sign and this estimation is

statistically significant.

Treatment	Light user (comments: 1-3)	Middle user (comments: 4-9)	Heavy user (comments: +10)	All users
Real name SNS	0.010** (0.005)	0.005 (0.008)	0.021*** (0.007)	0.010** (0.005)
Pseudonym SNS	0.009* (0.005)	0.001 (0.007)	0.022** (0.008)	0.009** (0.004)
Real name	0.012** (0.005)	0.005 (0.008)	0.018** (0.007)	0.006* (0.004)

Note: Standard errors are in parentheses. *** p<0.01; ** p<0.05; * p<0.1.

Table 8: Results: PSM Estimation

6.3 The Relationship between the Use of Offensive Words and Votes

In the previous sub-sections, we have examined the relationship between the degree of identifiability and commenting behaviors. Our results suggest that the use of SNS accounts could reduce the use of offensive words, and the effect could be augmented in the condition in which a user's true identity is more likely to be identifiable.

We now focus on the other aspect of social interaction by looking at feedback from other users in the commenting system. Binary social voting mechanisms have become pervasive on the web: most SNSs and commenting systems provide this simple function to facilitate more participation. Commenters may feel a higher level of satisfaction by receiving a number of positive votes from others. Such voting mechanisms are commonly designed so that voters remained anonymous, and as a result the aggregated votes from crowds could be biased. In this context, "bias" would refer to paradoxical outcomes (Mishra and Rastogi 2012): for instance, socially undesirable comments such as flames would receive a greater number of positive votes and high-quality comments may be relatively ignored. Resulting from crowdsourced value judgments (e.g., ordered from most liked to least), the inappropriate ranking scheme may cause critical problems unless a human moderator is engaged. In this respect, to examine this possible bias more explicitly, we set our dependent variables, $ln(\#of\,LIKES)$ and $ln(\#of\,DISLIKES)$, and our primary interest is how comments including offensive words (de-

fined as $d.ANTI$) are associated with the aggregated binary social feedback. Including other covariates that might affect the social feedback, results are reported in Table 9. One can easily notice that offensive comments receive both greater numbers of positive and negative votes, and the estimated coefficients are statistically significant at 0.01 level except Column (4). A more compelling result is that offensive comments still remained statistically significant in Column (2), indicating that the relationship between comments containing offensive words and positive votes are strongly positive, whereas the equivalent estimated result in Column (4) becomes not significant. The result is consistent with the finding by Mishra and Rastogi (2012) in which comments including offensive words may receive higher attentions and higher ratings.

	DV: ln(# of likes)		DV: ln(# of dislikes)	
Variable	(1)	(2)	(3)	(4)
d.ANTI	0.1771*** (0.0172)	0.0515*** (0.0135)	0.1891*** (0.0183)	0.0180 (0.0141)
ln(# of dislikes)	0.3508*** (0.0048)	0.3784*** (0.0058)		
ln(# of likes)			0.3503*** (0.0046)	0.3922*** (0.0058)
d.Real name SNS	0.2404*** (0.0201)	0.0018 (0.0214)	-0.2120*** (0.0199)	-0.0380* (0.0210)
d.Pseudonym SNS	0.1493*** (0.0169)	0.0488*** (0.0132)	-0.0819*** (0.0168)	-0.0035 (0.0135)
ln(LENGTH)		0.0128 (0.0093)		0.1450*** (0.0092)
NAME		0.0635*** (0.0212)		0.0350* (0.0209)
ln(ALL COMMENTS)		-0.0368*** (0.0090)		0.0099 (0.0090)
ln(AVG LENGTH)		-0.0146 (0.0112)		-0.1106*** (0.0112)
ln(AVG LIKES)		0.8716*** (0.0043)		-0.4201*** (0.0068)
ln (AVG DISLIKES)		-0.3642*** (0.0068)		0.8052*** (0.0048)
d.GROUP2		0.0370** (0.0172)		-0.0318* (0.0171)
d.GROUP3		0.0478* (0.0288)		-0.0861*** (0.0289)
Constant	2.0605*** (0.0177)	0.3214*** (0.0301)	0.8542*** (0.0184)	0.2735*** (0.0303)
R-squared	0.130	0.522	0.130	0.503
F-stat	1377.02	4507.45	1523.68	3081.44
Observations	35146	35146	35146	35146

Note: Standard errors are in parentheses. *** p<0.01; ** p<0.05; * p<0.1.

Table 9: OLS Results: Votes and Commenting Behavior

One additional noteworthy result from our main outcomes in Tables 5 and 6 is that comments with SNS accounts are likely to receive a greater number of positive votes and a smaller number of negative votes than comments with non-SNS accounts. This finding indicates that SNS account users might write more favorable and appropriate comments attached to a news

story. If this is true, adopting an alternative set of SNS accounts in commenting could be beneficial for news websites by motivating users to behave more normatively online. In addition, signs of group dummy variables are opposite in Columns (2) and (4), implying that more frequent users might post more elaborate and appropriate comments to a news article that receives more positive and less negative votes regardless of the fact that the comments by users in these groups are more likely to include offensive words.

It is widely believed based on previous theoretical and empirical literature (Lampe and Resnick 2004), that crowd-based moderation is effective. Such systems work by aggregating community members' moderating effect judgments in order to indicate the quality of comments based on collective intelligence (Malone et al., 2007). Our finding somewhat contradicts this conventional wisdom, suggesting that more comprehensive understandings and mechanisms would be required to manage a commenting system more effectively.

7 Conclusions

As social network functionality is growing exponentially, more users socialize through social network sites by interacting and exchanging their opinions. This function has also infiltrated news websites, facilitating the aggregation and public exposure of a wealth of user-contributed information and opinions. The value of a commenting system essentially relies on the participation of users and the quality of information they provide. Despite the fact that a growing number of websites has recently revamped its commenting system by connecting to social network sites, there have been few studies of how the emerging commenting system mechanism and the degree of identifiability affect the effectiveness of the system. Our study investigates these aspects by focusing on users' communicative behaviors according to the level of identifiability and the disclosure of their true identity. Guided by theories from socio-psychology and economics and empirical evidence from previous experiments, we identify characteristics of users by the degree of identifiability to examine users' behaviors. Our empirical results indicate that there are significant effects of the account choice on a user's commenting behavior. As in any econometric analysis, motivation to deviate from socially normative activities might be possibly

explained by spurious correlations, or unaccounted endogeneity. The main hypotheses tested in this paper are whether or not the degree of identifiability is an important motive for using offensive words. The use of SNS accounts and/or the disclosure of real name naturally lead to being a more identifiable condition and more convenient settings for commenting. That is, such motivation is associated with the use of socially connected accounts. Our finding documents that the use of SNS accounts and the disclosure of real name are less likely to be correlated to the use of offensive words. Considering the fact that commenters are not unwilling to use their SNS accounts for commenting, the adoption of social commenting system would be beneficial to reduce comments including aggressive expressions by naturally leading users to move to the higher degree of identifiability unresistingly.

These topics are not just of academic interests, but have practical implications for practitioners and policy makers. First, implementing a social commenting system seems to be beneficial in motivating users to behave less antisocially. Connection to SNSs may also create additional page views, which may increase online advertising revenues. Without providing a particular functionality to comment with an SNS account, we still recommend that it would be better if online news media allow users to take into account their social image in commenting, which might lead users to behaving more vigilantly and thoughtfully. Second, if news media websites establish a policy to reward heavy and frequent commenters, such as recognition as a "top commenter", this approach would be advantageous, because the small fraction of users seems to highly influence the entire commenting system. Finally, in managing the ranking system, a general ordering mechanism according to the number of positive votes received might cause a bias in which offensive comments could be acknowledged rather than other informative and useful comments. In this respect, more careful design and algorithm should be required in order to provide a more sensible ranking system. Finally, this study can be extended to the examination of crowdsourcing mechanism design, which is a rapidly growing area. While most studies focus on the advantages of crowdsourcing led by cheap and scalable giving, little research examined a dark side of crowdsourcing caused by unethical objectives, spreads of less trustable information and less transparency. Our findings shed light on designing better performing mechanisms

of crowdsourcing in the consideration of social interactions and anonymity.

Our study inevitably has limitations, and some of these limitations could be avenues for future research. First, we operationalize content analysis to separate out offensive comments from others. In spite of the considerable number of keywords applied to this study, more extensive approaches to analyze the given texts (e.g., sentiment analysis) could provide more interesting results in examining user behaviors. Another possible extension can be a replicated work with a new set of the U.S. data to conduct a comparative study. Second, we conjecture that the use of SNS accounts may augment voluntary commenting contributions. This prediction, however, was not explicitly measured in this study. By collecting additional commenting data prior to the introduction of social commenting system, we can test whether the revamped system would truly attract more participation. Another limitation is related to the impact of social effect. Our finding suggested that self-disclosure by using SNS accounts could alter their behaviors. Future study could seek more precise aspects of social interactions and pressure from social norms by tracking individual-level behavioral trends. Finally, our empirical analyses provide limited explanations on causal relationship between the account choice and the commenting behavior. Further structural analysis on the unit of individual commenter-level can be conducted to explore the causality.

References

Ariely, D., Bracha, A., Meier, S. 2009. Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *American Economic Review*, 99(1):544-555.

Andreoni, J., Petrie, R. 2004. Public Goods Experiments without Confidentiality: A Glimpse into Fund-Raising. *Journal of Public Economics*, 88(7-8): 1605–23.

Benabou, R, Tirole, J. 2006. Incentives and Prosocial Behavior. *American Economic Review*, 96(5): 1652–78.

Brekke, K.A., Kipperberg, G., Nyborg, K., 2010. Social Interaction in Responsibility Ascription: The Case of Household Recycling. *Land Economics* 86, 766–784.

Brynjolfsson, E., Hu, Y.J., and Simester, D. 2011. Goodbye Pareto Principle, Hello Long Tail: The Effect of Search Costs on the Concentration of Product Sales. *Management Science* (57:8), pp. 1373-1386.

Chiu, C.-M., Hsu,M.-H., E. T. G. Wang. 2006. Understanding knowledge sharing in virtual communities: An integration of social capital and social cognitive theories. *Decision Support Systems* 42(3) 1872–1888.

Cho, D., Kim, S.D. 2012. Empirical analysis of online anonymity and user behaviors: the impact of real name policy. *Proceedings of the 45th Hawaii International Conference on System Sciences (HICSS)*, pp. 3041-3050.

Cho, D. 2013. Real Name Verification Law on the Internet: A Poison or Cure for Privacy? in *Economics of Information Security and Privacy III* (eds.) Schneier, Bruce. Springer New York.

Christopherson, K.M. 2007. The positive and negative implications of anonymity in Internet social interactions: On the Internet, Nobody Knows You're a Dog. *Computers in Human Behavior*, 23(6): 3038-3056.

Cinnirella, M., Green, B. 2007. Does "cyber-conformity" vary cross-culturally? Exploring the effect of culture and communication medium on social conformity. *Computers in Human Behavior*, 23(4), 2011–2025.

Coffey, B., Woolworth, S. 2004. Destroy the scum, and then neuter their families: the web forum as a vehicle for community discourse? *The Social Science Journal*, 41, 1–14.

Cohen, J. 1995. Right to Read Anonymously: A Closer Look at Copyright Management in Cyberspace, A. *Connell Law Review*, vol. 28.

Dana, J., Cain, D.M., Dawes, R.M.. 2006. What You Don't Know Won't Hurt Me: Costly (But Quiet) Exit in Dictator Games. *Organizational Behavior and Human Decision Processes*, 100(2): 193–201.

Dellarocas, C. 2005. Reputation mechanism design in online trading environments with pure moral hazard. *Information Systems Research*. 16(2) 209–230.

Diakopoulos, N., Naaman, M. 2011. Towards Quality Discourse in Online News Comments, *ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW)*, March 19–23, 2011, Hangzhou, China.

Dubrovsky, V. J., Kiesler, B. N., Sethna, B. N. 1991. The equalization phenomenon: status effect in computer-mediated and face-to-face decision-making groups. *Human–Computer Interaction*, 2(2), 119–146.

Friedman, E. J., Resnick, P. 2001. The Social Cost of Cheap Pseudonyms. *Journal of Economics and Management Strategy* 10(2): 173-1999.

Froomkin, A. 1995. Flood Control on the Information Ocean: Living With Anonymity, Digital Cash, and Distributed Databases. *Journal of Law and Commerce*, vol. 15, pp. 396–507.

Froomkin, A. 1996. Anonymity and Its Enmities. *Journal of Online Law* vol. 4.

Gerstenfeld, P. B., Grant, D. R., Chiang, C. 2003. Hate online: A content analysis of extremist Internet sites. *Analyses of Social Issues and Public Policy*, 3(1), 29-44.

Golbeck, J. 2009. *Computing with Social Trust*. Springer.

Goldberg, I. 2000. A pseudonymous communications infrastructure for the internet. *UC Berkeley Working Paper*.

Greene, W. 2002. *Econometric Analysis, 4th Edition*. Prentice Hall, New Jersey.

Gross, R., Acquisti, A. 2005. Information Revelation and Privacy in Online Social Networks. in *Workshop on Privacy in the Electronic Society*, Alexandria, VA, ACM Press.

Grudin, J. 2002. Group dynamics and ubiquitous computing. *Communications of ACM* 45(12):74–78.

Hiltz, S.R. 1986. The Virtual Classroom: Using CMC for University Teaching. *Journal of Communications*, (36:2: 95-104.

Hollingshead, A. B. 1996. Information suppression and status persistence in group decision-making: the effects of communication media. *Human Communication Research*, 23, 193–219.

Jeppesen, L. B., Frederiksen, L. 2006. Why do users contribute to ÃŤÂÍÃĚrm-hosted user communities? The case of computercontrolled music instruments. *Organization Science*, 17(1) 45–63.

Jessup L., T. Connolly, & J. Galegher 1990, The Effects of Anonymity on GDSS Group Process with an Idea-Generating Task. *MIS Quarterly*, vol. 14.

Johnson, N. A., Cooper, R. B., & Chin, W. W. 2009. Anger and flaming in computer mediated negotiations among strangers. *Decision Support Systems*, 46, 660–672.

Joinson, A., McKenna, K., Postmes, T., Reips, U. 2009. *Internet psychology*. (eds.) Oxford University Press.

Keyes, S. 2009. Fiery forums. *The American Editor*. Winter: 10-12.

Kiesler, S., Kittur, A. Kraut, R., Resnick, P. 2010. Regulating behavior in online communities. In

Kraut, R. E., Resnick, P., eds., *Evidence based social design: Mining the social sciences to build online communities*. Cambridge, MA: MIT Press.

Lampe, C., Resnick, P. 2004. Slash(dot) and Burn: Distributed Moderation in a Large Online Conversation Space. *CHI* Vienna, Austria.

Lea, M., O'Shea, T., Fung P., Spears R. 1992. 'Flaming' in computer-mediated communicationâĂŽ observations, explanations and implications, in: M. Lea (Ed.), *Contexts of Computer Mediated Communication*, Harvester-Wheatsheaf, London, pp. 89–112.

Lerner, J., Tirole, J., 2003. Some simple economics of open source. *Journal of Industrial Economics* 50 (2), 197–234.

Ma, M., Agarwal, R. 2007. Through a glass darkly: Information technology design, identity verification, and knowledge contribution in online communities. *Information Systems Research*, 18(1) 42-67.

Malone, T. W., & Klein, M. 2007. Harnessing collective intelligence to address global climate change. *Innovations*, 2(3), 15–26.

Meier, S. 2007. A Survey on Economic Theories and Field Evidence on Pro-social Behavior. in *Economics and Psychology: A Promising New Cross-Disciplinary Field*, ed. Bruno S. Frey and Alois Stutzer, 51–88. Cambridge: MIT Press.

Millen, D.R. Patterson, J.F. 2003. Identity Disclosure and the Creation of Social Capital. *CHI* 2003.

Mishra, A. and Rastogi, R. 2012. Semi-Supervised Correction of Biased Commenting Ratings, *World Wide Web Conference Session: Frau and Bias in User Rating*, April 16–20, Lyon, France.

Nissenbaum, H. 1999. The meaning of anonymity in an information age. *The Information Society*, 15(2), 141–144.

Noonan, R. J. 1998. The Psychology of Sex: A Mirror from the Internet. pp. 143-168, in *Psychology and the Internet: Intrapersonal, Interpersonal and Transpersonal Implications*, edited by J. Gackenbach. San Diego: Academic Press.

Ostrom, E. 2000. Collective Action and the Evolution of Social Norms. *Journal of Economic Perspectives* 14, 137-158.

Postmes, T., Spears, R. 1998. Deindividuation and antinormative behavior: A meta-analysis. *Psychological Bulletin*, Vol 123(3), May 1998, 238-259.

Postmes, T., Spears, R., Sakhel, K., De Groot, D. 2001. Social influence in computer mediated groups: the effects of anonymity on social behavior. *Personality and Social Psychology Bulletin* 27 1243–1254.

Qian, H., & Scott, C. R. 2007. Anonymity and selfdisclosure on weblogs. *Journal of Computer-Mediated Communication*, 12(4), 1428-1451.

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

Reicher, S. D., Spears, R., & Postmes, T. 1995. A social identity model of deindividuation phenomenon. *European Review of Social Psychology*, 6, 161–198.

Ren, Y., Kraut, R. E. 2011. A simulation for designing online community: Member motivation, contribution, and discussion moderation, *Information Systems Research*.

Rosenbaum, P.R., and Rubin, D.B. 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* (70:1), pp. 41-55.

Resnick, P., Zeckhauser, R., Friedman, E., and K. Kuwabara. 2000. Reputation Systems. *Communications of the ACM*, 43(12):45–48.

Ruesch, M. A., Marker, O. 2012. Real Name Policy in E-Participation. *Conference for E-Democracy and Open Government*. Danube-University Krems, Austria.

Short, J., Williams E., Christie B. 1976. *The Social Psychology of Telecommunications*. Wiley, New York.

Sia, C., Tan, B. C. Y., & Wei, K. 2002. Group polarization and computer-mediated communications: effects of communication cues, social presence, and anonymity. Information Systems Research, 13(1), 70–90.

Spears, R., Lea, M. 1994. Panacea or Panopticon? The hidden power in computer-mediated communication. *Communication Research*, 21,427-459.

Sproull, L., Conley, C. A., & Moon, J. Y. 2005. Prosocial behavior on the net. In Y. Amichai-Hamburger (Ed.), *The social net: Understanding human behavior in cyberspace* (pp. 139–161). New York: Oxford University Press.

Sproull, L., Kiesler, S. 1986. Reducing social context cues: Electronic mail in organizational communication. *Management Science*. 32(11) 1492–1512.

Sproull, L., Kiesler, S. 1991. *Connections: New ways of working in the networked organization*. Cambridge, MA: MIT Press.

Strauss, S. G. 1996. Technology, group process, and group outcomes: testing the connections in computer mediated and face-to-face groups. *Human–Computer Interaction*, 12, 227–266.

Suler, John R. 2005. The online disinhibition effect. *International Journal of Applied Psychoanalytic Studies*, 2(2), 184-188.

Turner, Daniel Drew. 2010. Comments Gone Wilde: Trolls, Frames, and the Crisis at Online Newspapers, *mimeo*.

Wang, Z. 2010. Anonymity, Social Image, and the Competition for Volunteers: A Case Study of the Online Market for Reviews. *The B.E. Journal of Economic Analysis & Policy* 10(1).

Wasko, M., Faraj, S. 2005. Why should I share? Examining social capital and knowledge contribution in electronic networks of practice. *MIS Quarterly*. 29(1) 35-58.

Zarsky, T. 2004. Thinking Outside the Box: Considering Transparency, Anonymity, and Pseudonymity as Overall Solutions to the Problems of Privacy in the Internet Society. *University of Miami Law Review*, 58: 1028–1032.

Zhang, X.(M,), Zhu, F. 2011. Group Size and Incentives to Contribute: A Natural Experiment at Chinese Wikipedia. *American Economic Review* 101(4): 1601–1615.

Zimbardo, P. G. 1969. *The human choice: Individuation, reason, and order vs. deindividuation, impulse, and chaos.*