Who am I on Twitter? A Cross-Country Comparison

Wei Dong
Cornell University
wd96@cornell.edu

Minghui Qiu
Singapore Management University
minghui.qiu.2010@smu.edu.sg,

Feida Zhu
Singapore Management University
fdzhu@smu.edu.sg

ABSTRACT

Users often manage which aspects of their personal identities to be manifested on social network sites (SNS). Thus, the content of personal information disclosed on users' profiles can be influenced by a number of factors, such as motivation of using SNS and privacy concerns, both of which may vary depending on where users reside in. In this study, we compared the content of 2800 United States (US) and Singapore (SG) Twitter users' bios on their profile pages. We found US Twitter users were far more likely to disclose personal information that may reveal their true identity than SG users. The between country difference remained after we took bio length and user activity level into account. The results provide important insights on future studies to understand users' privacy concern in different regions of the world.

Categories and Subject Descriptors

J.4 [Computer Applications]: Social and Behavioral Science

Keywords

Self-Disclosure; Identity Management; Privacy; Culture; Twitter.

1. INTRODUCTION

What type of self-related information do people share in online SNS? Given the mutilfacetedness of one's identity [2] and the prevalence of identity management on SNS [1], users may want to choose which aspect of personal life to be disclosed for a particular account on a particular SNS platform. We examined users' self-information disclosure on Twitter in two countries: the United States (US) and Singapore (SG). Our results show that in addition to the sheer amount of personal information disclosed, whether the information reveals one's true identity in real life also matters. We are conducting follow up studies to understand the mechanism underlying the different self-disclosing behavior and to develop tools that better support the different privacy needs of users from different regions of the world.

According to [5], people from western countries (e.g., North America and West Europe) tend to develop an independent view of self. In these individualistic cultures, the ability of expressing oneself in a direct communication style is expected. In contrast, easterners (e.g., East Asians) tend to develop an interdependent view of self, emphasizing the ability to restrain one's self and to fit-in and maintain harmony with the social context. These cultural norms might make westerners more self-disclosive than easterners. On the other hand, people from individualistic cultures are also more likely to place value on maintaining a private life and protecting it from others' intrusion, whereas people from collectivistic cultures may be more acceptable to other group members' intrusion into their private life [3]. The heightened privacy concern may make westerners less likely to disclose self-related information than easterners. Seemingly, the cultural norms in self-expressiveness and in privacy concerns may influence SNS users' information disclosure behavior in opposite predictions.

To examine the question, we compared US and SG Twitter users' bios. We chose the two countries for the following reasons. First, similar as US, Twitter is the most frequently used micro-blogging platform in SG. Singaporeans are native English speakers, too. Thus, the differences we found cannot be attributed to language or interface features on different platforms. Second, Twitter users tend to have a more diverse set of goals for using the SNS platform (e.g., socialization, information foraging, broadcasting, etc.) [4], allowing us to collect a large variety of user profiles. Third, unlike other SNS such as Facebook where users can set different levels of privacy for each pieces of personal information, Twitter users' profile pages are visible to the general public, even when they have their tweets protected. Twitter users have the maximum freedom to choose what to write in the bio field, except for the length constraint of 160 characters. Thus, the information disclosure behavior is observed in the same privacy setting for all users, ruling out the potential confounding effect that US and SG users may be disclosing under different levels of privacy settings.

2. METHODS

We collected 2800 users' bios (1400/country) using Twitter's stream API. A user is included in our analysis if s/he satisfies all of the following criteria: 1) s/he mentioned any place in the US or in SG in the location field on profile page, 2) s/he has posted at least one tweet with geo-location in the US/SG, and 3) the bio s/he provided is in English, and with a length > 20 characters.

Table 1. Descriptive Statistics of the data.

	United States		Singapore	
	Mean	**Median**	**Mean**	**Median**
Follower #	4739.7	518	171.1	67
Followee #	1282.3	380	182.2	96
Tweet #	8555.5	3360	4327.9	1392
Bio Length	99.4	101	72.1	60

We used Amazon Mechanical Turk (MT) to categorize the bio contents. Workers need to have completed >500 HITs with >90% acceptance rate to work on our HITs. Each time a worker accepts a HIT, s/he will see a bio randomly chosen from the two groups. Workers are instructed to complete the HIT in 2 steps. First, they need to determine if the bio describes a person or a group. We ask workers to default selections to person unless there is evidence suggesting otherwise. Second, if the bio is coded as describing a person, we then ask workers to further categorize the bio content following a coding scheme adopted from [5] and modified to suit the content of Twitter bios (Table 2). The 8 categories are not mutually exclusive. Each bio was given to 4 different workers. To ensure the quality of coding, we used easily identifiable items (urls and email addresses) as criteria for accepting or rejecting HIT submissions (< 10% rejected and redistributed). For each category in the coding scheme, we considered a bio as mentioning this type of information if *2 or more* workers voted yes. As shown in Table 2, the first 5 categories of information are more closely related to one's true identity than the latter 3. For example, it is much easier to identify a person who is male, has graduated from university A and is now working in company B than a person who is Christian, happy, friendly and likes outdoor activities.

WWW 2014 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04

Table 2. Bio coding scheme.

Category		Description and Examples
Personal Identity	Contact information	Email, phone #, personal website/blog url, instant messaging or other SNS account, mailing address
	Demographic information	Age, gender, ethnicity, nationality, location, language, physical appearance
	Family/romantic relations	Father, mother, grandparents, daughter, son, siblings, boyfriend, girlfriend, and so on
	Education background	School/college/university attended, degree, major, and so on
	Career	Workplace, occupation, profession, career-related skills, and so on
Other	Personal interest	Preferences, interests, hobbies or celebrities that one likes
	Psychological attributes	Personality (e.g, easy-going, friendly), emotional status (e.g., happy)
	Values and attitudes	Religious or political views, values, attitudes, proverbs that convey similar information

3. RESULTS

Our analyses focused on the type of information mentioned by Twitter accounts owned by users as *individual persons* (1107 and 1217 in the US and SG, respectively) categorized by MT workers. Table 3 provides a summary of the proportion of users mentioning each type of information. Very distinctive patterns of self-disclosure can be observed in the two groups. US users are more likely to reveal information in all 5 categories that are more closely related to one's true identity. In contrast, SG users are more likely to describe their personality, emotional feelings, values and beliefs in the bios. The only exception is that US users are more likely to mention personal interests than SG users.

Table 3. Bio info mentioned by US and SG users. (*: p < .001)**

Category	US	Singapore	Pearson's χ^2
Contact	11.92%	3.20%	64.66***
Demographic	23.04%	14.22%	30.01***
Family	10.93%	4.68%	31.98***
Education	9.67%	3.62%	34.93***
Career	52.39%	15.45%	357.63***
Interest	53.57%	28.27%	154.23***
Psych. Attributes	16.89%	28.43%	43.66***
Values/Attitudes	17.89%	33.03%	69.43***

Since users in our US sample tend to write longer bios than those in our SG sample (Table 1), one possible alternative explanation is that longer bios give users more space to disclose more self-related information. Users in our US sample also had higher activity levels (i.e., more followers, followees, and tweets) than those in our SG sample. A second alternative explanation is that active users gain more experience, thus may be more willing to disclose self-identifiable information in their bios.

To test whether the between country difference still holds when bio length and activity level are taken into account, we conducted a set of Binary Logistic Regressions to predict the logit likelihood (logit(p) = ln[p/(1-p)]) of each type of information mentioned in the bio. The predictive variables were country (SG = 0, US = 1), bio length, and log transformations of users' number of followers, followees, and tweets (Table 4). As expected, longer bios are associated with more disclosure. The effects of activity levels are mixed. But most importantly, after controlling for bio length and activity levels, the effect of country remained highly significant and in the same direction as in Table 3, except for contact information, which remained in the same direction, but became non-significant. Thus, the observed culture differences are quite robust and cannot be fully accounted for by the bio length and activity level differences observed in our samples.

4. DISCUSSION AND FUTURE WORK

In the current study, we compared the content of information mentioned in US and SG Twitter users' bios. We found that US users tend to disclose more personal information that is closely related to real-life identities, whereas SG users tend to disclose information about their personality, emotion, values and attitudes.

The results point to important future directions in understanding users' self-disclosing behavior in different regions of the world. For example, do users in different countries tend to have different motivations when using the same SNS platform? Do the different user goals influence their self-disclosing behavior? Do users in different countries develop different levels of trust to other users? How does the level of trust vary when friends, acquaintances, or strangers are at the center of consideration of SNS users? With a better understanding of users' different privacy concerns in different parts of the world, future works could also look at their tweeting behavior and design an intelligent system that could remind users of their privacy concerns if they exhibit a different level of self-disclosure than what was observed in the bios.

5. ACKNOWLEDGMENTS

This research/project is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office and the Pinnacle Lab at Singapore Management University.

6. REFERENCES

[1] DiMicco, J.M. & Millen, D.R. Identity management: multiple presentations of self in facebook, in *GROUP*. 2007, ACM: Sanibel Island, FL. p.383-386.

[2] Farnham, S.D. & Churchill, E.F. Faceted identity, faceted lives: social and technical issues with being yourself online, in *CSCW*. 2011, ACM: Hangzhou, China. p. 359-368.

[3] Cho, H., Rivera-Sánchez, M., & Lim, S. A multinational study on online privacy: global concerns and local responses. *New Media & Society*, 2009. 11(3): p. 395-416.

[4] Java, A., Song, X., Finin, T., et al. Why we twitter: understanding microblogging usage and communities, in *Joint 9th WebKDD and 1st SNA-KDD workshop*. 2007, ACM: San Jose, CA. p. 56-65.

[5] Kanagawa, C., Cross, S.E., & Markus, H.R. "Who Am I?" The Cultural Psychology of the Conceptual Self. *Personality and Social Psychology Bulletin*, 2001. 27(1): p. 90-103.

Table 4. Logistic regression coefficients using culture, bio length and activity level to predict the presence of each category of information in users' bios. (*: p < .05, **: p < .01, ***: p < .001)

Outcomes Predictors	Contact	Demographic	Family	Education	Career	Interest	Psychological Attributes	Values & Attitudes
Intercept	-5.33***	-2.35***	-3.29***	-4.50***	-1.79***	-1.58***	-1.26***	-1.80***
Country (SG = 0, US = 1)	.28	**.59*****	**1.06*****	**1.06*****	**1.29*****	**.76*****	**-.53*****	**-.77*****
Bio Length	.02***	.01***	.01***	.01***	.01***	.01***	.00*	.00
Log (# of followers)	1.17***	-.35**	-.47*	-.87***	1.12***	.11	-.48***	-.42**
Log (# of followees)	-.51**	.11	.06	.78**	-.06	-.03	.09	.14
Log (# of tweets)	-.16	.04	.07	.08	-.98***	-.17**	.26***	.51***