© Springer International Publishing AG 2017

Laurie A. Schintler

and

Connie L. McNeely

Encyclopedia of Big Data

10.1007/978-3-319-32001-4\_153-1

**Online Identity**

Catalina L. Toma1

(1)Communication Science, University of Wisconsin-Madison, Madison, WI 53706, USA

**Catalina L. Toma**

**Email:**  [ctoma@wisc.edu](mailto:ctoma@wisc.edu)

Without Abstract

Identity refers to the stable ways in which individuals or organizations think of and express themselves. The availability of big data has enabled researchers to examine online communicators’ identity using generalizable samples. Empirical research to date has focused on personal, rather than organizational, identity, and on social media platforms, particularly Facebook and Twitter, given that these platforms require users to present themselves and their daily reflections to audiences. Research to date has investigated the following aspects of online identity: (1) *expression*, or how users express who they are, especially their personality traits and demographics (e.g., gender, age) through social media activity; (2) *censorship*, or how users suppress their urges to reveal aspects of themselves on social media; (3) *detection*, or the extent to which it is possible to use computational tools to infer users’ identity from their social media activity; (4) *audiences*, or who users believe accesses their social media postings and whether these beliefs are accurate; (5) *families*, or the extent to which users include family ties as part of their identity portrayals; and (6) *culture*, or how users express their identities in culturally determined ways. Each of these areas of research is described in detail below.

Identity Expression

In its early days, the Internet appealed to many users because it allowed them to engage with one another anonymously. However, in recent years, users have overwhelmingly migrated toward personalized interaction environments, where they reveal their real identities and often connect with members of their offline networks. Such is the case with social media platforms. Therefore, research has taken great interest in how users communicate various aspects of their identities to their audiences in these personalized environments.

One important aspect of people’s identities is their personality. Big data has been used to examine how personality traits get reflected in people’s social media activity. How do people possessing various personality traits talk, connect, and present themselves online? The development of the myPersonality Facebook application was instrumental in addressing these questions. myPersonality administers personality questionnaires to Facebook users and then informs them of their personality typology in exchange for access to all their Facebook data. The application has attracted millions of volunteers on Facebook and has enabled researchers to correlate Facebook activities with personality traits. The application, used in all the studies summarized below, measures personality using the Big Five Model, which specifies five basic personality traits: (1) extraversion, or an individual’s tendency to be outgoing, talkative, and socially active; (2) agreeableness, or an individual’s tendency to be compassionate, cooperative, trusting, and focused on maintaining positive social relations; (3) openness to experience, or an individual’s tendency to be curious, imaginative, and interested in new experiences and ideas; (4) conscientiousness, or an individual’s tendency to be organized, reliable, consistent, and focused on long-term goals and achievement; and (5) neuroticism, or an individuals’ tendency to experience negative emotions, stress, and mood swings.

One study conducted by Yoram Bachrach and his colleagues investigated the relationship between Big Five personality traits and Facebook activity for a sample of 180,000 users. Results show that individuals high in extraversion had more friends, posted more status updates, participated in more groups, and “liked” more pages on Facebook; individuals high in agreeableness appeared in more photographs with other Facebook users but “liked” fewer Facebook pages; individuals high in openness to experience posted more status updates, participated in more groups, and “liked” more Facebook pages; individuals high in conscientiousness posted more photographs but participated in fewer groups and “liked” fewer Facebook pages; and individuals high in neuroticism had fewer friends but participated in more groups and “liked” more Facebook pages. A related study, conducted by Michal Kosinski and his colleagues, replicated these findings on a sample of 350,000 American Facebook users, the largest dataset to date on the relationship between personality and Internet behavior.

Another study examined the relationship between personality traits and word usage in the status updates of over 69,000 English-speaking Facebook users. Results show that personality traits were indeed reflected in natural word use. For instance, extroverted users used words reflecting their sociable nature, such as “party,” whereas introverted users used words reflecting their more solitary interests, such as “reading” and “Internet.” Similarly, highly conscientious users expressed their achievement orientation through words such as “success,” “busy,” and “work,” whereas users high in openness to experience expressed their artistic and intellectual pursuits through words like “dreams,” “universe,” and “music.”

In sum, this body of work shows that people’s identity, operationalized as personality traits, is illustrated in the actions they undertake and words they use on Facebook. Given social media platforms’ controllable nature, which allows users time to ponder their claims and the ability to edit them, researchers argue that these digital traces likely illustrate users’ *intentional* efforts to communicate their identity to their audience, rather than being unintentionally produced.

Identity Censorship

While identity expression is frequent in social media and, as discussed above, illustrated by behavioral traces, sometimes users suppress identity claims despite their initial impulse to divulge them. This process, labeled “last-minute self-censorship,” was investigated by Sauvik Das and Adam Kramer using data from 3.9 million Facebook users over a period of 17 days. Censorship was measured as instances when users entered text in the status update or comment boxes on Facebook but did not post it in the next 10 min. The results show that 71% of the participants censored at least one post or comment during the time frame of the study. On average, participants censored 4.52 posts and 3.20 comments. Notably, 33% of all posts and 13% of all comments written by the sample were censored, indicating that self-censorship is a fairly prevalent phenomenon. Men censored more than women, presumably because they are less comfortable with self-disclosure. This study suggests that Facebook users take advantage of controllable media affordances, such as editability and unlimited composition time, in order to manage their identity claims. These self-regulatory efforts are perhaps a response to the challenging nature of addressing large and diverse audiences, whose interpretation of the poster’s identity claims may be difficult to predict.

Identity Detection

Given that users leave digital traces of their personal characteristics on social media platforms, research has been concerned with whether it is possible to *infer* these characteristics from social media activity. For instance, can we deduce users’ gender, sexual orientation, or personality from their explicit statements and patterns of activity? Is their identity implicit in their social media activity, even though they might not disclose it explicitly?

One well-publicized study by Michal Kosinski and his colleagues sought to predict Facebook users’ personal characteristics from their “likes” – that is, Facebook pages dedicated to products, sports, music, books, restaurant, and interests – that users can endorse and with which they can associate by clicking the “like” button. The study used a sample of 58,000 volunteers recruited through the myPersonality application. Results show that, based on Facebook “likes,” it is possible to predict a user’s ethnic identity (African-American vs. Caucasian) with 95% accuracy, gender with 93% accuracy, religion (Christian vs. Muslim) with 82% accuracy, political orientation (Democrat vs. Republican) with 85% accuracy, sexual orientation among men with 88% accuracy and among women with 75% accuracy, and relationship status with 65% accuracy. Certain “likes” stood out as having particularly high predictive ability for Facebook users’ personal characteristics. For instance, the best predictors of high intelligence were “The Colbert Report,” “Science,” and, unexpectedly, “curly fries.” Conversely, low intelligence was indicated by “Sephora,” “I Love Being a Mom,” “Harley Davidson,” and “Lady Antebellum.”

In the area of personality, two studies found that users’ extraversion can be most accurately inferred from Facebook profile activity (e.g., group membership, number of friends, number of status updates); neuroticism, conscientiousness, and openness to experience can be reasonably inferred; and agreeableness cannot be inferred at all. In other words, Facebook activity renders extraversion highly visible and agreeableness opaque.

Language can also be used to predict online communicators’ identity, as shown by Andrew Schwartz and his colleagues in a study of 15.4 million Facebook status updates, totaling over 700 million words. Language choice, including words, phrases, and topics of conversation, was used to predict users’ gender, age, and Big Five personality traits with high accuracy.

In sum, this body of research suggests that it is possible to infer many facets of Facebook users’ identity through automated analysis of their online activity, regardless of whether they explicitly choose to divulge this identity. While users typically choose to reveal their gender and ethnicity, they can be more reticent in disclosing their relational status or sexual orientation and might themselves be unaware of their personality traits or intelligence quotient. This line of research raises important questions about users’ privacy and the extent to which this information, once automatically extracted from Facebook activity, should be used by corporations for marketing or product optimization purposes.

Real and Imagined Audience for Identity Claims

The purpose of many online identity claims is to communicate a desired image to an audience. Therefore, the process of identity construction involves understanding the audience and targeting messages to them. Social media, such as Facebook and Twitter, where identity claims are posted very frequently, pose a conundrum in this regard, because audiences tend to be unprecedentedly large, sometimes reaching hundreds and thousands of members, and diverse. Indeed, “friends” and “followers” are accrued over time and often belong to different social circles (e.g., high school, college, employment). How do users conceptualize their audiences on social media platforms? Are users’ mental models of their audiences accurate?

These questions were addressed by Michael Bernstein and his colleagues in a study focusing specifically on Facebook users. The study used a survey methodology, where Facebook users indicated their beliefs about how many of their “friends” viewed their Facebook postings, coupled with large-scale log data for 220,000 Facebook users, where researchers captured the actual number of “friends” who viewed users’ postings. Results show that, by and large, Facebook users underestimated their audiences. First, they believed that any specific status update they posted was viewed, on average, by 20 “friends,” when in fact it was viewed by 78 “friends.” The median estimate for the audience size for any specific post was only 27% of the actual audience size, meaning that participants underestimated the size of their audience by a factor of 4. Second, when asked how many total audience members they had for their profile postings during the past month, Facebook users believed it was 50, when in fact it was 180. The median perceived audience for the Facebook profile, in general, was only 32% of the actual audience, indicating that users underestimated their cumulative audience by a factor of 3. Slightly less than half of Facebook users indicated they wanted a larger audience for their identity claims than they thought they had, ironically failing to understand that they did in fact have this larger audience. About half of Facebook users indicated that they were satisfied with the audience they thought they had, even though their audience was actually much greater than they perceived it to be. Overall, this study highlights a substantial mismatch between users’ beliefs about their audiences and their actual audiences, suggesting that social media environments are translucent, rather than transparent, when it comes to audiences. That is, actual audiences are somewhat opaque to users, who as a result may fail to properly target their identity claims to their audiences.

Family Identity

One critical aspect of personal identity is family ties. To what extent do social media users reveal their family connections to their audience, and how do family members publically talk to one another on these platforms? Moira Burke and her colleagues addressed these questions in the context of parent-child interactions on Facebook. Results show that 37.1% of English-speaking US Facebook users specified either a parent or child relationship on the site. About 40% of teenagers specified at least one parent on their profile, and almost half of users age 50 or above specified a child on their profile. The most common family ties were between mothers and daughters (41.4% of all parent-child ties), followed by mothers and sons (26.8%), fathers and daughters (18.9%), and least of all fathers and sons (13.1%). However, Facebook communication between parents and children was limited, accounting for only 1–4% of users’ public Facebook postings. When communication did happen, it illustrated family identities: Parents gave advice to children, expressed affection, and referenced extended family members, particularly grandchildren.

Cultural Identity

Another critical aspect of personal identity is cultural identity. Is online communicators’ cultural identity revealed by their communication patterns? Jaram Park and colleagues show that Twitter users create emoticons that reflect an individualistic or collectivistic cultural orientation. Specifically, users from individualistic cultures preferred horizontal and mouth-oriented emoticons, such as :), whereas users from collectivistic cultures preferred vertical and eye-oriented emoticons, such as ^\_^. Similarly, a study of self-expression using a sample of four million Facebook users from several English-speaking countries (USA, Canada, UK, Australia) shows that members of these cultures can be differentiated through their use of formal or informal speech, the extent to which they discuss positive personal events, and the extent to which they discuss school. In sum, this research shows that cultural identity is evident in linguistic self-expression on social media platforms.

Cross-References

[Anonymity](http://link.springer.com/search?facet-content-type=ReferenceWorkEntry&query=Anonymity)

[Behavioral Analytics](http://link.springer.com/search?facet-content-type=ReferenceWorkEntry&query=Behavioral%20Analytics)

[Facebook](http://link.springer.com/search?facet-content-type=ReferenceWorkEntry&query=Facebook)

[Privacy](http://link.springer.com/search?facet-content-type=ReferenceWorkEntry&query=Privacy)

[Profiling](http://link.springer.com/search?facet-content-type=ReferenceWorkEntry&query=Profiling)

[Psychology](http://link.springer.com/search?facet-content-type=ReferenceWorkEntry&query=Psychology)

[Twitter](http://link.springer.com/search?facet-content-type=ReferenceWorkEntry&query=Twitter)

Further Readings

Bachrach, Y., et al. (2012). Personality and patterns of Facebook usage. In *Proceedings of the 3rd Annual Web Science Conference* (pp. 24–32). Association for Computing Machinery.

Bernstein, M., et al. (2013). Quantifying the invisible audience in social networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 21–30). Association for Computing Machinery.

Burke, M., et al. (2013). Families on Facebook. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)* (pp. 41–50). Association for the Advancement of Artificial Intelligence.

Das, S., & Kramer, A. (2013). Self-censorship on Facebook. In *Proceedings of the 2013 Conference on Computer-Supported Cooperative Work* (pp. 793–802). Association for Computing Machinery.

Kern, M., et al. (2014). The online social self: An open vocabulary approach to personality. *Assessment, 21*, 158–169.

[CrossRef](http://dx.doi.org/10.1177/1073191113514104)

Kosinski, M., et al. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences, 110*, 5802–5805.

[CrossRef](http://dx.doi.org/10.1073/pnas.1218772110)

Kramer, A., & Chung, C. (2011). Dimensions of self-expression in Facebook status updates. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)* (pp. 169–176). Association for the Advancement of Artificial Intelligence.

Park, J., et al. (2014). Cross-cultural comparison of nonverbal cues in emoticons on twitter: Evidence from big data analysis. *Journal of Communication, 64*, 333–354.

[CrossRef](http://dx.doi.org/10.1111/jcom.12086)

Schwartz, A., et al. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS One, 8*, e73791.

[CrossRef](http://dx.doi.org/10.1371/journal.pone.0073791)