Writer Profiling Without the Writer's Text

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky

Stanford University, Stanford, CA 94095, USA,
{jurgens,tsvetkov,jurafsky}@stanford.edu

Abstract. Social network users may wish to preserve their anonymity online by masking their identity and not using language associated with any particular demographics or personality. However, they have no control over the language in incoming communications. We show that linguistic cues in public comments directed at a user are sufficient for an accurate inference of that user's gender, age, religion, diet, and even personality traits. Moreover, we show that directed communication is even more predictive of a user's profile than the user's own language. We then conduct a nuanced analysis of what types of social relationships are most predictive of users' attributes, and propose new strategies on how individuals can modulate their online social relationships and incoming communications to preserve their anonymity.

1 Introduction

Communication is the crux of online social platforms and the messages people write can reveal substantial information about their identity, such as their demographic attributes, personality, location, native language, or socioeconomic status. Knowledge of a person's identity benefits many downstream applications including commercial ones, which has led to a significant effort to develop methods that infer an author's latent attributes automatically from their writings. Most profiling and demographic inference methods focus on the text an individual writes. However, individuals also communicate directly with others, raising the question of how much *incoming* messages reveal about a recipient. Further, such directed speech also raises an important privacy concern: although people can opt to self-censor information to reveal less of their identity through the statements they make [63, 2, 62, 85], a person does not control what their friends say to them, potentially exposing much about their identity. Such directed speech can be highly revealing of the individual's identity and social relationships, as shown in Figure 1. Here, we measure to what degree incoming messages sent to an individual reveal their personal attributes and whether privacy-seeking individuals can obfuscate their own information when they cannot control the content they receive.

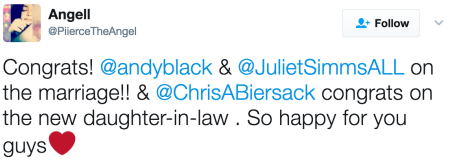


Fig. 1. A real demographically-revealing directed message that conveys age, gender, marital status, and familial relations of the recipients.

Prior work on demographic inference has largely focused on analyzing either the individual directly through the text they produce [20] or information from their social network, where friends are assumed to have similar attributes [1, 3]. Here, we consider a third source of information from directed communications *to* an individual. We hypothesize that directed communication affords multiple channels for revealing personal information, as it has been shown that peers construct a shared identity through language choice [17], that particular expressions are said more to particular social categories, e.g., biased language [33, 25], and that language expresses an asymmetric social relationship linked with social differences, e.g., condescension [92]. Our work offers an important complementary approach to profiling and demographic inference that is applicable even when individuals reduce the amount of personally-identifying statements they make on social media [82, 85] or even employ recent techniques to adversarially prevent their identity from being inferred [77, 114, 15, 112, 88].

Our work provides three main contributions:

- We establish that language in incoming communications is a highly-accurate source of demographic and personal information about an individual. To this end, we develop classifiers for five diverse demographic attributes: age, gender, religion, diet, and personality (§3). Our classifiers—trained on directed communications—learn to discriminate linguistic cues that are closely aligned with personal attributes of the individual to whom these messages are directed. These cues also have close correspondence with prior studies on profiling and demographic inference analyzing linguistic traits characteristic to the social group of the recipient.
- We establish, for the first time, a statistical relationship between (i) the efficacy of incoming communications in user profiling and demographic inference, and (ii) the strength of social ties of interlocutors (§4). We show that incoming communications from an individual's strong ties are more revealing of the individual's identity, but that this relationship only holds for publicly-visible aspects of the identity.
- We propose novel adversarial strategies for individuals to use for preserving their privacy (§5). We demonstrate that effective adversarial behavior is possible by strategic recruitment of new peers.

More broadly, our work captures the dual use of Natural Language Processing (NLP) techniques [50] by providing new demographic inference techniques that enable NLP for Social Good applications like mental health assessment [9], while also providing adversarial solutions for individuals who want to participate in social media but need to maximize anonymity for safety reasons [39].

2 Language and Identity

Individuals frequently reveal aspects of their social identity through their language choice and style. These choices help individuals associate themselves with a particular identity [37, 17] and in conversation, language choice helps establish a social common ground and implicitly denotes mutual membership in a common social category [55, 34]. Following, we outline some such demographic aspects and how they manifest in speech and social interactions.

2.1 Linguistic Signals of Personal Identity

Individuals make linguistic choices in order to signal their membership in social categories. These categories range from observable demographic attributes like age and gender to those associated with choice like diet and political affiliation. Here, we study five possible categories: gender, age, religion, diet, and extroversion. These categories capture a broad spectrum of possible attributes that affect language choice and include extrinsic, publicly-visible categories (age and gender) and intrinsic, private categories that are not necessarily publicly known by peers (diet, religion, extroversion). These attributes also reflect different computational tasks by including two binary variables (gender and extroversion[1]), two categorical variables (diet, religion) and a continuous-valued variable (age).

Gender Gender is known to be one of the most important social categories driving language choice and its study has a long tradition within sociolinguistics [104, 105, 52, 24, 37, 56, 25]. More recent work has begun to examine how gender is expressed in platforms such as Twitter, where the lack of prosodic and non-verbal cues gives rise to other forms of linguistic variation; these purely textual signals have enabled large-scale studies of gender signaling [87, 18, 78, 94, 10, 23, 41, 6, 111, 19, 43, *inter alia*]. These studies have found a broad range of style and content differences. For example, women are more likely to use pronouns, emotion words (like *sad*, *love*, and *glad*), interjections (*ah*, *hmmmm*, *ugh*), emoticons, and abbreviations associated with online discourse (*lol*, *omg*), while men tend to use higher frequency standard dictionary words, proper names (e.g., the names of sports teams), numbers, technology words, and links.

Age Individuals make language choices that signal their age, which may be intended to convey the speaker's maturity or express the stage in life [35]. As with gender, textual communication on social media has allowed building computational models for predicting age from language usage, including Twitter [87, 72, 73], blogs [90, 47], Facebook [93], and Netlog, a Dutch social network platform akin to Facebook [78]. The models show clear differences in linguistic choices, with younger individuals performing more stylistic variation like elongation and capitalization [47, 7], grammatical differences in sentence length and construction [49], and content choices to include more self-references, slang, and acronyms [90, 73, 87]. However, for older individuals these differences are less pronounced [74], which obstructs age inference solely from text.

Diet Individuals adopt self-imposed dietary restrictions for a variety of medical, religious, or ethical considerations. The impact of this choice may expressed in the topics they discuss. Prior work has primarily focused on identifying vegetarians [8, 38], suggesting such individual are identifiable from topical features.

Religion Religious affiliation provides a social construct that individuals may identify with, akin to race or ethnicity [5, 113]. Affiliations frequently provide metaphors and terminology that work their way into the regular lexicon, which automated methods have leveraged for classifying individuals in social media [75, 21].

Extroversion Extroversion is a strong predictor of social engagement with one's peers and increased social status [4], though the degree of extroversion or introversion has

[1] We note that both gender and extroversion may also be considered along a spectrum [37, 36]. We opt to study these as binary variables here due to lack of continuous-valued gender and extroversion ratings for social media users.

not been shown to affect the amount of self-disclosure [95]. Automated methods for personality detection have shown that extroverted individuals tend to use more terms describing social activities and concepts and colloquial language, whereas introverts refer to more solitary activities [30, 45, 86, 96, 53]

2.2 Social Identity in Communication

While an individual's own speech is predictive of their identity, the communication they receive is also potentially predictive due to the social processes that drive language selection. Following the principle of homophily, individuals tend to have social relationships with others similar to themselves in interests and demographics [66]; the communication within these relationships often focuses on the common ground [91], which can be used to identify their shared demographics. The process of revealing demographically-identifiable information is further supported by the tendency of individuals to reciprocate in self-disclosure in conversation, which provides more evidence of shared identity [29].

2.3 Online Identity and Privacy Protection

Public communication in social media can reveal significant information about a person, which has led some individuals to change their communication strategies to preserve privacy [102]. Individuals modulate these strategies relative to their closeness with the peer and are less likely to self-censor when talking with close friends [103]. In addition, anonymization strategies vary both demographically, e.g., females are more likely to use misinformation to preserve their privacy [77, 114], and culturally, e.g., cultures with collectivist tendencies tend to censor less [89, 109]. Despite these privacy efforts, an individual's identity may still be revealed by others' communication about or to them, e.g., parents compromising their children's privacy online by revealing their age and location [68], or friends revealing a person's religion or relationship status [108]. Our work extends this line of research by examining what can be inferred about a person from both explicit and implicit signals in the communications they receive.

3 Profiling via Incoming Communications

Given that a person's identity is expressed through language, we first examine to what degree *incoming* communications received by an individual are predictive of that user's personal attributes discussed in §2.1.

3.1 Data

Individuals for each demographic attribute were collected using targeted queries of the Twitter platform. For gender, we use fixed patterns on user profiles to find individuals who explicitly self-identify with a gender, e.g., "Writer in NY; she/her" or who identify with gendered social roles, e.g., "father to two girls." Age is identified using fixed patterns with aggressive filtering to remove noise. Diet was collected in a similar manner

Attribute	# of Tweets	Majority Class	%
Gender	59800	Male	52.5
Religion	19940	Christian	65.8
Extroversion	24576	Introvert	63.0
Diet	9001	Unrestricted	41.0
Age	38134	21.3 (mean)	5.7 (s.d.)

Table 1. Dataset sizes for each demographic attribute and frequencies of the majority classes.

as in El-Arini *et al.* [38] by identifying individuals who report themselves as vegetarian, vegan, or paleo;[2] we sample an equal number of individuals not reporting these diets and treat them as being examples of the unrestricted diet class. We follow Chen *et al.* [21] and identify individuals' religious affiliations by searching for a fixed set of terms in the user profile for the following affiliations : agnostic, atheist, Buddhist, Christian, Hindu, Jewish, Muslim.[3] For personality, we adopt the approach of Plank and Hovy [83] for gathering Myers-Briggs personality type indicators and its Introversion/Extroversion labels, which have been shown to strongly load on the Extroversion dimension of the more-commonly used Big Five personality assessment [51, 65].

Targeted queries were used to find all individuals with matching profiles or tweets during March 2016, except for Extroversion which was queried from January 2010 to December 2016 due to its relative sparsity. Tweets for identified individuals were then collected from a 10% random sample of Twitter from 2014 to 2016. The CLD2 language detector [64] was used to retain only English-speaking individuals. Finally, only those users with at least 100 tweets directed to them are included in the dataset. Table 1 summarizes the resulting dataset.

3.2 Personal Attribute Classifiers

Features An individual's associated text is represented using content and stylistic features drawn from prior work. The broad themes are represented using a 100-topic LDA model [12], capturing both the average topic distribution for a message and the maximum probability a topic ever receives. A lexicon learned for each attribute was constructed by ranking all unigrams and bigrams using the weighted log-odds-ratio with an informative Dirichlet prior [70]. To construct binary classes for computing the log-odds of multiclass attributes, we chose one attribute relative to all the other attributes in the class; for age, lexicons were created by discretizing age into decade ranges (e.g., age 20-29) and computing the log-odds for that decade relative to the others. Content was additionally categorized using the Linguistic Inquiry and Word Count (LIWC) dictionary [106] and the average GloVe vector computed from all words received by an individual [81].

Stylistic features include pronoun usage, disfluencies, laughing expressions, question frequencies, average word length, usage of capital letters, word lengthening, and punctuation [80, 7, 47, 87, 90, 73, 23]. We also include emoji usage, which are language-independent signals that carry social status [107, 60], a general lexicon for sentiment

[2] We note that other diets are possible, such as kosher or halal; however, these are closely related to religion, which we also study, so we intentionally exclude them.

[3] Additional queries were formed for Sikhism and Jainism which did not return sufficient numbers of English speaking individuals to be included.

[69] and second sentiment lexicon focused on the extremes [61], which was shown to be effective for distinguishing different personality types, and lexicons with concreteness and abstractness ratings [16]. Ultimately, 2,625 features are used.

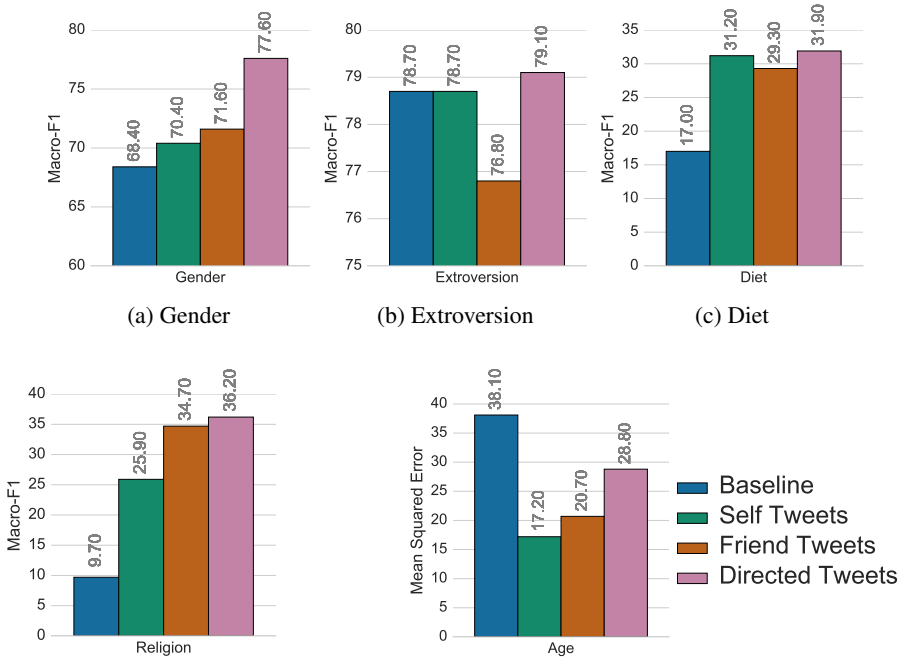
Models Categorical attributes are predicted using random forest classifiers [14], an ensemble of decision tree classifiers learned from many independent subsamples of the training data; age is predicted using a random forest regressor. Random forest ensembles are particularly suitable for imbalanced multilabel setups such as ours and have been shown to be robust to overfitting when using many features [40]. Separate models were trained for each attribute. Models are evaluated using ten-fold cross-validation using Macro- and Micro-F1 for categorical attributes (see Appendix A) and the numeric age predictions are evaluated using Mean Squared Error and Pearson's r.

Baselines We evaluate against two systems that use the same textual features but calculate them based on (1) the individual's outgoing language, rather than directed to them and (2) the language of the individual's peers, not necessarily directed towards the individual. For both baselines, we use the same ground truth individuals and recalculate the log-odds lexicons and topic models on their respective text. These models capture the information available through self-disclosure and from homophily, respectively, and are broadly representative of many related works for each attribute. To ensure a fair comparison with the incoming-text model, we evaluate each model on individuals that have 100 tweets authored by themselves or their peers, respectively. Finally, we include a baseline system that predicts the most frequent class or the mean numeric value.

3.3 Results

The language of incoming communications was highly predictive of recipients' demographic attributes, matching or surpassing the performance of the recipient's own speech and that of their peers for all attributes but age, as shown in Figure 2. The performance of the directed speech classifier for each attribute is close to current state-of-the-art methods, e.g., [22], which typically also include features from the social network, biography, and other sources. Performance across attributes varied widely; the highest improvement relative to other forms of communication was seen for gender. This is expected, as individuals may be referred to by gendered categories in discourse. For example, the following tweets provide a clear lexical signal: "@User bro u don't even know my squad lol tf" and "@User Sounds great! Are you ladies going to #DisruptHRTO?". Our results build upon multiple findings from sociolinguistics that directed speech can reveal significant information by individuals accommodating linguistic style of their peers [76, 26], that individuals sharing the same attribute are more likely to use a common vernacular that indicate in-group status [67], and that dialog is conducive to self-disclosure [29].

Examining the feature importance for each attribute's classifier, we find that features based on the log-odds bootstrapped lexicons, topical differences, and the average word vector of received messages account for the majority of the most discriminative features. This trend matches the prior observation that machine-learned topic features are highly effective at distinguishing between demographics [79]. Examining the log-odds bootstrapped lexicons, the words most biased towards particular demographic attributes mirrored categories seen in self-speech, such as speech towards women including more



(a) Gender (b) Extroversion (c) Diet

(d) Religion (e) Age

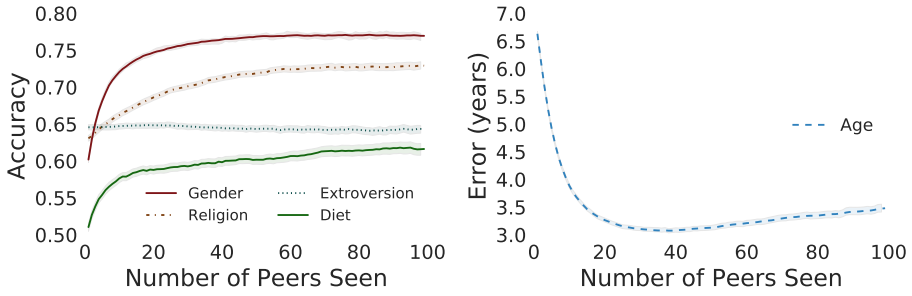
Fig. 2. Predictive accuracy for each attribute, reported as Macro-F1 and Mean Squared Error. Additional metrics are reported in Tables 4 and 5 in Appendix B.

emotion words and towards younger ages including more non-standard spellings and abbreviations. In addition, the log-odds lexicons captured topical preferences by gender, such as school terms for younger demographics, foods associated with the different diets, and religion-affiliated language. Table 2 shows examples of words from received communication that are most-biased towards particular demographic attributes. When an attribute's log-odds lexicon was highly weighted by the classifier, we found significant overlap with its most-biased words and the linguistic cues expected in self-speech when an individual signals their own social category. This overlap suggests that these demographically-biased words represent a common ground for two individuals with shared identity [37, 17] and the relative importance of such words for classification indicates they are key features as individuals signal their identity in dialog.

As a follow-up experiment, we calculated the learning curve for each classifier to test how many peers are needed to obtain high performance. Curves were estimated by repeatedly sampling up to 100 peers for each user and estimating performance. Learning curves, shown in Figure 3, reveal that most performance gains are seen with just a few peers and diminishing gains after 20 friends. These trends suggest that attributes can be reliably inferred from limited peer evidence, which substantially reduces the data collection effort needed to per user.

Attribute		Most Salient Words in Incoming Communications	Characteristic Cues Used By People to Signal Their Own Attributes
Gender	Male	*team, mate, coach, players, nfl, his, teams, games, player, football, man, matt, play*	Frequent words, names, sports teams.
	Female	*her, love, she, beautiful, gift, entered, girl happy, thank, lovely, mom, christmas, cute*	Pronouns, emotion words, interjections.
Religion	Christian	*pjnet, catholic, obama, deplorable, trump, christian, amen, bless, church, america, prolife*	Words from a particular religious affiliation.
	Atheist	*atheist, atheism, atheists, evolution, shit, fuck, science, evidence, fucking*	More words from scientific and political topics.
Age	10–19	*birthday, happy, whooo, via, iloveyou, coolest, hemmings, stepfather, wvu, thanks, gotham*	More stylistic and grammatical variation, self-references, slang, and acronyms.
	30–39	*trump, obama, great, verified, win, news, deplorable, daily, latest, john, book, read*	Less linguistic variation, speaker's maturity.

Table 2. Examples of the most salient words used in directed speech (incoming communications) towards people with a particular attribute, learned from log-odds with a Dirichlet prior [70]. The right column lists what linguistic cues are expected to be observed in individual's texts (outgoing communications), based on prior work, summarized in §2.1. Strong linguistic and topical correspondence between outgoing and incoming communications enables training effective profiling classifiers without any text produced by a person, as proposed in §3.



(a) Gender, Religion, Diet, and Extroversion

(b) Age

Fig. 3. Learning curves of how predictive accuracy changes relative to how many peers' communications are used for prediction. Shaded regions show bootstrapped 95% confidence intervals.

4 Demographic Inference and Tie Strength

Individuals engage with a variety of users on Twitter, from close friends to complete strangers; these different types of relationships can be categorized into strong and weak social ties [48]. Communication is driven in part by the strength of social ties [13]. For example, close friends may be more likely to discuss more intimate and immediate topics (e.g., dinner plans), whereas less-familiar acquaintances may use more formal language and relate to the person less as an individual and more as their social categories (e.g., gender, political affiliation) [100, 84, 57]. This observation motivates our research hypothesis that *the message content of close friends reveals more of the recipient's demographics than the more stereotypical language of socially-distant peers*. In this section, we use statistical inference to examine whether the directed speech generated from strong ties is more predictive of the recipient's identity.

	generalized logistic mixed-effects (Acc.)				*linear mixed-effects* (Err.)
	Gender	Religion	Diet	Extroversion	Age
Communication Frequency	0.032^{**}	-0.049	0.147	0.105	-0.190^{***}
Relationship: peer-follows	0.264^{***}	0.365	0.303	-0.179	-0.051
Relationship: ego-follows	0.270^{**}	0.336	-0.235	-0.263	0.085
Relationship: reciprocal	0.243^{***}	0.556^{***}	0.059	-0.056	-0.535^{***}
Intercept	0.687^{***}	0.384^{**}	0.510^{**}	1.584^{***}	4.539^{***}
Marginal R^2	1.713e-03	1.185e-03	8.123e-04	2.151e-04	3.617e-03
Conditional R^2	0.352	0.937	0.905	0.948	0.836
Observations	35,962	10,408	5,000	22,819	12,850

Note: *p<0.1; **p<0.05; ***p<0.01

Table 3. Regression models showing the impact of increased tie strength (calculated via proxies discussed in §4.1) on a peer's predictive accuracy. The categorical variable for Relationship has the reference coding of "No Relationship" for all models. Note that since the Age regression is on Error instead of accuracy, its coefficients' interpretations are reverse of those of other attributes.

4.1 Tie Strength

Tie strength in social relationships has many dimensions [54]. The four dimensions originally introduced by Granovetter are the amount of time in communication, intimacy, intensity, and reciprocity which characterize the tie [48]. Other dimensions were proposed later, for example social distance [58] and types of social relationships [42]. To capture the tie strength between the speaker and recipient, we use the following proxies, following prior work (also see Appendix C). **Communication frequency** [48] is operationalized by the number of tweets sent. **Explicit social relationships** [42] is captured by crawling the friend and followers edges of the individuals on the Twitter network. We consider four types of relationships: (1) the peer follows the individual, (2) the individual follows the peer, (3) both follow each other, (4) no edges.[4]

4.2 Statistical Inference

We test for the effect of social ties on predictiveness of an individual's identity by fitting mixed-effect models with random effects for the individual receiving the messages, and fixed-effects models for each proxy for tie strength. Random effects control for variation in the relative predictiveness across individuals. The number of messages sent by peers follows a power-law distribution, so we log-transform the message counts to avoid scaling issues during fitting. Generalized logistic mixed-effect models were fit to the binary dependent variable of whether the peer predicts the correct attribute; linear mixed effect models were fit to the absolute value of the error in age. We measure the amount of variance explained by the mixed-effect models by calculating R^2 using the method of Nakagawa and Schielzeth [71]; here, the marginal R^2 describes the proportion of variance explained by only the fixed effects (age and social distance); while the conditional R^2 measure variance from the fixed and random effects. To avoid data sparsity, we fit the models only with peers who send at least ten messages to a person.

[4] Due to Twitter API rate limits, full edge information was gathered only for 1.7M pairs.

4.3 Results

Results are summarized in Table 3. Our models reveal that increased tie strength is significantly associated with a peer's predictiveness for gender and age. However, no relationship was found for the other three attributes, religion, diet, and extroversion. We speculate that associations were found for gender and age because these attributes are more easily observable: they are frequently signaled by cues such as an individual's username or profile picture and therefore visible to both strong and weak ties. In contrast, religion, diet, and extroversion can be construed as more internal and not necessarily evident to an individual's ties, regardless of strength; because these attributes are less known, peers are less likely to modulate their speech on the basis of them.[5]

For gender and age, the majority of proxies for tie strength had a statistically-significant positive association with increased accuracy (gender) or reduced error (age), confirming our hypothesis that *stronger* ties are more predictive. For age and gender, more communication and having a reciprocal social relationship were consistently positively associated with demographic predictiveness, with additional significant effects seen for gender when individuals have any form of explicit relationship. The Marginal R^2 indicates that tie strength explains only a small part of the variance in predictive accuracy; however, the communications of a single peer alone are unlikely to be highly accurate (cf. the learning curve in Figure 3), which limits establishing a larger fit from the fixed effects.

5 Preserving Anonymity through Adversarial Behavior

Online social platforms can serve a critical need for individuals to engage with others and obtain social, physical, and mental support [11, 28]. When discussing sensitive or controversial topics, individuals often aim to maintain some degree of privacy online [27, 97, 115]. However, in §3 we have shown that even if individuals self-censor their content or employ adversarial strategies to mask their identity [63, 2, 62, 85], the messages they receive can still reveal significant information about them.[6] This loss of privacy is potentially disastrous for individuals discussing politically-sensitive topics, as multiple reports have shown governments to pursue individuals when their identity is revealed [39].

Given the potential loss of privacy from other's incoming communications, we consider here how a user may still minimize what might be inferred about them. Under the reasonable assumption that an individual has no control over what their existing peers communicate to them (especially weak-tie peers), an alternative option a user has is to adversarially recruit new peers whose messages to them will mask the existing demographic signal. Without the ability to control what is said to them, an individual can no longer rely on adversarial stylometrics to hide their identity [63, 2, 62, 85]. Following, we evaluate the effectiveness of adversarial strategies and then discuss key technical challenges for operationalizing these strategies in real platforms.

[5] One possibility for testing this hypothesis in future work is to identify a cohort of individuals who publicly signal these variables in an explicit way (e.g,. including religious imagery in their profile picture) and then test for effects of tie strength on their peers' predictiveness.

[6] This risk is valid even if the individual themselves does not engage with others, as platforms such as Twitter allow anyone to directly message another unless banned.

5.1 Adversarial Strategies

We model peer adoption strategies as a friend-of-a-friend recruitment, where individuals have the option of selecting a peer using two parameters: (i) who the peer is and (ii) who the peer is communicating with. For notational simplicity we refer to the individual recruiting a new peer as u_i; u_i can select another peer u_k who is communicating with peer u_j. Peer communication is simulated as if u_k communicated with u_i instead of u_j.

Four adversarial strategies are tested: three folk strategies that an individual might feasibly use on the basis of public information and one strategy that requires knowledge of the classifier.

1. **Random Peer**: the user chooses a random user u_j and then receives communication from a random friend of u_j.
2. **Different-Attribute Peer**: the user chooses the peer u_k of a user u_j who has a different demographic attribute than themself; e.g., a woman would choose the peer of a man.
3. **Topic Difference**: the user chooses the peer u_k whose messages to u_j are the most topically dissimilar from the topics seen in the current conversations to u_i.[7]
4. **Feature Difference**: This strategy has knowledge of the exact features used by the classifier and the feature vector for the current individual. A new peer is chosen by selecting u_k whose messages to u_j would produce a feature vector that is maximally different from the vector for u_i.

When sampling multiple new peers, all strategies sample peers without replacement and the Topic Difference and Feature Difference strategies sample in decreasing order of distance (i.e,. the most-different are chosen first).

5.2 Experimental Setup

We repeat the classifier and experimental setup from Section 3 with separate models for each attribute. The effectiveness of adversarial behavior is tested using ten-fold cross-validation where the test fold data alone is used for simulating adversarial behavior; i.e., all new peers selected by an adversarial strategy are chosen from within the test set, which prevents test-train leakage. During testing, we sample an increasing number of new peers for each user and compute the classifier's accuracy based on the percentage of new peers added relative to the number of prior peers.

5.3 Results

Classifier performance, shown in Figure 4, reveals that adversarial strategies can be effective at reducing performance to chance levels (denoted with a horizontal red line) and even at flipping the perceived attribute of the individual. However, the number of new peers needed to attain these goals varied substantially by strategy and attribute. With knowledge of the classifier's features, the Feature Difference strategy is able to

[7] We note that while we measure topical difference using our LDA model for messages, the peers selected by maximizing topical difference would be easily identified as such by the layperson (e.g., a peer discussing completely different topics).

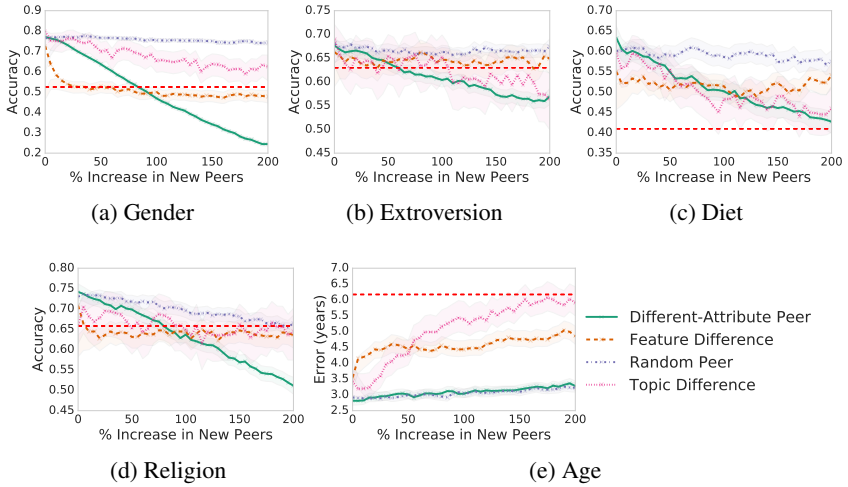


Fig. 4. Classifier accuracy for each adversarial strategy as new peers are chosen to communicate with an individual. Dashed red lines show the expected performance of randomly guessing for plots (a)-(d) and for guessing the mean age in (e). Shaded regions show 95% confidence intervals.

effectively reduce performance to near chance for all but the diet and age attributes; as Figures 4a, 4b, and 4d show, a 25% increase in the number of peers provides effective anonymity for those attributes.

Among the folk strategies, the Different-Attribute Peer and Topic Difference strategies are the most effective, though neither is able to consistently obtain at-chance performance. Both strategies are roughly equivalent for extroversion, diet, and religion, but diverge in their anonymization abilities for gender and age. For gender, Different-Attribute Peer is able to obtain at-chance performance with an 80% increase in new peers, while Topic Difference did not converge to at-chance even after tripling the number of peers a user has. However, for age, Topic Difference provides the highest effectiveness of any strategy. However, overall, folk strategies required far more new peers on average to be effective than the classifier-aware strategy, which is potentially prohibitive for users who already have a large cohort of peers messaging them.

Age and diet are difficult attributes to attain chance-level performance in our dataset. Age obfuscation is made difficult by the age distribution of the platform and textual differences between age groups. Individuals in Twitter are known to skew younger [31] (as also seen in our dataset) and older age groups are known to have fewer textual differences between them [73]; therefore, we speculate that it is difficult to select peers adversarially since there are fewer peers communicating with older individuals whose content is sufficiently dissimilar. We also note that the Different-Attribute Peer strategy could likely be improved by selecting u_j to maximize the difference in age, rather than simply selecting a u_j that has a different age. We speculate that obfuscating diet is difficult due to the nature of the unrestricted-diet class. Vegans, vegetarians, and paleo diet practitioners all have specific dietary restrictions and the topic of these restrictions serves as discriminating features; in contrast, the unrestricted diet is defined by the

absence of these, so new peers with this attribute are much less likely to have a strong lexical signal about their diet that would change the classifier's prediction.

5.4 Discussion

Our results demonstrate that adversarial strategies can be effective in theory. However, implementing these strategies in practice faces two key challenges. First, our experiment simulates adding new peers as if the content they directed to u_j had been directed to u_i. This process assumes that new peers would still communicate with the same type of content as if they had been talking to u_j; e.g., if a female adds the peer of a male, the new peer would need to talk to her as if she were that male. In addition, new peers are expected to message with the same frequency, which is potentially unlikely when engaging with strangers.

Second, optimal adversarial behavior should be undetectable by the observing party. However, structural properties of the underlying social network of Twitter can potentially reveal the adversarially-added peers. Specifically, adding new peers would likely add individuals from distant parts of the network making them easier to detect. As a result, an adversarial user must add new peers whose placement in the social network is similar to that of their current peers, which could significantly restrict the available poor of new peers.

We speculate that for the truly adversarial, one option is to create and use sock puppet accounts. These accounts can easily be managed to control their content and to follow existing peers in order to seamlessly integrate into a user's social network network. A second possibility for mitigating these challenges is changing the self-presentation signals of the account, such as selecting a gender-neutral profile picture and username. These strategies minimize the social signaling of identity in other domains (e.g., job applications, email) and are known to change peers' behaviors [101, 46].

Finally, our findings have important implications for related work in differential privacy where, when releasing data about users, their privacy is preserved by strategically manipulating (or adding noise to) text that each produces [32]. Our work demonstrates that in order to preserve the anonymity of a person's demographic attributes, a differential privacy system would need to modify any queries requesting the directed speech to an individual as well—not just the individual's own speech.

6 Conclusion

We have long known that what you say on social media reveals your identity and may compromise your privacy. Our work shows that even ignoring what you say, just looking at what your friends say to you is generally even more informative, and allows us to guess your gender with 80% accuracy, as well as your age (71.2%), and even your private attributes like your religion (74.4%) and personality traits (67.3%). Moreover, your closer friends reveal more about your publicly-visible identity than more distant ones, but not about your private attributes. Strategic selecting new peers to communicate with can obstruct profiling, but more work needs to be done to safeguard privacy. Code and data used in the experiments are at `https://github.com/davidjurgens/profiling-by-directed-speech`.

Acknowledgments

We thank the anonymous reviewers, SocInfo organizers, the Stanford Data Science Initiative, and Twitter and Gnip for providing access to part of data used in this study. This work was supported by the National Science Foundation through awards IIS-1159679 and IIS-1526745.

Appendix

A Classification Metrics

Macro-averaged F1 denotes the average F1 for each class, independent of how many instances were seen for that label. Micro-averaged F1 denotes the F1 measured from all instances and is sensitive to the skew in the distribution of classes in the dataset.

B Additional Classifier Results

	Baseline		Self Tweets		Friend Tweets		**Directed Tweets**	
Gender	52.0	68.4	75.2	70.4	71.7	71.6	**79.8**	**77.6**
Religion	63.2	9.7	71.5	25.9	**78.9**	34.7	74.4	**36.2**
Diet	51.3	17.0	55.7	31.2	54.2	29.3	**62.6**	**31.9**
Extroversion	64.8	78.7	65.3	78.7	64.2	76.8	**67.3**	**79.1**

Table 4. Predictive accuracy for categorical attributes, reported as Micro-F1 and Macro-F1.

	Baseline		Self Tweets		Friend Tweets		**Directed Tweets**	
Age	38.1	0.00	**17.2**	**0.27**	20.7	0.26	28.8	0.25

Table 5. Predictive accuracy for age, reported Mean Squared Error and Correlation.

C Additional Measures of Tie Strength

We initially considered two other potential proxies for tie strength based on textual analysis. First, we replicated the approach of Gilbert and Karahalios [44] which counted words occurring in ten LIWC categories to approximate intimacy in communication. Second, we attemped to measure social distance [58] by drawing upon Construal Theory [59, 110] which conjectures that individuals with low social distance typically use more concrete language, whereas those with high social distance use more abstract language [99, 98]; here, communication concreteness was measured using the word concreteness ratings of [16]. However, we found that the ratings for each approach did not match our judgments for their respective intended attributes and their use in the regression models produced non-significant results. Without ground truth for intimacy and social distance to validate their ratings, we therefore omitted these proxies based on our judgment of their unreliability to avoid drawing false conclusions about these dimensions of tie strength.

References

1. Al Zamal, F., Liu, W., Ruths, D.: Homophily and latent attribute inference: Inferring latent attributes of Twitter users from neighbors. In: Proc. ICWSM (2012)
2. Almishari, M., Oguz, E., Tsudik, G.: Fighting authorship linkability with crowdsourcing. In: Proc. COSN. pp. 69–82. ACM (2014)
3. Altenburger, K.M., Ugander, J.: Bias and variance in the social structure of gender. arXiv preprint arXiv:1705.04774 (2017)
4. Anderson, C., John, O.P., Keltner, D., Kring, A.M.: Who attains social status? effects of personality and physical attractiveness in social groups. Journal of personality and social psychology 81(1), 116 (2001)
5. Baker, W., Bowie, D.: Religious affiliation as a correlate of linguistic behavior. University of Pennsylvania Working Papers in Linguistics 15(2), 2 (2010)
6. Bamman, D., Eisenstein, J., Schnoebelen, T.: Gender identity and lexical variation in social media. Journal of Sociolinguistics 18(2), 135–160 (2014)
7. Barbieri, F.: Patterns of age-based linguistic variation in American English. Journal of sociolinguistics 12(1), 58–88 (2008)
8. Beller, C., Knowles, R., Harman, C., Bergsma, S., Mitchell, M., Van Durme, B.: I'm a Belieber: Social roles via self-identification and conceptual attributes. In: Proc. ACL. pp. 181–186 (2014)
9. Benton, A., Mitchell, M., Hovy, D.: Multitask learning for mental health conditions with limited social media data. In: Proc. EACL (2017)
10. Bergsma, S., Van Durme, B.: Using conceptual class attributes to characterize social media users. In: Proc. ACL (2013)
11. Best, P., Manktelow, R., Taylor, B.: Online communication, social media and adolescent wellbeing: A systematic narrative review. Children and Youth Services Review 41, 27–36 (2014)
12. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. Journal of Machine Learning Research (JMLR) 3(Jan), 993–1022 (2003)
13. Bogardus, E.S.: A social distance scale. Sociology & Social Research (1933)
14. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
15. Brennan, M., Afroz, S., Greenstadt, R.: Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. ACM Transactions on Information and System Security (TISSEC) 15(3), 12 (2012)
16. Brysbaert, M., Warriner, A.B., Kuperman, V.: Concreteness ratings for 40 thousand generally known English word lemmas. Behavior research methods 46(3), 904–911 (2014)
17. Bucholtz, M., Hall, K.: Identity and interaction: A sociocultural linguistic approach. Discourse studies 7(4-5), 585–614 (2005)
18. Burger, J.D., Henderson, J., Kim, G., Zarrella, G.: Discriminating gender on twitter. In: Proc. EMNLP. pp. 1301–1309 (2011)
19. Carpenter, J., Preotiuc-Pietro, D., Flekova, L., Giorgi, S., Hagan, C., Kern, M.L., Buffone, A.E., Ungar, L., Seligman, M.E.: Real men dont say "cute" using automatic language analysis to isolate inaccurate aspects of stereotypes. Social Psychological and Personality Science (2016)
20. Cesare, N., Grant, C., Nsoesie, E.O.: Detection of user demographics on social media: A review of methods and recommendations for best practices. arXiv preprint arXiv:1702.01807 (2017)
21. Chen, L., Weber, I., Okulicz-Kozaryn, A.: Us religious landscape on Twitter. In: Proc. SocInfo. pp. 544–560. Springer (2014)
22. Chen, X., Wang, Y., Agichtein, E., Wang, F.: A comparative study of demographic attribute inference in Twitter. Proc. ICWSM 15, 590–593 (2015)

23. Ciot, M., Sonderegger, M., Ruths, D.: Gender inference of Twitter users in non-English contexts. In: Proc. EMNLP. pp. 1136–1145 (2013)
24. Coates, J.: Language and gender: A reader. Wiley-blackwell (1998)
25. Coates, J.: Women, men and language: A sociolinguistic account of gender differences in language. Routledge (2015)
26. Danescu-Niculescu-Mizil, C., Gamon, M., Dumais, S.: Mark my words!: linguistic style accommodation in social media. In: Proc. WWW. pp. 745–754. ACM (2011)
27. De Choudhury, M., De, S.: Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In: Proc. ICWSM (2014)
28. De Choudhury, M., Kiciman, E.: The language of social support in social media and its effect on suicidal ideation risk. In: Proc. ICWSM. pp. 32–41 (2017)
29. Derlega, V.J., Harris, M.S., Chaikin, A.L.: Self-disclosure reciprocity, liking and the deviant. Journal of Experimental Social Psychology 9(4), 277–284 (1973)
30. Dewaele, J.M.: Individual differences in the use of colloquial vocabulary: The effects of sociobiographical and psychological factors. Learning vocabulary in a second language: Selection, acquisition and testing pp. 127–153 (2004)
31. Duggan, M.: Mobile messaging and social media 2015. Pew Research Center p. 13 (2015)
32. Dwork, C.: Differential privacy: A survey of results. In: Proc. TAMC. pp. 1–19. Springer (2008)
33. Eagly, A.H., Mladinic, A.: Gender stereotypes and attitudes toward women and men. Personality and Social Psychology Bulletin 15(4), 543–558 (1989)
34. Eckert, P.: Jocks and burnouts: Social categories and identity in the high school. Teachers College Press (1989)
35. Eckert, P.: Age as a sociolinguistic variable. The handbook of sociolinguistics pp. 151–167 (1997)
36. Eckert, P.: Variation and the indexical field. Journal of sociolinguistics 12(4), 453–476 (2008)
37. Eckert, P., McConnell-Ginet, S.: Language and gender. Cambridge University Press (2003)
38. El-Arini, K., Paquet, U., Herbrich, R., Van Gael, J., Agüera y Arcas, B.: Transparent user models for personalization. In: Proc. KDD. pp. 678–686. ACM (2012)
39. Elgin, B., Robison, P.: How despots use Twitter to hunt dissidents. *BloombergBusinessweek* https://www.bloomberg.com/news/articles/2016-10-27/twitter-s-firehose-of-tweets-is-incredibly-valuable-and-just-as-dangerous (2016)
40. Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D.: Do we need hundreds of classifiers to solve real world classification problems. J. Mach. Learn. Res 15(1), 3133–3181 (2014)
41. Flekova, L., Gurevych, I.: Can we hide in the web? large scale simultaneous age and gender author profiling in social media. In: Proc. CLEF (2013)
42. Friedkin, N.: A test of structural features of Granovetter's strength of weak ties theory. Social networks 2(4), 411–422 (1980)
43. Garimella, A., Mihalcea, R.: Zooming in on gender differences in social media. In: Proc. of the Workshop on Computational Modeling of Peoples Opinions, Personality, and Emotions in Social Media. pp. 1–10 (2016)
44. Gilbert, E., Karahalios, K.: Predicting tie strength with social media. In: Proc. CHI. pp. 211–220. ACM (2009)
45. Golbeck, J., Robles, C., Edmondson, M., Turner, K.: Predicting personality from twitter. In: Proc. SocialCom. pp. 149–156. IEEE (2011)
46. Goldin, C., Rouse, C.: Orchestrating impartiality: The impact of" blind" auditions on female musicians. Tech. rep., National bureau of economic research (1997)

47. Goswami, S., Sarkar, S., Rustagi, M.: Stylometric analysis of bloggers age and gender. In: Proc. ICWSM (2009)
48. Granovetter, M.S.: The strength of weak ties. American journal of sociology 78(6), 1360–1380 (1973)
49. Hovy, D., Søgaard, A.: Tagging performance correlates with author age. In: Proc. ACL. pp. 483–488 (2015)
50. Hovy, D., Spruit, S.L.: The social impact of natural language processing. In: Proc. ACL. vol. 2, pp. 591–598 (2016)
51. John, O.P., Srivastava, S.: The big five trait taxonomy: History, measurement, and theoretical perspectives. Handbook of personality: Theory and research 2(1999), 102–138 (1999)
52. Kendall, S., Tannen, D., et al.: Gender and language in the workplace. Gender and Discourse. London: Sage pp. 81–105 (1997)
53. Kosinski, M., Stillwell, D., Graepel, T.: Private traits and attributes are predictable from digital records of human behavior. Proceedings of the National Academy of Sciences (PNAS) 110(15), 5802–5805 (2013)
54. Krackhardt, D., Nohria, N., Eccles, B.: The strength of strong ties. Networks in the knowledge economy p. 82 (2003)
55. Labov, W.: Sociolinguistic patterns. University of Pennsylvania Press (1972)
56. Lakoff, R.T., Bucholtz, M.: Language and woman's place: Text and commentaries, vol. 3. Oxford University Press, USA (2004)
57. Lea, M., Spears, R., de Groot, D.: Knowing me, knowing you: Anonymity effects on social identity processes within groups. Personality and Social Psychology Bulletin 27(5), 526–537 (2001)
58. Lin, N., Ensel, W.M., Vaughn, J.C.: Social resources and strength of ties: Structural factors in occupational status attainment. American sociological review pp. 393–405 (1981)
59. Liviatan, I., Trope, Y., Liberman, N.: Interpersonal similarity as a social distance dimension: Implications for perception of others actions. Journal of experimental social psychology 44(5), 1256–1269 (2008)
60. Lu, X., Ai, W., Liu, X., Li, Q., Wang, N., Huang, G., Mei, Q.: Learning from the ubiquitous language: an empirical analysis of emoji usage of smartphone users. In: Proc. Ubicomp. pp. 770–780. ACM (2016)
61. Mairesse, F., Walker, M.A., Mehl, M.R., Moore, R.K.: Using linguistic cues for the automatic recognition of personality in conversation and text. Journal of Artificial Intelligence Research (JAIR) 30, 457–500 (2007)
62. Marder, B., Joinson, A., Shankar, A., Thirlaway, K.: Strength matters: self-presentation to the strongest audience rather than lowest common denominator when faced with multiple audiences in social network sites. Computers in Human Behavior 61, 56–62 (2016)
63. Marwick, A.E., Boyd, D.: I tweet honestly, i tweet passionately: Twitter users, context collapse, and the imagined audience. New media & society 13(1), 114–133 (2011)
64. McCandless, M.: Accuracy and performance of Google's compact language detector. Blog post (2010)
65. McCrae, R.R., Costa, P.T.: Reinterpreting the Myers-Briggs type indicator from the perspective of the five-factor model of personality. Journal of personality 57(1), 17–40 (1989)
66. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: Homophily in social networks. Annual review of sociology 27(1), 415–444 (2001)
67. Milroy, J.: Linguistic variation and change: on the historical sociolinguistics of English. B. Blackwell (1992)
68. Minkus, T., Liu, K., Ross, K.W.: Children seen but not heard: When parents compromise children's online privacy. In: Proc. WWW. pp. 776–786. ACM (2015)
69. Mohammad, S.M., Turney, P.D.: Crowdsourcing a word-emotion association lexicon. Artificial Intelligence 29(3), 436–465 (2013)

70. Monroe, B.L., Colaresi, M.P., Quinn, K.M.: Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. Political Analysis 16(4), 372–403 (2008)
71. Nakagawa, S., Schielzeth, H.: A general and simple method for obtaining r2 from generalized linear mixed-effects models. Methods in Ecology and Evolution 4(2), 133–142 (2013)
72. Nguyen, D., Smith, N.A., Rosé, C.P.: Author age prediction from text using linear regression. In: Proc. of the Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. pp. 115–123. Association for Computational Linguistics (2011)
73. Nguyen, D.P., Gravel, R., Trieschnigg, R., Meder, T.: "how old do you think I am?" a study of language and age in Twitter. In: Proc. ICWSM (2013)
74. Nguyen, D.P., Trieschnigg, R., Doğruöz, A.S., Gravel, R., Theune, M., Meder, T., de Jong, F.: Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In: Proc. COLING (2014)
75. Nguyen, M.T., Lim, E.P.: On predicting religion labels in microblogging networks. In: Proc. SIGIR. pp. 1211–1214. ACM (2014)
76. Niederhoffer, K.G., Pennebaker, J.W.: Linguistic style matching in social interaction. Journal of Language and Social Psychology 21(4), 337–360 (2002)
77. Oomen, I., Leenes, R.: Privacy risk perceptions and privacy protection strategies. In: Policies and research in identity management, pp. 121–138. Springer (2008)
78. Peersman, C., Daelemans, W., Van Vaerenbergh, L.: Predicting age and gender in online social networks. In: Proc. of the 3rd International Workshop on Search and Mining User-generated Contents. pp. 37–44. ACM (2011)
79. Pennacchiotti, M., Popescu, A.M.: A machine learning approach to Twitter user classification. In: Proc. ICWSM. pp. 281–288 (2011)
80. Pennebaker, J.W., Stone, L.D.: Words of wisdom: language use over the life span. Journal of Personality and Social Psychology 85(2), 291 (2003)
81. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proc. EMNLP. vol. 14, pp. 1532–1543 (2014)
82. Phelan, C., Lampe, C., Resnick, P.: It's creepy, but it doesn't bother me. In: Proc. CHI. pp. 5240–5251. ACM (2016)
83. Plank, B., Hovy, D.: Personality traits on twitterorhow to get 1,500 personality tests in a week. In: Proc. WASSA (2015)
84. Postmes, T., Spears, R., Lea, M.: Breaching or building social boundaries? SIDE-effects of computer-mediated communication. Communication research 25(6), 689–715 (1998)
85. Potthast, M., Hagen, M., Stein, B.: Author obfuscation: Attacking the state of the art in authorship verification. In: Proc. CLEF (Working Notes). pp. 716–749 (2016)
86. Quercia, D., Kosinski, M., Stillwell, D., Crowcroft, J.: Our Twitter profiles, our selves: Predicting personality with Twitter. In: Proc. SocialCom. pp. 180–185. IEEE (2011)
87. Rao, D., Yarowsky, D., Shreevats, A., Gupta, M.: Classifying latent user attributes in twitter. In: Proc. of the 2nd International Workshop on Search and Mining User-generated Contents. pp. 37–44. ACM (2010)
88. Reddy, S., Knight, K.: Obfuscating gender in social media writing. In: Proc. of Workshop on Natural Language Processing and Computational Social Science. pp. 17–26 (2016)
89. Reed, P.J., Spiro, E.S., Butts, C.T.: Thumbs up for privacy?: Differences in online self-disclosure behavior across national cultures. Social science research 59, 155–170 (2016)
90. Rosenthal, S., McKeown, K.: Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations. In: Proc. ACL. pp. 763–772. Association for Computational Linguistics (2011)
91. Rossi, L., Magnani, M.: Conversation practices and network structure in Twitter. In: Proc. ICWSM (2012)

92. Ryan, E.B., Hummert, M.L., Boich, L.H.: Communication predicaments of aging patronizing behavior toward older adults. Journal of Language and Social Psychology 14(1-2), 144–166 (1995)
93. Sap, M., Park, G., Eichstaedt, J., Kern, M., Stillwell, D., Kosinski, M., Ungar, L., Schwartz, H.A.: Developing age and gender predictive lexica over social media. In: Proc. EMNLP. pp. 1146–1151. Association for Computational Linguistics (2014)
94. Schnoebelen, T.J.: Emotions are relational: positioning and the use of affective linguistic resources. Ph.D. thesis, Stanford University (2012)
95. Schrammel, J., Köffel, C., Tscheligi, M.: Personality traits, usage patterns and information disclosure in online communities. In: Proc. HCI. pp. 169–174. British Computer Society (2009)
96. Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M.E.S., Ungar, L.H.: Personality, gender, and age in the language of social media: The open-vocabulary approach. PloS one 8(9), e73791 (2013)
97. Shelton, M., Lo, K., Nardi, B.: Online media forums as separate social lives: A qualitative study of disclosure within and beyond Reddit. Proc. iConference (2015)
98. Snefjella, B., Kuperman, V.: Concreteness and psychological distance in natural language use. Psychological science 26(9), 1449–1460 (2015)
99. Soderberg, C., Callahan, S., Kochersberger, A., Amit, E., Ledgerwood, A.: The effects of psychological distance on abstraction: Two meta-analyses. PSYCHOLOGICAL BULLETIN 141(3) (2015)
100. Spears, R., Lea, M.: Social influence and the influence of the "social" in computer-mediated communication. In: Lea, M. (ed.) Contexts of computer-mediated communication, pp. 30–65). Harvester Wheatsheaf (1992)
101. Steinpreis, R.E., Anders, K.A., Ritzke, D.: The impact of gender on the review of the curricula vitae of job applicants and tenure candidates: A national empirical study. Sex roles 41(7), 509–528 (1999)
102. Strater, K., Lipford, H.R.: Strategies and struggles with privacy in an online social networking community. In: Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction-Volume 1. pp. 111–119. British Computer Society (2008)
103. Stutzman, F., Vitak, J., Ellison, N.B., Gray, R., Lampe, C.: Privacy in interaction: Exploring disclosure and social capital in Facebook. In: Proc. ICWSM (2012)
104. Tannen, D.: You just don't understand: Women and men in conversation. Virago London (1991)
105. Tannen, D.: Gender and conversational interaction. Oxford University Press (1993)
106. Tausczik, Y.R., Pennebaker, J.W.: The psychological meaning of words: LIWC and computerized text analysis methods. Journal of language and social psychology 29(1), 24–54 (2010)
107. Tchokni, S.E., Séaghdha, D.O., Quercia, D.: Emoticons and phrases: Status symbols in social media. In: PRoc. ICWSM (2014)
108. Thomas, K., Grier, C., Nicol, D.M.: unfriendly: Multi-party privacy risks in social networks. In: Proc. PETS. pp. 236–252. Springer (2010)
109. Trepte, S., Reinecke, L., Ellison, N.B., Quiring, O., Yao, M.Z., Ziegele, M.: A cross-cultural perspective on the privacy calculus. Social Media+ Society 3(1), 2056305116688035 (2017)
110. Trope, Y., Liberman, N.: Construal-level theory of psychological distance. Psychological review 117(2), 440 (2010)
111. Volkova, S., Bachrach, Y., Armstrong, M., Sharma, V.: Inferring latent user properties from texts published in social media. In: Proc. AAAI. pp. 4296–4297 (2015)

112. Wienberg, C., Gordon, A.S.: Privacy considerations for public storytelling. In: Proc. ICWSM (2014)
113. Yaeger-Dror, M.: Religion as a sociolinguistic variable. Language and Linguistics Compass 8(11), 577–589 (2014)
114. Youn, S., Hall, K.: Gender and online privacy among teens: Risk perception, privacy concerns, and protection behaviors. Cyberpsychology & behavior 11(6), 763–765 (2008)
115. Zhang, K., Kizilcec, R.F.: Anonymity in social media: Effects of content controversiality and social endorsement on sharing behavior. In: Proc. ICWSM (2014)