Choi et al. *EPJ Data Science* (2024) 13:45
https://doi.org/10.1140/epjds/s13688-024-00483-0

EPJ Data Science
a SpringerOpen Journal

RESEARCH Open Access

# Profile update: the effects of identity disclosure on network connections and language

Minje Choi[1,2*], Daniel M. Romero[2,3,4] and David Jurgens[2,3]

*Correspondence: mchoi96@gatech.edu
[1] School of Interactive Computing, Georgia Institute of Technology, North Avenue, Atlanta, Georgia, USA
[2] School of Information, University of Michigan, 105 S State St, Ann Arbor, Michigan, USA
Full list of author information is available at the end of the article

**Abstract**

Our social identities determine how we interact and engage with the world surrounding us. In online settings, individuals can make these identities explicit by including them in their public biography, possibly signaling a change in what is important to them and how they should be viewed. While there is evidence suggesting the impact of intentional identity disclosure in online social platforms, its actual effect on engagement activities at the user level has yet to be explored. Here, we perform the first large-scale study on Twitter that examines behavioral changes following identity disclosure on Twitter profiles. Combining social networks with methods from natural language processing and quasi-experimental analyses, we discover that after disclosing an identity on their profiles, users (1) tweet and retweet more in a way that aligns with their respective identities, and (2) connect more with users that disclose similar identities. We also examine whether disclosing the identity increases the chance of being targeted for offensive comments and find that in fact (3) the combined effect of disclosing identity via both tweets and profiles is associated with a reduced number of offensive replies from others. Our findings highlight that the decision to disclose one's identity in online spaces can lead to substantial changes in how they express themselves or forge connections, with a lesser degree of negative consequences than anticipated.

**Keywords:** Identity disclosure; Twitter; Profile description; Natural language processing; Social networks

## 1 Introduction

The term "social identity" refers to an individual's self-concept derived from perceived membership in social groups, such as ethnicity, nationality, gender, and political leanings [1]. In the current digital age, identities have become increasingly dynamic and complex, shaped by interactions and expressions within the internet. Social media platforms, in particular, have emerged as important grounds for identity display, where users can shape and reveal their personas through text, images, and interactions. With online communication becoming an important element of social dynamics, understanding online identity construction and its impacts on social interaction is an important topic for un-

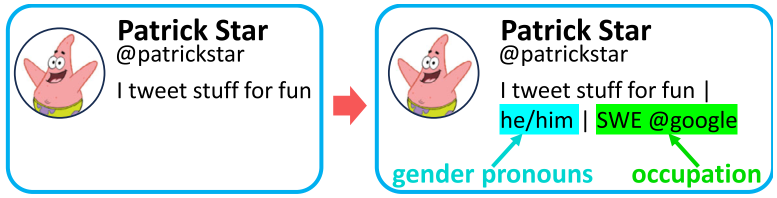

**Figure 1** Identity disclosure in Twitter profiles

derstanding our society. Twitter, with its diverse user base across the world, serves as an ideal stage for observing these digital identity transformations.

Identity disclosure and management is an essential part of online behavior [2, 3], as individuals navigate what aspects of themselves are salient to others. In more public platforms like Twitter, individuals must learn how to present themselves based on the mix of audiences they are exposed to [4–6]. People may explicitly express social identities in social media by including phrases related to the identity in *profile descriptions*, as shown in Fig. 1. Profile descriptions, similar to posts, contain rich textual features associated with the user's social identity, which has been covered in several studies [7–12]. These profiles are not static: individuals add and remove identity markers from their bios to emphasize new or specific aspects of themselves, such as political affiliations [13] or gender pronouns [14, 15]. While recent studies have shed light on the increasing importance of studying profile descriptions for understanding the social identities displayed within these bios [16–18], *how* a user changes when disclosing such identities remains unknown.

Decisions to disclose one's identity in online spaces can have profound effects on one's user experience and subsequent interactions. Consider the case of Alex, a fictional yet representative Twitter user, who one day updates their profile description to include their profession and LGBTQ+ identity. This action not only affects how Alex is perceived by their followers but may also influence the nature of the interactions they have on the platform, potentially increasing their engagement with certain communities while reducing it with others. Such real-life instances illustrate the potential of profile updates to act as pivotal moments of identity disclosure that can reshape online social landscapes. Prior studies have drawn connections between the disclosure of identities—especially marginalized or minority identities—and identity-based hate or cyberbullying, therefore hindering people from fully expressing themselves and sometimes even forcing them to hide identities online [19, 20]. However, the potential varied outcomes for identity disclosure are yet to be quantified.

This study aims to bridge the gap in the literature by focusing on the changes in Twitter user behavior following updates to their profile descriptions that disclose or emphasize aspects of their social identities. We conduct a large-scale quasi-experimental study on 283,793 Twitter users who updated their profiles to disclose a particular social identity. We observe that while overall tweet activity levels remain stable post-disclosure, their tweets contain significantly higher volumes of identity-relevant language, which we further dissect into topic and style properties. We demonstrate that this disclosure is also associated with social network changes: users actively engage more with similar-identity individuals following disclosure. Finally, we examine the number of offensive replies received from others during pre- and post-disclosure periods, where we show that contrary to prior work on offline and online settings [21–23], disclosing social identities in profiles did not lead to increased levels of received offensiveness. This holds even for identity categories known to be prone to targeted offensiveness such as sexual and gender minorities. Overall, our findings suggest that profile-based identity disclosure is an active process signaling future behavior changes in the priorities of a user.

## 2 Literature review

Prior work has examined identity disclosure from the perspectives of language, networks, and social interactions, particularly in online spaces. We build on existing theories and findings to formulate hypotheses that examine whether disclosure of social identities leads to changes in the behaviors of both the user themself and how they are perceived by others.

### 2.1 Social identities and language

Research in sociolinguistics has extensively focused on the relationship between language and social identity [24–26]. Language serves not only as a medium of communication but also as a tool for signaling social affiliations and identities. Individuals may adjust their linguistic style to align more closely with the norms and expectations of the identity they have chosen to present. This adaptation may include the use of specific terminologies, slang, or stylistic elements that resonate with the disclosed identity [27, 28]. Likewise, displaying identity through language can be seen as an intentional form of agency to meet certain social goals such as gaining acceptance within a desired community, maintaining solidarity among peers, or differentiating oneself from other groups [29].

A related behavior is impression management [30], which further explains how individuals strategically present themselves to be perceived in a desired manner. This management helps achieve socially desirable goals such as maintaining reputation [31, 32]. Especially for organizational accounts in online social networks such as Twitter or LinkedIn, strategically revealing certain aspects of identities while concealing others is associated with positive outcomes such as larger friend connections [33].

Through these perspectives, identity disclosure on platforms like Twitter can be understood as part of a broader communicative practice, where users selectively employ language to craft their personas to align with their social identities. Prior work has shown that users with identity signals in their profiles are more likely to share content that aligns with their presented identity [17]. We further examine this finding by hypothesizing that, even for the same user, adding one's identity further promotes this sharing behavior. Our first hypothesis examines the relationship between identity disclosed through language and profile updates. We hypothesize that the modification of one's profile to disclose a particular social identity will motivate the user to tune their linguistic style to accommodate their presented identity.

*H1 Disclosure of social identity on a Twitter profile will lead to posting more identity-aligned tweets compared to a reference group.*

### 2.2 Networked effects of identity disclosure

Several theories from social psychology offer insights into the motivations and processes underlying identity disclosure on social media. The social identity theory [1] suggests that individuals find part of themselves from their membership in social groups, influencing their attitudes and behaviors toward group members and outsiders. Similarly, the self-categorization theory [34] proposes that individuals identify with certain groups under specific contexts, which leads to a sense of belonging and conformity to group norms. Both theories suggest that individuals categorize themselves and others into social groups based on shared characteristics, which can guide behavior and social interactions. Such connections can lead to desirable outcomes including increased social support or access to information [35, 36]. Finally, the collective identity theory [37] emphasizes the formation of group identities, suggesting a framework for understanding increased connections among users sharing similar social identities. All of these theories point towards the act of identity disclosure being not merely a personal choice but a socially influenced consequence that shapes and is shaped by the interactions within online communities.

In the context of online social networks, we expect similar motivations for forming and maintaining social groups. We thus expect that the disclosure of social identity in one's profile reflects a desire to connect with like-minded others, which results in an increased effort to forge connections with people of the same identity.

*H2 Disclosure of social identity on a Twitter profile will directly lead to establishing more network connections with users of the same identity compared to a reference group.*

### 2.3 Consequences of identity disclosure

Public responses to identity disclosure on social media are varied and can lead to both positive support and negative backlash. Several studies have studied the effects of harmful and offensive messages targeted to minority groups or marginalized identities, including but not limited to gender [19, 38], ethnicity [39, 40], and religious beliefs [41]. Concerns of being exposed to such offensiveness can hurt not only a user's willingness to disclose themselves on online platforms but even to be active on the platform at all. This phenomenon is further exacerbated by online social networks providing anonymity and distance, enabling individuals to express prejudiced or hostile views without any social repercussions. The impact of such negativity not only affects individual users but can also influence the social dynamics within online communities, potentially reducing the overall level of open engagement and authentic self-expression.

Theoretical frameworks such as the spiral of silence theory [42] and online disinhibition effect [43] also offer insights into understanding the mechanisms behind negative responses towards certain social identities. The spiral of silence theory posits that individuals are less likely to express their opinions if they perceive them to be in the minority, for fear of isolation or backlash [44, 45]. This theory can be applied to understand how negative responses to identity disclosure on social media may silence other users with similar identities, especially when posts can be easily searched and can remain on the platform [46]. Meanwhile, the online disinhibition effect suggests that the anonymity and invisibility provided by online environments can lead to a reduction in social restrictions and norms governing behavior, resulting in more aggressive or hostile interactions [47, 48]. Together, these theories illuminate the challenges posed by negative responses to identity disclosure, highlighting the need for interventions and support mechanisms to foster more inclusive and respectful online communities.

Given these backgrounds, as our final hypothesis, we examine whether disclosure of one's identity truly leads to increased hostility directed at the user. Specifically, we measure

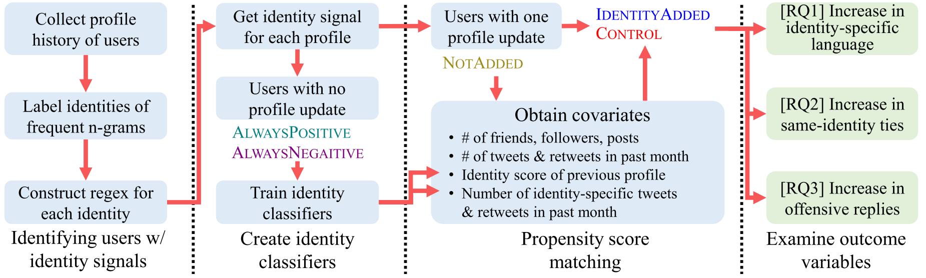
Collect profile history of users
Label identities of frequent n-grams
Construct regex for each identity
Identifying users w/ identity signals
Get identity signal for each profile
Users with no profile update
AlwaysPositive
AlwaysNegaitive
Train identity classifiers
Create identity classifiers
Users with one profile update
IdentityAdded
Control
NotAdded
Obtain covariates
• # of friends, followers, posts
• # of tweets & retweets in past month
• Identity score of previous profile
• Number of identity-specific tweets & retweets in past month
Propensity score matching
[RQ1] Increase in identity-specific language
[RQ2] Increase in same-identity ties
[RQ3] Increase in offensive replies
Examine outcome variables

**Figure 2** Schematic diagram of data collection, identity classifiers, propensity score matching, and analyses on research questions. AlwaysPositive, AlwaysNegative, IdentityAdded, NotAdded and Control represent sets of users with specific profile and identity statuses

if a user becomes a target of offensive content following the addition of their identity on the profile.

*H3 Disclosure of social identity on a Twitter profile will result in receiving more offensive replies compared to a reference group.*

## 3 Data collection

Here, we describe our pipeline for collecting a dataset of Twitter users disclosing social identities on their profiles. An overview of the data collection and processing is shown in Fig. 2.

### 3.1 Identifying Twitter profile changes

We first identify a set of users who have added signals of their social identity to their Twitter profiles. This information is unobtainable using just the Twitter API as it only returns a user's profile information at the time of the API call and does not provide a chronological timeline of profile changes. We instead use the Twitter Decahose dataset which contains a 10% sample of the entire Twitter activities for over 12 months. We identify all activities of every user between April 2020 and April 2021. Each tweet or retweet object includes various metadata, one of which is the user's profile description at the time of the tweet. We collect all instances of user profiles for our Twitter users and sort them in chronological order, enabling us to identify *when* a user changed their profile. We remove verified accounts and users whose language is set to a language other than English, resulting in 15,215,776 users and 73,048,466 unique profiles.

### 3.2 Categorizing social identities

Deciding what counts as a social identity can be challenging. Here, we start from an initial set of social categories based on two relevant studies. [8] categorize social identities into five groups based on the findings of [49]: personal relationships, vocations/avocations, political affiliations, ethnic/religious groups, and stigmatized groups. Meanwhile, [17] constructed identity categories based on [27]: age, ethnicity/nationality, fandoms, gender, interests, location, personality type, pronouns, relationship status, sexual orientation, and zodiac. Using this list of categories as a starting point, one of the authors manually inspected a list of frequent n-grams (n = [2,3]) extracted from the profiles of all users included in this study. We used `scikit-learn` to extract the top 30,000 n-grams, and further filtered only the top 3365 n-grams which appeared in the profiles of at least 1% of the users in our dataset. During the inspection, each n-gram was shown five random tweets containing the n-gram to verify whether it belonged to an identity category. Each n-gram was manually inspected and assigned to a category when applicable. Supplementary Table 1 (see Additional file 1) contains examples of five n-grams with their corresponding tweets and categories assigned. Furthermore, the n-grams within each category were additionally grouped into subcategory levels. For instance, the *gender* category consists of three subcategories: *men*, *women*, and *nonbinary*. A total of 221 n-grams were assigned to a category and subcategory. A list of the categories and subcategories can be found in Table 1.

Descriptions of the final categories are as follows:

- *Age* This category contains the disclosed age of the user. We grouped age into five bins to represent teenagers (13-17), college students (18-24), young adults at early stages of their careers (24-35), adults at the age of parenthood and advanced careers (35-49), and senior adults (50+). We acknowledge different categorizations of age could be used in this study, such as that of [10].
- *Education* We constructed a single-identity category *Education* to collect instances of students disclosing their education status, such as degree name, current university, or school year.
- *Ethnicity* This category contains the self-declared ethnicity of the user. We included words or phrases describing the user's ethnicity as well as nationality flag emojis which can be used to describe one's nationality. Our subcategories are limited to countries where there was at least one corresponding n-gram.
- *Gender pronouns* Following the work of [15], we use three subcategories of gender pronouns: men, women, and nonbinary.
- *Occupation* Occupation categories were obtained from the International Standard Classifications of Occupations (ISCO-08) list, where we selected all sub-major group categories that corresponded to any of the top n-grams we examined.
- *Political* This category corresponds to the disclosed political leaning of the user. Along with subcategories for conservative and liberal, we include another category related to activism, which in this case corresponds to phrases related to the Black Lives Matter movement.
- *Relationship* Based on the frequent n-grams, we identify three types of family relationship types mentioned in profiles: *partner*, *parent*, and *sibling*.
- *Religion* We identify n-grams containing religious terms, and create subcategories for each different religion that was mentioned. For Christianity and Catholism we discover that it is difficult to split out the two and thus combine them into a single category. Finally, n-grams genuinely mentioning 'God' are mapped to the *General* subcategory
- *Sexuality* We identified n-grams corresponding to LGBTQ+ identities and mapped them into a single subcategory. We remove phrases that signal only indirect membership (e.g. LGBT-ally)
- *Personal* We define a category for two additional types of self-disclosure. One is the disclosure of additional social media accounts, and the other is that of stigmatized identities such as joblessness, health issues, and trauma.

**Table 1** Count of users who added social identities to their Twitter profiles once in our observation period for each subcategory-level identity

| Category | Subcategory | # users |
|---|---|---:|
| Age | 13-17 | 871 |
| | 18-24 | 8872 |
| | 25-34 | 2449 |
| | 35-49 | 381 |
| | 50+ | 164 |
| Education | Student | 23,201 |
| Ethnicity | African | 575 |
| | American | 5397 |
| | British | 1487 |
| | Canadian | 2050 |
| | German | 636 |
| | Indian | 3045 |
| | Irish | 1023 |
| | Japanese | 349 |
| | Korean | 259 |
| | Mexican | 1686 |
| Gender pronouns | Men | 19,115 |
| | Women | 36,708 |
| | Non-binary | 4070 |
| Occupation | Administrative | 160 |
| | Art | 28,746 |
| | Business | 3284 |
| | Community | 635 |
| | Computer | 3031 |
| | Education | 6556 |
| | Engineering | 4765 |
| | Healthcare | 4109 |
| | Legal | 1117 |
| | Management | 13,646 |
| | Science | 2645 |
| Personal | Social Media | 39,310 |
| | Sensitive | 1797 |
| Political | Conservative | 2059 |
| | Liberal | 2347 |
| | Activism | 22,203 |
| Relationship | Partner | 6966 |
| | Parent | 12,233 |
| | Sibling | 968 |
| Religion | Catholic / Christian | 5954 |
| | Islam | 1255 |
| | Hinduism | 544 |
| | Atheism | 387 |
| | General | 3029 |
| Sexuality | LGBTQ+ | 3772 |
| Total | | 283,793 |

This process results in a total of ten categories and 44 subcategories of identities (Table 1).

After categorizing n-grams into identity categories and subcategories, we follow the approach from prior work [17, 18] and construct regular expressions for each category and subcategory based on the n-grams to improve precision. For example, when constructing regular expressions for age, we ensure that the corresponding phrases include identifiers such as 'years old' or 'y/o'.

Next, we identify a set of users who have changed their profiles to disclose their social identity. We run our regular expressions on every unique profile to determine whether a profile is associated with a particular identity. We assign multiple labels if a user's profile is associated with multiple identity categories (e.g. "18yo | he/him | father of two wonderful children"), but leave out profiles that our method labels as belonging to multiple subcategories within the same category when they are meant to be mutually exclusive (e.g. age - "18yo | 30y/o", political affiliation - "devout democrat | conservative"). Based on the mapped identities per profile, we can identify all users who satisfy the following two conditions: (1) each user has made only one change in their profile during the 1 year observation period, and (2) the only change is the addition of a new social identity—i.e., the phrase indicating identity should only exist in the changed profile and not the previous version. This filtering results in a set of 283,793 users who added a single new social identity through Twitter profiles, which we refer to as IdentityAdded. Table 1 contains subcategory-level counts.

We validate the quality of our pipeline for capturing instances of identity disclosure through an annotation task. For each subcategory, three annotators are provided twenty samples which each consist of two subsequent profiles, one pre- and one post-change. The twenty samples include ten positive samples from IdentityAdded as well as ten negative samples, which vary from (1) no disclosure in either, (2) disclosure in both, and (3) disclosure only in pre-change. The resulting Krippendorff's $\alpha$ was 0.74, indicating a high level of agreement that the changes detected by our approach do constitute meaningful self-disclosure of identity. We then evaluate our pipeline by evaluating it on the majority vote from the annotations, from which we saw that 41/44 identities achieved an F1 score higher than 0.5 (Supplementary Tables 3 and 4). We therefore removed the three identities with low performance: education:student, ethnicity:korean, and occupation:art.

## 4 Methods

Here we describe the methodological details of how the experiments were conducted.

### 4.1 Inferring identity-specific language in Twitter

Our analyses require models to quantify language that aligns with a particular social identity. We aim to achieve this by formulating classification tasks to distinguish the language patterns between two groups of users depending on their identity disclosure status. Based on existing findings that posts and profile descriptions in online platforms are reflective of one's social identity [e.g., 8, 50], we assume that if a user has disclosed a social identity on their profile description for a sufficiently long time, then the text created by the user contains topical and stylistic features indicative of the disclosed identity.

#### *4.1.1 Experiment setting*

We define AlwaysPositive users as those who (1) did not make any changes to their profiles during our observation period, and (2) contained phrases of a specific identity type in their profile. Similarly, we define AlwaysNegative users as those who (1) did not make any changes, and (2) did not include any identity-specific phrases in their profile (refer to Fig. 2). Here we assume that the tweets posted by a user with an identity-specific phrase in their profile are more likely to align with the listed identity, and so use the labels of the user as proxies for the tweets. However, it would be unrealistic to assume that all tweets contain such alignment. Therefore, for each user, we assign positive/negative labels at corpus level instead of the individual tweet-level, where each sample consists of a corpus of five randomly sampled tweets posted by a user. We remove tweets entirely consisting of URLs, and replace URLs within tweets with a `URL` token. To distinguish the different texts, a `</s>` separation token is inserted between each tweet.

For each identity subcategory, we sample up to 50K positive and 50K negative users, which we split into train/test/validation sets on an 8:1:1 ratio with balanced positive/negative samples. For identity classes with insufficient positive samples, we allow each user to be represented in up to ten different samples provided they have enough unique tweets. We allow for upsampling on identities with small sample sizes on the training set. We finetune each identity separately using a RoBERTa model [51] pretrained on a Twitter corpus [52] provided via the Hugging Face API. The training is done on Pytorch 1.13 and Pytorch Lightning 1.8.6 on an NVidia A5000 machine. We use a learning rate of 1e-6 after 100 initial warmup steps followed by linear decay and run for a maximum of 10 epochs where we stop if the validation performance measured in AUC does not increase after two consecutive epochs.

#### *4.1.2 Model performances*

We evaluate the performances of all models using two metrics: AUC and F1 score. AUC scores are generally high, with all models exceeding a performance of 0.7 (Supplementary Fig. 1). This indicates that the models are doing a good job at assigning higher scores to tweets that contain more signals of identity and vice versa. F1 scores are lower in general, with a few identities such as age:35-49 and occupation:administrative performing worse than random (Supplementary Fig. 2). The results from these two figures combined together indicate that while the model sometimes struggles predicting the correct label (positive/negative) for some identities, overall it does a decent job in producing continuous scores which we can use for measuring strong and weak associations of certain identities from texts. Therefore, we proceed with using all of the classifiers for subsequent experiments.

### 4.2 Cross-identity similarities in language

We compare the pairwise similarities between the identity-specific languages across different identities. We first sample a large corpus of one million random tweets from the history of tweets by AlwaysNegative users, so that we avoid biasing our tweets towards any particular identity. Next, we obtain the identity scores for each identity by running every classifier on the same corpus. We compute Spearman rank pairwise similarity between all identity pairs. Supplementary Fig. 3 contains the pairwise scores for all pairs. We can observe stronger similarity scores for within-category comparisons. This suggests that the language of users who disclose identity have some level of similarity regardless of identity type.

### 4.3 Estimating effects of identity disclosure

Throughout this study we aim to identify the impact of identity disclosure on interaction behaviors of Twitter users. As we only use observational data from the past history of users, we cannot measure the "true" causal effect which requires counterfactual information from the treated users had they not disclosed their identities. To address this, we adopted a methodology commonly utilized in observational studies to estimate causal effects. This involves selecting a comparable control group through propensity score matching. Subsequently, a propensity score analysis is conducted to quantify the differential impacts between the treatment and control groups. The following section elaborates on this methodology in greater detail.

#### *4.3.1 Propensity score matching*

The decision to disclose one's identity through profiles is associated with several behavioral characteristics which may not be prevalent in every Twitter user. Simply comparing the activity levels between a random sample of users who did not disclose their identities will not take into account such properties, leading to the inclusion of confounding factors. A meaningful measurement can be made by comparing against a control group that displays similar behaviors and demographic properties but does not disclose social identities through profile updates. We adopt propensity score matching (PSM), a quasi-experimental method widely adopted in observational studies involving observational social media data [53–55].

*Selecting covariates* The goal of PSM is to specify a subset of control samples which have properties similar to those of the treated samples, in this case the Twitter users who disclosed their identities. This is achieved by specifying a number of covariates which are relevant to the outcome variables, then ensuring that the covariates are balanced between the treated and control samples. We curate a list of covariates obtainable from Twitter data which we believe to be related to Twitter activity levels, which are the outcome variables of interest in this study.

- *Number of connections*: A higher number of connections indicates a wider audience and potentially a more diverse set of social interactions. Users with a larger follower base might feel more compelled to disclose their identity to establish a stronger, more authentic connection with their audience, or to leverage their social influence within certain communities. We include two types of network connections: the number of users one follows (*friends_count)* and is followed by (*followers_count*).
- *Past activity levels*: The frequency of a user's tweets or retweets is indicative of their level of activity and engagement on Twitter. Active users, who frequently post or retweet, are likely to have a more established presence and potentially a more defined online persona. This level of engagement suggests a user's comfort with, and commitment to, the platform, which may correlate with their willingness to disclose personal or social identities as part of their online narrative. To this extent, we include a user's total number of Twitter posts (*statuses_count)* as well as the number of days the account has been active. We also identify the number of tweets and retweets made by every user in one month before the date of the profile update, and include these count values as covariates.
- *Previous profile identity score*: A user's current profile can contain information indicative of the identity they plan to add during a profile update, such as keywords of interest. Along with total tweet activities, we also ensure that the level of identity disclosed by treated and control users remains at similar levels. To achieve this, we measure the level of identity disclosure from the profiles of each user before it was updated using a binary classifier to measure identity alignment from Twitter language,

and include the probability score as a covariate. Details of the classifiers can be found later in the Methods section.

*Selecting control users* To accurately measure the effect of identity disclosure through profile update, one confounder that should be especially controlled for is the effect of simply updating a profile, regardless of whether an identity is disclosed, which leads to specifying a condition for the potential control users. Apart from the IdentityAdded users we also identify 849,901 users who (1) made one profile update during the 1-year observation period but (2) did not include any phrases of social identity in their profiles before or after the update, which we refer to as NotAdded users. For each user in IdentityAdded and NotAdded, we identify the following covariates obtained at the date of the profile change: number of days since account creation, number of friends, number of followers, number of total posts, number of tweets and retweets posted during one month before the time of profile update. For each separate identity within the subcategory level, we use all covariates to train a logistic regression model using `scikit-learn`, which we use for assigning propensity scores to each sample. We stratify the scores into $N$ strata where $N$ equals the root number of positive samples. We use the Fisher Jenks natural break algorithm [56] to obtain the strata bins, which we use for binning both IdentityAdded and NotAdded users according to propensity scores. Within each strata, we assign matched pairs for each IdentityAdded user from the pool of NotAdded users with the following steps. We first limit to NotAdded users who changed their profiles in the same week as the IdentityAdded user. Next, we computed the Euclidean distance between the z-score normalized covariates to select up to 5 users with the shortest distance to the IdentityAdded user.

As a result of the matching process, we are left with 283,566 treated users and 1,228,945 matched users. We refer to the resulting matched set as Control users. Supplementary Figs. 4, 5, and 6 show that the distribution containing the standardized mean difference (SMD) of every covariate reduces sharply after matching, demonstrating the diminished effect caused by confounding covariates.

Along with the default logistic regression setting, we also test two variants. One is an extension of the logistic regression setting where the positive class is upsampled to match the size of the negative class. In another setting, we train a gradient tree boosting classifier instead of the logistic regression to produce propensity scores. We compare the methods across two metrics: AUC and SMD. Supplementary Table 2 compares the performance of each setting with the outcome scores averaged across all identities. When comparing the AUC of each setting, we notice that a balanced logistic regression outperforms the unbalanced logistic regression setting and is slightly inferior to gradient boosting. However, when we observe SMD, we see that both logistic regression settings produce smaller values compared to gradient boosting. We conclude that the performances of all settings are similar and choose to use the propensity scores from the logistic regression with class balance, which ensures both decent AUC and SMD scores. We also conduct ablation studies on variations of propensity scores where each type of covariate is excluded, and find that all covariates contribute to increased performance of the AUC of the propensity matches (Supplementary Fig. 7).

### *4.3.2 Estimating treatment effects*

Our resulting control group resembles the covariate distribution of the treated users. Using this setting, we can estimate the effect of profile disclosure on outcome variables such as tweet activity by comparing the levels of pre-post treatment gains between both groups. We can formulate this as a linear equation which can be fitted into a regression setting. Since our outcome of interest takes the form of count variables (e.g. tweet counts, number of connections), we can formulate this as a negative binomial regression task [57, 58]. Accordingly, we use the following equation:

$$\log(y_{i,t}) = \beta_0 + \beta_1 X_i + \beta_2(T = 1) + \beta_3(t \geq \mathrm{tr}) + \beta_4(T = 1)(t \geq \mathrm{tr})$$

where $y_{i,t}$ is the outcome variable at time $t$ for user $i$, $T = 1$ is a binary assignment status to treatment group, and $t >= \mathrm{tr}$ is whether time $t$ is beyond treatment period. $X_i$ is the time-invariant covariates of $i$, which consist of the number of friends, followers, and total posts. All experiments are modeled as a negative binomial regression using generalized estimating equations (GEE) in `statsmodels`. Because our hypothesis testing is done on multiple identities, we apply the Bonferroni-Holm correction [59] to account for false positives when reporting significance test results from the regressions.

## 5 How does identity disclosure affect language?

To understand behavioral changes following identity disclosure, we first study whether users change their language following profile updates to include a social identity. We examine H1 by measuring changes in the level of identity-relevant content users produce and engage with following identity disclosure.

### 5.1 Experiment setting

We use the scores from the classifiers trained on identity-specific tweets to measure levels of identity-revealing information from both the content that users post (tweets) and engage with through sharing (retweets). Using the identity classifiers, we obtain scores for every tweet and retweet generated by each IdentityAdded and Control user between one month before and after the profile update. We then count the number of tweets with an inferred identity score higher than 0.5 and aggregate them into two periods, before and after the profile update. We consider these as the total number of identity-relevant tweets the user tweeted or retweeted before or after treatment. We also count the number of total tweets regardless of identity score, which captures overall activity levels. We run separate regressions with the number of total tweets/retweets and identity-specific tweets/retweets as outcome variables, and include the number of total activities as a control variable when modeling identity-specific activities.

### 5.2 Results

Figure 3 shows the effects of adding profiles on four different types of tweet activity counts: the number of total tweets (Fig. 3(a)) and retweets (Fig. 3(b)) versus identity-aligning tweets (Fig. 3(c)) and retweets (Fig. 3(d)). We can first observe that, contrary to prior work [35], the additional disclosure of social identity via profiles does not lead to greater overall activity levels compared to profile updates without such disclosure (Figs. 3(a) and 3(b)). In fact, we observe the opposite for several types of identities, most notably drops

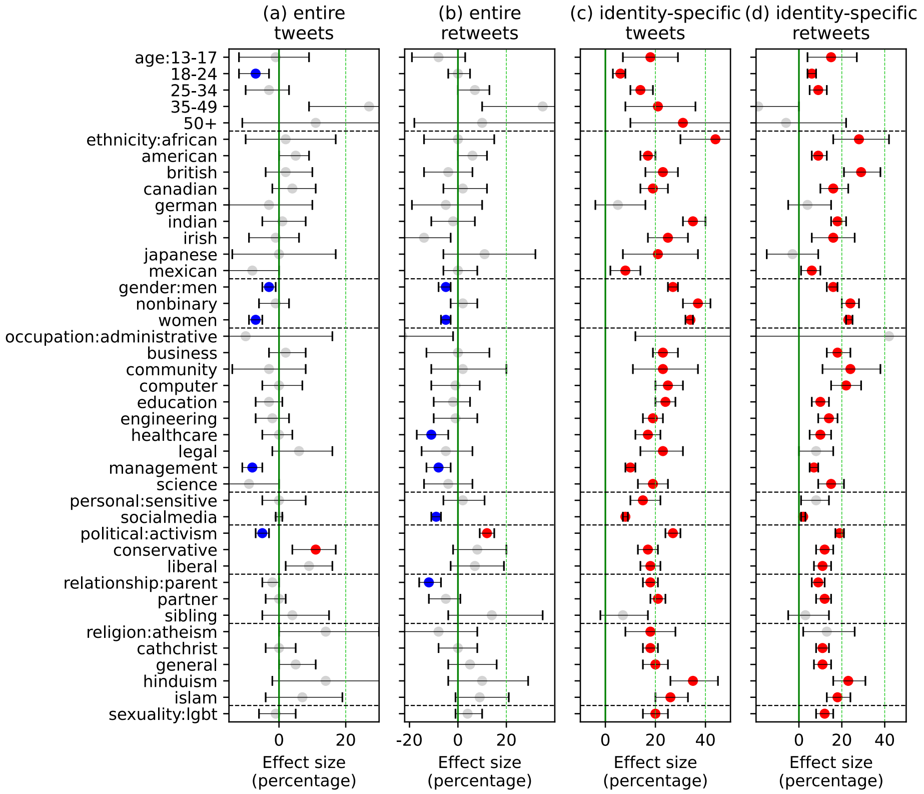

**Figure 3** Effect sizes of identity disclosure on tweet and retweet-level activities. The x-axis indicates a percentage increase in the number of tweets following identity disclosure. Significant positive and negative values that pass the correction test are marked in red (positive) and blue (negative). While identity disclosure does not lead to increased activity levels, there are significant increases in the number of tweets and retweets that contain identity-specific language

of both tweet and retweet levels in binary gender pronouns and student status. The only statistically significant increases we observe arise from disclosing political statuses.

On the other hand, we observe statistically significant increases in the number of tweets posted and retweeted that contain identity-specific language, across *almost every* category (Figs. 3(c) and 3(d)). Though there exists variance among categories, in general, we observe that identity-specific tweets increased by around 20-40% and identity-specific retweets increased by around 10-30%, indicating that though the content volume does not change, the percent of identity-related content within that volume increases substantially. Further comparisons within identity categories reveal interesting findings. For instance, we observe that for both tweets and retweets, the increase following identity disclosure of men is lower than that of women and nonbinary genders. One possible reason is women and nonbinary gender users may undergo harder decisions to disclose their identity, which results in a greater change in their behavior following disclosure. Similarly, our results on ethnicity disclosures show larger identity-specific activities for African identities compared to the American identity, suggesting the level of language change may differ by identity type. In conclusion, we confirm that H1 is valid in that Twitter users produced more tweets and retweets aligned to their identities after the profile update.

### 5.3 Identity-specific language: topic or style?

To further understand which aspects of language change following identity disclosure, we compare the tweets through two components of language: *topic* and *style*. We examine whether having a IdentityAdded user disclose their identity results in their language becoming more similar to that of a AlwaysPositive user regarding each component.

To measure topic distributions, for each identity we run zero-shot contextualized topic models [60, 61] on the tweets of AlwaysPositive users with 50 topics for 20 epochs, then obtain a 50-dimensional distribution which represents their topics $D^T_{AP}$. We then infer the topic distributions of the pre- and post-treatment tweets from IdentityAdded as $D^T_{pre}$ and $D^T_{post}$, which we use to measure the Jensen-Shannon distances of each distribution to $D^T_{AP}$. For style, we select five style variables from [62] as well as classifier models from the Hugging Face API trained on public datasets: offensiveness [52], formality [63, 64], sarcasm [65], toxicity [66, 67], and positive sentiment [68]. For each identity, we computed the binary style scores for every tweet of the AlwaysPositive users to obtain a $N \times 5$ dimension matrix of style scores with $N$ as the number of tweets. We fitted PCA on the matrix to obtain the projection of its principal component, $D^S_{AP}$, which we use to represent the stylistic distribution of AlwaysPositive users. Likewise, we obtained the same matrices for tweets from pre- and post-treatment periods of IdentityAdded users, and transformed these matrices into a single dimension using the principal component from fitted PCA of AlwaysPositive, resulting in $D^S_{pre}$ and $D^S_{post}$. We then used Cohen's d [69] to compute the difference between each of the style distributions to $D^S_{AP}$.

Figure 4 shows the changes in the distance between the language of users who change towards disclosing their identity to those who always have had the identity visible. A reduced difference score from 'pre' to 'post' indicates that after disclosing one's identity, the language one adopts became more similar to the group that always displayed such identity. For topic differences, we do observe a slight increase in topic difference after identity disclosure for some categories such as sexuality, religion, or political ideology. A possible explanation is that users may choose to explore a wider range of new topics after disclosure. However, this increase in difference is rather small (less than 10%) and does not hold uniformly for other categories, where the change is minimal. On the other hand, we observe a consistent trend of reduced difference in style after identity disclosure for all identity categories, with the exception of age. For some categories such as sexuality, relationship or occupation, we observe up to a 30% decrease in style difference after identity disclosure. We believe these results hint at some degree of stylistic matching after identity

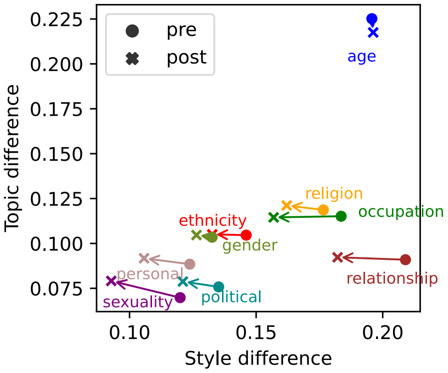

**Figure 4** Changes in style and topic differences between AlwaysPositive users and IdentityAdded users before and after identity disclosure. Style becomes more similar after the disclosure compared to topics, where the relative distances are much smaller

disclosure, indicating intentional adjustment of how one depicts their language to align with the identity they choose to disclose.

## 6 Does identity disclosure in profiles lead to network rewiring towards same-identity connections?

In our next analysis, we investigate whether the addition of identities leads to bridging more connections with like-minded others. To do so, we collect the ego networks of every IdentityAdded and Control user where an edge between two users $u$ and $v$ is defined when $u$ replies to or retweets a tweet posted by $v$. We divide a user's network activities by pre- and post-treatment where we look at a timespan of 12 weeks. We use the same set of regular expressions from the profiles of all users included in the networks and extract any social identities from their profiles during the 12-week period. The subset of connected users who have adopted the same identity as the ego user at any point will be considered same-identity nodes. Thus, in our subsequent analysis, the outcome variable is the number of same-identity nodes before and after the identity disclosure.

### 6.1 Results

Figure 5 displays the treatment effect on the out- and in in-degree of the network when restricted to users of the same identity. We can observe that across most categories, the out-degree of same-identity neighbors significantly increases after identity disclosure in profiles (Fig. 5(a)). This indicates that the users who choose to disclose their identities also choose to connect to more people that share the same identity.

We next look at the in-degree level changes, which is a stronger indicator of how the addition of identity is viewed by others (Fig. 5(b)). We observe that the in-degree of same-identity groups is less likely to increase compared to the out-degree, which indicates that inbound connections are less likely to be made compared to outbound connections, as the former requires others to actually be motivated to establish new connections with the user who has made a profile change. Therefore, we confirm that H2 holds in that a user's inward and outward connections with same-identity neighbors increase after identity disclosure.

Additional results highlight identity-specific changes. Supplementary Fig. 8 contains the effect sizes of the *total* out- and in-degree network sizes following disclosure, revealing that the overall network size only increases for political identities. These results support our claim that users choose to strategically rewire their connections more towards those of the same identity while keeping overall network sizes stable instead of merely being more open in general. Supplementary Fig. 9 shows changes in connection levels towards different identities in the *same* category. We find that gender pronouns are the only category to increase in both in-degree and out-degree for all identities, which is in line with existing work that showed tie clustering among such pronouns [14]. Last of all, we compare changes in cross-partisan connections for conservative and liberal users, where we observe significant increases of outbound connections from those who disclose their liberal identity to conservative users, but not the other way round (Supplementary Fig. 10).

## 7 Does identity disclosure lead to receiving more offensive content?

Finally, we test H3, our hypothesis on possible negative consequences of disclosing one's identity, namely whether identity disclosure leads to increased targeted offensive content.

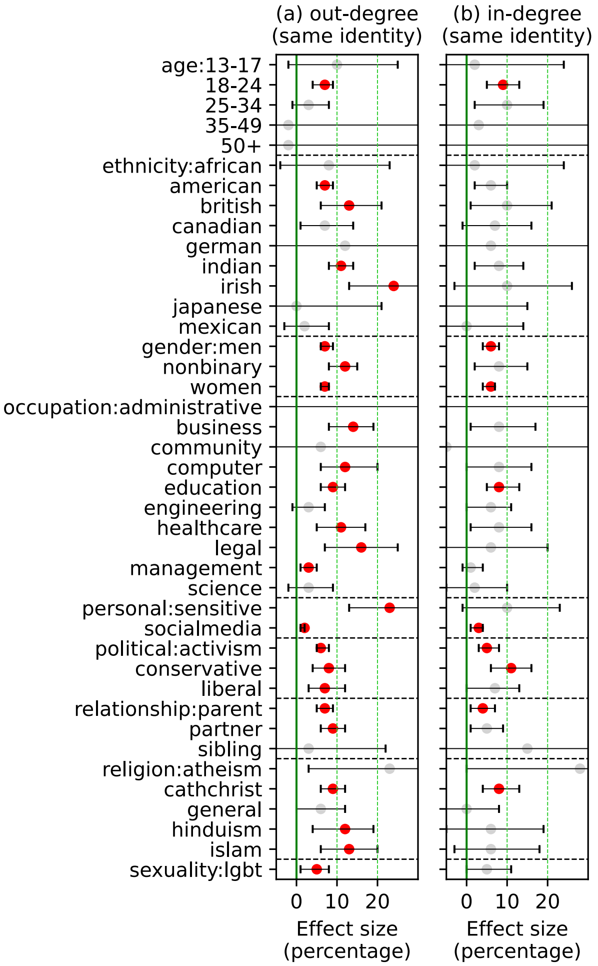

**Figure 5** Effect sizes of identity disclosure on out- and in-degree network sizes. Users reach out to those of the same identity following disclosure (out-degree), but not all identities receive increased attention from others in return (in-degree)

## 7.1 Experiment setting

For each IdentityAdded and Control user, we use the 10% sample dataset to collect a history of the tweets posted by the user during one month before and after the time of their profile update, as well as all replies received from other users during this period. Next, we use a publicly available classifier for detecting offensiveness from Hugging Face [52][1] to obtain offensiveness scores of both the tweets posted and the replies from others. We then formulate an equation to model the expected number of offensive replies

$$\log(y_{i,t}) = \beta_0 + \beta_1 X_i + \beta_2(T=1) + \beta_3(t \geq \text{tr}) + \beta_4(T=1)(t \geq \text{tr}) + \log(\beta_5 \text{n}_{id}) \\ + \log(\beta_6 \text{n}_{id})(T=1)(t \geq \text{tr}).$$

[1] https://huggingface.co/cardiffnlp/twitter-roberta-base-offensive.

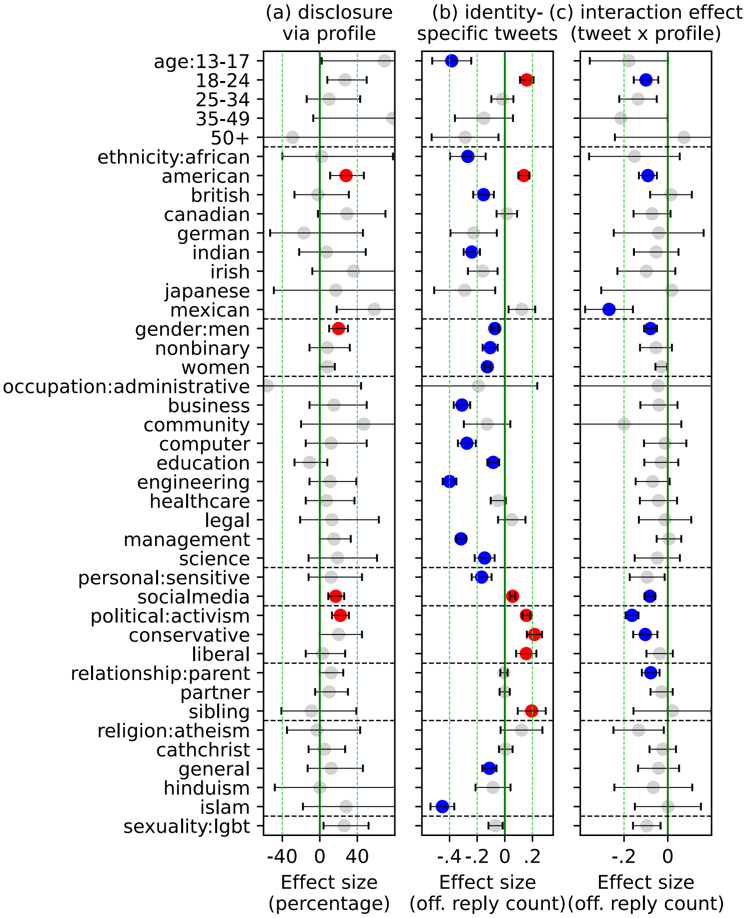

**Figure 6** Effect size of identity disclosure on the number of offensive replies received. (left) identity addition to profile, (middle) number of identity-specific tweets per week, (right) interaction effect of identity disclosure through profile and number of identity-specific tweets per week

The added term $\log(\beta_5 \mathrm{n}_{id})$ indicates the log-normalized number of identity-specific tweets posted by the user and $\log(\beta_6 \mathrm{n}_{id})(T = 1)(t \geq \mathrm{tr})$ is the interaction effect between identity disclosure via profile and identity-specific tweets.

### 7.2 Results

Figure 6(a) ($\beta_4$) first shows that identity disclosure through profiles increases offensiveness for only a handful of categories - ethnicity:American, gender:men, personal:socialmedia, and political:activism. However, when we observe changes in offensiveness levels caused by increased identity of tweets (Fig. 6(b) ($\beta_5$)), we can see that significant effects can be seen from several categories. Interestingly, the disclosure of identity through tweets leads to reduced levels of offensiveness from others for the three studied gender types, as well as for occupations and religion types. Meanwhile, we observe increased levels of offensive replies from all three types within the political category, hinting that this may be due to heated political conversations that often correlate with offensiveness. Lastly, the interaction effect of identity disclosure via both tweet and profile (Fig. 6(c)) ($\beta_6$) suggests that the combined effect from disclosure through both channels reduces levels of offensiveness for every category where increased identity disclosure through tweets was associated with increased offensiveness. One potential explanation is that disclosing identity through both profile and tweet could create a sense of consistency, which helps reduce levels of hostility towards that identity group. Therefore, we conclude that H3 does not hold, and in most cases the disclosure of identity through both profile and tweet language can even reduce the amount of offensive content.

## 8 Discussion

### 8.1 Changes in user behavior following identity disclosure

While general activity levels on Twitter remained largely unchanged across most categories following identity disclosure—apart from political identities—we observed a clear preference for users to tailor their posts to better reflect their identities. By further comparing the changes from two language features (topic and style) we discover that language adjustment mainly comes from stylistic accommodation towards the language of the targeted identity group. These findings resonate with the role of language in identity formation and signaling [24, 25], hinting that this alignment may stem from increased needs to connect to specific identity groups. Indeed, our subsequent findings on network connection volume show that while the overall size of users' networks did not expand, their connections significantly shifted toward individuals sharing the same identity. These observations highlight a deliberate change of user behavior where individuals not only adjust their linguistic and content choices but also strategically curate their social connections to interact with the identity they choose to publicize [17, 37]. Also, we observe an increase in identity-related retweets, which indicates that the disclosed users selectively increased engagement with posts that align with their identities, further disseminating them across their follower networks through retweeting. These behaviors align with the principles of self-categorization theory, which posits that individuals categorize themselves into groups to fulfill their need for social identity, and impression management theory, which suggests that individuals strategically present themselves in ways favorable to their desired group affiliations. Together, these findings highlight a comprehensive strategy by users to signal their identities in online spaces, involving not just the language they adopt and the content they interact with but also the social connections they cultivate.

### 8.2 Community reactions to disclosures of social identities

We also observe that as a user makes changes in how they engage on Twitter following identity disclosure, the surrounding community also responds, albeit to a much smaller extent. Regarding network connectivity of same-identity networks, an increase in-degree is significant for only a subset of categories which experienced an increase in outward connections. There may be various reasons behind the lesser degree of inward attention. One may be that some users may not have noticed the identity disclosure in the profile, a point we later address in the Limitations section. Another may be the inherent dynamics of social media interaction, where outward engagement efforts are more visible and immediate than the gradual accumulation of inward connections, reflecting how online communities form and evolve around shared identities. Nevertheless, we see these patterns as indicative of the complex interplay between individual identity expression and community recognition, where the act of disclosing one's identity leads to reconfiguring social ties, albeit with varying degrees of reciprocity and visibility within the community.

This highlights the importance of visibility and engagement in fostering community connections, suggesting that identity disclosure is a proactive, not just reactive, component of community integration and social support within online spaces.

Interestingly, our analysis on levels of offensive incoming tweets after identity disclosure reveals a complicated picture of identity-targeted offensiveness in online spaces. While existing literature reveal that disclosing one's identity in social media lead to increased hostility and personalized attacks, especially on gender or ethnicity [19, 39], our findings suggest a different narrative. Neither identity disclosure through profile updates nor increased usage of identity-aligned language were associated with increased levels of offensive comments. This observation suggests a cautiously optimistic perspective for individuals debating on disclosing their identities online. These findings, while preliminary, suggest hope on the resilience of online communities and the evolving norms of online social interactions, proposing further exploration in diverse contexts to fully understand the implications of these dynamics.

### 8.3 Different behaviors of identity disclosure by identity type

So far our analysis has applied a general framework to understanding identity disclosure across different categories. However, these identities have unique characteristics, especially with political identities standing out significantly. Notably, political identities were the sole category to show an increase in both tweet activity and network connections *irrespective* of the specific identity disclosed. The increase of political identity-aligned language usage was also associated with an increased volume of offensive replies. This observation aligns with prior research highlighting the dynamics of politically active users on Twitter who make their affiliations public [13]. Users who disclose their political leanings tend to be more active and establish networks characterized by strong homophily [70, 71]. Revealing one's political identity on their profile appears to act as a declaration of deeper engagement in political discourse and activities. Such findings underscore the distinct nature of political identity in the context of online social networks and have profound implications for understanding the mechanisms of political mobilization and the shaping of public opinion online.

## 9 Limitations and future directions

In this section, we address a number of limitations encountered during our study and propose avenues for future research that could address these points.

### 9.1 Generalization of findings across social media platforms

Our study primarily focused on the textual information obtained via profiles on Twitter to understand identity disclosure and its consequences. While insightful, this approach does not include the rich context that images or multimedia content might offer, especially given their role in capturing one's identity on social media. We argue that this was a methodological decision which was due to (1) profile descriptions being an effective method for capturing the disclosure of various types of identities which range from demographic attributes (e.g. age, gender, ethnicity) to societal constructs (e.g. occupation, political affiliation) and (2) access to historical changes in the profile text itself, which often not possible for a user's profile images. Nevertheless, the reliance on text leaves further room for understanding how visual elements contribute to identity signaling and interaction dynamics. Future work could extend this research by analyzing changes in profile images, which could highlight different dimensions of identity disclosure.

The focus on Twitter may also raise questions about the generalizability of our findings to other social media platforms, each with its own user dynamics and possible purposes for identity disclosure. For instance, network-based platforms like Twitter or Facebook, which emphasize personal connections, may exhibit different patterns of identity disclosure compared to community-driven platforms like Reddit or StackExchange, where the focus is on sharing knowledge and experiences. Investigating these platform-specific dynamics could enrich our understanding of online identity management, potentially revealing platform-specific strategies for self-presentation and community engagement [72].

### 9.2 Focus on English-centric users

Our analysis was limited to English-speaking users, predominantly from countries with Western cultures. This limitation raises questions about the applicability of our findings to non-English-speaking contexts or cultures with different social norms and online behaviors. Cross-cultural studies in social media use have highlighted significant differences between individualistic and collectivist cultures, particularly in aspects such as self-disclosure, privacy concerns, and community engagement [73, 74]. Future research could adopt a multilingual approach to explore how identity disclosure varies across cultures, offering insights into global patterns of how social identities are accepted online and the role of language and culture in shaping these processes.

### 9.3 Further inclusion of covariates

Throughout the study we prioritized certain user covariates, such as recent activity levels and network size, over others that could influence identity disclosure, such as personality traits or psychological factors. While this selection was a practical decision to ensure a large user pool could be considered for the matching process, future work may consider additional dimensions of online behavior and identity management. Specifically, personality traits, including the Big-5 (openness, conscientiousness, extraversion, agreeableness, and neuroticism), and self-esteem, have been shown to significantly affect social media use patterns and disclosure behaviors [75, 76]. While capturing personality traits of social media users on a reliable scale is a challenging task, incorporating these psychological covariates could provide a more detailed understanding of the factors that drive identity disclosure on social media.

Some of our covariates include temporal values that may change over time, such as tweet frequency or follower count. These covariates were selected to account for some of the temporal variation of such confounders by performing the matching at a weekly level. For instance, a treated user who updated their profile on week N will only be matched against control users who also updated their profile on the same week. We control for not only the user's total activity levels but also the activity levels around the period of the profile update. The inclusion of both the entire tweet count of a user (e.g., statuses_count) as well as the number of tweets one month before the profile update as covariates reduces these confounders, ensuring that both treated and matched users have been similarly active. However, we acknowledge that this may not capture the full temporal variability of activity levels which would be necessary to control for such analyses. Future studies could benefit from adding extra features to better account for such variances.

## 10 Conclusion

In this study, we explored the impact of social identity disclosure through Twitter profile updates on user behavior and interaction patterns. Our analysis of over 200K users revealed three principal findings: First, users who disclose their social identities within their profiles tend to use language that more explicitly reflects these identities in their tweets, suggesting a form of linguistic alignment with their disclosed social selves. Second, these users do not drastically change the size of their overall social network (Supplementary Fig. 9) but reshape their ties to foster stronger connections with individuals who share similar identity markers, indicating the formation of more homophilous online communities based on shared identities. Finally, our research indicates that disclosing identities in profiles does not lead to receiving more offensive content from others, providing a new perspective to the belief that visibility inherently increases vulnerability to online harassment.

The implications of these findings extend far beyond Twitter, offering insights into the broader dynamics of online identity management and social interaction. They highlight the potential of social media as platforms for positive identity expression and community building, while also reducing fears that such openness necessarily leads to negative consequences. This has important implications for designing social media policies and tools that support safe spaces for identity expression. Our research also has the potential to inform educators, policymakers, and technologists about the ways digital platforms can be structured to foster positive social interactions and minimize harm. In conclusion, our study contributes to a more holistic understanding of how online social networks can facilitate the expression and coexistence of diverse identities in a way that users can feel empowered to share and connect authentically. The code and annotated data for the study will be available at https://github.com/minjechoi/twitter_identity.

**Abbreviations**

SMD, standardized mean difference; NLP, natural language processing; AUC, area under ROC-curve; PSM, propensity score matching; PCA, principal component analysis.

## Supplementary information

**Supplementary information** accompanies this paper at https://doi.org/10.1140/epjds/s13688-024-00483-0.

**Additional file 1.** (PDF 254 kB)

**Acknowledgements**

The authors acknowledge the thoughtful comments and suggestions from the reviewers and editor, which helped improve the manuscript in numerous areas.

**Author contributions**

All authors contributed to the research design process. MC collected and processed the data, and conducted the experiments. All authors contributed to analyzing the results and writing the paper. All authors read and approved the final manuscript.

**Funding**

This work was supported by the National Science Foundation under Grants No. IIS-2007251 and IIS-2143529.

**Availability of data and materials**

Information of the users who disclosed their identities and their corresponding tweets can be provided upon request. Due to Twitter's policy, we will only be able to provide the user and tweet IDs instead of the raw values. The classifiers trained on the tweets of identity-disclosed users will be publicly uploaded on Hugging Face. The code can be accessed at https://github.com/minjechoi/twitter_identity.

## Declarations

### Competing interests

The authors declare that they have no competing interests.

**Author details**

[1]School of Interactive Computing, Georgia Institute of Technology, North Avenue, Atlanta, Georgia, USA. [2]School of Information, University of Michigan, 105 S State St, Ann Arbor, Michigan, USA. [3]Computer Science and Engineering Division, University of Michigan, 2260 Hayward Street, Ann Arbor, Michigan, USA. [4]Center for the Study of Complex Systems, University of Michigan, 500 Church St, Ann Arbor, Michigan, USA.

Received: 4 September 2023 Accepted: 7 June 2024 Published online: 28 June 2024

### References

1. Tajfel H (1982) Social psychology of intergroup relations. Annu Rev Psychol 33(1):1–39
2. Joinson AN, Reips U-D, Buchanan T, Schofield CBP (2010) Privacy, trust, and self-disclosure online. Hum-Comput Interact 25(1):1–24
3. Pavalanathan U, De Choudhury M (2015) Identity management and mental health discourse in social media. In: Proceedings of the 24th international conference on World Wide Web, pp 315–321
4. Marwick AE, Boyd D (2011) I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. New Media Soc 13(1):114–133
5. Bazarova NN, Choi YH (2014) Self-disclosure in social media: extending the functional approach to disclosure motivations and characteristics on social network sites. J Commun 64(4):635–657
6. Duguay S (2016) "he has a way gayer Facebook than I do": investigating sexual identity disclosure and context collapse on a social networking site. New Media Soc 18(6):891–907
7. Li J, Ritter A, Hovy E (2014) Weakly supervised user profile extraction from Twitter. In: Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: long papers). Association for Computational Linguistics, Baltimore, pp 165–174
8. Priante A, Hiemstra D, Van Den Broek T, Saeed A, Ehrenhard M, Need A (2016) # whoami in 160 characters? Classifying social identities based on Twitter profile descriptions. In: Proceedings of the first workshop on NLP and computational social science, pp 55–65
9. Wilson L-M, Wun C (2020) Automatic classification of students on Twitter using simple profile information. In: Proceedings of the 1st conference of the Asia-Pacific chapter of the association for computational linguistics and the 10th international joint conference on natural language processing: student research workshop. Association for Computational Linguistics, Suzhou, pp 30–36
10. Wang Z, Hale S, Adelani DI, Grabowicz P, Hartman T, Flöck F, Jurgens D (2019) Demographic inference and representative population estimates from multilingual social media data. In: The World Wide Web conference. WWW '19. Association for Computing Machinery, New York, pp 2056–2067
11. Nguyen D, Gravel R, Trieschnigg D, Meder T (2013) "how old do you think I am?" a study of language and age in Twitter. In: Proceedings of the international AAAI conference on web and social media, vol 7, pp 439–448
12. Pennacchiotti M, Popescu A-M (2011) A machine learning approach to Twitter user classification. In: Proceedings of the international AAAI conference on web and social media, vol 5, pp 281–288
13. Rogers N, Jones JJ (2021) Using Twitter bios to measure changes in self-identity: are Americans defining themselves more politically over time? J Soc Comput 2(1):1–13
14. Tucker L, Jones J (2023) Pronoun lists in profile bios display increased prevalence, systematic co-presence with other keywords and network tie clustering among us Twitter users 2015-2022. J Quant Descr: Digit Media 3:1–35
15. Jiang J, Chen E, Luceri L, Murić G, Pierri F, Chang H-CH, Ferrara E (2022) What are your pronouns? Examining gender pronoun usage on Twitter
16. Jones JJ (2021) A dataset for the study of identity at scale: annual prevalence of American Twitter users with specified token in their profile bio 2015–2020. PLoS ONE 16(11):1–17
17. Yoder MM, Shen Q, Wang Y, Coda A, Jang Y, Song Y, Thadani K, Rosé CP (2020) Phans, stans and cishets: self-presentation effects on content propagation in Tumblr. In: 12th ACM conference on web science. WebSci '20. Association for Computing Machinery, New York, pp 39–48
18. Pathak A, Madani N, Joseph K (2021) A method to analyze multiple social identities in Twitter bios. In: Proceedings of the ACM on human-computer interaction 5(CSCW2), pp 1–35
19. Haimson OL, Brubaker JR, Dombrowski L, Hayes GR (2015) Disclosure, stress, and support during gender transition on Facebook. In: Proceedings of the 18th ACM conference on computer supported cooperative work & social computing, pp 1176–1190
20. Jhaver S, Ghoshal S, Bruckman A, Gilbert E (2018) Online harassment and content moderation: the case of blocklists. ACM Trans Comput-Hum Interact 25(2):1–33
21. Chan RC (2022) Gender nonconformity, peer victimization, and internalizing problems among youth: differential moderating effects of school climate by sex assigned at birth. Psychol Violence 12:347–360
22. Meyer IH (2003) Prejudice, social stress, and mental health in lesbian, gay, and bisexual populations: conceptual issues and research evidence. Psychol Bull 129(5):674
23. Green M, Bobrowicz A, Ang CS (2015) The lesbian, gay, bisexual and transgender community online: discussions of bullying and self-disclosure in youtube videos. Behav Inf Technol 34(7):704–712
24. Labov W (1966) The social stratification of english in New York city
25. Eckert P (2000) Language variation as social practice: the linguistic construction of identity in belten high
26. Pomerantz A (2007) Language and social identity. J Linguist Anthr 17(1):146–148
27. Bucholtz M, Hall K (2005) Identity and interaction: a sociocultural linguistic approach. Discourse Stud 7(4–5):585–614
28. Schilling-Estes N (2004) Constructing ethnicity in interaction. J Sociolinguistics 8(2):163–195
29. Duranti A (2008) A companion to linguistic anthropology

30. Goffman E (1959) The presentation of self in everyday life
31. Schlenker BR, Britt TW (1999) Beneficial impression management: strategically controlling information to help friends. J Pers Soc Psychol 76(4):559
32. Zivnuska S, Kacmar KM, Witt LA, Carlson DS, Bratton VK (2004) Interactive effects of impression management and organizational politics on job performance. J Organ Behav 25(5):627–640
33. Krämer NC, Winter S (2008) Impression management 2.0: the relationship of self-esteem, extraversion, self-efficacy, and self-presentation within social networking sites. J Media Psychol 20(3):106–116
34. Turner JC, Hogg MA, Oakes PJ, Reicher SD, Wetherell MS (1987) Rediscovering the social group: a self-categorization theory. Basil Blackwell, London
35. Lampe CA, Ellison N, Steinfield C (2007) A familiar face (book) profile elements as signals in an online social network. In: Proceedings of the SIGCHI conference on human factors in computing systems, pp 435–444
36. Yan S, Altenburger KM, Wang Y-C, Cheng J (2022) What does perception bias on social networks tell us about friend count satisfaction? In: Proceedings of the ACM web conference 2022. WWW '22. Association for Computing Machinery, New York, pp 2687–2695
37. Melucci A, Keane J, Mier P (1989) Nomads of the present: social movements and individual needs in contemporary society. Temple University Press, Philadelphia
38. Haimson OL, Veinot TC (2020) Coming out to doctors, coming out to "everyone": understanding the average sequence of transgender identity disclosures using social media data. Transgend Health 5(3):158–165
39. Matthew Costello LC, Feng Luo HH, Song Liao NV (2021) Covid-19: a pandemic of anti-Asian cyberhate. J Hate Stud 17(1):108–118
40. Ziems C, He B, Soni S, Kumar S (2020) Racism is a virus: anti-asian hate and counterhate in social media during the covid-19 crisis. arXiv:2005.12423
41. Chandra M, Reddy M, Sehgal S, Gupta S, Buduru AB, Kumaraguru P (2021) "a virus has no religion": analyzing islamophobia on Twitter during the covid-19 outbreak. In: Proceedings of the 32nd ACM conference on hypertext and social media, pp 67–77
42. Noelle-Neumann E (1977) Turbulences in the climate of opinion: methodological applications of the spiral of silence theory. Public Opin Q 41(2):143–158
43. Suler J (2004) The online disinhibition effect. CyberPsychol Behav 7(3):321–326. PMID: 15257832
44. Neubaum G, Krämer NC (2018) What do we fear? Expected sanctions for expressing minority opinions in offline and online communication. Commun Res 45(2):139–164
45. Burnett A, Knighton D, Wilson C (2022) The self-censoring majority: how political identity and ideology impacts willingness to self-censor and fear of isolation in the United States. Soc Media Soc 8(3):20563051221123031
46. Neubaum G (2022) "it's going to be out there for a long time": the influence of message persistence on users' political opinion expression in social media. Commun Res 49(3):426–450
47. Christopherson KM (2007) The positive and negative implications of anonymity in internet social interactions: "on the Internet, nobody knows you're a dog". Comput Hum Behav 23(6):3038–3056
48. Nitschinsk L, Tobin SJ, Varley D, Vanman EJ (2023) Why do people sometimes wear an anonymous mask? Motivations for seeking anonymity online. Pers Soc Psychol Bull 0:01461672231210465
49. Deaux K, Reid A, Mizrahi K, Ethier KA (1995) Parameters of social identity. J Pers Soc Psychol 68(2):280
50. Preoţiuc-Pietro D, Ungar L (2018) User-level race and ethnicity predictors from Twitter text. In: Proceedings of the 27th international conference on computational linguistics. Association for Computational Linguistics, Santa Fe, pp 1534–1545
51. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: a robustly optimized bert pretraining approach. arXiv preprint. arXiv:1907.11692
52. Barbieri F, Camacho-Collados J, Espinosa Anke L, Neves L (2020) Tweeteval: unified benchmark and comparative evaluation for tweet classification. In: Findings of the association for computational linguistics: EMNLP 2020. Association for Computational Linguistics, Online, pp 1644–1650
53. Yuan Y, Saha K, Keller B, Isometsä ET, Aledavood T (2023) Mental health coping stories on social media: a causal-inference study of papageno effect. In: Proceedings of the ACM web conference 2023. WWW '23. Association for Computing Machinery, New York, pp 2677–2685
54. Choi M, Jurgens D, Romero DM (2023) Analyzing the engagement of social relationships during life event shocks in social media. In: Proceedings of the international AAAI conference on web and social media, vol 17, pp 149–160
55. Jiang J, Dotsch R, Triguero Roura M, Liu Y, Silva V, Bos MW, Barbieri F (2023) Reciprocity, homophily, and social network effects in pictorial communication: a case study of bitmoji stickers. In: Proceedings of the 2023 CHI conference on human factors in computing systems. CHI '23. Association for Computing Machinery, New York. https://doi.org/10.1145/3544548.3580990
56. Jenks GF (1967) The data model concept in statistical mapping. Int Yearb Cartogr 7:186–190
57. Cameron AC, Trivedi PK (2013) Regression analysis of count data **53**
58. Mark BA, Harless DW, Spetz J, Reiter KL, Pink GH (2013) California's minimum nurse staffing legislation: results from a natural experiment. Health Serv Res 48(2pt1):435–454
59. Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Stat 6:65–70
60. Bianchi F, Terragni S, Hovy D (2021) Pre-training is a hot topic: contextualized document embeddings improve topic coherence. In: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 2: short papers). Association for Computational Linguistics, Online, pp 759–766
61. Bianchi F, Terragni S, Hovy D, Nozza D, Fersini E (2021) Cross-lingual contextualized topic models with zero-shot learning. In: Proceedings of the 16th conference of the European chapter of the association for computational linguistics: main volume. Association for Computational Linguistics, Online, pp 1676–1683
62. Kang D, Hovy E (2021) Style is NOT a single variable: case studies for cross-stylistic language understanding. In: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers). Association for Computational Linguistics, Online, pp 2376–2387

63. Rao S, Tetreault J (2018) Dear sir or madam, may I introduce the GYAFC dataset: corpus, benchmarks and metrics for formality style transfer. In: Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies volume 1 (long papers). Association for Computational Linguistics, New Orleans, pp 129–140
64. Pavlick E, Tetreault J (2016) An empirical analysis of formality in online communication. Trans Assoc Comput Linguist 4:61–74
65. Misra R, Arora P (2023) Sarcasm detection using news headlines dataset. AI Open 4:13–18
66. cjadams, Sorensen J, Elliott J, Dixon L, McDonald M, nithum, Cukierski W (2017) Toxic comment classification challenge. Kaggle. https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge
67. cjadams, Borkan D, inversion, Sorensen J, Dixon L, Vasserman L (2019) nithum: jigsaw unintended bias in toxicity classification. Kaggle. https://kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification
68. Hartmann J, Heitmann M, Siebert C, Schamp C (2023) More than a feeling: accuracy and application of sentiment analysis. Int J Res Mark 40(1):75–87
69. Cohen J (2013) Statistical power analysis for the behavioral sciences
70. Colleoni E, Rozza A, Arvidsson A (2014) Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. J Commun 64(2):317–332
71. Halberstam Y, Knight B (2016) Homophily, group size, and the diffusion of political information in social networks: evidence from Twitter. J Public Econ 143:73–88
72. Park K, Kwak H, Song H, Cha M (2020) “trust me, I have a ph. d.”: a propensity score analysis on the halo effect of disclosing one’s offline social status in online communities. In: Proceedings of the international AAAI conference on web and social media, vol 14, pp 534–544
73. Dong W, Qiu M, Zhu F (2014) Who am I on Twitter? A cross-country comparison. In: Proceedings of the 23rd international conference on World Wide Web. WWW ’14 companion. Association for Computing Machinery, New York, pp 253–254
74. Liang H, Shen F, Fu K-W (2017) Privacy protection and self-disclosure across societies: a study of global Twitter users. New Media Soc 19(9):1476–1497
75. Seidman G (2013) Self-presentation and belonging on Facebook: how personality influences social media use and motivations. Pers Individ Differ 54(3):402–407
76. Chen JV, Widjaja AE, Yen DC (2015) Need for affiliation, need for popularity, self-esteem, and the moderating effect of big five personality traits affecting individuals’ self-disclosure on Facebook. Int J Hum-Comput Interact 31(11):815–831

## Publisher’s Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.