Journal of Quantitative Description: Digital Media 1(2023), 1–35 10.51685/jqd.2023.002

Pronoun Lists in Profile Bios Display Increased Prevalence, Systematic Co-Presence with Other Keywords and Network Tie Clustering among US Twitter Users 2015-2022

LIAM TUCKER[1]
The University of Alabama, USA

JASON J. JONES
Stony Brook University, USA

Over the past few years, pronoun lists have become more prevalent in online spaces. Currently, various LGBT+ activists, universities, and corporations encourage people to share their preferred pronouns. Little research exists examining the characteristics of individuals who do publicly share their preferred pronouns. Using Twitter bios from active, US-located accounts between early 2015 and June 30, 2022, we explored users' expression of preferred pronouns. First, we noted the prevalence of users with pronoun lists within their bio has increased substantially and that users with pronoun lists in their bios have tweeted more and created their accounts earlier than users without a pronoun list. Second, we observed that certain linguistic tokens systematically co-occurred with pronoun lists. Specifically, tokens associated with left-wing politics, gender or sexual identity, and social media argot co-occurred disproportionately often alongside pronoun lists, while tokens associated with right-wing politics, religion, sports, and finance co-occurred infrequently. Additionally, we discovered clustering among Twitter users with pronouns in their bios. Specifically, we found an above-average proportion of the followers and friends of Twitter users with

[1] Tucker: letucker@crimson.ua.edu
Jones: Jason.j.jones@stonybrook.edu
Date submitted: 2022-03-05

pronouns in their bio also had pronouns in their bios. Twitter users who did not share their preferred pronouns, on the other hand, were disproportionately unlikely to be connected with Twitter users who did.

Keywords: Twitter, pronoun list, preferred pronouns, identity

Over the past few years, preferred pronoun usage has greatly increased. One can find preferred pronoun lists expressed in email signatures, on nametags, in conversations when two people meet for the first time and in social media bios. There is evidence of a dramatic increase in web searches containing the terms "he/him", "she/her", or "they/them" (Google Trends, 2022), including the phrases "why do people put she/her" and "what does they/them pronouns mean." LGBT+ centers at various universities (The University of Maryland, n.d.; University of California, Davis, 2021), diversity centers ("Pronouns: a How-to", 2021), and companies (Chen, 2021) encourage people to share their preferred pronouns. An editorial supporting the sharing of pronoun lists was published by The New York Times (Galanes, 2021). In 2021, Linkedin (Arruda, 2021), Instagram (Instagram, 2021), and Zoom (Stewart, 2022) each added a separate field for users to specify their preferred pronoun lists. A YouGov poll conducted in June and July of 2022 found that 49% of Americans had encountered preferred pronouns in someone's social media bio (YouGov, 2022).

This increase in prevalence of pronoun lists has coincided with increases in the proportion of Americans who identify as nonbinary. The Williams Institute estimated that, in 2022, 1.4% of Americans aged 13-17 identified as transgender (Herman et al, 2022). A 2017 report by the same group estimated that only 0.7% of Americans aged 13-17 identified as transgender (Herman et al, 2017). Pew Research found that, in 2021, 26% of adults in the US knew someone who uses gender-neutral pronouns, up from 18% three years prior (Minkin & Brown, 2021).

While many groups of people benefit from sharing preferred pronouns (for instance, people with gender-ambiguous names), there is a particular benefit for nonbinary and transgender individuals. Much discourse related to expressing one's preferred

pronouns centers around being an ally for LGBT+ individuals. LGBT+ activists (Wamsley, 2021) encourage people to avoid misgendering others by asking for their pronouns.

The new popularity of preferred pronouns within personal identity expression marks a good opportunity for quantitative, descriptive research. Here we studied the Twitter profile biographies of US users. We estimated the prevalence of users with pronoun lists, contrasted the relative prevalence of words appearing alongside pronoun lists and detected clustering of pronouns lists within the Twitter follow network.

There is a small but growing set of research using social media bios as a measurement tool for personally expressed identity. Using Twitter bios from 2015 to 2018, Rogers and Jones (2021) argued that an increasing number of Americans consider their political affiliation a part of their identity. In 2021, Jones ranked 17,765 unique tokens based on growth over time within US user bios and found pronouns at the top. Using Twitter bios of US partisans, Eady et al demonstrated a decrease in "outward expressions of identification with the Republican Party and Donald Trump" in the wake of the US Capitol insurrection on January 6, 2021 (2021). However, apart from a preprint by Jiang et al (2022), which will be discussed in depth in the Discussion, little quantitative research exists concerning pronoun lists within user biographies on social media sites.

Data and Methods

In this work, we used the Longitudinal Online Profile Sampling method (Jones, 2021) and Twitter profile biographies to measure expressions of personal identity over time. This affords several advantages. First, the prompt for a Twitter bio is open ended, and users are not shown a template; thus, a user's bio is self-generated, self-descriptive text. The average Twitter bio is updated approximately once per year (Rogers & Jones, 2021). Thus, bios are relatively stable, but in large samples and over the course of years, meaningful variation can be observed. Unlike other social media sites, Twitter has never had a separate field for users to enter their preferred pronouns, so the presence or absence of pronoun lists within the bio remains a useful measure across time. Finally, Twitter data is easily accessible.

Sampling and Filtering

The target population was active, US-located Twitter accounts. Active was defined as: observed authoring a tweet. US-located was defined based on the account's profile

location text. The function classifying locations as US or not is available at https://osf.io/472sf. Developed iteratively over years, the function is a set of heuristics to capture common ways Twitter users indicated US and non-US locations. For example, state names and abbreviations indicate US locations; names of national capitals in isolation indicate non-US locations. *Cairo* is mapped non-US, while *Cairo, NY* and *Cairo Illinois* are mapped US.

We are characterizing Twitter *accounts* and their proclivities. The authors are all too aware that social media profiles do not map one-to-one with human individuals. A Twitter account could represent an organization, a software bot or one person pretending to be another. Users of Twitter are not a random sample of the US population. Nevertheless, we believe the study of temporal linguistic trends with Twitter account bios is an interesting, worthwhile endeavor. No matter who or what controls an account, their bios are observed by human individuals. It is impossible to verify, but we believe most accounts *do* represent individuals. There are vastly more people than there are companies. A representative sample of American adults could be had through other means. Traditional household sampling would be one. However, such methods would never reach the scale or temporal resolution of the current work without a gargantuan budget.

Thus, we believe US-located Twitter accounts comprise an interesting target population. First, there are many of them: 76.9 million as of January 2022 (Statista, 2023). Second, a single national context simplifies analysis and interpretation. The authors have a greater understanding of the US context (both are active, US-located Twitter users) than we anticipate we would for a global or multinational sample. Third, we speculate that active, US-located Twitter accounts have outsize influence on attitudes, beliefs and social norms. The Twitter activity we describe here has the potential to affect millions of individuals' attitudes, beliefs and social norms both online and off.

Constructing Annual and Daily Datasets

We constructed cross-sectional datasets at two different temporal resolutions: annual and daily. All data began from the 1% sample of all public tweets. We observed the tweet stream using the Twitter API version 1.1 GET statuses/sample endpoint (Twitter, 2023). We observed the user's biography at the time of posting. If multiple tweets in the

time period were from the same user, we selected one at random to keep and discarded the rest. Thus, a user observed tweeting 100 times in 2015 appears exactly once in the 2015 annual sample, as does a user observed tweeting once in 2015. Similarly, at daily resolution, a user observed tweeting 20 times on 2017-11-07 and a user observed tweeting 4 times both contributed exactly one record on that day in the daily resolution dataset.

We continued observing the stream, but do not include here data from July 30, 2022 onward due to a few potentially disruptive events. In August 2022, the version 1.1 sample stream became less reliable and delivered anomalously low volumes of tweets at times. In September 2022, we migrated to the Twitter API Version 2 sample stream endpoint. (It has been technically sound.) In October 2022, ownership of the platform changed from a public to a private company. Our analysis was performed in July 2022. We have not yet fully examined what effects and artifacts may be present in the late 2022 data.

We generated annual, cross-sectional datasets for every year 2015-2022. (2015 and 2022 were partial years.) In each year, one observation per active, US-located account was recorded. Table 1 lists tallies of unique accounts per year.

Table 1: Unique Accounts by Year

Year	2015	2016	2017	2018	2019	2020	2021	2022
Millions of Unique Accounts	8.56	10.23	10.64	10.31	9.82	10.18	8.17	5.45

We also constructed daily resolution, cross-sectional datasets. The procedure was exactly the same as above, except the time window was one day rather than one year. From daily data, one can more smoothly track the pace of change. One can also observe activity in temporal proximity to high-profile events (Jones & Cisternino, 2022).

It is not possible - to our knowledge - to randomly sample from the population of *all Twitter users*. Instead, we use the random sample of tweets to sample active accounts. This has implications. The more an account posts, the more likely they will appear in our data. This will be stronger (and a desired feature) in the daily resolution data. Higher

representation of frequent tweeters will still be present, but less pronounced, in the annual data. Notably, exclusively lurker accounts (who read but never post) will never be sampled.

Pronoun Lists

We chose to examine five pronoun lists: she/her, he/him, they/them, she/they, and he/they. Under our tokenization process, "he/him", "he/him/his", and "he/his" are each considered different pronoun lists. Because even similar pronoun lists, such as "she/they" and "they/she", can mean different things to the people who choose those labels, we chose to treat each pronoun list as its own category.

Each pronoun list we chose to consider was significantly more prevalent than the next-most-common similar pronoun list, as is demonstrated in Table 2. Apart from these five and similar pronoun lists, the most common pronoun list was a Portuguese-language pronoun list – "ela/dela" –with a prevalence of 2.1 occurrences per 10,000 Twitter bios. This was not frequent enough to warrant inclusion in our analysis. We also considered including neopronouns in our analysis, but no neopronoun list occurred frequently enough to reliably surpass a prevalence of 1 per 10,000 criterion. As a result, we focused only on the five most common pronoun lists: she/her, he/him, they/them, she/they, and he/they.

Table 2: Prevalence of pronoun lists from Jan 1 – June 30, 2022

Pronoun List	Prevalence (per 10,000 unique user bios)	Next-most-common Similar Pronoun List	Prevalence (per 10,000 unique user bios)
She/her	224.4	She/her/hers	11.6
He/him	157.5	He/him/his	14.2
They/them	36.7	They/them/theirs	0.45
She/they	30.1	They/she	5.8
He/they	17.7	They/he	3.5

Note also, if a bio contained two or more of the five pronoun lists we considered, we placed that bio into its own category. A bio containing both "he/him" and "they/them", for instance, would be considered a bio with a pronoun list, but would be counted as an instance of "Multiple pronoun lists" and not counted as an instance of either "he/him" nor "they/them". In 2022, the prevalence of "Multiple pronoun lists" was 4.20.

Tokenizing and Calculating Prevalence

To convert the unstructured biography text data to structured tabular data, we considered each bio to consist of a set of linguistic tokens. We used the regular expression "[^a-zA-Z0-9/'`-]" to tokenize bios. This regular expression split the Twitter bio at any character that was not alphanumeric, a forward slash (as used in pronoun lists), an apostrophe, a backtick (often used as an apostrophe mark), or a hyphen. The resulting list was reduced to the set of distinct tokens which appeared in the Twitter bio. From these sets we could tally the number of unique users whose bios contained any token.

Prevalence was defined as the number of bios a token appeared in per 10,000. We calculated prevalence rather than proportion for convenience. It is easier to understand that "student" has a prevalence of 84 per 10,000 in 2022 than that the proportion of bios that included "student" in 2022 was 0.0084, for instance. We also calculated the prevalence of tokens among subsets of bios. For instance, among Twitter bios containing a pronoun list in 2022, "student" has a prevalence of 202.

$$Prevalence(A) = \frac{\#\ bios\ A\ appears\ in}{total\ \#\ bios} * 10{,}000$$

Comparing prevalence in the total sample to subsets leads to discussion of relative prevalence. We defined the relative prevalence of Token A as the ratio of the prevalence of Token A in the subset to the prevalence of Token A among all bios. Consider an example. In 2022, the token "student" had a prevalence of 84 among all bios and a prevalence of 202 among bios that contained a pronoun list. Thus, the relative prevalence of "student" among bios with a pronoun list is 2.40, or 202/84.

If a relative prevalence was less than 1, we express it as its negative inverse. As an example, "brother" had a prevalence of 16.8 among all bios and a prevalence of 8.1 among bios with a pronoun list. Thus, its unadjusted relative prevalence would be 0.48. Instead, we say that the relative prevalence of "brother" is -1 / 0.48 = -2.1. We adjust the relative prevalence of these tokens for the sole purpose of improving data visualization. It is easier to graphically show the difference between -2 and -4 than between 0.5 and 0.25. Any further calculations involving relative prevalence (such as confidence interval calculations) used unadjusted relative prevalence. A positive relative prevalence means the token was

more common in bios with a pronoun list; a negative relative prevalence means the token was less common.

$$Relative\ Prevalence(A) = \frac{Prevalence(A)\ in\ subset}{Overall\ Prevalence(A)}$$

$$*If\ Rel.\ Prev.\ < 1,\ Relative\ Prevalence = \frac{-1.0}{Original\ Rev.\ Prev.}$$

We expect the average token to have a relative prevalence of roughly 1.1. Discussion of the expected relative prevalence can be found in Appendix A.

We also were interested in determining the prevalence and relative prevalence of bigrams and trigrams. To determine which n-grams appeared in any given bio, we again tokenized the Twitter bios using the same regular expression as before. We considered each group of n consecutive tokens to be a n-gram, so a Twitter bio that is tokenized into a list of 20 candidate tokens will have 19 candidate bigrams and 18 candidate trigrams. We discarded any bigrams or trigrams containing a pronoun list we study. We then converted the list of n-grams into a set of unique n-grams. Using the same methodology used to calculate the prevalence and relative prevalence of tokens, we calculated the prevalence and relative prevalence of each bigram and trigram.

Results

Daily Prevalence

We first wanted to understand how pronoun list usage had changed over time. To do this, we calculated the prevalence of each of our five pronoun lists each day, using our daily cross-sectional datasets. We then plotted each prevalence over time.

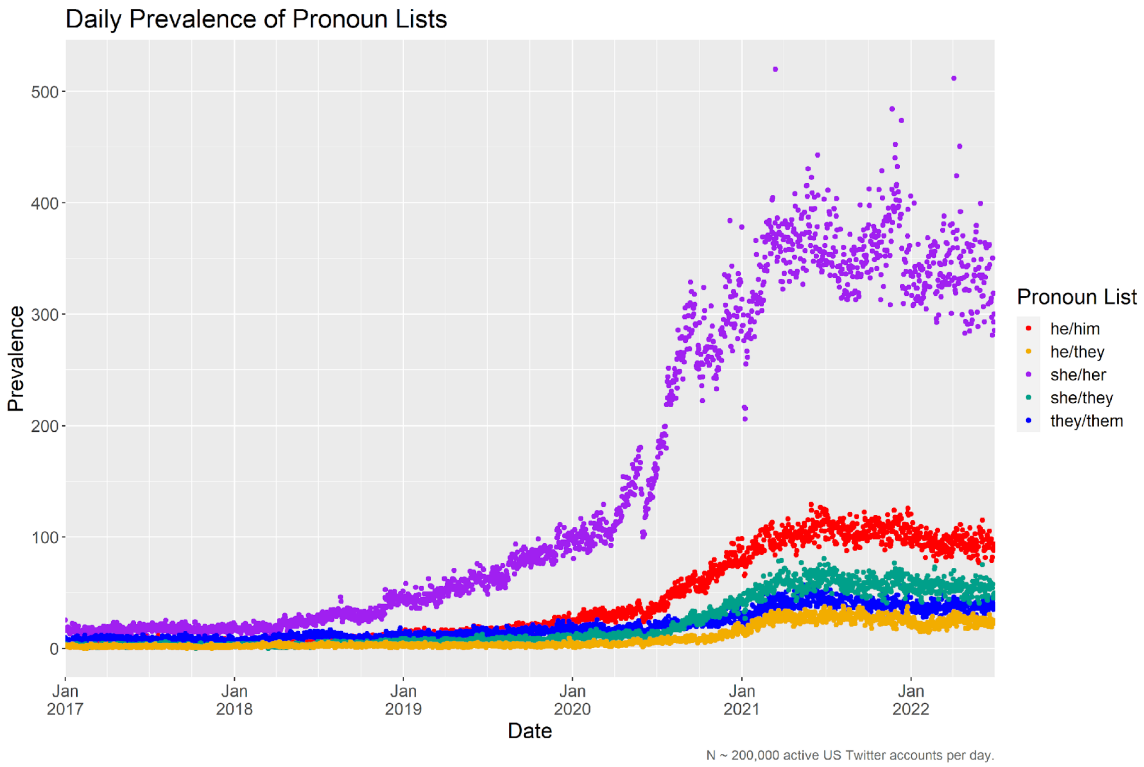

Figure 1: Daily Prevalence of Pronoun Lists

As is clear from the figure, pronoun lists appeared at very low rates until beginning to grow in prevalence in 2018. (We have omitted 2015 and 2016 in the figure, because the values are very similar to those in 2017.) Between 2018 and 2021, each pronoun list experienced relatively high rates of prevalence growth. Since 2021, prevalence has plateaued.

Consider what the values in Figure 1 imply. On each day, if one randomly selected 10,000 tweets from unique US users, then the prevalence is an estimate of how many authors of those tweets had she/her, he/him, they/them, he/they or she/they in the text of their profile bio. In all series, the numbers have increased substantially. From January 1, 2017 to June 30, 2022, the prevalence of she/her increased nearly 11x from 25.63 to 285.57. The prevalence of he/him increased 20x from 4.42 to 90.56. They/them increased 4x from 9.72 to 39.44. She/they increased 16x from 3.09 to 49.66, while he/they increased 14x from 1.77 to 24.83.

Co-occurring N-grams

Next, we investigated the extent to which other tokens, bigrams, and trigrams co-occurred with pronoun lists. In Figure 2, we examine the relative prevalence of various n-grams among bios with “she/her” pronouns and “he/him” pronouns, which together account for 81.6% of pronoun lists in our 2022 cross-sectional dataset. It should be noted that n-grams are only included in this chart if their prevalence both among all Twitter bios and among Twitter bios with a prevalence of at least 1.0.

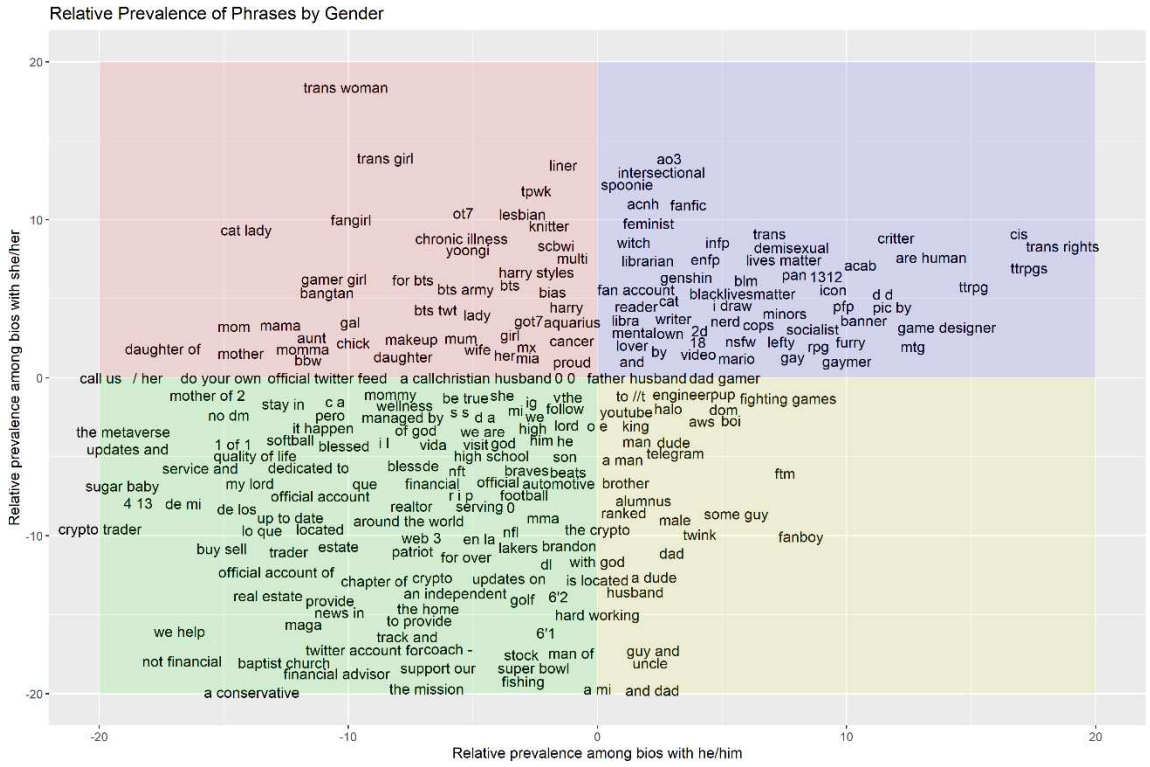

Figure 2: Relative Prevalence of N-Grams by Pronoun List

Note that Figure 2 depicts only n-grams of -20 <= relative prevalence <= 20 among both bios with “he/him” and “she/her”. This is for readability purposes; as the range of relative prevalences included in this chart increases, fewer n-grams can be included for any given range of relative prevalences, and the chart becomes less information dense. We

chose this cutoff as it maximizes the amount of token information that can be shared. Only three outlier tokens are excluded from this figure: “alumna”, which would appear at the point (-111.4, 3.1); “Latina”, at point (-63.4, 4.5); and “father”, at point (1.1, -58.7). Appendix E contains similar figures that contain (1) only tokens, (2) only bigrams, and (3) only trigrams. Appendix C links to code to visualize this figure with different ranges of relative prevalence. Additionally, Appendix D contains information about a glossary which explains the n-grams in this chart which are not necessarily intuitive.

Each quadrant in this figure represents a different category of n-gram. The top left quadrant contains n-grams that co-occur disproportionately frequently alongside “she/her” pronouns and disproportionately infrequently alongside “he/him” pronouns, while the bottom right quadrant represents the opposite pattern. They are dominated by likely-female terms and likely-male terms, respectively, which is consistent with the proposition that people generally use their Twitter bios to describe themselves (i.e. “Proud husband”) and not others (i.e. “My husband is …”). Of particular interest are the few n-grams in these quadrants that aren’t inherently gendered. “Engineer” occurs far more often alongside “he/him” than “she/her”, consistent with data showing that the vast majority of engineers and engineering graduates are male (De Brey et al, 2021; Employment, 2022). The top left quadrant contains references to BTS, a Korean pop music group, and to astrology.

The top right quadrant of this figure contains n-grams with positive prevalences both among bios with “she/her” and “he/him” pronouns. Most of these n-grams are related to either (1) left-wing politics, (2) gender or sexual identity, (3) video or tabletop gaming, or (4) slang words. The bottom left quadrant, on the other hand, contains n-grams with negative relative prevalences both among bios with “she/her” and “he/him” pronouns. These n-grams are generally related to (1) sports, (2) right-wing politics, (3) religion, (4) finance or cryptocurrency, or (5) are Spanish words.

Together, “she/her” and “he/him” account for just over 80% of the pronoun lists in our dataset. Thus, n-grams in the top right quadrant likely have positive relative prevalences among all bios with a pronoun list, while n-grams in the bottom left quadrant likely have negative prevalences among all bios with a pronoun list. To be certain, we must examine tokens, bigrams, and trigrams with particularly large or small relative prevalences

among all bios with a pronoun list. Figure 3 displays these tokens, and Appendix E contains similar figures generated from bigrams and trigrams.

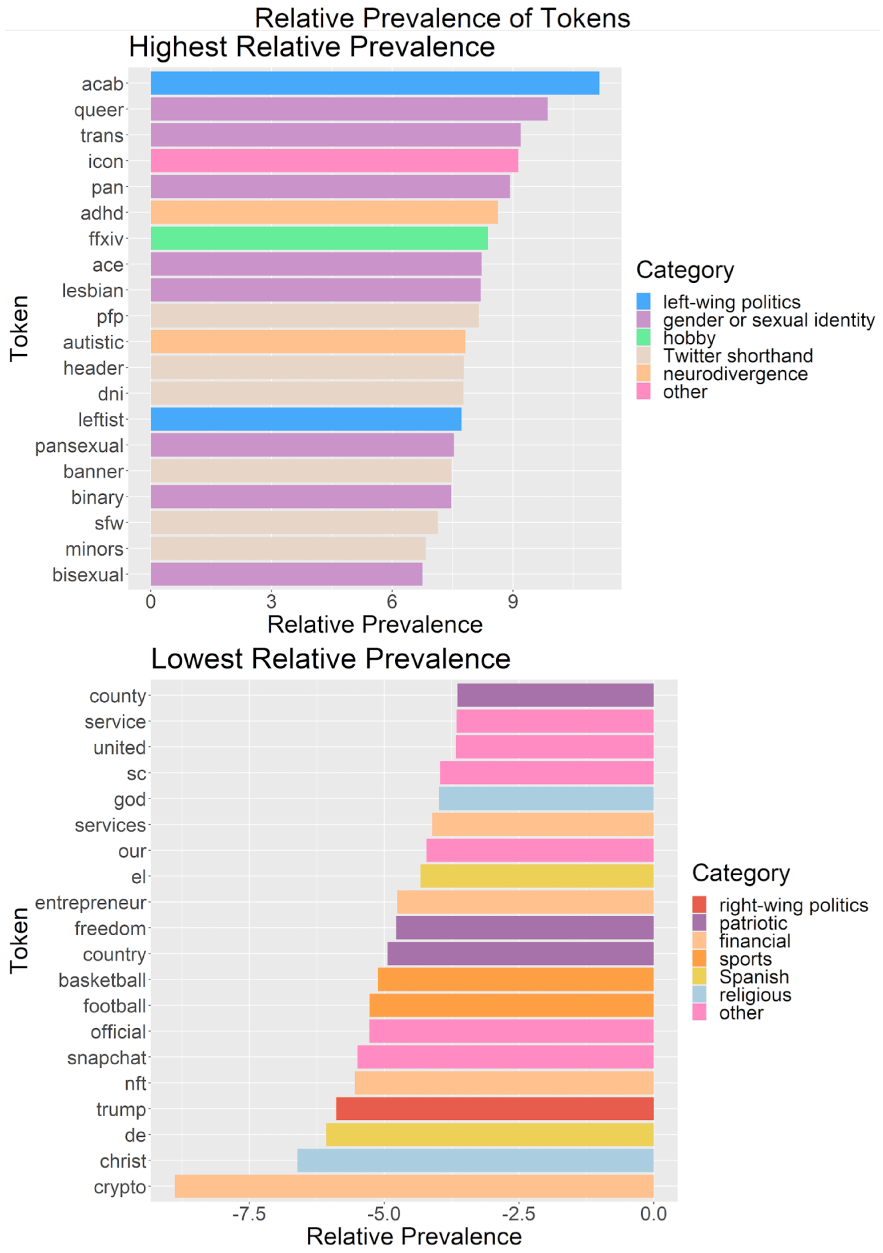

Figure 3: Tokens with Highest and Lowest Relative Prevalences

We see the same categories here as we saw in the bottom left and top right quadrants of Figure 2. Unsurprisingly, a large proportion of the tokens with the largest relative prevalence values are related to gender or sexual identity. Patriotic and religious tokens appear to be particularly unlikely to be listed alongside a pronoun list. Tokens we categorize as left-wing politics appear exclusively in the high relative prevalence list, while right-wing politics tokens appear exclusively in the low relative prevalence list. Two Spanish-language tokens have low relative prevalence; few Twitter bios contain both Spanish prepositions and English pronoun lists. Additionally, financial and particularly cryptocurrency tokens dominate the list of tokens with low relative prevalences.

Characteristics of Pronoun List Users

Also of interest were other characteristics of Twitter users who included a pronoun list in their bio. Did they create their Twitter accounts more or less recently? Were they more or less active than the average user in our sample? Were they more or less likely to be verified? We answer these questions in turn.

Here, we split Twitter bios without a pronoun list into two categories: users with a blank bio and users with a bio that is neither blank nor contains a pronoun list. We find that Twitter users with a blank bio tend to be particularly less active and less connected. By differentiating, one can observe both the rate of writing something and not including pronouns and the rate of simply leaving one's bio blank. In Figure 4, we examine the 5,444,623 Twitter profiles gathered between January 1 and June 30 of 2022.

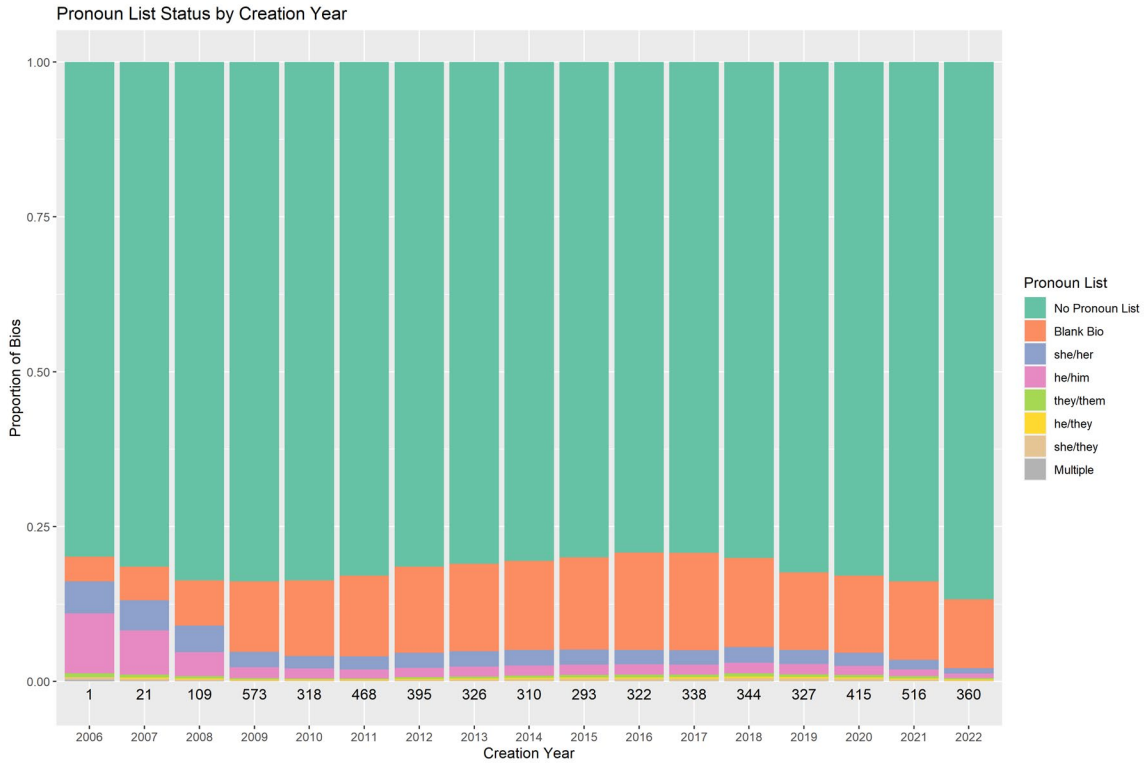

Figure 4: Distribution of Pronoun List Status in Twitter Bios by Join Year

(2022 active, US-located users)

Note that the n-values are in thousands of bios. This chart illustrates two general phenomena. Recently-created accounts (2021 and 2022) are less likely to include pronoun lists in their bios. Long-tenured, still-active accounts (2006-2008) and the 2018-join cohort are especially likely to include pronouns.

Next, we examined metrics regarding the influence of Twitter users—their friend count (the number of users they follow), their follower count, and whether they are verified.

Table 3: Influence of Twitter Users by Pronoun List Status

Pronoun List	Proportion of Bios Verified	Mean Follower Count	Mean Friend Count
She/her	1.65%	1809	829
He/him	1.60%	1756	894
They/them	0.46%	915	628
She/they	0.41%	1214	603
He/they	0.46%	997	702
Multiple	0.30%	993	725
No Pronoun List	1.55%	2742	805
Blank Bio	0.08%	519	406

We see a broadly similar pattern across all three metrics. Twitter users with "she/her" or "he/him" in their bio are roughly as likely to be verified as users without a pronoun list, while users with other pronoun lists are much less likely to be verified. Similarly, Twitter users with "she/her", "he/him", or no pronoun list have the largest mean follower and friend counts, while users with other pronoun lists have slightly fewer.

Finally, we investigated the activity level of Twitter users based on their pronoun list status. To do so, we examined status count (the total number of tweets and retweets posted by a user) of Twitter users, grouped by pronoun list status.

Table 4: Status Count of Twitter Users by Pronoun List Status

Pronoun List	1st Percentile	10th Percentile	50th Percentile	90th Percentile	99th Percentile
She/her	21	342	13,647	33,871	12,6046
He/him	20	316	11,936	29,205	11,4134
They/them	13	218	11,474	28,756	11,1799
He/they	13	214	11,183	27,306	11,8203
She/they	17	256	12,014	30,567	10,5242
Multiple	15	324	13,462	34,223	13,8315
Blank Bio	3	38	5,566	13,693	6,9664
No Pronoun List	5	87	9,993	24,963	11,1918

Once again, there are generally consistent results. Twitter users with blank bios are by far the least engaged. Twitter users without a pronoun list consistently have lower status count than Twitter users with a pronoun list, while the distributions of status count among Twitter users with a pronoun list are similar regardless of the pronoun list. These trends also hold when controlling for the year the Twitter account was created. This accords with our observation from Table 2 and Figure 1 that the prevalence of pronoun lists is larger among our Daily Cross Sectional Datasets than our Annual Cross Sectional Datasets.

Put together, this data suggests that Twitter users with a pronoun list other than "he/him" or "she/her" pronouns are less a part of the "in-group" of Twitter than users without a pronoun list: they have fewer followers and friends despite being more active, and they are much less likely to be verified than other users. Twitter users with "he/him" or "she/her" in their bio joined Twitter earlier and are more active than Twitter users without a pronoun list, but both groups are verified at similar rates and have similar mean friend counts. Twitter users without a pronoun list have a far larger average follower account than users with any given pronoun list, but this gap nearly disappears when excluding verified accounts: No Pronoun List has a mean follower account of 1232, compared to 1136 for "he/him", the next-most-common pronoun list status. Predictably, Twitter users with a blank bio are the least followed, verified and active.

The slight decrease in prevalence of pronoun lists among accounts created in 2021 or 2022 is of interest as it coincides with a slight decrease in pronoun list prevalence since 2021 as demonstrated in Figure 1. This may suggest that the phenomenon of pronoun lists in Twitter bios has peaked.

Pronoun List Clustering

We next explored whether users with a pronoun list in their bio are connected with other pronoun list users at a disproportionately high rate. Specifically, we examined "following" ties between users with and without a pronoun list in their bio. First, we randomly selected equal numbers of users from both categories (N approximately 3000 each). The presence of a pronoun list in the bio in 2022 defined the category membership. Next, we used the GET followers/list request to receive a list of their 1000 most recent Twitter followers (or, if the user had less than 1000 Twitter followers, all of their

followers). We then used the same regex expression as above, *"[^a-zA-Z0-9/'`’-]"*, to split each followers' Twitter bio into tokens, tallied how many followers had a pronoun list in their bio, and calculated the proportion. We repeated the above process using the GET friends/list request to calculate the proportion of Twitter friends with a pronoun list in their Twitter bio. (A Twitter user's friends are the accounts they follow).

We observed that users with a pronoun list in their bio were disproportionately likely to be friends with and followed by other users with pronoun lists in their bio. See Figures 5 and 6.

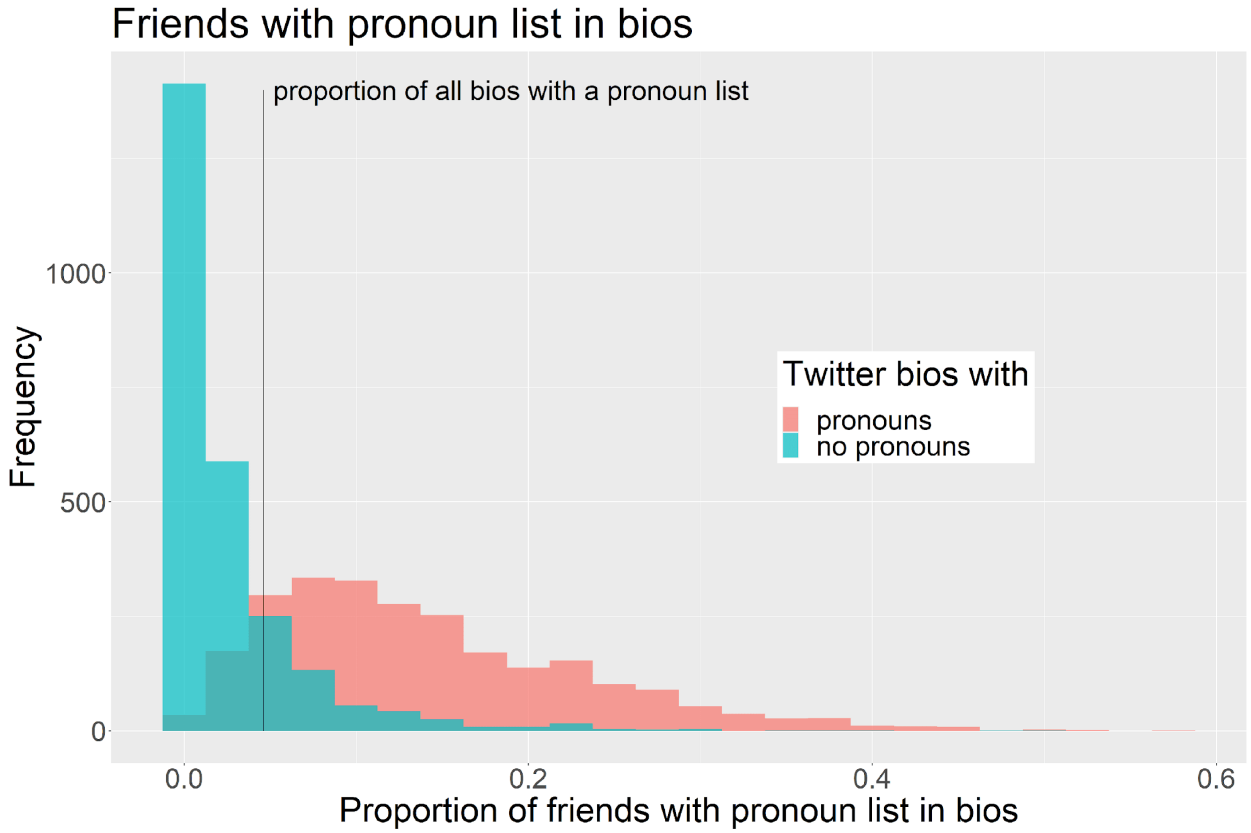

Figure 5: Distribution of Friends of Random Users with Pronouns in Bios

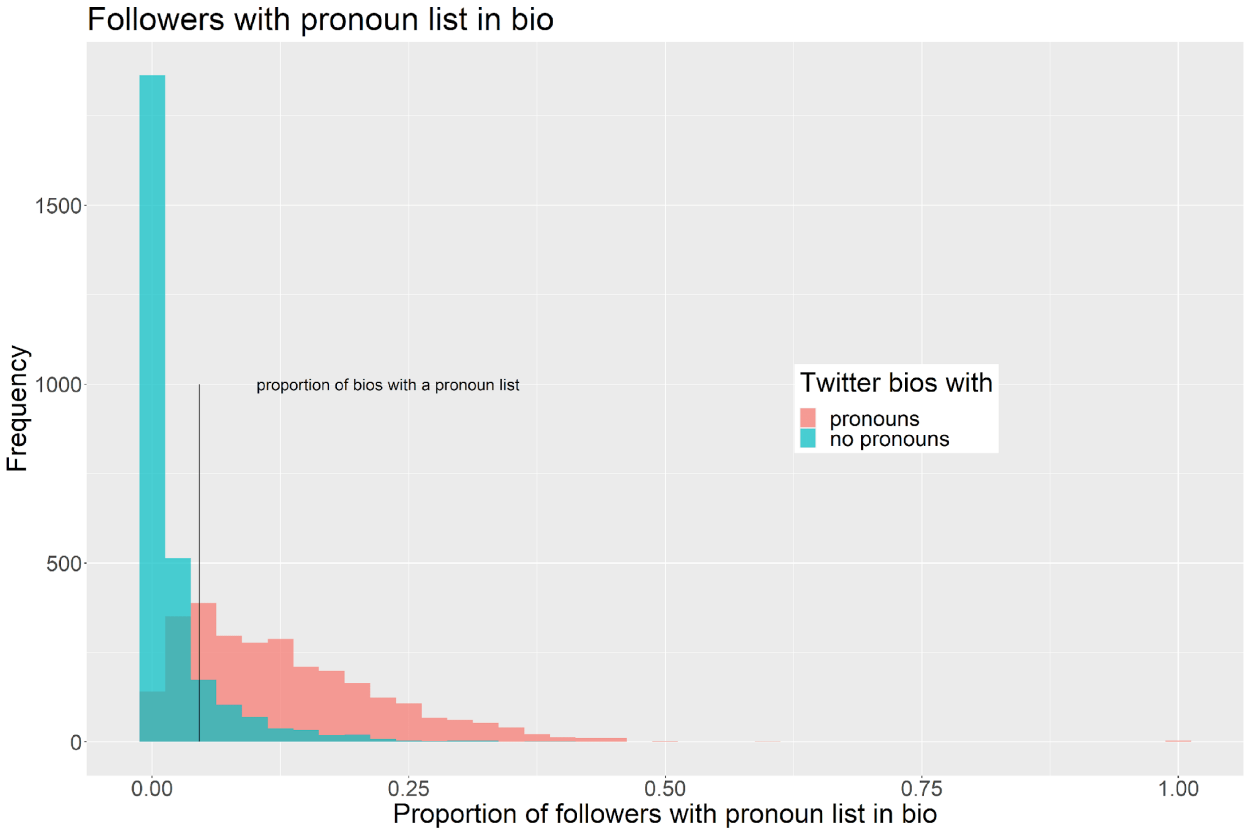

Figure 6: Distribution of Friends of Random Users with Pronouns in Bios

We found that, during the first six months of 2022, 4.61% of Twitter bios contained at least one pronoun list. If pronoun lists were uniformly distributed across Twitter bios, both distributions should be centered around 0.0461. Instead, we found that 13.0% of the followers and 14.1% of the friends of a random Twitter user with a pronoun list in their bio also have pronoun lists in their bio. By contrast, only 2.3% of the followers and 2.6% of the friends of a Twitter user without a pronoun list in their bio have a pronoun list in their bio. A t-test for a difference of proportions indicated $p < 0.001$ confidence of a statistically significant difference in both cases. This data supports the proposition that there is clustering among Twitter users with pronoun lists in their bio.

Discussion

Using Twitter bios from US users observed between early 2015 and June 30, 2022, we found evidence for increasing prevalence of users with pronoun lists, differences in non-bio attributes among users with different pronoun lists, systematic categories of tokens co-occurring (or not) with pronoun lists and clustering of pronoun list usage within the follow network. We will discuss each finding in turn.

The number and proportion of individuals listing pronouns in their bios increased substantially. As can be seen in Figure 1, recent years saw manyfold growth in prevalence for every pronoun list. This accords with other observations (Jiang et al., 2022; Jones 2021). Interestingly, the current data suggest this growth may have plateaued.

We find that the earliest Twitter users (those who created their account during 2006, 2007, or 2008) are more likely to include pronoun lists in their Twitter bios than Twitter users who created their account later. Perhaps the earliest Twitter adopters were younger, more educated, or politically liberal than later cohorts, and demographics would explain the difference. Or perhaps the difference is one of personality: the same type of person embraces new behaviors (Twitter and preferred pronouns) before mass adoption. Interestingly, “he/him” is the most popular pronoun list among bios created in 2006 or 2007, while “she/her” is most popular among bios created in any other year. This is in line with past research demonstrating that Twitter users between 2006 and 2007 were disproportionately male (Mislove et al, 2021).

We found that Twitter users with any pronoun list were more active on Twitter than users without a pronoun list in their bio. However, while Twitter users with “she/her” or “he/him” pronouns have comparable influence to Twitter users without a pronoun list (as measured by average follower count and proportion of accounts verified), Twitter users with “she/they”, “he/they”, or “they/them” pronoun lists have less influence than Twitter users without a pronoun list.

We note that corporate or brand accounts are a potential source of bias in this analysis. Corporate accounts are unlikely to contain a pronoun list (as they represent a non-human entity) but may be more likely to be verified or have large follower counts. Indeed, when removing verified accounts from our analysis, we found that the gap between the average follower count among Twitter users without a pronoun list and Twitter users with

"she/her" (the pronoun list with the largest average follower count) in their bio drops from over 900 followers to under 100 followers.

Users with a pronoun list in their Twitter bio are more likely to also include n-grams related to left-wing politics and gender or sexual identity and less likely to also include n-grams related to finance, sports, religion, patriotism, or right-wing politics. Polling indicates that individuals who support publicly sharing preferred pronouns tend to be younger, more liberal, and less religious (Kirzinger et al, 2021; Lipka & Tevington, 2022). Our results match these results and suggest new categories that may distinguish those who support and adopt preferred pronoun usage from those who do not.

Users with a pronoun list in their bio are more likely to follow, and be followed by, other Twitter users with a pronoun list in their bio. Similarly, users who do not cluster with others who do not. Here we moot three theories for why this would be:

1. **Pronoun-specific homophily.** Twitter users who publicly share their preferred pronouns presumably have a positive opinion of people sharing their preferred pronouns in Twitter bios. They may be more likely to follow another user if the other user has a pronoun list in their Twitter bio, all else equal.
2. **General homophily.** It is a human tendency to associate with similar others. All or much of the sorting we observe here may have been done based on other (including unobserved) common attitudes, affiliations, or demographics.
3. **Social contagion.** Perhaps encountering pronoun lists in the Twitter bios of social ties encourages one to adopt that same behavior. The clustering we observe is consistent with contagion-like spread as the mechanism for growth.

To be clear, any combination of the above (including none or all) could be true and accord with our demonstration of clustering. Theory 1 might be tested experimentally or with agent-based modelling just as racial segregation in tie formation has (Firmansyah & Pratama, 2021; Wimmer & Lewis, 2010). Regarding Theory 2, we have demonstrated that other affiliations besides sexual or gender identity co-occur alongside pronoun lists. For instance, FFXIV is an acronym for a videogame: Final Fantasy 14. Left-leaning political opinion and neurodivergence signifiers (i.e. adhd and autistic) also co-occur. The predictable co-occurrence of these tokens could be due to general homophily.

One interesting avenue to explore further would be the prominence of political spectrum signifiers associated with the presence or absence of pronouns lists. Here we have shown numerous signifiers (i.e. leftist, acab) were more likely to occur in the bios of Twitter users with pronoun lists than users drawn at random. It has been previously observed that Twitter users are more likely to be connected to copartisans (Colleoni et al., 2014). Additionally, Twitter users are more likely to form social ties with accounts with their same political affiliation (Mosleh et al, 2021). Users with pronouns in their bio may be more connected with other users with pronoun lists in their bios as an indirect result of shared political affiliations. Further analysis could attempt to determine which is the tail and which is the dog – is the pronoun clustering incidental to ideological sorting or its own phenomenon?

Alternatively (or additionally), clustering of pronoun list users could be the result of that behavior spreading online through social contagion. Just as infectious diseases spread when an uninfected individual encounters an infected individual, a social contagion spreads when an individual participating in the phenomenon encounters an individual not yet participating. Voting is one example of a social contagion; Facebook users were more likely to vote when it was made salient to them that their close ties had voted (Bond et al, 2012; Jones et al, 2017). Careful examination of the timecourse of pronoun list additions within the network would presumably reveal evidence for or against a contagion-like mechanism of growth.

After completing the work described here, the authors discovered a preprint manuscript with similar methods and (encouragingly) similar results. Jiang et al (2022) examined pronoun list usage within a set of Twitter bios associated with English-language tweets relating to COVID-19. Sampling is one major difference in the studies. Jiang et al used a dataset of just over two billion tweets related to COVID-19, gathered by Chen et al (2020). The data was collected based on keyword matching without restriction on geography. The sample consists of tweets posted between January 21, 2020 and November 5, 2021. 63.66% of tweets were in English.

The data used in this paper consisted of bios of US users whose tweets appeared in the 1% random sample stream provided by the Twitter API. About 200,000 unique users

met these criteria each day. The sample was collected from early 2015 through June 30, 2022. Thus, one would expect the Jiang et al sample to reflect those discussing COVID-19 within the global English-speaking set of users. The current sample reflects active users over a longer period with a US location listed in their profile.

Despite the differences in sampling, the two reports contain similar results. In their sample, Jiang et al. estimated 8% of tweets originated from users who included gender pronouns in the bio. In this work, we found 4.61% of unique users included any pronoun list. Jiang et al. found roughly double the number of tweets from she pronoun users as he pronoun users. Here, we found roughly triple the number of unique users with she/her in the bio as he/him. These numbers strike us as independent replication of the same patterns. The small differences in exact values likely are the result of counting tweets in a global English sample versus counting users in a US sample.

Further, our two investigations yielded remarkably similar results for co-occurring tokens. (Compare Figure 3 across both manuscripts.) Trump, god and country show up in both samples as reliable predictors for the absence of a pronoun list. The low-frequency tokens acab and icon appear in both lists of positive predictors. Evidence for the co-occurrence of pronouns and signifiers of gender or sexual identity and left-wing politics emerged from both datasets. We believe these consistencies reflect true relationships which exist within users' self-perceptions.

Finally, both manuscripts investigated social network effects but in different ways. Here, we presented descriptive evidence that pronoun use in the bio clustered in the follow network. Jiang et al. built a deep neural network to predict which users would add pronouns to their bio. They included many features in the model, but for the purpose of evaluating social network effects, it is most important to note they combined temporally-bounded adoption information (when users adopted pronouns) and temporally-bounded interaction information (who retweeted/mentioned whom and when did they do so). They drew two interesting conclusions. First, they argued the presence of users with non-binary pronouns in one's network was "linked" with subsequent addition of pronouns to one's bio. Second, they argued for selective effects: she/her neighbors beget she/her pronouns in the focal user, and corresponding linked behavior is present for he/him and non-binary pronouns.

Both investigations of social network effects provide tantalizing hints that Twitter bio data could and should be used to explore the spread of new social norms.

Studies of Twitter bios over time are valuable for another reason: the methods can be replicated nearly exactly and new data is generated constantly. Knowledge decays because the world is constantly changing (Munger, 2019). Temporal validity should thus be a constant concern of social science. The methods described here can be repeated on any self-descriptive short text data past or future. That does not guarantee that the results currently described shall hold true, but it does at least provide the opportunity for other researchers to compare results across time, geography and media.

Theorists of identity should engage with these new methods and results. Stets et al. (2020) recently posited a refined Identity Theory in which individuals define themselves with "role-related self-perceptions". That is, identity is the various roles in society one claims (e.g., mother, doctor, gardener, etc.). Furthermore, the theory states that people try out different identities and keep those that receive social validation (Burke & Stets, 2009). We posit that sharing preferred pronouns has only recently begun receiving validation. It would follow that the phenomenon would become more common following greater validation by others. While not considered here, one could imagine ways to test whether this process of trial-validation-spread best describes what one observes in Twitter bio data.

Conclusion

We have observed that many more US Twitter users included preferred pronouns in 2022 than did in previous years. Within those bios, we found systematicity in which words co-occurred with pronoun lists. Additionally, the evidence revealed clustering within the Twitter follow network for this expression – pronoun users were more likely to follow and be followed by a pronoun user than one would expect if ties were independent of bio content. Far beyond this single manifestation of personal identity expression, Twitter bios provide an opportunity to study how individuals perceive and present their selves.

Acknowledgements

We would like to thank Margot Hare for her comments on early drafts of this work.

Funding Information

This material is based upon work supported by the National Science Foundation under grants IIS-1927227, OAC-1950052 and CCF-2208663.

References

Arruda, W. (2021, April 1). *Major Changes to Your Linkedin Profile You Need to Know About*. Forbes. Retrieved October 23, 2022, from https://www.forbes.com/sites/williamarruda/2021/04/01/major-changes-to-your-linkedin-profile-you-need-to-know-about/?sh=1e00095e3efc

Ballard, J. (2022, August 2). *How Americans feel about gender-neutral pronouns in 2022*. YouGov America. Retrieved October 15, 2022, from https://today.yougov.com/topics/politics/articles-reports/2022/08/02/how-americans-gender-neutral-pronouns-2022-poll

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, *489*(7415), 295–298. https://doi.org/10.1038/nature11421

Burke, P. J., & Stets, J. E. (2009). *Identity theory*. Oxford University Press.

Chen, E., Lerman, K., & Ferrara, E. (2020). Tracking social media discourse about the COVID-19 pandemic: Development of a public coronavirus Twitter data set. *JMIR Public Health and Surveillance*, *6*(2). https://doi.org/10.2196/19273

Chen, T.-P. (2021, September 19). *Why Gender Pronouns are Becoming a Big Deal at Work*. The Wall Street Journal. Retrieved October 15, 2022, from https://www.wsj.com/articles/why-gender-pronouns-are-becoming-a-big-deal-at-work-11631797200

Colleoni, Elanor & Rozza, Alessandro & Arvidsson, Adam. (2014). Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data. Journal of Communication. 64. 10.1111/jcom.12084.

De Brey, C., Snyder, T.D., Zhang, A., and Dillow, S.A. (2021). Digest of Education Statistics 2019 (NCES 2021-009). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

Eady, Gregory, Frederik Hjorth, and Peter Thisted Dinesen. 2021. "Do Violent Protests Affect Expressions of Party Identity? Evidence from the Capitol Insurrection." https://osf.io/dvgj3/#!.

Employment. Society of Women Engineers. (2022, August 19). Retrieved January 10, 2023, from https://swe.org/research/2022/employment/

Firmansyah, F. M., & Pratama, A. R. (2021, September 6). Why Do Homogeneous Friendships Persist in a Diverse Population? Making Sense of Homophily. https://doi.org/10.31235/osf.io/zr7p3

Galanes, P. (2021, April 29). *Do I really need to state my pronouns?* The New York Times. Retrieved October 15, 2022, from https://www.nytimes.com/2021/04/29/style/pronouns-gender-work-social-qs.html

Google Trends. (2022). Google Trends. Google Trends. Retrieved October 19, 2022, from https://trends.google.com/trends/explore?date=today%205-y&geo=US&q=he%2Fhim,she%2Fher,they%2Fthem

Herman, J.L., Flores, A.R., Brown, T.N.T., Wilson, B.D.M., & Conron, K.J. (2017). Age of Individuals who Identify as Transgender in the United States. Los Angeles, CA: The Williams Institute.

Herman, J.L., Flores, A.R., O'Neill, K.K. (2022). How Many Adults and Youth Identify as Transgender in the United States? The Williams Institute, UCLA School of Law

Instagram [@Instagram]. (2021, May 11). Add pronouns to your profile ✨ The new field is available in a few countries, with plans for more. [Image Attached]. [Tweet]. Twitter. https://twitter.com/instagram/status/1392176784028749824

Jiang, J., Chen, E., Luceri, L., Murić, G., Pierri, F., Chang, H. C. H., & Ferrara, E. (2022). What are Your Pronouns? Examining Gender Pronoun Usage on Twitter. arXiv preprint arXiv:2207.10894.

Jones JJ, Bond RM, Bakshy E, Eckles D, Fowler JH (2017) Social influence and political mobilization: Further evidence from a randomized experiment in the 2012 U.S. presidential election. PLoS ONE 12(4): e0173851. https://doi.org/10.1371/journal.pone.0173851

Jones, J. J. (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). https://doi.org/10.1371/journal.pone.0260185

Jones, J. J., & Cisternino, I. (2022). LGBTQ Visibility Online Measured Consistently and Persistently from 2015 through Today. SocArXiv. https://doi.org/10.31235/osf.io/mcax8

Kirzinger, A., Kates, J., Dawson, L., Muñana, C., & Brodie, M. (2021, August 25). *Majorities Support Policies Banning Discrimination Against LGBTQ Individuals' Health Care Access*. KFF. Retrieved October 16, 2022, from https://www.kff.org/other/poll-finding/majorities-support-policies-banning-discrimination-against-lgbtq-individuals-health-care-access/

Lipka, M., & Tevington, P. (2022, July 7). *Attitudes about transgender issues vary widely among Christians, religious 'nones' in U.S.* Pew Research Center. Retrieved October 16, 2022, from https://www.pewresearch.org/fact-tank/2022/07/07/attitudes-about-transgender-issues-vary-widely-among-christians-religious-nones-in-u-s/

McClain, C., Widjaya, R., Rivero, G., & Smith, A. (2022, April 28). *The Behaviors and Attitudes of U.S. Adults on Twitter*. Pew Research Center. Retrieved October 15, 2022, from https://www.pewresearch.org/internet/2021/11/15/1-the-views-and-experiences-of-u-s-adult-twitter-users/

McGlashan, H., & Fitzpatrick, K. (2018). 'I use any pronouns, and I'm questioning everything else': Transgender youth and the issue of gender pronouns. *Sex Education*, *18*(3), 239–252. https://doi.org/10.1080/14681811.2017.1419949

Minkin, R., & Brown, A. (2021, July 27). *Rising shares of U.S. adults know someone who is transgender or goes by gender-neutral pronouns*. Pew Research Center. Retrieved October 15, 2022, from https://www.pewresearch.org/fact-tank/2021/07/27/rising-shares-of-u-s-adults-know-someone-who-is-transgender-or-goes-by-gender-neutral-pronouns/

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., & Rosenquist, J. (2021). Understanding the Demographics of Twitter Users. Proceedings of the International AAAI Conference on Web and Social Media, 5(1), 554-557. https://doi.org/10.1609/icwsm.v5i1.14168

Mosleh, M., Martel, C., Eckles, D., & Rand, D. G. (2021). Shared partisanship dramatically increases social tie formation in a twitter field experiment. *Proceedings of the National Academy of Sciences*, *118*(7). https://doi.org/10.1073/pnas.2022761118

Munger, K. (2019). The limited value of non-replicable field experiments in contexts with low temporal validity. Social Media+ Society, 5(3), 2056305119859294.

Pronouns: A how-to. The Diversity Center. (2021, August 13). Retrieved July 12, 2022, from https://www.diversitycenterneo.org/about-us/pronouns/

Rogers, N., & Jones, J. J. (2021). Using Twitter Bios to measure changes in self-identity: Are Americans defining themselves more politically over time? *Journal of Social Computing*, *2*(1), 1–13. https://doi.org/10.23919/jsc.2021.0002

Statista. (2023). Countries with most Twitter users 2022. Statista. https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

Stets, J. E., Burke, P. J., Serpe, R. T., & Stryker, R. (2020). Getting identity theory (IT) right. In *Advances in Group Processes* (Vol. 37, pp. 191-212). Emerald Publishing Limited.

Stewart, R. D. (2022, September 26). *New! More Easily Add and Manage Your Pronouns in Zoom*. Zoom Blog. Retrieved October 15, 2022, from https://blog.zoom.us/zoom-pronoun-sharing/

University of California, Davis. (2021, September 26). *Pronouns and inclusive language*. LGBTQIA Resource Center. Retrieved October 15, 2022, from https://lgbtqia.ucdavis.edu/educated/pronouns-inclusive-language

The University of Maryland. (n.d.). *Good practices: Names and pronouns*. Good Practices: Names and Pronouns | LGBT Equity Center. Retrieved July 12, 2022, from https://lgbtq.umd.edu/good-practices-names-and-pronouns

Twenge, J. M., Campbell, W. K., & Gentile, B. (2012). Changes in pronoun use in American books and the rise of individualism, 1960-2008. *Journal of Cross-Cultural Psychology*, *44*(3), 406–415. https://doi.org/10.1177/0022022112455100

Statista. (2023). Countries with most Twitter users 2022. Statista. https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

Wamsley, L. (2021, June 2). *A guide to gender identity terms*. NPR. Retrieved July 5, 2022, from https://www.npr.org/2021/06/02/996319297/gender-identity-pronouns-expression-guide-lgbtq

Wimmer, A., & Lewis, K. (2010). Beyond and below racial homophily: ERG models of a friendship network documented on Facebook. American journal of sociology, 116(2), 583-642.

Wojcik, S., & Hughes, A. (2021, January 7). *Sizing up Twitter users*. Pew Research Center. Retrieved October 15, 2022, from https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/

YouGov. (2022, July 8). *Viewpoints on pronouns: YouGov poll: June 29 - July 4, 2022*. YouGov. Retrieved July 8, 2022, from https://today.yougov.com/topics/lifestyle/articles-reports/2022/07/08/viewpoints-pronouns-yougov-poll-june-29-july-4-202

Appendices

Appendix A: Expected Relative Prevalence

If Twitter bios were filled out randomly, we would expect the relative prevalence of a random token among bios with a pronoun list to be 1.0. Because Twitter bios are not written randomly, we expect to see both large positive relative prevalences and large negative prevalences (which are the same as small positive unadjusted relative prevalences). One important factor is blank bios. In any given year between 2015-2022, roughly 15% of active Twitter users in the US had nothing listed in their bio. (In the 2022 dataset, 13.1% of bios are blank). Because empty bios cannot contain pronoun lists, we

speculated that the average token would have a relative prevalence of slightly greater than 1.

To test this, we plotted the prevalence of a given token in bios with a pronoun list over the prevalence of a given token in all bios and calculated the slope. The slope of the graph, which represents the average relative prevalence, is 1.11. This value is close enough to a theoretical average relative prevalence of 1.00 that it does not affect our analysis.

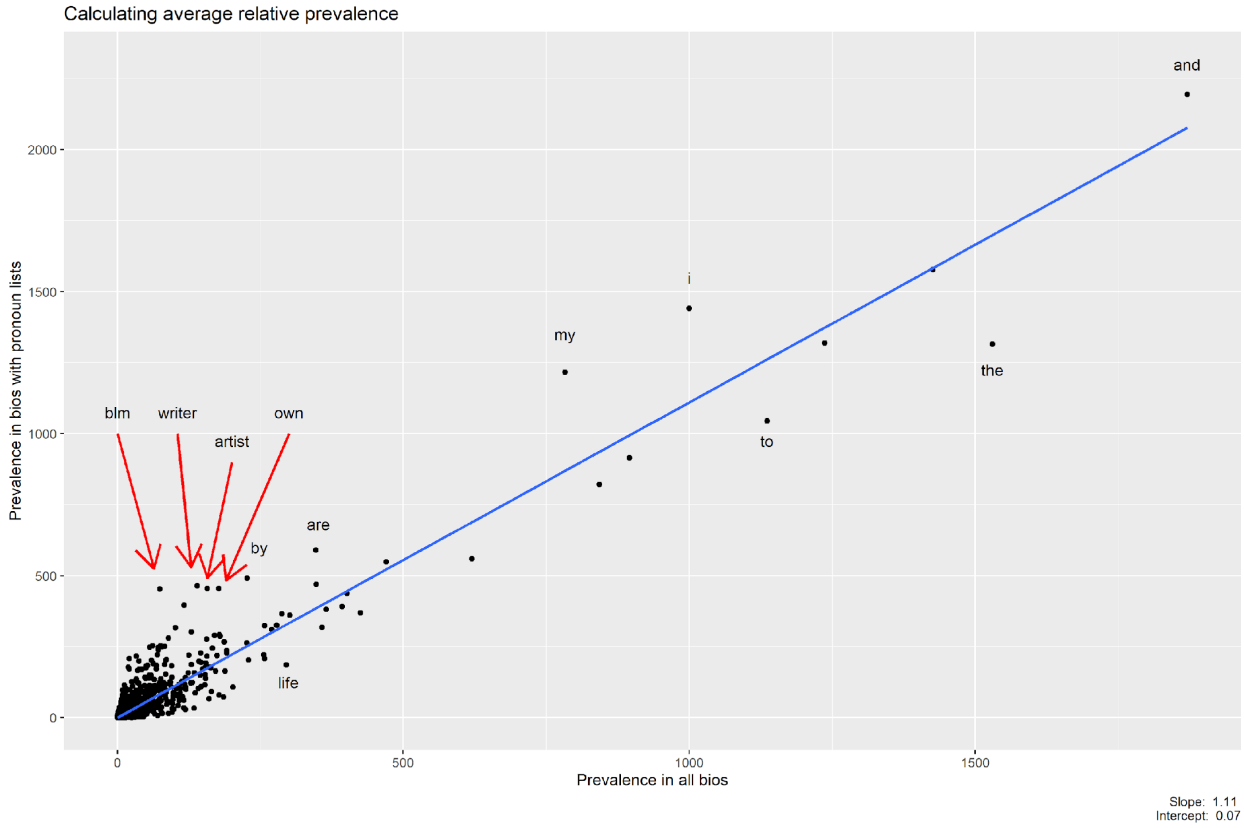

Figure 7: Prevalence of tokens in bios with a pronoun list vs. Prevalence of tokens in all bios

Appendix B: Regular Expressions

The first regular expression we considered using when tokenizing Twitter bios was "\b|\s+", which splits Twitter bios at whitespace or word boundaries. This expression has

been used previously (Jones, 2021). In most cases, this expression splits a bio into words as desired. This expression fails for pronoun lists; “she/her” is split into three different tokens by that expression: “she,” “/” and “her”.

We used the regular expression “[^a-zA-Z0-9/'`’-]” for this work. This expression matches for every character that is not alphanumeric, a forward slash, an apostrophe or character often used as an apostrophe, or a dash.

In many cases, one wants to split on any punctuation, including a slash. Consider as an example a token containing a slash: “singer/songwriter.” This ought to be split into two different meaningful tokens: “singer” and “songwriter”. However, we find that pronoun lists most often appear as one word without spaces (i.e. “she/her”) as opposed to three words with spaces (“she / her”). Thus we explicitly desire to not split on the slash character. As a result, using the regular expression from this work on the phrase “singer/songwriter” results in a single token “singer/songwriter”. Incidentally, “singer/songwriter” had a prevalence of 3.57 in 2022; this was the highest prevalence of any phrase consisting of two terms and a forward slash that is not a pronoun list. Thus, the decision not to split phrases at forward slashes does not have a major impact on results.

We also avoid splitting phrases at dashes and apostrophes because words that contain these characters (i.e. “can’t”, “non-binary”) are generally more meaningful when considered one token than when considered multiple tokens. We do tokenize phrases at any other punctuation mark. Therefore, “Singer,”, “singer.”, and “singer” would each be considered the same token: “singer”.

We did also observe users including pronoun lists within text fields other than the bio. Specifically, some users include pronoun lists in the “name” and “location” fields, but the prevalence was 15x larger in the biography than either the location or name field.

Appendix C: Code and Data

The data and code necessary to replicate this work can be found at https://osf.io/pjgr7/ .

Appendix D: Glossary

Figure 2 and Figure 3 (and their equivalents in Appendix E) contain a number of tokens, bigrams, and trigrams whose meanings aren't necessarily obvious. We created a glossary to define (and provide an example bio) for the possibly unclear terms in Figure 2, Figure 3, Figure 11, and Figure 12. The glossary can be accessed at https://osf.io/snz3v .

Appendix E: Additional Figures

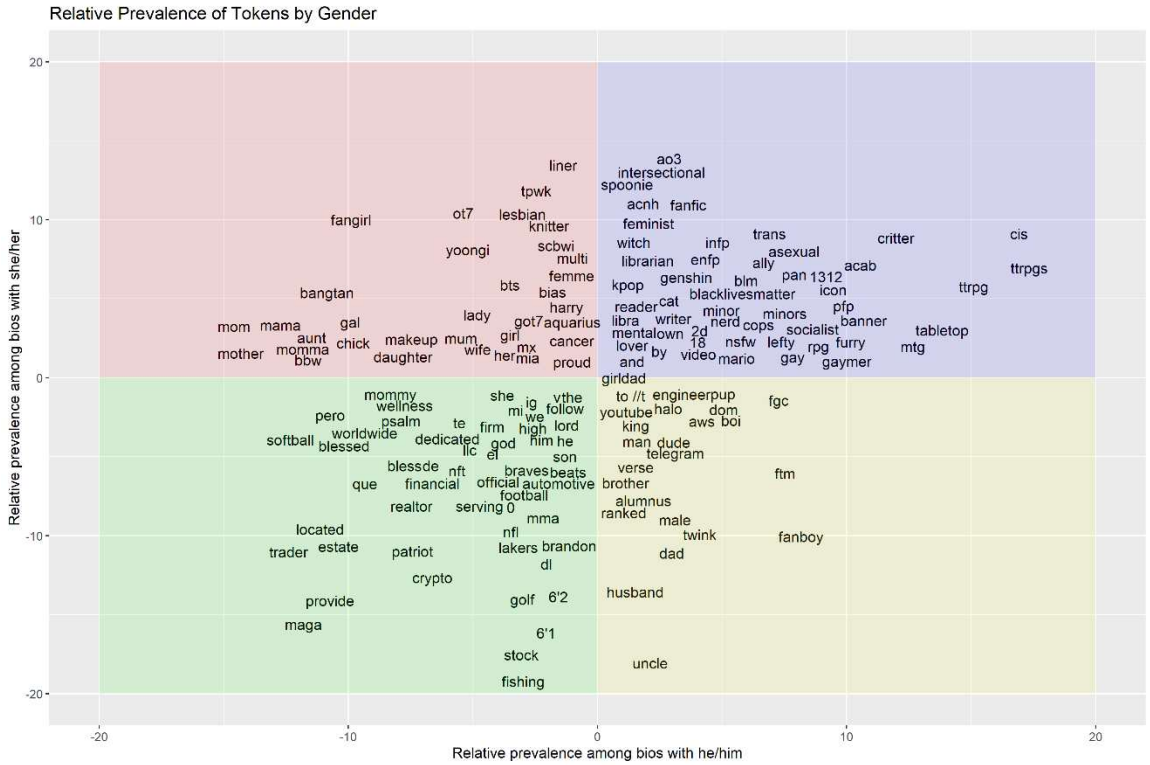

Figure 8: Relative Prevalence of Tokens by Pronoun List

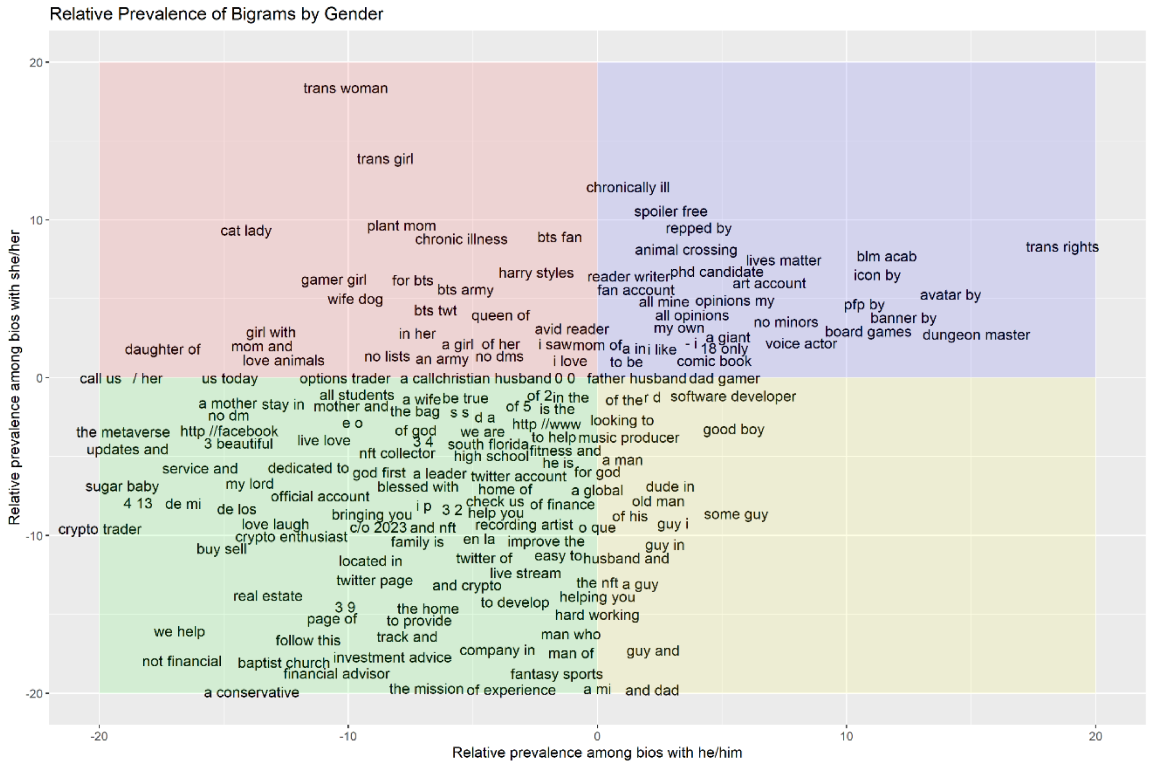

Figure 9: Relative Prevalence of Bigrams by Pronoun List

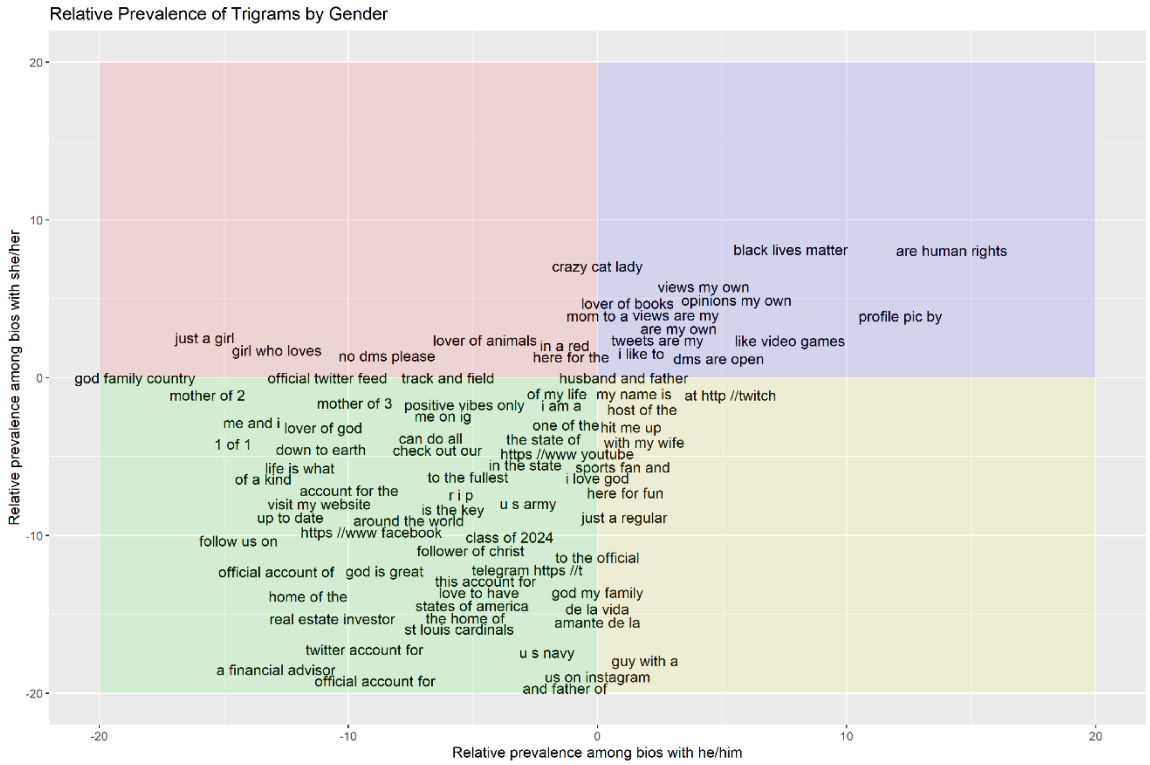

Figure 10: Relative Prevalence of Trigrams by Pronoun List

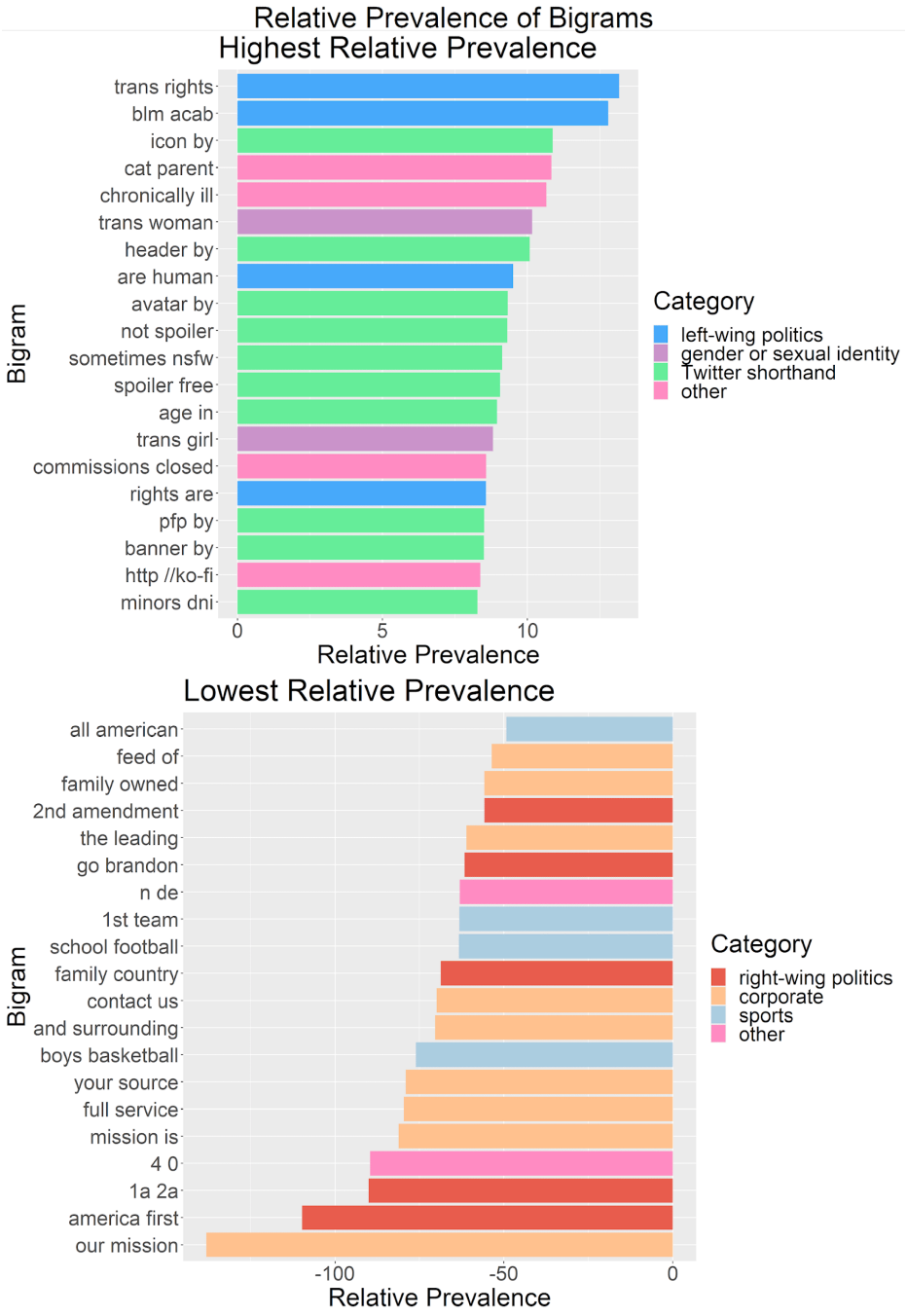

Figure 11: Bigrams with Highest and Lowest Relative Prevalence

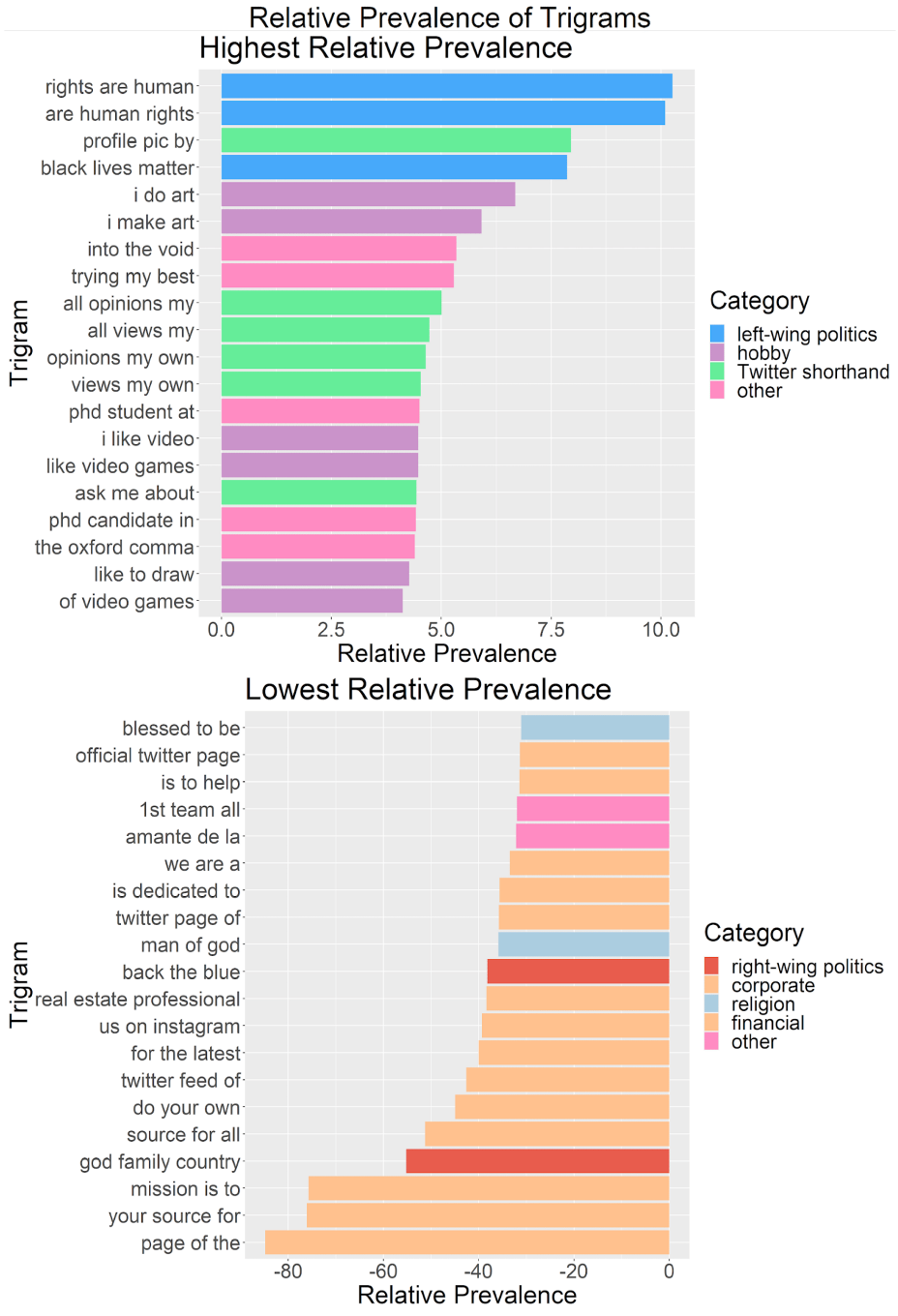

Figure 12: Trigrams with Highest and Lowest Relative Prevalence