Journal of Quantitative Description: Digital Media 1(2021), 1–n DOI

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 the US 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. 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

Liam Tucker: letucker@crimson.ua.edu
Jason J. Jones: jason.j.jones@stonybrook.edu
Date submitted: 2022-10-29

[1] We would like to thank Margot Hare for her comments on early drafts of this work. This material is based upon work supported by the National Science Foundation under grants IIS-1927227, OAC-1950052 and CCF-2208663.

above-average proportion of the followers and friends of Twitter users with 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, pronoun list usage has greatly increased. One can find preferred pronouns being 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 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 has been 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 age 13-17 identified as transgender (Herman et al, 2022). A 2017 report by the same group estimated that only 0.7% of Americans age 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. (We did also observe users including pronoun lists within the "name" and "location" profile fields, but the prevalence was 15x larger in the biography than either the location or name field.)

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.

We constructed cross-sectional datasets at two different temporal resolutions: annual and daily. To develop the 2021 cross-sectional dataset, we used the Twitter API to observe 1% of all tweets posted between January 1, 2021 and December 31, 2021. We discarded tweets whose location field did not reference the US. From each tweet, we observed the user's biography at the time of posting. If multiple tweets in our sample were from the same user, we selected one at random to keep and discarded the rest. The 2021 cross-sectional dataset consists of the Twitter bios associated with the tweets that were not discarded. Thus, a Twitter user can only have a Twitter bio appear in the 2021 cross-sectional dataset either zero times or once.

An annual cross-sectional dataset was generated for every year between 2015 and 2021 using the above process. We also generated a 2022 cross-sectional dataset using the same methodology but included only the dates available at the time (January 1, 2022

through June 30, 2022). This 2022 dataset contains 5,444,623 unique US users' Twitter bios.

Daily cross-sectional datasets were generated similarly, using only tweets from a single day. Daily cross-sectional datasets were compiled for nearly every day between February 16, 2015 and June 30, 2022[2].

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. Apart from these five and their corresponding pronoun lists, the most common pronoun list is a Portuguese-language pronoun list – "ela/dela" – with a prevalence of 2.1 occurrences per 10,000 Twitter bios. We considered including neopronouns in our analysis, but no neopronoun list occurred frequently enough to reliably surpass a prevalence of 1 per 10,000 criterion.

Table 1: 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

As is shown in Table 1, the five pronoun lists we considered were each far more prevalent than the corresponding pronoun list with the next largest prevalence[3]. Other combinations and permutations of lists might be considered, but for simplicity we will

[2] Days missing from this dataset are June 27, 2015; June 28, 2015; and January 9, 2016.
[3] For each year from 2015-2021, the five chosen pronoun lists were by far the most commonly used.

examine only these five. 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 4.26.

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, or a backtick (often used as an apostrophe mark). 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 the prevalence of tokens rather than the proportion for convenience. It is easier to understand that "student" has a prevalence of 84 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. A positive relative prevalence means the token is more common in bios with a pronoun list; a negative relative prevalence means the token is more common in all bios.

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

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

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.

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.

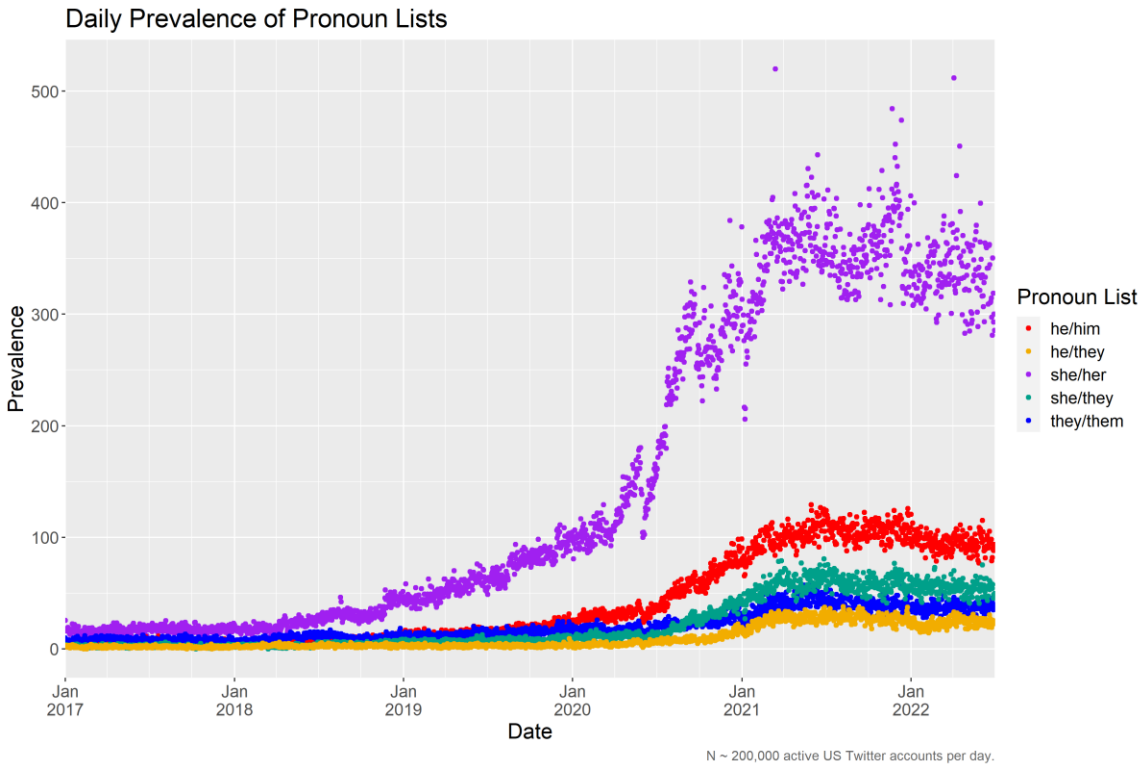

Figure 1: Daily Prevalence of Pronoun Lists

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 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 9.62 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 Tokens

Here we investigated the extent to which other tokens co-occured with pronoun lists. In Figure 2, we present tokens with extreme high and low relative prevalence values. It should be noted that tokens were only included if their prevalence within all Twitter bios was at least 5.0 and their prevalence within pronoun list bios was also at least 5.0.

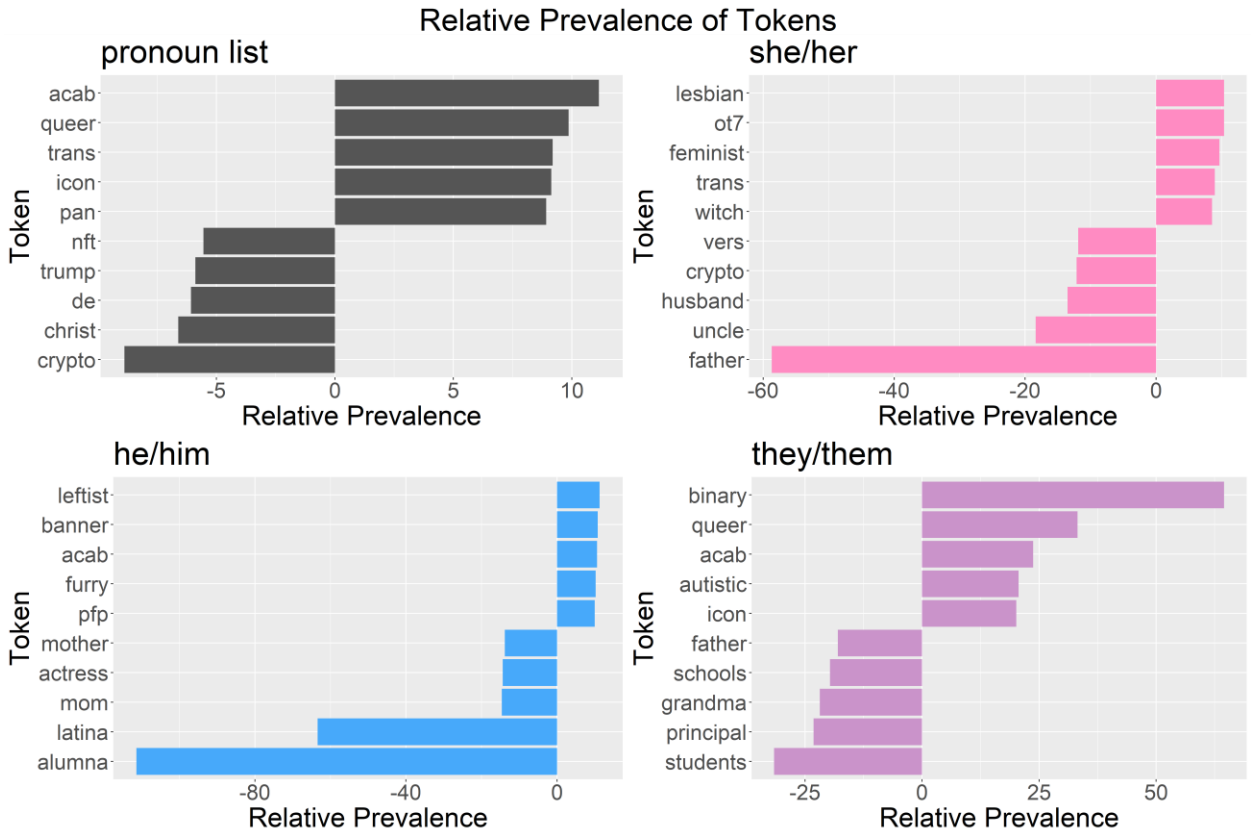

Figure 2: Tokens with Highest and Lowest Relative Prevalence Among Various Pronoun Lists

Many of the tokens in Figure 2 are unsurprising. Likely-female terms such as “alumna”, “Latina”, “mom”, “actress”, and “mother” display low relative prevalence

among Twitter bios containing “he/him.” Similarly, “father”, “uncle”, and “husband” display low relative prevalence among Twitter bios containing “she/her”. These results are 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 …”). Other tokens prompted further inquiry. In Appendix C, we explain that “binary” has the highest relative prevalence of any token among bios with “they/them” because it is used in expressions that indicate than an individual is *non*binary.

Consider that “acab” appears in Figure 2 for both he/him and they/them. An acronym for “all cops are bastards”, “acab” spiked in popularity during the summer of 2020, during protests in the aftermath of the murder of George Floyd. This term and others that do not seem directly related to gender identity prompted us to explore more deeply relative prevalence with pronoun lists. To do so, we ranked tokens by relative prevalence and explored those with the most extreme values. Specifically, we looked at the tokens with the twenty highest and twenty lowest relative prevalence values among bios with a pronoun list and manually categorized them. See Figure 3.

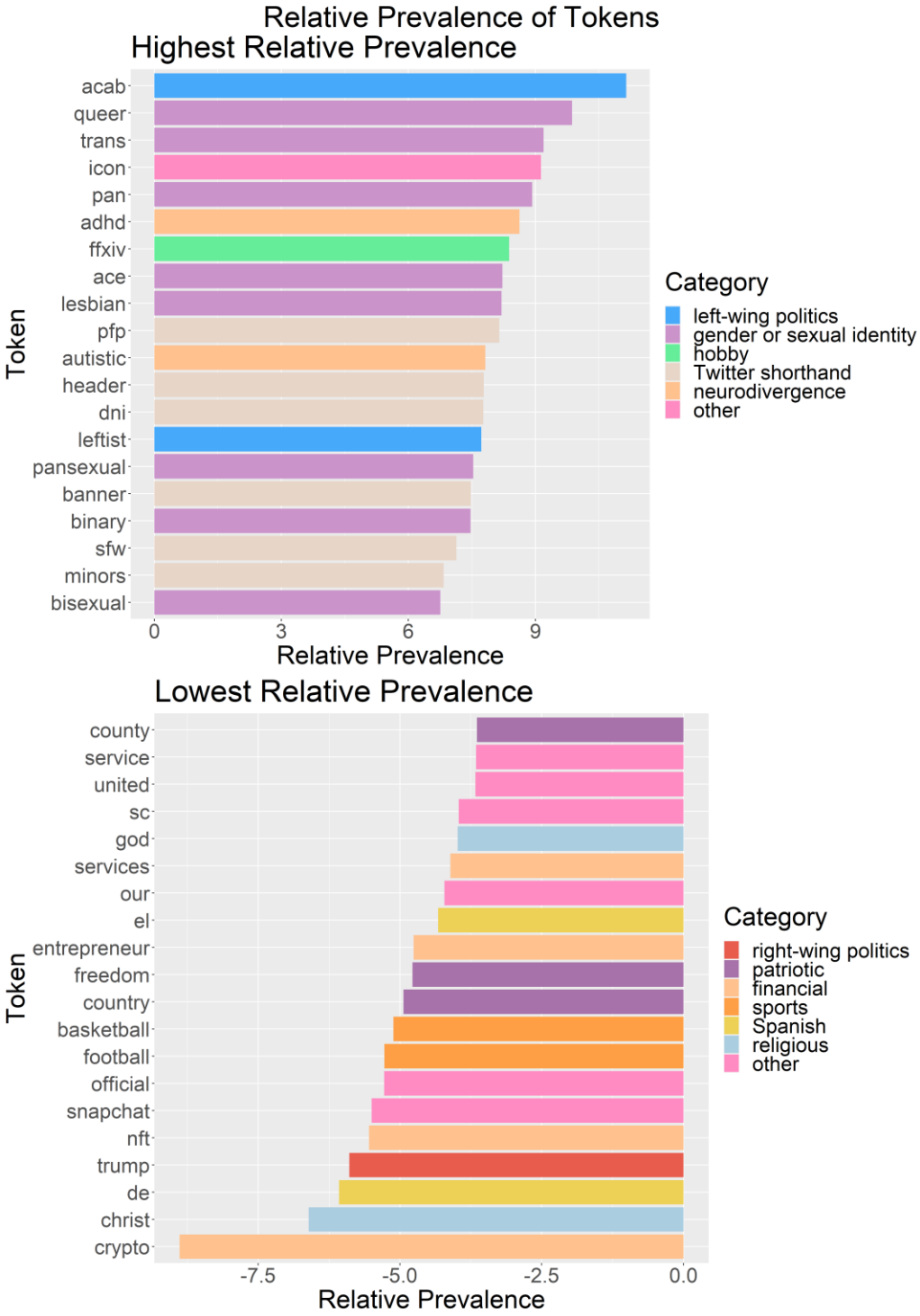
Relative Prevalence of Tokens
Highest Relative Prevalence
acab
queer
trans
icon
pan
adhd
ffxiv
ace
lesbian
pfp
autistic
header
dni
leftist
pansexual
banner
binary
sfw
minors
bisexual
Token
0
3
6
9
Relative Prevalence
Category
left-wing politics
gender or sexual identity
hobby
Twitter shorthand
neurodivergence
other
Lowest Relative Prevalence
county
service
united
sc
god
services
our
el
entrepreneur
freedom
country
basketball
football
official
snapchat
nft
trump
de
christ
crypto
Token
-7.5
-5.0
-2.5
0.0
Relative Prevalence
Category
right-wing politics
patriotic
financial
sports
Spanish
religious
other

Figure 3: Tokens with Highest and Lowest Relative Prevalences

A few clear categories emerge. 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.

Pronoun List Clustering

We next explored whether users with a pronoun list in their bio were 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 4 and 5.

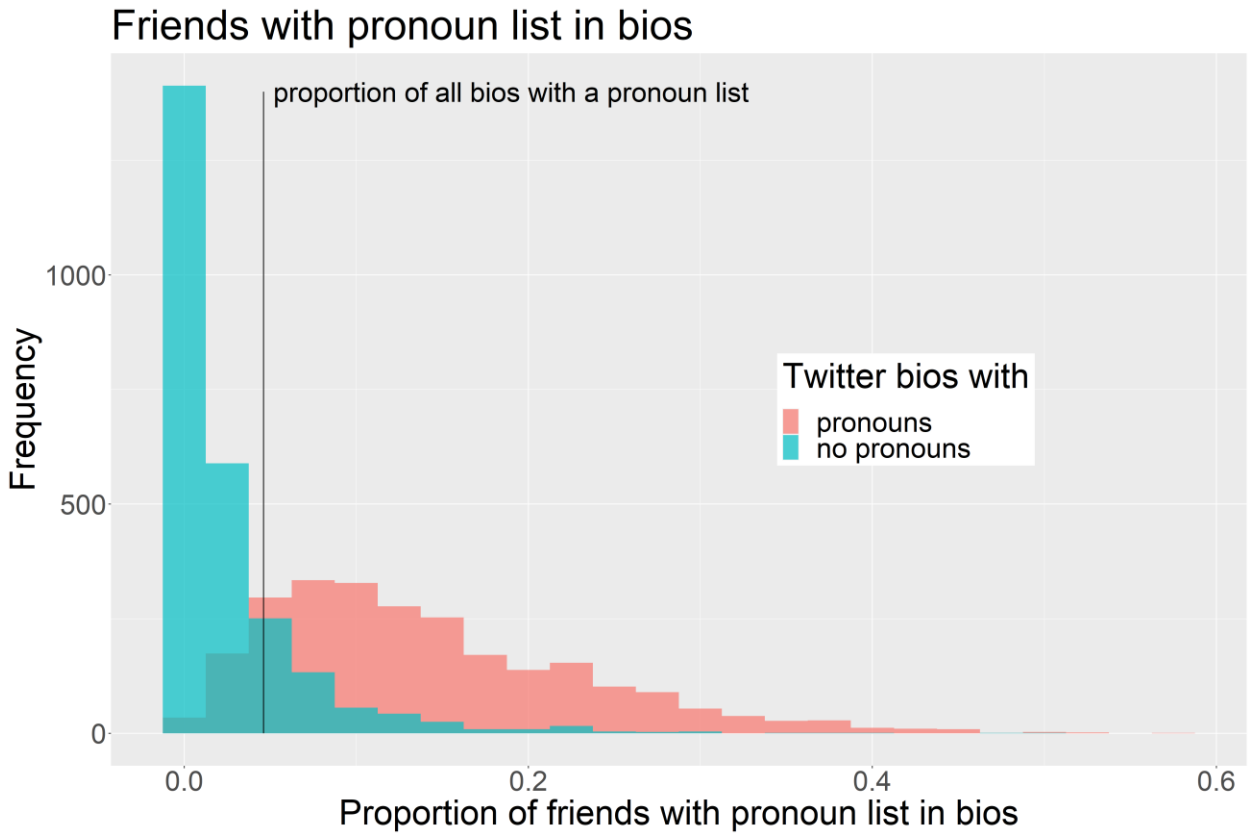

Figure 4: Distribution of Friends of Random Users with Pronouns in Bios

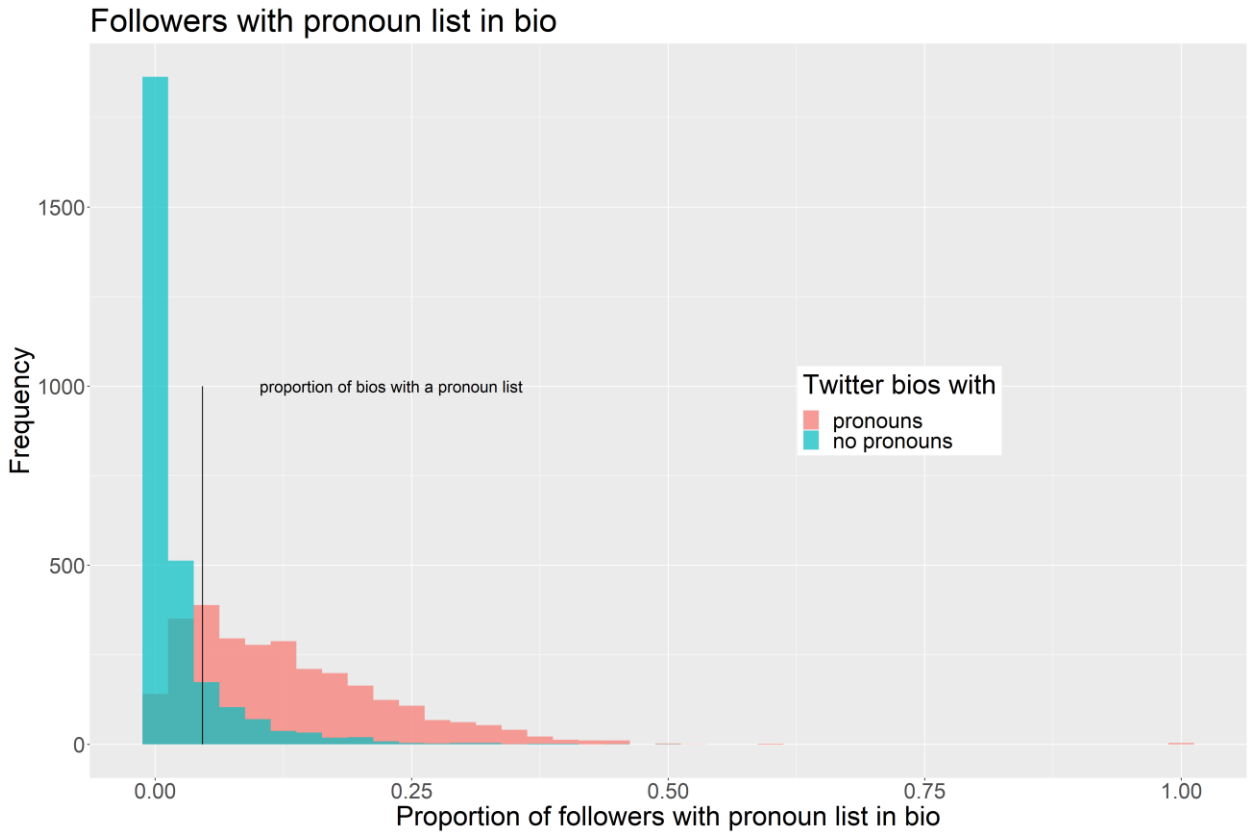

Figure 5: Distribution of Friends of Random Users with Pronouns in Bios

During the first six months of 2022, we found that 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 had 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, 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 have seen 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.

Users with a pronoun list in their Twitter bio were more likely to also include tokens related to left-wing politics and gender or sexual identity and less likely to also include tokens 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 were more likely to follow, and be followed by, other Twitter users with a pronoun list in their bio. Users who did not similarly clustered with others who did 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. Further research might be able to recover the weights of each factor on US Twitter users' decisions to include or not include pronoun lists and to form or not form ties with others who do. 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). 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. If Twitter users disproportionately follow users with similar interests, and interests systematically coincide with pronoun list usage, then even if Theory 1 were false, one would still expect to see clustering among users with a pronoun list in their bio.

One interesting comparison to make would be between pronouns lists and political spectrum signifiers in the bio. Here we have shown two left-wing signifiers (i.e. leftist and acab) were more likely to occur in the bios of Twitter users with pronoun lists than a user 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 argue 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. It is difficult (and in some cases unethical) to experimentally intervene to create new social norms in the field. The Twitter bio data, though it is observational, is rich in all the ways one might hope if one's goal were to separately attribute the spread of a behavior to homophily, social contagion and spontaneous adoption as mechanisms. The data is publicly available for millions of users, changes frequently without being ephemeral, comes timestamped and geocodable, and describes something important: how individuals see and present themselves. Many new insights await those who use Twitter bio data to study shifting trends in personally expressed identity.

Studies of this sort 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. As one instance, Stets et al. (2020) posit 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 sequence 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.

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.

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/#!.

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

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/

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

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

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 in expectation. 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 have nothing listed in their bio. (In the 2022 dataset, 13.1% of bios were blank). Because empty bios cannot contain pronoun lists, we speculated that the average token will 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.

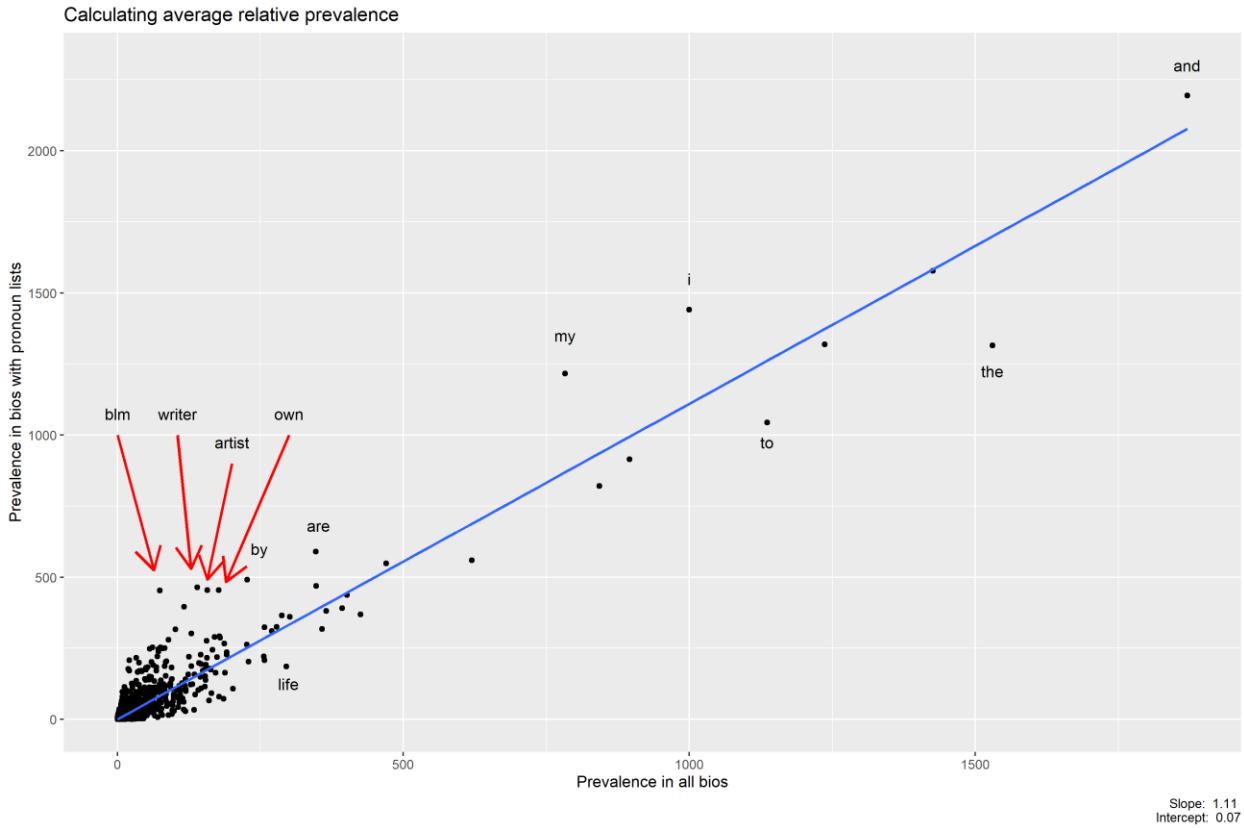

Figure 6: Prevalence of tokens in bios with a pronoun list vs. Prevalence of tokens in all bios

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 two 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, or an apostrophe or character often used as an apostrophe. In many cases, one wants to split on any punctuation, including a slash. In the current case, however, we explicitly desired to not split on the slash character. Consider as an example a token containing a slash: "singer/songwriter." Splitting on space or word boundaries would deliver three tokens: singer, / and songwriter. In many cases that would be desired behavior. Using the regular expression from this work, however, "singer/songwriter" would be the resulting token. However, each of "singer-songwriter," "singer,", and "Singer," would be counted as an instance of "singer." 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 was not a pronoun list.

Appendix C: Relative Prevalence of "binary"

In Figure 2, which displayed tokens with the highest and lowest relative prevalences among bios containing various pronoun lists, "binary" was the term with the highest relative frequency among bios containing "they/them". We found it likely that "binary" was most often used to signal being "non-binary".

To test this idea, we counted the total number of times “binary” appeared, in any form, in the Twitter bios. Note that, for the analyses in this Appendix, we did not tokeniz the Twitter bios. That means “binary”, “non-binary”, “BinaryX”, and “#binaryoptions” would each be counted as an instance of “binary” appearing in a bio. Separately, we counted the number of bios that contained, in some form, “non-binary”, “nonbinary”, “non binary”. If instances of “binary” actually refer to non-binary, the number of bios containing one of “non-binary”, “nonbinary”, or “non binary” should be nearly as large as the number of bios containing “binary”.

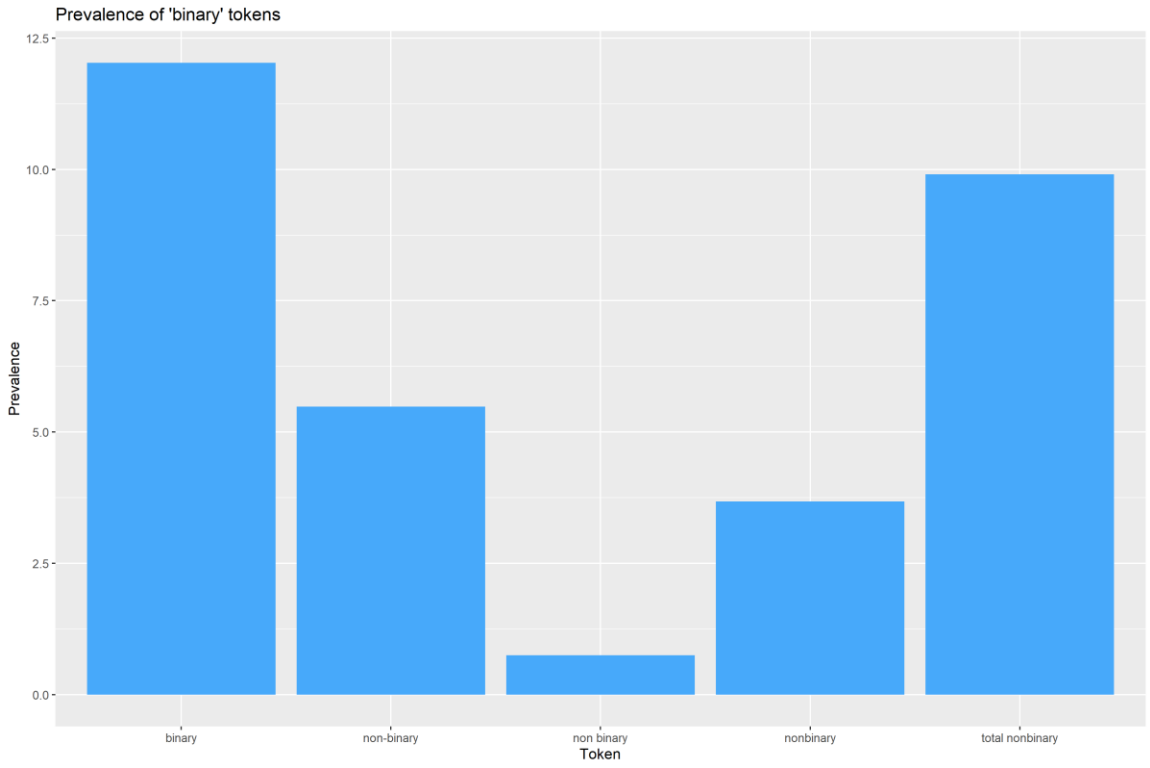

Figure 7: Frequency of Various Tokens Containing “binary”

Figure 7 displays the prevalence of various tokens. “non-binary” is the most common way for someone to signal that they are non-binary, followed by “nonbinary”

and “non binary”. Collectively, these tokens that signify “non-binary” comprise a majority of the appearances of “binary” in Twitter bios.

As is shown in this chart, “non-binary” is the most frequent way to indicate being non-binary, followed by “nonbinary” and “non binary”. During our main analysis, our regex expression split these words such that “nonbinary” is counted as an appearance of the word nonbinary while “non-binary” and “non binary” are each counted as an appearance of the word binary. Each time that “non-binary”, “non binary”, or “nonbinary” is detected in a bio and included in this chart, “binary” is also detected and included. The prevalence of “binary” in this analysis is 12.03, while 9.91 per 10,000 Twitter bios contain at least one of “non-binary”, “nonbinary”, or “non binary”. Thus, 82.4% of the time the word “binary” appears in a Twitter bio, it signifies “non-binary”.

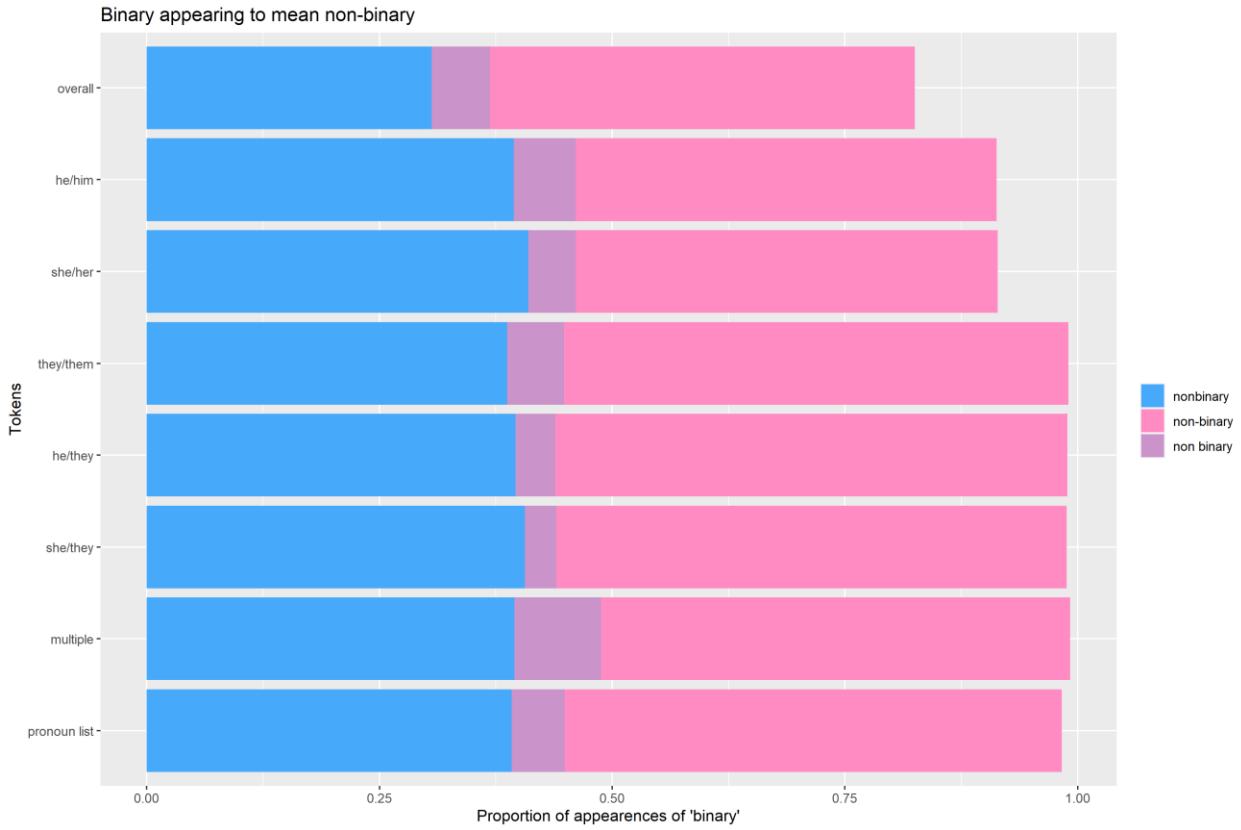

Figure 8: Appearances of "Binary" by Token

Figure 8 shows the proportion of appearances of "binary" in twitter bios that are used to express nonbinary. The total length of a bar is the proportion of the time that the appearance of "binary" in a bio is part of the phrase "nonbinary", "non binary", or "non-binary". The total length of the "overall" bar is .824, meaning 82.4% of the times that "binary" appears in a bio, it is used to signal being non-binary. The length of the "they/them" bar is .990.

We were more specifically curious how "binary" was used by Twitter users with "they/them" pronouns in their Twitter bio. To explore this, we plotted the proportion of

appearances of “binary” in Twitter bios by pronoun list status. We found that, while 82.4% of instances of “binary” indicate being non-binary in all Twitter bios, among Twitter bios with they/them pronouns, that number is 99.0%. (The confidence interval is [98.3%, 99.4%]). This strongly supports the idea that the reason why “binary” has the highest relative prevalence of all tokens among bios containing “they/them” is to signify identifying as non-binary.

Appendix D: Code and Data

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