SUPPLAMENTAL MATERIAL: VISUALIZING PREFERRED PRONOUNS AND IDENTITY

DEFINITIONS

Term	**Definition**
Token	A keyword that appears in a Twitter bio
Prevalence	The number of bios per 10,000 that a token appears in
Relative Prevalence	The prevalence of a token in bios that contain a specific pronoun list divided by the prevalence of the token in all bios
Pronoun list	Five pronoun lists were considered: “he/him”, “she/her”, “they/them”, “he/they”, and “she/they”. We considered various neopronouns, such as “xe/xem”, “per/per”, or “ve/ver”. Each neopronoun list had a prevalence of less than 1, so we chose to exclude them from our analysis.
Regex Expression	"[^a-zA-Z0-9/'`’]” is the regex expression used for tokenization. This splits up the bio at any character that isn’t a number, a letter, the forward slash, or an apostrophe. This means a bio that contains the phrase “he / him” is counted as three separate tokens: “he”, “/”, and “him”. The phrase “he/him” would be counted as one token, while “football/basketball” would also be counted as one token. For this project, this regex expression best balanced simplicity with accuracy (i.e. splitting a bio into tokens the way a human would).
Twitter bio	The Twitter field in which users can enter any information they want publicly associated with their account. When creating a Twitter account, a user is prompted to: “Describe yourself. What makes you special? Don't think too hard, just have fun with it.” A user can also edit their Twitter bio at any time, though the average Twitter user in a similar dataset updates their bio only once per year (Rogers 2021).

Table 1: Definitions of tokens that appear in the main portion of this article

SUPPLAMENTAL MATERIAL: VISUALIZING PREFERRED PRONOUNS AND IDENTITY

TOKEN DEFINITIONS

Token	**Definition**
acab	“all cops are bastards”, a political catchphrase associated with left-wing politics
pan	“short for pansexual”, a sexual orientation
ffxiv	“Final Fantasy XIV”, an online multiplayer role-playing game
ace	“asexual”
pfp	“profile pic”
dni	“do not interact”
sfw	“safe for work
nft	“non-fungible token”
de	a Spanish preposition
ot7	“one true 7”, a slang term most commonly used to describe K-pop groups
vers	a term most often used to describe the sexual preference of certain gay men

Table 2: Definitions of tokens appearing in Figure 1 in the main section of this article

CONFIDENCE INTERVALS

Because of the large number of Twitter bios used in this dataset, the confidence intervals are generally very small. The most common token, “and”, has a prevalence of 1871.1, with a 95% confidence interval of [1867.8, 1874.3]. Among only bios with a pronoun list in them, “and” has a prevalence of 2194.5, with a confidence interval of [2178.4, 2210.8]. The token with the highest relative frequency among bios with a pronoun list, “acab”, has a relative frequency of 11.1, with a 95% confidence interval of [10.6, 11.7]. The probability that a bio contains a pronoun list, given that it contains the token “acab”, is 51.5%, with a 95% confidence interval of [50.1%, 52.9%].

GATHERING THE DATASET

Dr. Jason J Jones compiled a dataset of Twitter bios of active users in the USA. One percent of tweets are gathered using the Twitter API. In his own words:

SUPPLAMENTAL MATERIAL: VISUALIZING PREFERRED PRONOUNS AND IDENTITY

The Twitter Streaming API was used to observe a random sample of 1% of all tweets posted between January 1, 2022 and June 30, 2022. Any tweets whose profile location field did not reference the US were removed from the dataset. If one user had multiple tweets observed during this time period, we randomly selected one of their tweets to use and discarded the rest. Thus, at most one Twitter bio per user appears as part of this dataset.

BLANK BIOS

It should be noted that, in any given year, roughly 15% of active Twitter users in the US do not anything listed in their bios. (In this 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. Indeed, for every point increase in prevalence of a token, we can expect the prevalence of the token to increase by **1.10** points among bios that contain a pronoun list.

SUPPLAMENTAL MATERIAL: VISUALIZING PREFERRED PRONOUNS AND IDENTITY

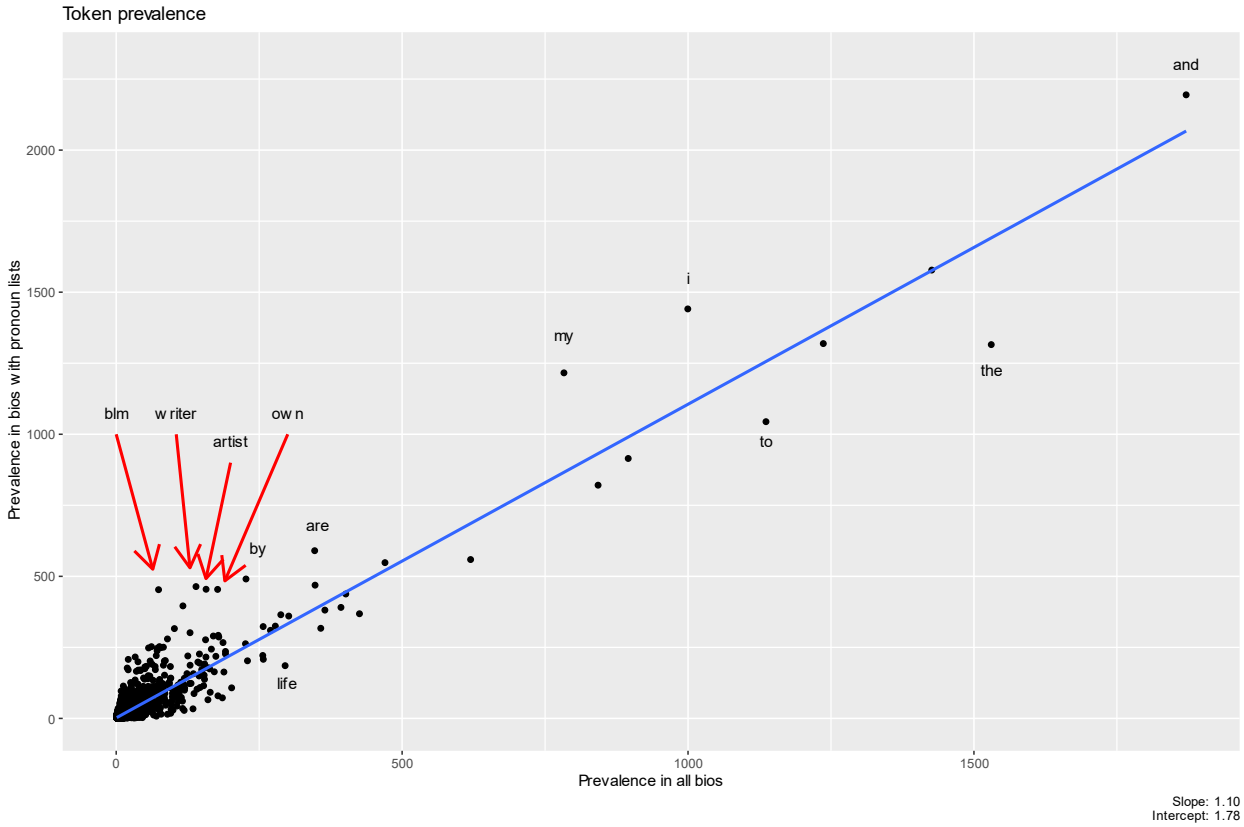

Figure 1: The majority of tokens have a prevalence of less than 50, both among all bios and among all bios that contain a pronoun list. Notable outliers are labeled, and the best fit line is plotted. The slope of the best fit line is 1.10, which is in line with our expectations.

BINARY

The portion of Figure 1.B looking at the relative frequency of tokens in bios containing the token “they/them” shows that “binary” is the term with the highest relative frequency. We found it likely that “binary” was most often used to signal being “non-binary”.

To test this theory, we counted the total number of times “binary” appears, in any form, in Twitter bios. This means “binary”, “non-binary”, “BinaryX”, and “#binaryoptions” would each be counted as an instance of “binary” appearing. We also checked the number of bios that contain, in some form, “non-binary”, “nonbinary”, “non binary”, and the distinct number of bios

that contain one of those three tokens. If our theory is correct, 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”.

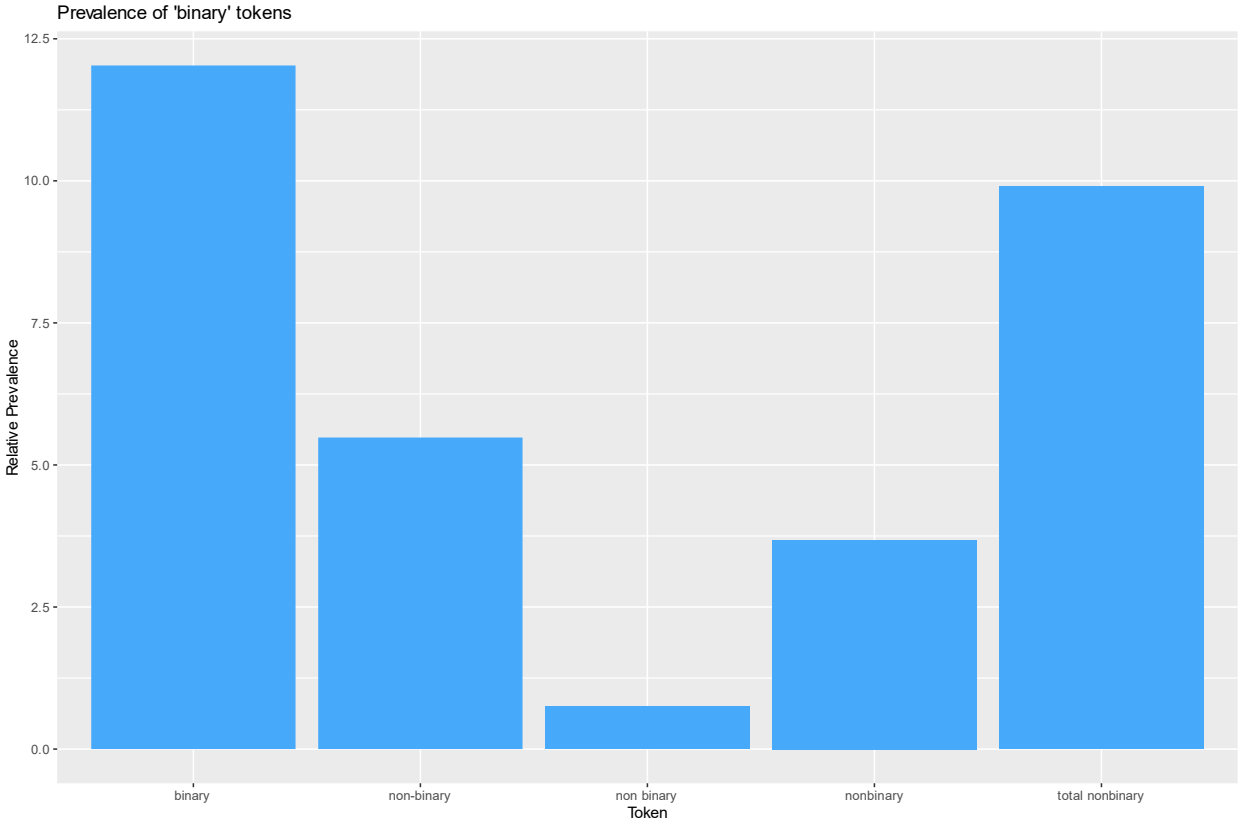

Figure 2: The prevalences of various tokens in all bios are displayed. “non-binary” is the most common way for someone to signal that they are non-binary, followed by “nonbinary” and “non binary”. Collectively, these tokens that signify “non-binary” comprise a majority of the appearances of “binary” in Twitter bios.

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” 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".

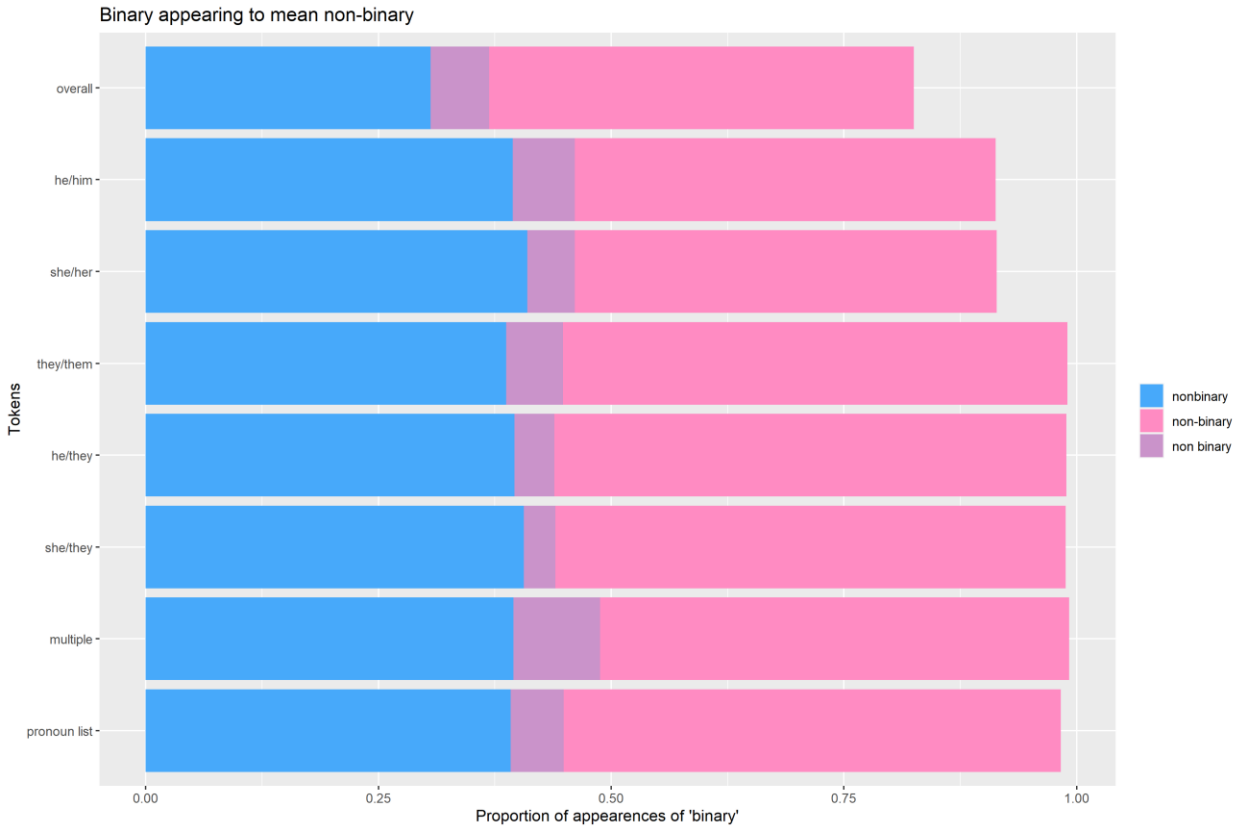

Figure 3: This chart shows the proportion of appearances of "binary" in twitter bios that are used to express "self-identity". The total length of a bar is the proportion of the time that the appearance of "binary" in a bio is part of the phrase "nonbinary", "non binary", or "non-binary". The total length 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.

As stated above, among all Twitter bios, "binary" really means non-binary 82.4% of the time. Among Twitter bios with they/them pronouns, this number is 99.0%. (The confidence interval is [98.3%, 99.4%]. This strongly supports our theory 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.

CRYPTO

The portion of Figure B looking at the relative frequency of tokens in bios with any pronoun list stands out as particularly interesting. The five tokens with the lowest relative frequency are “crypto”, “christ”, “de”, “trump”, and “nft”. Some of these make intuitive sense; people who are more religious or more conservative tend to be less likely to support displaying preferred pronouns. “de” is a Spanish preposition, and presumably most bios that contain prepositions in Spanish are mostly or entirely in Spanish and thus unlikely to contain an English pronoun list.

When we expand our scope to the 20 tokens with the lowest relative prevalence, many tokens are fall into one of five categories: common Spanish words (“por”, “que”, etc.); words associated with right-wing politics (“trump”, “patriot”, “united”); religious words (“christ”, “god”); sports words (“football”, “basketball”); and financial terms (“bitcoin”, “investor”, “trading”). When we consider any tokens with a prevalence greater than 1 in both bios with pronoun lists and all bios, 12 of the 20 most common tokens are financial.

SUPPLAMENTAL MATERIAL: VISUALIZING PREFERRED PRONOUNS AND IDENTITY

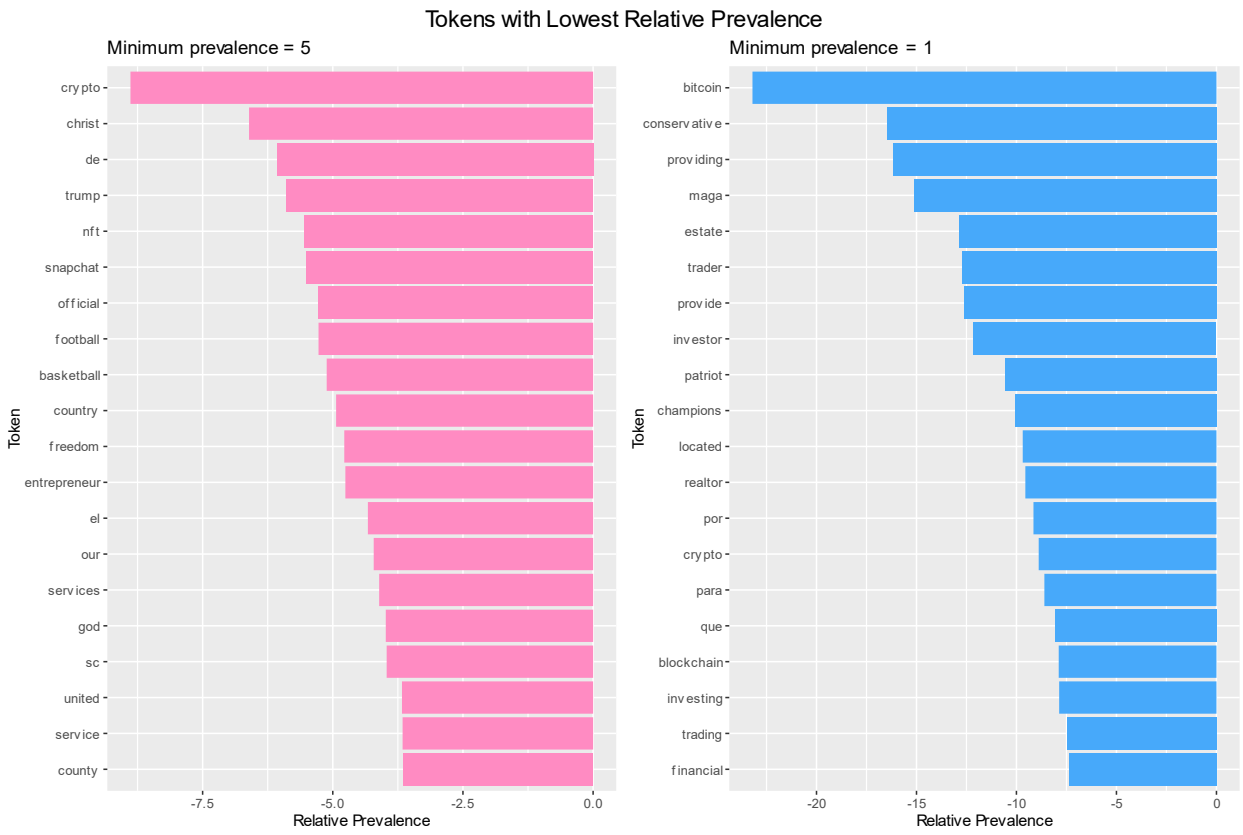

Figure 4: The 20 tokens with the lowest relative prevalence among bios with a pronoun list are plotted here. Most of the terms are either (1) associated with right-wing politics, (2) religious terms, (3) sports terms, (4) in Spanish, or (5) related to finance or cryptocurrency.

We considered two basic theories to explain why the relative prevalence of financial terms, and especially cryptocurrency terms, are so low:

1. Financial or cryptocurrency tokens are disproportionately used in Twitter bios in ways other than for self-identification. That could be because of a disproportionately large number of bots containing financial tokens or accounts dedicated only to financial purposes.

2. Financial or cryptocurrency tokens are not disproportionately used in Twitter bios in ways other than for self-identification. This would imply that people who consider crypto a part of their identity are less likely to consider pronouns a part of their identity, and vice versa.

To test this, we created a dataset similar to the one used in the main article. This time, instead of tracking how often tokens appear alongside various pronoun lists, we tested how often tokens appeared alongside various tokens related to cryptocurrency: “crypto”, “nft”, “bitcoin”, “blockchain”, “trader”, “entrepreneur”, “investor”.

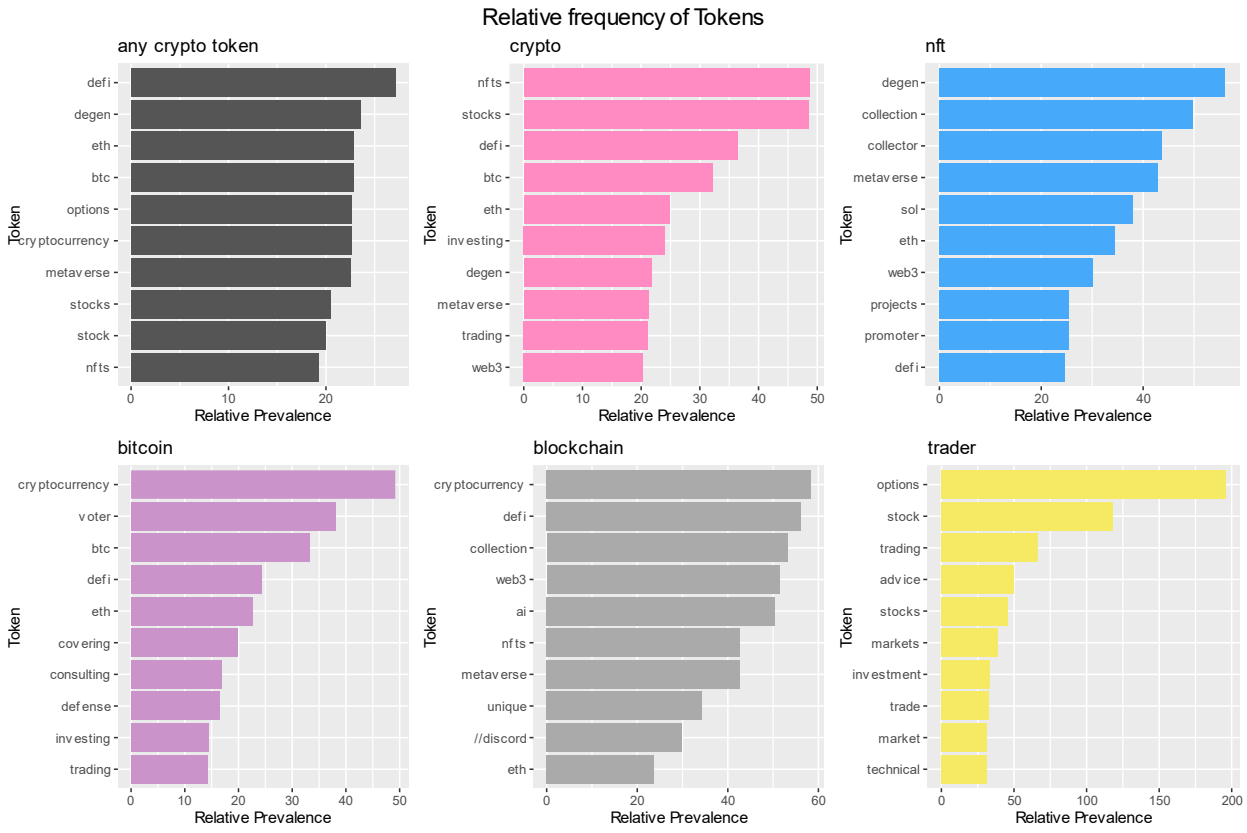

Figure 5: The tokens with the largest relative prevalences, among bios that contain a selected financial token, are displayed here. In each graph, many of the tokens are related to finance or

cryptocurrency. Of the tokens not related to finance or crypto currency, many are related to technology (“web3”, “//discord”, “defense”, “metaverse”).

As is apparent on the graph, the tokens with the highest relative prevalences are nearly all related to finance. In fact, cannabis, the non-finance token with the highest relative prevalence among bios containing any crypto/finance related term, has a smaller relative prevalence than 44 other tokens. This data could support either theory. It is clear that Twitter bios that contain crypto-related tokens are extremely likely to contain other crypto-related tokens. This could be because the accounts are professional or scam accounts, or that users who populate their bios with crypto-related tokens are simply less likely to include other types of tokens, including pronouns, in their bios.

DATA AND CODE

The data and code necessary to replicate these results are available at https://osf.io/pjgr7/.

REFERENCES

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