# Measuring Trends in Identity Transition Using Social Media Bios: A Methodology and Proof of Concept

Nick Rogers[1[0000-0003-0624-8413]] and Jason J. Jones[2[0000-0002-4140-0268]]

[1] University of Pittsburgh, Pittsburgh PA, USA
[2] Stony Brook University, Stony Brook NY, USA
nick.rogers@pitt.edu

**Abstract.** "Social Identity Theory" is one of the most influential ideas in social science, but identity is notoriously difficult to operationalize and observe. In a previous project, we introduced a method for studying identity at massive scale by longitudinally observing a close proxy: social media bios. Here, we discuss a new variation on the method, which focuses on identity *transitions*: the amendment of an identity following a particular event. We illustrate the method through a test case taken from millions of Twitter bios from 2015-2023: patterns in the addition and deletion of the acronym "MAGA" ("Make America Great Again"). We show that certain categories of bio words at one time point can predict the addition of "MAGA" at a later point. We also examine the sorts of words that tended to *replace* "MAGA" when users amended their bios following the insurrection of January 6th, 2021. Finally, we discuss potential future applications of the method, focusing on the topic of "stigma exits."

**Keywords:** Social Identity, Social Media, Political Identity, Identity Transition, Stigma Exit.

## 1 Introduction

The concept of *identity* is among the most important and ubiquitous ideas in modern social theory. An individual's identity may be thought of as how they define themselves in the context of the society in which they are embedded. According to the influential *social identity theory*, a person's identity is commonly shaped by the groups to which they belong, e.g. their religious affiliation, profession, or family roles. As such, a given individual's identity is multiple; each of us, as Whitman penned, contain multitudes.

Individuals tend to seek identities that reflect favorably on their character, and avoid identities that are discrediting. Sometimes, a given identity might initially be socially beneficial, but later become an unwanted burden. For example, a person might take pride in defining themselves as an employee of a particular company, but if that company should be revealed as a fraud or bad actor, the person might regret making that relationship so foundational to their sense of self. There is substantial research interest in the question of *how individuals transition away from stigmatized identities*. For example, scholars have examined how former violent political extremists amend their self-image when leaving their respective movements , and the ways that identity changes – and doesn't change – when formerly-obese individuals lose substantial weight.

The practical application of these insights are myriad. Society might, for example, seek ways to draw people out of dangerous religious cults by activating alternative identities, or stop people from joining in the first place by understanding the sorts of identities that are likely to morph into extremism.

Despite the urgency of the topic, it is notoriously difficult to study empirically the concept of social identity. Identity is internal, subjective, amorphous, fluid. Yet we think it is possible not only to operationalize it, but even examine it at massive scale. In a previous paper, we made a case that the concept of social identity is tidily analogous to the *social media bio*: a brief autobiographical statement that often takes the form of a series of social roles. For example, the bio for former President Barack Obama on the social media site X (née Twitter) reads: "Dad, husband, President, citizen."

## 1.1 *Ipseology*: The Empirical Study of Identity at Scale

By studying social media bios at the scale of millions, we have argued, we can measure societal trends in social identity. When people are adding descriptions to their bios, or taking descriptions away, they are in some sense amending their identities; they are announcing that they wish to be socially defined in a different way than they were before. When a mass of people are amending their identities in the same sort of way, then something sociological is happening.

This idea undergirds an approach we call *ipseology*: the study of identity using large datasets and computational social science methods. Ipseology stands in contrast to how social identity is normally observed: through small-sample, long-form qualitative interviews. While the traditional approach can be useful for studying the identities derived from niche, static subcultures, we think bigger questions require larger datasets and more powerful methods.

## 1.2 A Method for Studying Identity *Transition*, and a Test Case

This paper introduces a new iteration of ipseology: a method that can be used to study the contexts in which a person *amends* their sense of self by adding a new, beneficial identity or abandoning one that has become stigmatized. We explain this method by applying it to a particular context: the adoption of a "Make America Great Again" identity during the initial rise to power of American President Donald Trump, and the later abandonment of that identity in the period following the political insurrection of January 6th, 2021. We use millions of longitudinal Twitter/X bios to pose two research questions: (1) Which pre-existing identities most strongly predicted – and protected against – adoption of a "MAGA" identity?; and (2) Following January 6th, which new identities most frequently replaced an abandoned MAGA identity?

# 2 Methodology and Results

The original source dataset from which all the below were derived is a nearly-complete, contemporarily collected 1% random sample of all published tweets from 2012-2023. These were collected with the GET statuses/sample endpoint of the Twitter API [1]. Hundreds of millions of users' profiles were attached to the tweets they authored. Within the profile, the text of the bio field was our object of study. We filtered users to those whose profile location indicated a US location. In every downstream dataset, if a user was observed tweeting more than once in the temporal period, we chose exactly one observation at random to represent that user. (For more details on these methods, please consult [2], [3].)

For simplicity's sake, we define a "MAGA identity" as any bio that contains the token "maga". Of course, there are any number of *other* tokens that might announce an identity that is defined by support for Donald Trump's political movement. And conversely, not every bio that contains the token "maga" is intended to convey such support. (Imagine, e.g., a bio that states "Fuck MAGA!".). Future applications of our method may wish to define their identity of interest by a *collection* of tokens, rather than any one in particular. And perhaps the validity of those tokens can be assessed by manually auditing a small random subset of bios to determine, e.g., whether users seem to be using those tokens to convey the sorts of ideas that the researcher expects.

## 2.1 An Annual, Cross-Sectional Perspective

As an initial matter, we find that the prevalence of various MAGA signifiers rose sharply in bios from the years 2015-2020, and declined sharply in 2021-2022 (cf. [4]). Then certain of the signifiers began to rise again.

To determine this, we first created a cross-sectional set of US users' bios for each year. Each bio was then tokenized, and the set of observed tokens became the representation of each user. The cross-sectional nature of this technique means that every qualifying author had their bio included. This maximizes observation count and statistical power, but also means the set of profiles fluctuates each year. New users will show up that were unobserved in the past, and old users observed in the past will attrite away.

In Figure 1, we present the prevalence of users displaying four different MAGA signifier within each year. Note that each point represents a prevalence estimate based on millions of unique users in the denominator and thousands to tens-of-thousands in the numerator. When a point is missing from a series, it means the prevalence did not reach the criterion of 1.0 or greater prevalence. When a year is missing from the graph, that means no signifiers reached criterion in that year. One observes that the popularity of all signifiers increased from first detection (above threshold) to 2020. From these peaks, each signifier saw a decline. "Maga" and "trump" rebounded in 2023, while "qanon" and "wwg1wga" did not.

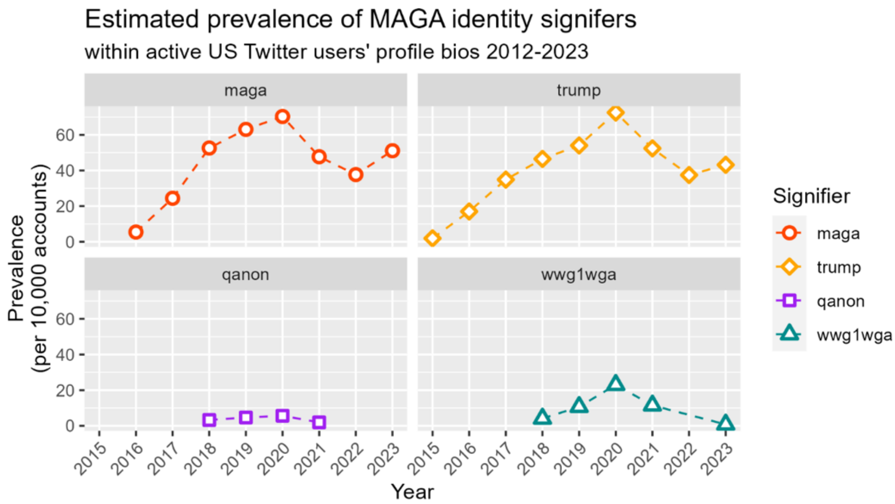


Figure 1. The plots depict how many users' bios contained each MAGA signifier per 10,000 unique US users observed tweeting each year.

## 2.2 A Daily, Cross-Sectional Perspective

What happened in 2021 to cause the sharp MAGA decline? Perhaps it was Trump's electoral defeat and departure from office. Or perhaps it was the widespread social condemnation of his efforts to remain in office by inciting the January 6th insurrection. Our annual data lacks the granularity to answer that questions, but we also created a cross-sectional set of tweeting US users' bios observed *each day*. Then, we estimated prevalence within these daily samples. In Figure 2, we zoom in on the period January 1, 2020 through December 31, 2021 and present the daily estimated prevalence for each relevant signifier.

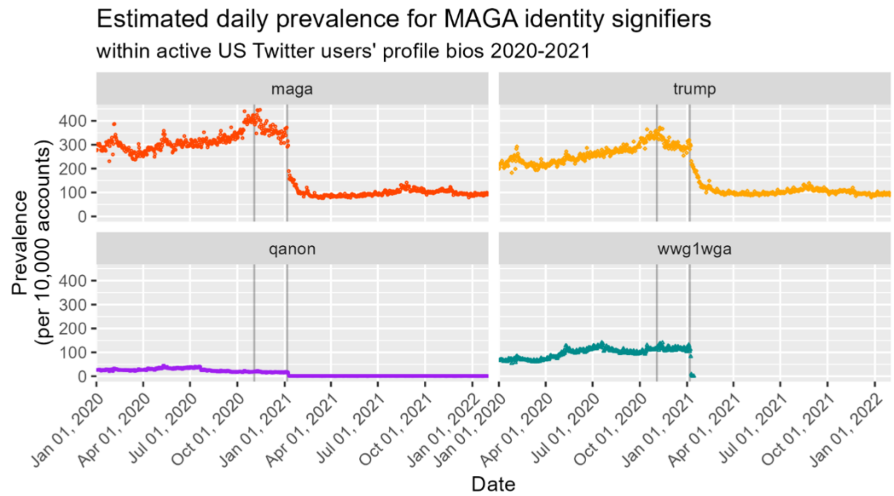


Figure 2. Daily prevalence estimates based on daily, cross-sectional samples. Vertical lines mark Election Day (2020-11-03) and the Attack on the Capitol (2021-01-06). Compare the y-axis scale to Figure 1. Until January 6th, 2021, users with MAGA identity bios were consistent, vociferous tweeters.

The 2020 peak and 2021 decline observed in Figure 1 are dramatically apparent in these results. More specifically, each series reached its zenith just after Election Day 2020, but dropped precipitously only in the immediate aftermath of the January 6th insurrection. This suggests that the identity abandonment was a primary consequence not of electoral defeat, but rather of the post-election attack on the U.S. Capital.

### 2.3 A Year-Over-Year Longitudinal Perspective

In the blooming, buzzing confusion of a random tweet stream, is *longitudinal* analysis possible? Yes. Each observed tweet had an author whose bio we observed, and thankfully, each observation was stamped with a user_id that uniquely identifies the user and remains exactly the same no matter how often the user edits their bio, screen-name, or contact information. Next, we found those users observed in both years for every pair of consecutive years in the period 2014 through 2023. The procedure yielded nine year-over-year longitudinal samples.

As a preliminary matter we find that the longitudinal sample is quite similar to the cross-sectional sample, with respect to our relevant trend. The inclusion of "maga" in bios was not merely due to new users joining the platform who were more political than the old users; rather, users who *were active all along* were adding the token "maga" to bios that had not previously included them.

This longitudinal sample, then, has the features necessary for us to pursue answers to our research questions. We start with RQ1: *What specific bio words tend to precede the addition of a MAGA identity?*

### 2.4 RQ1: Words that Predict – and Protect Against – a MAGA Identity Adoption

Using our longitudinal subset, we compared the probability of adding a MAGA identity *given a pre-existing token was present in the bio* to the probability given that token did not exist in the bio. This is the statistical concept known as "relative risk" [5]. We computed relative risk values for every token over each year, and our MAGA signifier.

An illustrative example is helpful. Let's treat the tokens "christian" and "snapchat" as the pre-existing identities. We will examine the set of users we observed in both 2015 and 2016. We will compute probabilities of adding "trump". (The online Supplement contains all the data needed for these computations.).

Among only those users with "snapchat" in their bio in 2015, the probability of adding "trump" in 2016 was: 38 / 72,402 = 0.00052. (Prevalence change of +5 per 10,000 for "snapchat".) For those users *without* "snapchat" in their bio in 2015, the probability of adding "trump" in 2016 was: 3,311/ 4,854,412 = 0.00068. The ratio of the two probabilities yields a relative risk of 0.77. The value below 1.0 implies that having "snapchat" in the bio is a *protective* factor against later adding "trump". (This may be because Snapchat users tend to be younger, and younger people are less likely to support Donald Trump).

Next, let's consider the 29,113 users who included "christian" in their bio in 2015 (and the many more who did not). Among "christian" bios, 166 added "trump". Thus, the probability was 0.00570. (Note the decimal points; this rate is more than 8 times the overall add rate. The prevalence change was +57 per 10,000!) Among "christian"-absent users, the probability of adding "trump" was 3,183 / 4,897,701 = 0.00065. The ratio of the two probabilities yields a relative risk of 8.77. The value well above 1.0 implies that having "christian" in the bio is a *risk* factor (or *predictive* factor) for adding "trump." (This is unsurprising, as Donald Trump enjoys substantial support among evangelicals). Clearly (in retrospect) "christian" in the 2015 bio was a predictive signal of a "trump" Add in the 2016 bio.

We decided to rank *many* tokens by the strength and direction of their association with Trump-related identity events. (Specifically, we included 21,172 tokens that had previously met a prevalence threshold of 1 per 10,000 users.) We computed relative risk for every token over each year by reference to the token "MAGA". (The entire dataset is available at https://osf.io/7trsh/files/). In the next section, we will discuss results for only the year 2018, because that is the year in which the "maga" signifier recorded its maximum number of Add events.

**Predictors and Protectors of a "MAGA" Add in 2018**

The peak incidence of "maga" additions in year-over-year longitudinal samples occurred when comparing 2017 to 2018. (You can confirm this by examining Figure 1.) Let us focus on this sample and quantify the relative risk of a "maga" Add for many tokens. Figure 3 is a histogram presenting the distribution of relative risk values for pre-existing 2017 tokens. The figure displays token count on the y-axis and a log10 transformation of the relative risk on the x-axis. Values lower than 0.0 imply a protective factor (predicting a no-Add event). A value close to 0.0 indicates the token provides little information regarding whether “maga” was added. Values greater than 0.0 imply that the existence of the token in 2017 predicts the addition of “maga” in 2018. For a few tokens, we have highlighted where they fall in the distribution.

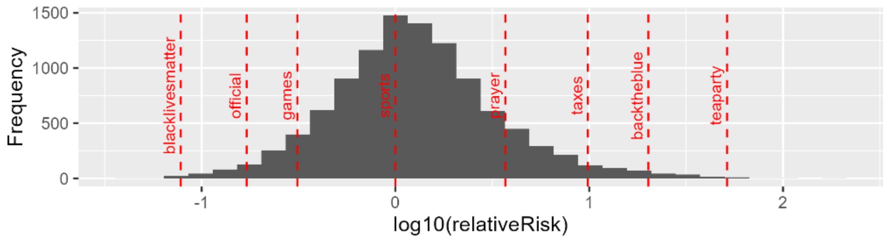


Figure 3. The distribution of “maga”-Add relative risk values for all 2017 tokens for which we observed at least one “maga” Add. Many tokens have neither protective *nor* predictive value, but some tokens strongly predict for or against.

Table 1 below lists the words that are the most predictive and most protective, from our filtered wordlist.

| Twitter Bio Words and the Subsequent Adoption of a "MAGA" Identity | | | | | | | |
|---|---|---|---|---|---|---|---|
| Most Predictive Words | Count | "MAGA" Adds | Risk Ratio | Most Protective Words | Count | "MAGA" Adds | Risk Ratio |
| "deplorable" | 2846 | 338 | 45.27 | "features" | 7883 | 0 | 0 |
| "tcot" | 1312 | 152 | 39.11 | "bookings" | 7695 | 0 | 0 |
| "nra" | 4282 | 430 | 38.5 | "amosc" | 7353 | 0 | 0 |
| "conservative" | 25384 | 2135 | 36.22 | "n*gga" | 5683 | 0 | 0 |
| "constitutionalist" | 1492 | 142 | 35.82 | "professionals" | 5551 | 0 | 0 |
| "liberals" | 1329 | 124 | 35.08 | "bts" | 5364 | 0 | 0 |
| "amendment" | 2052 | 174 | 32 | "chs" | 5342 | 0 | 0 |
| "2a" | 5260 | 436 | 31.79 | "careerarc" | 5318 | 0 | 0 |
| "israel" | 2947 | 246 | 31.63 | "communication" | 4832 | 0 | 0 |
| "prolife" | 1679 | 138 | 30.93 | "positivity" | 4549 | 0 | 0 |
| "vets" | 1175 | 92 | 29.38 | "grind" | 4282 | 0 | 0 |
| "1a" | 1463 | 114 | 29.28 | "region" | 4119 | 0 | 0 |
| "patriot" | 8338 | 623 | 29 | "academic" | 4077 | 0 | 0 |
| "lists" | 2761 | 170 | 23.21 | "una" | 4022 | 0 | 0 |
| "supporter" | 14530 | 758 | 20.41 | "downtown" | 3838 | 0 | 0 |
| "republican" | 5841 | 2308 | 20.05 | "shs" | 3836 | 0 | 0 |
| "flag" | 1501 | 79 | 19.73 | "brands" | 3778 | 0 | 0 |
| "reagan" | 1330 | 69 | 19.44 | "analytics" | 3712 | 0 | 0 |
| "military" | 7907 | 390 | 18.85 | "advocacy" | 3662 | 0 | 0 |
| "ret" | 1327 | 64 | 18.06 | "awkward" | 3623 | 0 | 0 |
| Includes only words with a usage count > 1,000. | | | | | | | |
| Excludes words with duplicate roots, and words explicitly referencing Donald Trump. | | | | | | | |

Table 1. The most predictive – and protective – words for a MAGA identity. The predictive words tend to be overtly political; the protective words vary but frequently seem to convey that the account is “public-facing,” i.e. engaged in commerce, education, government, or influencing.

It is interesting to look at specific words like these, but there are approximately 1,200 that had a raw usage count of at least 1,000 and a risk ratio of at least 1.5 (“predictive” words), and about 1,900 words that had a raw count greater than 1,000 and a risk ratio less than 0.5 ("protective" words) – that's too many words to easily discuss! It is more helpful to talk about broad *patterns* among the words, and to do that we have manually sorted the words into *categories*. In Table 2 below, we show qualitative categories which we manually constructed for our two types of words. (A full list of constituent words for each category are available in the online supplement). While this was an admittedly subjective process, future projects may devise a more objective method of category construction. (It may, e.g., be a task well suited – with supervision – for artificial intelligence).

Categories of Predictive and Protective Words

| Predictive Categories | Examples | # of Words | Protective Categories | Examples | # of Words |
|---|---|---|---|---|---|
| Sports Team Names | dodgers; dallascowboys | 53 | Digital Culture & Social Media | podcast; esports; insta | 48 |
| Politics & Government | obama; tcot; 2a; GOP | 34 | Commerce | services; boutique; premium | 31 |
| Rural Life | hunting; nature; farmer | 18 | Geek Culture | kpop; cosplay; jedi; fandom | 20 |
| Religion | faith; believer; bible | 17 | Academia | university; faculty; stanford | 19 |
| Military & Law Enforcement | veteran; usmc; police | 16 | Fashion & Beauty | chic; aesthetic; skincare | 18 |
| Technical Professions | pharmicist; accountant | 15 | Spanish Words | dios; familia; mas; amigo | 17 |
| Family Roles | mother; father; husband | 14 | Arts & Creativity | filmmaker; literary; mixtape | 16 |
| Negativity & Insults | hate; idiot; wicked; insane | 14 | Niche Sports | volleyball; futbol; lacrosse | 16 |
| Strength & Conflict | strong; survivor; battle; enemy | 12 | Hip Hop Culture | rapper; beats; sneakers | 14 |
| Emotions | angry; excited; tired | 10 | Food | taco; foodie; recipes; baker | 13 |
| Personality & Identity | independent; brave; loyal | 10 | Activism & Social Justice | blacklivesmatter; advocacy | 11 |
| Philosophy & Debate | wisdom; logic; ethics; values | 10 | Vulgarity | fucked; shitty; hoe; n*gga | 11 |
| Guns | shooting; nra; target; hunt | 10 | Urban Life | downtown; urban; memphis | 10 |
| Patriotism & Nationalism | usa; american; flag; patriotic | 8 | Drugs & Alcohol | cocktails; drunk; stoner | 10 |
| Crime | criminal; law; police; junkie | 8 | Facilitation & Cooperation | outreach; teamwork | 8 |
| Race & Ethnicity | white; jewish; hispanic; irish | 8 | Sex & Sexuality | bisexual; queer; sexy; poly | 8 |
| Vehicles | trucks; motorcycle; NASCAR | 6 | Globalism | diverse; diaspora; global | 8 |
| Finance | banking; trading; economics | 5 | Humor | goofy; hilarious; pun | 8 |
| Extremity of Degree | extreme; diehard; totally | 3 | Mental Health | depression; mindfulness | 7 |
| | | | Environmental Protection | eco; sustainability | 5 |
| | | | Coffee | caffeine; barista; café | 3 |

Table 2: Categories of Predictive and Protective words. Some words belong to more than one category, and most words were not assigned to a category at all.

## 2.5 RQ2: Words that Replace a "MAGA" Identity Following Abandonment

Our second research question is, which new identities most frequently replace an abandoned MAGA identity? Put a bit more precisely: Among the users who removed "maga" from their bio following January 6$^{th}$, what new words are most likely to enter the bio? In this section, we will further inspect the year-over-year longitudinal data to identify words that were *added* to bios coincident with a MAGA identity deletion. Specifically, we will note the incidence of words newly-added in 2021 among those users who deleted "maga" in 2021.

In ipseology, a *transmutation* occurs when a newer bio reveals that both an Add and a Delete have occurred. For instance, the sequence of bios "WUSTL Senior" followed by "WUSTL Alumnus" would be recorded as a transmutation from Senior to Alumnus. In the current analysis, we observed transmutations in the 2020-2021 longitudinal bios with "maga" as the source. Fewer than thirty distinct words were targets with 100 or more users adding the target coincident with deleting "maga". For 10 or more transmuting users, the count increases to 645 words.

The full list of these words is available in the online supplement. Broadly speaking, we find 11 categories among them, some of which suggest a true identity change, some that do not, and some ambiguously in between.

| Categories of Transmutational Words | | |
|---|---|---|
| Category | # of Words | Examples |
| MAGA Lingo | 30 | kag; stopthesteal; trumpwon; maga2020; savethechildren; swamp |
| Platform Exodus | 9 | parler; gab; telegram; gettr; freespeech; blocked; censorship; account |
| Anti-Liberalism | 32 | notmypresident; pronouns; letsgobrandon; commies; antifa; masks |
| General Conservatism | 8 | libertarian; prolife; gop; conservative; reagan |
| Patriotism/Nationalism | 10 | usa; american; freedom; patriot; flag; 1776; ifbap; godblessamerica |
| Military & Law Enforcement | 15 | allvetsmatter; usaf; semper; combat; backtheblue; police; leo |
| Guns | 4 | guns; nra; 2a; 2nd |
| Religion | 18 | christian; pray; godwins; evil; hell; believer; seeker; faith |
| Cryptocurrency | 7 | dogecoin; bitcoin; btc; crypto; investor; trader; amc |
| Family Roles | 15 | married; family; father; mother; grandma; kids |
| Hobbies & Interests | 19 | enthusiast; lover; enjoy; golf; yankees; avid; music; writer; photographer |

Table 3: Among individuals who removed “maga” from their bio following the events of January 6th, some seemed to retain the identity, some seemed to amend it to a more general iteration, and others apparently replaced it with something more benign and apolitical.

## 3 Discussion and Conclusion

Substantively, our analysis reveals insight into the identity dynamics within the MAGA movement. For example, our data suggests that myriad paths led to a MAGA identity in 2018; pre-existing identities were from such diverse categories as gun advocacy, sports team allegiance, rural lifestyle, and an affinity for big, loud vehicles. Certain pre-existing identities, by contrast, insulated *against* MAGA-fication, e.g.: involvement in public-facing professions; immersion in arts and popular culture; and concern for social justice.

There may also be multiple paths *away from* an established identification with the MAGA movement. When removing the “maga” identifier from their bio after the insurrection of January 6th, some users conveyed emphasis on a new identity: religious affiliation; or interest in cryptocurrency; or a family role; or as a particular type of hobbyist.

Perhaps none of this is surprising to those who closely follow current politics. To the extent that our substantive findings are “obvious,” we think this provides evidence for the facial validity of our methodology. Put another way, if this application of our process leads to exactly the sort of discoveries that common sense would suggest, then we might be confident that the discoveries from future applications are also grounded in reality, even when they are less intuitive. Here, boring output might be the best evidence of an effective tool.

This brings us to discuss the potential for broader methodological utility. What sorts of questions might our approach help answer? As mentioned in the Introduction, there is great interest in “stigma exits”: how individuals with stigmatized identities might transition to new, benign self-concepts. For example, as cited previously, Granberg (2011) examined the strategies that people use to improve their self esteem, after a dramatic weight loss changes their body from "obese" to "normal". Ferguson et al (2015) and Altier et al (2020) explored the identity dynamics that help former violent political extremists disengage from conflict and reacclimate to mainstream life. Both these topics have obvious practical importance. But each project, while valuable, depended upon a small number of in-depth interviews with a non-random sample of subjects. As such, the research methods were laborious, highly subjective, and severely limited in generalizability. By contrast, our ipseological approach can examine millions of cases quickly and remotely, with less need for interpretive work and a much higher degree of generalizability.

Of course, there are limitations to our approach as well. Social media sites are not perfect demographic microcosms of society. The inclusion of a certain word in a bio doesn't necessarily indicate it is part of the user's social identity. And manually grouping thousands of words into a small number of discrete categories is a rather subjective task. Despite these limitations, we hope future researchers might fruitfully employ our process to examine a variety of questions regarding identity transition. How can societies provide a psychological "off-ramp" to citizens who have come to define themselves by affiliation with a gang, cult, hate group, or other problematic faction? How might survivors of trauma, abuse, or chronic illness be taught to amend their self-concept so that they are no longer defined by their pain? In the vast, noisy sea of social media, an ipseological method might help us find a signal and chart a course.

## References

[1] Inc. Twitter, "Sample Realtime Tweets," available at https://developer.twitter.com/en/docs/twitter-api/v1/tweets/sample-realtime/overview.

[2] J. J. Jones, "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*, vol. 16, no. 11 November, Nov. 2021, doi: 10.1371/journal.pone.0260185.

[3] J. J. Jones, *Ipseology: A New Science of the Self*. Long Island: Jason Jeffrey Jones Productions , 2023.

[4] G. Eady, F. K. Hjorth, and P. T. Dinesen, "Do Violent Protests Affect Expressions of Party Identity? Evidence from the Capitol Insurrection," *American Political Science Review*, vol. 117, no. 3, pp. 1151–1157, Aug. 2023, doi: 10.1017/S0003055422001058.

[5] C. A'Court, R. Stevens, and C. Heneghan, "Against all odds? Improving the understanding of risk reporting," *British Journal of General Practice*, vol. 62, no. 596, pp. e220–e223, Mar. 2012, doi: 10.3399/bjgp12X630223.