Journal of Computational Social Science (2025) 8:28
https://doi.org/10.1007/s42001-025-00358-y

RESEARCH ARTICLE

Identity diversification and homogenization: evidence from frequent estimates of similarity of self-authored, self-descriptive text

Danial Vahabli[1] · **Jason Jeffrey Jones**[1]

Received: 4 August 2024 / Accepted: 15 January 2025

Abstract
For more than a decade, individuals composed and edited self-authored self-descriptions as social media biographies. Did these identities become more diverse over time because of a "rise in individualism" and increasing tolerance or did they become more homogenous through social learning, conformity, and fear of isolation? We analyzed longitudinal and cross-sectional Twitter bio samples with a variety of lexical and semantic methods for the 2012–2022 interval. We show that longitudinally, users diversified on lexical and semantic levels. On a cross-sectional sample—representing the state of the platform at any time point—we again observed a trend of diversification at the lexical level, but a trend of diversification reversed toward re-homogenization on the semantic level. Further, by focusing on local maxima and minima of identity similarity we identified "coordination shocks"—temporally confined intervals where similar users became overactive on the platform and drove short-term deviations from longer-term trends.

Keywords Identity · Ipseology · Individualism · Natural language processing · Identity embeddings

1 Introduction

When an individual expresses who they are, they face a dual task: highlight uniqueness, but avoid social rejection. One force pushes toward diversification – i.e. dissimilar self-descriptions. Another force pulls toward homogenization. Which force

✉ Danial Vahabli
danial.vahabli@stonybrook.edu
https://scholar.google.com/citations?user=jswGuY0AAAAJ&hl=en

Jason Jeffrey Jones
https://scholar.google.com/citations?user=erSGiqUAAAAJ&hl=en

[1] Department of Sociology and Institute for Advanced Computational Science, Stony Brook University, Stony Brook, USA

Published online: 06 February 2025

is stronger when millions of individuals begin authoring and editing persistent, public autobiographies? Here we explore answers to that question with eleven years of observations from millions of US Twitter accounts.

1.1 The argument for identity diversification

According to some, the *Rise of Individualism* is a global, long-term trend. Santos, et al. [26] examined decades of census and survey data regarding individualist practices (e.g. living alone) and individualist values (e.g. the importance of teaching children to be independent versus obedient). They found an upward trend in both measures. Thirty-four (out of 41) countries showed a rise in individualist practices, and thirty-nine (out of 53) showed a rise in individualist values. In the US specifically, estimates for change over time were increasing on both measures.

Further evidence comes from studies of language use in cultural products. Comparative analysis of cultural discourse in media such as magazine and TV shows in United States and East Asia suggests that American culture values individualism and uniqueness [19, 23, 32]. On Twitter, it has been found that tweet authors mention other users less often in more individualistic countries such as the United States [9]. Twenge [28] has framed recent trends in the regard for self versus others as a shift from "Generation We" to "Generation Me". This change is especially apparent in language. Singular pronouns are more frequent than plural pronouns in American books [30]. Parents have increasingly shunned common names (e.g. Mary) for their children and instead bestowed unique names (e.g. X Æ A-Xii) [29].

From this line of argument, one can perceive at least two paths to identity diversification. First, in response to the value placed on individuality and uniqueness, individuals may prefer to craft and edit their self-descriptions to differ noticeably from others. They might accomplish this by saying essentially the same thing in a different way (e.g. progressive versus liberal) or by inventing novel titles (e.g. dogecoin holder).

Second, users may respond to a (real or perceived) increase in tolerance for previously denigrated or ignored identities. Twenge, Carter, and Campbell [31] have argued that younger Americans possess more tolerance for diversity. As individuals express themselves more freely, previously absent (or silent) subcultures and sub-communities can emerge. Either or both mechanisms should drive increasing identity diversification – i.e. more dissimilarity between any two randomly-chosen individuals.

1.2 The argument for identity homogenization

In contrast to observations pointing toward the rise of individualism, it could be argued, that in group settings, humans tend to conform [1]. It is a reasonable hypothesis that individuals will learn over time what to include in online biographies, and they may do so through imitation learning [3]. Specifically, they will observe others' biographies and edit their own to be similar. Prior research has shown behavior – specifically voting – spreading across online social ties [4, 16]. In addition, using

marginalized identities in a public space can open the door to harassment, hence pushing individuals towards conformity [8, 10, 18].

There is evidence that elements of identity cluster in the social network; Tucker and Jones [27] demonstrated that US Twitter users with pronoun lists (e.g. she/her) in their biography were especially likely to follow and be followed by other users also with pronouns. Further, they found that the prevalence of users with pronoun lists in their bios had dramatically increased. This pattern suggests innovation (adding pronouns) followed by contagion (others copied the behavior). When such a process occurs, bios of the participating users will necessarily increase in similarity.

Finally, population shifts in salient subjects (even without any learning or conformity) could drive increasing similarity. For instance, it has been observed that Americans increasingly mention political affiliations in their biographies [25]. Following the Russian invasion of Ukraine, the number of US users with a Ukrainian flag emoji in their bio grew from near-zero to tens of thousands within days [11]. As an increasing proportion of the population narrows in on one Topic, their biographies will necessarily become more similar.

1.3 Optimal distinctiveness theory

A synthesis of these competing forces may be found in optimal distinctiveness theory. Optimal distinctiveness theory [21] posits that while individuals are creating their identity, there are two competing pressures. On one side, individuals want to be included in groups and be similar to their members. On the other side, they want to be distinct and different from others so as not to be ignored. Everyone plays an equilibrium game by finding the optimal distinctiveness towards others. Individuals strive to be similar to but different enough from others.

On social media, individuals describing themselves face the dilemma posed by optimal distinctiveness theory. We argue that the rise of individualism and the imitation learning process are additional, external forces which push similarity in opposing directions. Being completely unique may lead to rejection (or at least the fear of rejection) for being too weird. But standing out from others in the population is a valued trait in individualist cultures. The argument for diversification delineates reasons bios should have become increasingly different, whereas the argument for homogenization provides paths for the bios to become increasingly similar. The rest of this paper evaluates the evidence regarding the relative strength of these dynamics.

2 The current study

Here we endeavor to estimate individual-to-individual similarity in *personally expressed identity*. Personally expressed identity is self-authored, visible, self-descriptive text. It is personal—the individual is describing themselves. It is expressed—these are words the individual emits, where others might see them. And

it describes identity—the explicit purpose of the text is description of the author. Discussion of how to estimate *similarity* is deferred to the methods.

We chose Twitter profile biographies as the source of personally expressed identity text. This choice afforded us the opportunity to estimate similarity at high resolution, drawing from millions of observations spread over more than a decade. It accords with the ipseological philosophy [15]: consistent, persistent, precise estimation is and should be the foundation upon which a science of identity is built.

2.1 A model of personally expressed identity text production

When producing personally expressed identity text, individuals must make many choices. Influenced by oral and written language production models, we provide our own personally expressed identity text production model as depicted in Fig. 1. (For a review of language production models read [7]). Many language production models have 3 main steps: a semantic level where individuals decide the meaning they wish to convey, a lexical level where they translate that meaning into grammatically sound word sequences, and a motor level where they use their body to execute the act of production. Whether these steps are serial or parallel is a matter of much debate, but outside the scope of our work. Similarly, the stage of motor function is not of much interest to the present work. Personally expressed identity text production also differs from general models in two more aspects. First, Twitter users have a limited length of characters to define themselves. Second, they have as much time as they care to expend on composing their bio and the opportunity to edit it whenever they wish.

We include three levels in our model: Topic, Stance and Lexicon. At the Topic level, individuals choose whether to include (or not) each potential category of identity. There are *many* potential categories, and the length restriction of a Twitter biography means only a few can be included. At the Stance level, individuals instantiate their position within the Topic. Finally, at the Lexicon level, individuals choose the

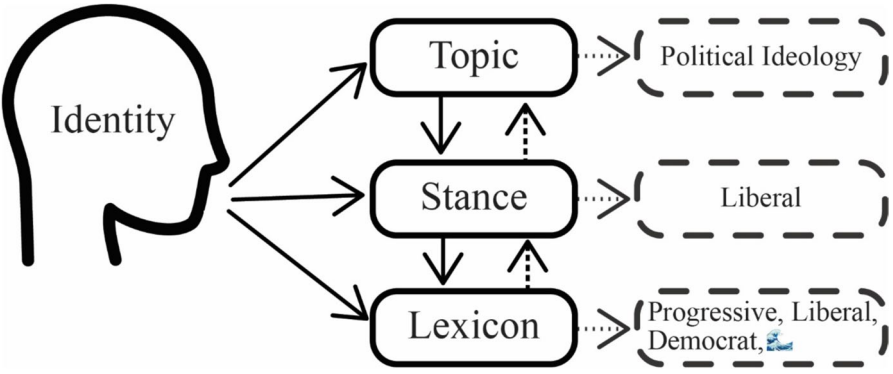

Fig. 1 A model of personally expressed identity text production. Individuals choose which categories to include in the bio at the Topic level. At the stance level, they signal their position within the Topic. At the lexicon level, individuals choose specific words to communicate their stance within each Topic

exact wording to appear in the text. For example, a user might decide religious affiliation is a Topic they should include in their bio. As a Stance, they may wish to communicate their identity as a Christian. At the Lexicon level, they might choose to say "follower of Jesus" or "Baptist" or quote a passage from the Bible.

We argue that individuals will choose Topics that are most central to their identity. In many instances, their Stance is determined – one is expected to have only one religious affiliation. But in other cases, it is not. One could be a fan of multiple celebrities or media products, and be forced to choose only a subset for which to express that fandom. At the Lexicon level, both individual choice and quirks of language will arise. Consider, for instance, the fact that there are fewer words for positive emotions than negative [2]. While these specific details may constrain and influence the space of potential bios, we contend that individuals still exert considerable control over how they are described. Social trends, online fads, the changing composition of active users and their demographics and many more factors all will affect the similarity of bios in the sampled population. The aim of the current work is to estimate individual-to-individual similarity within the population. The intent of this model is to facilitate analysis at separate levels of expression.

3 Methods

3.1 Data: personally expressed identity text

On the Twitter social media platform, users constructed a profile. They entered their name, uploaded an image and chose a screenname. Crucially, for our purposes, they were prompted to enter a bio. This was a short, free-form text field. The length was limited to 160 characters. These profile bios were the source of our personally expressed identity text data. Users were also prompted to enter a location. We used the text of the location field to geocode profiles. We sorted bios to US locations (included in analysis) versus non-US locations (excluded from analysis) using the same algorithm as Tucker and Jones [27].

To sample profile bios, we relied on the 1% sample of all public tweets. Jones [17] details the process of constructing datasets of profiles from the tweet stream, but we will review it here briefly. There was no mechanism to randomly sample the Twitter userbase. Instead, we examine the random sample of tweets. Each tweet had a snapshot of the author's current profile (including bio and location) attached to it. Here we use tweets (and thus profiles) collected continuously from 2012 through 2022. It should be understood, then that each day's sample consists of actively tweeting users on that day. Our sample of tweets provides a cross-sectional sample of tweet authors over time.

From the daily, US cross-sectional sets of profiles, we randomly sampled N number of unique users' bios every other day. We chose to not alter the text of the bio in any way (no stemming, deleting stop words or links, etc.), because we wanted to compare the raw text as presented by individuals. The similarity of each pair of bios was calculated using the methods described below. This yielded N/2 unique pair estimates of individual-to-individual similarity. In every iteration of the code

at every time point, we sample the N bios and N/2 random pairs again. This makes sure that we capture a large number of unique bios. For further robustness checks about bio selection see the supplementary material.

Additionally, we prepared a longitudinal dataset. This dataset had annual resolution. For each year, one instance of every observed US profile was chosen at random and saved. (That means that each user contributed one bio per year, whether they were observed tweeting once or one hundred times.) The intersection of each annual set of users served as the longitudinal set. In total we have 280,635 users who consistently show up at least once in the 2012–2022 interval every year; we made use of one observation for each user in each and every year 2012–2022.

3.2 Quantifying similarity

It is not at all clear how one *ought* to estimate the similarity of one self-authored, self-descriptive text to another. Therefore, we implemented multiple methods – each of which we reasoned could legitimately be said to quantify similarity. Relying on our 3-level bio production model, we argue that similarity can emerge in each of the 3 different levels. First, two users may decide to mention the same Topic in their bio; for example, both decided to mention their university degree, political ideology, and favorite artist (Topic level). Further, they may have a similar Stance on a Topic. Perhaps they have the same favorite artist, same university degree, but dissimilar political ideologies. Lastly, each user must choose particular words to represent each element of their identity (Lexical level).

Consider the two bios "MBA, conservative, Bruce Springsteen fan" and "Liberal MBA who loves The Boss." These bios are identical on the Topic level, very similar at the Stance level, but overlap only a small amount at the Lexical level. To capture similarity at multiple levels, we opted to use multiple measures of similarity. We do not argue that each measure maps cleanly to one and only one level. Instead, we rank them on a continuum (we believe to be self-evident) from more lexical to more topical.

We have 3 measures of similarity we place nearer the Lexical pole in our production model. First, we simply counted the number of unique tokens on each day per N bios. A token can be an emoji, a word, or other expression such as "BLM." Second, we sampled two users' bios, tokenized each and then calculated the Jaccard index over the two sets of tokens. The Jaccard index calculates the proportion of tokens which were shared in the pair. The mean over N/2 pairs per day was used for this similarity score. Third, we compressed N bios and calculated the compression ratio: original file size in bytes to compressed file size. Higher compression ratios correspond to higher similarity. The lexical methods make no attempt at understanding the underlying meaning of the tokens. For such measures, "science" and "scientist" are completely different.

We have 3 measures further from the Lexical pole and closer to the semantic, meaning-based, Stance and Topic levels in our production model. These are based on the distributional hypothesis [12] and contemporary language modelling algorithms. First, we calculated a vector representation of bios by averaging over the

tokens comprising the bio using word2vec [22] and emoji2vec [6] token vectors. We then calculated the pairwise cosine similarity of bio pairs (where each bio was the mean of its constituent tokens). Second, we used the BERTscore algorithm [33] with two pretrained language models—BERTweet-Base [24] and deberta-xlarge-mnli [14]. For these measures, "science" and "scientist" are very similar.

Lastly and separately, we trained our own custom bio embedding models on diachronic samples. We computed vector representations of tokens within bios using the word2vec algorithm. The training data consisted exclusively of bios observed within the temporal period (daily or annual). For an intuitive explanation, imagine the algorithm must newly learn the meanings of every word at each datapoint only from reading the bios from that time interval (day or year). The vector embeddings the algorithm produces thus represent the "space" of personally expressed identity on that time interval. We aimed to estimate the size of that space to see if it was growing or shrinking over time. For each diachronic we sampled N unique words, made N/2 random pairs and calculated the cosine similarity of each pair. The average cosine similarity is reported as the similarity score of that time interval. Our measure estimates the average distance in the space between two randomly selected vectors. We intend for this to estimate the overall size of identity space for the time period.

Our bio embeddings' semantic space is different from the pre-trained models'. Our custom embeddings result from learning the meanings of words by reading bios. Similar to other embedding models, our process honed vector representations of words through context and reliance on the distributional hypothesis. By training our own model, we made a language model sensitive to the typical (and atypical) co-occurrences of words *in the context of bios*. In many large text corpora, words of a similar Topic have many opportunities to co-exist or be used in a similar context. However, Twitter bios are severely limited in length compared to texts in other corpora—books or Wikipedia articles, for example.

Consider the words "republican" and "democrat." In books, news articles and probably most documents, these words must frequently coexist. Similarly, their surrounding contexts are likely quite similar. Contrast this with Twitter bios, which have two constraints: short length and the intention of describing one individual. Rarely will republican and democrat co-occur. Due to polarization in the United States, its likely their contexts are dissimilar. We find evidence for this. In a pretrained embedding, republican is in the top ten words most similar to democrat and vice-versa. In our bio embeddings that is not the case; "democrat" is more similar to the words "liberal", "equality" and "resiter" than "republican." We will argue below that this makes our custom bio embeddings especially sensitive to contrasting meanings at the Stance level. (See the supplementary material for details.)

4 Results

We visualized the temporal trends in diversification/homogenization over 11 years in Figs. 2, 3, 4, 5, 6. Panels within each Figure present results for different measures. We applied our methods to three differing datasets. First, we estimated

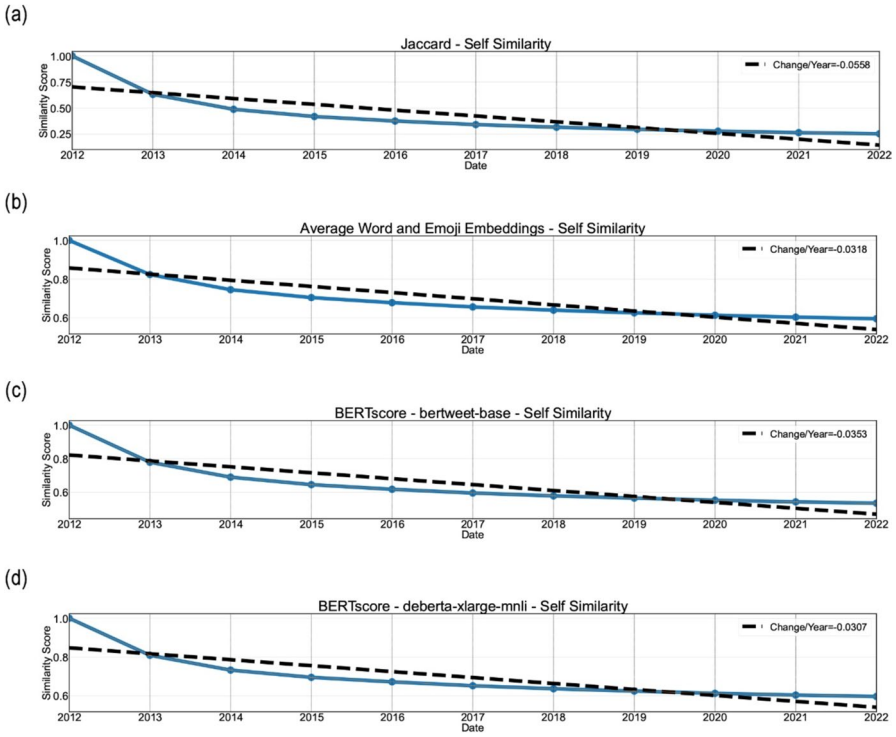

Fig. 2 Self-to-self similarity. For 100,000 users, each bio was compared to the 2012 bio. Every measure displayed decreasing similarity over time

self-to-self similarity using the longitudinal dataset. By calculating the similarity of each user's bio every year to their bio in 2012, we provide a baseline for similarity (N = 100,000). Next, we selected random pairs from the longitudinal dataset per each year and calculated similarity scores (N = 100,000). Finally, we used the cross-sectional, daily dataset to calculate bio similarity on every other day for 50,000 bios (N = 50,000). While the longitudinal results follow the same set of users over time, the cross-sectional results reflect the bio-to-bio similarity of a churning sample of users. That churning sample is representative of the active users per each day. Of course, these cross-sectional results are thus affected by users dropping out of the platform and new users joining the platform. One would expect they will also be temporarily affected by current events both online and offline.

4.1 Self-to-self similarity trends

Figure 2 depicts the results for self-to-self similarity across all applicable measures. (We cannot apply compression ratio and custom bio embeddings to measure self-similarity.) Every user's bio was compared to their 2012 bio. In the Figure, we present the mean similarity over randomly selected 100,000 users. All users' similarity

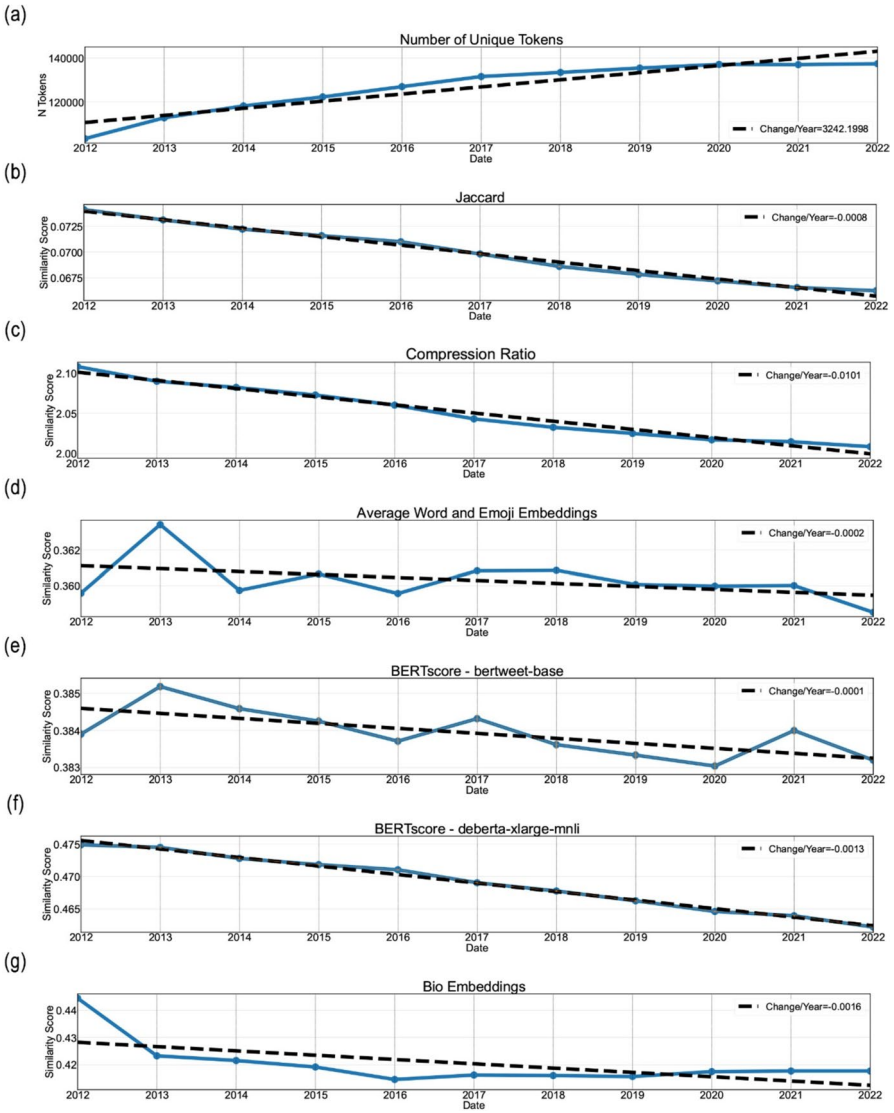

Fig. 3 Longitudinal sample, similarity estimates for random pairs or sets. Every measure indicates diversification—decreasing similarity—over time

of their 2012 bio to their 2012 bio was 1.00. From 2013 forward, we observed steady diversification.

The results simply reflect that users revise their bios and diversify from themselves over the years. It would not be possible for this trend to show homogenization – i.e. rise above 1.0. But there was no guarantee the results must have shown the smooth monotonic decrease that they do. The mean similarity could have gone down then back up again or flatlined at a floor or ceiling value. We believe the trend

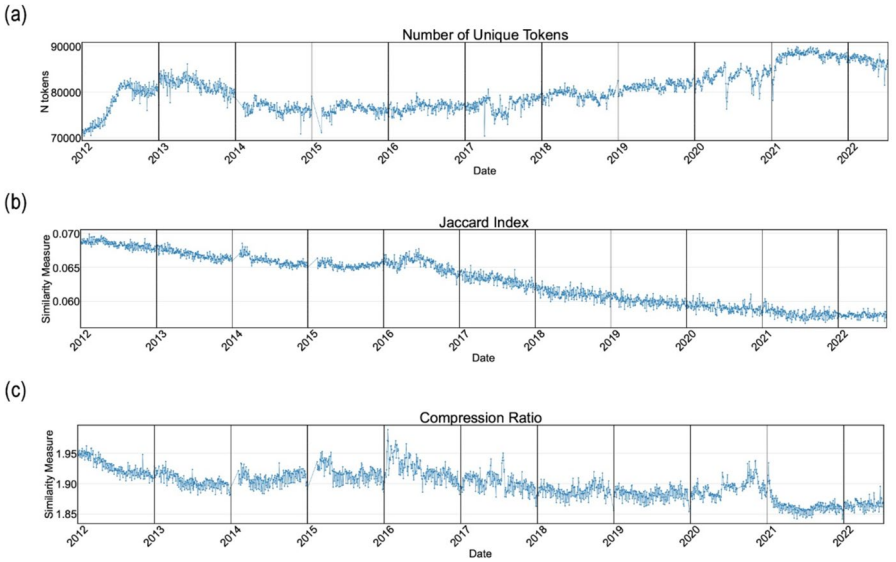

Fig. 4 Lexical similarity estimates for cross-sectional, daily samples. Broadly, we observed diversification—decreasing similarity

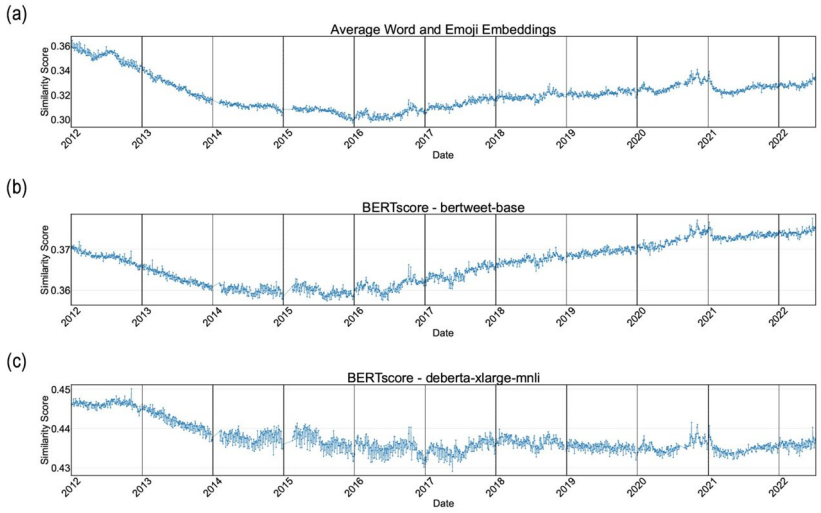

Fig. 5 Semantic similarity estimates for cross-sectional, daily samples. The evidence consistently points to diversification in the early years (pre-2016)

we observe demonstrates the results one should expect: our current self drifts away from our past self – quickly at first, then more gradually. For results with different years as the reference year see the supplementary material.

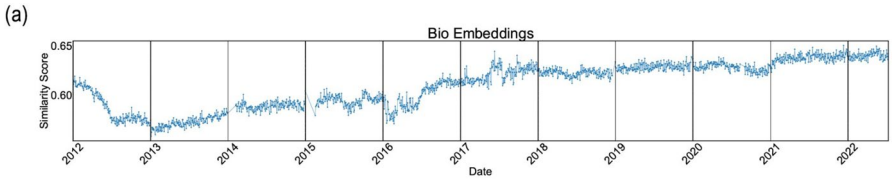

Fig. 6 Custom bio embeddings estimated daily similarity as the size of identity space. Every day, a model learned word embeddings anew from 50,000 sampled bios. Then we estimated the size of identity space by randomly sampling vectors and computing the distance between them

4.2 Longitudinal set similarity trends

Figure 3 summarizes the results of calculating similarity estimates for random pairs of users chosen from the 11-year longitudinal set. (The compression ratio and bio embeddings are actually computed on a subsample of bios – not pairs.) Over the entire interval, all measures indicate diversification. Lexical measures present an almost perfectly linear diversification trend. Users increasingly used different terms to express themselves. Semantic measures also revealed an overall diversification over the interval with temporary fluctuation. Years of presidential elections (2012, 2016, 2020) are all local minima in average word and emoji embeddings (Panel d) and BERTscore-BERTweet-base (Panel e). Each is also followed by a sudden increase in similarity (2013, 2017, 2021). These results might reflect polarization of the platform prior to each presidential election.

As expected, the average similarity scores are lower than self-to-self similarity scores across the measures. Within this compressed range, the magnitude of estimated slopes were smaller, but still reliably signaled diversification in every case. This is evidence supporting the rise of individualism argument – at least within consistently-active Twitter users in the United States over these 11 years.

4.3 Cross-sectional set similarity trends

Finally, Figs. 4, 5, 6 depict the results for the cross-sectional dataset. Intuitively, the cross-sectional results represent the user-to-user similarity on a given day on the platform as it would be experienced by someone reading tweets chosen at random. As discussed earlier, this cross-section is a churning mixture of new users entering, old users leaving and remaining users editing their bios. Current events of the day also likely affect who is active. For example, we suspect the days surrounding presidential elections are saturated with users interested in the election. Hence, the results are more complicated than previous results. Across different measures, the results signify different trajectories, and we discuss them separately.

The small daily fluctuations represented in the results are in part due to the randomness introduced in the subset selection process. For further information see the supplementary material.

4.4 Lexical similarity

Figure 4 depicts the results for lexical measures. Panel (a) denotes the number of unique tokens in total over 50,000 bios per day. Panel (b) depicts the Jaccard index per day for 50,000 bios (25,000 pairs) and Panel (c) shows the compression ratio for a file containing the text of 50,000 bios.

We argue that these measures estimate similarity at the Lexical level of our production model. Panel (a) shows that the number of unique tokens increased in the 2012–2022 interval. The decline in Jaccard Index in panel (b) shows that users' bios overlapped less—they shared a lower proportion of tokens over time. Further, panel (c) shows that the compression ratio decreased; at a byte-level there were fewer repeated patterns in later compared to earlier days.

All three measures eschew latent semantic similarity; they focus on the co-presence of tokens. High similarity scores would only arise if bios shared the same tokens. The results in Fig. 4 unanimously indicate growing diversification at the lexical level. The results are also aligned with longitudinal results reported in Fig. 3.

4.5 Semantic similarity

Figure 5 summarizes results from measures meant to estimate semantic similarity. Semantic meaning was translated to a numeric form using a vector representation of bios' textual content. In these three cases, we borrowed word vectors from pre-trained models. Panel (a) and Panel (b) show a trend of diversification, an interval of stable similarity, and then a rise in similarity. Panel (c) shows a trend of diversification followed by an interval of stable similarity.

Contrasting Figs. 4 and 5 demonstrates that the answer to the question, "Are identities diversifying or homogenizing?" depends upon operationalization. Individuals chose different words to include in their bios, but when we swapped words for their numerical counterparts in machine-learned, latent semantic spaces, no clear and consistent temporal trend emerged. Perhaps early Twitter users were very similar in self-representation, then a growing userbase pushed toward diversification, reaching a nadir in 2016 before similarity began climbing again. Panels (a) and (b) are consistent with this story. However, panel (c) breaks from the story in 2016, as similarity flatlines rather than pivots.

Comparing the cross-sectional results with longitudinal results in Fig. 3 shows the homogenization trend post 2016 is mainly apparent in cross-sectional results. This is not a contradiction, because the results draw from different sets of users. While the longitudinal dataset follows a subset of users who consistently used the platform for at least 11 years, the cross-sectional results include a wider range of users which were active at their recorded time. Further research might contrast users by their join cohort (i.e. when they first created their profile) and the number of years spent on the site, but we can safely draw one conclusion from the present results: the later homogenization trends were not due to veteran users becoming more similar.

4.6 Custom bio embeddings and "identity space"

Lastly, Fig. 6 summarizes results from our custom bio embeddings language model trained on 50,000 bios per day. Similarly to Average Word and Emoji Embeddings (Fig. 5, Panel a) and BERTscore – BERTweet-base (Fig. 5, Panel b), one observes a temporal trend of diversification followed by homogenization. But the inflection point is much earlier – a minimum value in early 2013.

Because these models were trained on the bios themselves rather than any other external data, the results reflect only how tokens are used in Twitter bios. Keeping in mind the increase in the number of unique tokens reported in Fig. 4, the measure suggests that while the number of unique tokens was increasing, the underlying latent semantic spaces was not expanding apace. In fact, it was contracting over most of the period of observation. It surprised us that the custom bio embeddings (Fig. 6) were not tightly correlated with all the other embeddings-based results (Fig. 5). Table 1 contains the correlation matrix for the four embedings-based cross-sectional daily similarity estimates. In every row, at least one measure shows a non-positive correlation. This is evidence that similarity estimates (and therefore the trends observed) depend upon choice of word vector representation.

We posit that similarity estimates based on pre-trained embeddings are driven more by Topic similarity than Stance similarity. The converse is true for similarity estimates based on our custom embeddings – they are weighted more toward measuring Stance distance than Topic. In the following section, we illustrate this supposition with an example.

4.7 Coordination shocks

Across similarity measures, we observed certain days on which similarity scores have a sudden increase (For example, the end of 2020 on Fig. 5). We call these sudden increases *coordination shocks*. We speculate coordination shocks occur when users with similar identities become over-active in response to or anticipation of an event. For example, the coordination shock on 11/6/2020 corresponds to Election Day and a tweet from Donald Trump: "Joe Biden should not wrongfully claim the office of the President. I could make that claim also. Legal proceedings are just now beginning!". See Fig. 7. Further, there was a coordination shock on January 7, 2021, a day after the attack on the United States Capitol. In addition, there was a peak on

Table 1 Correlation matrix for all daily similarity estimates for the four embeddings-based measures

	Average Word + Emoji	BERTscore BERTweet	BERTscore deberta	Custom Bio Embedding
Average word + emoji	1.000	0.707	0.469	− 0.020
BERTscore BERTweet		1.000	− 0.019	0.689
BERTscore deberta			1.000	− 0.015
Custom bio embedding				1.000

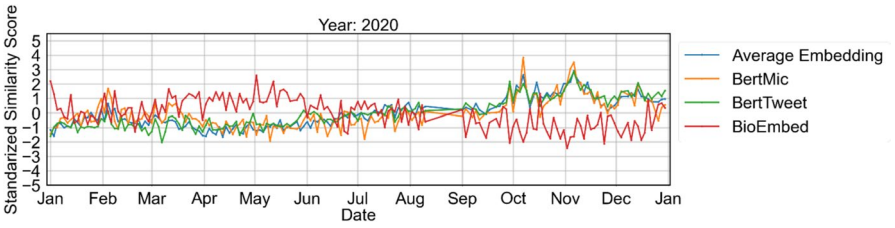

Fig. 7 Estimated daily similarity within 2020 using the four embeddings-based measures. Notice the divergence of the Custom Bio Embeddings and other series surrounding the Election Day coordination shock

24th June 2022 we speculate was related to the Supreme Court's abortion ruling overturning Roe v. Wade. These are visualized in the supplement.

These coordination shocks were marked by a sudden rise in similarity in all measures except the Custom Bio Embedding measure. Coordination shocks showed a sudden *decrease* in similarity in the Custom Bio Embedding measure. See the divergence between series in November in Fig. 7. Based on the behavior of these series, the centrality of the Twitter platform in the Biden-Trump election, and the differing representations of political affiliation words we explored previously, we put forward a potential explanation. Pretrained embedding models recognized that politicized bios were at a heightened prevalence in the days surrounding Election Day. The coordination on the Topic of politics drove an increase in similarity. Our Custom Bio Embeddings, however, learn language each day anew during the same period. We posit this allows (one could say forces) these similarity estimates to go down as they recognize the polarized Stances of tweet authors' bios.

5 Discussion

Our evidence – based on millions of observations of online public self-descriptions – supports multiple conclusions. First, at the Lexical level, bios diversified. In 2022, individuals used a broader vocabulary to describe themselves than they did in 2012. They shared fewer tokens in common with others. Second, the preponderance of the evidence suggests that diversification at the Semantic level occurred during early years, but then stopped or reversed. We can't say *why* with the current analysis, but we can support the claim that US tweet authors became more cosmopolitan early in the observation period, but that trend did not hold.

Where does that leave us on the broad question of identity diversification versus homogenization? The academic answer is: it depends on operationalization and more research is necessary. Subsample and out-of-sample replications of our work would provide more information. How did bio similarity change among elected officials? Among kpop fans? Among professional athletes? How did bio similarity change pre- to post- Musk acquisition? Did other nations' bio similarities follow the same temporal trends as the US?

For now, we offer this as our best interpretation of the evidence: One may claim that identities diversified only if one constrains analysis to early in the observation period or only at the superficial level of word choice. Whether named *X Æ A-Xii* or *Mary*, there is evidence that the bounds of identity space for our users actually constricted over time.

At the same time, individuals did change. Self-to-self comparisons revealed current selves drifting away from past selves. Further analysis could reveal the extent to which conformance to platform norms drove this drift. A hypothesis one could explore is that users presented their own weird authentic selves on Twitter in 2013, but then became politicized, well-sorted group members with their rough edges sanded off over the course of two contentious Twitter-centric Presidential elections.

Readers should not take away the wrong conclusion that Twitter users are a narrow and homogenous group. Weird Twitter and weird users still exist. There has always been a long tail of idiosyncratic, infrequently-used words in bios [17]. Longtime, consistent users of the platform (Fig. 3) ended up less similar to each other than they began in 2012. We argue that the conformance and restricted identity space we observe in the daily series is driven by later-joining users. Later-joining new users received more information on what was "allowed" or "normal" to mention in one's bio. Existing popular users set norms on what a twitter profile should look like and what Topics were to be included. These standards were adopted by newly joining users. They learned which Topics to include in a bio while still differentiating from others by expressing their own stance and choosing their own words.

We were surprised and pleased to discover coordination shocks, moments when similarity sharply deviated from its longer-term trend. Coordination shocks appeared in conjunction with highly-salient US political events. This comports with previous research demonstrating the importance of politics on the platform [5, 25]. We have speculated that the divergence of our custom bio embeddings measurement from measures using pre-trained embeddings indicates models "seeing similarity" at different levels. Custom bio embeddings realized that a *Democrat Stance* and a *Republican Stance* differ on many axes, while pre-trained models are tuned to recognize two bios both containing a *Politics Topic*. We believe there is a great deal of fruitful research to be done testing this speculation.

6 Limitations

Some Twitter users may be bots. Because of hyperbolic claims from early studies, many believe that bots are easily distinguished from users representing an individual, existing human. More recent, less-conflicted work shows this was never the case [13]. It is possible "Dead Internet Theory" is true and we have presented here the similarity of bots to other bots. The reader is free to believe that if they wish.

Twitter users—the human ones, should the reader allow—are not a representative sample of Americans. However, it is a large and available sample. One in four Americans claim to have a Twitter profile, and we found millions of unique user bios including a few hundred-thousand with full, annual, longitudinal data. We are

not aware of any repeated, representative sample of individual self-descriptions of any size.

This study is only a first step towards an exhaustive understanding of identity diversification and homogenization on social media. In this study, we focused on documenting trends within United States users of Twitter. It might be the case that the trajectory of identity diversification is contingent on country and social media platform. For example, our results might be affected because Twitter suspended many accounts following the 2016 U.S election in the aftermath of congressional investigations about Russian interference [20]. Future work should implement similar studies on different social media platforms—such as Instagram and TikTok—to see if the trends we document here are Twitter specific or replicate across platforms. Future studies can also focus on users in different countries.

We have avoided an overarching narrative to explain "why" identity diversification took the course it did. Over short time intervals, we observed and speculated on the reason for coordination shocks – i.e. high-profile political events attracted users with political bios. However, we believe the multiyear trends, the difference across measures and the specific inflection points are better documented than "explained." Many reasons for rising and falling diversification are simultaneously plausible. We do not believe our data affords immediate means to apportion credit or blame. We hope the current results spur future work testing the mechanism(s) driving active users' self-descriptions to become more similar or more divergent.

7 Conclusions

Identity diversification (consistent with arguments for increasing individualization) and identity homogenization (consistent with arguments for conformity and social learning) are opposite trends. One could reasonably expect one or the other to dominate given consistent, persistent measurement of identity similarity. We argue that we have presented such evidence and are justified in several conclusions. Over time, individuals became less similar to their past selves. US Twitter bios were most homogenous (by most measures) at the beginning of our observation period. At a semantic (but not lexical) level, we found evidence for a reversal away from diversification and toward re-homogenization. Prospective, longitudinal studies of representative samples of individual humans are critically necessary to determine if our selves are defining a contracting or expanding identity space.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s42001-025-00358-y.

Acknowledgements This material is based upon work supported by the National Science Foundation under grant CCF-2208664 (JJJ). The Center for Advanced Internet Studies provided JJJ a fellowship in the Fall of 2023 that allowed the focus of time and attention on this work. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. DV wants to thank the Institute for Advanced Computational Science at Stony Brook University for their graduate fellowship and computational resources.

Data availability The data that support the findings of this study are available upon reasonable request.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Asch, S. E. (1951). "Effects of group pressure upon the modification and distortion of judgments." In *Groups, leadership and men; research in human relations*. Oxford, (Pp. 177–90) England: Carnegie Press.
2. Averill, J. R. (1980). On the paucity of positive emotions. In K. R. Blankstein, P. Pliner, & J. Polivy (Eds.), *Assessment and modification of emotional behavior* (pp. 7–45). Springer. https://doi.org/10.1007/978-1-4684-3782-9_2
3. Bandura, A., & Richard H. W. (1977). *Social Learning Theory*. Vol. 1. Englewood cliffs Prentice Hall.
4. Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., 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
5. Eady, G., Hjorth, F., & Dinesen, P. T. (2023). Do violent protests affect expressions of party identity? Evidence from the capitol insurrection. *American Political Science Review, 117*(3), 1151–1157. https://doi.org/10.1017/S0003055422001058
6. Eisner, B., Tim R., Isabelle A., Matko B., & Sebastian R. (2016). "Emoji2vec: learning emoji representations from their description." In *Proceedings of the fourth international workshop on natural language processing for social media*. Austin, (Pp. 48–54) TX, USA: Association for Computational Linguistics.
7. Fayol, M. (1991). From sentence production to text production: investigating fundamental processes. *European Journal of Psychology of Education, 6*(2), 101–119.
8. Francisco, S. C., & Felmlee, D. H. (2022). What did you call me? An analysis of online harassment towards black and latinx women. *Race and Social Problems, 14*(1), 1–13. https://doi.org/10.1007/s12552-021-09330-7
9. Garcia-Gavilanes, R., Quercia, D., & Jaimes, A. (2013). Cultural dimensions in twitter: time, individualism and power. *Proceedings of the International AAAI Conference on Web and Social Media, 7*(1), 195–204. https://doi.org/10.1609/icwsm.v7i1.14419
10. Green, M., Bobrowicz, A., & Ang, C. S. (2015). The lesbian, gay, bisexual and transgender community online: Discussions of bullying and self-disclosure in youtube videos. *Behaviour & Information Technology, 34*(7), 704–712. https://doi.org/10.1080/0144929X.2015.1012649
11. Hare, M., & Jason J. (2023). "Slava Ukraini: Exploring identity activism in support of ukraine via the ukraine flag emoji on twitter." *Journal of Quantitative Description: Digital Media* 3. https://doi.org/10.51685/jqd.2023.005.
12. Harris, Z. S. (1954). Distributional structure. *WORD, 10*(2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520
13. Hays, C., Zachary S., Manish R., Erin W., & Philipp Z. (2023). "Simplistic collection and labeling practices limit the utility of benchmark datasets for twitter bot detection." In *Proceedings of the ACM Web Conference 2023*, *WWW '23*. (Pp. 3660–69) New York, NY, USA: Association for Computing Machinery.
14. He, P., Liu X., Gao J., & Chen W. (2021). "DeBERTa: Decoding-enhanced BERT with disentangled attention."
15. Jones, J. J. (2024). *Ipseology: A New Science of the Self*. Jason Jeffrey Jones Productions.
16. Jones, J. J., Bond, R. M., Bakshy, E., Eckles, D., & Fowler, J. H. (2017). Social Influence and political mobilization: further evidence from a randomized experiment in the 2012 US presidential election. *PLoS ONE, 12*(4), e0173851.
17. 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), e0260185. https://doi.org/10.1371/journal.pone.0260185

18. Katz-Wise, S. L., & Hyde, J. S. (2012). Victimization experiences of lesbian, gay, and bisexual individuals: A meta-analysis. *The Journal of Sex Research, 49*(2–3), 142–167. https://doi.org/10.1080/00224499.2011.637247
19. Kim, H., & Markus, H. R. (1999). Deviance or uniqueness, harmony or conformity? A cultural analysis. *Journal of Personality and Social Psychology, 77*, 785–800. https://doi.org/10.1037/0022-3514.77.4.785
20. Le, H., Boynton G. R., Shafiq Z., & rinivasan P. (2020). "A postmortem of suspended twitter accounts in the 2016 U.S. Presidential election." In *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining*, *ASONAM '19*. New York, NY, USA: association for computing machinery. (Pp. 258–65)
21 .Leonardelli, G. J., Cynthia L. P., & Marilynn B. B. 2010. "Chapter 2—Optimal distinctiveness theory: A framework for social identity, social cognition, and intergroup relations." In *Advances in Experimental Social Psychology*. Vol. 43, edited by M. P. Zanna and J. M. Olson. Academic Press. (Pp. 63–113)
22. Mikolov, T., Chen K., Corrado G., & Dean J. (2013). "Efficient estimation of word representations in vector space."
23 Morling, B., & Lamoreaux, M. (2008). Measuring culture outside the head: A meta-analysis of individualism—Collectivism in cultural products. *Personality and Social Psychology Review, 12*(3), 199–221. https://doi.org/10.1177/1088868308318260
24. Nguyen, D. Q., Vu T., & Nguyen A. T. (2020). "BERTweet: A pre-trained language model for english tweets." In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*. online: association for computational linguistics. (Pp. 9–14)
25. 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
26. Santos, H. C., Varnum, M. E. W., & Grossmann, I. (2017). Global Increases in Individualism. *Psychological Science, 28*(9), 1228–1239. https://doi.org/10.1177/0956797617700622
27. Tucker, L., & Jason J. (2023). "Pronoun lists in profile bios display increased prevalence, systematic co-presence with other keywords and network tie clustering among US twitter users 2015–2022." *Journal of Quantitative Description: Digital Media* 3. https://doi.org/10.51685/jqd.2023.003.
28. Twenge, J. M. (2013). The evidence for generation me and against generation we. *Emerging Adulthood, 1*(1), 11–16. https://doi.org/10.1177/2167696812466548
29. Twenge, J. M., Abebe, E. M., & Keith Campbell, W. (2010). Fitting in or standing out: trends in american parents' choices for children's names, 1880–2007. *Social Psychological and Personality Science, 1*(1), 19–25. https://doi.org/10.1177/1948550609349515
30. Twenge, J. M., Keith Campbell, W., & Gentile, B. (2013). 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
31. Twenge, J. M., Carter, N. T., & Keith Campbell, W. (2015). Time period, generational, and age differences in tolerance for controversial beliefs and lifestyles in the United States, 1972–2012. *Social Forces, 94*(1), 379–399.
32. UhlsYalda, T., & Greenfield, M. P. (2011). "The Rise of fame: an historical content analysis. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace, 5*, 1.
33. Zhang, T., Kishore V., Wu F., Weinberger K. Q., & Yoav A. (2020). "BERTScore: evaluating text generation with BERT."