Pre-Analysis Plan: Who to Follow? An experiment about discrimination on Twitter[*]

Nicolás Ajzenman Bruno Ferman Pedro Sant'Anna

May 27, 2022

Abstract

We design an experiment to identify discrimination in users' following behavior on Twitter. We create fictitious bot accounts that resemble humans and that claim to be PhD students in economics. The accounts differ on three observable characteristics: gender (male or female), race (black or white) and university affiliation (highly ranked or not). The bot accounts randomly follow Twitter users that are part of the *#EconTwitter* community. We measure how many follow-backs each account obtains after a given period. This allow us to identify if users from this social media community exhibit discrimination in their following behavior in each of the three dimensions we vary in the bot accounts. Finally, we consider if there is heterogeneity in the discrimination among sub-groups of users.

Keywords: Discrimination; Economics Profession; Gender; Race; Social Media.

JEL Codes: A11; C93; I23; J15; J16.

[*] We thank Leonardo Bursztyn, Gregorio Caetano, Fernanda Estevan, Mohsen Mosleh and Vitor Possebom for their helpful comments and suggestions. This research was approved by the Ethical Compliance Committee on Research Involving Human Beings at Fundação Getulio Vargas with a waiver of informed consent (CEPH/FGV, IRB approval n. 034/2022).

Contents

1 Introduction

We design an experiment to identify discrimination in Twitter users' following behavior, focusing on the economics academic community on this social media. We create fictitious bot accounts that resemble humans and that claim to be PhD students in economics. The accounts differ on three observable characteristics: gender (male or female), race (black or white), and university affiliation (top-ranked or lower-ranked university).[1] The accounts randomly follow Twitter users who are part of the academic economics community. After twelve days of activation, we measure how many follows-backs each account obtained.

To define the subjects in the experiment, we obtain data on all Twitter accounts that either posted a status ('tweet') or a retweet containing the term *'#EconTwitter'* between January 1st and February 28th, 2022 – which represents almost 15,000 accounts. Then, we will create 240 fictitious accounts claiming to be PhD students and who differ on their gender, race and university affiliation (overall, we have 30 accounts of each group). Each account will be active for a period of twelve days and follow approximately 100 subjects from the 'EconTwitter' sample. Our main outcome is whether or not each subject followed-back the bot accounts. Given that the accounts are identical in all respects but the three dimensions we varied, differences in follow-back associated with the groups will be due to discrimination. Specifically, we define discrimination as the unequal treatment of persons or groups on the basis of some group-based characteristic (Pager and Shepherd, 2008). In our case, these group-based characteristics are their gender, race or university ranking.

Apart from our main analysis of differences in the probability of follow-back caused by the bot account's gender, race or university ranking, we will also explore possible heterogeneous effects, taking advantage of our large sample of Twitter users. Specifically, using public information from Twitter, we classify our subjects on several characteristics and study if there is heterogeneity on the following-back behavior based on these characteristics. In particular, we study heterogeneities in terms of subject's gender; account quality and concern about the lack of diversity in the profession.

Finally, we also explore possible interactions between bot's gender, race and university ranking. The objective of this exercise is to test whether the discrimination against members of groups who are at the intersection of different group-based characteristics (for instance, women from lower-ranked universities) is different from the discrimination suffered by individuals who are not in this type of intersection.

The rest of this plan proceeds as follows. In section 2, we describe the experiment's design; in section 3, we present the experiment's hypotheses and data; then, in section 4,

[1] We decided to focus on these three dimensions of discrimination for several reasons, but we highlight that there are several other extremely relevant dimensions of disparities both within economics and in a broader context, and that would have been interesting to study. For instance, there has been a recent debate in the United States about how Asian-Americans may be discriminated against in contexts such as college admissions (Arcidiacono et al., 2022). We could also have considered other dimensions, such as discrimination against LGBTQ+ individuals or due to country of origin. We decided to focus on the three dimensions listed before because they are the most salient in the debates about the lack of diversity in the economics profession (see Allgood et al. (2019)) and due to concerns about reducing the power of the experiment by increasing the number of treatment groups.

we discuss the empirical methodology used to identify and estimate discrimination in our setting.

2 Experimental Design

The objective of this experiment is to assess if Twitter users from the "*#EconTwitter*" community display a discriminatory behavior in their decision to follow or not other accounts. To test this hypothesis, we will create Twitter accounts of economics PhD students that vary in 3 dimensions: gender (male or female), race (black or white) and university affiliation (highly ranked university or not). Hence, we will have 8 different types of "profiles". For each type of profile, we will create 30 human-like accounts.

2.1 Making credible profiles

One of the main challenges in this experiment is to create credible Twitter profiles. We want our profiles to be a good representation of a student at the beginning of the PhD (since students in the final years of a PhD tend to be relatively well-known in the economics community, have websites, etc, which would make it more difficult to create a credible profile). To understand more clearly what is an average Twitter profile of a PhD student in economics at the beginning of his or her graduate studies, we analysed the Twitter profiles of first and second year PhD students from three programs that publicly list their students on the program's website. Specifically, we considered the Twitter profiles of the 2020 and 2021 PhD cohorts of Harvard, University of British Columbia (UBC) and Yale. Out of the 59 students in the 2021 cohort (considering the three programs together), we were able to encounter the Twitter profile of 24 (40.67%) students as of January 6th, 2022. Similarly, for the 2020 cohort, we encountered the profile of 18 students out of 56 (32.14%).

Table 1 summarizes the characteristics of these profiles. The median profile of a first-year PhD student in these universities has 94.5 tweets, follows 331.5 accounts, is followed by 152.5 accounts, has a background image, usually of a landscape, does not have a website (over 70% of the profiles do not have a website) and does not publicly include a location. Moreover, the vast majority of the accounts has a profile picture that is a portrait of the person's face (in two cases, the profile picture is a landscape and, in one case, a dog).

Considering these characteristics, we believe that the profiles created for the bot accounts are credible. In terms of profile pictures, we use AI generated images made available by Generated Media Inc. The data-set provided by the company allows to vary only one parameter when choosing images (specifically, we can control the following attributes when creating an image: gender, head pose, age, emotion, skin tone, hair color and length, and whether or not the avatar will be wearing glasses or make-up). This reduces the concern that the images would be significantly different in other dimensions other than the ones we are interested in varying (race and gender).

Moreover, all profiles have a background image (the landscape of the city in which they

Table 1: Summary Statistics of Real-life Twitter profiles of first and second year economics PhD students

	2021 Cohort			2020 Cohort		
	Median	Mean	Std.Dev	Median	Mean	Std.Dev
Tweets	94.5	357.33	917.26	47	174.58	252.40
Following	331.5	471.29	495.88	279	382.32	327.60
Followers	152.5	372.96	507.11	115	401.00	807.61
Website	0.0	0.29	0.46	0	0.26	0.45
Background Image	1.0	0.71	0.46	0	0.47	0.51
Location	0.0	0.42	0.50	1	0.53	0.51
Profile Pic is a Self-Portrait	1.0	0.83	0.38	1	0.89	0.32

Notes: The table shows descriptive statistics of Twitter profiles from first and second-year PhD cohorts from three universities. For the first year (2021) cohort, 24 out of 59 students had profiles we could find (40.67%) as of January 6th, 2022. For the second year (2020) cohort, we could find 18 out of 56 students (32.14%).

claim to be doing their PhD), do not have a website and do not include a location. As discussed above, this is similar to the average profile of a first or second-year PhD student. We also asked a small group of economists and students who have Twitter to follow the profiles, so that all profiles have from the start a certain number of followers. All of this helped making the profiles more credible.

We also created a list of common names in the US and designed a list of possible Twitter bios. The list of names is made up of the most common names and surnames in the United States, excluding names that are race or ethnicity-specific (including Hispanic names and first names disproportionately more likely to be used by whites or by blacks) and names that are gender-neutral.[2] Apart from university affiliation, all Twitter bios contain a field of interest (since most real Twitter profiles we evaluated contained more information than just university affiliation). Moreover, upon the creation of each bot account, the account will tweet a status presenting itself. We also created a list of "presentation" status, and randomly selected the status each bot account would tweet.

During the experimental waves, all active accounts will also randomly retweet from an account that was excluded from our pool of subjects for having a high follows/friends ratio.[3] In some random cases, the retweet will contain some simple text (a "quote tweet"). We will do this as a way of keeping the account active and more authentic (both for other users and for Twitter's algorithm).

The fictional bot accounts we create will differ in three dimensions: gender (male or female), race or skin color (black or white) and university affiliation (highly ranked – or

[2] Specifically, we used the NamSor tool to predict the gender of the names in our list, and excluded those with less than 85% accuracy in the gender prediction. To define race-specific names, we used data from Tzioumis (2018).

[3] "Friends" is the term used by Twitter to designate the accounts you follow.

"elite" – versus lower ranked). The dimensions of gender and race will be signaled by the profile image, which, as already discussed, will be generated using Artificial Intelligence. The university will be informed in the account's Twitter bio. Specifically, to select the universities used in our experiment, we first considered the set of 10 highest ranked universities in the 2017 USNews Ranking of universities in terms of economics graduate programs, and the last ten ranked universities in the same ranking[4] that make public available their list of students.[5] To avoid concerns related to exposing specific universities, out of this set of 20 universities, we randomly selected 5 high-ranked and 5 lower-ranked universities to be used in the experiment. Then, for each bot account, we randomly selected either one of the 5 highly-ranked or one of the 5 lower-ranked universities.

Overall, considering that the fictional profiles are different in three dimensions, we have eight types of profiles. For each type, we will create 30 accounts, so that we end up with 240 accounts. Table 2 summarizes the procedures we used when creating the accounts.

2.2 Sample Selection and Assignment into Treatment

As stated earlier, this study will focus on the "*#EconTwitter*" community on Twitter, so it is natural that the subjects of the experiment will be accounts that are part of this community. Using Twitter's API, we obtained a dataset of all accounts that tweeted or re-tweet statuses containing the term *#EconTwitter* during January and February 2022. We restricted our sample to only unprotected ("public") accounts. Moreover, given that we want to maximize the chance that the subject accounts interact with our bot accounts, we also excluded from the sample all accounts with a follows/friends ratio above 15.[6] This procedure is helpful, since it is likely to exclude institutional accounts, which generally have many followers, but follow only a few users, and also profiles that are too selective in their choice of who to follow. After these procedures, we ended up with a sample of 13,373 subjects. See the Appendix for a complete description of how we obtained our subject pool.

For each of these subjects, we obtained a set of variables using Twitter's API. Specifically, we have information on the number of tweets, number of followers and number of friends (accounts being followed by the subject). We also have information on location for the accounts that choose to let this information public. Moreover, we know whether the account is verified, the number of likes ("favorites") it performed and its date of creation. From the Twitter bio of the subjects, we are also able to infer some other more specific information: we create a dummy variable equal to one if the bio contains the name of one highly ranked university; we also create indicator variables for the user's occupation.

[4] This rank is highly correlated with both the IDEAS/RePEc and the Tilburg university ranking, but has some advantages: first, it is a rank exclusively of US institutions, and focus on universities, not differentiating specific departments within universities. Second, the rank's methodology is based on a survey of academics in peer institutions, so it more accurately represents the perceptions academics themselves have of the universities (by contrast, the other two ranks are based on citations and publications).

[5] We restrict our analysis to universities that make publicly available their list of PhD students because that is a practice taken by all highly ranked universities. Therefore, using universities that don't have public list of students could bias our results.

[6] This is approximately the follows/friends ratio of the 95th percentile of our subject pool.

Table 2: Procedures used to create the bot accounts

Element of Profile	**Procedure**
Profile Picture	Use AI generated images from the Generated Media Inc. database. The dataset allows us to control several parameters when generating each picture: gender, head pose, age, emotion, skin tone, hair color, hair length, glasses and make-up. For each set of four profile pictures, we start from the same "base" face and vary gender (male or female) and skin tone (black or white).
Name	Randomly generated by matching a list of the most common first names and surnames in the US. We exclude from the list all names that are gender-neutral (specifically, we used the NamSor tool to predict the gender of the names in our list, and excluded those with less than 85% accuracy in the gender prediction). We also exclude names that are race or ethnicity specific (to define race and ethnicity specific names, we used data from Tzioumis (2018)).
Bio	The Bio from the bot accounts contains two information: first, the university where they claim to be doing their PhD; second, their research interests. To select the universities, we first considered the ten highest ranked universities and the last ten ranked universities of the USNews rank of graduate universities in economics. We randomly selected 5 universities from each of these two sets. . The interest is also randomly assigned from a list designed by the authors. Generally, the bio from a bot account will be something like: "PhD student at University X. Interested in labor economics and economics of education.". We decided not to use the university's Twitter handles (for instance, @UniversityX) because this would likely affect follow recommendations made by Twitter's algorithm, which could bias the experiment (if Twitter recommended the bot accounts to users from the same university, bots from some universities could get a disproportional volume of followers for reasons unrelated with discrimination). It is harder for the algorithm to target recommendations when the Twitter handle is not used. Importantly, we do not explicitly say in the Bio that the student is doing his or her PhD in economics. This is implicit from the interests listed.
Background Image	A landscape from the city where the student claims to be doing the PhD. We have a single landscape for each city.
Location	The bot accounts' profiles do not include a location.
Website	The bot accounts' profiles do not include an website.
First Tweet and Retweets	The first tweet of the bot account is a personal presentation. It includes the same information as the Bio (university affiliation and research interests), as well as some kind of greeting. We randomly selected the first tweet from a sample designed by us. The first tweet is published a day before the bot account follows the accounts assigned to it. After the first tweet, the bot account will also retweet one status from an user with a follows/friends ratio above 15.
Followers	We asked a group of economic professors and graduate students to follow the bot accounts one day before the bot account follows the accounts randomly assigned to it. We will randomly define the number of followers the accounts will have at the beginning: it will either be lower than 15 or around 100.
Following	One day before following the accounts randomly assigned to it, the bot account will follow all professors and graduate students we asked to follow the account. It will also follow the accounts in our sample that have a follows/friends ratio above 15, which are excluded from the experiment.

Notes: The table summarizes the procedures used to create the bot accounts.

Following the suggestion of Athey and Imbens (2017), we perform block randomization as a way to improve balance. For the block randomization, we will use the following variables: gender (male, female, missing); profession (professor; graduate student; other; missing); number of followers (above or below median). This gives us 24 strata. We will sample randomly from within each strata, assigning the same proportion of users in each strata to each bot account. Specifically, each bot account will be assigned to approximately 100 accounts to follow.[7]. Misfits will be reassigned globally, i.e., we will create a "misfits strata" and sample from there (see Carril (2017)). The bot accounts will always follow the designated accounts on a Thursday (see the Timeline on the Appendix). The follows are done manually to minimize the chance that Twitter considers the accounts' behavior suspicious. At each wave, we will also randomize the order in which we will create the bots and that we will follow the subjects, so there is no concern that a specific type has a timing advantage to receive follow backs.

Apart from following the experimentally assigned accounts in this moment, each bot account will also follow one account from someone who knows about the experiment. This person will then inform us whether they received a notification of the follow. The objective of doing so is to guarantee that the follow is being notified to the users.[8] If an account is shadow-banned, we will simply drop it from the analysis.

We won't allow subjects to be selected in waves that happen less than one month apart, but re-sampling will be permitted between more distant waves.

2.3 Timeline

For each experimental wave, we will activate 8 accounts (one of each type). Within one wave, we will use the following timeline (which is illustrated on the Appendix on Table A.1):

(i) **Day 0:** Creation of accounts according to the procedures described in Table 2. The account posts a tweet presenting itself and re-tweets two posts from accounts from academic journals in economics. The posts are chosen randomly among recent posts from these accounts[9] that already have more than 3 retweets.

(ii) **Day 1:** Each bot account follows the users assigned to it.

[7] There will be some variation in this number to reduce the number of misfits.

[8] On Twitter, a concern we have is with the so-called "shadow-ban". This is a type of punishment Twitter may deploy against users whose behavior on the platform seems suspicious. In practice, what happens is that all activity from a shadow-banned user is "hidden" to other users, including notifications of follows. Therefore, we guarantee that no bot account is shadow-banned before using the results from any experimental wave.

[9] The accounts we sample from to make the re-tweets are: @AEAJournals (journals from the American Economic Association), @ecmaEditors (Econometrica), @JPolEcon (Journal of Political Economy), @QJE-Harvard (Quarterly Journal of Economics), @RevEconStudies (Review of Economic Studies), @restatjournal (Review of Economics and Statistics), @JEEA_News (Journal of the European Economic Association), @EJ_RES (Economic Journal), @JPubEcon (Journal of Public Economics), @nberpubs (National Bureau of Economics Working Papers), @qe_editors (Quantitative Economics), @EconTheory (Theoretical Economics), @J_HumanResource (Journal of Human Resources), @RevOfFinStudies (Review of Financial Studies, @JofFinance (Journal of Finance) and @J_Fin_Economics (Journal of Financial Economics).

(iii) **Day 13:** After twelve days active, we will compute the number of followers for each account and delete them.

Therefore, the experimental waves will have a twelve-days span.[10] We plan on running 30 waves between May 23rd, 2022, and December 20th, 2022. In each wave, we will compute follows twice a day: first at 8 am and then at 8 pm. We call each moment in which we collect the bot's followers a 'period'. Thus, each experimental wave will last 24 periods (or twelve days).

3 Hypotheses and Data

This section presents the outcomes and hypothesis we test in the experiment.

3.1 Primary outcome: Follow-Backs

Our primary outcome of interest an indicator equal to one if a subject experimentally assigned to be followed by one of the bot accounts follows back that account. Another possible outcome would be the total number of follows obtained by the bot accounts at the end of the active period. However, a potential concern with this outcome – total number of follows – is that the decision to follow or not one of the bot accounts may be related to Twitter's algorithm instead of due to a discriminatory behavior. Twitter's algorithm targets following suggestions to users based on "on your activity on Twitter, such as your Tweets, who you follow, and accounts and Tweets that you view or otherwise interact with" (Twitter, 2022). Therefore, it is possible that some of the follows given to a bot account happen simply because the account was suggested by the algorithm to some user.

This is particularly problematic if our bot accounts are target to systematically different users due to some characteristic of the account, and those groups of users display different following behaviors. If this is the case, considering the entire sample of follows could introduce bias in our main analysis, as the selection of treated users would not be random. We tried to minimize the precision of the following suggestion algorithm by reducing the information that it could use – specifically, we do not use university's Twitter handle on bot account's profiles, which makes it harder for the algorithm to recognize the university as a characteristic it could use to suggest the bot to other users. We also guaranteed that the bot accounts have identical characteristics in terms of accounts they follow at first (some accounts of institutions related to economics, such as academic journals, newspapers and multilateral institutions, plus the same set of accounts from colleagues) and that follow them. To the eye of the algorithm, this means that accounts are similar and should be suggested to the

[10] During the pilots, we noticed that all follow-backs happened almost immediately after the bot accounts follow the subject accounts. Out of the 290 follow-backs received in the first two pilots, 80.3% happened before two days since the follow; and 96.9% of them happened before six days. In our longest pilot, which lasted 14 days, a single follow back happened after a week had passed since our follow. In the Appendix A.5, we plot the evolution of follows in our three pilots, showing that they become flat after a few days. For all these reasons we believe that this period of activation is more than sufficient to the study.

same set of users. Therefore, we do not expect, at first, that the bot accounts are suggested to systematically different users by the algorithm[11].

Nevertheless, it is still possible that considering the total number of follows may lead us to obtain biased estimates for the discrimination effect, since the selection of users that see the bot accounts could be no-random. By restricting the analysis to follow-backs, we guarantee that we are comparing groups that were randomly assigned to "find-out" about the bot accounts' existence. We will also report the results for total follows, but keeping in mind that our main analysis is that of Follow-Backs.

Hence, our primary outcome will be an indicator equal to one if each experimentally-assigned subject followed-back the bot account it was assigned to. We are primarily interested in assessing whether or not Twitter users exhibit discrimination on the basis of gender, race and university affiliation. Therefore, our main hypothesis are:

A1 ***(Gender disparities in Follow-Backs).*** *The probability of follow-back of the bot accounts at the end of the active period will be different between bot accounts representing male and female PhD students;*

A2 ***(Racial disparities in Follow-Backs).*** *The probability of follow-back of the bot accounts at the end of the active period will be different on average between bot accounts representing black and white PhD students;*

A3 ***(Rank disparities in Follow-Backs).*** *The probability of follow-back of the bot accounts at the end of the active period will be different on average between bot accounts claiming to be affiliated to top-ranked and to lower-ranked universities.*

Overall, we expect *a priori* the difference in Follow Backs to be consistent with the patterns of discrimination on the basis of gender, racial, and university affiliation observed in previous audit studies and within the economics profession; i.e., we imagine that bot accounts of female PhD students, black PhD students, and affiliated to lower-ranked universities receive less total follows than accounts of male PhD students, white PhD students and students affiliated to top-ranked universities. However, it is possible that we observe an effect in the opposite direction, for example if Twitter users in the "Econtwitter" community are actively trying to engage more with underrepresented minorities, or with students from less prestigious universities.

Note that all the hypotheses above consider exclusively the marginal effect of each group (gender, race and university) on follow backs, but do not take into account possible interactions among these effects. Still, it is possible that the pattern of total follows exhibit intersectionality. Therefore, we will also test the following hypotheses:

[11] We also experimentally observed this by creating two identical accounts who claimed to have different university affiliations. Both accounts posted the same tweet and we observed the reactions, views and follows to these accounts for two weeks. There was no evidence of any differential treatment in terms of impressions (views), which indicates that the algorithm does not target differently when accounts are not very active.

A4 ***(Differences in gender disparities by rank).*** *The gender differential on the probability of follow-back of the bot accounts at the end of the active period will be different on average between bot accounts claiming to be affiliated to top-ranked and to lower-ranked universities;*

A5 ***(Differences in racial disparities by rank).*** *The racial differential on the probability of follow-back of the bot accounts at the end of the active period will be different on average between bot accounts claiming to be affiliated to top-ranked and to lower-ranked universities;*

A6 ***(Differences in gender disparities by race).*** *The gender differential on the probability of follow-back of the bot accounts at the end of the active period will be different on average between bot accounts representing black and white PhD students;*

A7 ***(Differences in gender disparities by race × rank).*** *The gender differential on the probability of follow-back of the bot accounts at the end of the active period will be different on average between bot accounts representing black and white PhD students who claim to be affiliated to top-ranked or to-lower ranked universities;*

We do not have strong priors about the directions of each of the differences described in hypotheses **A4** through **A7**.

3.2 Secondary Outcome: Total Follows

As discussed above, our main outcome of interest is Follow Backs (thus, in our main analysis we restrict the analysis to the experimentally-assigned pairs subjects-bots). However, it is still interesting to consider the effect of gender, race and university affiliation on the Total number of follows received by the bot accounts at the end of the active period. This is relevant for a few reasons: first, the main objective of this experiment is to verify if PhD students belonging to different groups are treated differently on Twitter. This must take into account both follows that happen in response to a follow (the "follow-backs") and follows that happen organically. Moreover, what really matters in terms of a user's experience on Twitter is the total number of followers (this is what will determine, for instance, how many people see the user's posts). Therefore, considering the effect of the groups on total follows is important. In this sense, we will test the same hypotheses listed in the previous section for this secondary outcome. Our priors for these hypotheses are the same as the ones for the main outcome.

4 Research Plan

4.1 Identification and Estimation

4.1.1 Primary Outcome: Follow Backs

As discussed in section 3.2, our primary outcome of interested is follow-backs, i.e., follows that happen after the bot account randomly followed the subjects. Thus, we restrict our analysis to the experimentally assigned pairs users-bot accounts – in other words, instead of considering follows performed by any user on Twitter (including "organic" follows), we consider only the behavior of users that were experimentally assigned to be followed by a bot account.

Our outcome of interest in this section is $Y_{ijst} := \mathbb{1}\{\text{user } i \text{ followed bot account } j \text{ at wave } t\}$, where s denotes the strata user i belongs to. Y_{ijst} is an indicator equal to 1 if the user "followed-back" the bot account, in response to the bot account following the account. Given that the follows from bot accounts were randomly assigned, the causal effects of gender, race and university ranking of bot accounts is identified and can be estimated by comparing the probability of follow back between each group of bot accounts. Specifically, we estimate by OLS the following equation:

$$Y_{ijst} = \alpha + \beta_1 \times RACE_j + \beta_2 \times GENDER_j + \beta_3 \times RANK_j + X_{it}\lambda + \delta_t + \theta_s + \phi_{st} + \varepsilon_{ijst} \quad (1)$$

where $RACE_j$ is a dummy variable equal to one if the bot account represents a black PhD students; $GENDER_j$ is equal to one if the bot account represents a women, and $RANK_j$ is equal to one if the bot account claims to be affiliated with a top-ranked university. X_{it} is a vector of pre-treatment characteristics of subject i's account;[12] δ_t, θ_s and ϕ_{st} represent, respectively, wave, strata and strata × wave fixed effects[13] and ε_{ijt} is the error term. Since our outcome is an indicator variable, we are now considering a linear probability model, so the coefficients β_1, β_2 and β_3 can be interpreted as the difference (in percentage points) on the probability of a bot account being followed back caused by its race, gender, or university affiliation (respectively). Thus, to test hypotheses **A1** through **A3**, we will test whether the coefficients β_1 through β_3 are statistically different from zero.

Apart from the main specification from equation (1), we also estimate similar models with slightly different features, as a way of testing our additional hypotheses and of verifying the robustness of our results:

[12] The variables included in the vector X_{it} are: gender; profession; continent; an indicator for affiliation to a top-ten university or an important association (NBER, IZA or CEPR); indicator for whether the account is verified; indicator for whether the account has a background picture; number of followers; number of friends; number of statuses; and year the account was created. For the categorical variables, we include a category for missing data.

[13] We include strata fixed effects following the suggestion from Bruhn and McKenzie (2009). We also include strata × wave fixed effects to account for possible differences in behavior of subjects from different stratas in differents moments on time. Moreover, note that, among the strata fixed effects, there will be a misfits dummy.

- Adding interaction terms between the variables $RACE_j$, $GENDER_j$, and $RANK_j$. This will allow us to consider the average treatment effect of each of the eight possible "treatments" (considering that there are 8 types of bot accounts), and to consider how these three dimensions interact with each other in the following decision. With this new specification, we will be able to test the hypothesis **A4** through **A7**.
- Verify if the results are robust to the exclusion of the control variables.[14]

4.1.2 Secondary Outcome: Total Follows

Apart from Follow Backs, we are also interested in studying discrimination on the total follows obtained by each bot account at the end of the ten-day active period. Assuming that there is no systematic difference in the Twitter users for whom the bot accounts are suggested to[15], the causal impact of the bot account's gender, race and university ranking is identified and can be studied by comparing the total follows of each group of bot accounts. Specifically, let j denote a bot-account and t represent the wave in which account j was active. Our outcome of interest, Y_{jt}, is the total number of followers obtained by bot account j at the end of experimental wave t. We will estimate the following equation by OLS:

$$Y_{jt} = \alpha + \beta_1 \times RACE_j + \beta_2 \times GENDER_j + \beta_3 \times RANK_j + \delta_t + \varepsilon_{jt} \tag{2}$$

where $RACE_j$, $GENDER_j$ and $RANK_j$ have the same definition as before. Moreover, δ_t represents a wave fixed-effect, and ε_{jt} is an idiosyncratic error term. We interpret the coefficients β_1, β_2 and β_3 as the average effect on the number of follows that are caused by differences in race, gender and university affiliation, respectively. Therefore, testing if each of these coefficients is different from zero is equivalent to testing hypothesis **A1**, **A2** and **A3** respectively (but for the secondary outcome).

We will also estimate equation (2) including interaction terms among the variables $RACE_j$, $GENDER_j$ and $RANK_j$. This will allow us to test hypotheses **A4**–**A7** for the secondary outcome. We will also consider alternative specifications that include base-image fixed effects.

4.2 Inference and Power

We will present standard errors clustered at the bot account level.

Using the results from the three pilot waves we conducted, we can compute the Minimum Detectable Effect of our experiment. The baseline mean follow backs among the three pilots is of 0.167, implying a standard deviation of 0.37 for this outcome. Using this information and considering that we will run 30 waves with 8 accounts each – and each account follows approximately 100 subjects –, we estimate via simulations that our MDE for a power of 80%

[14] Given that we performed block randomization, we do not expect the control variables to enhance precision of estimates significantly.

[15] We discussed this hypothesis on section 3.1. We took several steps to minimize the possibility that Twitter's algorithm suggests accounts differently, and found no evidence that this happened in a pilot.

and a significance level fixed at 5% is of 1.4 percentage points for the marginal discrimination effects. When we add the covariates listed in section 4.1 (using the information from the pilots to estimate the probability that subjects of different types follow the bot accounts) our MDE becomes of 1.3 pp approximately. For the interactions, our power is lower: for the two-way interactions (for instance, the interaction between gender and race) our MDE is of 4.5 pp (4.0 with covariates), while for the three-way interaction (gender × race × university rank) the MDE is of 7.0 pp (6.0 with covariates).

We also perform the simplest assessment proposed by Ferman (2022) to verify if our inference method is reliable. We simulate our data under the null hypothesis of no treatment effects, using Bernoulli draws with parameter equal to the average follow-back rate in the three pilots to input our outcome (follow-back). Reassuringly, we obtained a rate of rejection of the null under a nominal significance level of 5% that was very close to 5%.

4.3 Heterogeneity

We leverage the data we have on characteristics of the subjects to explore possible heterogeneities on the discriminating behavior. Given that we only have this data on our experimental subject pool (not on all Twitter users), we will in principle focus the analysis on this sample. Specifically, we will consider the following heterogeneities:

H1 Gender (male *vs.* female)[16];

H2 Account Quality (above or below median number of followers);

We note that we divide our heterogeneities in two groups: first, in heterogeneity **H1**, we are more interested in studying differences in behavior among groups of subjects. Therefore, we will study this heterogeneities by estimating equations similar to (1) and (2) with interactions between the subjects' and the bots' characteristics, including controls for the remaining characteristics of the subjects (the list of controls is on footnote 12). Specifically, we interact the covariates with the bot's characteristics. This is important because we are interested in verifying if – for instance – male and female subjects with similar accounts and characteristics (apart from their gender) have different behaviors on Twitter. Our prior for **H1** is that women will be more likely to follow female bots and practice less discrimination in the other two dimensions.

For the second group of heterogeneities – heterogeneity **H2**, we are not interested in analysing differences in behavior, but rather differences in the implications of the behavior. We view subjects with many followers as "strong" accounts. Indeed, being followed by someone with many followers will arguably bring more benefits to your Twitter presence than being followed by someone less well-known online. Hence, if subjects with many followers exhibit discrimination, this will be particularly problematic.

[16] Since some Twitter users disclose their preferred pronouns in their bios, we are able to classify some non-binary users as well (those users who indicate the pronouns "they/them"). However, the number of observations with this characteristic is too small to allow for a rigorous analysis, so we restrict ourselves to check heterogeneity between male and female users.

Given that we are interested in analysing implications, our preferred specification to analyse this last heterogeneity will not include covariates. We will also report the results with covariates, but we point out that even unconditional on subjects characteristics, it is relevant to know whether those gatekeepers exhibit discrimination.

Moreover, we highlight that, in interpreting the results for this heterogeneities, we are not necessarily interested in comparing the degree of discrimination exhibited by each of the two groups (above median followers *vs.* below median followers). We are merely interested in testing if subjects with many followers exhibit discrimination in their behavior. If they do (even if this discrimination is comparable or lower than that of the other subjects), we would interpret this as evidence that those with more power (either in academia or online) are being discriminatory.

References

Allgood, Sam, Lee Badgett, Amanda Bayer, Marianne Bertrand, Sandra E Black, Nick Bloom, and Lisa D Cook, "AEA Professional Climate Survey: Final Report," Technical Report, American Economic Association's Committee on Equity, Diversity and Professional Conduct 2019.

Arcidiacono, Peter, Josh Kinsler, and Tyler Ransom, "Asian American discrimination in Harvard admissions," *European Economic Review*, 2022, *144*, 104079.

Athey, Susan and Guido W Imbens, "The econometrics of randomized experiments," in "Handbook of economic field experiments," Vol. 1, Elsevier, 2017, pp. 73–140.

Bruhn, Miriam and David McKenzie, "In pursuit of balance: Randomization in practice in development field experiments," *American economic journal: applied economics*, 2009, *1* (4), 200–232.

Carril, Alvaro, "Dealing with misfits in random treatment assignment," *The Stata Journal*, 2017, *17* (3), 652–667.

Ferman, Bruno, "Assessing inference methods," *arXiv preprint arXiv:1912.08772*, 2022.

Hridoy, Syed Akib Anwar, M Tahmid Ekram, Mohammad Samiul Islam, Faysal Ahmed, and Rashedur M Rahman, "Localized twitter opinion mining using sentiment analysis," *Decision Analytics*, 2015, *2* (1), 1–19.

Pager, Devah and Hana Shepherd, "The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets," *Annu. Rev. Sociol*, 2008, *34*, 181–209.

Sebo, Paul, "Performance of gender detection tools: a comparative study of name-to-gender inference services," *Journal of the Medical Library Association: JMLA*, 2021, *109* (3), 414.

Twitter, "About Twitter's account suggestions," `https://help.twitter.com/en/using-twitter/account-suggestions` 2022. Accessed: 03-04-2022.

Tzioumis, Konstantinos, "Demographic aspects of first names," *Scientific data*, 2018, *5* (1), 1–9.

Appendix

A Additional Information

A.1 Procedure to obtain the Subject Pool

(i) From January 1st, 2022, to February 28th, 2022, obtain all Twitter users that either tweeted or retweeted a status containing the term *“#econtwitter”*[17] → 14,449 accounts.

(ii) Remove accounts that no longer exist, accounts that are clearly bots, and protected accounts[18] → 14,055 accounts.

(iii) Compute follows/friends ratio for the remaining account. Remove accounts with a follows/friends ratio above 15 and accounts with less than 10 friends and institutional accounts → 10,226 accounts[19]. This is our final subject pool.

A.2 Experimental Wave Timeline

Table A.1: Wave Timeline

Monday	**Tuesday**	**Wednesday**	**Thursday**	**Friday**	**Saturday**	**Sunday**
...	Create Accounts Presentation Tweet	Follow Friends Rt (x2)	$d = 0$ Follow subjects $d = 1$	$d = 2$ $d = 3$	$d = 4$ $d = 5$	$d = 6$ $d = 7$
Monday	**Tuesday**	**Wednesday**	**Thursday**	**Friday**	**Saturday**	**Sunday**
$d = 8$ $d = 9$	$d = 10$ $d = 11$	$d = 12$ $d = 13$	$d = 14$ $d = 15$	$d = 16$ $d = 17$	$d = 18$ $d = 19$	$d = 20$ $d = 21$
Monday	**Tuesday**	**Wednesday**	**Thursday**	**Friday**	**Saturday**	**Sunday**
$d = 22$ $d = 23$	$d = 24$ Delete Accounts	...	...	...	...	...

Notes: The table shows the timeline of an experimental wave. Accounts will be active for 12 days after following the experimental subjects. Each wave starts on a Tuesday, so there will always be accounts from two waves active in the same period (8 of them in the second week and 8 in the first week of the wave).

[17] The search considered all variations of capital and small letters for the term.

[18] Note that an account that tweeted a status containing “#econtwitter” at the beginning of January, for instance, may no longer exist at the beginning of March (the account owner may have deleted the account). We identify accounts that are clearly bots by analysing the accounts’ Twitter bios. In Twitter, “Protected” accounts are the ones that choose not to be public, restraining their information and interaction to the account’s friends.

[19] 15 is approximately the follows/friends ratio of the 95th percentile of the subject pool sample after step (ii). We removed accounts with too few friends because those accounts are likely to be inactive or (at least) are extremely unlikely to follow an unknown account.

A.3 Description of subject-level variables

Table A.2: Description of variables at the subject level

Variable	Description	N (%)
Gender	Whether the account belongs to someone identified as male or female. To obtain this information, we use the information on the first name of the user to predict its gender, using the **NamSor** tool,[1] which accurately predicts gender based on full names. We only considered predictions done with above 90% accuracy, and assigned as missing the gender information for the accounts with accuracy below this threshold. We manually checked a randomly selected subsample of 100 accounts, and obtained 98% accuracy.	8,316 (58.4%)
Nfavorites	Number of tweets marked as "favorite" (i.e., "liked") by the user.	14,055 (100%)
Nfollows	Number of accounts the user follows.	14,294 (100%)
Nfriends	Number of friends the user has, i.e., number of accounts that follow the user.	14,055 (100%)
Verified	Indicator variable equal to one if the account is verified, a "badge" provided by Twitter to signal that the account is authentic.	14,055 (100%)
Continent	The continent in which the user lives. We obtain this information via the "location" information from Twitter. This information is given by the user and can, in principle, be anything (it does not have to be a real location and does not have to be correct). We classify the "real" location given by region: North America, South and Central America, Europe, Asia, Africa, Oceania. If a person indicates more than one place from different continents, we classify location as missing. At the end of the procedure, we manually checked a random subsample of 100 accounts and obtained 100% accuracy.	8,931 (62.7%)

Continued on next page

Table A.2: Description of variables at the subject level (Continued)

Variable	Description	N (%)
Profession	The user's profession. We classify professions using the user's account description (or "bio"). The list of professions/areas of work is: professor (which is subdivided into "assistant", "associate" and "other); PhD student; Post-Doc; Other academic position (for instance, Research Fellow, Research Assistance, etc.); Industry/Tech; Government; Non-profit/Multi-lateral Organization; Journalist. We first search for keywords related to each profession, and then manually verify the matches. At the end of the procedure, we checked a random subsample of 100 accounts and obtained 99% accuracy.	8,555 (60.1%)
University Affiliation	Indicator variable equal to one if a user is affiliated to a highly ranked university. To obtain this information, we also consider the user's account description ("bio") and search for keywords associated with the highly ranked universities. We obtain this variable for top-ten and top-twenty US universities according to the USNews Ranking.	14,055 (100%)
Affiliation to Prestigious Economics Institution	Indicator variable equal to one if a user is affiliated with the National Bureau of Economics Research (NBER), the Centre for Economic Policy Research (CEPR), the The Abdul Latif Jameel Poverty Action Lab (JPAL) or the IZA Institute of Labor Economics. As with the other variables, we obtain this one by searching for these keywords in the user's account description ("bio")	14,055 (100%)
Year of Account Creation	The year in which the account was created. This information is provided by Twitter's API and is therefore precise.	14,055 (100%)
Profile Picture	Indicator variable equal to one if the user has a profile picture.	14,055 (100%)
Background Picture	Indicator variable equal to one if the user has a background picture (banner).	14,055 (100%)

[1] We chose this tool for a few reasons: first, it has already been used in academia, including to predict names using Twitter data (e.g., Hridoy et al. (2015)); second, it has been shown to be more or equally accurate than similar tools (Sebo, 2021); third, its database includes names from a variety of countries, and allows the analysis of full names.

Notes: The table lists and describes the variables obtained for the users in the subject pool. The column N (%) shows the number of accounts and the percentage out of the total pool for which we were able to obtain each information.

A.4 Descriptive statistics of subjects

Figure A.1: Number of New Accounts Created by Year Among the *#EconTwitter* Community

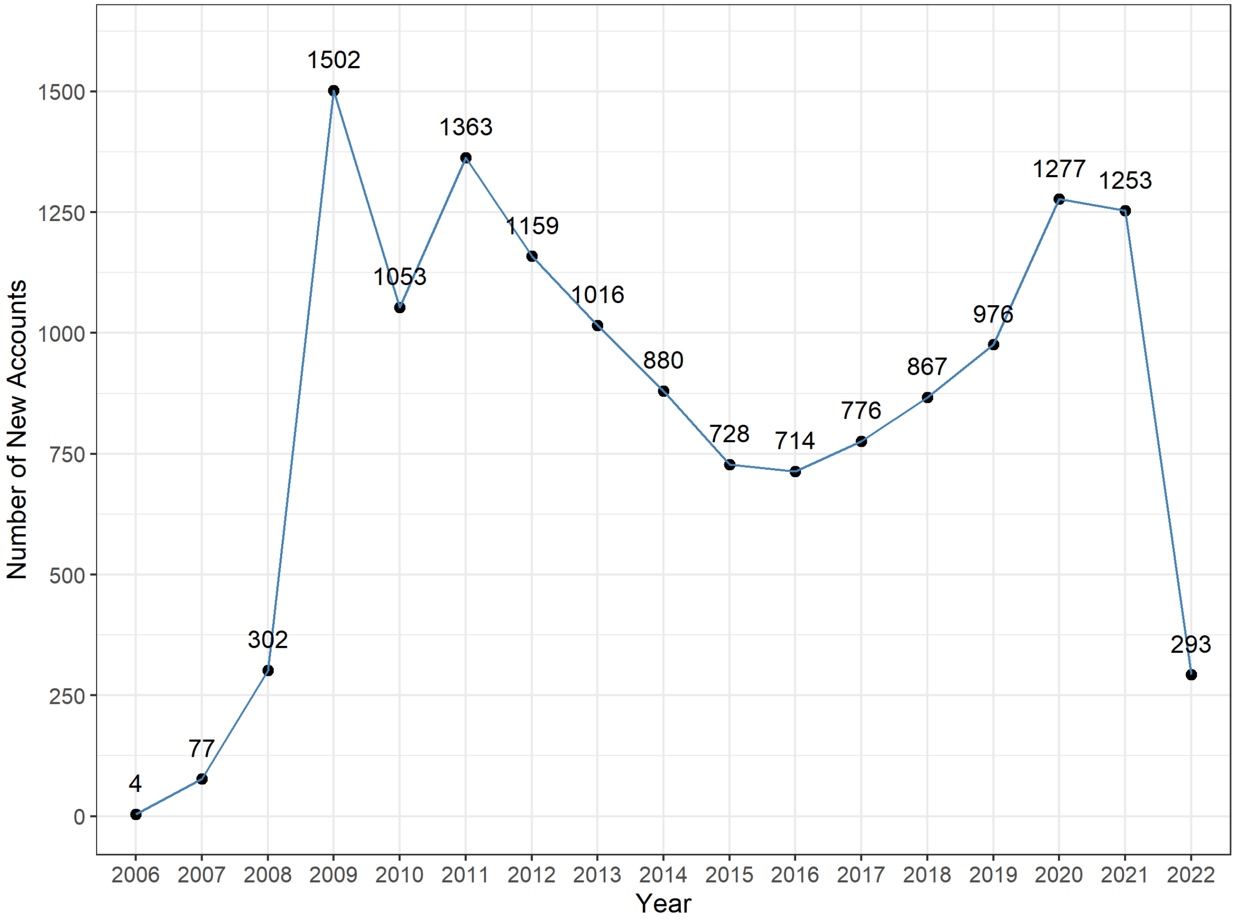


Table A.3: Descriptive Statistics of the Subject Pool - Qualitative Variables

Variables	% Classified	N	%
Gender	58.68		
Female		2,200	26.67
Male		5,976	72.45
Non-binary		72	0.87
Continent	63.05		
Africa		344	3.88
Asia		793	8.95
Europe		3,443	38.86
Latin America		612	6.91
North America		3,492	39.41
Oceania		177	2
Profession	60.45		
Professor		2,911	34.26
Assistant Prof.		627	7.38
Associate Prof.		301	3.54
Undefined Prof.		1,983	23.34
Government		426	5.01
Industry/Tech		1,121	13.19
Institution		1,021	12.02
Journalist		221	2.6
Non-profit/Multilateral Org.		269	3.17
PhD Student		1,156	13.61
Post-Doc		272	3.2
Other Researcher		1,099	12.94
Affiliated to top-ten university	100		
No		13,436	95.6
Yes		619	4.4
Follows Twitter account(s) addressing diversity in economics	100		
No		11,926	84.85
Yes		2,129	15.15
Verified	100		
No		13,684	97.36
Yes		371	2.64
Has background picture	100		
No		3,367	23.96
Yes		10,688	76.04

Notes: The table shows the distribution of gender, continent, and profession from our subject pool. The procedure to obtain each variable is described on Table A.2.

Table A.4: Descriptive Statistics of the Subject Pool - Quantitative Variables

Variables	Mean	Std. Deviation	Min	Max	Obs.
Number of followers	3,958.06	37,378.96	0	2,437,589	14,055
Number of friends	1,245.91	2,477.13	0	113,267	14,055
Number of statuses ('tweets')	22,067.7	83,014.79	0	2,696,665	14,055
Number of favourites ('likes')	21,361.06	62,001.76	0	1,250,869	14,055

Notes: The table shows summary statistics for the sample of experimental subjects – all accounts that tweeted or re-tweeted the term *EconTwitter* between January 1st and February 28th, 2022.

A.5 Evolution of Follow Backs

In this appendix, we give additional information from our pilots to justify our decision to run relatively short experimental waves (with a time span of six days). Before the experiment, we ran three pilots with the following characteristics:

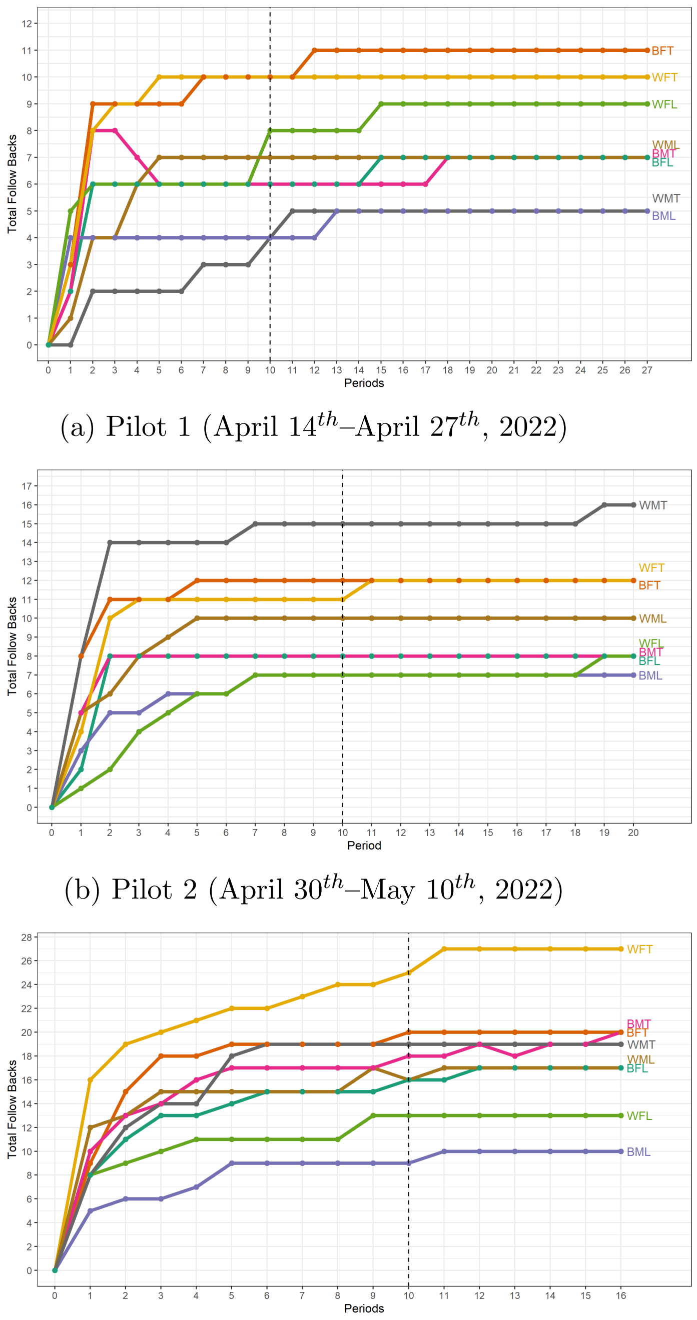
1. **Pilot 1:** eight bot accounts (one of each group), 14-day span;
2. **Pilot 2:** eight bot accounts (one of each group), 10-day span;
3. **Pilot 3:** sixteen bot accounts (two of each group), 10-day span.

In these pilots, we followed the subjects in two moments: the beginning of the wave (period zero) and the middle (after period 14 in the first pilot and after period 10 in the other two). Figure A.2 shows the evolution of follow backs in each of the three pilots considering the subjects that were followed in the beginning of each pilot (i.e., in period zero). We see that follow-backs reach a plateau after period 10 (five days). Specifically, of all 288 follow-backs received from subjects followed in period 0, only 16 (5.56%) happened after the tenth period.

Considering all accounts that followed-back one of the bots (a total of 434 subjects across the three pilots), we have that 79.03% of the follow-backs happened within 24h of the moment in which the bot followed the subject; 83.89% happened 2 days or less after the follow, and 95.62% after 5 days or less. Therefore, the overwhelming majority of follow-backs happen within the first few days after a follow.[20]. In our first pilot, we kept the accounts active for two weeks (fourteen days), and not a single follow-back happened more than a week after the subject was followed. Therefore, we believe that a period of twelve days is more than enough to study the follow back behavior of our subjects.

[20] In particular, we always follow the subjects on Thursdays, so that a period of 12 days includes two weekends.

Figure A.2: Evolution of Follow Backs - Pilots



(a) Pilot 1 (April 14^{th}–April 27^{th}, 2022)

(b) Pilot 2 (April 30^{th}–May 10^{th}, 2022)

(c) Pilot 3 (May 14^{th}–May 24^{th}, 2022)

Notes: The figures show the evolution of Follow Backs in the three pilots we ran, for each type of fictional account. The types are written as Race-Gender-Rank, where B and W stand, respectively, for "Black" and "White"; F and M are, respectively, "Female" and "Male"; and T and L stand for "Top-Ranked" and "Lower-Ranked" universities, respectively. For example, WMT is a bot account representing a white, male PhD student from a Top-Ranked university. The dashed vertical line represents the moment we will de-activate the accounts in our actual experiment.