Rooting for the same team: Shared social identities in a polarized context*

Nicolás Ajzenman† Bruno Ferman‡ Pedro C. Sant'Anna§

This version: March 21, 2025
First draft: January 16, 2023

Abstract

Can shared social identities help overcome online political divides? We investigate this question through a field experiment with 4,620 unique Twitter users conducted over six months during the 2022 Brazilian elections. Although both political congruence (supporting the same candidate) and social non-political congruence (rooting for the same football team) increase follows and reduce blocks, the positive effect of shared social identity weakens substantially when political identity information becomes available. The effect of political congruence remains strong even after the election and is unaffected by the Brazilian national team's positive results during the 2022 FIFA World Cup, despite the team being a quintessential national symbol. Text analysis of live-streamed tweets of Brazilian nationals during the tournament suggests that this shared national experience failed to reduce political polarization in our setting because polarization had extended to the players themselves. Overall, our results indicate that political polarization can undermine the potential of other shared identities to reduce political divides and foster social cohesion.

Keywords: Social Media; Social Identity; Affective Polarization; Brazilian Elections.
JEL Codes: D72; D91; C93; Z20.

*We thank Filipe Campante, Fernanda Estevan, Claudio Ferraz, Thomas Fujiwara, Lorenzo Lagos, Horacio Larreguy, Mario Macis, Mohsen Mosleh, Marcos Nakaguma, Ben Olken, Frank Schilback, Romain Wacziarg, Ekaterina Zhuravskaya, and seminar participants at Brown, CEPR/Warwick/Princeton/Yale Polecon Symposium 2023, EEA, Exeter, FGV-EESP, Insper, JPAL-SEA, MIT Behavioral Lunch, Monash, NEUDC 2023, NUS, NOVA SBE, NYU, PUC-Chile, Queens, UCSD, UofM, UofT, Universidad de San Andres, Universidad Torcuato Di Tella, and Warwick for their helpful comments and suggestions. We are grateful to Livia Haddad, Luis Lins, and Nícolas de Moura for superb research assistance. This research was approved by the Ethical Compliance Committee on Research Involving Human Beings at Fundação Getulio Vargas (CEPH/FGV, IRB approval n. 208/2022). The experiment was pre-registered at the AEA RCT Registry under ID AEARCTR-0009982.

†McGill University. E-mail: nicolas.ajzenman@mcgill.ca
‡São Paulo School of Economics - FGV. E-mail: bruno.ferman@fgv.br
§MIT Department of Economics. E-mail: p_stanna@mit.edu

1 Introduction

Political identity has recently become a crucial divisive cleavage (Iyengar et al., 2019; Boxell et al., 2022), negatively affecting interpersonal relations (Huber and Malhotra, 2017; Chen and Rohla, 2018), democratic norms, and social cohesion (Iyengar et al., 2019). Digital technology is frequently pointed out as an environment that amplifies this societal divide through echo chambers (Sunstein, 2001, 2018). More than 70% of young Americans consume news daily from social media platforms (American Press Institute, 2022) and a massive proportion use social media as their main source to consume political news; in such a context, sorting in terms of political preferences in social media (commonly called homophily) may escalate misinformation sharing (Del Vicario et al., 2016), affect the type of content and news consumed by these individuals (Levy, 2021; Halberstam and Knight, 2016) and reduce exposure to dissenting views (Bursztyn et al., 2022; Zhuravskaya et al., 2020), potentially reinforcing polarization and affecting even offline social cohesion (Enikolopov et al., 2024).

While this phenomenon stems from the fact that individuals increasingly consider their political preference as a core element of their social identity (Huddy et al., 2015; Van Bavel and Packer, 2021), there are several other non-political identities with which individuals can identify (Tajfel and Turner, 1986; Akerlof and Kranton, 2000). A recent literature (Voelkel et al., 2024) posits that, when shared, these common identities may be important to reduce partisan sorting and, more generally, ingroup biases (Depetris-Chauvin et al., 2020). A crucial question is whether the cohesive power of sharing other common identities is enough to soften partisan homophily; or whether political identities are so strong as to overshadow those other identities, thus preventing the formation of social ties in social media that otherwise could have flourished. This paper examines these questions through experimental and observational methods implemented on the social media platform Twitter (now X) in the highly polarized context of Brazil (Ortellado et al., 2022; Wagner, 2021).

We explore two different types of non-political identities with plausible cohesive power: rooting for a Brazilian football club, and rooting for the Brazilian football national team, which is a relevant national symbol, during a high-stakes tournament such as the World Cup. First, in our experimental analysis, we study the interplay between congruence in political identity and congruence in preference for a Brazilian football club in forming social ties among Twitter users (follow-back and block rates).[1] We conducted a pre-registered trial on Twitter in the second semester of 2022, before, after, and during the 2022 Brazilian presidential election campaign.[2] We created fictional accounts that signaled their preferred candidate in this election (either Luiz Inácio Lula da Silva or Jair Messias Bolsonaro, the two

[1] We use the term "football" instead of "soccer" to refer to the sport "association football". This is the usual practice in most of the literature in social sciences studying this sport in the Brazilian context. Notably, football in Brazil is characterized by clubs with historical rivalries, and the set of supporters of rival clubs creates a division in society that is relatively uncorrelated with political preferences or other societal divides such as income levels (Ronconi, 2022). We interpret preference for a club as a social non-political identity, as football is a crucial element of Brazilian culture (DaMatta, 1994)

[2] AEA RCT Registry ID AEARCTR-0009982. The experiment 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. 208/2022).

candidates that have been the symbols of opposite sides of the political spectrum in Brazil in the last few years) and their preferred football club. We call the non-political identity "affective identity" in opposition to the political one. We also created neutral accounts in one of the two dimensions to study the effect of shared identity without conditioning on the other one. The accounts then randomly followed Twitter users with congruent and non-congruent identities across these two dimensions (political and affective). Our sample comprises 4,620 politically active Twitter users, who we followed over 43 waves in the second semester of 2022 for a total of 30,194 observations.[3] Finally, we computed the proportion of follow-backs and blocks each fictional account received as measures of social ties between Twitter users and our experimental accounts.

We document two main experimental results. First, using accounts that signal a single dimension of identity (either affective or political), we find that identity congruence increases the probability of follow-backs and reduces the probability of blocks. Sharing affective identity causes an increase of 13.4 percentage points in follow-backs (or a 58.5% increase) relative to the case of opposite affective identities, while sharing political identity increases follow-backs by 20 pp (or a 119% increase) relative to opposite political identities. In addition, sharing affective identity decreases the probability of blocks by 1.4 pp compared to supporting opposite clubs, and sharing political identity decreases the probability of blocks by 12.3 pp compared to preferring opposite candidates.

Second, using the experimental accounts that signal both dimensions of identity, we find that, although both dimensions are relevant to forming ties, the political dimension overshadows a large part of the positive effects of sharing an affective identity. Indeed, when a fictional account is politically neutral, sharing affective identity increases the likelihood of follow-backs by 13.4 pp, but only by 4.3 pp when the fictional account and subject politically disagree and 8.5 pp when they agree. These results indicate that political identities can partially undermine social connections that could have been formed due to other shared identities in a context of intense political polarization. While disclosure of political preferences weakens the effects of sharing an affective identity, sharing affective identity still fosters (some) ties among counter-partisans and is particularly powerful in preventing blocks. Together, these results suggest that political identities significantly undermine the cohesive potential of shared non-political identities, yet these other identities still play a role in reducing cross-partisan animosity.

We then explore the fact that our experiment was conducted over six months with several repeated waves to study how natural shifts in the salience of elections impact the formation of cross-partisan ties on Twitter. When comparing cross-partisan follow-back rates before and after the official electoral period, we find a small but significant decrease in political homophily, consistent with a decrease in the salience of the political social identity after the electoral period.

Furthermore, we investigate how shared collective experiences surrounding the Brazilian national team's World Cup performance, a national symbol distinct from team allegiance, influence social tie formation with counter-partisans on Twitter. The World Cup timing,

[3]As explained in the Experimental Design Section, the same subject could be treated more than once, but never in subsequent experimental waves.

beginning shortly after the elections, created a natural shift in public discourse from partisan identities to a shared national identity. This timing provided an opportunity to test whether cross-partisan connections increased in response to shared collective experiences: moments that could unite fans in celebration (when Brazil won matches or advanced further in the tournament) or in shared disappointment (when Brazil lost matches and was eliminated from the competition), consistent with what was previously observed in a different context by Depetris-Chauvin et al. (2020). While political homophily decreased slightly after the election, our analysis reveals no significant impact of these World Cup-related experiences— whether moments of collective celebration or shared disappointment—on cross-partisan follow-back rates.

Our results suggest more limited potential for non-political shared identities to reduce political segregation compared to studies like Voelkel et al. (2024). Even a powerful shared national experience like the World Cup proved largely ineffective in reducing political homophily, contrasting with Depetris-Chauvin et al. (2020)'s findings regarding the bridging of ethnic divisions in Africa. This raises important questions about why shared national identity failed to foster cohesion in this context. Anecdotal evidence worldwide suggests that national symbols may lose their power to foster nationalism as they become politicized. In a good example, President Trump claimed that "Americans were happy" about the women's soccer team defeat against Sweden in 2021—which, he affirmed, was a consequence of their "wokeness," a position that echoed that of many conservative sectors.[4] We hypothesize that political polarization had become so intense in our setting that it permeated national symbols (specifically, the national football team), diminishing their unifying potential.

To test this hypothesis, we collected live Twitter data during the World Cup, taking advantage of the fact that some Brazilian players had publicly expressed political preferences, which could weaken the identification of the team for those with opposing political views. We analyzed Brazilian nationals' Twitter reactions to World Cup events affecting the national team (both positive events like goals and negative ones like injuries), examining how these reactions varied based on political alignment between supporters and the specific players involved. Our findings indicate that Brazilians' responses to both positive and negative events were strongly influenced by their political alignment with the players involved. A striking example of this phenomenon was the partisan divide in reactions to a severe injury suffered by Brazil's star player Neymar: while politically aligned fans expressed distress, those with opposing political views celebrated the injury. In Section 6.2, we demonstrate that this example reflects a systematic pattern in Brazilian supporters' behavior during the World Cup rather than an anecdotal example.

Together, the experimental and observational results indicate that political polarization can overshadow the potential of shared non-political identities to enhance cohesion and foster ties. Our results contribute to several strands of the literature.

First, our paper relates to the literature on echo chambers in social media (Zhuravskaya et al. (2020). Digital technology is frequently pointed out as part of the cause of polariza-

[4]See https://www.vox.com/22600500/olympics-conservatives-simone-biles-anti-american and https://www.washingtonpost.com/sports/interactive/2024/american-sports-grievance-culture/. Accessed in February 2025.

tion (Gentzkow, 2016), as well as an environment that amplifies polarization through echo chambers (Sunstein, 2001, 2018). While many authors highlight that echo chambers can be created by algorithms' recommendations (Epstein and Robertson, 2015), our results suggest that, in part, echo chambers are created by individuals choosing to sort with those who share their identities, even among individuals that are congruent in terms of non-political identities (and that thus could have formed ties in the absence of political polarization). This sorting in terms of political preferences may have implications on the type of content and news consumed by these individuals (Levy, 2021; Halberstam and Knight, 2016) and on how likely it is for individuals to be exposed to dissenting views (Bursztyn et al., 2022), potentially increasing polarization.

Second, we contribute to the literature on intergroup animosity and the role of shared identities in reducing it. Prior work suggests that shared identities can play a vital role in mitigating partisan divisions and reducing intergroup prejudices (Voelkel et al., 2024). More specifically, an important strand of this literature focuses on how contact through sport can foster cohesion between conflicting groups (Lowe, 2021; Mousa, 2020). We are more closely related to two papers in this line. First, Depetris-Chauvin et al. (2020) show that individuals in Sub-Saharan Africa are more likely to identify with their nation than their ethnic group following important victories of their national football teams. Second, Ronconi (2022) finds that, in the days following a match between rival football clubs in Latin America, social cohesion tends to improve for those in regions where the match is relevant (not only for football fans), except when players behave violently. We contribute to this literature by studying the interplay between political identity and football club preference. Our result shows that, in a context of intense political polarization, the positive effects of sharing a football-related affective identity on cohesion could be severely weakened. Moreover, the observational evidence from Twitter during the World Cup (a setting more akin to Depetris-Chauvin et al., 2020) suggests that even the identification with the national football team may have limited power to increase social cohesion in a polarized setting if polarization also permeates the players.

In a more optimistic interpretation of our results, the fact that sharing an affective identity can still have a positive effect, even if relatively small, in the formation of ties emphasizes the importance of sports (football in particular) in fostering integration, in line with Depetris-Chauvin et al. (2020). It also indicates that there is demand (albeit small) for cross-partisan interactions, sometimes pointed out as a potential strategy to reduce political polarization (Santoro and Broockman, 2022), when subjects share football interests. Thus, this result invites us to think of ways to make these commonalities more salient in order to increase their power to build cohesion and reduce polarization (Hartman et al., 2022).

Our paper also relates to the literature on affective polarization, the extent of out-group animosity and in-group favoritism based on political preferences (Iyengar et al., 2012, 2019). While most of the literature on this topic uses surveys to measure polarization (e.g., Iyengar et al., 2012; Boxell et al., 2022; Wagner, 2021; Reiljan, 2020), we interpret our metric as a measure of affective polarization in a natural setting, providing a behavioral, revealed-preference measure of this phenomenon. Similar to our case, some papers documented affective polarization using behavioral measures in different contexts: online dating (Huber

and Malhotra, 2017), family gatherings during holidays (Chen and Rohla, 2018), and connections on social media (Mosleh et al., 2021). In particular, our paper is closely connected with Mosleh et al. (2021), who use a similar methodology to ours to study the effect of shared partisanship in the US on the formation of ties on Twitter. Our contribution relative to that paper is threefold. First, our paper explores the interplay between political and non-political identities. This is particularly relevant as it advances our understanding of how and to what extent political identity overshadows other dimensions of shared identity. Second, by measuring follow-backs and blocks, we can separate the two dimensions of affective polarization: willingness to form a tie with a congruent account (akin to in-group favoritism) and willingness to impede a potential tie with an incongruent account (akin to out-group derogation).[5] Third, by conducting the experiment over time with a large sample, we can study the extent to which patterns of social connections change in response to exogenous changes in the salience of social categories.

The remainder of this paper is organized as follows. In Section 2, we provide relevant background on polarization in Brazil, football, and Twitter, focusing on information relevant to the understanding of the experimental design and its results; then, in Section 3, we present our conceptual background, drawing upon Social Identity Theory; in Section 4, we detail our experimental design and empirical strategy; then, in Section 5 we present our experimental results; finally, in Section 6, we document how cross-partisan interactions changed over time, focusing on explaining why the shared experience of the World Cup did not significantly reduce political homophily.

2 Background

2.1 Political Polarization in Brazil

For many analysts, Brazil's democracy is currently "caught up in the sharpest and most polarizing moment in its history" (Kingstone and Power, 2017). In 2022, Brazilian citizens chose between Jair Bolsonaro (the right-wing incumbent) and Luis Inácio Lula da Silva, the country's former president and member of the Workers Party. The two candidates obtained over 90% of valid votes in the first election round—for comparison, in the three previous presidential elections, the two most voted candidates obtained less than 80% of votes in the first round. Moreover, in the 2022 election, the distance in valid votes between the two candidates was less than two percentage points in the run-off election, also much closer

[5] Our result of political incongruence overshadowing other affective identities is consistent with Chen and Rohla (2018), who show, using anonymized data from cellphones, that Thanksgiving dinners attended by individuals from opposing-party precincts were shorter on average than same-party dinners in 2016 (after the presidential election in the US), suggesting that political incongruence can even affect family cohesion. A challenge in ascribing a causal interpretation to the pattern they document is that partisan mismatch could be correlated with other individual characteristics that might lead to shorter gatherings. By creating experimental accounts that are identical apart from their political identity signal, we can more clearly study the causal effect of political mismatch on undermining affective ties derived from other shared identities. Moreover, by comparing accounts that do not signal political identity with accounts that do, we can isolate the effect of political polarization in overshadowing other identities.

than in previous elections.[6] Polarization also manifested itself in violence between counter-partisans. During the 2022 presidential campaign, at least three cases of politically-motivated homicides involving common citizens were reported.[7]

To give a sense of the level of affective polarization currently experienced in Brazil, we use data from the Brazilian Electoral Study (BES), a nationally representative post-electoral survey part of the Comparative Study of Electoral Systems project. Following Boxell et al. (2022), we measure affective polarization among those who report identifying with a party as the distance between the affect towards this party and all other parties and show results in Appendix Figure B.1. Consistent with analysts' views, affective polarization in Brazil seems to have reached a new high since 2018. Boxell et al. (2022) provide measures of affective polarization in the United States and other OECD countries, which allow us to compare polarization in this dimension in Brazil with that in other settings. The mean level of affective polarization in Brazil in 2022 (59.1) is comparable to (and even slightly greater than) that of the United States in 2020 (56.3) and higher than that of countries such as France (52.6 in 2017), Canada (37.7 in 2020), and Germany (28.5 in 2018). Moreover, Brazil experienced a positive trend in affective polarization, smaller in magnitude than the US (which has an estimated slope of 0.56) but comparable to France.

Therefore, Brazil has recently experienced an increase in affective polarization. Furthermore, Brazil's affective polarization level is comparable to that of the US and Latin American countries but greater than some OECD countries. This polarization pattern can impact the formation of social ties among Brazilians of opposite ends of the political spectrum, which we will study in this paper. Moreover, while survey-based indicators of affective polarization can be informative, they have several limitations as they can be susceptible to intentional exaggeration (Iyengar et al., 2012). In contrast, measures of polarization based on follow-backs and blocks in a real-world setting provide behavioral metrics of affective polarization in a natural environment.

2.2 Football

Football is by far the most popular sport in Brazil. For instance, 65% of the country's population claim to be interested in this sport (Nielsen Sports, 2022). An even larger fraction of the population claims to support a football team: 73.1% of the Brazilian population (85.1% of men and 62.5% of women) support a football club (IPEC and O Globo, 2022).

The fact that a larger fraction of the Brazilian population claims to support a football club than to be interested in the sport suggests that football has a distinctive role in Brazilian society. Indeed, more than being a mere entertaining or recreational activity, football is a fundamental and constitutive element of Brazil's national identity (Murad, 1995). Many anthropologists and sociologists have pointed out that a football club is an important element of an individual's identity: DaMatta (1994) argues that, in the process of socialization in Brazil, there are "complex ties that entangle us [Brazilians] to a football team (...), recreating

[6] Lula was elected with 50.90% of valid votes, against 49.10% for Bolsonaro in 2022.

[7] See, for instance, Reuters (09-09-2022), CNN (07-11-2022) and BBC News Brasil (10-05-2022).

in a modern level the idea of family as a community (...) that is chosen voluntarily" (see also DaMatta, 1982). This constitutive role of football in Brazilian society manifests itself not only in a positive way (e.g., by fostering a sense of community) but also negatively, as episodes of football-related violence are not uncommon in the country.[8] Hence, a preferred football club is a relevant dimension of social identity in Brazil. This central role of football in identity is not exclusive to Brazil but is also common in many other Latin American countries (Alabarces, 2003).

Football in Brazil is characterized by teams with traditional rivalries (Ronconi, 2022). Those rivalries are usually constituted historically and create a sense of antagonism between clubs and, by extension, supporters of those clubs. Furthermore, most rivalries are between clubs from the same region of the country; for instance, some famous rivalries are those between Palmeiras and Corinthians (from the city of São Paulo) or between Flamengo and Vasco (from Rio de Janeiro).

A relevant feature of those rivalries is that the characteristics of club supporters are relatively uncorrelated with other societal divides such as income, gender, or political affiliation. Indeed, Appendix Figure B.2 shows that supporters of the six most popular Brazilian clubs and their rivals are mostly similar in terms of age, gender, race, education, income, and religion. Crucially, there is no case of a club whose supporters are associated almost exclusively with one characteristic. Moreover, all clubs we analyze have millions of supporters so that even "minorities" across some characteristics are numerous. Therefore, no club is associated with the characteristics of the majority of its supporters.

This feature is relevant in our context, as it suggests that socialization through the preferences of the Brazilian football clubs has the potential of creating ties among individuals who would not necessarily share other identities. In consonance with this, in our sample of Twitter users (which we will describe in Section 4), supporters of specific clubs are not disproportionately associated with a political affiliation (Appendix Table B.2). For nine out of the ten clubs we analyze, at least 38% of the supporters in our sample prefer the candidate preferred by the minority of those clubs' supporters. Even for Corinthians, the club that has a more substantial majority of supporters with a political identity, the minority is still numerous: at least 27% of the club's supporters in our sample prefer the minority candidate. Hence, the set of supporters of a given Brazilian football club is highly heterogeneous. This heterogeneity creates the opportunity for the formation of ties across income or partisan lines.

2.3 Twitter

The setting of this experiment is Twitter, one of Brazil's most popular social media platforms and one of the biggest Twitter markets worldwide. In 2021, right before our experiment, over 17 million Brazilians used Twitter (Statista, 2022b), making it the fourth country in usage of this platform (Brazil's population size ranks 7th worldwide). Twitter is a microblogging

[8]During the first semester of 2023, at least seven people were killed as a result of fights between football club supporters in Brazil (G1, 07-11-2023). Unexpected results in football matches in Brazil are also causally linked to episodes of domestic violence (Arabe, 2022).

platform where users can share content in short posts (tweets) of at most 280 characters. On this platform, it is common to use hashtags—short expressions beginning with the symbol #—to signal a post's topic. Through hashtags, it is easy for users to find others tweeting about their topics of interest. Users can also re-tweet or like posts from others, amplifying this content by making it visible to their followers.

On Twitter, the default configuration is for an account to be public and thus the vast majority of users have public profiles, which implies that their posts are publicly visible. Each user with a public account has a profile page visible to all other users, including a profile picture, a background picture, and a short description (called *bio*) provided by the users. Moreover, the profile page shows the account's history of tweets and usage metrics, such as the number of tweets, followers, and friends (the profiles the user follows).

Users can connect via follows, which do not need to be reciprocated, differently from other social media platforms such as Facebook. Indeed, to follow a public account, a user merely needs to click on "follow" on the account's profile page. Right after the follow, the user who has been followed usually receives a *follow notification* on their account, informing them that a new account has followed their profile. This notification shows the profile of who followed the user, and this user may decide to follow that account back, do nothing, or block it. Once someone follows another account, its new tweets, re-tweets, and likes may appear on this person's *timeline* (Twitter's main page). In contrast, users can also *block* others' accounts if they do not want those accounts to be able to interact with them. When an account is blocked, it cannot follow the user who blocked it or see its tweets. Importantly, the blocked account is not notified of the block, but if it visits the profile of an account that has blocked it, it can see that it was blocked.[9]

We interpret follows and blocks as two opposite measures of the willingness to establish social ties with other accounts. On the one hand, following an account signals a desire to connect with that account (for instance, by seeing its posts or being able to send direct messages to it). On the other hand, blocks signal derogation or a desire to be as distant as possible (in the Twitter environment) from the account that is the object of the block. In fact, a block is an active measure taken by an account that prevents any contact between that account and the blocked one.

A notable feature of Twitter in Brazil and other countries, such as the US, is that it plays an increasingly relevant role in shaping political discourse, particularly during campaign periods (Jungherr, 2016). Candidates and the general public have increasingly used Twitter to comment and gather information about politics in Brazil and elsewhere. Moreover, in countries such as the US, it has been shown that using Twitter had a causal effect on voter's decisions during the 2016 and 2020 elections (Fujiwara et al., 2021). While such direct evidence does not exist for Brazil, some statistics suggest that this platform is indeed relevant to elections in the country. Using data from the 2019 Latin American Public Opinion Survey (LAPOP), we see that among Brazilians who used Twitter in 2018, 75% claimed to use the platform to see political information at least sometimes a year, a similar rate to that of Facebook (80%) and above that of WhatsApp, of 65% (LAPOP, 2019). These

[9]This description corresponds to how blocks used to work on Twitter until 2024, when some changes where put in place.

numbers are particularly relevant considering that, in the 2018 presidential elections, social media influenced the vote of 45% of Brazilians, according to a recent survey by DataSenado (DataSenado, 2019). Therefore, social media in general—and Twitter in particular—is increasingly relevant for politics worldwide and in Brazil specifically, making this platform an ideal setting for our experiment on political identity and the formation of social ties.

3 Conceptual Framework

In our experiment, individuals (Twitter users) who prefer a political candidate in the Brazilian presidential election and support a football club are followed on Twitter by a fictional account with the same or different preferences as theirs. The individual must then decide how to interact with that account, either by following it back (thereby creating a social tie), ignoring it, or blocking it (demonstrating its desire to be as far apart as possible from that account in the social media environment).

We interpret these decisions in light of social identity theory (Tajfel and Turner, 1986). Identity—or a person's "sense of self" as Akerlof and Kranton (2000) put it—represents the idea that, in many situations, people do not see themselves as independent individuals but rather as belonging to certain social groups, with a membership they value. This theory starts from the assumption that society encompasses several social categories (Tajfel, 1981)—such as "male", "female", "democrat", "republican", "supporter of football club X", etc. These categories are constructed through historical, cultural, and sociological processes and can evolve or be relatively fluid (Kalin and Sambanis, 2018).

At different points in their lives, individuals may belong to some of these social groups. However, this does not imply that the individual identifies with all of those groups at all times. Indeed, an individual's sense of self may change depending on situational cues or the salience of certain groups. For instance (adapting an example from Shayo, 2020), someone who is male, supports Brazilian football club Palmeiras, and intended to vote for Lula in the 2022 presidential election may identify as a man, as a Palmeiras supporter, as a Lula voter, as a combination of some or all of these categories, or even with none of the above depending on the context. Social identity may be an important determinant of networks, since those identifying with a particular group tend to evaluate in-group members positively while being relatively hostile towards the out-group (Tajfel, 1974, 1981). Therefore, given their identity, people may form social ties with those perceived as more similar to them, leading to homophily in social interactions (McPherson et al., 2001; Currarini et al., 2009).

We evaluate this hypothesis in our experiment by considering two dimensions of social identity: political and football club preference (which is part of a person's "sense of self" in the Brazilian context). Each one of these dimensions contains, in principle, several social categories: for instance, someone can be pro-Lula, pro-Bolsonaro, or favor another candidate or party (or none). In the experiment, we focus on subjects belonging either to the pro-Lula or pro-Bolsonaro social categories. Similarly, in the football dimension, a person's social category is the club they support. Since we are interested exclusively in whether fictional account and subject share identities, we will focus on whether fictional account and subject

share identities in each dimension (and not on how subjects belonging to specific social categories behave).

Throughout this paper, we call the football club dimension of identity "affective" identity. We do this in opposition to political identity as a way to highlight that political identity may overshadow other dimensions of identity in general, not just the one we analyze. Moreover, this terminology highlights that, historically, political preferences did not have such a significant "affective" content, as the literature on (political) affective polarization suggests (Iyengar et al., 2019). Indeed, this literature argues that people with opposing political identities increasingly consider their political preference as a core element of their social identity (Huddy et al., 2015; Van Bavel and Packer, 2021), leading them to evaluate positively those from the same political group while being relatively hostile towards the out-group. Therefore, by using the term "affective identity" in opposition to political identity, we stress that the other dimensions of identity which we analyze —and which we show are overshadowed by political preference in a context of polarization—are dimensions within which people would traditionally socialize.

Finally, as we pointed out before, an individual's social identities are not fixed. Given the social categories they belong to, someone may identify with one or a subset of these categories at different times. Shayo (2020) models these decisions as endogenous, depending on the status, salience, and costs of identifying with a given group. In our setting, these changes in identity could have important implications for an individual's behavior. For instance, when elections are close—and potentially more salient—the identity weight people assign to their political identity may increase relative to the weight assigned to other dimensions. This would lead to more homophily in the political dimension and a decrease in the relative importance of the other dimension to the formation of ties.

4 Experimental Design and Data

4.1 Experimental Design

We conducted our pre-registered experiment on Twitter between July and December 2022. We created fictional accounts on Twitter that signaled their preferred candidate in the 2022 Brazilian election and/or their preferred Brazilian football club. The fictional accounts randomly followed Twitter users who shared or did not share each identity with it. After five days of activity, we computed the number of follow-backs and blocks obtained by each bot. These are our two outcomes of interest in the experiment.

We ran the experiment on waves of five days each. On each wave, we activate three types of fictional accounts: (1) fictional accounts that signal both dimensions of identity (political and affective); (2) fictional accounts that only signal political identity; (3) fictional accounts that only signal affective (football-related) identity. Specifically, for each wave, we randomly chose two Brazilian football clubs (say, clubs A and B). We then created eight fictional accounts: pro-Lula, supporter of club A; pro-Bolsonaro, supporter of club A; pro-Lula, supporter of club B; pro-Bolsonaro, supporter of club B; supporter of club A

(politically neutral); supporter of club B (politically neutral); pro-Lula (no club preference); pro-Bolsonaro (no club preference). The objective of creating fictional accounts that were neutral in one of the two identity dimensions is to evaluate the importance of each one of these two identities to the formation of ties without having to condition on the other identity.

4.1.1 Fictional Accounts

Appendix Table A.1 describes the elements used in the accounts. Each account is characterized by its preference for a political candidate (Lula, Bolsonaro, or neutral), and by its preference for a football club (which can be one of the six Brazilian clubs with the largest number of supporters, or neutral).[10] The political and affective identities of each fictional account are chosen randomly using a procedure described in the following subsection.

Given the assigned identity of the fictional account, we signal political identity by including, in the account's bio, either the hashtag *#Lula2022* or *#Bolsonaro2022*, and by re-tweeting one post from its supported candidate.[11] If the fictional account is politically neutral, we do not include either hashtag and do not retweet a political post. On the other hand, we signal affective identity through its profile picture (a picture of a flag with its preferred team logo in a stadium) and by adding the text "Supporter of team X" in the account's bio. For fictional accounts that are neutral in the football-related dimension, we use a photo of a football stadium outside Brazil (for which it is not possible to identify the teams) instead of a specific team's logo as the profile pic and include the text "Football fan" in the bio. Accounts that are football team-neutral still signal that they are interested in football (the only difference is that they do not signal a preference for a specific team). Figure 1 shows examples of fictional accounts.

Therefore, for the fictional accounts that signal both dimensions of identity, the affective identity—preferred football club—is more salient than political identity (which is signaled exclusively on the fictional account's bio). To assess the robustness of our results, we replicate our experiment with fictional accounts that signal political identity more saliently (see discussion in Subsection 5.3).

4.1.2 Sample Selection and Assignment into Treatment

The most important feature of our sample is that we must be able to identify the political identity (either pro-Lula or pro-Bolsonaro) and the preferred football team of each subject.

[10]The six clubs with the largest number of supporters in Brazil are C.R. Flamengo, S.C. Corinthians Paulista, São Paulo F.C., S.E. Palmeiras, Grêmio F.B.P.A. and C.R. Vasco da Gama. The ranking of club supporters comes from a 2022 survey by Sport Track and XP (Sport Track and XP, 2022). While the fictional accounts only support one of these six teams, the subject pool includes individuals who support rivals of these teams—specifically, apart from the six teams listed, we include subjects who support S.C. Internacional (Grêmio's rival), Botafogo F.R. and Fluminense F.C. (Flamengo and Vasco's rivals), and Santos F.C. (Palmeiras, São Paulo and Corinthians' rival).

[11]To alleviate concerns that the fictional accounts may be amplifying political content, we only re-tweet posts that already have more than 500 re-tweets and that do not include misleading information or hate speech, as agreed with our Institutional Review Board.

Figure 1: Examples of Fictional Accounts

(a) Pro-Bolsonaro; Flamengo supporter

(b) Pro-Lula; Palmeiras supporter

(c) Pro-Lula; Neutral-Team

(d) Politically Neutral, Flamengo supporter

Notes: The figures show examples of fictional accounts used in the experiment. Political identity is signaled by the hashtag *#Lula2022* or *#Bolsonaro2022* on the account's bio. Football club identity is signaled by the profile picture and the text "Supporter of [club's official Twitter account]" on the bio. When football club identity is not signaled (Panel c), the account still signals interest in football through its profile picture (a neutral football stadium) and the text "Football lover" on its bio.

Appendix Figure A.2 represents schematically the procedure used to obtain the subject sample. First, we use Twitter's API to obtain a sample of users who either tweeted or re-tweet a status containing pro-Lula or pro-Bolsonaro hashtags between May 31^{st} and July 11^{th}, 2022. The list of hashtags we considered is in Appendix Table A.2. Hence, our sample comprises politically engaged individuals, who were already actively discussing politics a couple of months before the election and the official campaign period (which started on August 16^{th}). Then, we inspected if the user's Twitter bio (the short description that the user writes in their profile) signaled the user's preferred Brazilian football club. To do this, we first use a simple algorithm that detects terms associated with the six most popular Brazilian football clubs and their rivals in the bios and then manually check if the matches are correct. We then remove accounts that were created in 2022 (that are more likely to be inauthentic), accounts that are clearly automated, accounts with less than 10 followers and accounts with a ratio of followers to friends above 20. The objective of doing this is to remove accounts that are very unlikely to follow back the experimental accounts and

accounts that are not authentic.[12] After these procedures, we are left with a sample of 4,652 individual accounts. We note that, due to query restrictions of Twitter's API, this is only a random sample of the Brazilian accounts that signal political and football club preferences on Twitter.

We obtained a set of variables for each subject using Twitter's API. We have information on the number of tweets, followers, and friends and the location of the accounts that choose to make this information public, which we recode to a regional level. Moreover, we know whether the account is verified, the number of likes ("favorites") performed, and its creation date. From the names of the users, we predicted their gender using information from the Brazilian Census (tabulated by Meireles, 2021). Appendix Tables B.1 and B.2 present descriptive statistics of subjects. First, our sample is not heavily skewed towards the Bolsonaro and Lula supporters (45 and 55%, approximately). We also show that, for all football clubs we consider, there is a significant group that supports each of the two political candidates. In some cases, the distribution is skewed towards one candidate, but there is always at least 27% of users who support each candidate. This is consistent with the observation that, in Brazil, football clubs are not specifically associated with political preferences and that the set of supporters of every mass club is heterogeneous. In subsection 5.3.1, we show that our results are robust to excluding any single team, including those whose fans are skewed towards one of the politicians, from the sample.

In each experimental wave, we activated eight fictional accounts: four accounts that signal both their preferred football team and their political preference; and four accounts that are neutral in one of the two dimensions (i.e., two accounts that are "football fans", but do not signal a specific team; and two accounts that signal a specific team, but not a political identity). In each wave, we randomly choose two football clubs for the fictional accounts.[13] Then, within a wave, three fictional accounts signaled a preference for each of these two teams.

Each fictional account then follows approximately 100 subjects during each wave. Following the suggestion of Athey and Imbens (2017), we perform block-randomization to define the treatment assignment. Specifically, the treatment assignment to each fictional account is done by stratifying the subjects on their political identity, preferred football team, and number of followers (above or below the sample median). For the fictional accounts that signal their preferred football club, we restrict the sample of subjects to those who either support the same team as the fictional account or support a rival team. We only consider regional (intra-state) rivalries; the list of rivalries is in Appendix Table A.3. Given that we

[12]Before the experiment, as pre-registered, we inspected manually to identify automated ("bot") accounts and removed them from the subject pool. After running the experiment, we also used the *Botometer* API to estimate the probability that each of our potential subjects was an automated account. This API uses several publicly available information from Twitter accounts to estimate the probability that the account is automated (for details, see Sayyadiharikandeh et al., 2020; Yang et al., 2020). Reassuringly, we only excluded from the final subject pool accounts with a Botometer score above 0.85. However, we did miss 39 accounts with more than 85% chance of being automated (less than 1% of our final sample). The median Botometer score of the subjects is 0.13. In Section 5.3, we show that our results are identical for a sub-sample of subjects that are unlikely fictional accounts according to the Botometer classification.

[13]Throughout the experiment, we randomly sample teams with a probability equal to the proportion of each team's supporters in our sample.

are interested in studying the effect of matching fictional accounts and subjects' identities on follow-backs and blocks, we have four strata in terms of fictional account-subjects identity pairs (congruence in both dimensions, incongruence in both dimensions, or congruence in a single dimension), and each pair is further divided into two smaller strata (above or below the median number of followers). We sample the same proportion of subjects from each stratum. Each subject may be treated (i.e., followed by a fictional account) more than once, but never in subsequent waves: after being treated in a wave, a subject only returns to the subject pool after three waves. Hence, concerns about subjects "learning" about the experiment are alleviated.

Therefore, the "treatment" in our experiment is to receive a follow notification from one of the experimental accounts on Twitter. The experimental variation comes from whether the subject and the fictional account agree or disagree in their political and/or affective dimension of identity. Appendix Figure A.3 illustrates such notification. How a user sees the notification depends on whether he or she is using Twitter from a mobile application or a desktop computer.[14] In both cases, the user immediately sees the fictional account's photo. In the mobile app, he or she also sees the description (which indicates the political affiliation). The user only sees the description on a computer when they click (or hover the mouse's cursor) over the profile. However, to follow back or block the account, every desktop user will inevitably need to either click on the profile or hover the mouse's cursor over it, thus seeing the fictional account's description and, therefore, its political affiliation.

Apart from following the experimentally assigned accounts, each fictional account followed one account from someone aware of the experiment. This person then informed 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.[15] If an account is shadow-banned, we drop it from the analysis, as determined in our pre-analysis plan. Over the entire experiment, we had 12 shadow-banned accounts (5.1% of the accounts we created). Shadow-banning was not correlated with the bot's political identity (specifically, out of the 12 shadow-banned accounts, 3 were pro-Lula, 4 pro-Bolsonaro, and 5 were politically neutral).

4.1.3 Timing

As described in the previous section, the experiment was carried out in waves. In each wave, eight fictional accounts were activated: four signaling both their political identity and football club preference, and four neutral in one of the two dimensions of social identity. Within each wave, we used the following timeline:

[14] Overall, 80% of Twitter users access the platform via their mobile device (Statista, 2022a). In our sample, by live-streaming tweets using Twitter API during the experimental period, we find that 72% of subjects exclusively tweeted and re-tweeted through the Mobile App.

[15] 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 follow notifications. Therefore, we guarantee that no fictional account is shadow-banned before using the results from any experimental wave.

(i) **Day 0:** Creation of accounts according to the procedures described in the Appendix Table A.1. The account re-tweets a post related to its sportive identity (either a post from its preferred club official account—if the fictional account has a preferred club—or a general post about football that does not favor any club), and then a post from its preferred political candidate. The fictional accounts also follow a set of 15 "elite" accounts related to their interest (for instance, the official account of their preferred candidate and club), and is followed by a set of five colleagues who were aware of the experiment.

(ii) **Day 1:** Each fictional account follows the subjects assigned to it according to the procedure described in the previous section.

(iii) **Day 5:** After five active days, we compute the number of followers and blocks for each account and delete all the information in the account, relaunching it for use in the next wave. We defined the five days account life based on previous research on Twitter showing that over 95% of follow-backs tend to occur within five days of the treatment period (Ajzenman et al., Forthcoming).

We started with one wave every Tuesday and every Friday, which means that we had two overlapping waves at each moment. The timeline is displayed in Appendix Figure A.1. We ran 43 experimental waves between July and December 2022. The Brazilian presidential election of 2022 was held during the second semester of 2022 (specifically, the first round occurred on October 3^{rd} and the second round on October 29^{th}). We use the differential timing of the experimental waves to study the heterogeneous effect of shared identity on the formation of social ties when political identity is more or less salient.

On each wave, we compute follow-backs once a day using Twitter's API. In our main analysis, we will use the final follow-back measure, computed on the fifth day since the fictional account followed the subjects. On the other hand, we only compute blocks at the end of each wave (i.e., on the fifth day), because Twitter's API does not allow us to directly compute blocks. The procedure we use to compute blocks is as follows: first, we use Twitter's API to obtain, for each fictional account, the set of accounts followed by it. We then compare this set with the set of accounts assigned to be followed by the fictional account. The difference between the two sets can be due to three mutually exclusive reasons: (i) the fictional account was indeed blocked by a subject; (ii) the subject was suspended or deactivated their account; (iii) the subject removed the fictional account from its followers. To assess which one of the three happened for each subject in this difference set, we manually enter these subjects' profiles from the fictional account's Twitter account. From the profiles, we can easily see which of the three cases happened. We only classify the subject as having blocked the fictional account if we see, on the fifth day, a block using this procedure.[16]

[16] A fourth possibility is that a subject blocked a fictional account but then unblocked it. We do not treat this as a block but as a follower removal. Thus, in our measure of blocks, there are only subjects that block a subject and keep it this way until the end of the wave.

4.2 Empirical Strategy

We are interested in studying the effect of identity congruence in the formation of social ties on Twitter. In most of our analysis, we present results that pool all experimental waves, comparing the follow-back and block rates of subjects who shared or did not share political and/or affective identity with the fictional account.

To formally test the significance of our results, we use the following pre-registered specifications, which include wave and strata fixed effects. First, we focus on experimental accounts that signal a single dimension of identity (either political or affective). These accounts follow subjects with whom they agree or disagree in this dimension. Our outcomes of interest (follow-backs and blocks) measure how subjects interact with the fictional accounts in response to being followed by them. Thus, we restrict our analysis to the experimentally assigned pairs of subjects-fictional accounts. We denote our outcome of interest by Y_{ijst}, an indicator equal to one if subject i from strata s interacted with fictional account j during wave t. Here, "interacted" can either represent a follow-back or a block. We then estimate an equation of the form:

$$Y_{ijst} = \alpha + \beta \times \text{identity\_congruence}_{ij} + X_{ijt}\delta + \varphi_{st} + \epsilon_{ijst} \quad (1)$$

were $\text{identity\_congruence}_{ij}$ is an indicator equal to one if a fictional account and a subject share identity (in the dimension we are studying), φ_{st} represents strata $\times$ wave fixed effects and ϵ_{ijst} is the error term. X_{ijt} is a vector of control variables from the fictional account, subjects, and waves (interacted with the treatment dummies).[17] This vector includes the number of followers and tweets from the subject; the year he or she created the account; the subject's gender and location; and the Google trend index of the fictional account's j football club at wave t (interacted with the identity congruence indicator, in the case of fictional accounts that signal their preferred football club). The purpose of controlling for this trend is to control for the salience of the football-related identity across waves.

We also analyze accounts that signal both dimensions. In this case, there are four possible pairs of subjects and fictional accounts (congruent in both dimensions, congruent either on affective or political identity, but incongruent in the other dimension, and incongruent in both). To study these treatment arms, we estimate the following equation:

$$\begin{aligned} Y_{ijst} = \beta_0 + \beta_1 \times \text{political\_congruence}_{ij} + \beta_2 \times \text{affective\_congruence}_{ij} + \\ \beta_3 \times \text{political\_congruence}_{ij} \times \text{affective\_congruence}_{ij} + \\ X_{ijt}\lambda + \phi_{st} + \varepsilon_{ijst} \end{aligned} \quad (2)$$

where $\text{political\_congruence}_{ij}$ is an indicator equal to one if fictional account j and subject i share political preferences, $\text{affective\_congruence}_{ij}$ equals one if fictional account j and subject i share preference for football club, and the other variables have the same definition as before. We include the same covariates as before.

[17] We include strata $\times$ wave fixed effects following the suggestion from Bruhn and McKenzie (2009). Among the strata fixed effects, there is a misfit dummy—"misfits" are subjects assigned to treatment that violate the proportional assignment to each treatment arm within their stratum. These misfits are re-assigned globally following Carril (2017).

Coefficient β_1 is the effect (in percentage points) on follow-backs or blocks of sharing political identity for subjects who do not share affective identity with the bot. Similarly, β_2 is the effect of sharing affective identity for subjects who do not share political identity with the fictional account. Finally, β_3 can be interpreted as the difference in the effect of sharing affective identity between subjects who share or do not share political identity with the bot.

We present standard errors clustered at the fictional account level. We performed the simplest assessment proposed by Ferman (2022) to verify if our inference method is reliable, given the number of clusters. 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 pilot. Reassuringly, we obtained a rejection rate of the null under a nominal significance level of 5% that was very close to 5% in all cases.

4.3 Balance and Attrition

Appendix Table B.3 present summary statistics of the treated subjects in the eight treatment arms. In all cases, pre-treatment subject characteristics are balanced across treatments: for all pre-treatment variables, we cannot reject the null hypothesis of equality across all treatment arms for standard significance levels (we perform a joint test of equality and report its F-statistic in the last column of the table).

The table also shows attrition rates for each treatment arm. We consider that a subject suffered attrition if it was assigned to be treated, but we were unable to treat it. This could happen for three reasons: the subject's account was suspended (a punishment inflicted by Twitter when the account's use violates the platform's policy); the subject deactivated their account; or the subject made its account private. In the first two cases, we would not be able to find the account on Twitter. In the third case, we could find the account but did not follow it as agreed with our IRB.

Overall, there was no differential attrition in the experiment. For all treatment arms, the attrition rate was close to 9%, and the characteristics of the attrited subjects are not different between the treatment arms, as can be seen in Appendix Table B.4. Therefore, when analyzing results by pooling the results of different waves and estimating Equation (1) or (2) in a cross-section, attrition will not be a concern. Moreover, since we include wave fixed effects and attrition is observed at the beginning of each wave, we always compare statistically similar accounts when doing this type of "statical" specification.

Attrition could be a concern in our analysis of heterogeneous effects over time because this analysis is, by construction, dynamic: we would like to compare the behavior of subjects across different waves. However, given that attrition can happen, the pool of subjects we observe in different waves may differ. This is indeed the case, as shown in Appendix Table B.5. This table compares the characteristics of subjects who never suffered attrition in the experiment with those who suffered it at some point. Subjects that suffered attrition are disproportionately more likely to support candidate Jair Bolsonaro, have more followers and more Twitter activity than those that never suffered attrition. Therefore, when analyzing heterogeneous effects over time, we present results restricting the analysis to the sample of

subjects who remained active throughout the experimental period.

5 Experimental Results

5.1 Effects of Political or Affective Congruence on the Formation of Ties

We start by examining whether sharing each identity—affective or political—impacts the formation of social ties on Twitter. To do that, we restrict our analysis to the experimental accounts that signal a single dimension of identity. Results for this subset of experimental accounts are displayed in Table 1, on Panel A for follow-backs and Panel B for blocks.

We first find that football clubs, unconditionally, are a relevant dimension of socialization in our setting. The first three columns of Table 1 show estimates of β_1 in Equation (1) for the experiment using politically neutral accounts. For completeness and following our pre-registration, we show similar results without controls, fixed effects of the wave and strata, and the additional controls listed in Section 4.2.

A subject is about 14 percentage points (pp) more likely to reciprocate a follow from a fictional account supporting their team rather than a rival—more than a 50% increase in the likelihood relative to the follow-back rate for rival teams, of 22.9%. The results for blocks tell a similar story, even though the block rate of politically neutral accounts is low. Subjects block 2.3% of rival accounts, and sharing a football club reduces this likelihood by 1.26 pp (significant at the 1% level). Therefore, the preferences of football clubs are a relevant determinant of forming social ties in our setting. This result provides quantitative evidence in favor of the observation, made by several sociologists and anthropologists, that football club preferences are a relevant dimension of socialization in Brazil (e.g., Murad, 1995; DaMatta, 1982).

We also find that political identity plays an important role in forming social ties. Columns 4-6 of Table 1 show estimates of the effect of shared political identity, now considering experimental accounts that only signal political preference. Recall that although these experimental accounts are neutral regarding football club preference, they still signal interest in football. We find that sharing political identity causes an increase in the probability of follow-back of 20 pp (from 16.8% to 36.8%) and a decrease in the probability of blocks of approximately 12 pp (from 0.7% to 13%), both significant at the 1% level and large in magnitude. Overall, this first set of results validates that both identity dimensions—political and football-related—are independently relevant for socialization in our setting.

5.2 The Interplay between Political and Affective Congruence

So far, we only discussed the results for the accounts that signal either political or affective identity (exclusively). Analyzing results for the experimental accounts that signal both dimensions of identity allows us to study their interplay on the formation of social ties.

Table 1: Effect of Shared Identity on Follow-Backs and Blocks—Experimental Accounts Signaling a Single Dimension of Identity

Panel A: Follow Backs, Fictional Accounts Signaling a Single Dimension of Identity						
	Dependent Variable: Follow-Backs (1 = Yes)					
	Politically-neutral Fictional accounts			Football-neutral Fictional accounts		
	(1)	(2)	(3)	(4)	(5)	(6)
Football club congruence	0.1337	0.1413	0.1406			
	(0.0133)	(0.0134)	(0.0130)			
Political congruence				0.2000	0.1994	0.1979
				(0.0148)	(0.0147)	(0.0133)
Average (incongruent pairs)	0.229			0.168		
Wave, Strata Fixed Effects	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes
Observations	7,388	7,388	7,388	7,678	7,678	7,678
Panel B: Blocks, Fictional accounts Signaling a Single Dimension of Identity						
	Dependent Variable: Blocks (1 = Yes)					
	Politically-neutral Fictional accounts			Football-neutral Fictional accounts		
	(1)	(2)	(3)	(4)	(5)	(6)
Football club congruence	-0.0126	-0.0126	-0.0132			
	(0.0032)	(0.0032)	(0.0043)			
Political congruence				-0.1225	-0.1225	-0.1224
				(0.0062)	(0.0061)	(0.0060)
Average (incongruent pairs)	0.023			0.130		
Wave, Strata Fixed Effects	No	Yes	Yes	No	Yes	Yes
Controls	No	No	Yes	No	No	Yes
Observations	7,199	7,199	7,199	7,492	7,492	7,492

Notes: The table presents regression estimates for the effect of sharing identities on follow-backs (Panel A) and blocks (Panel B), for the subset of experimental accounts that signaled a single dimension of identity (either political or affective). The table presents estimates for β_1 in Equation (1), first with no controls, then with wave and strata fixed effects, and then with additional controls interacted with the treatment indicator (bot's football club, Google trend index of the clubs, and subject's number of followers and friends). The sample excludes shadow-banned accounts, as pre-registered and discussed in the text. The number of observations for blocks is slightly smaller because we were unable to collect block information in one of 43 waves due to a technical error. Standard errors clustered at the fictional account level are in parentheses.

Results are displayed in Figure 2 for follow-backs and Figure 3 for blocks, using the raw data (i.e., without any additional controls). The top panel of either figure shows the average follow-back or block rate for all eight treatment arms (the four arms with fictional accounts signaling both identities and the four arms with fictional accounts signaling a single dimension). All of these treatment arms are represented simultaneously in the plot. In the *x-axis*, we represent whether the fictional account and subjects share political identity: the left-most three columns show cases where the fictional account and subject disagree politically, while the right-most show cases where they agree. Moreover, the two bars in the center represent the two cases in which political identity is not signaled, and therefore, the only dimension of interest is the affective (these are the same results as columns 1 of Table 1). Finally, the *bar colors* indicate the relationship between the subject and the fictional account's football club preference. As in the case of political identity, there are three possibilities: either the fictional account and subject share football club preference,

support rival clubs, or the fictional account does not signal its preferred club (in that case, it only signals political preferences, and the results, in gray, are the same as column 4 of Table 1). For simplicity of exposure, we report estimated differences and p-values (in brackets) for the most relevant two-by-two comparisons between treatment arms in the bottom panel of both figures (with standard errors clustered at the fictional account level). Appendix Tables B.6 and B.7 report tests of difference in means for follow-backs and blocks (respectively) for every pair of treatment arms, including with wave and strata fixed effects and additional controls (results are virtually unchanged).

The figures reveal that sharing either dimension of identity significantly increases the probability of follow-backs and decreases the probability of blocks. Regarding follow-backs, the subjects least likely to reciprocate a fictional account's follow are those who share neither a political identity nor a preferred football club with the bot. In this case, there is only a 16% chance of follow-back. For blocks, the result is qualitatively similar since those most likely to block a fictional account are subjects who do not share either dimension of identity with it (14.6% chance of blocking). By sharing either dimension of identity, there is an increase in the follow-back probability and a decrease in the blocking probability. However, the magnitudes of these effects are different once we condition on the other dimension of identity, suggesting that political identities overshadows most of the positive effect of affective identities. We analyze this idea in more depth in the following sub-sections.

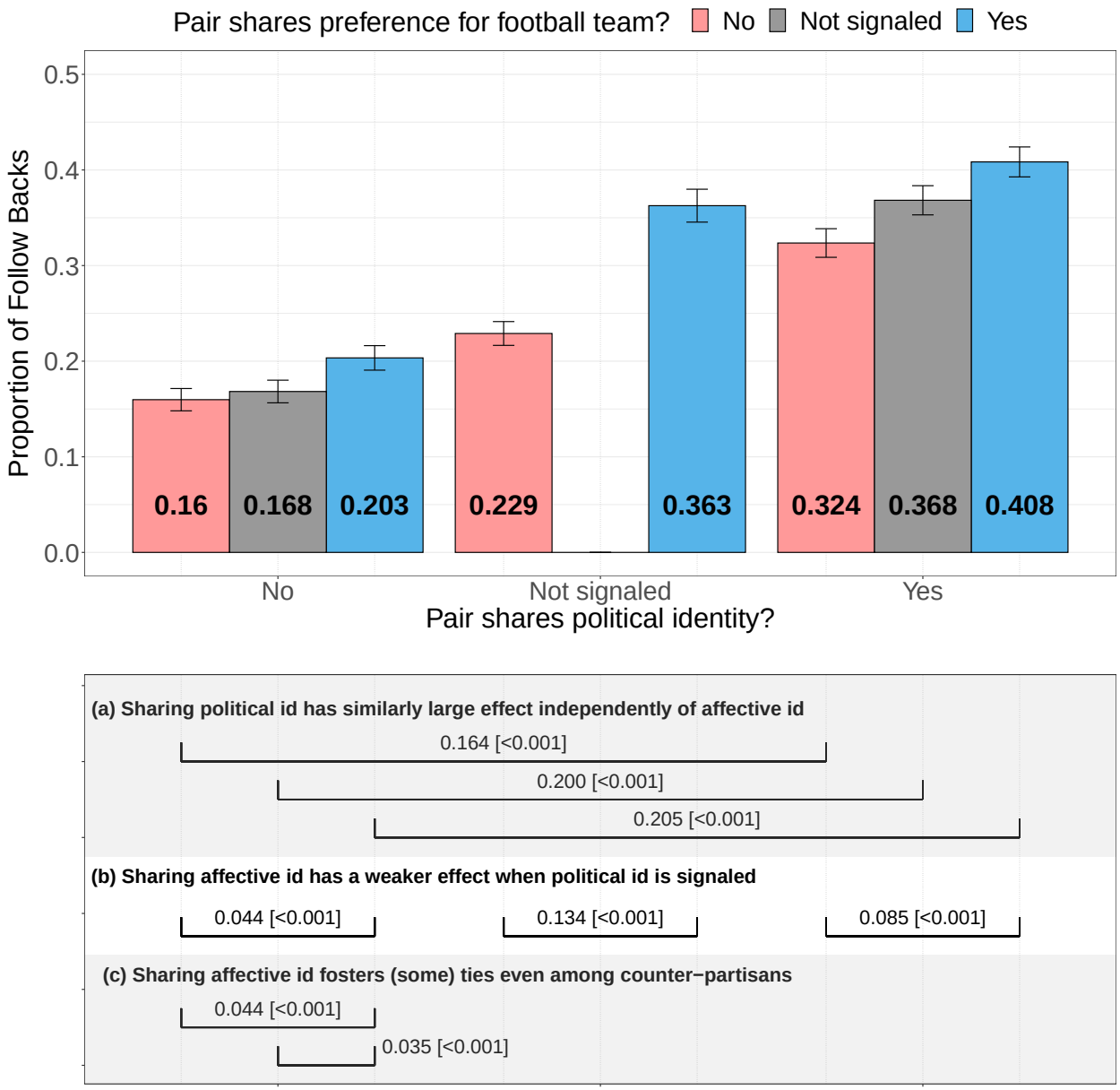
5.2.1 Effects of shared identity when the other identity is revealed

We start by studying whether congruence in a dimension significantly impacts the probability of tie formation conditional on sharing or not the other identity.

First, among subject-fictional account pairs who support the same football club (blue bars), the likelihood of follow-backs is 20.3% when the fictional account and subject disagree politically, against 40.8% when they agree. The difference of 20.5 percentage points is highly statistically significant (p-value < 0.001), and very similar to the effect of political congruence considering football-neutral fictional accounts. Likewise, among pairs who support rival clubs (red bars), the follow-back probability is 16% for subjects and fictional accounts who disagree politically and 32.4% for subjects and fictional accounts who agree in this dimension (the difference of 16.4 pp is also significant at the 1% level). In all cases, sharing political identity roughly doubles the probability of follow-back. Figure 2, comparison (a) reflects this pattern.

Similarly, the effect of political identity on blocks is substantial, independent of information about the fictional account's preferred club. When this information is unavailable, the probability of blocking a fictional account is 12.3 pp smaller when the fictional account and subject share political identity compared to when they have opposite identities in this dimension (p-value < 0.001). When the fictional account and subjects support rival football clubs, this difference is 13.5 pp (p-value < 0.001). Finally, when the fictional account and subject support the same club, the difference is relatively smaller but still large: 7.9 pp (p-value < 0.001). While it is true that, when the fictional account and subject support the same club, there is a significant reduction in the blocking probability—which we will discuss later in this section—, the difference is still sizable. As Figure 3 shows, blocking happens

Figure 2: Effect of Shared Political and Affective Identity on Follow-Backs



Notes: The figures show the effect of sharing political and affective (football club) identity on the rate of follow-backs for all eight treatment arms in the main experiment (fictional accounts that signal both or a single dimension of identity). The sample comprises the subject-fictional account pairs in the experiment, pooling all waves and excluding shadow-banned accounts (as pre-registered) for a total of 30,194 observations (of 4,620 unique subjects). The x-axis shows whether the fictional account and subject share political identity (or show that this dimension is not signaled by the fictional accounts), while the colors show whether the fictional account and subject share a preference for a football club (or show that this dimension is not signaled by the bot). Each bar shows the average follow-back rate for each treatment arm. The error bars represent 95% confidence intervals. The bottom panel shows the estimated average difference in follow-backs and p-values (in brackets) for selected pairs of treatment arms, computed using standard errors clustered at the fictional account level. Tests of difference for all treatment pairs and results with controls are reported in Appendix Table B.6.

almost exclusively against politically opposite accounts. Therefore, counter-partisans tend to avoid each other (through blocks).

Overall, the effect of sharing political identity is large regardless of whether the fictional account and subject support the same or rival clubs or if there is no information on the bot's football club preference. We interpret this as evidence that the effect of political identity on follow-backs is not offset significantly (nor reinforced) by information on affective identity.

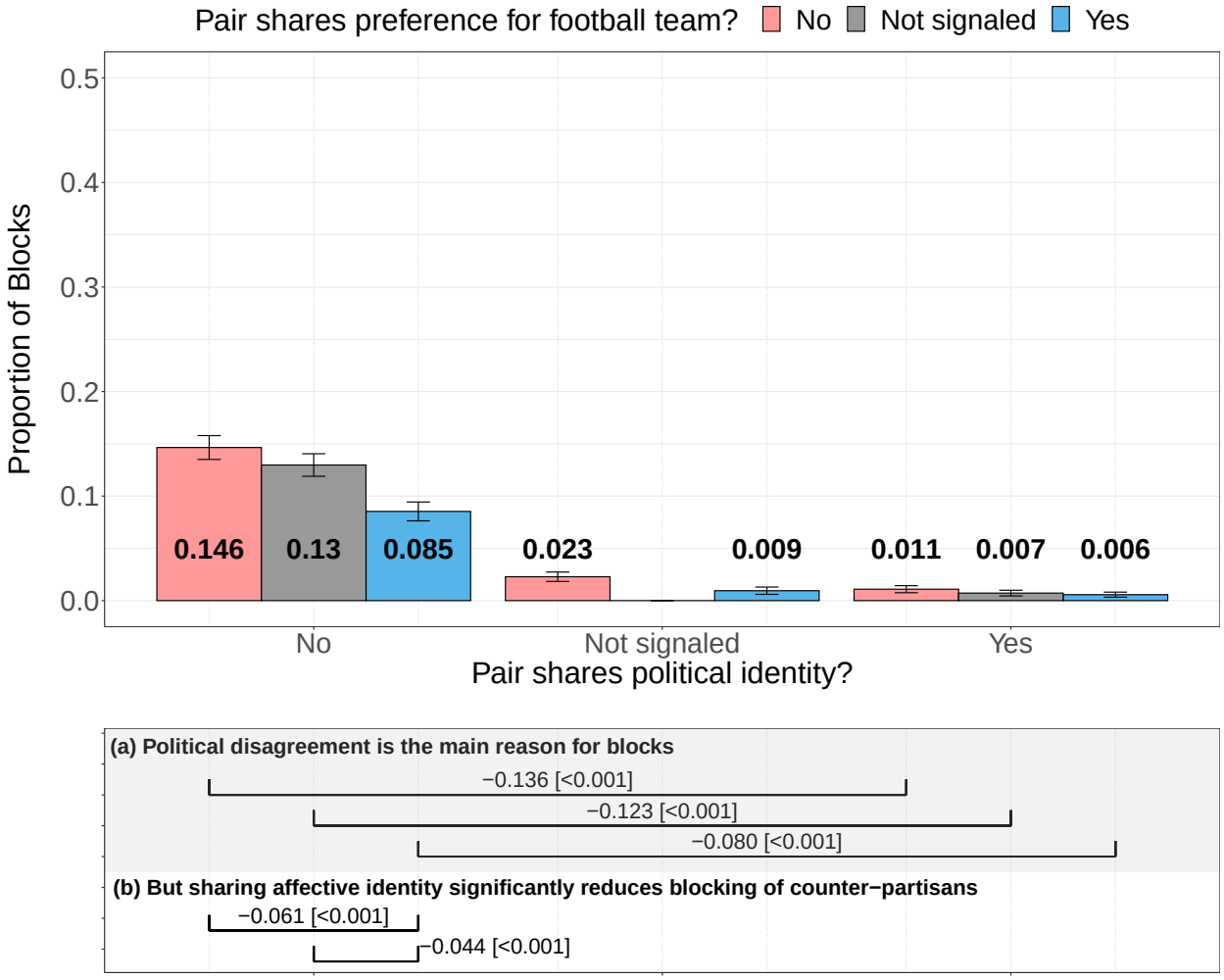
The same cannot be said of the effect of sharing affective identity conditional on information on the bot's political identity. When fictional accounts do not signal their political identity, the effect of sharing a football club was to increase the probability of follow-backs by 13.4 pp. This effect is considerably smaller both when the fictional account and subject agree or disagree politically. First, conditional on sharing political identity (Figure 2, right-most bars), the probability of follow-back when fictional account and subject support rival clubs is 32.4%, against 40.8% when they support the same one, a difference of 8.4 pp (p-value < 0.001). This effect of sharing affective identity conditional on agreeing politically is significantly smaller than the effect of sharing affective identity when the fictional accounts do not give information about political preferences. This difference is quantitatively meaningful, as it implies a reduction of almost 40% in the effect of affective congruence when political identity is shared compared to when it is not signaled. The reduction in the effect of affective identity is even more striking when we consider pairs that disagree politically. In this case, sharing a football club raises by 4.3 pp the probability of follow-backs (from a baseline of 16%), which is significantly less than the effect when the bot's political identity was not informed. This represents a reduction of approximately 68% in the effect of sharing affective identity relative to this effect when fictional accounts did not signal their political identity. Figure 2, comparison (b) reflects this pattern.

Hence, during the period we analyzed, political identity overshadowed other dimensions of identity (namely, football club preference) in forming social ties. Indeed, information on the fictional account's political identity offsets the effect of shared affective identity, particularly among politically-opposite individuals, undermining social ties that could be formed if the fictional account did not signal their political identity. This is evidence that at least in contexts of high polarization—such as the one we analyze—political preferences can reduce the potential of other shared identities to foster connections among individuals and cause the destruction of social ties that would otherwise be formed.

5.2.2 Congruence in affective identity and the formation of social ties among counter-partisans

Nevertheless, it is important to highlight that even though political divergence makes the effects of shared affective identity substantially smaller, this dimension of identity can still create ties among politically opposite individuals. In fact, among politically opposite individuals, sharing the preference of the football club increases the probability of follow-backs by 4.4 percentage points relative to subject-fictional account pairs supporting rival clubs (p-value < 0.001). Shared football identity also increases the probability of follow-back among politically opposite individuals relative to the case in which fictional accounts do not

Figure 3: Effect of Shared Political and Affective Identity on Blocks



Notes: The figures show the effect of sharing political and affective (football club) identity on the rate of blocks for all eight treatment arms in the main experiment (fictional accounts that signal both or a single dimension of identity). The sample comprises the subject-fictional account pairs in the experiment, pooling all waves and excluding shadow-banned accounts (as pre-registered) for a total of 30,194 observations (of 4,620 unique subjects). The x-axis shows whether the fictional account and subject share political identity (or show that this dimension is not signaled by the fictional accounts), while the colors show whether the fictional account and subject share a preference for a football club (or show that this dimension is not signaled by the bot). Each bar shows the average follow-back rate for each treatment arm. The error bars represent 95% confidence intervals. The bottom panel shows the estimated average difference in blocks and p-values (in brackets) for selected pairs of treatment arms, computed using standard errors clustered at the fictional account level. Tests of difference for all treatment pairs and results with controls are reported in Appendix Table B.7.

signal their preferred club. In this case, the effect is smaller (3.5 pp) but still significant at the 1% level. Therefore, even among politically opposing individuals, shared football clubs foster ties.

For blocks, the effect is even greater. Among politically opposite individuals, sharing affective identity reduces the probability of blocking by 6.1 pp compared to the case of rival clubs (p-value < 0.001). This represents a substantial reduction of 42% in the likelihood of blocking. Sharing a club also reduces the probability of blocking among counter-partisans relative to the case in which the information of the football club is not given by 4.5 pp (p-value < 0.001). Hence, sharing affective identity significantly reduces the probability of blocks—even if this probability remains relatively large.[18]

Therefore, despite indicating that political identity overshadows affective identity in the formation of ties, our results also suggest that, even in a context of intense polarization and among politically engaged individuals, sharing a dimension of identity such as football club can foster ties and reduce avoidance among politically opposing individuals. This result is consistent with evidence that football can promote social cohesion (Depetris-Chauvin et al., 2020; Ronconi, 2022). Hence, our findings indicate that highlighting a shared common interest—in this case, preference for a football club—can help reduce politically-induced societal divides. This result is also consistent with the evaluations of interventions to reduce affective polarization, particularly cross-partisan conversations (i.e. conversations between supporters of opposite parties). Santoro and Broockman (2022) show that the effectiveness of such conversations is conditional on the conversation's topic. Our experiment suggests that highlighting shared identities may be effective in this type of intervention, even though the effect may be small in a polarized context.[19]

Overall, we find that both dimensions of identity are relevant to forming social ties. Comparing the results for the accounts signaling both or a single dimension, our main finding is that political identity overshadows affective identity, reducing the importance of sharing a preference for a football club on the decision to follow back an account or not. Nevertheless, even when political identity is signaled, congruence in affective identity can generate social ties. This is surprising, particularly considering that our sample of subjects is composed of politically engaged individuals (who were using political hashtags at least three months before the election). The fact that congruence in affective identity plays a role even in this context suggests that similarities in dimensions relatively uncorrelated with politics can reduce political divides despite the overshadowing effect.

5.3 Validity, Interpretation, and Robustness

In this section, we report some robustness tests and results of additional exercises that may provide additional validity to the results and interpretations discussed so far.

[18] The positive effect of affective identity in preventing blocks among counter-partisans remains large even when political identity is signaled more saliently by fictional accounts. See Appendix B.6.

[19] We also show that there is *demand* for cross-partisan interactions (albeit small), while most of the literature on cross-partisan conversations focuses on experimentally assigning cross-partisan to talks.

5.3.1 Effect of Specific Clubs

A potential concern with our results could be that the specific football clubs we used signal other characteristics, changing subjects' decisions not because of congruence in football club preference but because some other characteristic signaled by a club preference is valued (or not) by them. This could have been a concern if, for instance, some clubs were particularly associated to one politician. To show that this is not the case, we repeat our analysis, excluding one fictional account club (and its rivals) at a time. This analysis, displayed in Appendix Table B.10, shows that results are not driven by specific clubs since the point estimates of the effect of sharing identities are stable for all sub-samples.

Moreover, some of the subjects in the experiment support football clubs that were not signaled by any fictional account throughout the experiment. This may be a concern since those subjects can only be assigned to be followed by a fictional account supporting a rival club, i.e., they can only be from the out-group in the affective dimension. We repeat our analysis excluding those subjects. Reassuringly, results are qualitatively similar and almost identical numerically to the main analysis, as reported in Appendix Table B.11.

5.3.2 Automated Accounts

Some concerns may also arise about automated accounts (or "bots") on Twitter. First, a potential concern is that part of our subject pool may include automated accounts. Following our pre-analysis plan, we manually excluded from the potential subject pool accounts that seemed likely to be fictional accounts before conducting the experiment. However, after the experiment, we also used the *Botometer* API to estimate the probability that each of our subjects was an automated account.[20] Reassuringly, the median score in our final sample was 0.13 (i.e., the API classifies the account as having a 13% chance of being automated), and only 39 accounts—less than 1%—had a score above 0.85 (in our manual classification, we did not remove any account with a score lower than 0.85). The final columns of Appendix Tables B.8 and B.9 display results for a subsample of subjects with below median *Botometer* score, i.e., for subjects whose accounts are extremely unlikely to be automated. The results are almost identical to the results for the main sample. Moreover, estimates are stable for other (both more and less conservative) *Botometer* scores thresholds.

A related concern is that some subjects may perceive the experimental accounts as inauthentic. Overall, given the high take-up of the experiment, it is clear that many users considered the accounts realistic. Nevertheless, the interpretation of our results could be challenged if users perceived accounts that did not share their political identity as more likely to be fictional accounts than accounts that shared their political identity. While we cannot directly assess this type of perception, we provide some indirect evidence suggesting this was not the case. Specifically, we use a Bayesian Classifier algorithm to classify whether the subject's most recent tweets before being followed by an experimental account had political content or not. Compared to a subject that tweeted about politics, a subject that tweeted about some other topic may expect to receive followers from other users who

[20]For details on *Botometer*'s algorithm, see Sayyadiharikandeh et al. (2020) and Yang et al. (2020).

do not share their political identity with a higher probability. In that sense, we consider that subjects that had just tweeted about a topic other than politics may be less suspicious when receiving a follow from someone identifying with the opposite political group, i.e., they may have less reason to (differentially) believe that an experimental account with opposite political identity is inauthentic. Hence, if the follow-back and blocking behavior of users who just tweeted about politics and who just tweeted about other topics is similar, we would have indirect evidence that suspicions about the experimental account's lack of authenticity are not driving our results.

Appendix Figure B.7 reports the results for this analysis. Due to data constraints, we only report results starting at wave 11 (the first one for which we collected the subject's most recent tweet before treatment). Appendix Figure B.8 reports the full sample results for this specific time frame. In all cases, we restrict the analysis to accounts that tweeted seven days before treatment. Reassuringly, we find that the behavior of subjects who had just tweeted something political is remarkably similar to that of subjects who tweeted something about another topic (and, more generally, to the overall behavior we documented in the main analysis). We also perform a similar heterogeneity analysis by classifying users' bios (the short profile description) according to their political content. In this case, we search for keywords associated with politics to classify bios as political or not. Again, the idea is that subjects who do not disclose their political preferences in their bios should be less suspicious when they are followed by a politically divergent account. We consider users' bios from before the beginning of the experiment. Approximately half of our subjects had political references in their pre-experiment bios. Appendix Figure B.9 shows that subjects whose bio did not contain political references behave similarly to those whose bio contained this type of content. As discussed in the previous paragraph, these two pieces of evidence suggest that suspicions about the experimental accounts' lack of authenticity are not driving our results.

These exercises also help rule out another alternative story: that follows by counter-partisans lead subjects to differentially update about some other (undesired) characteristic of this potential friend. For instance, a pro-Lula user might think that someone with a pro-Bolsonaro hashtag who follows them is not smart, even compared to other potential pro-Bolsonaro Twitter friends. This type of story would only be a concern if the updated belief is specific to the act of following a counter-partisan that signals this characteristic. To the extent that less openly political subjects have less reason to make this type of judgement, the two checks described above also suggest this type of concern is unwarranted in our setting.

5.3.3 Demand for follow-backs: Information versus social connections

Why do the subjects in our experiment establish (or reject) ties with the experimental accounts? Our preferred explanation is that Twitter ties are social connections (i.e., subjects want to become virtual "friends" with the experimental accounts). An alternative interpretation is that subjects demand information and believe that the experimental accounts may be good sources of information with the preferred slant. To shed light on this, we conducted an auxiliary experiment to help disentangle these two types of effect by creating accounts that explicitly state that they are automated and share information (see examples in Ap-

pendix Figure B.10). These accounts follow randomly selected subjects as in the original experiment. We then compare the follow-back and block rates between these "information" accounts and accounts similar to the original experiment. We conducted the auxiliary experiment one year after the original experiment, which explains the lower level of follow-backs and blocks. This auxiliary experiment is similar to the one by Mosleh et al. (2023), who document that users in the United States prefer to follow same-party accounts that identify as human rather than those that identify as fictional accounts and disseminate information, suggesting that users have a social motivation to follow back rather than simply wanting to see partisan information.

We find that the subjects in our sample do not only care about information. Appendix Figure B.11 and Table B.12 show the results. For follow-backs, subjects who share political and affective identities with the fictional account are 10 pp (56%) more likely to follow the original fictional accounts than the informational fictional accounts. At the same time, subjects who disagree with the fictional account in the affective dimension are almost equally likely to follow fictional accounts of the original or the informational type. This suggests that, particularly for the in-group, social motivation (rather than a demand for information) plays a large role in determining follow-back decisions. The results for blocks mirror those for follow-backs: in this case, subjects who disagree in both dimensions are four pp (65%) more likely to block the original fictional accounts rather than the informational fictional accounts (though this difference is insignificant). This result suggests that social motivations are an important aspect of blocking decisions as well. Overall, this exercise points towards the conclusion that, if subjects only cared about receiving information in their follow-backs and blocking decisions, the level of polarization in social ties formation that we document would be lower (members of the in-group would be less likely to follow the experimental accounts, and out-group members would be less likely to block them).

6 Effects of political congruence over time & the National Football Team as an Affective Identity

6.1 Experimental effects over time

During our experimental period (August to December 2022), two key events naturally changed the perceived salience of political and non-political identities for the Brazilian public: the official electoral calendar (the second round of voting happened on October 30th, and the campaign period ended a day earlier) and the 2022 FIFA World Cup, which started twenty-one days after the election ended. This allows us to conduct two tests to complement our main results. First, we hypothesize that the salience of the political dimension of identity would be at its apex during electoral times, potentially increasing political homophily online. Second, the World Cup, which is widely popular and represents a major national symbol in Brazil, could have made a common national identity salient, reducing political homophily. Arguably, and based on established literature (Depetris-Chauvin et al., 2020), sharing a national identity could be a more powerful cohesive force in comparison to sharing

a football team, especially when there is a positive shared experience, such as winning a match.

In this section we answer two questions to test our hypothesis. First, did online political homophily in our experiment decrease after the election, when the salience of this dimension declined? Second, did the shared national experience of the World Cup foster cross-partisan connections, and if so, did this effect depend on whether the results of the national football team were positive or negative?

We show follow-back results over time in Figure 4. We plot the estimated effect of political congruence when pooling all experimental accounts within a given period: before, during, or after the official electoral period. As explained in Section 4, we restrict the sample to subjects who remained active throughout the experiment (i.e., tweeted on the seven days before each wave in which they were treated) to avoid a mechanical fall in follow-backs due to a fall in Twitter activity. With this restriction, our sample has 27,701 observations. Equivalent results for blocks are in Appendix Figure C.1.[21]

We first document a small but significant decline in the effect of political congruence after the official electoral period, compared to before and during that period. In the earlier period (before and during the election), sharing political identity increased the likelihood of follow-backs by 21.2 pp, while this effect is 16.8 pp after the election. The implied difference of 4.4 pp is significant at the 5% level. We interpret this result as evidence that the decline in the salience of politics—which naturally happens after the election—reduced political homophily.[22] We find a similar fall in the relevance of political congruence for blocks, reinforcing the interpretation that after the election there is a reduction in the role of political congruence in forming social media ties.

Nevertheless, political congruence still has a large effect on both follow-backs and blocks even after the election. Does the effect of political congruence diminish when a shared national identity becomes salient through the World Cup? Fostering inter-group cohesion (in particular, cross-partisan) is an important focus of the literature, which often points to other shared identities as a way of creating such inter-group ties (Voelkel et al., 2024; Depetris-Chauvin et al., 2020). Contrary to the prediction of this literature, we find that the World Cup—at most—had a very small effect on reducing political homophily or preventing politically motivated blocks.

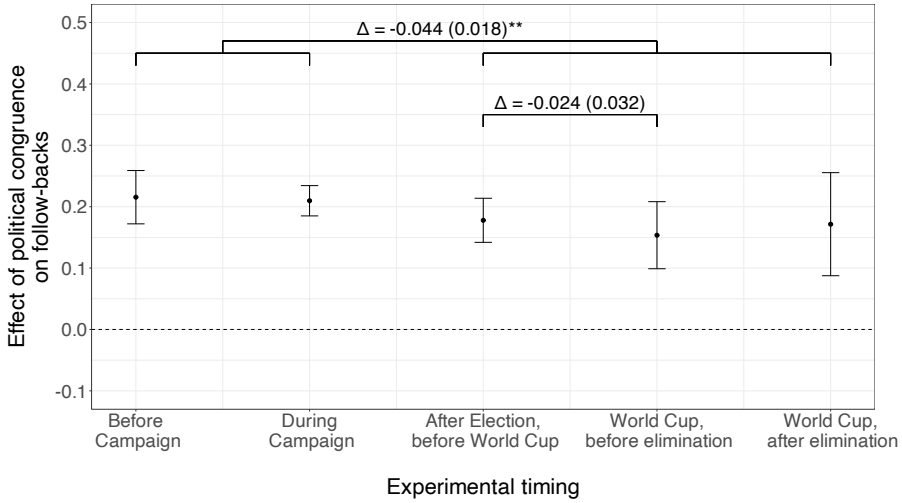
To show this, we focus on the post-election period. The World Cup started 21 days after the election ended. First, comparing the effect of political congruence on follow-backs before and after the World Cup started, we find a remarkably stable effect: political congruence increases follow-backs by 17.8 pp after the elections but before the World Cup, and by 16.3 pp after the beginning of the World Cup (p-value = 0.51).

However, as argued by Depetris-Chauvin et al. (2020), it could be that the cohesive effect

[21] Appendix Table B.8 shows that the static results from the previous section are unchanged when we do this and alternative sample restrictions, considering that some subjects naturally became inactive on Twitter throughout the experimental period.

[22] Appendix Figure C.2 shows that Google searches of election-related terms indeed track the official campaign period, supporting the claim that politics is more salient during the campaign period.

Figure 4: Effect of Congruence in Political Identity on Follow-Backs at Different Times



Notes: The figure displays point estimates and 95% confidence intervals for the effect of congruence in political identity on follow-backs for different sets of experimental waves, ordered by period: before the official electoral period; during the electoral period; and after the electoral period. The after-election period is further divided into before the beginning of the World Cup, during the World Cup and when results for Brazil were positive, and after Brazil's elimination from the Tournament. Timing details are in Appendix Figure A.1. The sample pools data from all experimental waves within each period, restricting the analysis to subjects who were always active during the experimental period (i.e., who tweeted in the seven days before being treated every time they were treated). This gives us a total of 27,701 observations. The brackets above the point estimates display estimates and standard errors (in parentheses) for the difference in the effect of political congruence between the signaled periods. Standard errors are clustered at the bot-account level. Significance codes: $^{***}: p < 0.01$, $^{**}: p < 0.05$, $^{*}: p < 0.1$.

of shared national identity only occurs while the results of the national team are positive. We can test this hypothesis using Brazil's 2022 World Cup performance, which included a positive streak (ranking first in its group and winning the round of 16) before elimination in the quarterfinals. If shared national identity reduces political homophily during successful periods, we would expect to see decreased political polarization during Brazil's winning streak. However, comparing the effect of political congruence on follow-backs between two periods—after the election but before the World Cup versus during Brazil's positive World Cup performance—we find no significant reduction in political homophily. Specifically, the effect of political congruence on follow-backs decreased by only 2.4 percentage points during Brazil's successful World Cup run, a difference that is not statistically significant (p-value = 0.45). This evidence, along with similar findings for blocks, suggests that even during periods of national team success, shared national identity had limited impact on fostering cross-partisan connections.

Given the importance attached to the World Cup by Brazilian nationals,[23] we interpret

[23]Over 85% of Brazilians claimed to be interested in the 2022 edition (TGMResearch, 2022). The national

this result as additional evidence that, in a polarized setting, political identities can overshadow other dimensions of identity (in this case, national identity), hindering the potential of these identities to foster cross-partisan ties.

6.2 Why did the World Cup not foster cross-partisan connections?

Why, contrary to the prediction in the literature, did the shared national identity made prominent by the World Cup not significantly foster cross-partisan connections and cohesion in our experiment? We hypothesize that the answer to this question is that, in our setting, political polarization also permeated the players and staff of the national team, affecting supporters' reactions to events in the tournament and preventing cross-partisan connections. As anecdotal evidence worldwide suggests, national symbols may have lost their power to foster nationalism as they become politicized. For example, President Trump claimed that "Americans were happy" about the women's soccer team defeat against Sweden in July 2021—which, he sustained, was because of their "wokeism",[24] and Brazil was not the exception.

In this section, we present evidence consistent with this explanation by analyzing observational data from Twitter during the tournament. We document that the level of criticism or cheering for players and the team coach during the tournament depended on the congruence or incongruence between supporters and players' political identity. We interpret this result as additional evidence that, in a polarized setting, political identities can overshadow other dimensions of identity (in this case, national identity), affecting social interactions in domains seemingly unrelated to politics.

6.2.1 Background, Data and Methods

We focus on tweets from users in Brazil during this country's World Cup matches. Using Twitter's API, we collect all Twitter users based in Brazil who tweeted or re-tweeted a status containing a pro-Lula or pro-Bolsonaro hashtag in the week before the first round of the 2022 presidential election (from September 25^{th} to October 1^{st}, 2022). The list of hashtags used is the same as the one used to obtain the subject pool in the field experiment (see Appendix A). This procedure gives us approximately 200 thousand individual accounts, of which 49.8% are classified as pro-Lula and 50.2% as pro-Bolsonaro.

After obtaining these accounts, we randomly sampled 10% of the pro-Lula and 10% of the pro-Bolsonaro accounts. Then, using Twitter's API, we collected all tweets and re-tweets sent by these accounts in intervals spanning two hours before and two hours after Brazil's matches in the World Cup. For Brazil's debut game against Serbia, this gives us 230,953 tweets by 17,701 users. We classify tweets' content using two methods: for straightforward categories (such as tweets about specific players), we rely on generic keyword search; for more

football team is a constitutive element of Brazilian identity (DaMatta, 1994)

[24] See https://www.vox.com/22600500/olympics-conservatives-simone-biles-anti-american and https://www.washingtonpost.com/sports/interactive/2024/american-sports-grievance-culture/ as two examples.

abstract categories (such as tweets about politics), we use a Bayesian Classifier algorithm.[25] For each of the games we are interested in, we estimate the difference in the likelihood that a pro-Lula user (versus a pro-Bolsonaro user) posts at least one tweet of a particular topic at every five-minute interval within the game period. The equation we estimate is:

$$\mathbb{1}\{\text{Tweeted about topic}\}_{it} = \lambda_t + \sum_{k=\underline{t}}^{\bar{t}} \beta_k \times \mathbb{1}\{k = t\} \times \mathbb{1}\{\text{Pro-Lula}\}_i + \varepsilon_{it} \tag{3}$$

where the dependent variable is an indicator equal to one if user i at time interval t tweeted about the topic under study (for instance, tweets about Neymar), λ_t represents interval fixed effects (which account for variation on the frequency of tweets during different moments in the game), and ε_{it} is an idiosyncratic error term. We are interested in the β_t, which represents the difference in the likelihood (in percentage points) that a pro-Lula user tweeted about the topic under study at interval t, relative to a pro-Bolsonaro user. We present results with standard errors clustered at the user level, and also report uniform confidence bands (using the plug-in method from Montiel Olea and Plagborg-Møller, 2019).

We focus our analysis on two of Brazil's games in the 2022 World Cup: the country's debut game against Serbia and its last game against Croatia. Both games are interesting case studies to illustrate how political identities shaped interactions with the national team during this tournament. First, in Brazil's opening game, Richarlison (a player publicly against President Bolsonaro) scored the two winning goals, while Neymar (who had backed Bolsonaro during the electoral campaign) left the game injured.[26] Given the opposite political affiliations of these two players—who had prominent roles in the match we analyze—we ask whether reactions to events related to them differed depending on supporters' political affiliations. Analyzing how pro-Lula and pro-Bolsonaro Twitter users reacted to these events may be informative about how political identity shaped interactions during the World Cup. Similarly, the game between Brazil and Croatia, which happened in the knock-out stage, presents interesting opportunities to study this topic. After a draw in regular time, Neymar scored a potentially winning goal at the end of the first half of overtime; Brazil would then suffer a goal and lose in the penalty shootout, getting disqualified from the tournament. Many Brazilians blamed the team's coach, Tite—who was considered a Lula supporter—, for the defeat.[27] Comparing fans' partisan reactions in this game (with a negative outcome)

[25] These algorithms are standard in the literature (e.g., Alrababa'h et al., 2021). Specifically, for each category of interest, we manually categorized a random sample of 2,000 tweets per match. We then used this data to train a Naïve Bayesian Classifier algorithm, predicting whether the remaining tweets in our dataset are in the same category.

[26] Before the World Cup, Richarlison's political positions were well-known and featured on important news outlets in the country, such as the newspaper O Globo (O Globo, 09-13-2022) and the news website UOL (UOL, 11-22-2022). Richarlison would also frequently post political content on his social media accounts and publicly adopt positions contrary to Jair Bolsonaro's government, such as becoming an Ambassador from the University of São Paulo in the fight against COVID and criticizing deforestation of the Amazon rainforest. In contrast, before the first election round, Neymar posted a video demonstrating his support for Bolsonaro (G1, 09-29-2022). During his campaign, Bolsonaro also rallied at Neymar's Institute in the city of Santos, and Neymar promised to dedicate his first World Cup goal to the former president.

[27] In the months preceding the World Cup, Tite had tried to adopt a politically-neutral position, arguing that "his activity is not mixed with politics" (Folha de S. Paulo, 12-04-2018). However, his avoidance

with the initial game (with a positive outcome) will allow us to understand whether such partisanship is similarly present in both types of events.[28]

6.2.2 Results

Overall, we find that players' and staff's political identities affected the way fans reacted to World Cup events involving them, which may prevent the cohesive effect of national identity that this football tournament could have sparked.

We start by discussing the first game, between Brazil and Serbia. Appendix Figure C.3 plots the number of tweets or re-tweets (on any topic) sent by pro-Lula or pro-Bolsonaro users in our sample in intervals of five minutes, highlighting the timing of the two goals of the game—both scored by Richarlison. Before the game started, the number of tweets sent by the two groups of Twitter users was remarkably similar, fluctuating around a mean of approximately 800 tweets every five minutes. After the game begins, there is a slight increase in the average number of tweets in both groups, but the trajectories of both groups remain the same. However, after Richarlison scored his first goal, we see a substantial spike in the number of tweets sent by pro-Lula accounts, not accompanied by a comparable increase in the number of tweets by pro-Bolsonaro accounts. The difference remains until after the game ends—the number of tweets by pro-Lula accounts only returns to the same level as those by pro-Bolsonaro accounts over one hour after Richarlison scored his second goal.

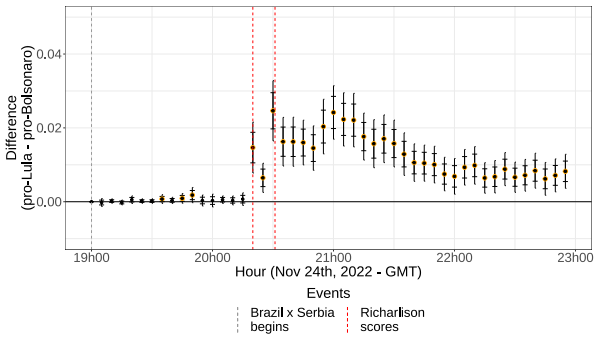
These trends suggest that the reactions to Richarlison's goals—which led Brazil to a victory in their first World Cup match—significantly differed between pro-Lula and pro-Bolsonaro Twitter users. Given that Richarlison was publicly critical of Jair Bolsonaro's government, one explanation for the phenomenon we document is that political identities mediated Brazilians' interactions with the national team, leading to fewer celebrations by individuals with political identities opposite to those of the players. We further investigate this hypothesis by estimating a version of Equation (3) for tweets about Richarlison. Figure 5a plots the difference in the likelihood that a pro-Lula user tweets about Richarlison relative to a pro-Bolsonaro user for every five-minute interval in our time window. The plot shows that, after Richarlison's first goal, pro-Lula accounts were between 1 and 3 pp more likely than pro-Bolsonaro accounts to tweet about Richarlison. Since before the goal, the likelihood of a tweet in each five-minute interval is 7.8%, this implies a sizable increase of 25%. This difference is statistically significant and remained throughout the entire time frame we analyzed (until two hours after the match).

This result reinforces the interpretation that Twitter users who shared political identities with Richarlison were likelier to engage with his goals. But what about the content of the tweets sent? We use two strategies to present some evidence in this regard. First, Appendix Figures C.4a and C.4b present word clouds with the most used words in tweets

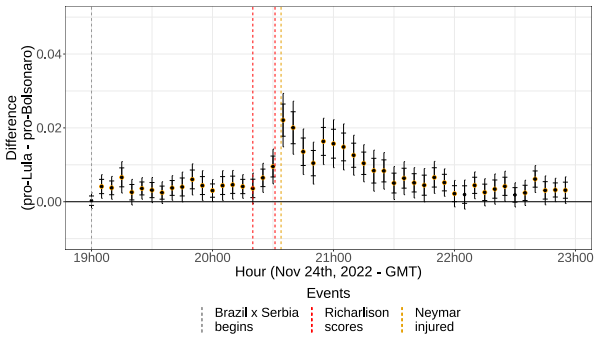
of meeting Bolsonaro (Folha de S. Paulo, 07-07-2019) and a previous photograph with Lula (Poder 360, 05-06-2021) led many Brazilians to consider the coach politically aligned with the left-wing candidate.

[28] Brazil's other three games offer fewer opportunities for analyses of partisan reactions, as there were either relevant events involving players of opposite affiliations happening too close in time, or no relevant events directly involving key polarized players.

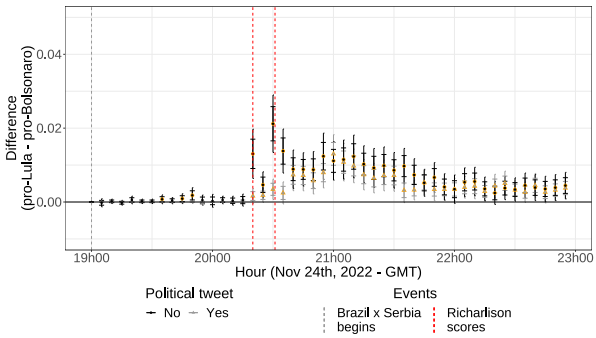
Figure 5: Difference in the number of tweets about Neymar and Richarlison between pro-Lula and pro-Bolsonaro Twitter users during Brazil × Serbia



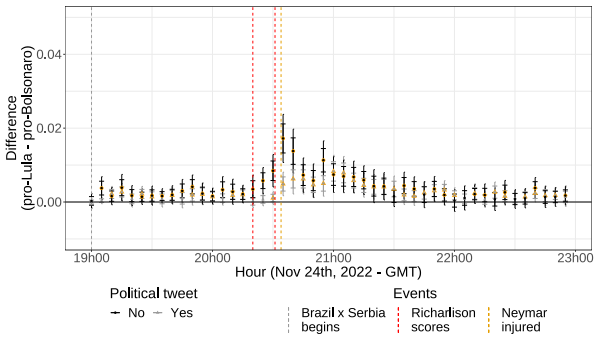
(a) Tweets about Richarlison



(b) Tweets about Neymar



(c) Tweets about Richarlison, Heterogeneity by Political Content of Tweet



(d) Tweets about Neymar, Heterogeneity by Political Content of Tweet

Notes: The top two figures plot the difference in the likelihood that a pro-Lula and pro-Bolsonaro account posts a tweet about a specific topic for every five-minute interval around the 2022 World Cup game between Brazil and Serbia. Figure 5a displays results for tweets about Richarlison, while 5b displays results for tweets about Neymar. We estimate Equation (3) as described in the text. Figures 5c and 5d plot a similar exercise but separate the analysis between tweets with or without political content. To classify tweets according to their content, we use a Bayesian Classifier algorithm. In all cases, data comes from a 10% random sample of all Brazilian Twitter users that tweeted or re-tweeted a status containing a pro-Lula or pro-Bolsonaro hashtag in the week before the first round of the 2022 presidential election. The error bars with ticks represent 95% confidence intervals, while the extended bars represent 95% uniform sup-t confidence bands, estimated using Montiel Olea and Plagborg-Møller (2019)'s plug-in estimator. Standard errors are clustered at the user level. Point estimates marked in orange denote estimates significant at the 5% level (point-wise).

related to Richarlison during the game for pro-Lula and pro-Bolsonaro users, respectively. Comparing the words used by pro-Lula and pro-Bolsonaro accounts, we see that words related to Richarlison's social activism, such as "science" and "ambassador" (he was an ambassador in the fight against Covid for the University of São Paulo) appear relatively more frequently among pro-Lula users. Moreover, the words "Lula" and "voter" also appear frequently among pro-Lula accounts, revealing that these users highlighted the player's political affiliation in

their tweets.[29] Second, to analyze content more systematically, we use a Bayesian classifier algorithm to predict whether tweets in our sample had political content. Figure 5c reports the difference in the rate of tweets about Richarlison posted by pro-Lula and pro-Bolsonaro accounts, this time dividing tweets between those with or without political content. This analysis reveals that pro-Lula accounts are likelier than pro-Bolsonaro accounts to tweet about Richarlison in general and specifically by highlighting politics. Therefore, pro-Lula users were not only more likely to tweet about Richarlison after his goals but also often mentioned his political identity when doing so.

While Richarlison was critical of Bolsonaro, Neymar—possibly the most famous active Brazilian footballer at the time—publicly supported this candidate during the elections. Apart from not scoring in the first match, Neymar suffered an injury that caused him to miss the tournament's next two games. How did pro-Lula and pro-Bolsonaro Twitter users react to Neymar during this game, especially after his injury? To answer this question, we first repeat the previous analysis focusing on tweets about Neymar. Results are displayed in Figures 5b (overall differences between pro-Lula and pro-Bolsonaro users), 5d (heterogeneity by political content of tweets). Pro-Lula users were slightly more likely to tweet about Neymar since the beginning of the game. Moreover, after Neymar's injury, pro-Lula users were likelier to tweet about this player, including with political content.

In particular, we find that pro-Lula users were likely to celebrate Neymar's injury. Using the same Bayesian classification method as before, we classify tweets about Neymar in our sample according to whether the tweets express positive sentiments about this event (which could have prevented one of the team's key players from playing in the remainder of the tournament).[30] Figure 6 shows the percentage of users in the sample that posted any tweet classified as celebrating Neymar's injury, according to their political affiliation. By the end of the match, 6% of pro-Lula accounts had posted a tweet celebrating the injury, against

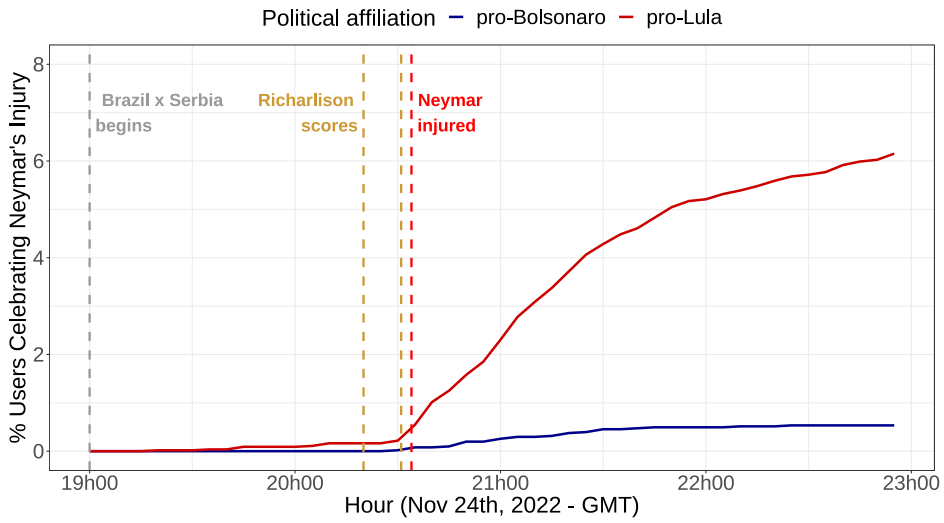
[29] We reproduce here a few illustrative examples of tweets sent by pro-Lula accounts in this context (translated by us):

- "Richarlison: 2 goals, not *bolsonarista* and doesn't owe to the IRS, I am so happy";
- "The only one who isn't *bolsominion*! Wonderful Richarlison"
- "And who scored? One of the few decent players of this tiny national team, Richarlison c'mon! He voted for Lula, knows where he comes from, and honors his country's jersey!"

[30] We reproduce below some of the tweets explicitly celebrating Neymar's injury as examples:

- "Who needs Neymar when they have Richarlison? Apart from being against Bolsonaro, he has no debt with the IRS."
- "Neymar is crying, I'm smiling."
- "Neymar got hurt, cries, and supporters shout: 'So what? I'm not an orthopedist'."
- "Brazil won, Richarlison scored, Neymar left the game crying. I couldn't be happier!"
- "The game got so good without Neymar, the tax evader who supports a coup. I hope he doesn't return until the end of the World Cup."
- "The tax evader is out of the World Cup? I can't believe God can be that good."

Figure 6: Cumulative percentage of Twitter users that posted tweets celebrating Neymar's injury by user's political affiliation



Notes: The figure plots the cumulative percentage of Twitter users in our sample that posted any tweet classified as celebrating Neymar's injury, by their political affiliation (pro-Lula or pro-Bolsonaro), during Brazil's debut match against Serbia in the 2022 FIFA World Cup. Data comes from a 10% random sample of all Brazilian Twitter users that tweeted or re-tweeted a status containing a pro-Lula or pro-Bolsonaro hashtag in the week before the first round of the 2022 presidential election. To classify tweets according to whether they celebrated Neymar's injury, we use a Bayesian Classifier algorithm as described in the text.

0.5% of pro-Bolsonaro users. Therefore, we find that pro-Lula users are disproportionately more likely to tweet celebrating Neymar's injury. Since Neymar was a public supporter of Bolsonaro at the time, this result shows in an even more surprising way that polarization affected how Brazilians interacted with the national football team—in this case, by celebrating an event that had the potential to negatively affect the performance of the national team.[31]

The analyses of tweets about Neymar and Richarlison point to the same conclusion. In both cases, we find significant differences in interactions between pro-Lula and pro-Bolsonaro users, suggesting that, in a polarized setting such as Brazil at the time, political identities mediated how Brazilians interacted with players of the national team. Combined with the results from the experiment, this case study illustrates that, in contexts of high affective polarization, political identities may overshadow other dimensions of identity, leading to an erosion of social ties across partisan lines and, in this case, to changes in how individuals engage with the national team. In particular, political identities may reduce the potential of collective experiences—support of the national team—to increase social cohesion.

This game symbolizes a positive result for Brazil—which the literature argues tends to be more effective in fostering national cohesion. An opposite and interesting case is the match between Brazil and Croatia, which happened during the knockout stage (i.e., the

[31] Word clouds of tweets about Neymar by each of the two political groups reinforce these conclusions (Appendix Figures C.4c and C.4d).

loser would be eliminated from the World Cup). After a draw in regular time, Neymar scored a potentially winning goal in the first half of overtime. However, Croatia would draw the game and win in the penalty shoot-out, eliminating Brazil. Appendix Figure C.5 shows that, while pro-Lula users tweeted more frequently than pro-Bolsonaro accounts since the beginning of the game, there is no clear pattern after the most relevant events of the match (Brazil's goal and ultimate elimination).

To investigate patterns more deeply, we analyze how pro-Lula and pro-Bolsonaro users reacted to the elimination. We focus on criticism towards the Brazilian coach, Tite, who was widely considered a Lula supporter. Analyzing tweets about Tite by pro-Lula and pro-Bolsonaro accounts during and after the game, it is clear that pro-Bolsonaro accounts disproportionately posted tweets and re-tweets about Tite immediately after Brazil lost the penalty shoot-out, as shown in Appendix Figure C.6a. After the game ended, pro-Bolsonaro accounts were, on average, 1-2 pp more likely to tweet about Tite than pro-Lula accounts every five minutes. This difference remained for approximately two hours after the game ended. The word clouds of tweets posted after the end of the game, displayed in Appendix Figure C.7, also suggest that both pro-Lula and pro-Bolsonaro accounts posting about Tite were critical of the coach. For example, both groups frequently cite the word "vestiário" (locker room), criticizing the fact that Tite went to the locker room immediately after the game instead of talking to the players. In addition, there are other generic negative words (such as arrogant, hate, and dumb) in both groups. However, the word cloud of pro-Bolsonaro accounts contains several words related to the coach's (alleged) political affiliation: words such as "comunista" (communist) and "Lula" appear frequently and do not appear among pro-Lula tweeters. On the other hand, "Neymar" frequently appears among pro-Lula tweets, suggesting that this pro-Bolsonaro player was a target of criticism from this group. To analyze this pattern more systematically, we divide tweets about Tite between those with or without political content. Once again, this analysis reveals that political identities shaped reactions to the World Cup. In this case, pro-Bolsonaro criticism of Tite was more frequent and often focused on political differences.

6.3 Discussion

Overall, the two case studies discussed in this section illustrate situations in which political identities mediated interactions with the Brazilian national team during the 2022 World Cup—which happened in a context of intense affective polarization. Importantly, we show that this effect happened both in a celebratory context (after Richarlison's goals) and a loss (after the elimination). In the case of a defeat, the mediating role of political identity on criticism against the Brazilian coach is expected, given that negative results tend to reinforce out-group animosity (Hewstone et al., 2002). On the other hand, the fact that shared political identities also impact reactions in a win is particularly relevant given the evidence that national teams' victories enhance social cohesion (Depetris-Chauvin et al., 2020), or that the success of out-group football players may reduce animosity towards members of this group (Alrababa'h et al., 2021). Our results suggest that this effect may be reduced in the context of affective polarization, given that supporters only identify with players who

share their political identity. This evidence helps to explain why the political homophily in the experiment did not fall during the World Cup despite the prediction derived from the literature. It also complements our experimental analysis by documenting another context in which political identity hinders interactions based on another shared identity. Indeed, while the World Cup could be seen as an opportunity to increase cohesion through shared experience, political differences made it harder for supporters to identify with the national team.

7 Conclusion

Political homophily in online settings seems to have been growing recently, and there is an intense debate about the potential consequences of this phenomenon on social interactions. We contribute to this debate by conducting a field experiment and analyzing observational data on Twitter to study the interplay between political identity and football club preference—a relevant dimension of identity for many Brazilians—in forming social ties in a natural environment: Twitter. Both dimensions of identity are relevant to forming ties, but the effect of sharing political identity is relatively more important.

Our results show that in a setting of intense affective polarization, political identities are capable of overshadowing other dimensions of identity in the formation of ties, to the point that signaling political identity undermines connections that could have been formed if such identity had not been signaled.

Online political homophily and animosity towards counter-partisans did not fall significantly even with a shared national experience—the World Cup. Even in this context, we find that political identities overshadowed the shared national identity and provide suggestive evidence that the reason for that is that polarization also permeated the players of the national team. While sportive events have the potential to foster national identity and social cohesion, this potential is hindered in a polarized context since interactions with the national team become mediated by political identities.

This observation has important implications. First, by showing that one potential consequence of affective polarization is to overshadow other dimensions of identity, we suggest one mechanism through which affective polarization may affect social interactions. In our setting, we show that people not only sort in terms of their political preference but also reduce the relative importance they attach to other dimensions of identity in forming ties. This behavior would lead people to have fewer opportunities to be in contact with dissenting views or have collaborative contact with politically opposite individuals, potentially changing people's attitudes and values, ultimately increasing segregation and polarization.

Moreover, this result has implications for the debate on the relationship between social media and polarization. Many analysts argue that social media amplifies polarization by creating echo chambers (Sunstein, 2018). Our experiment shows that online echo chambers are created not only via algorithmic suggestions or the reproduction of relationships outside of social media but also via individuals actively choosing to connect with those politically similar. This type of sorting may also reduce the exposure of individuals to dissenting views,

further contributing to polarization.

However, although limited, we also show that sharing affective identity—a preference for the same football club—still fosters ties in our setting, even among politically opposite individuals. This finding may seem at odds with the previous one, but they are consistent with each other. Signaling political identity reduced the effect of affective congruence, overshadowing this dimension of identity. However, the positive effect of sharing an affective identity was still present despite being small. This observation suggests that other dimensions of shared identity—in particular, preference for a football club—have the potential to reduce politically-induced societal divides. This result is particularly relevant considering that the subjects in our sample are politically engaged and that the experiment took place during and right after an election period. Such potential of other shared identities could be explored in interventions to reduce affective polarization, such as in conversations between supporters of opposing parties. Moreover, the positive effect of shared football clubs appeared even when a fictional account signaled political identity more saliently. Thus, highlighting similarities across other identities may be an avenue to reduce political animosity and foster ties across partisan lines. An interesting direction for future research would be to analyze how shared identities can be best used to reduce political divides.

Finally, this paper has some limitations that suggest other possible directions for future research. First, one important question is whether the behaviors we documented are restricted to periods close to elections—where political identities are salient. Further research would be needed to understand under what conditions the type of political overshadowing we discuss is absent. Moreover, since our sample is made up of politically engaged individuals in Brazil, we are unable to assess whether the type of behavior we document would generalize to other individuals and other contexts. Yet, the main objective of the experiment was to study whether political identities could undermine the formation of ties due to other shared identities, particularly in a context of affective polarization. Demonstrating that this is indeed the case is fundamental to advancing our understanding of the consequences of affective polarization and the mechanisms that can reinforce or reduce such polarization.

References

Ajzenman, Nicolás, Bruno Ferman, and Pedro C Sant'Anna, "Discrimination in the Formation of Academic Networks: A Field Experiment on #EconTwitter," *American Economic Review: Insights*, Forthcoming.

Akerlof, George A and Rachel E Kranton, "Economics and identity," *The Quarterly Journal of Economics*, 2000, *115* (3), 715–753.

Alabarces, Pablo, *Futbologías: fútbol, identidad y violencia en América Latina*, CLACSO, Consejo Latinoamericano de Ciencias Sociales, 2003.

Alrababa'h, Ala, William Marble, Salma Mousa, and Alexandra A Siegel, "Can exposure to celebrities reduce prejudice? The effect of Mohamed Salah on islamophobic behaviors and attitudes," *American Political Science Review*, 2021, *115* (4), 1111–1128.

American Press Institute, "The news consumption habits of 16- to 40-year-olds," https://americanpressinstitute.org/the-news-consumption-habits-of-16-to-40-year-olds/# August 2022. Accessed: 2024-12-26.

Arabe, Isadora Bousquat, "Own goal: impact of soccer matches on domestic violence in Brazil," Master's thesis, Universidade de São Paulo 2022.

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

BBC News Brasil, "Petista mata amigo bolsonarista a facadas em discussão política; veja outros casos," https://www.bbc.com/portuguese/brasil-63152266 10-05-2022. Acessed: 07-06-2023.

Boxell, Levi, Matthew Gentzkow, and Jesse M Shapiro, "Cross-country trends in affective polarization," *The Review of Economics and Statistics*, 2022, pp. 1–60.

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.

Bursztyn, Leonardo, Georgy Egorov, Ingar K Haaland, Aakaash Rao, and Christopher Roth, "Justifying dissent," Technical Report, National Bureau of Economic Research 2022.

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

Chen, M Keith and Ryne Rohla, "The effect of partisanship and political advertising on close family ties," *Science*, 2018, *360* (6392), 1020–1024.

CNN, "Fatal shooting at a party in Brazil highlights soaring political tensions," https://edition.cnn.com/2022/07/11/americas/brazil-shooting-lula-bolsonaro-intl-latam/index.html 07-11-2022. Acessed: 07-06-2023.

Currarini, Sergio, Matthew O Jackson, and Paolo Pin, "An economic model of friendship: Homophily, minorities, and segregation," *Econometrica*, 2009, *77* (4), 1003–1045.

DaMatta, Roberto, "Esporte na sociedade: um ensaio sobre o futebol brasileiro," *Universo do futebol: esporte e sociedade brasileira. Rio de Janeiro: Pinakotheke*, 1982, pp. 19–42.

_ , "Antropologia do óbvio: Notas em torno do significado social do futebol brasileiro," *Revista USP*, 1994, (22), 10–17.

DataSenado, "Redes Sociais, Notícias Falsas e Privacidade na Internet," 2019.

Depetris-Chauvin, Emilio, Ruben Durante, and Filipe Campante, "Building nations through shared experiences: Evidence from African football," *American Economic Review*, 2020, *110* (5), 1572–1602.

Enikolopov, Ruben, Maria Petrova, Gianluca Russo, and David Yanagizawa-Drott, "Socializing Alone: How Online Homophily Has Undermined Social Cohesion in the US," *Available at SSRN*, 2024.

Epstein, Robert and Ronald E Robertson, "The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections," *Proceedings of the National Academy of Sciences*, 2015, *112* (33), E4512–E4521.

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

Folha de S. Paulo, "Tite se recusa a falar sobre Bolsonaro e cobra de Messi respeito," https://www1.folha.uol.com.br/esporte/2019/07/tite-se-recusa-a-falar-sobre-bolsonaro-e-cobra-de-messi-respeito.shtml 07-07-2019. Acessed: 02-03-2023.

_ , "Tite se recusa a encontrar Bolsonaro antes da disputa da Copa América," https://www1.folha.uol.com.br/esporte/2018/12/tite-se-recusa-a-encontrar-bolsonaro-antes-da-disputa-da-copa-america.shtml 12-04-2018. Acessed: 02-03-2023.

Fujiwara, Thomas, Karsten Müller, and Carlo Schwarz, "The effect of social media on elections: Evidence from the United States," Technical Report, National Bureau of Economic Research 2021.

G1, "Levantamento mostra que sete torcedores morreram durante brigas de torcidas em 2023," https://g1.globo.com/sp/sao-paulo/noticia/2023/07/11/levantamento-mostra-que-sete-torcedores-morreram-durante-brigas-de-torcidas-em-2023.ghtml 07-11-2023. Acessed: 07-26-2023.

_ , "Neymar declara apoio a Bolsonaro," https://g1.globo.com/sp/santos-regiao/eleicoes/2022/noticia/2022/09/29/neymar-declara-apoio-a-jair-bolsonaro.ghtml 09-29-2022. Acessed: 02-01-2023.

Gentzkow, Matthew, "Polarization in 2016," *Toulouse Network for Information Technology Whitepaper*, 2016, pp. 1–23.

Halberstam, Yosh and Brian Knight, "Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter," *Journal of Public Economics*, 2016, *143*, 73–88.

Hartman, Rachel, Will Blakey, Jake Womick, Chris Bail, Eli J Finkel, Hahrie Han, John Sarrouf, Juliana Schroeder, Paschal Sheeran, Jay J Van Bavel et al., "Interventions to reduce partisan animosity," *Nature Human Behaviour*, 2022, *6* (9), 1194–1205.

Hewstone, Miles, Mark Rubin, and Hazel Willis, "Intergroup bias," *Annual Review of Psychology*, 2002, *53* (1), 575–604.

Huber, Gregory A and Neil Malhotra, "Political homophily in social relationships: Evidence from online dating behavior," *The Journal of Politics*, 2017, *79* (1), 269–283.

Huddy, Leonie, Lilliana Mason, and Lene Aarøe, "Expressive partisanship: Campaign involvement, political emotion, and partisan identity," *American Political Science Review*, 2015, *109* (1), 1–17.

IPEC and O Globo, "Pesquisa de Opinião Pública sobre Torcidas de Futebol," 2022.

Iyengar, Shanto, Gaurav Sood, and Yphtach Lelkes, "Affect, not ideologya: A social identity perspective on polarization," *Public Opinion Quarterly*, 2012, *76* (3), 405–431.

_ , Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J Westwood, "The origins and consequences of affective polarization in the United States," *Annual Review of Political Science*, 2019, *22* (1), 129–146.

Jungherr, Andreas, "Twitter use in election campaigns: A systematic literature review," *Journal of Information Technology & Politics*, 2016, *13* (1), 72–91.

Kalin, Michael and Nicholas Sambanis, "How to think about social identity," *Annual Review of Political Science*, 2018, *21*, 239–257.

Kingstone, Peter and Timothy J Power, *Democratic Brazil divided*, University of Pittsburgh Press, 2017.

LAPOP, "AmericasBarometer," 2019.

Levy, Ro'ee, "Social media, news consumption, and polarization: Evidence from a field experiment," *American Economic Review*, 2021, *111* (3), 831–70.

Lowe, Matt, "Types of contact: A field experiment on collaborative and adversarial caste integration," *American Economic Review*, 2021, *111* (6), 1807–44.

McPherson, Miller, Lynn Smith-Lovin, and James M Cook, "Birds of a feather: Homophily in social networks," *Annual Review of Sociology*, 2001, pp. 415–444.

Meireles, Fernando, *genderBR: Predict Gender from Brazilian First Names* 2021. R package version 1.1.2.

Mosleh, Mohsen, Cameron Martel, and David Rand, "Psychological underpinnings of partisan bias in tie formation on social media," 2023.

_ , _ , Dean Eckles, and David G Rand, "Shared partisanship dramatically increases social tie formation in a Twitter field experiment," *Proceedings of the National Academy of Sciences*, 2021, *118* (7).

Mousa, Salma, "Building social cohesion between Christians and Muslims through soccer in post-ISIS Iraq," *Science*, 2020, *369* (6505), 866–870.

Murad, Maurício, "O lugar teórico da sociologia do futebol," *Revista Pesquisa de Campo-Núcleo de Sociologia do Futebol-UERJ*, 1995, (2).

Nielsen Sports, "World Football Report," 2022.

O Globo, "Richarlison critica uso político de camisa da Seleção: 'Faz perder a identidade'," https://oglobo.globo.com/esportes/noticia/2022/09/richarlison-evita-divisoes-politicas-de-camisa-da-selecao-faz-a-gente-perder-a-identidade.ghtml 09-13-2022. Acessed: 01-30-2023.

Olea, José Luis Montiel and Mikkel Plagborg-Møller, "Simultaneous confidence bands: Theory, implementation, and an application to SVARs," *Journal of Applied Econometrics*, 2019, *34* (1), 1–17.

Ortellado, Pablo, Marcio Moretto Ribeiro, and Leonardo Zeine, "Existe polarização política no Brasil? Análise das evidências em duas séries de pesquisas de opinião," *Opinião Pública*, 2022, *28*, 62–91.

Poder 360, "Aliados de Bolsonaro publicam fotos de Tite com Lula," https://www.poder360.com.br/brasil/aliados-de-bolsonaro-publicam-fotos-de-tite-com-lula/ 05-06-2021. Acessed: 02-03-2023.

Reiljan, Andres, "'Fear and loathing across party lines' (also) in Europe: Affective polarisation in European party systems," *European Journal of Political Research*, 2020, *59* (2), 376–396.

Reuters, "Bolsonaro backer kills Lula fan as Brazil election tensions mount," https://www.reuters.com/world/americas/bolsonaro-fan-kills-lula-backer-brazil-election-tensions-mount-2022-09-09/ 09-09-2022. Acessed: 07-06-2023.

Ronconi, Juan Pedro, "Divided for Good: Football Rivalries and Social Cohesion in Latin America," 2022.

Santoro, Erik and David E Broockman, "The promise and pitfalls of cross-partisan conversations for reducing affective polarization: Evidence from randomized experiments," *Science Advances*, 2022, *8* (25).

Sayyadiharikandeh, Mohsen, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer, "Detection of novel social bots by ensembles of specialized classifiers," in "Proceedings of the 29th ACM international conference on information & knowledge management" 2020, pp. 2725–2732.

Shayo, Moses, "Social identity and economic policy," *Annual Review of Economics*, 2020, *12*.

Sport Track and XP, "Convocados/XP Football Report," 2022.

Statista, "Twitter: Statistics & Facts," 2022.

_ , "Twitter users in Brazil from 2017 to 2025," https://www.statista.com/forecasts/1146589/twitter-users-in-brazil#statisticContainer 2022. Accessed: 10-24-2022.

Sunstein, Cass R, *Echo chambers: Bush v. Gore, impeachment, and beyond*, Princeton University Press Princeton, NJ, 2001.

_ , *# Republic: Divided democracy in the age of social media*, Princeton University Press, 2018.

Tajfel, Henri, "Social identity and intergroup behaviour," *Social science information*, 1974, *13* (2), 65–93.

_ , *Human groups and social categories*, Cambridge university press Cambridge, 1981.

_ **and John C Turner**, *The social identity theory of intergroup behavior*, Chicago: Nelson-Hall, 1986.

TGMResearch, "TGM Global World Cup Survey 2022," https://tgmresearch.com/football-world-cup-2022-in-brazil.html 2022. Acessed: 01-30-2023.

UOL, "Destaque da Seleção, atacante Richarlison vira voz política entre jogadores," https://congressoemfoco.uol.com.br/temas/esporte/destaque-da-selecao-atacante-richarlison-vira-voz-politica-entre-jogadores/ 11-22-2022. Acessed: 01-30-2023.

Van Bavel, Jay J and Dominic J Packer, *The power of Us: Harnessing our shared identities to improve performance, increase cooperation, and promote social harmony*, Little, Brown Spark, 2021.

Vicario, Michela Del, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi, "The spreading of misinformation online," *Proceedings of the National Academy of Sciences*, 2016, *113* (3), 554–559.

Voelkel, Jan G, Michael N Stagnaro, James Y Chu, Sophia L Pink, Joseph S Mernyk, Chrystal Redekopp, Isaias Ghezae, Matthew Cashman, Dhaval Adjodah, Levi G Allen et al., "Megastudy testing 25 treatments to reduce antidemocratic attitudes and partisan animosity," *Science*, 2024, *386* (6719), eadh4764.

Wagner, Markus, "Affective polarization in multiparty systems," *Electoral Studies*, 2021, *69*, 102199.

Yang, Kai-Cheng, Onur Varol, Pik-Mai Hui, and Filippo Menczer, "Scalable and generalizable social bot detection through data selection," in "Proceedings of the AAAI conference on artificial intelligence," Vol. 34 2020, pp. 1096–1103.

Zhuravskaya, Ekaterina, Maria Petrova, and Ruben Enikolopov, "Political effects of the internet and social media," *Annual Review of Economics*, 2020, *12* (1), 415–438.

Online Appendix to

"Rooting for the same team: Shared social identities in a polarized context"

Nicolás Ajzenman Bruno Ferman Pedro C. Sant'Anna

March 21, 2025

A Additional Information on Experimental Design

A.1 Procedures to Create Experimental Accounts

Table A.1: Procedures Used to Create the Experimental Accounts

Element of Profile	Procedure
Profile Picture	For the accounts that signal their preferred team, the profile picture is a photo of the team's logo in a flag inside a stadium; for the team-neutral accounts, the profile picture is a photo of the interior of a foreign football stadium during a football game (we chose photos in which the teams that were playing could not be identified). In all cases, we have a set of possible images, which are randomly chosen to construct each bot.
Name	Randomly generated by matching a list of the most common male first names and surnames in Brazil.
Bio	The Bio from the fictional account accounts contains two pieces of information: first, it either says "Supporter of team X" (if the account signals her preferred team) or "football fan" (if the account is team-neutral); second, it includes either the hashtag "#Lula2022" or "#Bolsonaro2022" (depending on the bot's political identity). For the politically-neutral accounts, we merely remove this second part.
Background Image	A landscape from the city where the account's preferred football team plays its home matches (and random city landscape for the football team-neutral accounts).
Location	The fictional account accounts' profiles do not include a location. 25.5% of subjects' profiles do not include a location.
Website	The fictional account accounts' profiles do not include a website. 82.7% of subjects' profiles do not include a website url.
Retweets	The fictional account account first re-tweets a post from an account related to her preferred football team or, in the case of team-neutral accounts, a general tweet about football (that isn't specific about any football team). Then, the account re-tweets a post from its preferred political candidate. The post must necessarily have more than 500 re-tweets and not include any misleading information or hate speech. This way, the first post that is seen when someone accesses the bot's profile is the one that signals political identity.
Followers	We asked a group of colleagues to follow the fictional account accounts before each experimental wave so that the fictional account accounts have some followers when subjects receive the notifications.
Following	One day before following the accounts randomly assigned to it, the fictional account account will follow a set of "elite" accounts related to its political identity and preferred team (for instance, it will follow the team's official profile, the profile of its preferred candidate and of some of its allies).

Notes: The table summarizes the procedures used to create the fictional account accounts. Figure 1 shows examples of accounts.

A.2 Pro-Lula and Pro-Bolsonaro Hashtags

Table A.2: List of pro-Lula and pro-Bolsonaro hashtags used to build the subject pool

Pro-Lula	Pro-Bolsonaro
#Lula2022	#Bolsonaro2022
#Lula22	#Bolsonaro22
#Lula13	#FechadoComBolsonaro
#LulaPresidente	#BolsonaroReeleito
#LulaNoPrimeiroTurno	#BolsonaroNoPrimeiroTurno
#VamosJuntosPeloBrasil	#BolsonaroOrgulhoDoBrasil
#JuntosComLula	#JuntosComBolsonaro
#BrasilComLula	#BrasilComBolsonaro

A.3 Football Club Rivalries

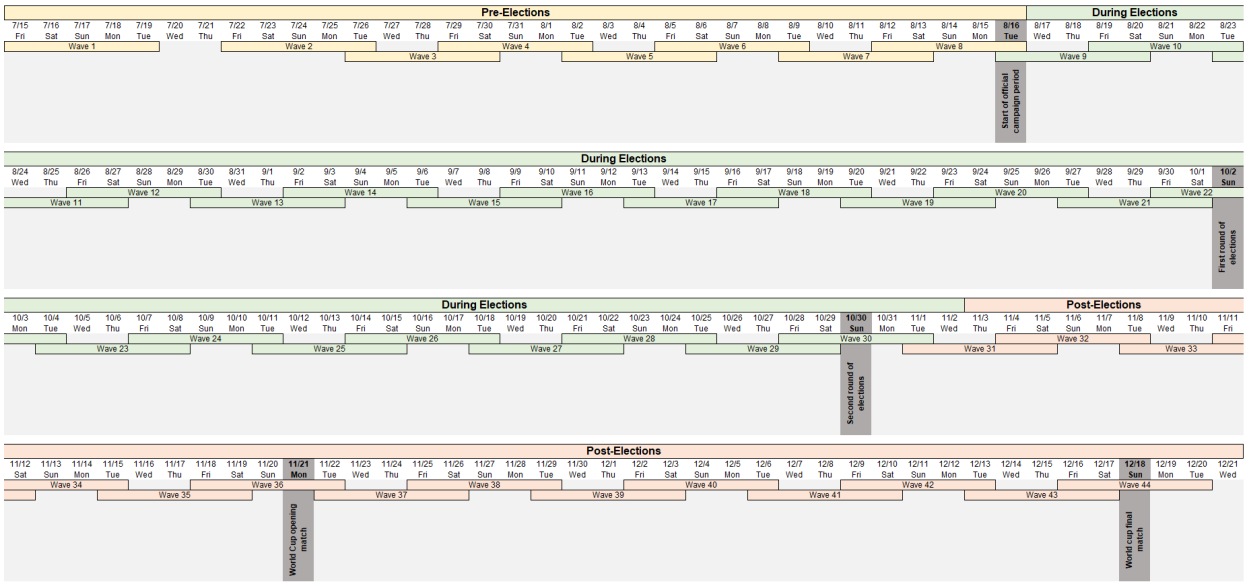
Table A.3: Football club rivalries

	Botafogo	Flamengo	Fluminense	Vasco	Corinthians	Palmeiras	Santos	São Paulo	Grêmio	Internacional
Flamengo	X	✓	X	X						
Vasco	X	X	X	✓						
Corinthians					✓	X	X	X		
Palmeiras					X	✓	X	X		
São Paulo					X	X	X	✓		
Grêmio									✓	X

Notes: The table displays the football club rivalries we considered when constructing the sample of subjects. The X mark indicates a rivalry. A fictional account that signals support for team A will only follow subjects whose preferred football club is either team A or team A's rival. We restricted ourselves to regional (inter-state rivalries). The clubs in the rows are the ones that a fictional account may support, while the clubs in the columns are the ones that subjects may support.

A.4 Experimental Timeline

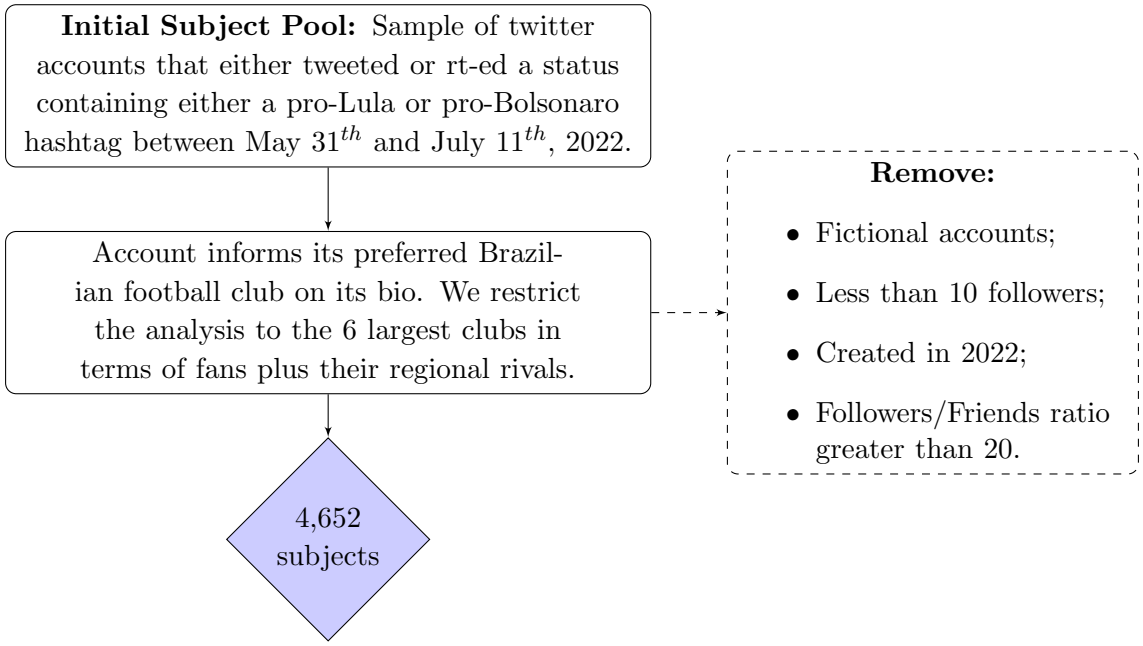
Figure A.1: Experimental Timeline



Notes: The table shows the experimental timeline. We consider that each wave starts at the moment in which the fictional account accounts follow the subjects. The table also shows the periods we define as before, during, and after the election period, along with relevant dates.

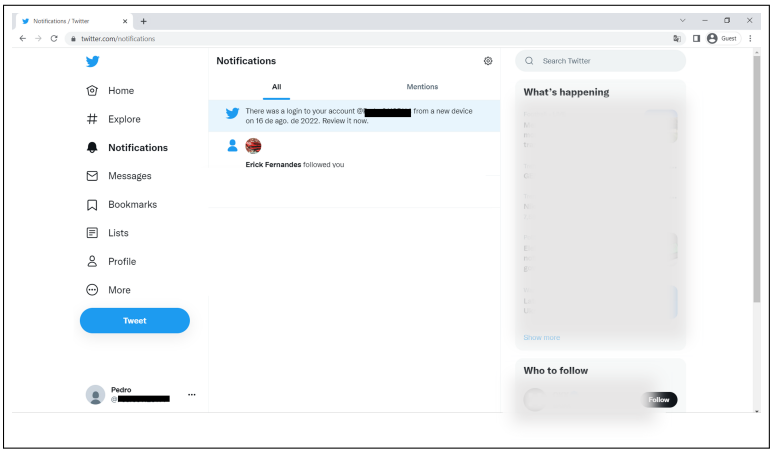
A.5 Procedure to Obtain the Subject Pool

Figure A.2: Procedure to obtain the subject pool

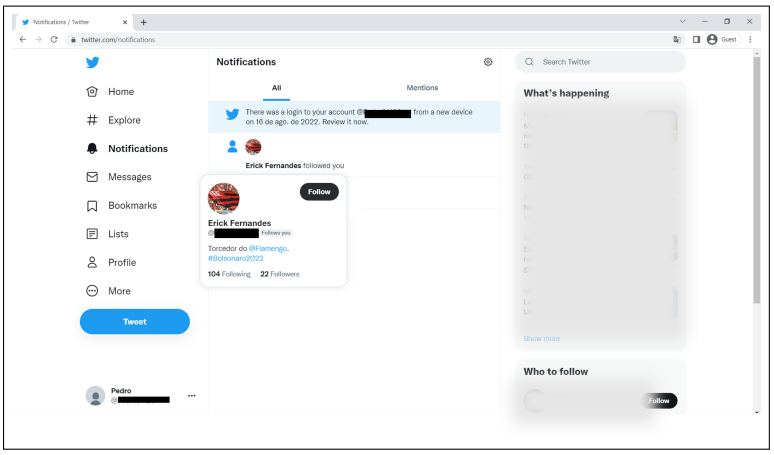


A.6 Follow Notification

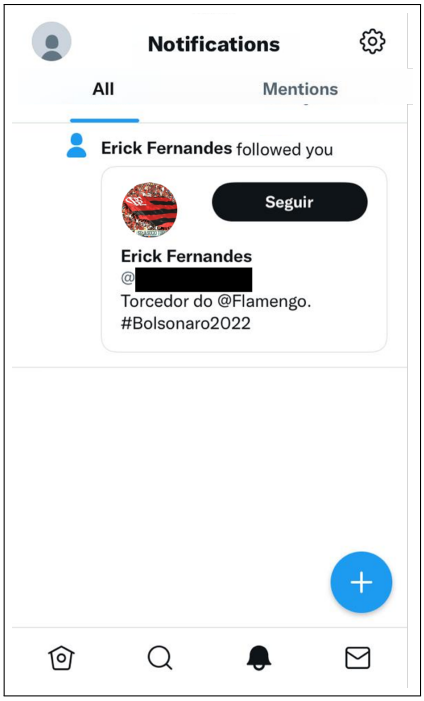
Figure A.3: Example of treatment notifications on desktop and mobile Twitter apps



(a) Desktop Notification



(b) Desktop Notification (after hovering the mouse's cursor over the bot's profile)

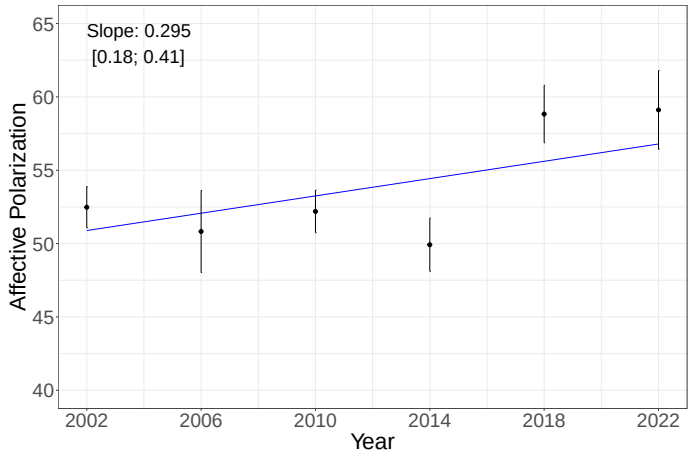


(c) Mobile app notification

B Additional Figures and Tables: Twitter Experiment

B.1 Affective Polarization in Brazil: Comparative Electoral Studies Survey Measure

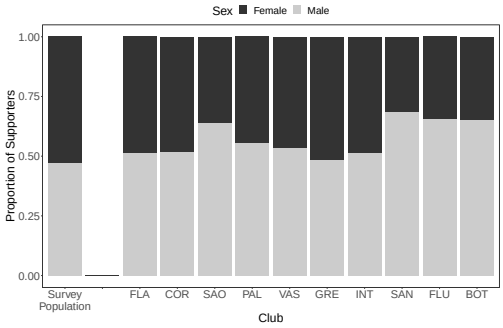
Figure B.1: Trends in Affective Polarization, Brazil (Boxell et al. (2022)'s method)



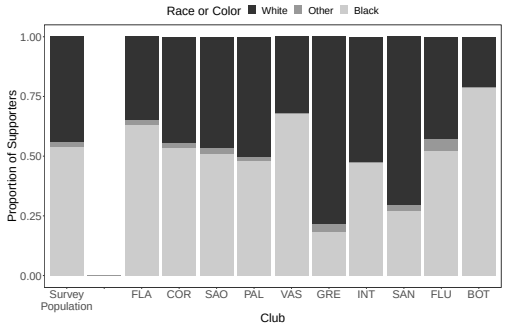
Notes: The figure presents trends in affective polarization in Brazil, using data from the Brazilian Electoral Study (BES), a national post-electoral survey undertaken since 2002. Following Boxell et al. (2022), we estimate affective polarization as the mean difference between in-party and out-party feeling among respondents who claim to identify with a given party. The question from which we construct the measures of in- and out-party feeling is "I'd like to know what you think about each of our political parties. After I read the name of a political party, please rate it on a scale from 0 to 10, where 0 means you strongly dislike that party and 10 means that you strongly like that party. If I come to a party you haven't heard of or you feel you do not know enough about, just say so. The first party is PARTY A." Error bars display 95% confidence intervals for the affective polarization index in each election year, and the blue line displays a fitted bivariate linear regression line with affective polarization as the dependent variable and the survey year as the independent one. The plot reports the slope (change per year) and estimated 95% confidence interval computed using heteroskedasticity-robust standard errors in the top-left.

B.2 Characteristics of Brazilian Football Club Supporters

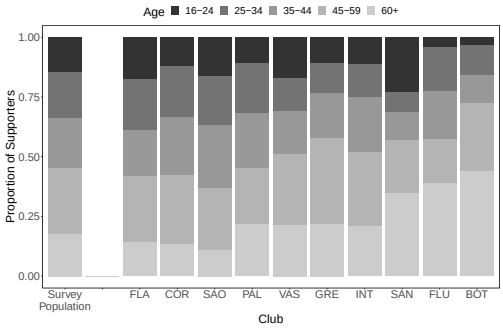
Figure B.2: Characteristics of Brazilian Football Club Supporters



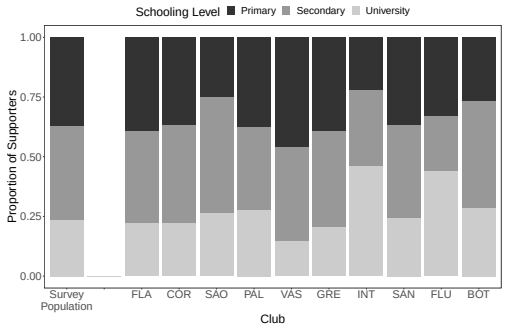
(a) Sex



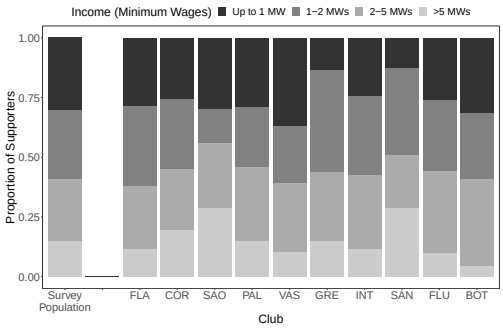
(b) Race



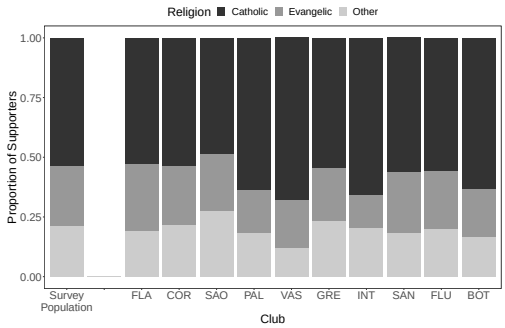
(c) Age



(d) Education Level



(e) Income



(f) Religion

Notes: The figures show the proportion of supporters of each of the six most popular Brazilian clubs and its rivals across socio-economic characteristics. Data comes from IPEC and O Globo (2022). The left-most bar in each plot shows the proportion with each characteristic in the survey population. Clubs are ordered by number of supporters.

B.3 Descriptive Statistics of the Subject Pool

Table B.1: Descriptive Statistics of the Subject Pool - Numerical Variables

Variables	Mean	Median	Std. Deviation	Min	Max	Obs.
Number of followers	2047.66	662	5685.92	11	141490	4652
Number of friends	2289.28	1057	5078.55	8	137451	4652
Number of statuses (tweets + rts)	25439.26	8050	58732.38	4	1665213	4652
Number of favorites (likes)	42152.19	17398	72984.66	0	1618281	4652
Year of account creation	2015.28	2016	4.66	2006	2022	4652
Botometer score	0.2	0.13	0.2	0	0.98	3878

Notes: The table shows summary statistics for the subject pool in the experiment. '*Botometer* score' is a number between 0 and 1 generated by the *Botometer* API, which determines the probability that each subject is classified as an automated account. A higher score means that the account is more likely to be automated.

Table B.2: Descriptive Statistics of Subject Pool

Variables	% Classified	N	%
Political Identity	100		
Bolsonaro		2069	44.48
Lula		2583	55.52
Affective Identity			
Corinthians	100	566	12.17
pro-Bolsonaro		156	27.56
pro-Lula		410	72.44
Palmeiras	100	485	10.43
pro-Bolsonaro		293	60.41
pro-Lula		192	39.59
São Paulo	100	403	8.66
pro-Bolsonaro		219	54.34
pro-Lula		184	45.66
Santos	100	165	3.55
pro-Bolsonaro		74	44.85
pro-Lula		91	55.15
Flamengo	100	1342	28.85
pro-Bolsonaro		641	47.76
pro-Lula		701	52.24
Vasco	100	447	9.61
pro-Bolsonaro		179	40.04
pro-Lula		268	59.96
Botafogo	100	245	5.27
pro-Bolsonaro		102	41.63
pro-Lula		143	58.37
Fluminense	100	172	3.70
pro-Bolsonaro		69	40.12
pro-Lula		103	59.88
Grêmio	100	258	5.55
pro-Bolsonaro		118	45.74
pro-Lula		140	54.26
Internacional	100	210	4.51
pro-Bolsonaro		80	38.10
pro-Lula		130	61.90

Variables	% Classified	N	%
Region	64.23		
Center-West		216	7.23
pro-Bolsonaro		117	54.17
pro-Lula		99	45.83
Northeast		379	12.68
pro-Bolsonaro		122	32.19
pro-Lula		257	67.81
North		123	4.12
pro-Bolsonaro		58	47.15
pro-Lula		65	52.85
Southeast		1746	58.43
pro-Bolsonaro		760	43.53
pro-Lula		986	56.47
South		418	13.99
pro-Bolsonaro		199	47.61
pro-Lula		219	52.39
Foreign		106	3.55
pro-Bolsonaro		67	63.21
pro-Lula		39	36.79
Gender	81.17		
Female		844	22.35
pro-Bolsonaro		268	31.75
pro-Lula		576	68.25
Male		2932	77.65
pro-Bolsonaro		1462	49.86
pro-Lula		1470	50.14
Has background pic.	100	3930	84.48
pro-Bolsonaro		1689	42.98
pro-Lula		2241	57.02
Has website	100	804	17.28
pro-Bolsonaro		253	31.47
pro-Lula		551	68.53

Notes: The table displays summary statistics for the subject pool. Figure A.2 describes the procedure used to obtain the subjects. The variable political identity is obtained accordingly to the hashtag used by the subject, while affective identity is obtained from information in the subject's bios. Region is created using self-declared information in the "location" field of the profile, which we recode to the regional level. % Classified is the percentage of all subjects for which we were able to obtain the variable. For each variable, we indicate the number of subjects (N) and the proportion of subjects in each category (the proportion is relative to the number of classified subjects). Finally, for each category, we show the proportion of subjects who are pro-Lula or pro-Bolsonaro. The variable Gender is obtained by using Brazilian Census data (organized by Meireles (2021)) to compute the proportion of men and women with each given name in the sample. A gender is assigned to a subject if at least 90% of his or her name's occurrences in the 2010 census were of an specific gender.

B.4 Balance, Attrition, and Take-up

Table B.3: Balance Table

	Treatment Arm								
	Both Dimensions				Affectively Neutral Accounts		Politically Neutral Accounts		
Variable	In-politics; In-affective	In-politics; Out-affective	Out-politics; In-affective	Out-politics; Out-affective	In-politics; Neutral-affective	Out-politics; Neutral-affective	Neutral-politics; In-affective	Neutral-politics; Out-affective	F Stat [p-value]
Number of followers	1,858.1	1,939.5	1,826.5	2,032.1	2,077.4	1,962	1,839	2,032.2	0.0137 [1.00]
	(4,899.3)	(4,816.2)	(4,584.2)	(5,387.3)	(6,220)	(5,001.6)	(5,067.6)	(5,987.4)	
Number of friends	2,190.1	2,191.7	2,074.4	2,302.9	2,313.7	2,221.4	2,132.8	2,312.9	0.0146 [1.00]
	(4,548.3)	(3,898.6)	(3,958.3)	(4,779.3)	(5,556)	(4,368.3)	(4,587.9)	(5,494.9)	
Number of statuses ('tweets + rts')	24,448	24,873.2	25,061.3	24,909.8	24,775.2	25,720.6	24,168.4	26,130.1	0.0055 [1.00]
	(55,867.4)	(51,507.3)	(56,480.8)	(53,622.5)	(51,148.7)	(50,935.8)	(60,306.6)	(64,734.1)	
Number of favorited statuses ('likes')	43,139.1	43,136.6	44,731.2	40,517.3	44,915.1	41,968.2	42,112	42,084.4	0.0154 [1.00]
	(87,867.7)	(73,385.3)	(83,492.9)	(63,641.3)	(82,595.3)	(69,698.8)	(81,036.1)	(71,119.9)	
Number of lists	4.024	4.164	4.133	4.33	4.056	3.619	3.157	4.184	0.0119 [1.00]
	(24.8)	(20.1)	(28.8)	(25.2)	(20.5)	(13.4)	(10)	(19.1)	
Account is verified	0.001	0.001	0.002	0.001	0.002	0.002	0	0.002	0.0129 [1.00]
	(0.033)	(0.028)	(0.043)	(0.023)	(0.039)	(0.046)	(0.018)	(0.04)	
Year of account creation	2,015.1	2,015.1	2,015.2	2,015.1	2,015	2,015.2	2,015.1	2,015	0.0104 [1.00]
	(4.599)	(4.689)	(4.582)	(4.653)	(4.585)	(4.59)	(4.599)	(4.655)	
Has background picture	0.839	0.843	0.841	0.838	0.839	0.83	0.833	0.838	0.0054 [1.00]
	(0.368)	(0.363)	(0.366)	(0.368)	(0.368)	(0.376)	(0.373)	(0.368)	
Gender (1=Female)	0.173	0.172	0.175	0.184	0.169	0.186	0.188	0.179	0.0138 [1.00]
	(0.378)	(0.377)	(0.38)	(0.387)	(0.375)	(0.389)	(0.391)	(0.383)	
Region									
Center-West	0.043	0.036	0.041	0.031	0.042	0.041	0.045	0.041	0.0214 [1.00]
	(0.202)	(0.186)	(0.198)	(0.172)	(0.2)	(0.199)	(0.207)	(0.198)	
Northeast	0.065	0.064	0.075	0.064	0.067	0.07	0.082	0.06	0.0311 [1.00]
	(0.246)	(0.244)	(0.264)	(0.245)	(0.25)	(0.255)	(0.275)	(0.238)	
North	0.021	0.017	0.024	0.023	0.023	0.021	0.032	0.016	0.0443 [1.00]
	(0.144)	(0.128)	(0.154)	(0.15)	(0.15)	(0.143)	(0.177)	(0.126)	
Southeast	0.311	0.335	0.303	0.329	0.311	0.295	0.329	0.335	0.0458 [1.00]
	(0.463)	(0.472)	(0.459)	(0.47)	(0.463)	(0.456)	(0.47)	(0.472)	
South	0.082	0.072	0.082	0.07	0.072	0.071	0.09	0.071	0.0283 [1.00]
	(0.274)	(0.258)	(0.274)	(0.255)	(0.259)	(0.258)	(0.286)	(0.257)	
Foreign	0.02	0.02	0.02	0.02	0.021	0.018	0.015	0.016	0.0113 [1.00]
	(0.139)	(0.141)	(0.141)	(0.14)	(0.144)	(0.133)	(0.12)	(0.126)	
Number of treated observations	3783	3761	3790	3794	3845	3833	3003	4385	
%	0.125	0.125	0.126	0.126	0.127	0.127	0.099	0.145	
Attrition (not treated)	379	415	396	384	346	363	356	378	0.0367 [1.00]
% of assigned to treatment	0.091	0.099	0.095	0.092	0.083	0.087	0.106	0.079	
Always active (tweeted every week)	2863	2830	2900	2923	2901	2908	2300	3353	0.0091 [1.00]
% of treated	0.757	0.752	0.765	0.77	0.754	0.759	0.766	0.765	
Active 1 day before treatment	2965	2948	2947	2994	3030	2969	2253	3435	0.0319 [1.00]
% of treated	0.784	0.784	0.778	0.789	0.788	0.775	0.75	0.783	

Notes: The table displays average and standard deviations for subject-level variables across the eight treatment arms in the experiment. The F-statistic is computed from a regression of the pre-treatment variable on the treatment indicators. For all pre-treatment variables, we cannot reject the null hypothesis of equality of means across all eight treatments. The row "Number of treated obs." shows the number of treated observations (i.e., accounts followed by a bot) for each treatment arm, while "%" shows the percentage treated among all treated participants. The row "Attrition" shows the number of participants assigned to each treatment that could not be treated (either because they de-activated their account, were suspended by Twitter, or chose to make their profile private). The row "Always active" show the number and proportion of subjects that tweeted at least once in the seven days before every experimental wave (not only those in which they were specifically treated), while "Active 1 day before treatment" show the number of subjects who had Twitter activity (tweets or re-tweets) one day before treatment.

Table B.4: Balance Table - Attrited subjects

	Treatment Arm								
	Both Dimensions				Affectively Neutral Accounts		Politically Neutral Accounts		
Variable	In-politics; In-affective	In-politics; Out-affective	Out-politics; In-affective	Out-politics; Out-affective	In-politics; Neutral-affective	Out-politics; Neutral-affective	Neutral-politics; In-affective	Neutral-politics; Out-affective	F Stat [p-value]
Number of followers	2,632.7	2,230.8	2,639.5	2,037.3	3,466.9	1,919.2	2,968.2	3,806.9	0.3768 [0.916]
	(6,739.1)	(5,147.4)	(6,707.1)	(4,794.8)	(8,311.2)	(3,758.9)	(8,108.2)	(9,896.4)	
Number of friends	2,913.8	2,474.6	2,944.9	2,357.5	3,666.6	2,281.9	2,975.4	3,832.6	0.3385 [0.936]
	(5,836.6)	(4,405.6)	(6,389.1)	(4,600.4)	(7,612.2)	(3,614.7)	(7,165.1)	(8,813.5)	
Number of statuses ('tweets + rts')	32,189.6	27,035.1	29,585.8	25,845.2	34,110.6	20,760.1	23,967.5	25,613.6	0.1369 [0.995]
	(102,192.6)	(54,890.2)	(97,306.3)	(58,804.4)	(105,093.3)	(35,136.7)	(55,960.2)	(48,499)	
Number of favorited statuses ('likes')	44,705.6	45,693	39,562.4	39,206.7	38,030.3	40,212.8	34,530.7	44,454.2	0.1215 [0.997]
	(68,391)	(83,868.8)	(65,426.9)	(66,475.1)	(62,098.3)	(66,975.1)	(66,181.8)	(72,858.7)	
Number of lists	3.011	2.949	4.869	2.227	4.107	1.683	2.408	2.705	0.1077 [0.998]
	(12.4)	(11.7)	(48.4)	(8.2)	(16)	(5.7)	(7.5)	(8.5)	
Account is verified	0	0	0	0	0	0	0	0.003	0.1114 [0.998]
	(0)	(0)	(0)	(0)	(0)	(0)	(0)	(0.052)	
Year of account creation	2,017.4	2,017.5	2,017.6	2,017.4	2,017.1	2,017.3	2,018	2,017.7	0.1194 [0.997]
	(4.514)	(4.545)	(4.687)	(4.653)	(4.742)	(4.704)	(4.527)	(4.455)	
Has background picture	0.868	0.913	0.866	0.854	0.874	0.889	0.832	0.862	0.2075 [0.984]
	(0.339)	(0.282)	(0.341)	(0.353)	(0.333)	(0.315)	(0.374)	(0.345)	
Gender (1=Female)	0.108	0.137	0.096	0.112	0.126	0.13	0.168	0.16	0.2175 [0.981]
	(0.311)	(0.345)	(0.295)	(0.316)	(0.333)	(0.336)	(0.374)	(0.367)	
Region									
Center-West	0.018	0.036	0.025	0.023	0.031	0.026	0.064	0.039	0.2406 [0.975]
	(0.135)	(0.187)	(0.157)	(0.151)	(0.173)	(0.161)	(0.244)	(0.193)	
Northeast	0.063	0.036	0.056	0.049	0.045	0.058	0.046	0.061	0.0716 [0.999]
	(0.244)	(0.187)	(0.229)	(0.217)	(0.207)	(0.234)	(0.21)	(0.239)	
North	0.026	0.024	0.033	0.026	0.014	0.026	0.029	0.033	0.0537 [1.00]
	(0.16)	(0.154)	(0.178)	(0.159)	(0.118)	(0.161)	(0.168)	(0.179)	
Southeast	0.311	0.328	0.263	0.339	0.287	0.304	0.26	0.298	0.1537 [0.993]
	(0.464)	(0.47)	(0.441)	(0.474)	(0.453)	(0.461)	(0.439)	(0.458)	
South	0.071	0.063	0.076	0.102	0.104	0.087	0.061	0.085	0.1438 [0.995]
	(0.258)	(0.243)	(0.265)	(0.302)	(0.306)	(0.283)	(0.239)	(0.28)	
Foreign	0.021	0.034	0.018	0.036	0.02	0.032	0.026	0.028	0.0751 [0.999]
	(0.144)	(0.181)	(0.132)	(0.188)	(0.139)	(0.176)	(0.159)	(0.164)	
Attrition (not treated)	379	415	396	384	356	378	346	363	
% of assigned to treatment	0.091	0.099	0.095	0.092	0.106	0.079	0.083	0.087	

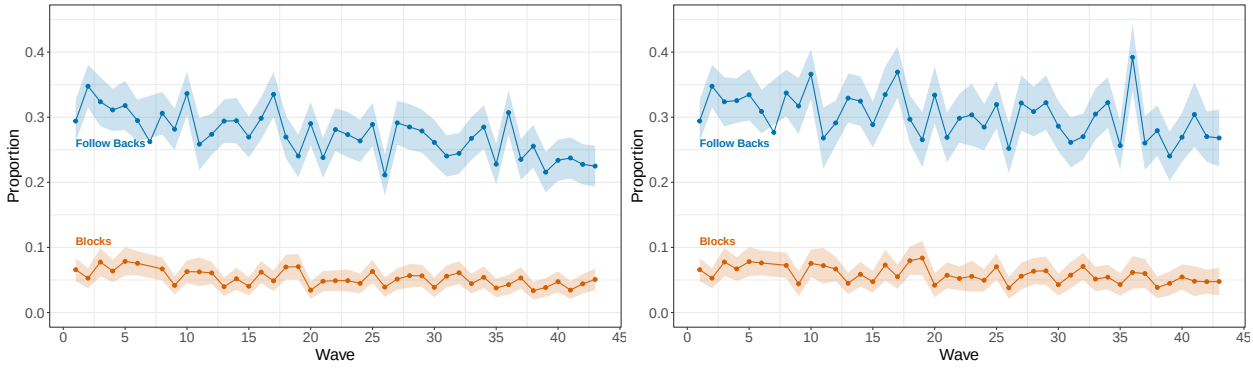
Notes: This table shows the average and standard deviations (in parentheses) of pre-treatment variables for subjects that suffered attrition at some point of the experiment. The last column in the table reports a F-test of joint equality of means across all treatment arms.

Table B.5: Differences between accounts that ever suffered attrition or did not

Variable	Never Attrited	Ever Attrited	T Stat [p-value]
Political identity (1=pro-Bolsonaro)	0.415	0.582	8.9246 [0.00]***
	(0.493)	(0.494)	
Number of followers	1,888.5	2,745.1	3.3879 [0.001]***
	(5,357.4)	(6,923.8)	
Number of friends	2,159	2,886.7	3.2766 [0.001]***
	(4,833.2)	(6,059.9)	
Number of statuses ('tweets + rts')	24,035.8	31,123.1	2.3951 [0.017]**
	(50,720.5)	(82,900.1)	
Number of favorited statuses ('likes')	41,183.4	46,212.4	1.7953 [0.073]*
	(72,139.2)	(74,208.5)	
Number of lists	4.143	4.235	0.0751 [0.94]
	(19.9)	(34.4)	
Account is verified	0.002	0.001	-0.3066 [0.759]
	(0.04)	(0.034)	
Year of account creation	2,014.9	2,016.8	10.3188 [0.00]***
	(4.568)	(4.748)	
Has background picture	0.841	0.861	1.5468 [0.122]
	(0.366)	(0.346)	
Gender (1=Female)	0.23	0.19	-2.3476 [0.019]**
	(0.421)	(0.393)	
Region			
Center-West	0.073	0.065	-0.662 [0.508]
	(0.261)	(0.247)	
Northeast	0.134	0.094	-2.7021 [0.007]***
	(0.34)	(0.292)	
North	0.038	0.057	1.68 [0.093]*
	(0.192)	(0.232)	
Southeast	0.578	0.615	1.5306 [0.126]
	(0.494)	(0.487)	
South	0.143	0.126	-1.0113 [0.312]
	(0.35)	(0.332)	
Foreign	0.033	0.043	0.9493 [0.343]
	(0.18)	(0.203)	
Number of observations	3782	851	
%	0.816	0.184	

Notes: The table compares average characteristics of subjects that never suffered attrition throughout all experimental waves ("never attrited") and those that suffered attrition at some point ('ever attrited'). Standard deviations are in parentheses. A subject is considered to have suffered attrition if we cannot find its account or cannot follow it on Twitter, which can happen if the user is suspended, deactivated its accounts, or made it private. The last column of the table displays the t-statistic and p-value of a test of difference in means for the respective variable between the two groups. Significance codes: $^{***}: p < 0.01$, $^{**}: p < 0.05$, $^{*}: p < 0.1$.

Figure B.3: Evolution of Treatment Take-up



(a) All Subjects (b) Active Subjects

Notes: The figures display the evolution of experimental take up across experimental waves. The first figures considers all subjects, while the second is conditional on subjects who were active (i.e., tweeted or re-tweeted) at least 24 hours before treatment. The shaded areas correspond to 95% confidence intervals.

B.5 Main Results: Comparison of Results across Treatment Arms and Robustness

Table B.6: Differences in Average Follow-Back Rate Across Treatment Arms

i/j	Out; Out		Out; No Signal		Out; In		No Signal; Out		No Signal; In		In; Out		In; No Signal		In; In	
Out-politics; Out-affective	$\Delta_{raw}(j-i)$ (Std. Error)	$\Delta_{FE,Controls}(j-i)$ (Std. Error)	0.009 (0.011)	0.012 (0.01)	0.044*** (0.009)	0.044*** (0.009)	0.069*** (0.011)	0.045*** (0.01)	0.203*** (0.014)	0.188*** (0.013)	0.164*** (0.011)	0.164*** (0.011)	0.209*** (0.013)	0.21*** (0.011)	0.249*** (0.012)	0.244*** (0.012)
Out-politics; No signal affective					0.035*** (0.012)	0.035*** (0.012)	0.061*** (0.012)	0.037*** (0.011)	0.194*** (0.014)	0.18*** (0.013)	0.155*** (0.012)	0.151*** (0.011)	0.2*** (0.015)	0.199*** (0.015)	0.24*** (0.012)	0.228*** (0.012)
Out-politics; In-affective							0.026** (0.012)	-0.002 (0.012)	0.159*** (0.014)	0.145*** (0.013)	0.12*** (0.013)	0.121*** (0.012)	0.165*** (0.014)	0.162*** (0.013)	0.205*** (0.012)	0.204*** (0.012)
No signal politics; Out-affective									0.134*** (0.013)	0.141*** (0.013)	0.095*** (0.012)	0.117*** (0.011)	0.139*** (0.014)	0.163*** (0.012)	0.179*** (0.012)	0.193*** (0.013)
No signal politics; In-affective											-0.039*** (0.015)	-0.033*** (0.013)	0.006 (0.016)	0.021 (0.014)	0.046*** (0.015)	0.042*** (0.013)
In-politics; Out-affective													0.045*** (0.014)	0.046*** (0.012)	0.085*** (0.011)	0.079*** (0.011)
In-politics; No signal affective															0.04*** (0.014)	0.031** (0.013)

Notes: The table displays differences in average follow-back rate between treatment arms. Each column or row represents one of the eight treatment arms in the experiment (the same ones displayed in Figure 2). The treatment arms are defined by whether fictional account and subject have congruent or incongruent identities in the political and affective (football club preference) dimensions. For each dimension (political or affective) we denote congruence using the term "in", and incongruence with the term "out" (as in "in-group" and "out-group" ties). A third option is that the fictional account does not signal the dimension. For each treatment arm, we first inform the relationship between fictional account and subject's political identity, and then affective (for example, "in; out" means that fictional account and subject share political identity and support rival clubs). Each table cell shows estimates and standard deviations for the difference in the average follow-back rate between the column and the row-treatment arm. In each cell, we report the raw difference between the groups, and the estimate including wave and strata fixed effects. Standard errors clustered at the bot-account level are in parentheses. Significance codes: $^{***}: p < 0.01$, $^{**}: p < 0.05$, $^{*}: p < 0.1$.

Table B.7: Differences in Average Blocking Rate Across Treatment Arms

i/j	Out; Out		Out; No Signal		Out; In		No Signal; Out		No Signal; In		In; Out		In; No Signal		In; In	
Out-politics; Out-affective	$\Delta_{raw}(j-i)$ (Std. Error)	$\Delta_{FE,Controls}(j-i)$ (Std. Error)	-0.017* (0.009)	-0.018** (0.009)	-0.061*** (0.008)	-0.06*** (0.008)	-0.124*** (0.008)	-0.123*** (0.008)	-0.137*** (0.007)	-0.127*** (0.007)	-0.136*** (0.007)	-0.135*** (0.007)	-0.139*** (0.007)	-0.141*** (0.008)	-0.141*** (0.007)	-0.14*** (0.007)
Out-politics; No signal affective					-0.044*** (0.008)	-0.044*** (0.007)	-0.107*** (0.006)	-0.104*** (0.006)	-0.12*** (0.006)	-0.121*** (0.006)	-0.119*** (0.006)	-0.119*** (0.006)	-0.123*** (0.006)	-0.123*** (0.006)	-0.124*** (0.006)	-0.123*** (0.006)
Out-politics; In-affective							-0.062*** (0.006)	-0.063*** (0.006)	-0.076*** (0.005)	-0.072*** (0.006)	-0.074*** (0.005)	-0.074*** (0.005)	-0.078*** (0.005)	-0.077*** (0.005)	-0.08*** (0.005)	-0.08*** (0.005)
No signal politics; Out-affective									-0.013*** (0.003)	-0.013*** (0.003)	-0.012*** (0.003)	-0.013*** (0.003)	-0.016*** (0.003)	-0.015*** (0.002)	-0.017*** (0.003)	-0.016*** (0.003)
No signal politics; In-affective											0.001 (0.003)	0 (0.003)	-0.002 (0.002)	-0.002 (0.002)	-0.004 (0.002)	-0.003 (0.002)
In-politics; Out-affective													-0.004* (0.002)	-0.003 (0.002)	-0.005** (0.002)	-0.006*** (0.002)
In-politics; No signal affective															-0.002 (0.002)	-0.002 (0.002)

Notes: The table displays differences in average blocking rate between treatment arms. Each column or row represents one of the eight treatment arms in the experiment (the same ones displayed in Figure 3). The treatment arms are defined by whether fictional account and subject have congruent or incongruent identities in the political and affective (football club preference) dimensions. For each dimension (political or affective) we denote congruence using the term "in", and incongruence with the term "out" (as in "in-group" and "out-group" ties). A third option is that the fictional account does not signal the dimension. For each treatment arm, we first inform the relationship between fictional account and subject's political identity, and then affective (for example, "in; out" means that fictional account and subject share political identity and support rival clubs). Each table cell shows estimates and standard deviations for the difference in the average blocking rate between the column and the row-treatment arm. In each cell, we report the raw difference between the groups (column − row), and the estimate including wave and strata fixed effects. Standard errors clustered at the bot-account level are in parentheses. Significance codes: $^{***}: p < 0.01$, $^{**}: p < 0.05$, $^{*}: p < 0.1$.

Table B.8: Main Results for Different Sub-samples: Experimental accounts that signal both dimensions of identity

Panel A: Follow Backs							
	Dependent Variable: Follow Backs (1 = Yes)						
	Full Sample			Never attrited	Tweeted every week	Active (1 day)	Unlikely to be automated
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Political congruence	0.1639***	0.1643***	0.1476***	0.1439***	0.1622***	0.1606***	0.1398***
	(0.0108)	(0.0106)	(0.0139)	(0.0139)	(0.0165)	(0.0166)	(0.0172)
Affective congruence	0.0437***	0.0424***	0.0512***	0.0473***	0.0597***	0.0551***	0.0532***
	(0.0087)	(0.0087)	(0.0114)	(0.0129)	(0.0145)	(0.0136)	(0.0154)
Political congruence × Affective congruence	0.0411***	0.0387***	0.0503***	0.0531***	0.0364*	0.0521**	0.0461*
	(0.0129)	(0.0127)	(0.0170)	(0.0184)	(0.0211)	(0.0200)	(0.0267)
Wave, Strata Fixed Effects	No	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	No	Yes	Yes	Yes	Yes	Yes
Observations	15,128	15,128	15,128	13,257	9,953	11,854	6,814
R^2	0.04856	0.08886	0.09909	0.09795	0.10527	0.10199	0.10824
Panel B: Blocks							
	Dependent Variable: Blocks (1 = Yes)						
	Full Sample			Never attrited	Tweeted every week	Active (1 day)	Unlikely to be automated
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Political congruence	-0.1355***	-0.1354***	-0.1267***	-0.1062***	-0.1193***	-0.1469***	-0.1329***
	(0.0072)	(0.0073)	(0.0081)	(0.0080)	(0.0092)	(0.0092)	(0.0115)
Affective congruence	-0.0611***	-0.0609***	-0.0652***	-0.0518***	-0.0623***	-0.0797***	-0.0859***
	(0.0076)	(0.0076)	(0.0093)	(0.0093)	(0.0115)	(0.0115)	(0.0139)
Political congruence × Affective congruence	0.0559***	0.0553***	0.0578***	0.0457***	0.0534***	0.0707***	0.0722***
	(0.0078)	(0.0078)	(0.0097)	(0.0096)	(0.0120)	(0.0121)	(0.0150)
Wave, Strata Fixed Effects	No	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	No	Yes	Yes	Yes	Yes	Yes
Observations	14,737	14,737	14,737	12,945	9,718	11,501	6,645
R^2	0.05768	0.06426	0.06790	0.05552	0.06102	0.07730	0.07583

Notes: The table presents regression estimates for the effect of sharing identities on follow-backs (Panel A) and blocks (Panel B), for different sub-samples of subjects, considering only the accounts that signaled both dimensions of identity. The sample excludes shadow-banned accounts, as pre-registered and discussed in the text. The first three columns show estimates using the full sample, estimating Equation (2) with and without wave and strata fixed effects and additional controls. The controls used are bot's football club, clubs' Google Trends index, subjects' region, gender, number of followers and number of tweets. The remaining columns perform similar estimates using sub-samples of subjects. A subject suffers attrition if we cannot follow it during a wave (because its account was de-activated, suspended, or made private). The sample of "never attrited" subjects is composed exclusively of subjects that did not suffer this type of attrition at any wave. Subjects that tweeted at least once in the seven days before every treatment wave are considered always active. Active subjects are those who tweeted or re-tweeted a status one day before treatment. Finally, the last column considers the sub-sample composed of subjects with below median score from the *Botometer* API (specifically, subjects with less than 13% chance of being automated accounts), which estimates the probability that a Twitter account is automated. Standard errors clustered at the fictional account account level are in parentheses. Significance codes: $^{***}: p < 0.01$, $^{**}: p < 0.05$, $^{*}: p < 0.1$.

Table B.9: Main Results for Different Sub-samples: Experimental accounts that signal a single dimension of identity

Panel A: Follow Backs, Affective Identity Only							
	Dependent Variable: Follow Backs (1 = Yes)						
	Full Sample			Never attrited	Tweeted every week	Active (1 day)	Unlikely to be automated
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Affective congruence	0.1337***	0.1413***	0.1454***	0.1548***	0.1747***	0.1604***	0.1483***
	(0.0133)	(0.0134)	(0.0187)	(0.0196)	(0.0212)	(0.0213)	(0.0211)
Wave, Strata Fixed Effects	No	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	No	Yes	Yes	Yes	Yes	Yes
Observations	7,388	7,388	7,388	6,583	4,983	5,688	3,500
R^2	0.02123	0.06732	0.08339	0.09017	0.09770	0.08595	0.08507
Panel B: Blocks, Affective Identity Only							
	Dependent Variable: Blocks (1 = Yes)						
	Full Sample			Never attrited	Tweeted every week	Active (1 day)	Unlikely to be automated
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Affective congruence	-0.0134***	-0.0126***	-0.0132***	-0.0113**	-0.0129**	-0.0135***	-0.0183**
	(0.0031)	(0.0032)	(0.0043)	(0.0046)	(0.0053)	(0.0048)	(0.0070)
Wave, Strata Fixed Effects	No	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	No	Yes	Yes	Yes	Yes	Yes
Observations	7,199	7,199	7,199	6,424	4,859	5,516	3,423
R^2	0.00253	0.01003	0.01773	0.01529	0.01757	0.02072	0.02001
Panel C: Follow Backs, Political Identity Only							
	Dependent Variable: Follow Backs (1 = Yes)						
	Full Sample			Never attrited	Tweeted every week	Active (1 day)	Unlikely to be automated
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Political congruence	0.2000***	0.1994***	0.1979***	0.1880***	0.1982***	0.2076***	0.1797***
	(0.0148)	(0.0147)	(0.0133)	(0.0135)	(0.0164)	(0.0162)	(0.0185)
Wave, Strata Fixed Effects	No	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	No	Yes	Yes	Yes	Yes	Yes
Observations	7,678	7,678	7,678	6,823	5,079	5,999	3,418
R^2	0.05092	0.08798	0.10159	0.09892	0.10616	0.10787	0.10123
Panel D: Blocks, Political Identity Only							
	Dependent Variable: Blocks (1 = Yes)						
	Full Sample			Never attrited	Tweeted every week	Active (1 day)	Unlikely to be automated
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Political congruence	-0.1225***	-0.1225***	-0.1224***	-0.1075***	-0.1188***	-0.1386***	-0.1236***
	(0.0062)	(0.0061)	(0.0060)	(0.0058)	(0.0074)	(0.0068)	(0.0085)
Wave, Strata Fixed Effects	No	Yes	Yes	Yes	Yes	Yes	Yes
Controls	No	No	Yes	Yes	Yes	Yes	Yes
Observations	7,492	7,492	7,492	6,668	4,961	5,830	3,324
R^2	0.05894	0.06775	0.07063	0.06323	0.07162	0.08179	0.08295

Notes: The table presents regression estimates for the effect of sharing identities on follow-backs (Panel A and C) and blocks (Panel B and D), for different sub-samples of subjects, considering only the accounts that signaled either affective (top two panels) or political (bottom two panels) identity. The sample excludes shadow-banned accounts, as pre-registered and discussed in the text. The first three columns show estimates using the full sample, estimating Equation (2) with and without wave and strata fixed effects and additional controls. The controls used are bot's football club, clubs' Google Trends index, subjects' region, gender, number of followers and number of tweets. Controls involving bot's football club are not included for the treatment arms with fictional accounts that only signal political identity. The remaining columns perform similar estimates using sub-samples of subjects. A subject suffers attrition if we cannot follow it during a wave (because its account was de-activated, suspended, or made private). The sample of "never attrited" subjects is composed exclusively of subjects that did not suffer this type of attrition at any wave. Subjects that tweeted at least once in the seven days before every treatment wave are considered always active. Active subjects are those who tweeted or re-tweeted a status one day before treatment. Finally, the last column considers the sub-sample composed of subjects with below median score from the *Botometer* API (specifically, subjects with less than 13% chance of being automated accounts), which estimates the probability that a Twitter account is automated. Standard errors clustered at the fictional account account level are in parentheses. Significance codes: *** : $p < 0.01$, ** : $p < 0.05$, * : $p < 0.1$.

B.6 Experiment with Fictional accounts with more Salient Political Identity

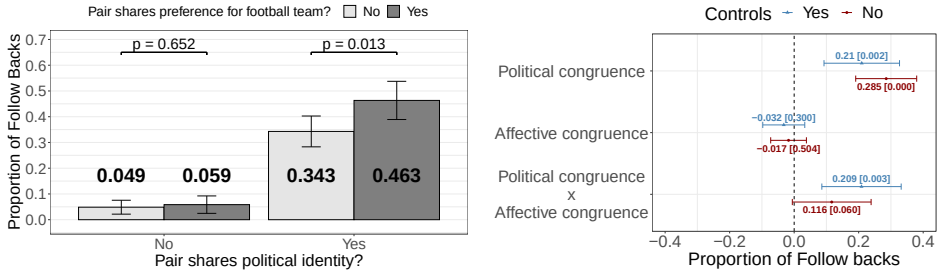
Figure B.4: Examples of Fictional Accounts - More salient political identity

(a) Pro-Bolsonaro; São Paulo supporter

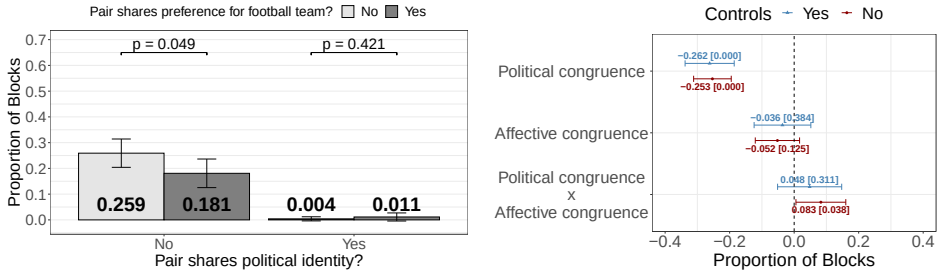
(b) Pro-Lula; Palmeiras supporter

Notes: The figures show examples of fictional account accounts used in the extra experiment, in which the political identity signal was more salient.

Figure B.5: Effect of shared political and affective identity on the formation of social ties: Fictional accounts with more salient political identity



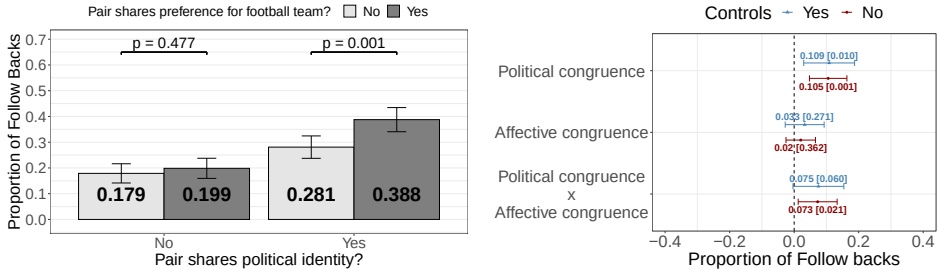
(a) Follow Backs



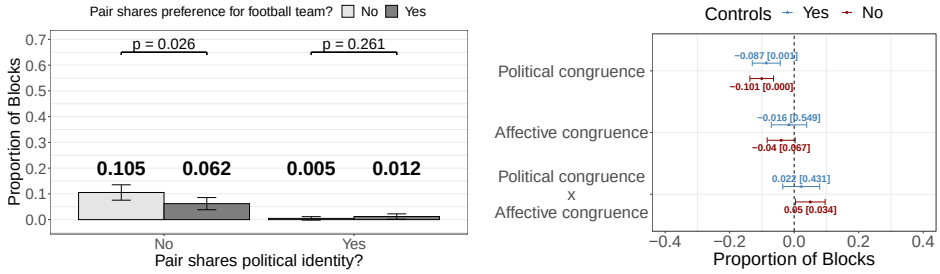
(b) Blocks

Notes: The figures show the effect of sharing political and affective (football club) identity on the rate of follow-backs and blocks for the experiment with fictional account accounts with a more salient political identity. The figure on the left shows the average rate of follow-backs or fictional accounts for the entire experiment, excluding shadow-banned accounts. The p-value on these plots is the p-value of a simple t-test of difference in means between the two groups indicated by the bracket. The left-hand side plot shows the coefficients estimated from equation (2), which includes wave and strata fixed effects. The controls used are the bot's football club, the google trend index of the clubs, subject's number of followers and statuses, interacted with the treatment indicator. The plots show 95% confidence intervals (error bar), coefficient estimates and p-values (in brackets). Confidence intervals and p-values are computed using standard errors clustered at the fictional account account level.

Figure B.6: Results of the main experiment for the same waves as experiment with more salient political identity



(a) Follow Backs



(b) Blocks

Notes: The figures show the effect of sharing political and affective (football club) identity on the rate of follow-backs and blocks for the fictional account accounts of the original experiment, restricting the analysis for the waves in which we conducted the extra experiment with fictional accounts with more salient political identity. The figure on the left shows the average rate of follow-backs or fictional accounts for the entire experiment, excluding shadow-banned accounts. The p-value on these plots is the p-value of a simple t-test of difference in means between the two groups indicated by the bracket. The left-hand side plot shows the coefficients estimated from equation (2), which includes wave and strata fixed effects. The controls used are the bot's football club, the google trend index of the clubs, subject's number of followers and statuses, interacted with the treatment indicator. The plots show 95% confidence intervals (error bar), coefficient estimates and p-values (in brackets). Confidence intervals and p-values are computed using standard errors clustered at the fictional account account level.

B.7 Other Robustness Exercises

Table B.10: Main Results Excluding Fictional accounts' Football Clubs

Panel A: Follow Backs, Affective Identity Only							
	Dependent Variable: Follow Backs (1 = Yes)						
Excluded Club:	-	Flamengo	Corinthians	São Paulo	Palmeiras	Vasco	Grêmio
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Affective congruence	0.1454***	0.1402***	0.1596***	0.1519***	0.1057***	0.1617***	0.1419***
	(0.0187)	(0.0236)	(0.0197)	(0.0202)	(0.0217)	(0.0193)	(0.0200)
Wave, Strata Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	7,388	5,167	6,567	5,756	6,148	6,649	6,653
R^2	0.08339	0.08352	0.08533	0.09057	0.08148	0.08547	0.08590
Panel B: Blocks, Affective Identity Only							
	Dependent Variable: Blocks (1 = Yes)						
Excluded Club:	-	Flamengo	Corinthians	São Paulo	Palmeiras	Vasco	Grêmio
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Affective congruence	-0.0132***	-0.0159***	-0.0119**	-0.0142***	-0.0145***	-0.0134***	-0.0116**
	(0.0043)	(0.0048)	(0.0049)	(0.0046)	(0.0053)	(0.0047)	(0.0046)
Wave, Strata Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	7,199	4,978	6,473	5,661	5,959	6,460	6,464
R^2	0.01773	0.01653	0.01857	0.02276	0.01953	0.01848	0.01998
Panel C: Follow Backs, Both Dimensions of Identity							
	Dependent Variable: Follow Backs (1 = Yes)						
Excluded Club:	-	Flamengo	Corinthians	São Paulo	Palmeiras	Vasco	Grêmio
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Political congruence	0.1476***	0.1429***	0.1434***	0.1497***	0.1520***	0.1500***	0.1493***
	(0.0139)	(0.0169)	(0.0153)	(0.0155)	(0.0149)	(0.0155)	(0.0151)
Affective congruence	0.0512***	0.0266*	0.0520***	0.0542***	0.0605***	0.0619***	0.0469***
	(0.0114)	(0.0150)	(0.0119)	(0.0134)	(0.0131)	(0.0118)	(0.0118)
Political congruence × Affective congruence	0.0503***	0.0693***	0.0583***	0.0466**	0.0331*	0.0387**	0.0522***
	(0.0170)	(0.0203)	(0.0183)	(0.0206)	(0.0187)	(0.0181)	(0.0176)
Wave, Strata Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	15,128	10,464	13,205	11,928	12,755	13,647	13,641
R^2	0.09909	0.09980	0.10082	0.10415	0.09794	0.09836	0.10029
Panel D: Blocks, Both Dimensions of Identity							
	Dependent Variable: Blocks (1 = Yes)						
Excluded Club:	-	Flamengo	Corinthians	São Paulo	Palmeiras	Vasco	Grêmio
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Political congruence	-0.1267***	-0.1159***	-0.1227***	-0.1283***	-0.1293***	-0.1284***	-0.1268***
	(0.0081)	(0.0101)	(0.0086)	(0.0096)	(0.0087)	(0.0090)	(0.0088)
Affective congruence	-0.0652***	-0.0530***	-0.0592***	-0.0682***	-0.0728***	-0.0639***	-0.0674***
	(0.0093)	(0.0107)	(0.0100)	(0.0109)	(0.0101)	(0.0103)	(0.0101)
Political congruence × Affective congruence	0.0578***	0.0437***	0.0527***	0.0599***	0.0636***	0.0590***	0.0607***
	(0.0097)	(0.0112)	(0.0105)	(0.0113)	(0.0106)	(0.0106)	(0.0103)
Wave, Strata Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	14,737	10,073	13,011	11,731	12,364	13,256	13,250
R^2	0.06790	0.06653	0.06637	0.07301	0.07053	0.06778	0.06943

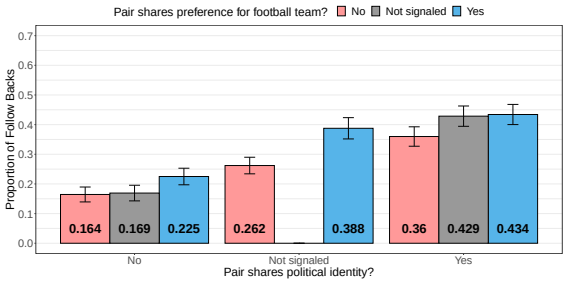
Notes: The table presents regression estimates for the effect of sharing affective identity on follow-backs (Panel A and C) and blocks (Panel B and D), considering only the accounts that signaled only affective identity (top two panels), or accounts that signaled both dimensions (bottom two panels). Specifically, it shows OLS estimates of specification 2, excluding one of the bot's clubs at a time. The sample excludes shadow-banned accounts, as pre-registered and discussed in the text. Standard errors clustered at the fictional account account level are in parentheses. Significance codes: *** : $p < 0.01$, ** : $p < 0.05$, * : $p < 0.1$.

Table B.11: Experiment Results Excluding Clubs Not Signaled by Fictional accounts

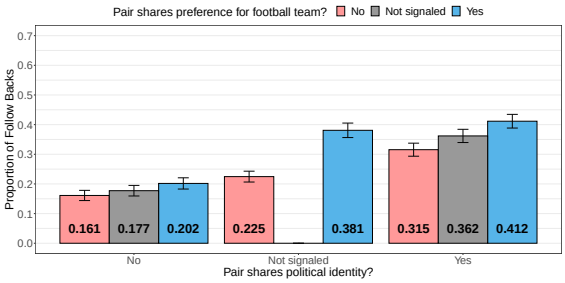
Panel A: Fictional accounts signaling affective Identity Only				
Dependent Variables:	Follow Backs (1 = Yes)		Blocks (1 = Yes)	
Sample:	Full	Excluding non-signaled Clubs	Full	Excluding non-signaled Clubs
	(1)	(2)	(3)	(4)
Affective congruence	0.1454***	0.1636***	-0.0132***	-0.0123**
	(0.0187)	(0.0204)	(0.0043)	(0.0048)
Wave, Strata Fixed Effects	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Observations	7,388	5,949	7,199	5,784
R^2	0.08339	0.08361	0.01773	0.01779
Panel B: Fictional accounts Signaling both Dimensions of Identity				
Dependent Variables:	Follow Backs (1 = Yes)		Blocks (1 = Yes)	
Sample:	Full	Excluding non-signaled Clubs	Full	Excluding non-signaled Clubs
	(1)	(2)	(3)	(4)
Political congruence	0.1476***	0.1454***	-0.1267***	-0.1209***
	(0.0139)	(0.0171)	(0.0081)	(0.0096)
Affective congruence	0.0512***	0.0455***	-0.0652***	-0.0577***
	(0.0114)	(0.0131)	(0.0093)	(0.0103)
Political congruence × Affective congruence	0.0503***	0.0529***	0.0578***	0.0515***
	(0.0170)	(0.0195)	(0.0097)	(0.0106)
Wave, Strata Fixed Effects	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
Observations	15,128	12,326	14,737	11,964
R^2	0.09909	0.09614	0.06790	0.06223

Notes: The table presents regression estimates for the effect of sharing identity on follow-backs and blocks, considering treatment arms with fictional account accounts that signaled affective identity only (Panel A) or both dimensions of identity (Panel B). Columns (2) and (4) present results for a subsample of subjects that exclude those who support a club that was not among the six clubs signaled by fictional accounts during the experiment. Standard errors clustered at the fictional account account level are in parentheses. Significance codes: *** : $p < 0.01$, ** : $p < 0.05$, * : $p < 0.1$.

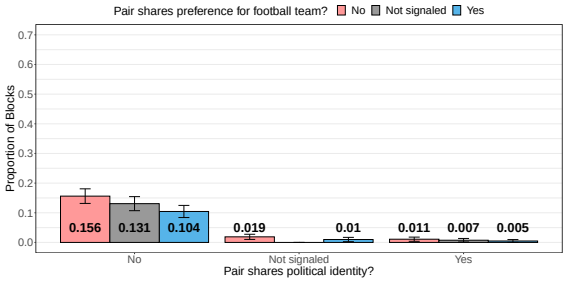
Figure B.7: Heterogeneity on type of content posted before treatment



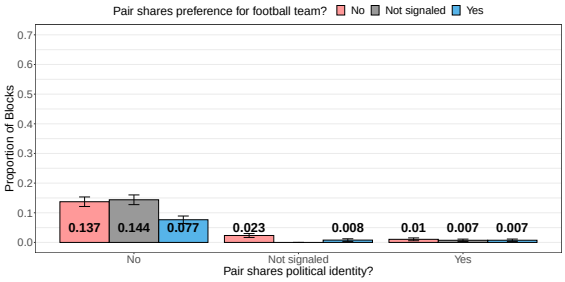
(a) Follow Backs, subjects whose last tweet before treatment had political content



(b) Follow Backs, subjects whose last tweet before treatment did not have political content



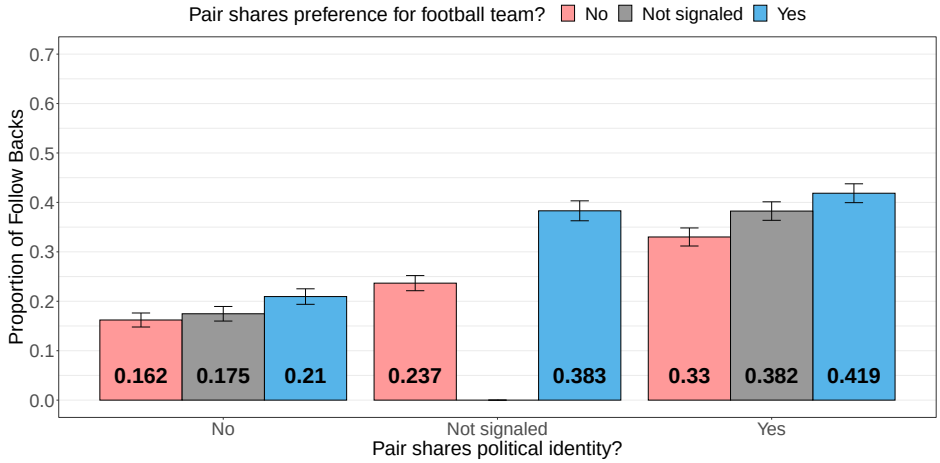
(c) Blocks, subjects whose last tweet before treatment had political content



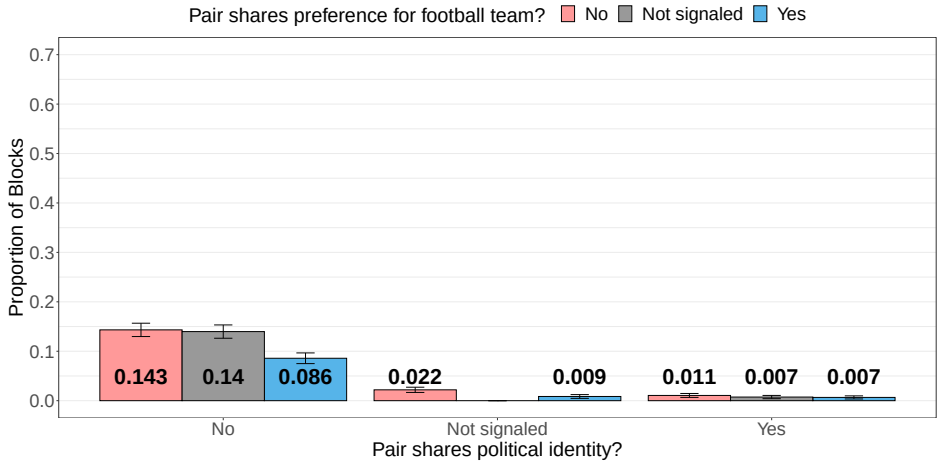
(d) Blocks, subjects whose last tweet before treatment did not have political content

Notes: The figures show the effect of sharing political and affective (football club) identity on the rate of follow-backs and blocks, for all eight treatment arms in the main experiment (fictional accounts that signal both or a single dimension of identity). The x-axis shows whether fictional account and subject share political identity (or show that this dimension is not signaled by the fictional accounts), while the colors show whether fictional account and subject share preference for football club (or show that this dimension is not signaled by the bot). Each bar shows the average follow-back rate (panels a and b) and block-rate (panels c and d) for each of these treatment arms. The figures report results for two sub samples of subjects: the ones whose last tweet before treatment had political content and the ones whose last tweet before treatment had other type of content. To classify tweets' content, we use a Naive Bayesian Classifier Algorithm. This analysis is restricted to waves 11 to 43 due to data constraints. We also restrict the analysis to subjects who tweeted at most one week before treatment. The error bars represent 95% confidence intervals.

Figure B.8: Effect of shared political and affective identity on the formation of social ties, Waves 11-43



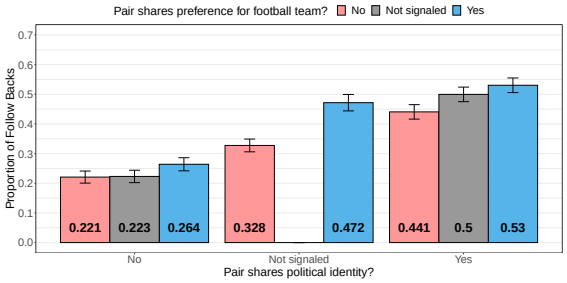
(a) Follow Backs



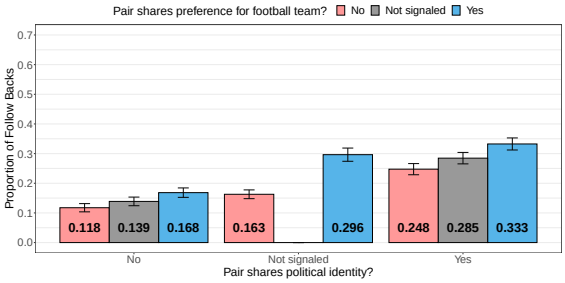
(b) Blocks

Notes: The figures show the effect of sharing political and affective (football club) identity on the rate of follow-backs and blocks, for all eight treatment arms in the main experiment (fictional accounts that signal both or a single dimension of identity). The x-axis shows whether fictional account and subject share political identity (or show that this dimension is not signaled by the fictional accounts), while the colors show whether fictional account and subject share preference for football club (or show that this dimension is not signaled by the bot). Each bar shows the average follow-back rate (panel a) and block-rate (panel b) for each of these treatment arms. This analysis is restricted to waves 11 to 43, and to subjects who tweeted at most one week before treatment, in order to allow comparisons with the heterogeneity analysis of Appendix Figure B.7. The error bars represent 95% confidence intervals.

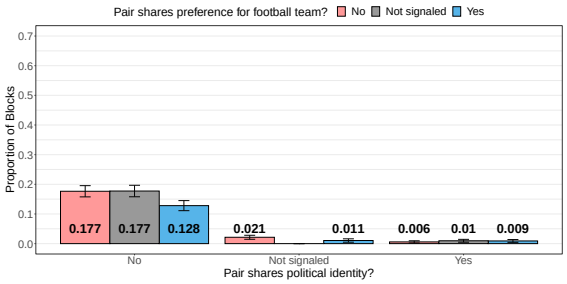
Figure B.9: Heterogeneity on type of content in user's pre-treatment bios



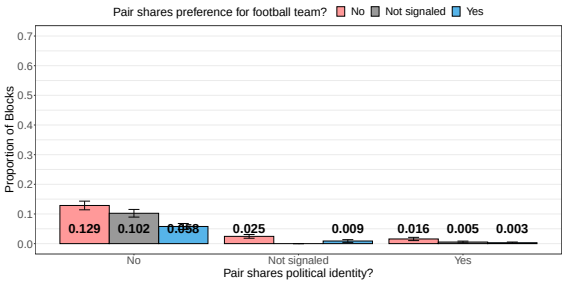
(a) Follow Backs, subjects whose bio had political content



(b) Follow Backs, subjects whose bio did not have political content



(c) Blocks, subjects whose bio had political content



(d) Blocks, subjects whose bio did not have political content

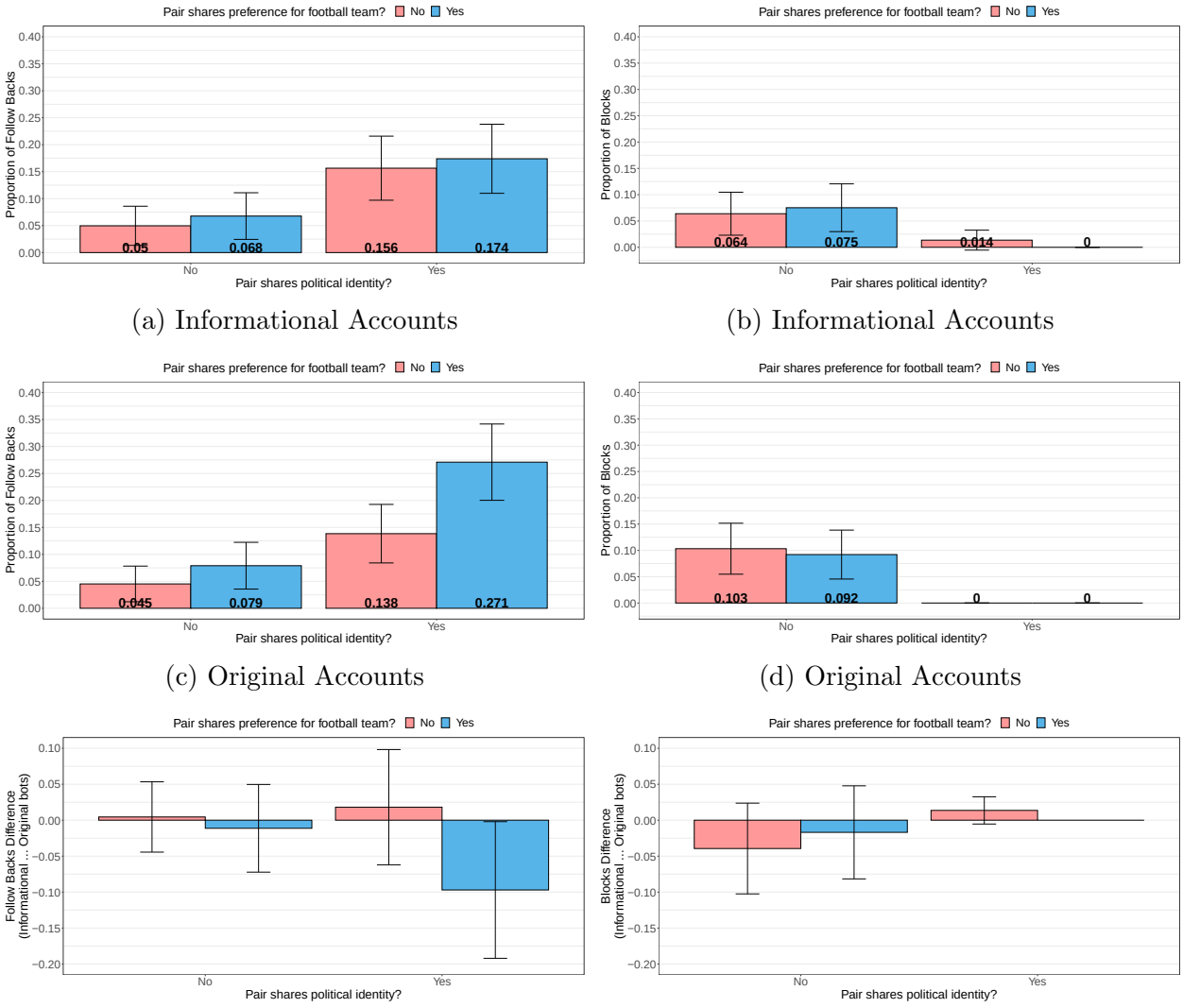
Notes: The figures show the effect of sharing political and affective (football club) identity on the rate of follow-backs and blocks, for all eight treatment arms in the main experiment (fictional accounts that signal both or a single dimension of identity). The x-axis shows whether fictional account and subject share political identity (or show that this dimension is not signaled by the fictional accounts), while the colors show whether fictional account and subject share preference for football club (or show that this dimension is not signaled by the bot). Each bar shows the average follow-back rate (panels a and b) and block-rate (panels c and d) for each of these treatment arms. The figures report results for two sub samples of subjects: the ones whose bio (before treatment) had political content and the ones whose bio (before treatment) had other type of content. To classify bios' content, we use a simple keyword search in a dictionary of words related to the Brazilian elections. The error bars represent 95% confidence intervals.

B.8 Demand for Information *versus* Social Connections

Figure B.10: Examples of Informational Fictional Accounts

Notes: The figure shows an example of informational fictional account account used in the experiment. The informational fictional account accounts explicitly signal they are automated in their profile name (with the word "Bot" in parentheses) and that they share information on their political and affective identity in their bio. The bio reads *"fictional account that re-tweets news pieces about Corinthians [its football team] and #Bolsonaro22."*

Figure B.11: Effect of shared political and affective identity on Follow Backs, information versus original accounts



(a) Informational Accounts

(b) Informational Accounts

(c) Original Accounts

(d) Original Accounts

(e) Difference (Informational − Original)

(f) Difference (Informational − Original)

Notes: The figures show the effect of sharing political and affective (football club) identity on the rate of follow-backs (left) and blocks (right), separately for the fictional accounts that explicitly say they will share information (top) and the original fictional accounts (middle), as well as the differences (bottom panel). Data comes from four experimental waves conducted between December 13th, 2023, and February 14th, 2024. The plots show 95% confidence intervals (error bars). The bottom plot displays differences between informational and original accounts.

Table B.12: Motivation to establish ties: information versus social ties

Panel A: Follow Backs

	Dependent Variable: Follow Backs (1 = Yes)			
	(1)	(2)	(3)	(4)
Informational Bot	-0.0258**	-0.0175	-0.0145	-0.0165
	(0.0097)	(0.0135)	(0.0115)	(0.0155)
Political congruence		0.1282***		0.1098***
		(0.0177)		(0.0218)
Informational Bot × Political congruence		-0.0174		0.0029
		(0.0268)		(0.0276)
Affective congruence			0.0630***	0.0260
			(0.0179)	(0.0180)
Informational Bot × Affective congruence			-0.0447**	-0.0042
			(0.0217)	(0.0268)
Political × Affective congruence				0.0731
				(0.0441)
Informational Bot × Political × Affective congruence				-0.0801
				(0.0667)
Wave, Strata Fixed Effects	Yes	Yes	Yes	Yes
Observations	2,433	2,433	2,433	2,433
R^2	0.02311	0.05665	0.02639	0.06115

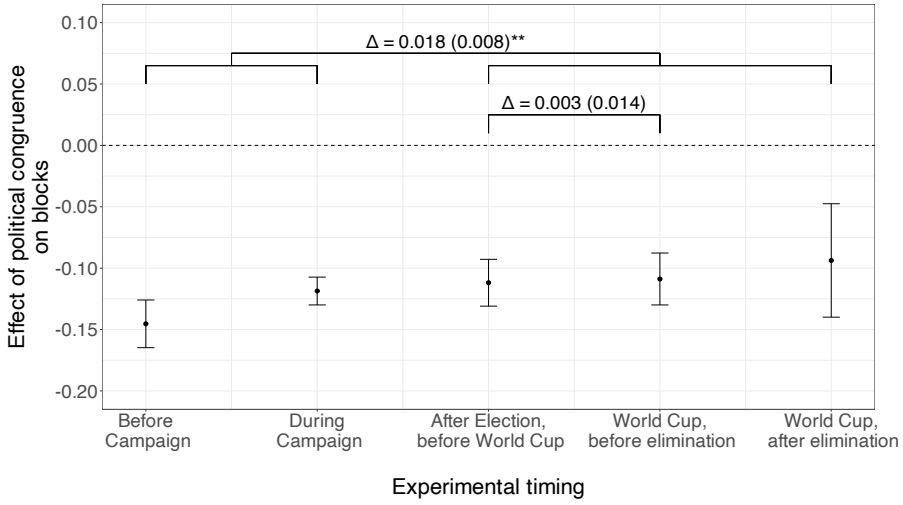
Panel B: Blocks

	Dependent Variable: Blocks (1 = Yes)			
	(1)	(2)	(3)	(4)
Informational Bot	0.0011	0.0032	0.0044	0.0099
	(0.0091)	(0.0201)	(0.0102)	(0.0223)
Political congruence		-0.0902***		-0.0890***
		(0.0164)		(0.0188)
Informational Bot × Political congruence = 1		-0.0033		-0.0099
		(0.0227)		(0.0248)
Affective congruence			-0.0008	0.0014
			(0.0168)	(0.0354)
Informational Bot × Affective congruence = 1			-0.0147	-0.0292
			(0.0229)	(0.0478)
Political × Affective congruence				-0.0045
				(0.0368)
Informational Bot × Political × Affective congruence = 1				0.0284
				(0.0498)
Wave, Strata Fixed Effects	Yes	Yes	Yes	Yes
Observations	2,433	2,433	2,433	2,433
R^2	0.00408	0.04974	0.00450	0.05044

Notes: The table presents regression estimates for the effect of explicitly informational-sharing accounts on follow-backs (Panel A) and blocks (Panel B). The sample excludes shadow-banned accounts. "Informational Bot" is an indicator equal to one for fictional accounts that explicitly state that they are automated and will share information about their preferred politician and football club. "Political identity" and "Affective identity" are indicators equal to one if fictional account and subject share political or affective identity (respectively). Data for this table comes from four experimental waves conducted between December 2023 and February 2024 (i.e., one year after the original experiment). Standard errors clustered at the fictional account account level are in parentheses. Significance codes: $^{***}: p < 0.01$, $^{**}: p < 0.05$, $^{*}: p < 0.1$.

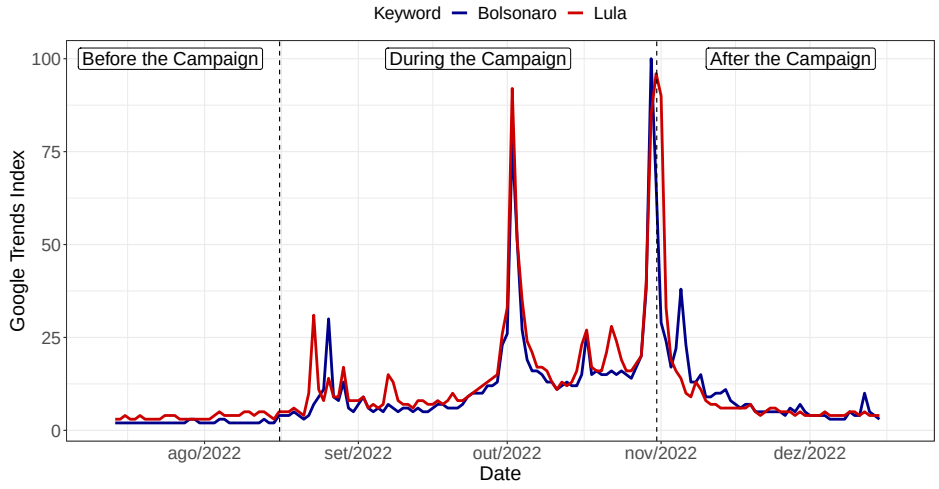
C Additional Results for the Analysis of Formation of Ties over Time

Figure C.1: Effect of Congruence in Political Identity on Blocks at Different Times



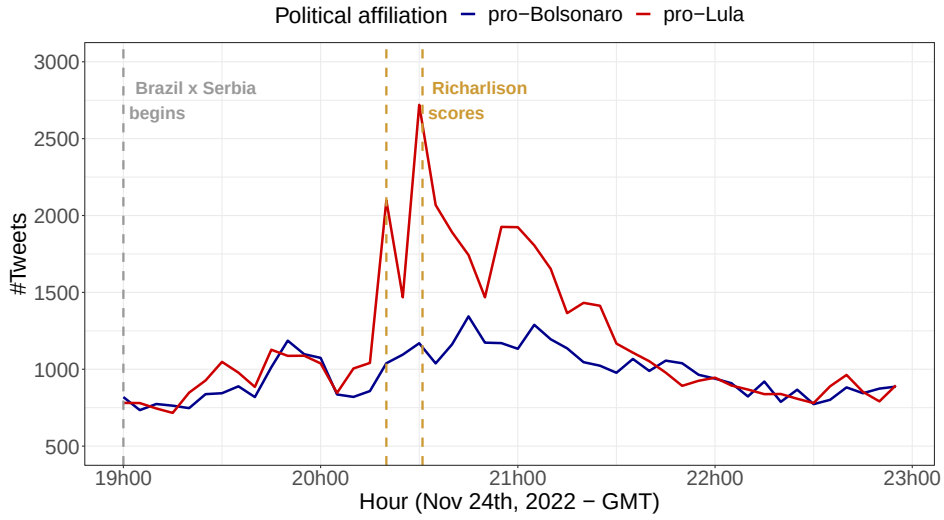
Notes: The figure displays point estimates and confidence intervals for the effect of congruence in political identity on blocks for different sets of experimental waves, ordered by period: before the official electoral period; during the electoral period; and after the electoral period. The after-election period is further divided into before the beginning of the World Cup, during the World Cup and when results for Brazil were positive, and after Brazil's elimination from the Tournament. Timing details are in Appendix Figure A.1. The sample pools data from all experimental waves within each period, restricting the analysis to subjects who were always active during the experimental period (i.e., who tweeted in the seven days before being treated every time they were treated). This gives us a total of 27,701 observations. The brackets above the point estimates display estimates and standard errors (in parentheses) for the difference in the effect of political congruence between the signaled periods. Standard errors are clustered at the bot-account level. Significance codes: $^{***}: p < 0.01$, $^{**}: p < 0.05$, $^{*}: p < 0.1$.

Figure C.2: Google Trend Index during the Experimental Period for the Two Main Presidential Candidates in the Brazilian 2022 Presidential Elections



Notes: The figure displays the Google Trends Index for searches of the terms "Lula" and "Bolsonaro" in Brazil during the experimental period. The periods denoted as "before", "during", and "after" the campaign correspond to official campaign periods as determined by Brazil's Superior Electoral Court.

Figure C.3: Number of tweets during Brazil *vs.* Serbia, 2022 FIFA World Cup



Notes: The figure displays the number of tweets sent by pro-Lula and pro-Bolsonaro users in the day of the match between Brazil and Serbia in the 2022 FIFA World Cup. Data comes from a 10% random sample of all Brazilian Twitter users that tweeted or re-tweeted a status containing a pro-Lula or pro-Bolsonaro hashtag in the week before the first round of the 2022 presidential election. Tweets are aggregated into intervals of five minutes.

Figure C.4: Word Clouds of Tweets by pro-Lula and pro-Bolsonaro users during Brazil x Serbia

(a) Tweets about Richarlison, pro-Lula users

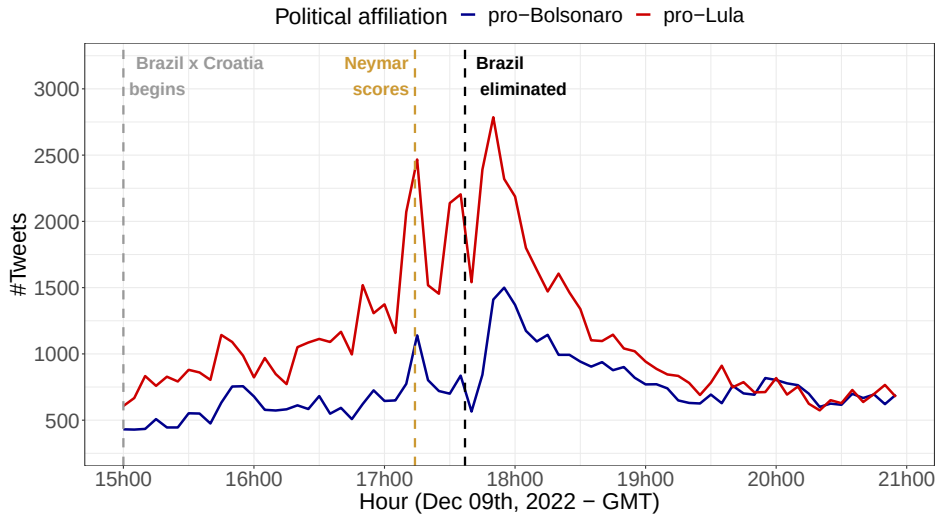
(b) Tweets about Richarlison, pro-Bolsonaro users

(c) Tweets about Neymar, pro-Lula users

(d) Tweets about Neymar, pro-Bolsonaro users

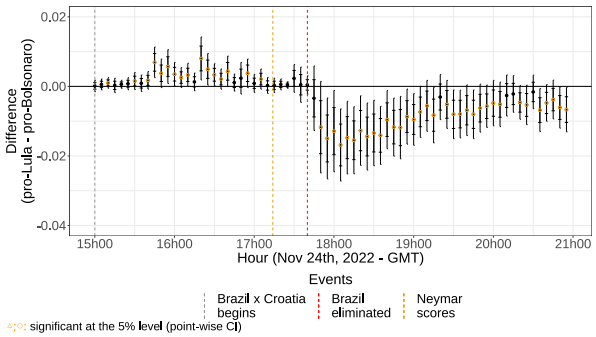
Notes: The figures show word clouds for tweets and re-tweets posted during Brazil's debut World Cup match against Serbia, from our random sample of users.

Figure C.5: Number of tweets during Brazil *vs.* Croatia, 2022 FIFA World Cup

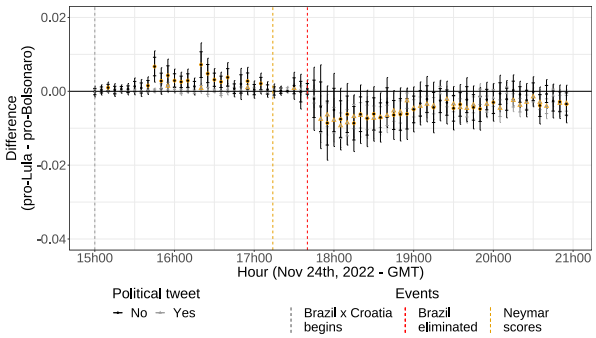


Notes: The figure displays the number of tweets sent by pro-Lula and pro-Bolsonaro users in the day of the match between Brazil and Croatia in the 2022 FIFA World Cup. Data comes from a 10% random sample of all Brazilian Twitter users that tweeted or re-tweeted a status containing a pro-Lula or pro-Bolsonaro hashtag in the week before the first round of the 2022 presidential election. Tweets are aggregated into intervals of five minutes.

Figure C.6: Difference in the number of tweets about Tite between pro-Lula and pro-Bolsonaro Twitter users during Brazil × Croatia



(a) Tweets about Tite



(b) Tweets about Tite, Heterogeneity by Political Content of Tweet

Notes: The left figure plots the difference in the likelihood that a pro-Lula and pro-Bolsonaro account posts a tweet about Brazil's coach Tite for every five minute interval around the 2022 World Cup game between Brazil and Croatia. The right figure plots a similar exercise, but separating the analysis between tweets with political content or not. In both cases, we estimate Equation (**??**) as described in the main text. To classify tweets according to their content, we use a Bayesian Classifier algorithm. In all cases, data comes from a 10% random sample of all Brazilian Twitter users that tweeted or re-tweeted a status containing a pro-Lula or pro-Bolsonaro hashtag in the week before the first round of the 2022 presidential election. The error bars with ticks represent 95% confidence intervals, while the extended bars represent 95% uniform sup-t confidence bands, estimated using Montiel Olea and Plagborg-Møller (2019)'s plug-in estimator. Standard errors are clustered at the user level. Point estimates marked in orange denote estimates significant at the 5% level (point-wise).

Figure C.7: Word Clouds of Tweets by pro-Lula and pro-Bolsonaro users after Brazil x Croatia

(a) Tweets about Tite, pro-Lula users

(b) Tweets about Tite, pro-Bolsonaro users

Notes: The figures show word clouds for tweets and re-tweets posted in the two hours after Brazil's match against Croatia (when Brazil was eliminated), from our random sample of users.