RedNIE Red Nacional de Investigadores en Economía

Rooting for the Same Team: On the Interplay between Political and Social Identities in the Formation of Social Ties

Nicolás Ajzenman (McGill University)

Bruno Ferman (São Paulo School of Economics - FGV)

Pedro C. Sant'Anna (São Paulo School of Economics - FGV)

DOCUMENTO DE TRABAJO N° 231

Abril de 2023

Citar como:
Ajzenman, Nicolás, Bruno Ferman y Pedro C. Sant'Anna (2023). Rooting for the Same Team: On the Interplay between Political and Social Identities in the Formation of Social Ties . *Documento de trabajo RedNIE N°231*.

Rooting for the Same Team: On the Interplay between Political and Social Identities in the Formation of Social Ties*

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

January 16, 2023

Abstract

We study the interplay between political and other social identities in the formation of social ties in a setting of intense affective polarization. We created fictional accounts on Twitter that signaled their political preference for one of the two leading candidates in the Brazilian 2022 Presidential election, their preference for a Brazilian football club, or both. We interpret preference for a football club as an affective dimension of identity. The bots randomly followed Twitter accounts with congruent and incongruent identities across these two dimensions, and we computed the proportion of follow-backs and blocks they received. Both dimensions of identity are relevant in forming ties, but the effect of sharing a political identity is significantly greater. Moreover, affective identity becomes substantially less relevant when information about political identity is available, indicating that political identity can overshadow other dimensions of identity. Still, shared affective identity has a positive effect in fostering ties even among politically opposite individuals. This result suggests that shared identities such as preference for a football club have the potential to reduce politically induced societal divides, despite the evidence that affective polarization may diminish this effect.

Keywords: Social Identity; Affective Polarization; Brazilian Elections; Social Media.

JEL Codes: D72; D91; C93; Z20.

*We thank Claudio Ferraz, Thomas Fujiwara, Lorenzo Lagos, Horacio Larreguy, Mario Macis, Mohsen Mosleh, Romain Wacziarg and seminar participants at PUC-Chile for their helpful comments and suggestions. We are grateful to Livia Haddad and Luis Lins for superb research assistance. This research was approved by the Ethical Compliance Committee on Research Involving Human Beings at Fundação Getulio Vargas with a waiver of informed consent (CEPH/FGV, IRB approval n. 208/2022). The experiment was pre-registered at the AEA RCT Registry under ID AEARCTR-0009982.

†McGill University, São Paulo School of Economics and IZA. E-mail: nicolas.ajzenman@mcgill.ca

‡São Paulo School of Economics - FGV. E-mail: bruno.ferman@fgv.br

§São Paulo School of Economics - FGV. E-mail: pedro.santanna@fgv.br

1 Introduction

Affective polarization, the degree to which people display animosity towards those with political preferences different from their own, has been growing recently in many countries (Iyengar et al., 2019; Boxell et al., 2022). This phenomenon is related to the fact that, in many instances, people seem to consider their political preference as a core element of their social identity (Huddy et al., 2015), leading them to evaluate positively those from the same political group while being relatively hostile towards the out-group. Such polarization can have substantial negative consequences on interpersonal relations (Huber and Malhotra, 2017; Chen and Rohla, 2018) and, more broadly, on democratic norms (Iyengar et al., 2019). In particular, affective polarization can lead people to disproportionately establish ties with those politically similar to them. Such political homophily may have powerful negative implications for the type of information consumed by individuals and the attitudes and values they form (McPherson et al., 2001), reinforcing polarization and reducing social cohesion. However, an individual's social identity is not composed exclusively of their political preference; rather, as pointed out by social identity theory (Tajfel and Turner, 1986), there are several social categories with which one can identify. Individuals with opposite political preferences could, in principle, identify with one another through other shared identities, such as belonging to the same family or nation, supporting the same football club, among others. One crucial question, therefore, is to what extent differences in political identity undermine or overshadow other dimensions of social identity, particularly in contexts of intense polarization. At the same time, it is relevant to investigate whether other shared identities have the potential to reduce polarization, fostering ties across partisan lines.

In this paper, we investigate these questions by analyzing the interplay between congruence in political identity and congruence in other dimensions of identity (that would traditionally promote affective ties between individuals) in the formation of social ties. Our setting is Brazil, a country with high affective polarization (Ortellado et al., 2022; Wagner, 2021). Specifically, we study the interplay between congruence in political identity and congruence in preference for a Brazilian football club.[1] Football is a fundamental cultural element in Brazil (DaMatta, 1994). Thus, we interpret preference for a club as a possible dimension of social identity. 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 preference or other societal divides such as income levels (Ronconi, 2022). Hence, football club preference is an ideal dimension to study how political identity interacts with other affective identities in a context of polarization.

To study this interaction, we conducted a pre-registered experiment 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

[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 context of Brazil or Latin America.

[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).

election (either Luiz Inácio Lula da Silva or Jair Messias Bolsonaro, the two 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 football club identity "affective" identity in opposition to political identity. We also created accounts that signaled only one of the two dimensions, allowing us to study the effect of shared identity without conditioning on the other dimension. The accounts then randomly followed Twitter users with congruent and non-congruent identities across these two dimensions (political and affective). Importantly, our sample is composed of politically engaged Twitter users. While this sample is not representative of the Brazilian Twitter population, it is a relevant group when studying political polarization since those users are arguably more susceptible to polarization and actively produce or share political content online (potentially influencing less engaged users). Finally, we computed the proportion of follow-backs and blocks that each bot received as measures of social ties between Twitter users and our experimental accounts.

Our measures of social ties are interesting since they capture opposite intentions of Twitter users. On the one hand, follow-backs — a standard measure in the literature (e.g., Ajzenman et al., 2022; Mosleh et al., 2021) — represent an intention to connect with the experimental account. It is, therefore, a measure of positive affect. On the other hand, the use of blocks is novel in the literature, and we interpret it as indicative of avoidance or derogation. Through a block, users forbid the blocked user from contacting them, seeing their posts, or commenting on their feed. Hence, we see a block as signaling that the user wants to be as far apart as possible from the fictional account.

Overall, we find that sharing identities — either political or football related — shapes the formation of social ties, with political identity taking a more prominent role. First, using accounts that signal a single dimension of identity (either affective or political), we show that subjects who share a football club with a bot are significantly more likely to follow it back and less likely to block it relative to when they support rival clubs. Nevertheless, sharing political identity is an even more relevant determinant of ties. Indeed, while sharing affective identity causes an increase of 13.4 percentage points in follow-backs (or a 58.5% increase), sharing political identity increases follow-backs by 20 pp (or a 119% increase). The results for blocks are even more striking: while sharing affective identity decreases the probability of blocks by 1.4 pp, sharing political identity decreases the probability of blocks by 12.3 pp.

Looking at the experimental accounts that signal both dimensions of identity, we find, consistent with the analysis of the accounts signaling a single identity, that both dimensions are relevant to the formation of ties, but political identity is relatively more important. For instance, relative to following someone who disagrees in both dimensions with the bot, following someone who shares a political identity but supports a rival football club increases the likelihood of having a follow reciprocated approximately four times more than following someone who supports the same club but has the opposite political identity. More importantly, we obtain evidence that, when information about political identity is available, the effect of sharing affective identity decreases. Indeed, while sharing affective identity increases the likelihood of follow-backs by 13.4 pp when a bot does not signal its political identity, the effect is only 8.4 pp when the bot and subject politically agree, and 4.3 pp when the bot and subject politically disagree. These results suggest that, in a context of intense affective po-

larization, political identities can undermine social connections that could have been formed due to other shared identities, overshadowing these other identities.

However, we note that congruence in the affective identity can still foster ties, even among politically incongruent individuals. Among politically opposite individuals, sharing affective identity reduces the blocking probability by 42% and increases the follow-back probability by 27%. These results suggest that, while affective polarization may reduce the effect of other identities and Twitter users sort more strongly on political preferences than on other dimensions, shared identities such as preference for the same football club can help reduce politically induced divides.

Finally, taking advantage of the different treatment timings across the experiment, we study how these behaviors change depending on how salient the election (and, consequently, the political identity) is. We hypothesize that, during the election, sorting on the political dimension would increase, and the affective identity would become less relevant to the formation of ties. We find some suggestive evidence that the effect of sharing political identity becomes smaller after the election, when this dimension is arguably less salient.

Our results have implications for the study of polarization and social identity. First, studying the formation of ties in a social media environment is particularly relevant because digital technology is frequently pointed out as part of the cause of polarization (Gentzkow, 2016), as well as an environment that amplifies polarization through echo chambers (Sunstein, 2001, 2018). Hence, it is relevant to understand to what degree political congruence (or lack thereof) causes the formation of social media ties even when people can identify with each other via other shared identities. 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. In particular, we document that political identities can undermine the formation of ties due to other shared identities in a context of intense affective polarization. Thus, we contribute to the study of the consequences of affective polarization on interpersonal relations (Iyengar et al., 2019). The sorting on political preferences that we document can 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.

However, on a more positive interpretation, it is notable that even politically engaged individuals are willing to establish social media ties with those with opposite political identities when the two sides share another identity (in our case, football club preference). The effect of shared affective identity is small relative to the effect of shared political identity but is still sizeable. This result is consistent with the literature on determinants of social cohesion, which points to the importance of sports (football in particular) in fostering integration (Depetris-Chauvin et al., 2020). Thus, our result suggests that highlighting similarities in other dimensions instead of political differences may help create ties across partisan lines.

Our results also contribute to several strands of the literature. First, our paper relates to the literature on affective polarization (Iyengar et al., 2012, 2019). Most of the literature on this topic uses survey evidence to measure polarization (e.g. Iyengar et al., 2012; Boxell et al., 2022; Wagner, 2021; Reiljan, 2020). We contribute by studying affective polarization

in a natural setting, providing a behavioral measure of this phenomenon. Some papers have started documenting affective polarization in behavioral measures in different contexts for the United States, such as behavior in online dating platforms (Huber and Malhotra, 2017), family gatherings during holidays (Chen and Rohla, 2018), and connections on social media (Mosleh et al., 2021). Among these papers, we are more closely connected with Mosleh et al. (2021). They use a similar methodology to ours to study the effect of shared partisanship in the US on the formation of ties on Twitter. Compared to this paper, we contribute by studying a different institutional setting, showing that affective polarization in the formation of social media ties is not exclusive to the US. More importantly, we contribute by studying how political identity interacts with another dimension of identity. This second contribution is particularly relevant as it advances our understanding of how and to what extent political identity overshadows other dimensions of shared identity. A phenomenon of this nature has been documented in the US by 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). 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.

More generally, we connect to the literature on social media and politics, particularly the strand studying the welfare effects of social media (Zhuravskaya et al., 2020). The literature on this topic documents that social media has effects, among others, on protest participation (Enikolopov et al., 2020; Acemoglu et al., 2018), expression of xenophobic views (Bursztyn et al., 2019), political polarization (Allcott et al., 2020; Di Tella et al., 2021), and news consumption (Levy, 2021). By documenting homophily in the political dimension in the formation of social media ties, we contribute to this literature by providing evidence of a potential mechanism that could amplify the political effects of social media. At the same time, our results suggest a mechanism for fostering ties across partisan lines, possibly reducing polarization.

Second, we relate to the literature on the determinants of social cohesion, particularly the strand studying how social ties among groups who perceive each other as different can be formed when those groups have some other shared identity or experiences.[3] 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 studying how football can affect social ties. First, Depetris-Chauvin et al. (2020) show that individuals in Sub-Saharan Africa are more likely to identify with their nation than with 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,

[3]We also relate to the literature on the relationship between social media and social cohesion (González-Bailón and Lelkes, 2022).

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 first contribution is to show that shared football club preference has a causal effect on social ties, even among politically opposite individuals. However, we also show that, while congruence in football club preferences can increase the likelihood of establishing a social tie, the effect is small relative to the divisiveness created by opposite political preferences, and affective polarization may diminish this effect even further. Hence, our results suggest that football club identification may have limited power to increase social cohesion (when it comes to the formation of ties), particularly in a polarized setting.

Third, we are related to the literature on social identity, which started in psychology with the works of Tajfel and Turner (e.g. Tajfel and Turner, 1986) and was introduced into economic analysis by Akerlof and Kranton (2000) and Shayo (2009, 2020). More recently, some papers have focused on studying the implications of social identity to a variety of relevant economic outcomes, such as trade policy (Grossman and Helpman, 2021), teamwork (Charness and Chen, 2020), and acceptance of bonus payments (Bursztyn et al., 2020). In this context, we contribute by studying the interplay between two social categories, providing evidence on how different dimensions of identity interact in a natural setting and their effects on establishing social ties. Moreover, one relevant open question in this literature is how identity choice changes over time — for instance, due to changes in the salience of a given social category. By running the experiment during different moments of a presidential election, we find some evidence that subjects' behavior changed depending on the salience of the election. These changes can be explained with the help of social identity theory.

Finally, we are related to the literature on homophily, the tendency of people to establish contact with those perceived as similar to themselves (McPherson et al., 2001; Currarini et al., 2009). There is considerable evidence that politically engaged individuals tend to have more social media connections with like-minded individuals (Halberstam and Knight, 2016). However, it is less clear whether this is caused by shared political identity or if sharing political identity is correlated with other characteristics that cause sorting (an exception is Mosleh et al., 2021). We first contribute by providing causal evidence that shared political identity causes homophily in social networks. Second and more importantly, by studying the relative importance of shared political identity compared to sharing other dimensions of identity. Moreover, we contribute by showing that homophily in a social media setting happens not only by social media users choosing to be closer to those politically similar to them (through follow-backs) but also by users avoiding contact with those who are different (through blocks).[4]

The remainder of this paper is organized as follows. In Section 2, we provide relevant background on polarization in Brazil, football, and Twitter; 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; finally, Section 5 presents our results.

[4]In that sense, we also contribute to the literature on experiments on social media. For comprehensive reviews, see Mosleh et al. (2022) and Guess (2021).

2 Background

2.1 Political Polarization in Brazil

For many analysts, Brazil's relatively young democracy is currently "caught up in the sharpest and most polarizing moment in its history" (Kingstone and Power, 2017). This trend started at least in 2013, when millions of Brazilians went to the streets to protest against the political establishment. Since then, the country experienced an impeachment process against left-wing President Dilma Rousseff (of the Worker's Party), who was accused of breaking budget laws, in 2016, and saw the election of far-right candidate Jair Bolsonaro as president in 2018. In the 2018 elections run-off, as Mignozzetti and Spektor (2019) argue, the country was presented with two choices — Jair Bolsonaro and Fernando Haddad, from the Worker's Party — that represented opposite ends of the political spectrum, in sharp contrast with previous elections (that always featured at least one representative of traditional centrist parties). In 2022, Brazilian citizens faced a similar decision, this time having to choose between Jair Bolsonaro 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 than in previous elections.[5] These numbers are perhaps the clearest possible evidence that the country's population was divided between those who support Bolsonaro and those who support Lula.

Given the elevated level of political turmoil in recent years, it would not be surprising if polarization among the general Brazilian population had increased. Interestingly, while there is evidence that Brazil currently has low levels of ideological polarization both compared to other countries and to its history (Ortellado et al., 2022; Mignozzetti and Spektor, 2019), affective polarization has indeed increased recently in the country. This is similar to countries such as the United States, where ideological polarization (i.e., polarization in terms of issues or opinions) evolves much slower than affective polarization – possibly because affection is not always anchored on policy preferences (Iyengar et al., 2012).

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 undertaken by the Center of Studies on Public Opinion (CESOP) since 2002. This survey is part of the Comparative Study of Electoral Systems project. We consider answers to the question "how much do you like each party?"as measures of the respondent's affect towards each party.[6] The results from our analysis are displayed in Figure 1. First, we plot the evolution of answers to this question over the five BES modules for the most preferred party (in-party) and the remaining parties. This analysis mirrors that of Iyengar et al. (2019)

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

[6] More precisely, the question asked 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."

for the US. In this case, *affective polarization* is the mean difference between in-party and out-party scores. Using this measure, we see that affective polarization in Brazil has been at a high level since 2002 (for comparison, Iyengar et al. (2012), using the same "feeling thermometer" method, finds that affective polarization in the US was at 40.87 degrees in 2016, while in Brazil we estimate it to be at 50.3 degrees). Moreover, there is evidence of a weak growing trend in polarization, mostly driven by an increase in dislike towards out-parties. Indeed, while in 2002 the average respondent gave a score of 32.7 for parties different from their most preferred party, in 2018, this score was 23.9.

We also report another measure of affective polarization using the BES data. 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. 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. Using this method, we find that the mean level of affective polarization in Brazil in 2018 (58.8) is comparable to 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 great political turmoil and — consistently — an increase in affective polarization (even though there is less evidence of an increase in partisan or ideological 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). Our experiment contributes by studying polarization in a real-world setting, providing behavioral measures of affective polarization obtained in a natural environment.

2.2 Football

Football is by far the most popular sport in Brazil. 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 to a football team (...), recreating in a modern level the idea of family as a community (...) that is chosen voluntarily" (see also

DaMatta, 1982). 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).

Importantly, 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.1 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. Although there is some variation across some clubs, supporters are generally similar. 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 characteristic of the majority of its supporters. This feature is interesting as it suggests that socialization through football club preferences in Brazil has the potential of creating ties among individuals who would not generally share other identities. Finally, we also show that, 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 analyse, at least 38% of the supporters in our sample prefers the candidate preferred by the minority of that club's 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, which we will explore in this paper. Specifically, we ask if sharing a political club can generate social connections even when individuals have opposing political views.

2.3 Twitter

The setting of this experiment is Twitter, one of Brazil's most popular social media platforms. In 2021, over 17 million Brazilians used Twitter (Statista, 2022b), making it one of the country's most-used social media platforms. Twitter is a microblogging 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, most users have public profiles, which implies that their posts are publicly visible. Although it is also possible to have *protected* — i.e., private — accounts, the default configuration is for an account to be public. 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 receives a *follow notification* on their account, informing them that a new account has followed their profile.[7] 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 nor 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 they were blocked.

Hence, 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. Indeed, a block is an active measure taken by an account, preventing 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). This social media has been increasingly used both by candidates and the general public 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 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 —

[7] Twitter sends this follow notification in most cases, but this does not always happen. In some cases, Twitter may consider that an account is acting suspiciously and *shadow-ban* it by making it invisible to other users. In this case, a followed user would not receive a follow notification.

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 who identify 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, as previously discussed, is often 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 bot and subject share identities, we will focus on whether bot 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 analyse. 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 dislike and distrust 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 analyse — 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 belongs to, someone may identify with one or a subset of these categories in different points in time. 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 in 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. We study this possibility in section 5.4.

4 Design and Data

4.1 Experimental Design

We conducted our pre-registered experiment on Twitter between July and December 2022. We created fictional accounts (called 'bot' accounts) on Twitter that signaled their preferred candidate in the 2022 Brazilian election and/or their preferred Brazilian football club. The bot accounts randomly followed Twitter users who shared or not each identity (political and football club preference) with it. After five days active, 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 bots: (1) bots that signal both dimensions of identity (political and affective); (2) bots that only signal political identity; (3) bots 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 bots: 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 bots 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. Moreover, by comparing bots that signal a single or both dimensions of identity, we are able to access the relative importance of each identity signal. We now give additional details on each aspect of the experimental design.

4.1.1 Bot Accounts

Table 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).[8] The political and affective identity of each bot are chosen randomly, using a procedure that we describe in the following subsection.

Given an assigned political and affective identity, we can describe how these are signaled in the bot's profile. We signal political identity by including, in the bot's bio, either the hashtag *#Lula2022* or *#Bolsonaro2022*, and by re-tweeting one post from the candidate supported by the bot.[9] If the bot is politically neutral, we simply do not include either hashtag and do not retweet a political post. On the other hand, we signal affective identity through the bot's profile picture (which is a picture of a flag with the bot's preferred team logo in a stadium) and by adding the text "Supporter of team X" in the bot's bio. For bots that are neutral in the football-related dimension, we use a photo of a football stadium outside Brazil (and 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. Therefore, the accounts that are football team-neutral are still signalling that they are interested in football (the only difference is that they do not signal preference for a specific team). Figure 2 shows examples of bot accounts.

Therefore, for the 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 bot's bio). Nevertheless, we also created, in three experimental waves, extra accounts that signaled their political preference more saliently. Examples of such accounts can be seen on Appendix Figure B.3. In the case of these accounts, the profile picture is the official campaign photo of their preferred candidate. Thus, affective identity is only signaled on the account's bio. We created these accounts with the objective of testing how subjects' behavior changes in a more extreme case, where political preference is more salient.

[8]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 bots 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).

[9]To alleviate concerns that the bots 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.

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. 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 displayed in Appendix Table A.1. Hence, our sample is composed of 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) signalled the user's preferred Brazilian football club. To do this, we first use a simple algorithm that detects terms associated with the 6 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 bots, 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 real people. 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 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. Specifically, we have information on the number of tweets, followers, and friends. We also have information on location for the accounts that choose to let this information public, which we recode to a regional level. Moreover, we know whether the account is verified, the number of likes ("favorites") it performed, and its date of creation. From users' names, 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, we see that our sample is balanced across 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 25% 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 each experimental wave, we activated eight bot accounts: four accounts that are part of the main experiment, and 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). The objective of the four neutral accounts is to study if each dimension of identity (football preference and political identity) are relevant to the formation of social ties when they are the only signaled identity; and to compare users behavior when information on bot's political preferences is or is not available. In each wave,

we randomly choose two football clubs for the bot accounts.[10] Then, within a wave, three bots signaled preference for each of these two teams. For example, in the pilot wave, we had three bots who supported Flamengo (one pro-Lula, one pro-Bolsonaro and one politically neutral), and three bots who supported Palmeiras (again, one pro-Lula, one pro-Bolsonaro and one politically neutral), plus one pro-Bolsonaro and one pro-Lula bot that did not signal a preferred football club.

Each bot 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 bot is done by stratifying the subjects in terms of their political identity, preferred football team and whether the subjects' number of followers is above or below the median. First, for the bots who signal their preferred football club, we restrict the sample of subjects to the ones who either support the same team as the bot, or who support a rival team. We only consider regional (intra-state) rivalries; the list of rivalries is described in Appendix Table A.2. Given that we are interested in studying the effect of matching bot and subjects' identities on follow-backs and blocks, we have four strata in terms of bot-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 bot) more than once, but never in subsequent waves: after being treated in a wave, a subject only returns to the subject pool after 3 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 subject and bot agree or disagree in their political and/or affective dimension of identity. Figure 3 illustrates such notification. The way a user sees the notification depends on whether he or she is using Twitter from the mobile application or a desktop computer.[11] In both cases, the user immediately sees the bot'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 bot's description and, therefore, its political affiliation.

Apart from following the experimentally assigned accounts, each bot account also followed one account from someone who is aware about 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.[12] If an account is shadow-banned, we simply

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

[11]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.

[12]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

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 run in waves. In each wave, 8 bots were activated: 4 signaling both their political identity and football club preference, and 4 neutral in one of the two dimensions of social identity. Within each wave, we used the following timeline:

(i) **Day 0:** Creation of accounts according to the procedures described in Table 1. The account re-tweets a post related to its sportive identity (either a post from its preferred club official account — if the bot 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 political post must have at least 500 re-tweets and not contain misleading information or hate speech. The bot 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 bot account follows the subjects assigned to it according to the procedure described in the previous section.

(iii) **Day 5:** After five days active, we compute the number of followers and blocks for each account and delete all information in the account, rebooting it to be used in the next wave.

We started one wave every Tuesday and every Friday, which means that we had two overlapping waves at each moment. The specific timeline is displayed in Appendix Figure A.1. We ran 43 experimental waves between July and December 2022. This period is particularly interesting because the Brazilian presidential election of 2022 was held during the second semester of 2022 (specifically, the first round happened 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 bot followed the subjects. On the other hand, we only compute blocks at the end of each wave (i.e., on the fifth day). This happens 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

that all activity from a shadow-banned user is "hidden" to other users, including notifications of follows. Therefore, we guarantee that no bot account is shadow-banned before using the results from any experimental wave.

Twitter's API to obtain, for each bot account, the set of accounts followed by it. We then compare this set with the set of accounts assigned to be followed by the bot. The difference between the two sets can be due to three mutually exclusive reasons: (i) the bot was indeed blocked by a subject; (ii) the subject was suspended or deactivated their account; (iii) the subject removed the bot 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 bot'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 bot if we see, on the fifth day, a block using this procedure.[13]

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 pooling all experimental waves, comparing the follow-back and block rates of subjects who shared or not political and/or affective identity with the bot.

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

$$Y_{ijst} = \alpha + \beta_1 \times \text{identity\_congruence}_{ij} + X_{ijt}\lambda + \delta_t + \theta_s + \phi_{st} + \varepsilon_{ijst} \quad (1)$$

were $\text{identity\_congruence}_{ij}$ is an indicator equal to one if bot and subject share identity (in the dimension we are studying), δ_t, θ_s and ϕ_{st} represent, respectively, wave, strata and strata $\times$ wave fixed effects[14] and ε_{ijst} is the error term. X_{ijt} is a vector of control variables from the bot, subjects and waves (interacted with the treatment dummies). Specifically, we include in this vector 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 bot's j football club at wave t (which is included, interacted with the identity congruence indicator, in the case of bots 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.

[13] A fourth possibility is that a subject blocked a bot, 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 blocked a subject and kept it this way until the end of the wave.

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

Apart from studying the treatment arms with accounts that signal a single dimension of identity, we also study the accounts that signal both dimensions. In this case, there are four possible pairs of subjects and bots (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:

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

where political_congruence$_{ij}$ is an indicator equal to one if bot j and subject i share political preferences, affective_congruence$_{ij}$ equals one if bot j and subject i share preference for football club, and the other variables have the same definition as before. Importantly, we control by the bot's football club salience by including, in the vector of controls, the google trends index of the bot during each wave (interacted with the treatment indicators).

Since our outcome is an indicator variable, Equation (2) represents a linear probability model. Coefficient β_1 can be interpreted as the effect (in percentage points) in 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 bot. Finally, β_3 can be interpreted as the difference in the effect of sharing affective identity between subjects who share or not political identity with the bot.

We present standard errors clustered at the bot-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 rate of rejection 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 Tables B.3 and B.4 present summary statistics of treated subjects across all treatments arms in which bots signalled both dimensions of identity and in the treatment arms experiments in which bots signalled a single identity dimension, respectively. In all cases, pre-treatment subject characteristics are balanced across treatments. For the experiment with accounts signaling both dimensions, there are four treatment arms (the combinations of in and out-group for political and affective identity), so we test balance by performing an F-test of equality across the four means. In the other experiments with accounts signaling a single dimension, there are only two treatment arms (in and out-group), and we therefore report results of t-tests of difference in means across the two arms. For all pre-treatment variables, we cannot reject the null hypothesis of equality in the main experiment for standard significance levels. Similarly, we do not reject the null of equality for most variables of the two auxiliary experiments.

Moreover, we also cannot reject the null hypothesis of equality if we jointly tested the

equality of means across all eight treatment arms (the four signaling both identities, plus the four with a single identity signal). This observation allows us to consistently perform analyses both within similar types of bots (i.e., those that signal the same categories), or across types of bots (for instance, comparing results for bots that signal both dimensions with results for bots signaling only political identity).

Both tables also show attrition rates for each treatment arm. In this experiment, 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 subjected deactivated their account; or the subject made its account private. In the first two cases, we would be unable 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, attrition rate was close to 9%, and characteristics of attrited subjects are not different across treatment arms, as can be seen on Table B.5 in the Appendix. Therefore, when analysing "statical" results, by pooling results of different waves and estimating Equation (1) or (2), attrition will not be a concern. Indeed, 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.

However, attrition could be a problem in our analysis of heterogeneous effects across different moments in time. This happens 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 be different. This is indeed the case, as shown in Table B.6 in the Appendix. This table compares characteristics of subjects that never suffered attrition in the experiment with those that 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, particularly for the analysis of heterogeneous effects, we present results both for the full sample of subjects and restricting the analysis to the sample of never attrited subjects. This second exercise allows us to compare the behavior of similar subjects across the entire experimental period.

5 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 for our sample of politically engaged individuals. To do that, we restrict our analysis to the experimental accounts that signal a single dimension of identity.

We first show that football clubs are indeed a relevant dimension of socialization in our setting. Results for the experiment using the politically neutral accounts (i.e., the accounts that only signaled preferred football club) are displayed in Figure 4. Our figures follow a similar pattern: each figure displays results for follow backs (top panel) and blocks (lower panel). For each of these two outcomes, we plot on the left-hand side the average rate of the outcome across the entire experimental period for pairs of subjects and bot that are congruent or incongruent in the dimension of identity analysed. For instance, the top-left plot of Figure 4 shows the average follow-back rate from subjects who shared or not a preferred football club with the bot. Finally, the right-hand side plot shows coefficient estimates and 95% confidence intervals for an indicator of congruence between bot and subject in the relevant identity dimension, using our pre-registered specification with strata and wave fixed-effects, and with or without controls (the list of controls used is in Section 4.2).

Figure 4 shows that individuals in our sample are more likely to establish ties with experimental accounts that support the same football club as them rather than their rivals. Indeed, a subject has a 36.3% chance of reciprocating a follow from a bot supporting their team, against a 22.9% chance of reciprocating a follow from a supporter of a rival team. The difference, of 13.4 percentage points, is highly significant and meaningful (a subject is more than 50% more likely to follow a bot from the same group than from the out-group in this affective dimension). The results for blocks tell a similar story, even tough the block rate of politically-neutral accounts is low. Subjects block 2.3% of rival accounts, against only 0.9% of accounts with a shared preference. This difference is again highly significant, but quantitatively small since the baseline rate of blocks is very low in this case.

Therefore, football club preferences are a relevant determinant of the formation of social ties in our setting. Using our pre-registered specification, we estimate that shared identity in this dimensions causes an increase in the probability of follow backs of approximately 14.1 pp, and a decrease in the probability of blocks of 1.3 pp, both significant at the 1% level. 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).

Nevertheless, while football seems to be a relevant determinant of socialization, we find that political identity plays a greater role in the formation of social ties, particularly when it comes to avoiding contact with the out-group. Figure 5 shows the effect of shared political identity on the formation of ties, considering the accounts that only signal political preference. Recall that, even tough these experimental accounts are neutral in terms of 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%). Therefore, the effect of shared political identity was greater than the effect of shared affective identity, even tough the bots signaled more saliently their preferred football club (in their profile picture and bio) than their political preference (only in their bio).

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. However, analyzing results for the experimental accounts that signal both dimensions of identity allows us to study their interplay on the formation of social ties.

Results for this analysis are displayed in Figure 6. The top panel of this figure shows the average follow-back rate, while the bottom panel shows the average block rate, for all eight treatment arms (the four arms with bots signaling both identities and the four arms with bots signaling a single dimension). These treatment arms are all represented at the same time in the two plots. In the *x-axis*, we represent whether bot and subjects share political identity: the left-most three columns show cases in which bot and subject disagree politically, while the right-most show cases in which 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 Figure 4). Finally, the *bar colors* indicate the relationship between subject and bot's football club preference. As in the case of political identity, there are three possibilities: either bot and subject share football club preference, support rival clubs, or the bot does not signal its preferred club (in that case, it only signals political preferences and the results, in grey, are the same as in Figure 5).[15]

The figures reveal that sharing either dimension of identity significantly increases the probability of follow-backs and decreases the probability of blocks. However, the magnitude of these effects are different once we condition on the other dimension of identity. We discuss these differences, and potential interpretations, in what follows, dividing the analysis in four main findings: first, while congruence in both identities have a positive effect on ties, the effective of political identity is greater; second, conditioning on football club preference does not change much the effect of political identity, which remains large; in contrast, by conditioning on information about bot's political identity, the effect of affective identity becomes smaller; yet, there is still a significant effect of sharing football club preference, suggesting that, even in a polarized setting, sharing identities such as preference for a football club may help reduce political divides.

5.2.1 The exchange rate between shared political and affective identities

Overall, consistent with our findings in the previous section, we find that both congruence in political identity and in affective identity have a positive effect on follow-backs and a negative effect on blocks. When it comes to follow-backs, the subjects who are least likely to reciprocate a bot's follow are those who do not 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 bot are

[15] Apart from the graphical results shown in Figure 6, Appendix Tables B.7 and B.8 report tests of difference in means, with standard errors clustered at the bot account level, for follow-backs and blocks (respectively) for each pair of treatment arms. For simplicity of exposure, we focus on discussing the results as shown in Figure 6, and report hypothesis test results in the text when relevant.

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 effect of sharing political identity is substantially larger than the effect of sharing affective identity.

An interesting way of visualizing this effect is by considering the exchange rate between shared political and affective identities implied by the experiment's results. Consider someone who follows a subject with whom they disagree in both dimensions of identities (and, therefore, has a 16% chance of being followed-back by it). In this case, by sharing political identity (but not affective), the follow-back probability would increase to 32.4% (an increase of 16.4 pp relative to the case of disagreement in both dimensions). However, by sharing affective identity only (but not political) this probability would increase to 20.3% (an increase of 4.3 pp relative to the case of disagreement in both dimensions). Therefore, the increase in the probability of having a follow reciprocated, relative to the case of following someone who disagrees with a bot in both identity dimensions, is roughly four times larger if the bot follows someone who shares political identity but supports a rival club than if it follows someone who supports the same club, but has an opposite political preference. Hence, political identity is significantly more relevant to the formation of social ties than football club in our setting, despite the political identity being signaled less saliently than the affective identity.

5.2.2 Effect of political congruence conditional on congruence in affective identity

We now analyze the effects of congruence in political identity conditional on congruence (or lack there-off) in affective identity. In the previous section, we found that, when information about football club preferences was not available, subjects were 20 pp more likely to follow-back bots with whom they shared political preferences (compared to bots who preferred the opposite candidate). When both dimensions of identity are signaled by the bots, the effect of congruence in political identity remains quantitatively similar in the case of follow-backs.

First, among subject-bot pairs who support the same football club (green bars in Figure 6), the likelihood of follow-backs is of 20.3% when bot 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). Similarly, among bot-subject pairs who support rival clubs (red bars), the follow-back probability is of 16% for subjects and bots who disagree politically, and of 32.4% for subjects and bot who agree in this dimension. The difference, of 16.4 pp (also significant at the 1% level), is relatively smaller than the difference in the case in which football club is not informed by the bot, but is still substantial. In all cases, sharing political identity roughly doubles the probability of follow-back.

Similarly, the effect of political identity on blocks is substantial independently of information about bot's preferred club. When this information is unavailable, the probability of blocking a bot is 12.3 pp smaller when bot and subject share political identity compared to when they have opposite identities in this dimension. When bot and subjects support rival football clubs, this difference is of 13.5 pp (from 1.1% chance of blocking when the two polit-

ically agree to 14.6% when they disagree). Finally, when bot and subject support the same club, the difference is relatively smaller, but still large: 7.9 pp (from 0.6% chance of blocking when the two politically agree to 8.5% when they disagree). While it is true that when bot and subject support the same club there is a significant reduction in the blocking probability — which we will further discuss later — it is noticeable that the difference is sizeable even in this case. Indeed, as Figure 6b shows, blocking happens almost exclusively against politically opposite accounts. Therefore, we find that individuals who disagree politically have a tendency to avoid each other (through blocks).

Overall, the effect of sharing political identity is large regardless of whether bot and subject support the same or rival clubs, or if there is no information on 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.

5.2.3 Effect of affective congruence conditional on political identity

The same cannot be said of the effect of sharing affective identity conditional on information on bot's political identity. When bots do not signal their political identity, we saw that the effect of sharing a football club was to increase the probability of follow-backs by 13.4 pp. This effect is smaller both when bot and subject agree or disagree politically. First, conditional on congruence in the political identity (right-most bars), the probability of follow-back when bot and subject support rival clubs is of 32.4%, against 40.8% when they support the same one, a difference of 8.4 pp (significant at the 1% level). This effect of sharing affective identity conditional on agreeing politically is significantly smaller than the effect of sharing affective identity when no information about political preferences is given by the bots. This difference is also quantitatively meaningful, as it implies a reduction of almost 40% in the effect of affective congruence.

The reduction in the effect of affective identity is even more striking when we consider bot-subject 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 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 bots did not signal their political identity.

Hence, during the period we analysed, political identity overshadowed other dimensions of identity (namely, football club preference) in the formation of social ties. Indeed, information on bot's political identity offsets the effect of shared affective identity, particularly among politically-opposite individuals, undermining social ties that could be formed if the bot did not signal their political identity. This is evidence that, at least in contexts of high polarization — such as the one we analyse — 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, leading to a less integrated society.

5.2.4 Congruence in affective identity and the formation of social ties

Nevertheless, it is important to highlight that, even tough political divergence makes the effects of shared affective identity substantially smaller, this dimension of identity is still able to create ties among politically opposite individuals. Indeed, among politically-opposite individuals, sharing football club preference increases the probability of follow-backs by 4.3 percentage points relative to subject-bot pairs supporting rival clubs (the effect is significant at the 1% level). Perhaps more surprisingly, shared football identity also increases the probability of follow-back among politically-opposite individuals relative to the case in which bots do not signal their preferred club. In this case, the effect is somewhat smaller (of 3.5 pp), but is still significant at the 1% level. Therefore, even among politically-opposite individuals, shared football club fostered ties.

For blocks, the effect is even larger. Among politically-opposite individuals, sharing affective identity reduces the blocking probability by 6.1 pp relative to the case of rival clubs. This represents a substantial reduction of 42% in the likelihood of blocking. Sharing a club also reduces the blocking probability among politically-opposite individuals relative to the case in which football club information is not given, by 4.5 pp. Both differences are significant at the 1% level. Hence, sharing affective identity significantly reduces the probability of blocks — even if this probability remains relatively large.

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-opposite individuals. This result is consistent with the evidence that football can foster social cohesion (Depetris-Chauvin et al., 2020; Ronconi, 2022). Hence, our findings indicate that highlighting a shared common interest — in this case, preference for football club — can help reduce politically-induced societal divides.

Overall, our results suggest that both dimensions of identity are relevant to the formation of social ties, but that political identity has a significantly larger role than the affective dimension, at least in our sample of subjects. Comparing the results for the accounts signaling both or a single dimension, we find that political identity overshadows affective identity, reducing the importance of sharing a preference for football club on the decision to follow-back an account or not. Nevertheless, even when political identity is signaled, congruence in affective identity is capable of generating social ties. This is surprising, particularly considering that our sample of subject 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 is suggestive of the fact that similarities in dimensions relatively uncorrelated with politics can reduce political divides.

5.3 Bots with more salient political identity

The results discussed so far focused on accounts that signaled political identity only in their bio. One caveat in the interpretation of these results is that the bot accounts we constructed

signal more saliently their affective identity (the football club's symbol is on their profile picture) than their political identity. What would results look like if the salience of those signals in the bot accounts were inverted?

To answer this question, we created, during a few experimental waves, another treatment arm with bots signaling political identity more saliently.[16] As discussed in the experimental design section, instead of having the football club's symbol in their profile picture, these bots had the official campaign photo of their preferred candidate as their profile picture, only signaling affective identity in their bio (examples of these accounts can be seen in Appendix Figure B.3). Considering accounts constructed in this way is relevant as it allow us to consider a more extreme case, in which the bot highlights its political affiliation. Interestingly, we find that even when bots signal political identity more saliently, affective identity plays a role in the formation of ties, particularly when it comes to avoiding blocks.

Figure 7 shows results for this extra treatment arm, while Appendix Figure B.4 shows similar results for the accounts in the main experiment, restricting the analysis to the waves in which we also conducted the extra experiment. We find that, when bots signal political identity more saliently, congruence in political identity becomes more relevant to the formation of ties. Indeed, the effect of sharing political identity on follow-backs is roughly twice as large when political identity is more salient than in the standard case. Yet, congruence in affective identity remains having a role on the formation of ties. For follow-backs, this effect is positive and statistically significant among politically congruent individuals. More importantly, for blocks, we see that the likelihood of blocks when subjects do not share either identity is of 25.9%, against only 18.1% when subjects share affective identity (but not political). This effect — of 7.8 percentage points — is comparable to the effect of congruence in affective identity in the main experiment, where political identity is signaled less saliently by the bots. Therefore, affective congruence still plays a role even when political identity is signalled with higher salience by the bots.

Overall, the result of this extra experiment reinforces our main result: both identities continue to have an impact on the formation of social ties. Importantly, even when political identity is signaled very saliently, sharing preference for a football club has a positive impact on ties, particularly by reducing the probability of blocks. This suggests that the potential positive effect of shared affective identity does not hinge on the degree on which this identity is highlighted.

5.4 Formation of ties and salience of elections

Apart from studying the interplay between sharing different dimensions of identity, our experimental design allows us to study how the follow-back and blocking behavior changes over time. This is particularly interesting since we ran the experiment during the second semester of 2022, both before, during, and after the campaign period of the 2022 presidential election in Brazil. We hypothesize that, during the election, the salience of the political

[16] This extra experiment was not pre-registered, and was conducted as a way of evaluating the robustness of the previous (pre-registered) analysis.

dimension of identity would increase relative to the affective identity, reducing the importance of sharing this dimension of identity on the formation of social ties.

We use two pre-registered strategies to study this hypothesis. First, we consider the official campaign period as defined by Brazil's Superior Electoral Court. The campaign period is the period in which candidates can legally present themselves to the population as presidential candidates, and in which their advertisement can be broadcast. Thus, we divide our experimental period on waves that happened before, during, and after this period. The timeline can be seen on Appendix Figure A.1. Specifically, we create indicators for each period (before, during, and after), and run a specification similar to equation (2) including interactions between the identity congruence indicators and the campaign period indicators.

Results using this methods, for follow-backs and blocks, are displayed in Figure 8, which restricts the analysis to subjects that were active on Twitter throughout the experimental period.[17] The left-hand side figures shows results for follow-backs, and the right-hand side, for blocks. All figures plot the average effect of each identity congruence before, during, and after the election, considering the treatment arms of bots that signaled affective identity only (top), political identity only (middle) or both identities (bottom). These figures also show, for each dimension of identity, the estimated difference in behavior between periods (always relative to waves that happened during the official campaign period).

Reassuringly, when we consider bots that signal affective identity only, we find no difference in the effect of congruence in the affective dimension for waves that happened before, during or after the election. This is expected because the election period should only change the salience of the political dimension (moreover, we control for specific club salience). Therefore, it is interesting to see that the official campaign periods did not change subject's behavior when they were followed by bots that did not signal their political preference.

On the other hand, there is some evidence of differential behavior for treatment arms with bots that signaled either their political identity or both dimensions. First, considering follow-backs, subject's behavior before and during the campaign is similar, both considering accounts signaling political identity only or signaling the two dimensions. Given that we begun the experiment approximately one month before the official campaign period, this could indicate that polarization was already at a high level at this pre-campaign moment. However, we find some evidence that, after the election, the effect of congruence in the political dimension decreased relative to its effect during (or before) the campaign. Indeed, for bots that signal political identity only, the effect of sharing political identity with the bot fell by 5.5 percentage points after the election (compared to during). This difference, of 6 percentage points, is marginally insignificant at the 10% level (p-value of 11%). A similar pattern is observed for the accounts that signal both dimensions of identity: relative to during the election, the effect of sharing political identity fell by 4.1 pp (p-value of 11.3%) after the election for these accounts.

In particular, Appendix Figure B.5 reports some suggestive evidence that the decrease in the effect of political congruence we observe after the election was more pronounced during

[17]This sample restriction was pre-registered and was done to guarantee that we compared similar sets of subjects over time. Results for the full sample are reported in Appendix Tables B.11 and B.12.

the 2022 FIFA World Cup, an extremely popular international football tournament that happened approximately one month after the Brazilian Presidential Election.[18] In this figure, the post-election waves are divided between those before or during the tournament. Focusing on follow-backs for accounts that signaled both dimensions of identity, there is evidence that the effect of sharing political identity was smaller during the World Cup than before it (but also after the election). The difference, of 6.2 pp, is substantial, despite being marginally insignificant at usual levels (p-value of 13.2%). When we focus on accounts signaling a single dimension of identity, there is no evidence of difference in behavior during the World Cup, suggesting that the change we observe post-election did not happen exclusively during the tournament.

While imprecise, these results suggest that, after the election, the relative importance of political identity to the formation of ties was somewhat reduced. This result is consistent with an interpretation related to social identity theory (as discussed in Section 3): when the political identity was more salient (during the campaign period), its relative importance to an individual's "sense of self" was greater, and therefore sharing political preference mattered more to the formation of ties was greater. The larger reduction observed during the World Cup (at least for accounts signaling both identities) reinforces this interpretation, as this tournament, widely followed in Brazil, may have further reduced the perceived salience of political identity. However, there are other alternative explanations to the result we document. For instance, it could be that the meaning of the political identity signal changed over time after the election (for instance, being pro-Lula or pro-Bolsonaro had a different meaning before or after the election). It could also be that some subjects changed their political preference over time. Therefore, while interesting, these results must be interpreted with caution.

Moreover, the results for blocks do not fully support the conclusions of the results for follow-backs. Indeed, for the bots that signal both dimensions of identity, there is some evidence that sharing political identity had a larger effect on preventing blocks before the official campaign period. This result might reflect that, when the experiment started, political polarization was already high or may be related to subjects becoming less active on Twitter over time, as discussed in the following section. Yet, comparing results during and after the campaign, we do not have evidence that the blocking behavior changed significantly.

The official campaign period may be unable to capture fluctuations in the salience of elections caused by relevant events such as debates or voting days, that can happen during the official campaign period. To take this into account, we also report heterogeneity results using an electoral salience measure based on the Google Trends Index of search volume for the names of the two presidential candidates we analyzed (Lula and Bolsonaro). This measure may vary at every wave, capturing more subtle changes in electoral salience than the official campaign periods. We report this analysis in Appendix C. We find some evidence that, when the election was more salient, sharing affective identity becomes relatively less important among those who share political identity.

Overall, results using both methods suggest that there has been some change in behavior

[18]The World Cup is the most watched sport tournament worldwide. Over 3.5 billion people watched the 2018 edition (FIFA, 2018).

due to changes in electoral salience (either measured directly, by Google Trends, or indirectly considering official campaign periods). Those effects, however, are small and not always point to the same direction — particularly when we consider blocks vs follow-backs. The results can be interpreted in light of social identity theory, as discussed in Section 3. When the political dimension is more salient, individuals may give a larger weight to their political identity, overshadowing other social categories with which they identify. Our follow-back results suggest that this is indeed happening: after the election, it is expected that the salience of politics diminishes, and we observe a decrease in the relative importance of political congruence on causing follow-backs. More directly, we find that an increase in salience — as measured by Google Trends — reduces the relative importance of sharing affective identity among politically congruent individuals. These results may reflect changes in the relative weight people give to their political identity. However, for this interpretation to be valid, we must assume that the meaning of the social categories we analyse were fixed, which may not be the case (for instance, being pro-Lula before the election might signal different characteristics than being pro-Lula after the election). We must also assume that individuals in our sample did not change their political preference (which is more likely given that they were politically engaged). Thus, our results most likely capture a combination of these different channels.

5.5 Robustness and Validity

In this section, we report some robustness tests and results of additional exercises that may provide further validity to the results and interpretations discussed so far. One relevant concern, particularly in the analysis across time, is that the pool of subjects may change over time due to attrition or reduction on Twitter activity. To address these issues, we repeat our analysis for different sub-samples of subjects, discussed in what follows.

5.5.1 Attrition and Activity

First, following our pre-analysis plan, we repeat both our static and dynamic analyses for a sample of subjects that never suffered attrition. Results of the static analysis are displayed on Column (4) of Appendix Table B.9, for the case of bots that signal both dimensions of identity. Similar analyses are reported in Appendix Table B.10 for the treatment arms with accounts signaling a single dimension of identity. We estimate Equation (2) for the sample of subjects that did not suffer attrition at any point in the experiment. Results are remarkably similar to the main results (reported on the third column of Appendix Table B.9).

However, suffering attrition is not the only way an account may, in practice, "leave" the experiment. It is possible that individuals who used Twitter frequently at the beginning of the experiment become less active on the platform over time, leading to a fall in experimental take-up (and, potentially, a change in the characteristics of the pool of effectively treated subjects). Indeed, Figure B.2 in the Appendix shows that, for our entire pool of subjects, there is a fall in take-up (both in terms of follow-backs and of blocks) over time. To address this concern, we create two indicators of Twitter activity, using information on the exact

date of tweets and re-tweets (the only type of activity we observe directly). First, we create a wave-specific indicator, equal to one if subjects that were assigned to treatment on that wave had tweeted or re-tweeted a status on the 24 hours before treatment. Considering only those subjects, we see that take-up is fairly constant (second panel of Appendix Figure B.2), suggesting that the fall on take-up is driven by a fall on Twitter activity. We also create a second measure of activity at the subject level, equal to one if the subject tweeted in all weeks the experiment was running. This second measure is important for the dynamic analysis, as the type of subjects who are active in each wave may change. For both sub-samples of active users (active in every week and active the day before treatment), Appendix Table B.9 show that results for our main static analysis are also unchanged (columns (6) and (7), respectively).

It is more important, however, to verify whether the dynamic results are similar for these different sub-samples. This is done in Tables B.11 (for follow-backs) and B.12 (for blocks), in the Appendix. For follow-backs, results are extremely similar across all specifications, suggesting that the results discussed on the previous section are robust. On the other hand, for blocks, the differences in behavior between waves that happened before and during the campaign period we find using the sample of active users is smaller than when using the full sample. For some of the sub samples, the effects even lose statistical significance. This suggests that those results are at least partially driven by the fact that, in the beginning of the experiment, there were more active users who were likely to block the experimental accounts.

5.5.2 Effect of Specific Clubs

Another 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. To show that this is not the case, we repeat our analysis excluding one bot club (and its rivals) at a time. This analysis, displayed in Appendix Table B.13, 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 bot throughout the experiment. This may be a concern since those subjects can only be assigned to be followed by a bot 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.14.

Conclusion

Affective polarization has been growing recently in many countries, 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 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. Both dimensions of identity are relevant to forming ties, but the effect of sharing political identity is considerably greater.

Our main contribution is to show that, in a setting of intense affective polarization, such as Brazil (particularly during an election), political identities are capable of overshadowing other dimensions of identity in the formation of ties. Indeed, the effect of supporting the same football club is much smaller when a bot signals its political preference than when it does not. This suggests that signaling political identity undermines connections that could have been formed if such identity had not been signaled. Thus, we document that a potential consequence of affective polarization is to reduce social interactions that could be formed based on other shared interests.

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. Indeed, we show that, in our setting, 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, and ultimately increasing segregation and polarization.

Moreover, this result has implications for the debate about 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, we also show that sharing affective identity — 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 did reduce the effect of affective congruence, overshadowing this dimension of identity. Nevertheless, 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. Moreover, the positive effect of shared football club appeared even when a bot 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. One interesting direction of future research would be to analyze how these shared identities can be best exploited to reduce political divides.

Finally, this paper does have some limitations that suggest other possible directions for future research. First, while the use of follow-backs and blocks is interesting as these outcomes

represent a willingness to connect or to be as far apart as possible from other social media users, it would be interesting to see if the type of phenomena we document is also present in other contexts. For instance, it would be particularly relevant to document whether affective polarization changes how people interact in other contexts or even how they react to different non-political events. Second, while the experiment was conducted over a relatively long period, it started in a context where the Brazilian presidential elections were relatively close — and, therefore, salient. Thus, it would be interesting to analyze the behaviors we studied further away from an election period to verify how they change when the political identity is even less salient. Finally, since our sample is composed of politically-engaged individuals and our experiment happened during a highly polarized period, we are unable to assess whether the type of behavior we document would generalize to other individuals and other periods. 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 advance our understanding of the consequences of affective polarization, and the mechanisms that can reinforce or reduce such polarization.

References

Acemoglu, Daron, Tarek A Hassan, and Ahmed Tahoun, "The power of the street: Evidence from Egypt's Arab Spring," *The Review of Financial Studies*, 2018, *31* (1), 1–42.

Ajzenman, Nicolás, Bruno Ferman, and Pedro C Sant'Anna, "Discrimination in the Formation of Academic Networks: A Field Experiment on #EconTwitter," 2022.

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.

Allcott, Hunt, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow, "The welfare effects of social media," *American Economic Review*, 2020, *110* (3), 629–76.

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

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.

_ , _ , Ruben Enikolopov, and Maria Petrova, "Social media and xenophobia: evidence from Russia," Technical Report, National Bureau of Economic Research 2019.

_ , Michael Callen, Bruno Ferman, Saad Gulzar, Ali Hasanain, and Noam Yuchtman, "Political identity: Experimental evidence on anti-Americanism in Pakistan," *Journal of the European Economic Association*, 2020, *18* (5), 2532–2560.

Charness, Gary and Yan Chen, "Social identity, group behavior, and teams," *Annual Review of Economics*, 2020, pp. 691–713.

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

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.

Di Tella, Rafael, Ramiro H Gálvez, and Ernesto Schargrodsky, "Does Social Media cause Polarization? Evidence from access to Twitter Echo Chambers during the 2019 Argentine Presidential Debate," Technical Report, National Bureau of Economic Research 2021.

Enikolopov, Ruben, Alexey Makarin, and Maria Petrova, "Social media and protest participation: Evidence from Russia," *Econometrica*, 2020, *88* (4), 1479–1514.

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.

FIFA, "2018 FIFA World Cup Russia Global broadcast and audience summary," https://digitalhub.fifa.com/m/2589b77c20849beb/original/njqsntrvdvqv8ho1dag5-pdf.pdf 2018. Accessed: 12-27-2022.

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.

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

González-Bailón, Sandra and Yphtach Lelkes, "Do social media undermine social cohesion? A critical review," *Social Issues and Policy Review*, 2022.

Grossman, Gene M and Elhanan Helpman, "Identity politics and trade policy," *The Review of Economic Studies*, 2021, *88* (3), 1101–1126.

Guess, Andrew M, "Experiments using social media data," *Advances in experimental political science*, 2021, *184*.

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.

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.

Mignozzetti, Umberto and Matias Spektor, "Brazil: when political oligarchies limit polarization but fuel populism," in Thomas Carothers and Andrew O'Donohue, eds., *Democracies divided*, Brookings Institution Press, 2019, chapter 9, pp. 228–256.

Mosleh, Mohsen, Cameron Martel, 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).

_ , Gordon Pennycook, and David G. Rand, "Field Experiments on Social Media," *Current Directions in Psychological Science*, 2022.

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.

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.

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.

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

Shayo, Moses, "A model of social identity with an application to political economy: Nation, class, and redistribution," *American Political science review*, 2009, *103* (2), 147–174.

_ , "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.

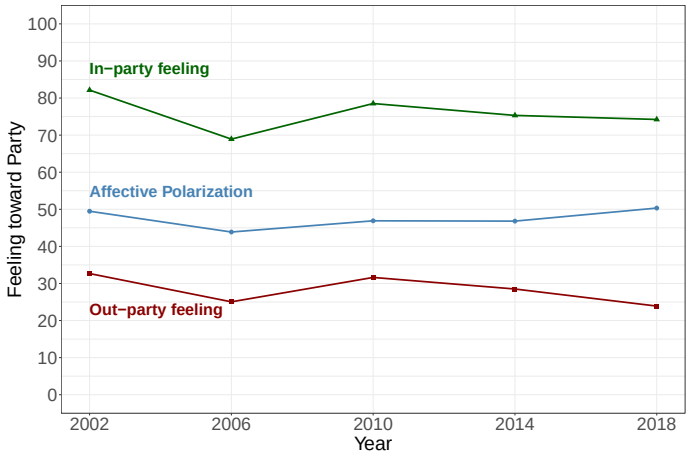
_ **and John C Turner**, *The social identity theory of intergroup behavior*, Chicago: Nelson-Hall, 1986.

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

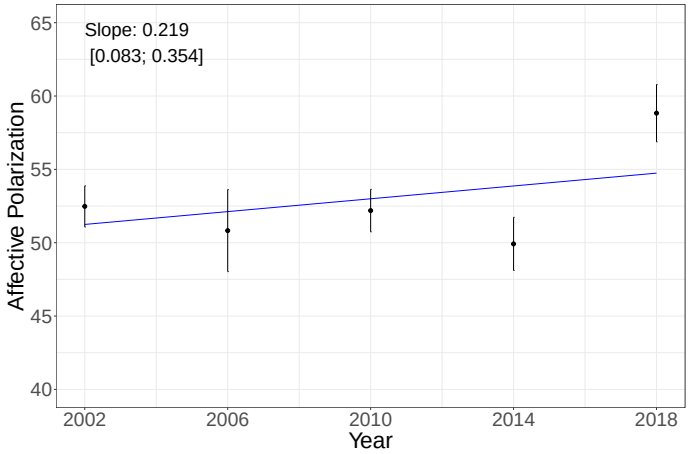
Zhuravskaya, Ekaterina, Maria Petrova, and Ruben Enikolopov, "Political effects of the internet and social media," *Annual review of economics*, 2020, *12*, 415–438.

Figures

Figure 1: Trends in Affective Polarization, Brazil



(a) Trends in Affective Polarization



(b) Affective Polarization (Boxell et al. (2022)'s method)

Notes: The figures present trends in affective polarization in Brazil, using data from the Brazilian Electoral Study (BES), a national post-electoral survey undertaken since 2002. The top panel (Panel 1a) shows trends in average feeling for the party respondents like the most (in-party feeling) and for the other parties (out-party feeling). Affective polarization in this case is the mean difference between in-party and out-party feeling. Figure 1b estimate affective polarization using the same data, but using the method of Boxell et al. (2022). In this cases, 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 report the slope (change per year) and estimated 95% confidence interval computed using heteroskedasticity-robust standard errors in the top-left.

Figure 2: Examples of Bot Accounts

(a) Pro-Bolsonaro; Flamengo supporter

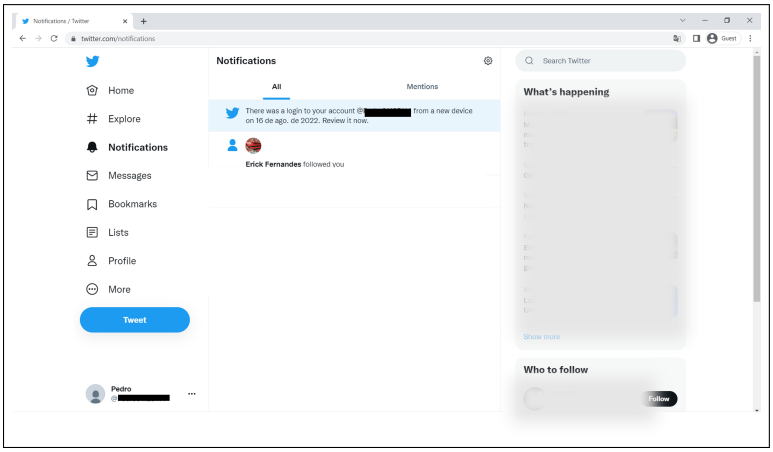
(b) Pro-Lula; Palmeiras supporter

(c) Pro-Lula; Neutral-Team

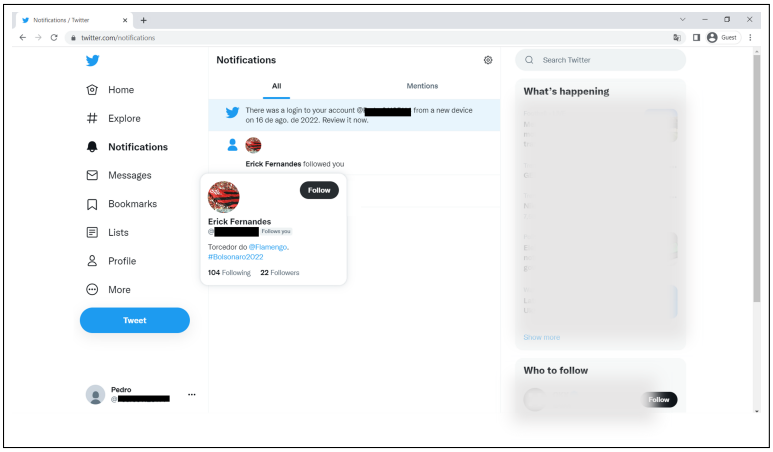
(d) Politically Neutral, Flamengo supporter

Notes: The figures show examples of bot accounts used in the experiment.

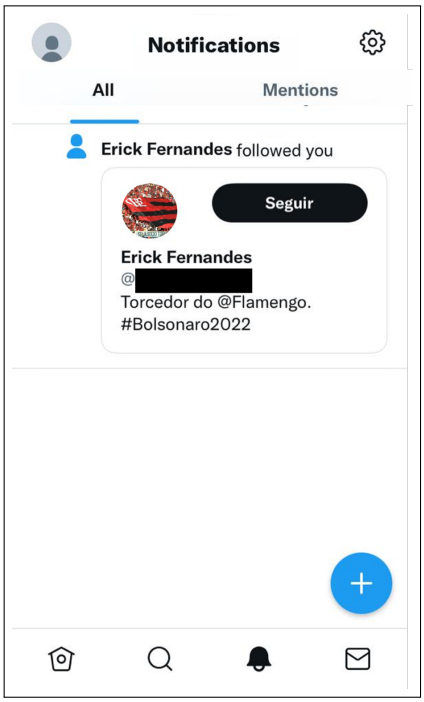
Figure 3: Example of treatment notifications on desktop and mobile Twitter apps



(a) Desktop Notification

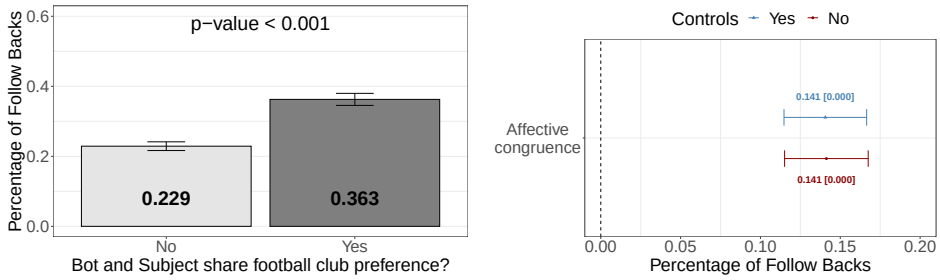


(b) Desktop Notification (after hovering the mouse's cursor over the bot's profile)

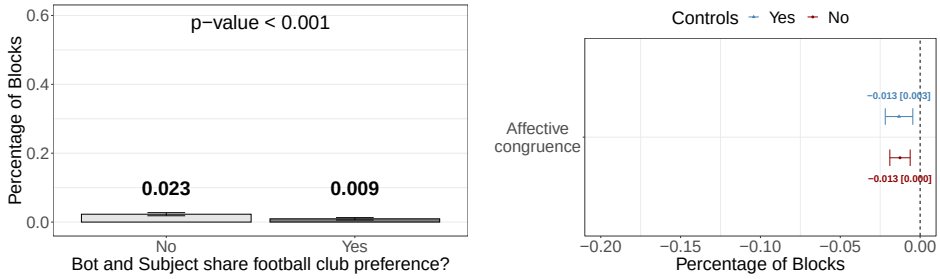


(c) Mobile app notification

Figure 4: Effect of shared affective identity (football club preference) on the formation of social ties – Experimental accounts that do not signal political preference



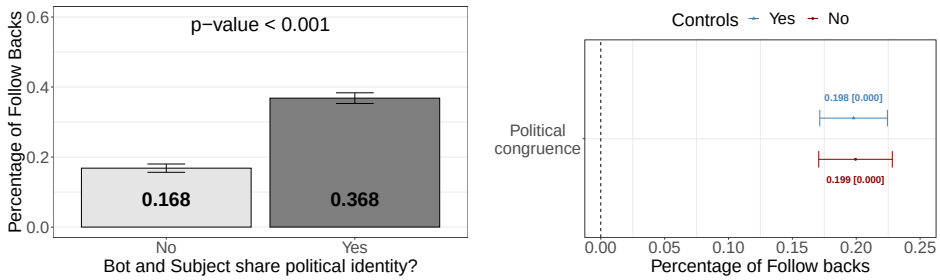
(a) Follow Backs



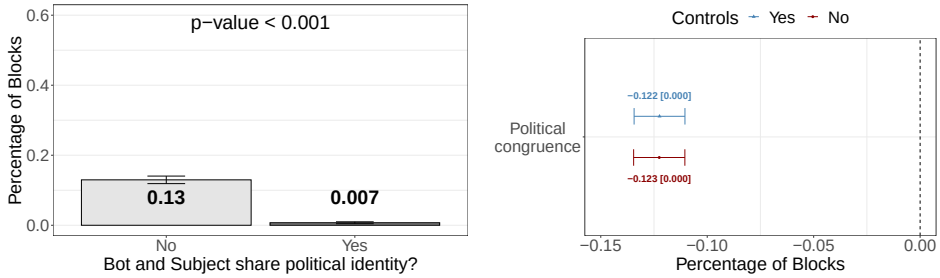
(b) Blocks

Notes: The figures show the effect of sharing affective identity (football club preference) on the rate of follow-backs and blocks. The sample is composed of the subject-bot pairs of politically neutral bots (i.e., we only consider bots that do not signal political identity). The figure on the left shows the average rate of follow-backs or bots 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. The left-hand side plot shows the coefficients estimated from equation (1), 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 bot account level.

Figure 5: Effect of shared political identity on the formation of social ties – Experimental accounts that do not signal football club preference



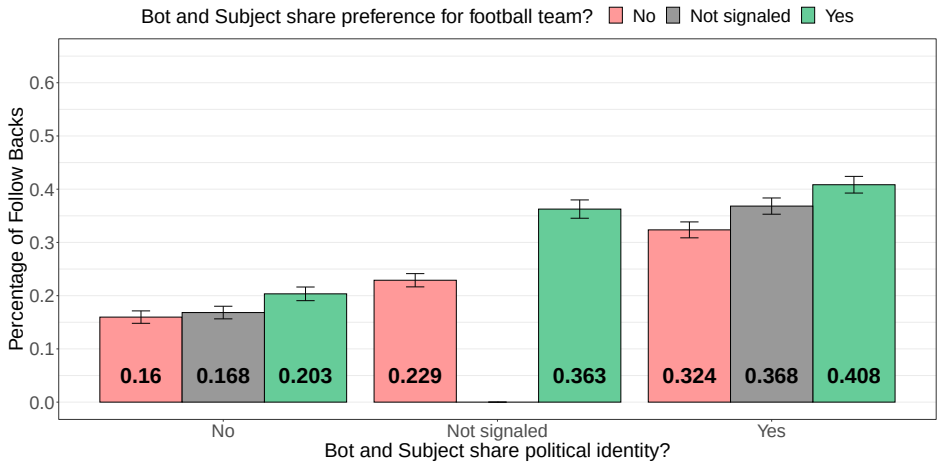
(a) Follow Backs



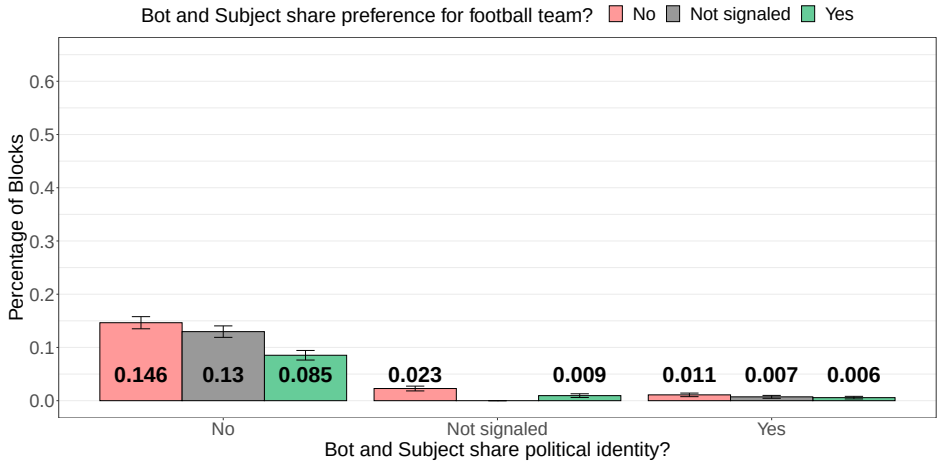
(b) Blocks

Notes: The figures show the effect of sharing political identity on the rate of follow-backs and blocks. The sample is composed of the subject-bot pairs of affectively neutral bots (i.e., we only consider bots that do not signal football club preference). The figure on the left shows the average rate of follow-backs or bots 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. The left-hand side plot shows the coefficients estimated from equation (1), 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 bot account level.

Figure 6: Effect of shared political and affective identity on the formation of social ties



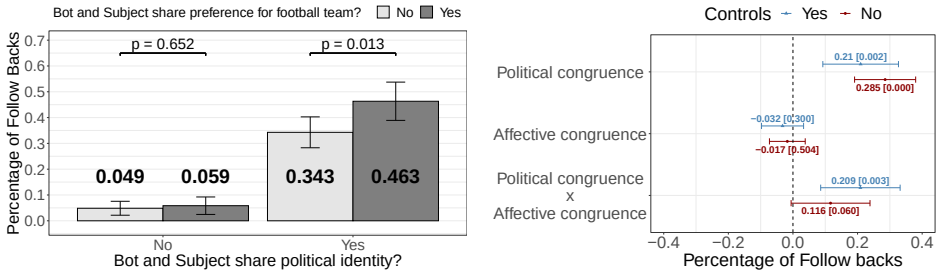
(a) Follow Backs



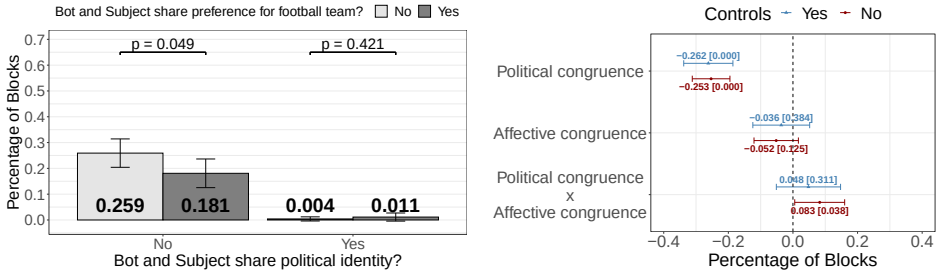
(b) Blocks

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 (bots that signal both or a single dimension of identity). The sample is composed of the subject-bot pairs in the experiment, pooling all waves and excluding shadow-banned accounts (as pre-registered). The x-axis shows whether bot and subject share political identity (or show that this dimension is not signaled by the bots), while the colors show whether bot 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. The error bars represent 95% confidence intervals.

Figure 7: Effect of shared political and affective identity on the formation of social ties: Bot accounts with more salient political identity



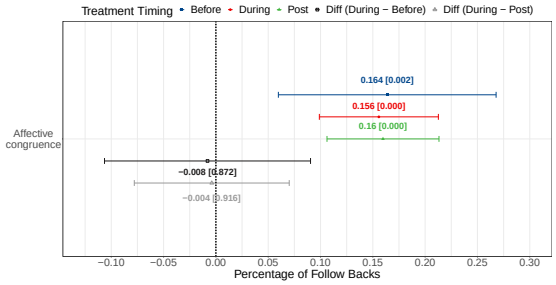
(a) Follow Backs



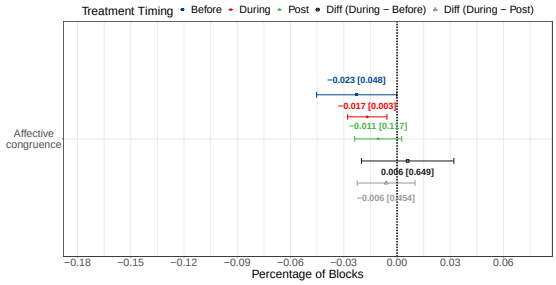
(b) Blocks

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 bot accounts with a more salient political identity. The figure on the left shows the average rate of follow-backs or bots 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 bot account level.

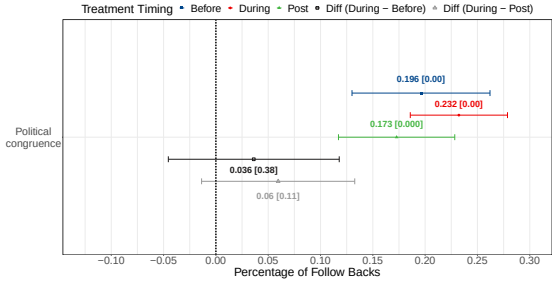
Figure 8: Heterogeneity on Treatment Timing: Official Campaign Periods



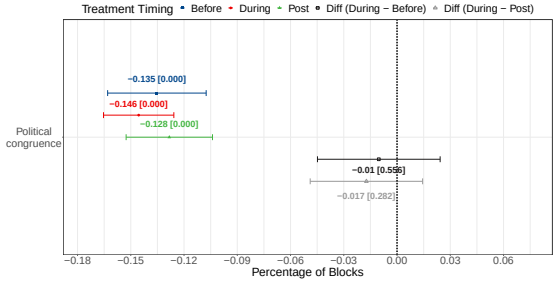
(a) Effect on Follow Backs, bots signaling affective identity only



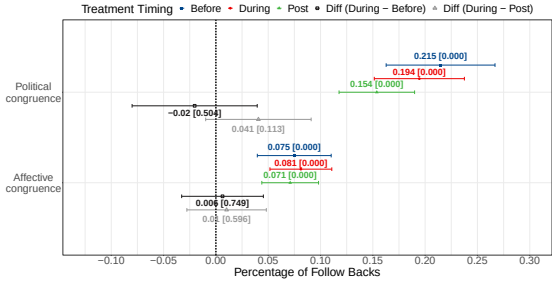
(b) Effect on Blocks, bots signaling affective identity only



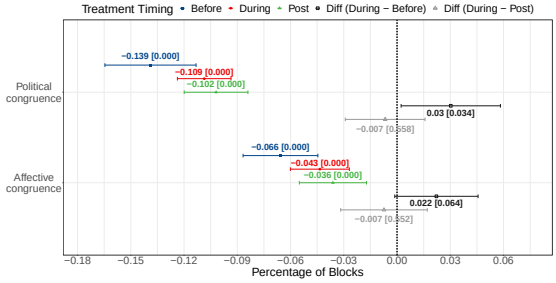
(c) Effect on Follow Backs, bots signaling political identity only



(d) Effect on Blocks, bots signaling political identity only



(e) Effect on Follow Backs, bots signaling both dimensions of identity



(f) Effect on Blocks, bots signaling both dimensions of identity

Notes: The figures show the effect of shared political or affective identity on follow-backs (left) or blocks (right) depending on the treatment timing, dividing the experimental period according to the official campaign calendar of the Brazilian 2022 Presidential Election, as shown in Appendix Figure A.1. The sample consider treatment arms with bots that signaled political identity only (top), affective identity only (middle) or both dimensions of identity (bottom). We restrict the analysis to subjects that were active before every wave in which they were assigned to treatment. Results for the whole subject pool are available in Appendix Tables B.11 and B.12. We create indicators for the official campaign period and estimate an equation similar to (2) including interactions between each campaign period indicators and the identity congruence indicators. From these estimates, we obtain the average effect of identity congruence before, during and after the campaign, as well as the estimated difference between each two periods. We plot the average effect of sharing identity in each period, as well as the difference in behavior pre and post campaign relative to during. All plots show 95% confidence intervals, coefficient estimates and and p-values in brackets. Confidence intervals and p-values are computed using standard errors clustered at the bot account level.

Tables

Table 1: Procedures used to create the bot 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 bot accounts contains two 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 bot accounts' profiles do not include a location.
Website	The bot accounts' profiles do not include an website.
Retweets	The bot 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 access the bot's profile is the one that signals political identity.
Followers	We asked a group of colleagues to follow the bot accounts before each experimental wave.
Following	One day before following the accounts randomly assigned to it, the bot 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 bot accounts. Figure 2 on the Appendix shows examples of accounts.

Appendix

A Additional Information on Experimental Design

A.1 Pro-Lula and Pro-Bolsonaro Hashtags

Table A.1: 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.2 Football Club Rivalries

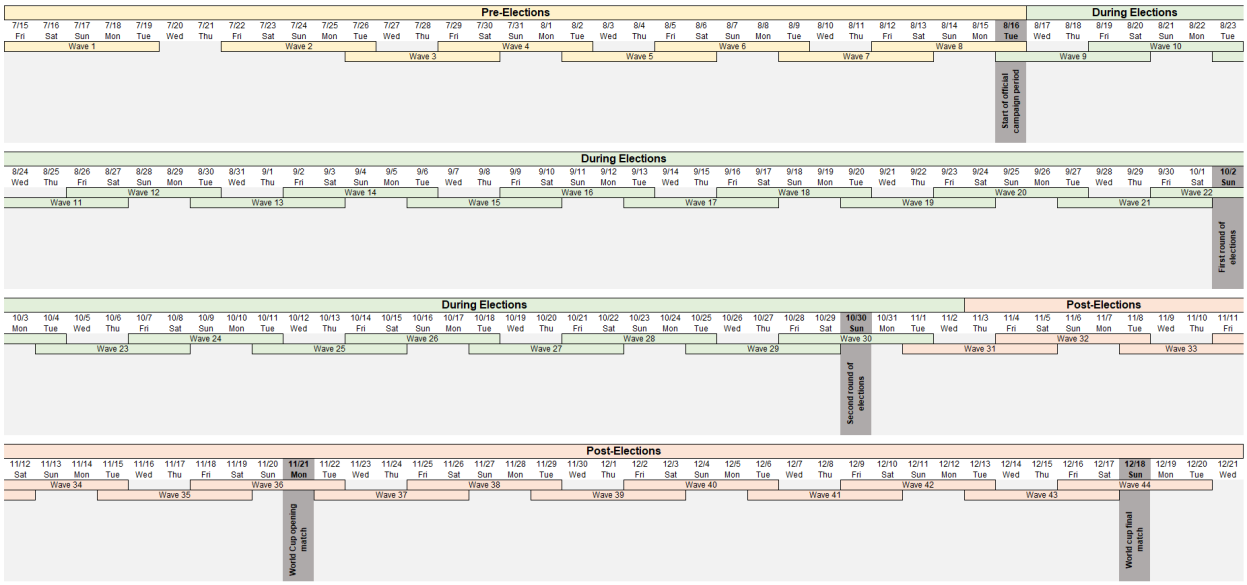
Table A.2: Football club rivalries

	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 bot 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 bot may support, while the clubs in the columns are the ones that subjects may support.

A.3 Experimental Timeline

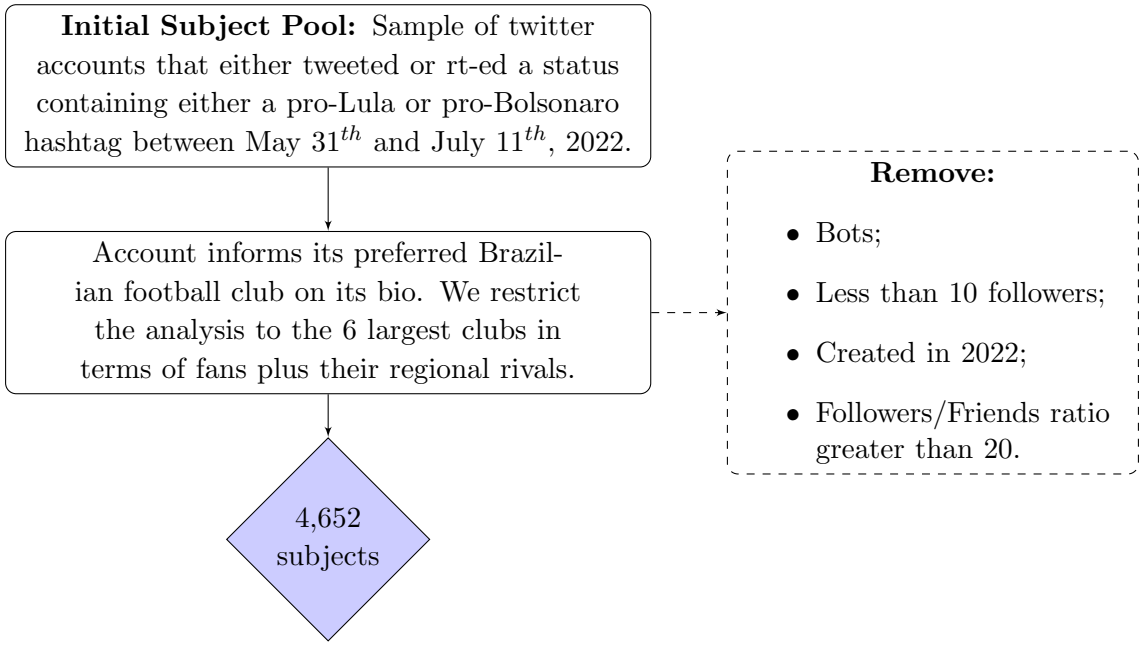
Figure A.1: Experimental Timeline



Notes: The table shows the experimental timeline. We consider that each wave starts at the moment in which the bot 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.4 Procedure to Obtain the Subject Pool

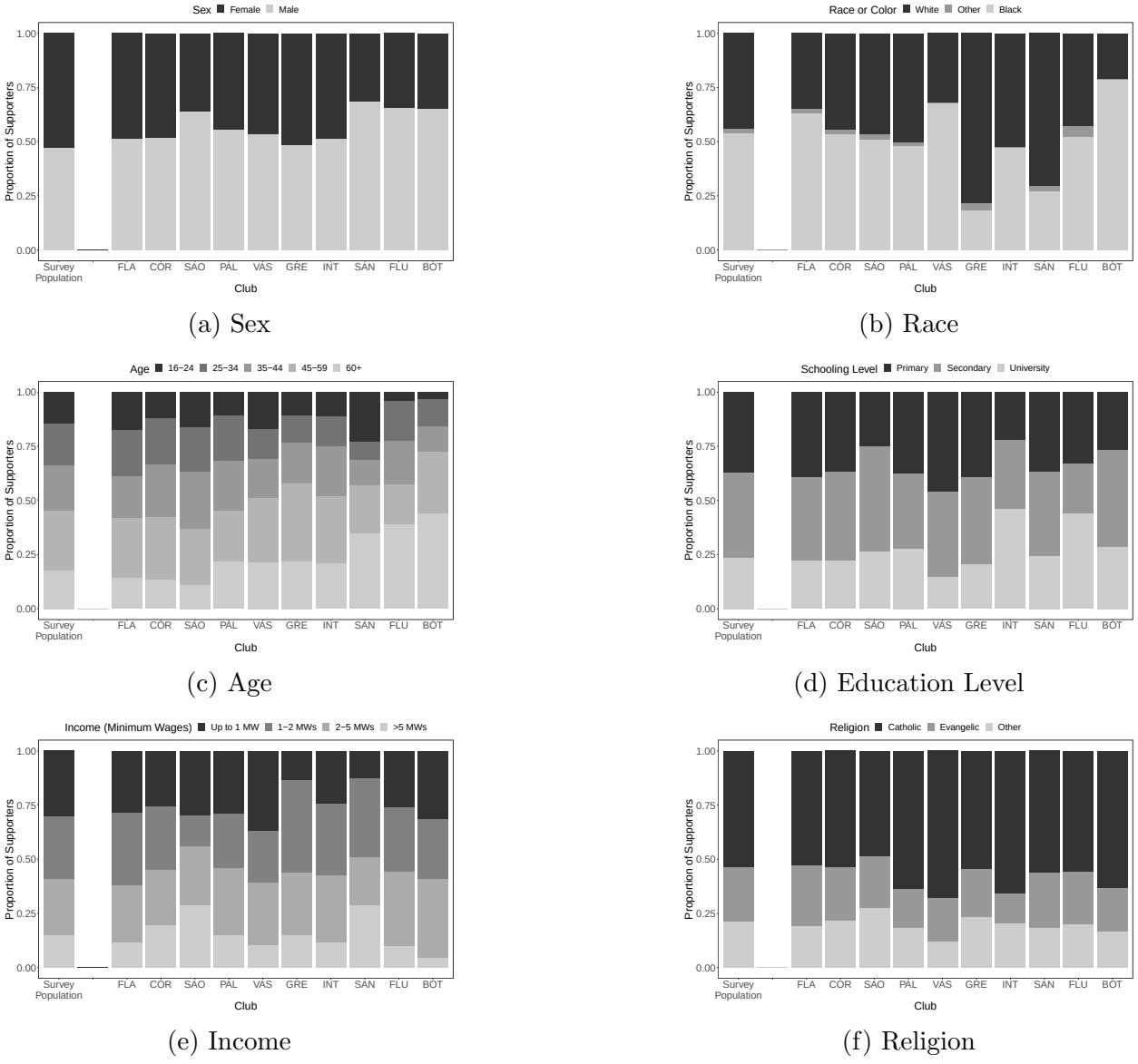
Figure A.2: Procedure to obtain the subject pool



B Additional Figures and Tables

B.1 Characteristics of Brazilian Football Club Supporters

Figure B.1: 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.2 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 accounts followed ('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	2021	4652

Notes: The table shows summary statistics for the subject pool in the experiment.

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
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.3 Balance, Attrition, and Take-up

Table B.3: Balance Table - Experiment with accounts signaling both dimensions of identity

	Treatment Arm				
Variable	In-politics; In-affective	In-politics; Out-affective	Out-politics; In-affective	Out-politics; Out-affective	F Stat [p-value]
Number of followers	1,858.1	1,939.5	1,826.5	2,032.1	0.0144 [0.998]
	(4,899.3)	(4,816.2)	(4,584.2)	(5,387.3)	
Number of friends	2,190.1	2,191.7	2,074.4	2,302.9	0.0193 [0.996]
	(4,548.3)	(3,898.6)	(3,958.3)	(4,779.3)	
Number of statuses ('tweets + rts')	24,448	24,873.2	25,061.3	24,909.8	0.001 [1.00]
	(55,867.4)	(51,507.3)	(56,480.8)	(53,622.5)	
Number of favorited statuses ('likes')	43,139.1	43,136.6	44,731.2	40,517.3	0.0209 [0.996]
	(87,867.7)	(73,385.3)	(83,492.9)	(63,641.3)	
Number of lists	4.024	4.164	4.133	4.33	0.0011 [1.00]
	(24.8)	(20.1)	(28.8)	(25.2)	
Account is verified	0.001	0.001	0.002	0.001	0.0127 [0.998]
	(0.033)	(0.028)	(0.043)	(0.023)	
Year of account creation	2,015.1	2,015.1	2,015.2	2,015.1	0.006 [0.999]
	(4.599)	(4.689)	(4.582)	(4.653)	
Has background picture	0.839	0.843	0.841	0.838	0.0016 [1.00]
	(0.368)	(0.363)	(0.366)	(0.368)	
Gender (1=Female)	0.173	0.172	0.175	0.184	0.0082 [0.999]
	(0.378)	(0.377)	(0.38)	(0.387)	
Region					
Center-West	0.043	0.036	0.041	0.031	0.0335 [0.992]
	(0.202)	(0.186)	(0.198)	(0.172)	
Northeast	0.065	0.064	0.075	0.064	0.0214 [0.996]
	(0.246)	(0.244)	(0.264)	(0.245)	
North	0.021	0.017	0.024	0.023	0.0212 [0.996]
	(0.144)	(0.128)	(0.154)	(0.15)	
Southeast	0.311	0.335	0.303	0.329	0.0437 [0.988]
	(0.463)	(0.472)	(0.459)	(0.47)	
South	0.082	0.072	0.082	0.07	0.023 [0.995]
	(0.274)	(0.258)	(0.274)	(0.255)	
Foreign	0.02	0.02	0.02	0.02	0.0001 [1.00]
	(0.139)	(0.141)	(0.141)	(0.14)	
Number of treated observations	3783	3761	3790	3794	
%	0.25	0.249	0.251	0.251	
Attrition (not treated)	379	415	396	384	0.0068 [0.999]
% of assigned to treatment	0.091	0.099	0.095	0.092	
Always active (tweeted every week)	2863	2830	2900	2923	0.0149 [0.998]
% of treated	0.757	0.752	0.765	0.77	
Active 1 day before treatment	2965	2948	2947	2994	0.0055 [0.999]
% of treated	0.784	0.784	0.778	0.789	

Notes: The table displays average and standard deviations for subject-level variables across the four treatment arms in the experiment with accounts that signal both dimensions of identity (political and affective). 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 four 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 - Experiment with Accounts signaling a single dimensions of identity

	Treatment Arm					
	Politically Neutral Accounts			Affectively Neutral Accounts		
Variable	In-affective	Out-affective	T Stat [p-value]	In-politics	Out-politics	T Stat [p-value]
Number of followers	1,839	2,032.2	-1.4932 [0.135]	2,077.4	1,962	0.896 [0.37]
	(5,067.6)	(5,987.4)		(6,220)	(5,001.6)	
Number of friends	2,132.8	2,312.9	-1.5283 [0.126]	2,313.7	2,221.4	0.8093 [0.418]
	(4,587.9)	(5,494.9)		(5,556)	(4,368.3)	
Number of statuses ('tweets + rts')	24,168.4	26,130.1	-1.3327 [0.183]	24,775.2	25,720.6	-0.8115 [0.417]
	(60,306.6)	(64,734.1)		(51,148.7)	(50,935.8)	
Number of favorited statuses ('likes')	42,112	42,084.4	0.0151 [0.988]	44,915.1	41,968.2	1.6897 [0.091]
	(81,036.1)	(71,119.9)		(82,595.3)	(69,698.8)	
Number of lists	3.157	4.184	-3.0164 [0.003]	4.056	3.619	1.1033 [0.27]
	(10)	(19.1)		(20.5)	(13.4)	
Account is verified	0	0.002	-1.8341 [0.067]	0.002	0.002	-0.5407 [0.589]
	(0.018)	(0.04)		(0.039)	(0.046)	
Year of account creation	2,015.1	2,015	1.1946 [0.232]	2,015	2,015.2	-1.1322 [0.258]
	(4.599)	(4.655)		(4.585)	(4.59)	
Has background picture	0.833	0.838	-0.5218 [0.602]	0.839	0.83	1.1046 [0.269]
	(0.373)	(0.368)		(0.368)	(0.376)	
Gender (1=Female)	0.188	0.179	1.0291 [0.303]	0.169	0.186	-1.9758 [0.048]
	(0.391)	(0.383)		(0.375)	(0.389)	
Region						
Center-West	0.045	0.041	0.8577 [0.391]	0.042	0.041	0.143 [0.886]
	(0.207)	(0.198)		(0.2)	(0.199)	
Northeast	0.082	0.06	3.5348 [0.00]	0.067	0.07	-0.444 [0.657]
	(0.275)	(0.238)		(0.25)	(0.255)	
North	0.032	0.016	4.2984 [0.00]	0.023	0.021	0.6795 [0.497]
	(0.177)	(0.126)		(0.15)	(0.143)	
Southeast	0.329	0.335	-0.5273 [0.598]	0.311	0.295	1.4495 [0.147]
	(0.47)	(0.472)		(0.463)	(0.456)	
South	0.09	0.071	2.8831 [0.004]	0.072	0.071	0.0946 [0.925]
	(0.286)	(0.257)		(0.259)	(0.258)	
Foreign	0.015	0.016	-0.5298 [0.596]	0.021	0.018	1.0492 [0.294]
	(0.12)	(0.126)		(0.144)	(0.133)	
Number of treated observations	3003	4385		3845	3833	
%	0.406	0.594		0.501	0.499	
Attrition (not treated)	356	378		346	363	
% of assigned to treatment	0.106	0.079		0.083	0.087	
Always active (tweeted every week)	2300	3353		2901	2908	
% of treated	0.766	0.765		0.754	0.759	
Active 1 day before treatment	2253	3435		3030	2969	
% of treated	0.75	0.783		0.788	0.775	

Notes: The table displays average and standard deviations for subject-level variables across the treatment arms in the experiment with accounts that signal a single dimensions of identity (either affective or political). For each variable, we compute a t test of difference in means between the two groups of subjects. In most cases, we cannot reject the null hypothesis of equality of means across 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.5: Balance Table - Attrited subjects

	Treatment Arm (Dimensions of identity signaled by bots)								
	Both dimensions				Only affective identity		Only political identity		
Variable	In-politics; In-affective	In-politics; Out-affective	Out-politics; In-affective	Out-politics; Out-affective	In-affective	Out-affective	In-politics	Out-politics	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	

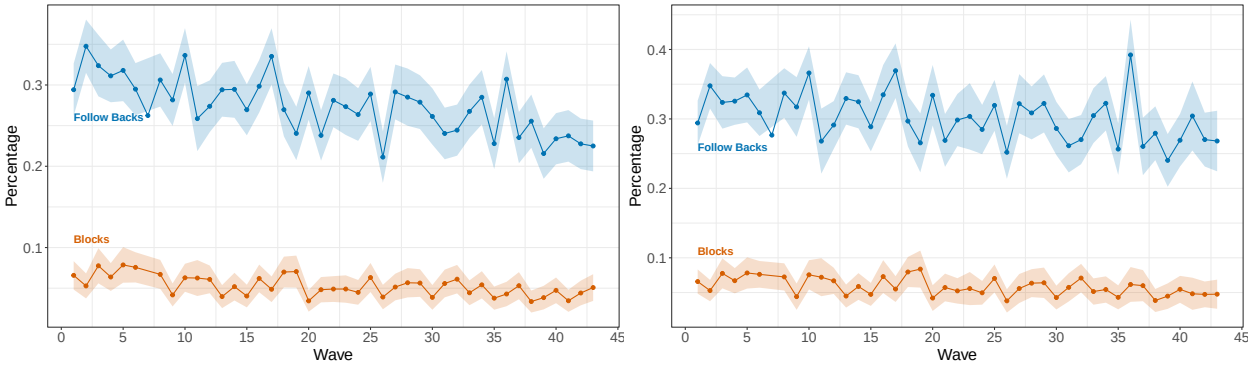
Notes: This table shows the average and standard deviations (in parentheses) of pre-treatment variables for subjects that suffered attrition at some point of the experiment. The last column in the table reports a F-test of joint equality of means across all treatment arms.

Table B.6: 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.2: 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.4 Main Results: Comparison of Results across Treatment Arms and Robustness

Table B.7: 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 6). The treatment arms are defined by whether bot 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 bot does not signal the dimension. For each treatment arm, we first inform the relationship between bot and subject's political identity, and then affective (for example, "in; out" means that bot 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.8: 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 6). The treatment arms are defined by whether bot 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 bot does not signal the dimension. For each treatment arm, we first inform the relationship between bot and subject's political identity, and then affective (for example, "in; out" means that bot 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.9: 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)
	(1)	(2)	(3)	(4)	(5)	(6)
Political congruence	0.1639***	0.1643***	0.1476***	0.1439***	0.1622***	0.1606***
	(0.0108)	(0.0106)	(0.0139)	(0.0139)	(0.0165)	(0.0166)
Affective congruence	0.0437***	0.0424***	0.0512***	0.0473***	0.0597***	0.0551***
	(0.0087)	(0.0087)	(0.0114)	(0.0129)	(0.0145)	(0.0136)
Political congruence × Affective congruence	0.0411***	0.0387***	0.0503***	0.0531***	0.0364*	0.0521**
	(0.0129)	(0.0127)	(0.0170)	(0.0184)	(0.0211)	(0.0200)
Wave, Strata Fixed Effects	No	Yes	Yes	Yes	Yes	Yes
Controls	No	No	Yes	Yes	Yes	Yes
Observations	15,128	15,128	15,128	13,257	9,953	11,854
R^2	0.04856	0.08886	0.09909	0.09795	0.10527	0.10199
Panel B: Blocks						
	Dependent Variable: Blocks (1 = Yes)					
	Full Sample			Never attrited	Tweeted every week	Active (1 day)
	(1)	(2)	(3)	(4)	(5)	(6)
Political congruence	-0.1355***	-0.1354***	-0.1267***	-0.1062***	-0.1193***	-0.1469***
	(0.0072)	(0.0073)	(0.0081)	(0.0080)	(0.0092)	(0.0092)
Affective congruence	-0.0611***	-0.0609***	-0.0652***	-0.0518***	-0.0623***	-0.0797***
	(0.0076)	(0.0076)	(0.0093)	(0.0093)	(0.0115)	(0.0115)
Political congruence × Affective congruence	0.0559***	0.0553***	0.0578***	0.0457***	0.0534***	0.0707***
	(0.0078)	(0.0078)	(0.0097)	(0.0096)	(0.0120)	(0.0121)
Fixed-effects						
Wave, Strata Fixed Effects	No	Yes	Yes	Yes	Yes	Yes
Controls	No	No	Yes	Yes	Yes	Yes
Observations	14,737	14,737	14,737	12,945	9,718	11,501
R^2	0.05768	0.06426	0.06790	0.05552	0.06102	0.07730

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. Standard errors clustered at the bot account level are in parentheses. Significance codes: $^{***}: p < 0.01$, $^{**}: p < 0.05$, $^{*}: p < 0.1$.

Table B.10: 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)
	(1)	(2)	(3)	(4)	(5)	(6)
Affective congruence	0.1337***	0.1413***	0.1454***	0.1548***	0.1747***	0.1604***
	(0.0133)	(0.0134)	(0.0187)	(0.0196)	(0.0212)	(0.0213)
Wave, Strata Fixed Effects	No	Yes	Yes	Yes	Yes	Yes
Controls	No	No	Yes	Yes	Yes	Yes
Observations	7,388	7,388	7,388	6,583	4,983	5,688
R^2	0.02123	0.06732	0.08339	0.09017	0.09770	0.08595
Panel B: Blocks, Affective Identity Only						
	Dependent Variable: Blocks (1 = Yes)					
		Full Sample		Never attrited	Tweeted every week	Active (1 day)
	(1)	(2)	(3)	(4)	(5)	(6)
Affective congruence	-0.0134***	-0.0126***	-0.0132***	-0.0113**	-0.0129**	-0.0135***
	(0.0031)	(0.0032)	(0.0043)	(0.0046)	(0.0053)	(0.0048)
Wave, Strata Fixed Effects	No	Yes	Yes	Yes	Yes	Yes
Controls	No	No	Yes	Yes	Yes	Yes
Observations	7,199	7,199	7,199	6,424	4,859	5,516
R^2	0.00253	0.01003	0.01773	0.01529	0.01757	0.02072
Panel C: Follow Backs, Political Identity Only						
	Dependent Variable: Blocks (1 = Yes)					
		Full Sample		Never attrited	Tweeted every week	Active (1 day)
	(1)	(2)	(3)	(4)	(5)	(6)
Political congruence	0.2000***	0.1994***	0.1979***	0.1880***	0.1982***	0.2076***
	(0.0148)	(0.0147)	(0.0133)	(0.0135)	(0.0164)	(0.0162)
Fixed-effects						
Wave, Strata Fixed Effects	No	Yes	Yes	Yes	Yes	Yes
Controls	No	No	Yes	Yes	Yes	Yes
Observations	7,678	7,678	7,678	6,823	5,079	5,999
R^2	0.05092	0.08798	0.10159	0.09892	0.10616	0.10787
Panel D: Blocks, Political Identity Only						
	Dependent Variable: Blocks (1 = Yes)					
		Full Sample		Never attrited	Tweeted every week	Active (1 day)
	(1)	(2)	(3)	(4)	(5)	(6)
Political congruence	-0.1225***	-0.1225***	-0.1224***	-0.1075***	-0.1188***	-0.1386***
	(0.0062)	(0.0061)	(0.0060)	(0.0058)	(0.0074)	(0.0068)
Wave, Strata Fixed Effects	No	Yes	Yes	Yes	Yes	Yes
Controls	No	No	Yes	Yes	Yes	Yes
Observations	7,492	7,492	7,492	6,668	4,961	5,830
R^2	0.05894	0.06775	0.07063	0.06323	0.07162	0.08179

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 bots 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. Standard errors clustered at the bot account level are in parentheses. Significance codes: *** : $p < 0.01$, ** : $p < 0.05$, * : $p < 0.1$.

B.5 Experiment with Bots with more Salient Political Identity

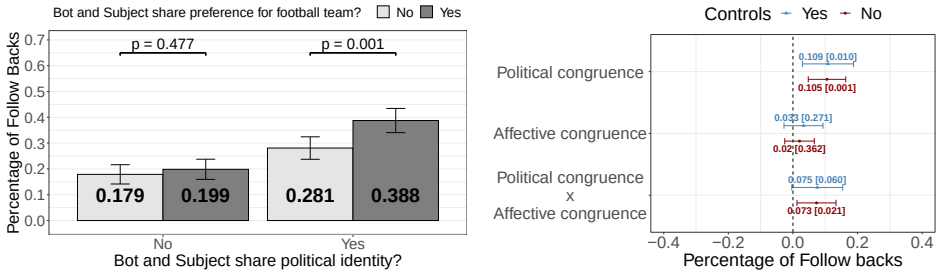
Figure B.3: Examples of Bot Accounts - More salient political identity

(a) Pro-Bolsonaro; São Paulo supporter

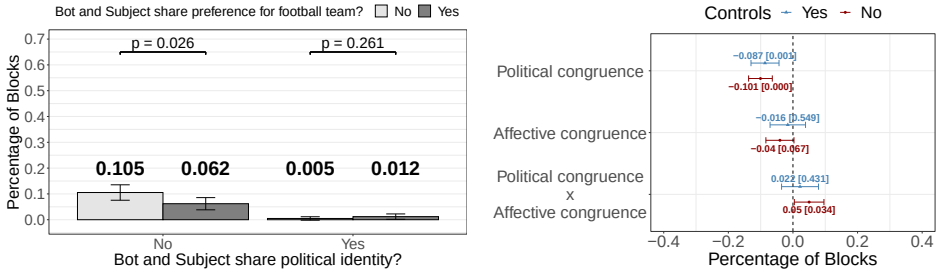
(b) Pro-Lula; Palmeiras supporter

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

Figure B.4: 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 bot accounts of the original experiment, restricting the analysis for the waves in which we conducted the extra experiment with bots with more salient political identity. The figure on the left shows the average rate of follow-backs or bots 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 bot account level.

B.6 Heterogeneity on Treatment Timing: Robustness and Effects during the FIFA World Cup

Table B.11: Heterogeneity on Treatment Timing and Election Salience for Different Subsamples: Follow Backs

Dependent Variable:	Follow Backs (1 = Yes)							
	Full Sample		Never attrited		Always active		Active (1 day)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Google Trends Index of Electoral Salience								
Political congruence × Political GTI	0.0001		0.0003		0.0004		0.0005	
	(0.0008)		(0.0008)		(0.0009)		(0.0008)	
Affective congruence × Political GTI	0.0005		0.0007		0.0007		0.0008	
	(0.0009)		(0.0009)		(0.0008)		(0.0009)	
Political congruence × Affective congruence × Political GTI	-0.0020**		-0.0019*		-0.0016		-0.0019*	
	(0.0009)		(0.0010)		(0.0010)		(0.0011)	
Official Campaign Period								
Political congruence × Before Campaign		0.0079		-0.0168		-0.0134		-0.0020
		(0.0293)		(0.0317)		(0.0344)		(0.0322)
Affective congruence × Before Campaign		-0.0189		-0.0475*		-0.0389		-0.0288
		(0.0192)		(0.0245)		(0.0285)		(0.0205)
Political congruence × Post Campaign		-0.0427*		-0.0493**		-0.0381		-0.0514*
		(0.0225)		(0.0229)		(0.0295)		(0.0287)
Affective congruence × Post Campaign		-0.0148		-0.0138		-0.0148		-0.0209
		(0.0190)		(0.0203)		(0.0251)		(0.0239)
Political congruence × Affective congruence × Before Campaign		0.0431		0.0780*		0.0565		0.0449
		(0.0325)		(0.0399)		(0.0443)		(0.0333)
Political congruence × Affective congruence × Post Campaign		0.0223		0.0216		0.0159		0.0211
		(0.0237)		(0.0235)		(0.0299)		(0.0299)
Wave, Strata Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	15,128	15,128	13,257	13,257	9,953	9,953	11,854	11,854
R^2	0.09854	0.09839	0.09695	0.09708	0.10252	0.10257	0.10113	0.10118

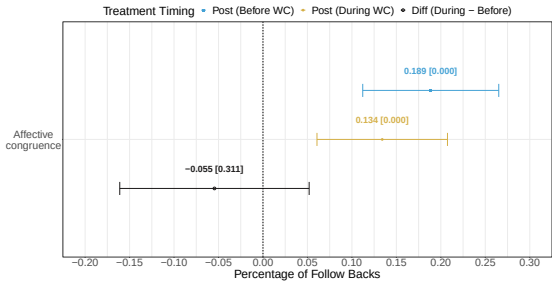
Notes: The table displays estimates for the effect of political salience (as measured by the Google Trends Index explained in the text or by the official campaign dates) on follow-backs for different sub-samples of subjects: full sample, never attrited and active at least one day before treatment. We restrict the analysis to the main experiment (bots that signaled both dimensions of identity), excluding shadow-banned accounts as discussed in text. 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. A subject is "always active" if they tweeted at least once in the seven days before the beginning of every single wave. Active subjects are those who tweeted or re-tweeted a status one day before being treated. Standard errors clustered at the bot account level are in parentheses. Significance codes: *** : $p < 0.01$, ** : $p < 0.05$, * : $p < 0.1$.

Table B.12: Heterogeneity on Treatment Timing and Election Salience for Different Subsamples: Blocks

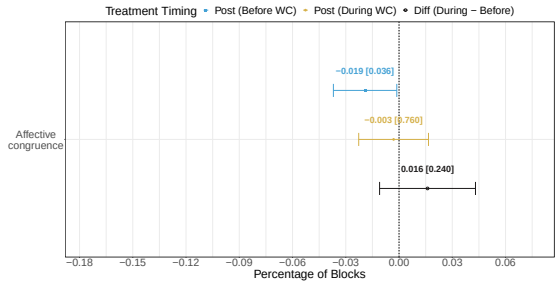
Dependent Variable:	Blocks (1 = Yes)							
	Full Sample		Never attrited		Always active		Active (1 day)	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Google Trends Index of Electoral Salience								
Political congruence × Political GTI	-0.00005		-0.00009		-0.0003		-0.0002	
	(0.0007)		(0.0008)		(0.0010)		(0.0008)	
Affective congruence × Political GTI	-0.0005		-0.0007		-0.0014*		-0.0005	
	(0.0005)		(0.0006)		(0.0008)		(0.0009)	
Political congruence × Affective congruence × Political GTI	0.0005		0.0006		0.0012		0.0005	
	(0.0005)		(0.0007)		(0.0009)		(0.0009)	
Official Campaign Period								
Political congruence × Before Campaign		-0.0679***		-0.0586**		-0.0643**		-0.0497**
		(0.0207)		(0.0234)		(0.0263)		(0.0212)
Affective congruence × Before Campaign		-0.0464**		-0.0317		-0.0421		-0.0417*
		(0.0222)		(0.0263)		(0.0295)		(0.0236)
Political congruence × Post Campaign		-0.0039		-0.0054		-0.0106		-0.0006
		(0.0152)		(0.0146)		(0.0171)		(0.0178)
Affective congruence × Post Campaign		0.0005		0.0003		-0.0085		0.0007
		(0.0177)		(0.0174)		(0.0215)		(0.0227)
Political congruence × Affective congruence × Before Campaign		0.0423*		0.0290		0.0408		0.0388
		(0.0232)		(0.0268)		(0.0302)		(0.0246)
Political congruence × Affective congruence × Post Campaign		0.0125		0.0098		0.0210		0.0129
		(0.0178)		(0.0178)		(0.0221)		(0.0231)
Wave, Strata Fixed-effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	14,737	14,737	12,945	12,945	9,718	9,718	11,501	11,501
R^2	0.06773	0.06978	0.05533	0.05684	0.06116	0.06222	0.07706	0.07835

Notes: The table displays estimates for the effect of political salience (as measured by the Google Trends Index explained in the text or by the official campaign dates) on blocks for different sub-samples of subjects: full sample, never attrited and active at least one day before treatment. We restrict the analysis to the main experiment (bots that signaled both dimensions of identity), excluding shadow-banned accounts as discussed in text. 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. A subject is "always active" if they tweeted at least once in the seven days before the beginning of every single wave. Active subjects are those who tweeted or re-tweeted a status one day before being treated. Standard errors clustered at the bot account level are in parentheses. Significance codes: $^{***}: p < 0.01$, $^{**}: p < 0.05$, $^{*}: p < 0.1$.

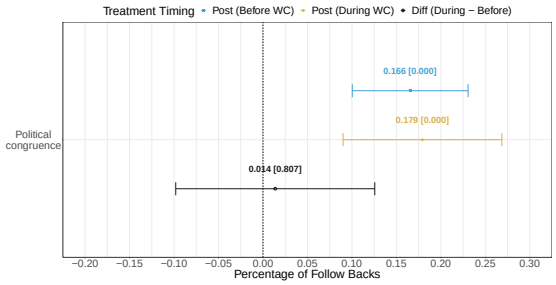
Figure B.5: Heterogeneity on Treatment Timing After the Election: Before and During the FIFA World Cup



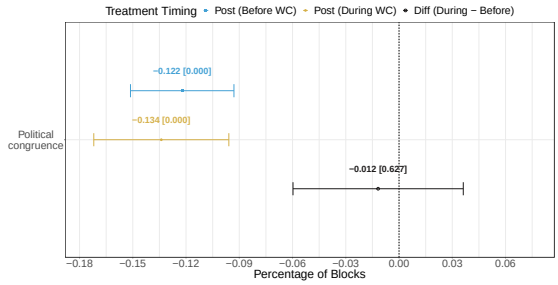
(a) Effect on Follow Backs, bots signaling affective identity only



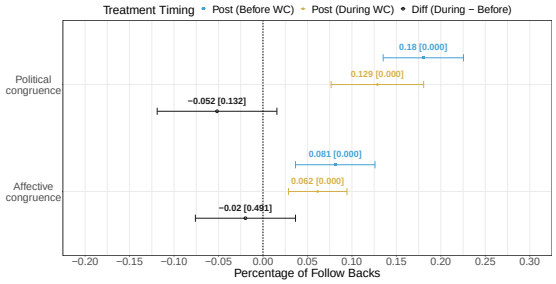
(b) Effect on Blocks, bots signaling affective identity only



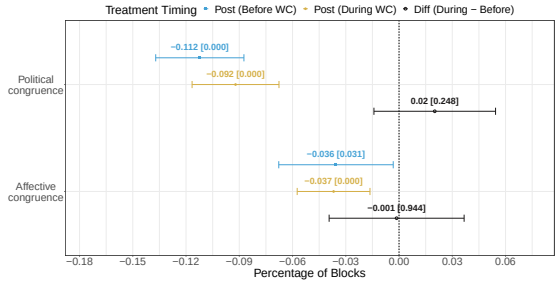
(c) Effect on Follow Backs, bots signaling political identity only



(d) Effect on Blocks, bots signaling political identity only



(e) Effect on Follow Backs, bots signaling both dimensions of identity



(f) Effect on Blocks, bots signaling both dimensions of identity

Notes: The figures show the effect of shared political or affective identity on follow-backs (left) or blocks (right) after the 2022 Brazilian Presidential Election, either before or during the 2022 FIFA World Cup. The sample consider treatment arms with bots that signaled political identity only (top), affective identity only (middle) or both dimensions of identity (bottom). We restrict the analysis to subjects that were active before every wave in which they were assigned to treatment. We create indicators for the official campaign and the World Cup periods and estimate an equation similar to (2) including interactions between each period indicator and the identity congruence indicators. From these estimates, we obtain the average effect of identity congruence before and during the World Cup, as well as the estimated difference between each two periods. We focus exclusively on waves happening after the end of the election. We plot the average effect of sharing identity in each of these two periods, as well as the difference in behavior between them. All plots show 95% confidence intervals, coefficient estimates and and p-values in brackets. Confidence intervals and p-values are computed using standard errors clustered at the bot account level.

B.7 Other Robustness Exercises

Table B.13: Main Results Excluding Bots' 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 bot account level are in parentheses. Significance codes: *** : $p < 0.01$, ** : $p < 0.05$, * : $p < 0.1$.

Table B.14: Experiment Results Excluding Clubs Not Signaled by Bots

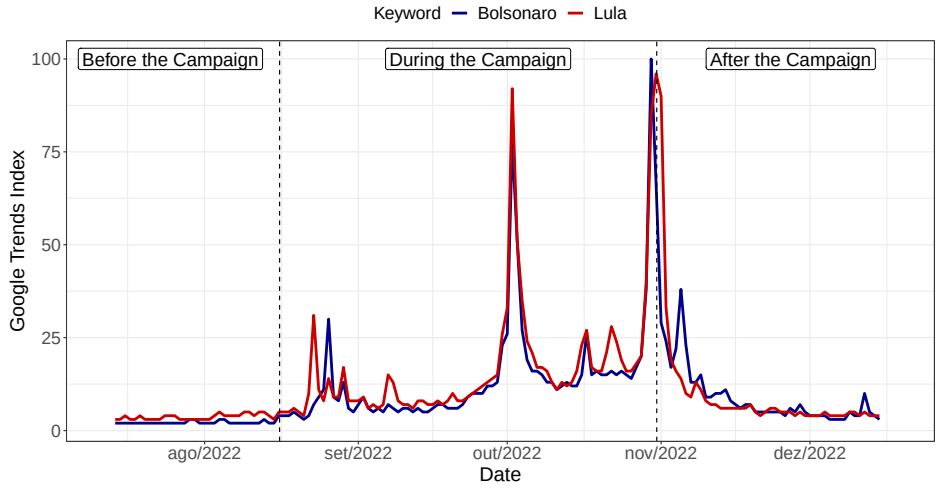
Panel A: Bots signaling affective Identity Only				
Dependent Variables:	Follow Backs (1 = Yes)		Blocks (1 = Yes)	
Sample:	Full	Excluding non-signaled Clubs	Full	Excluding non-signaled Clubs
	(1)	(2)	(3)	(4)
Affective congruence	0.1454***	0.1636***	-0.0132***	-0.0123**
	(0.0187)	(0.0204)	(0.0043)	(0.0048)
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: Bots 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 bot 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 bots during the experiment. Standard errors clustered at the bot account level are in parentheses. Significance codes: *** : $p < 0.01$, ** : $p < 0.05$, * : $p < 0.1$.

C Formation of ties and salience of elections: alternative analysis using Google Trends

As an alternative method to study the effect of electoral salience on the follow-back and blocking behavior (Section 5.4), we constructed a political salience index using Google Trends (GTI). For each experimental wave, we collected the GTI in Brazil for the names of the two presidential candidates we analysed (Lula and Bolsonaro). The evolution of these index during our experimental period can be seen in Appendix Figure C.1.

Figure C.1: 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.

The figure shows that the correlation between the GTI and the official campaign calendar exists, but is not perfect. During the official campaign period (from mid-August to the end of October) the average GTI is higher, but there are moments when we observe a spike in interest on the candidates (specifically, the first debates and the days around the first round and run-off election). Using the GTI allows us to capture this type of fluctuations in interest that would not be observed considering the official calendar. Our final GTI is the average between the index for each of the two candidates, in the first two days of each wave — since these are the days in which the majority of follow-backs and blocks happen. We then run a specification similar to equation (2) including interactions between the GTI and the identity

congruence indicators. Specifically, we estimate the following equation:

$$\begin{aligned} Y_{ijst} = \alpha + \beta_1 \times \text{political\_congruence}_{ij} + \beta_2 \times \text{affective\_congruence}_{ij} + \\ \beta_3 \times \text{political\_congruence}_{ij} \times \text{affective\_congruence}_{ij} + \\ \beta_4 \times \text{political\_congruence}_{ij} \times GTI_t + \beta_5 \times \text{affective\_congruence}_{ij} \times GTI_t + \\ \beta_6 \times \text{political\_congruence}_{ij} \times \text{affective\_congruence}_{ij} \times GTI_t \\ X_{ijt}\lambda + \delta_t + \theta_s + \phi_{st} + \varepsilon_{ijst} \quad (3) \end{aligned}$$

where GTI_t is the google trends index of wave t and the remaining variables have the same definition as before. Note that the GTI is an index from 0 to 100, where higher values represent a greater google search volume for the terms "Lula" and "Bolsonaro". We are interested in coefficients β_4, β_5 and β_6, which represent the marginal effect of an increase of the GTI by one unit on the effect of congruence in the respective identity dimension. Importantly, as in all previous analyses, we control for the salience of the football clubs used in the experiment with a similar google trends index. Hence, our estimates control for fluctuations in the relative salience given to specific clubs during the experimental timeline.

Table C.1: Heteregeneity by Treatment Timing: Google Trends Index

Dependent Variables:	Follow Backs			Blocks		
Treatment Arm	Affective only	Political only	Both	Affective only	Political only	Both
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Political congruence		0.2006***	0.1456***		-0.1222***	-0.1252***
		(0.0180)	(0.0156)		(0.0078)	(0.0107)
Affective congruence	0.1438***		0.0452***	-0.0135***		-0.0591***
	(0.0217)		(0.0137)	(0.0048)		(0.0108)
Political congruence × Affective congruence			0.0721***			0.0518***
			(0.0200)			(0.0114)
Political congruence × GTI		-0.0003	0.0001		-1.83×10^{-5}	-5.33×10^{-5}
		(0.0011)	(0.0008)		(0.0003)	(0.0006)
Affective congruence × GTI	8.52×10^{-5}		0.0005	-1.75×10^{-5}		-0.0005
	(0.0008)		(0.0009)	(0.0002)		(0.0005)
Political congruence × Affective congruence × GTI			-0.0020**			0.0005
			(0.0009)			(0.0005)
Wave, Strata Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Observations	7,388	7,678	15,128	7,199	7,492	14,737
R^2	0.08183	0.10160	0.09854	0.01689	0.07063	0.06773

Notes: This table displays results of the effect of electoral salience, as measured by a Google Trends Index, on follow-back and blocking behaviors. We create a Google Trends Index (GTI) for each experimental wave, equal to the average search volume on Google in Brazil for the terms "Lula" and "Bolsonaro", on the first two days of each wave. We then use this index to estimate specifications similar to Equation (3). The table shows estimates of this model for treatment arms containing bots that signaled both dimensions of identity or a single one. Standard errors clustered at the bot account level are in parentheses. Significance codes: *** : $p < 0.01$, ** : $p < 0.05$, * : $p < 0.1$.

Appendix Table C.1 shows results using this methodology, both for treatment arms with bots that signaled a single dimension of identity and with bots that signaled both. Overall, we find little evidence of differences in behavior depending on the salience of the election as measured by Google Trends. First, for bots that signal a single dimension of identity (either political or affective), changes in election salience do not seem to change the relative effect of sharing identities.

On the other hand, election salience does have some effect for bots signaling both dimensions, at least when it comes to follow-backs. First, conditional on not sharing the other

identity, there is no effect of electoral salience on the effect of political congruence, nor on the effect of affective congruence. This is seen by the fact that we estimate null effects for β_4 and β_5 in column (3) of Appendix Table C.1. However, for the interaction between political and affective congruence, we obtain a negative effect of the political GTI. This can be interpreted as follows: conditional on bot and subject sharing political identity, an increase of 10 units in the GTI causes a decrease in the effect of affective congruence of approximately 2 percentage points. In other words, when politics is more salient, sharing preference for football clubs becomes relatively less important among those who share political identity. The results for blocks, on the other hand, are slightly different, since we find that changes in the GTI did not impact the effect of congruence in each dimension of identity.