Measuring Alignment of Online Grassroots Political Communities with Political Campaigns

XXXX
XXXX
XXXX

XXXX
XXXX
XXXX

XXXX
XXXX
XXXX

Abstract

Social media reduces barriers for the formation of large, self-organizing grassroots communities. For political campaigns this poses significant opportunities to address declining party membership, but also reputational risks and potential loss of campaign coherence. While balancing these factors is often done informally, this study implements implement a data-driven method for measuring the alignment between online political communities and their corresponding campaigns. We apply this technique to the 2020 U.S. Democratic presidential primaries, providing novel insights into the important tension between campaigns and third-party actors.

Adopting a behavioural approach, we apply neural embedding techniques, creating a "community embedding", to evaluate grassroots communities along various cultural, political, and demographic dimensions. Our embedding demonstrates that these communities align with the views of a candidate's supporters. However, this does not necessarily reflect the campaign's policy platforms. Finally, we introduce temporal aspects to our community embedding to evaluate the stability of political communities and their interrelations. Our methods provide new insights into how politics on Reddit operates in relation to its offline analogs.

Introduction

Online forums are becoming an increasingly central conduit through which political communication is organized (Stier et al. 2018). Their adoption is prevalent among both political campaigns, who take advantage of the reach and control that online platforms methods afford (Stier et al. 2018; Bossetta 2018; Enli and Skogerbø 2013), and grassroots movements, who share information, discuss key issues, and solve collective action problems (Mills 2018; Bennett 2012). This is especially the case during election cycles, when political actors are faced with opportunities and challenges as to how best harness the energy, authenticity, and new forms of political content that third-party actors introduce.

One approach to engaging with third-party actors is top-down organization. MyBO, a platform created by then-Senator Barack Obama, was utilized during his 2008 US presidential bid to organize and cultivate his support base (Gibson 2015; Penney 2017). The open platform was one of the first to distribute many key campaign operations online. Through MyBO, users were able to generate resources with a donation mechanism, mobilize get-out-the-vote initiatives independent of the campaign, and share information to other platforms like Facebook and Twitter (Gibson 2015). While the Obama campaign still oversaw, monitored and moderated MyBO — ensuring that the platform was used in ways that aligned with the campaign — it offloaded critical organizational roles to semi-autonomous individuals who were not officially affiliated with the campaign (Penney 2017).

On the other end of the spectrum, online political communities can be completely separate from the campaigns they support. Bernie Sanders' Dank Meme Stash (BSDMS) was a Facebook page, formed during Senator Bernie Sanders' 2016 US presidential bid, dedicated to sharing humorous pro-Sanders images and content (Penney 2017). Formed by a college student with no affiliation to the Sanders campaign, it quickly amassed over 400,000 members, any of whom could contribute content. BSDMS became a large hub of political content, allowing for more experimentation in messaging and an 'edgier' campaign (Penney 2017). However, BSDMS wasn't without its warts. While the Sanders campaign was able to benefit from BSDMS, there were no mechanisms for reigning in problematic content. Harmful content directed at Sanders' opponent, Hillary Clinton, contributed to the "Bernie Bros" image that portrayed Sanders' supporters as misogynistic (Penney 2017). BSDMS's lack of alignment on such messaging posed substantial risk to Sanders, and was never integrated with the campaign.

These opposing cases are examined by Penney, who proposed a taxonomy of digital grassroots movements with respect to political campaigns (2017). Penney contrasts "official" digital community structures with "unofficial" structures that operate in a more independent manner and as a result offer different benefits, and pose different dangers, to a campaign. The more aligned the third-party actors are with a campaign's messaging, the less risk they pose to the campaign, and as a result the more integrated they tend to be. However, grassroots organizations can aid political campaigns by introducing new energy, authenticity and forms of political content. Alignment can be spurious, as a result of the third-party's existing norms, or it can be enforced

through campaign oversight. However, offloading responsibility can result in a lack of campaign cohesion. If the third party is ill-aligned, this can pose risks to the campaign, as was the case with BSDMS.

However, characterizing the alignment between online political communities and the campaigns they support is made difficult by the rapid proliferation of political communities and a dearth of formal methods for analysis. This is not only the case in academia, but also for political campaigns who often decide questions of alignment on an informal basis (Penney 2017; Dommett and Temple 2018; Gibson 2015). Thus, as third-party actors become increasingly relevant in electoral cycles, it is important to better understand their relationship with official political campaigns.

We propose a methodology to understand online political communities' alignment with official campaigns in a principled, data-driven way, and apply our methodology to online political communities centered around the 2020 U.S. Democratic presidential primaries. While Facebook and Twitter are most commonly studied, they are both platforms where user activity is strongly guided by algorithmic recommendation. As such, empirical studies of activity on them can be algorithmically confounded. To avoid this problem, we analyse political communities on Reddit, where users have more granular control over their user experience. With over 430 million monthly users, Reddit is used by over 7% of Americans as a primary news source, providing a large and rich dataset for analysis (Pew Research Center 2018). Using our method, we measure the alignment between political communities on Reddit dedicated to political campaigns and the campaigns themselves. In doing so we answer the question: Do these Reddit communities align with the campaign's they support?, which within Penney's taxonomy would warrant closer integration, or do they deviate in meaningful ways?

The Present Work: Measuring Alignment of Online Grassroots Political Communities with Political Campaigns. Our methodology builds on neural embedding techniques to represent communities on Reddit as dense vectors. This allows us to make full use the Reddit ecosystem and utilize the community itself as a unit of analysis, overcoming previous hurdles to the analysis of online communities. Our method differs from previous work quantifying political and ideological alignment online in two main ways. First, prior work has largely been focused on assessing individual-level political leanings, which does not easily translate to community-level analysis like the one presented in this paper. Second, previous attempts largely rely on explicit markers of partisan or ideological association, such as survey responses, self-identified party affiliation (Bakshy, Messing, and Adamic 2015), or textual analysis, to characterize political alignment. Given the pseudonymous nature of Reddit, administering surveys can result in response bias and poses difficulties when scaling to tens of thousands of unique communities (Van de Mortel and others 2008). Additionally, given that norms vary wildly across communities on Reddit, textual analyses — such as building word clouds and training topic or sentiment models — may struggle to fully capture the constellation of communities and how they each relate to one another (Rajadesingan, Resnick, and Budak 2020). As such, we adopt a behavioural method that positions each community in relation to one another based on where users decide to spend their time, as indicated by their comment patterns.

Within our embedding vector space we construct eight cultural, political and demographic dimensions — ideology, birth control, gun control, trade, hawkishness, religiosity, age and gender — that we measure candidate focused communities on Reddit against. Our projections are analyzed against survey data and policy analysis, giving two metrics for ways in which communities can be aligned, or deviate from the campaign itself.

Finally, the relationship between political campaigns and external actors is not uni-directional. One mechanism that allows grassroots communities to shape the broader political discourse is through public sentiment analysis. This provides a messaging feedback loop to political campaigns (McGregor 2020). Public sentiment analysis is often operationalized via social media — which, relative to administering public opinion surveys, gives the ability to measure sentiment over granular time increments. In line with this literature, we introduce a temporal variant of our embedding applied to the 2020 Democratic primaries. We then also analyze the stability of political communities on Reddit, and whether they become more isolated as the primaries drew on. We find that despite substantial drift among the subreddits overtime, they retain high levels of alignment and actually increase inter-community activity.

Related Work

Web 2.0 ushered in greater interactivity online and a shift away from passive reception of information towards active engagement and production of online content. With this development, and the following proliferation of social media platforms the implications of online platforms for political campaigns have grown. Some see online platforms as a means for parties to be shaped from the "outside in," primarily by viewing the Internet as outside the realm of traditional political actors (Chadwick and Stromer-Galley 2016). Others observe how political actors can co-opt online communication and question whether the Internet serves to reinforce campaign organizational hierarchies, in what has come to be known as the "normalization thesis" (Gibson and Ward 2012; Margolis, Resnick, and Levy 2003). Others yet take a middle ground (Chadwick 2007), arguing that online platforms allow for greater organizational fluidity. These effects have been studied in the US (Penney 2017; Scott and Johnson 2005), the United Kingdom and broadly in political campaigns as well as social movements (Abbott 2012). Focusing on Reddit in particular, recent studies have provided large-scale analyses of political subreddits with respect to different tolerances for toxic content (Rajadesingan, Resnick, and Budak 2020) and comparative analyses of communities from opposing ideological vantages (Soliman, Hafer, and Lemmerich 2019).

While Scott and Johnson noted the ability for grassroots movements to formulate and self-organize online, these

technologies were not initially adopted by official political campaigns. The 2008 US presidential election is viewed as a inflection point in this respect (Gibson 2015). Gibson cites a decline in party membership as a key driver of online adoption for political campaigns, where campaigns opt for more flexible models of organization in order to maximize their strength (Gibson 2015; Penney 2017). This presents a key tension, noted in the literature of digital grassroots movements, where political campaigns balance tapping in to grassroots energy and maintaining campaign coherence. How and when campaigns decide to interact with or integrate third party actors online is a critical question in this respect. Third party organizations can be thought of as an emergent, decentralized organizational network when loosely integrated with the official campaign, termed "connective action" (Bennett and Segerberg 2013), or as a form of "satellite campaign" (Dommett and Temple 2018).

Through interviews with campaign staffers, Penney notes that the degree of integration is often a function of how aligned the third party actor is with the official campaign (Penney 2017). This presents a taxonomy of "official" and "unofficial" forms of grassroots mobilization. The official form encompasses top-down applications which are controlled by the campaign, as seen with MyBO. The unofficial form of grassroots mobilization contains two distinct subcategories. The first are 'amplifiers': communities who are aligned with the campaign, and thus while autonomous are "still within the confines of organizational oversight and curation" (Penney 2017). This contrasts with the other form of campaign group, crowdsourced community pages like BSDMS. Unbound by campaign norms, these groups allow for unrestricted experimentation and production of political communication. While campaigns are happy to benefit from third-parties promoting their campaign independently, these online communities have the potential to deviate from the campaign's ethos in problematic ways. Penney's taxonomy, specifically the two forms of "unoffical" groups, is the primary framework this study is situated in.

Methodology

Data

Our core dataset is derived from Reddit, a decentralized collection of online, topic specific sub-forums called *subreddits*. Within subreddits, users are able to share posts containing text, images or videos from across the Web. Each post also contains a comment section where users are able to discuss that post's content and the subreddit's broader themes. Subreddits can serve different purposes; there are humorous subreddits like *r/funny*, geographic-specific subreddits like *r/toronto*, and a healthy ecosystem of political subreddits. In this work, we use the Pushshift Reddit archive of all comments made on the site (Baumgartner et al. 2020). Given this study's focus on the U.S. 2019-2020 Democratic presidential primaries, we limit the time frame of our dataset to user comments made in 2019. After restricting the Pushshift dataset to the top 10,000 subreddits by activity, our dataset contains 1.59 billion comments from 19.34 million distinct users. This accounts for 95.6% of comments made in 2019.

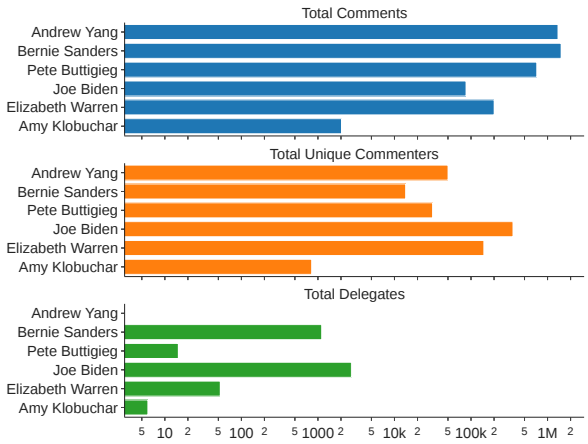

Figure 1: Top plot: total number of comments in candidate subreddit. Middle plot: total number of unique commenters in candidate subreddit. Bottom plot: Total number of delegates achieved for each candidate.

Candidate Subreddits. While our dataset includes all user comments from the 10,000 most active subreddits, we focus on "candidate subreddits": communities dedicated to candidates running to be the 2020 Democratic presidential nominee. While most political communities on Reddit are what Edwards would term an information intermediary, candidate subreddits are interactional intermediaries (Edwards 2006). Information intermediaries are third-party organizations that distribute political information, while interactional intermediaries facilitate political participation (Edwards 2006). Candidate subreddits differ from other political communities on Reddit because they are created with the explicit aim to facilitate political participation. Unlike large political communities on Reddit like *r/Politics*, candidate subreddit raise money for political campaigns – with Sander's candidate subreddit raising over a million dollars for his presidential bid, Biden, Buttigieg, Yang and Warren's subreddits raising in the tens of thousands of dollars for their respective candidates (Reddit 2020c; ActBlue 2020a; Reddit 2020b; 2020a; ActBlue 2020b). With the exception of *r/BaemyKlobaechar* each subreddit organized their own donation tag to keep track of donations made through the subreddit. In addition, candidate subreddits organize canvassing events and other forms of offline political participation, again a key differentiator. Thus, within the context of grassroots movements, they serve as a clear unit of analysis.

However, not all candidates were in the race for a substantial period of time – and given Reddit's decentralized nature, not every candidate had a dedicated subreddit. As such, we limit our analysis to major candidates, i.e. candidates who participated in at least 7 of the 11 nationally televised presidential debates: Joe Biden (*r/JoeBiden*), Bernie Sanders (*r/SandersForPresident*), Elizabeth Warren (*r/ElizabethWarren*), Amy Klobuchar

(*r/Baemy Klobaechar*), Pete Buttigieg (*r/Pete_Buttigieg*) and Andrew Yang (*r/YangForPresidentHQ*).

Bernie Sanders, an especially popular figure on Reddit, had multiple subreddits beyond *r/SandersFor President* that could be considered interactional intermediaries: *r/WayOfTheBern* and *r/OurPresident*. However, we found that either merging all three subreddits in our embedding, or including all three separately, had a negligible impact on our findings. Other candidates, like Andrew Yang, had multiple state-level communities as well. Thus, for simplicity we only retain the largest candidate subreddit for each campaign.

To supplement our core dataset, which only includes comments made during 2019, we used the Pushshift application programming interface (API) to collect all the comments on candidate subreddits from January 1, 2020 to May 1, 2020 – by which point Joe Biden had become the presumptive nominee. Due to API constraints this was not feasible for non-candidate subreddits. As a result, these data are only used in Figures 1 and 7, which only relied on data pertaining to the candidate subreddits and no other Reddit communities. Please note that all other analyses incorporate data from not only candidate subreddits but all of the 10,000 most active subreddits, to situate candidate subreddits within the broader context of political discussion on Reddit. This is accomplished using community embeddings, as we describe below.

Community Embedding

As previously mentioned, Reddit contains many political subreddits other than those dedicated to political campaigns, including broader ideological subreddits and individual-issue oriented subreddits such as *r/prochoice*. By comparing the user-base of a candidate subreddit (i.e. the set of users who comment in it) to that of an issue or ideology subreddit, we can measure how associated the two communities are. This provides a quantitative metric for how 'aligned' that community is with a certain political concept.

To facilitate such comparisons between subreddits, we apply community embeddings, an adaptation of word embeddings trained on behavioural data. Such embeddings generate community vectors such that communities with similar user-bases have similar vector representations. We create an embedding of Reddit communities based on the users who comment in each subreddit, using the skip-gram with negative sampling method as performed in Waller and Anderson (2019). We use our complete Reddit dataset to train the embedding (since a large collection of communities are required, this section does not make use of any data from 2020). After hyperparameter tuning, our optimal embedding correctly answered 82.99% of the 2462 simple analogy problems posed (Waller and Anderson 2019; XXXXX 2020).

In the resulting embedding, the similarity of any two subreddits can be computed simply by taking the cosine similarity of their two vectors. This similarity score represents the strength of association of the user-bases of the two subreddits. Subreddits with many users in common will have a high cosine similarity, whereas subreddits with very few users in common will have a low similarity (Waller and Anderson 2020). For example, the subreddit *r/Conservative* is far more similar to *r/guns* than *r/SandersForPresident*.

Ideology		Religiosity	
Conservative	Progressive	Religious	Secular
r/Conservative	r/progressive	r/Christianity	r/exchristian
r/Republican	r/democrats	r/TraditionalCatholics	r/excatholic
r/conservatives	r/SocialDemocracy	r/lds	r/exmormon
r/TheNewRight	r/WeAreNotAsking	r/mormon	r/exmormon
r/neoliberal	r/dsa	r/islam	r/exmuslim
Gun Control		Birth Control	
Less Regulation	More Regulation	Pro-Life	Pro-Choice
r/progun	r/GunsAreCool	r/prolife	r/prochoice
r/Firearms	r/GunsAreCool	r/prolife	r/birthcontrol
War		Trade	
Hawkish	Pacifistic	Pro-Globalism	Anti-Globalism
r/CredibleDefense	r/EndlessWar	r/neoliberal	r/LateStageImperialism
r/WarCollege	r/EndlessWar	r/Economics	r/capitalism_in_decay
r/Intelligence	r/EndlessWar	r/Libertarian	r/EnoughLibertarianSpam
Age		Gender	
Older	Younger	Masculine	Feminine
r/RedditForGrownups	r/teenagers	r/AskMen	r/AskWomen
r/RedditForGrownups	r/teenagersnew	r/daddit	r/Mommit

Table 1: Political, cultural and demographic dimension definitions

Temporal Community Embedding. Building on recent work in temporal analysis of word embeddings (Kim et al. 2014), we generate additional embeddings for smaller time periods to measure how these associations change over time. Our temporal community embedding is built by first segmenting our dataset into n time steps. For this study, an appropriate sized window was to segment our dataset by month. We then train our skip-gram model over all pairs (c_{jt}, u_i) of user u_i commenting in a subreddit c_j at time t. Thus, in an embedding with n time steps, there are n vector representations for each community. We can evaluate the temporal community embedding by recording the mean performance on the same analogy set for each of the n time steps, avoiding duplicate subreddits in the prediction problem. Using the same hyper-parameters as our static community embedding, our model achieved a mean accuracy of 67.98% with a standard deviation of 0.05 percentage points.

Cultural, Political and Demographic Dimensions

It has been well-established that the relation-preserving properties of word embeddings allow for the definition of arbitrary dimensions in terms of their polarities; for example, class, using "rich" and "poor" (Kozlowski, Taddy, and Evans 2019). Previous work has demonstrated that this principle can be applied to community embeddings as well to construct *social dimensions* that accurately represent dimensions of social identity; for example, age, gender, and political affiliation (Waller and Anderson 2020). We apply this method to our embedding by creating cultural, political, and demographic dimensions in our embedding that represent a variety of relevant political concepts. The dimensions we analyze in this study are: ideology (conservative, progressive), religiosity (religious, secular), gun control (less regulation, more regulation), birth control (pro-life, pro-choice), war (hawkish, pacifistic), trade (pro-globalism, anti-globalism), age (older, younger) and gender (masculine,

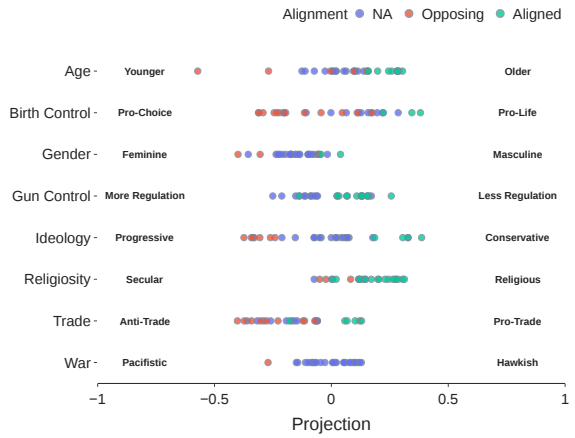

Figure 2: Dimension validation against subreddits with representative words.

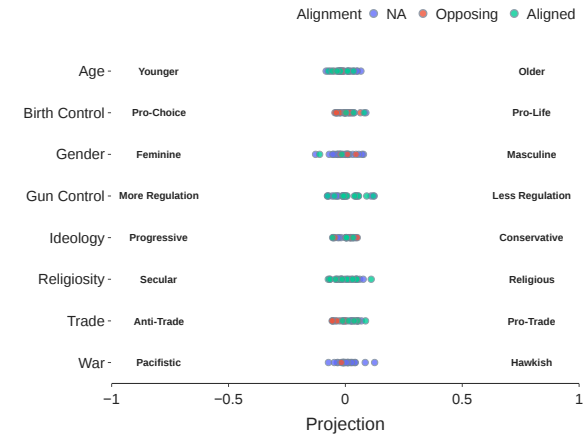

Figure 3: A version of Figure 2 generated using a null hypothesis embedding. In this embedding, the author of each comment was shuffled randomly. Validation subreddits do not vary as much in a null hypothesis embedding, demonstrating that the variation we see in Figure 2 is representative of real patterns in behaviour.

feminine). While far from exhaustive, we select these dimensions as an interesting cross-section from which future work can build upon and as a demonstration of the flexibility of our methodology to pick up on nuanced characteristics at the community level.

We first define pairs of subreddits which differ in their alignment with our chosen political concepts to create various dimensions to evaluate candidate subreddits against. Table 1 shows the definitions for our dimensions. Each subreddit chosen explicitly designates themselves as aligned with or against that definition through their "about" section, wikis, and subreddit specific rules. We then take the arithmetic difference of the vectors for each pair of communities (a, b) to obtain a vector representing the dimension $\vec{d} = \vec{b} - \vec{a}$. Any subreddit c can then be 'scored' on the dimension by computing the cosine similarity of its vector representation $\vec{a}$ with that of the dimension $\vec{d}$. This score is equivalent to the difference between the cosine similarity of a and c and the cosine similarity of b and c. As previously established, the cosine similarity of two subreddits is related to the similarity in their user-bases. Therefore, the score of a subreddit on a dimension is directly related to the association of its user-base with the desired political concept. If a subreddit has a user-base far more similar to a than b, it will receive a highly positive score, while if it has a user-base far more similar to b than a, its score will be highly negative. If a subreddit is approximately equidistant between a and b, it will receive a score close to 0. When constructing the dimensions it is helpful to reduce subreddit specific variance by averaging across multiple pairs of subreddits. In the scenario where there were more subreddits aligned with one end of the dimension relative to another, we repeated the minority subreddits to achieve this reduction in variance; however, this was only done in cases where the semantic relationship of the pair would be preserved.

To validate our dimensions, we first selected a list of representative words and phrases for that dimension. For example, this list for the ideology dimension is progressive, conservative, socialism and libtard, a common pejorative on Reddit. We then collect the thirty subreddits that use those phrases at the highest rate, and hand label them as aligned with, opposing, or not relevant to the dimension. The full list of phrases, the resulting subreddits, and their classifications can be seen on Github.com (XXXXX 2020). We project those subreddits onto the dimension and calculate the mean projection for all classes of subreddit. The projection of these baseline subreddits onto our dimensions are visualized in Figure 2. For all dimensions, there is a strong alignment between our hand labelled classifications and our model's predictions. This alignment is prominently visible when compared to a random baseline embedding, in which the author of each comment was shuffled randomly (Figure 3). Using subreddit projection scores as a prediction for their label 'Opposing' or 'Aligned' (excluding NA subreddits), the area under the receiver operating characteristic curve (ROC AUC) is 0.94, as opposed to 0.6 in the random baseline (using completely random projection scores, we see ROC AUC values ranging from 0.4 to 0.6.) Across the dimensions, the mean difference between projections aligned with, and opposed to, a dimension is 0.509, indicating substantial separation that is congruent with the hand classifications. Thus, our model has a strong basis for analyzing candidate subreddits moving forward.

Results

We now apply our community embedding and our eight political, cultural, and demographic dimensions to candidate subreddits during the 2020 U.S. Democratic presidential primaries. Our embedding makes use of the entire Reddit dataset, positioning each subreddit in relation to every other subreddit, and allows for a logical system of manipulation and analysis through vector algebra. In doing so,

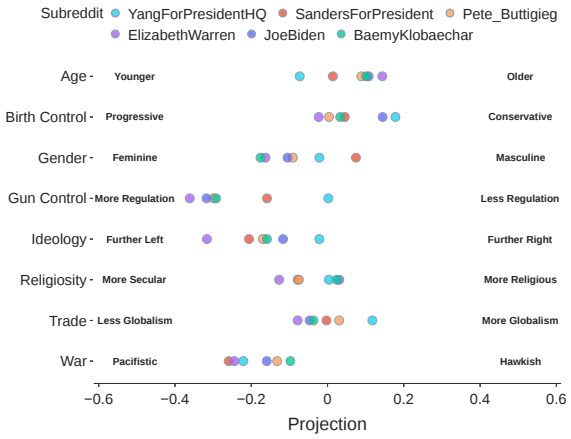

Figure 4: Projection of subreddits on to dimensions.

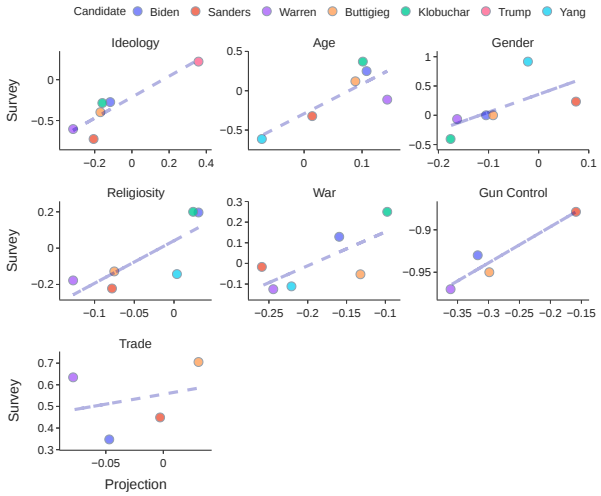

Figure 5: Supporter alignment as a function of Reddit dimensions. Y-axis: normalized likert scale survey responses. X-axis: dimension projection.

we compare our model's performance against various survey data and policy analyses. This will demonstrate where within Penney's taxonomy candidate subreddits may lie.

Dimension Analysis

The score of a subreddit on a dimension is a measure of the relative similarity of its userbase to the political, cultural, and demographic-related subreddits used in its definition. Therefore, by projecting candidate subreddits on these dimensions, the resulting scores can be used as an accurate measure of the alignment of a candidate's online community with each of our chosen political concepts. With our embedding dimensions serving as a proxy for how aligned each grassroots community is along that dimension, our study now turns to methods of comparison to offline politics. We posit that there are two relevant metrics of comparison: the first compares our dimensions against the views of a candidate's supporters. If a large faction of a political figure's supporters identify as progressive, it may be the case that the political campaign's subreddit, whose user base is presumably a subset of those supporters, may have progressive characteristics as well. This would indicate a grassroots community's alignment with the supporters of a candidate. If the intersection between a candidate's Reddit supporters and offline supporters is very small, then the community may be ill-aligned with the campaign, placing the community towards the far-end of Penney's taxonomy. An example of such a community would be the BSDMS Facebook page analyzed by Penney, which represents a subset of Sanders supporters whose norms substantially deviate from the campaign's (Penney 2017). Conversely, the second metric would measure the community's alignment with the candidate and their campaign. This would serve as a measure with how aligned the subreddits are with officially-sanctioned policy and messaging, which within Penney's taxonomy would warrant closer integration. Our study conducts both an supporter and candidate alignment analysis in order to fully parse out the ways in which Reddit acts within the grassroots ecosystem. The projection of each candidate subreddit on to one of the dimensions is visualized in Figure 4.

Comparison One: Candidate Supporter Alignment with Grassroots Communities. Our supporter alignment comparison makes use of two surveys, both sourced from the Pew Research Centre, which first ask the respondent which of the primary candidates they support, and as well as a variety of questions about the respondent's belief. We calculate the respondent dimension score for a dimension d, as the weighted average of responses normalized on the range $[-1, 1]$, where -1 indicates that all respondents answered in one extreme, and 1 indicates that all respondents answered in the other. We can regress the survey responses as a function of the dimension projections and visualize their relationship. This provides a simple metric for how well survey data is reflected in Reddit. These relationships are visualized in Figure 5. Not all survey data included all relevant candidates, in which case they were omitted.

As a whole, our embedding dimensions mirrored nuanced characteristics of a candidate's supporters. We analyze our ideology dimension with respect to a Pew Research Center poll which asked individuals which candidate they supported and whether they identified as "very liberal," "liberal," "moderate" or "conservative" (Pew Research Center 2020). Therefore, a higher score along the y-axis in Figure 5 indicates that a higher proportion of individuals responded with "moderate" or "conservative" — and we would expect a larger alignment with our ideology dimension in the conservative direction. A similar survey was conducted with respect to Donald Trump in 2016; given the availability of this data the subreddit *r/The_Donald* was also included for this dimension (Pew Research Center 2016). We can then regress this survey data on our embedding dimensions as a goodness-of-fit test. There was a significant positive coefficient when regressing our ideology dimension on the ideological preferences of a candidate's supporters ($\beta_1 = 1.284, t(4) = 4.528, p = 0.011$). The same poll was used

	Ideology		Age		Gender		Religiosity		War		Birth Control		Gun Control		Trade	
	Supporter	Candidate	Supporter	Candidate	Supporter	Candidate[1]	Supporter	Candidate	Supporter	Candidate	Supporter	Candidate	Supporter	Candidate	Supporter	Candidate
Constant (β_0)	-0.214**		-0.292*	57.313***	0.358	55.337	0.041		0.317	1.208		-0.8656***	-0.8111**	-0.475	0.557	-0.326
	(-0.067)		(0.134)	(9.939)	(0.213)	(3178.589)	(0.062)		(0.152)	(0.859)		(0.084)	(0.028)	(0.243)	(0.113)	(0.363)
Dimension Projection (β_1)	1.284**		3.776*	70.926	3.044	431.771	2.344*		1.640	7.405		0.999	0.426**	-0.362	0.921	2.683
	(0.284)		(1.390)	(103.026)	(1.810)	(3.01e+04)	(0.890)		(0.778)	(4.404)		(0.868)	(0.096)	(0.907)	(2.337)	(5.684)
R-Squared	0.837		0.649	0.106	0.414	1.000[2]	0.634		0.526	0.414		0.249	0.907	0.038	0.072	0.053
Adj. R-Squared	0.796		0.561	-0.118	0.268		0.543		0.408	0.268		0.061	0.861	-0.202	-0.392	-0.184
No. Observations	6		6	6	6	6	6		6	6		6	4	6	4	6

Standard errors are reported in parentheses.
$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$
[1] Used logistic regression to model candidate gender.
[2] McFadden's pseudo-R squared.

Table 2: OLS Regression results: political, cultural and demographic dimensions. Note that while the number of observations in the regression represents the number of subreddits examined, each subreddit projection value is derived from a far greater number of comments. This confidence in the projection values is not taken into account in this statistical test, reducing its power.

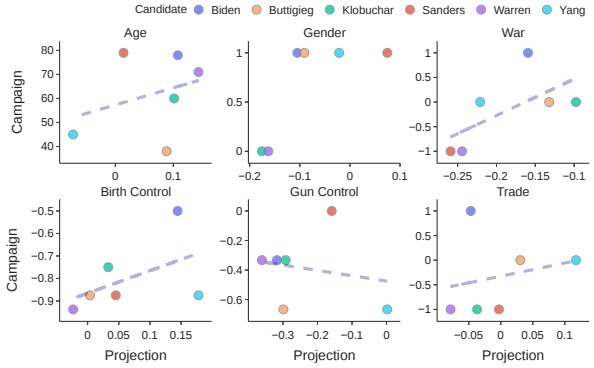

Figure 6: Campaign/candidate alignment as a function of Reddit dimensions. Y-axis: normalized policy index + candidate age/gender. X-axis: dimension projection.

to analyze the age, religiosity, hawkishness, trade, gun control and gender dimensions by asking respondents their age ($\beta_1 = 3.776, t(4) = 2.717, p = 0.053$), religious attendance ($\beta_1 = 2.344, t(4) = 2.634, p = 0.058$), whether they come from a military family ($\beta_1 = 1.640, t(4) = 2.108, p = 0.103$), whether free trade agreements have had a positive impact on them ($\beta_1 = 0.921, t(2) = 0.394, p = 0.732$), their opinion on the current level of gun control ($\beta_1 = 0.426, t(4) = 4.419, p = 0.048$), and their gender ($\beta_1 = 3.044, t(4) = 1.681, p = 0.168$), respectively (Pew Research Center 2020). The full OLS regression results, which summarize the statistical relationships shown in Figure 5, are shown in Table 2. Figure 5 thus, also contains the exact populations used in the regression results.

It should be noted that despite the high alignment of our dimension vectors, and the relative similarity of the candidate vectors, the rank order of candidate subreddits against these dimensions often differs. Additionally, despite Reddit being predominately made up of users age 18-29, our model is able to extrapolate dimensions like age beyond Reddit's user-base relatively well (Pew Research Center 2018).

Comparison Two: Candidate Alignment with Grassroots Communities. Our second comparison technique asks if candidate subreddits are reflective of the political figure themselves and their campaign. To do so, this study compares our dimensions to policy analysis and characteristics of the candidate. For each dimension, the number of polices aligned with and opposing are tallied and normalized between -1 and 1, where a score of 1 indicates all policies aligned with the dimension and a score of -1 indicates all policies opposing that dimension. The full OLS regression results, which summarize the statistical relationships shown in Figure 6, are shown in Table 2. Figure 6 thus, also contains the exact populations used in the regression results.

The hawkish dimension is validated with analysis by Politico on whether each candidate would increase the military budget, and how quickly they would terminate overseas deployments ($\beta_1 = 7.405, t(4) = 1.681, p = 0.168$) (Sudeep Reddy and Bland 2019). Similarly, the gun control dimension is evaluated using Politico analysis on how each candidate would regulate assault rifles, whether they would implement universal background checks on gun purchases, and whether they would create a national gun registry or licensing program ($\beta_1 = -0.362, t(4) = -0.399, p = 0.710$) (Forgey 2019). The birth control dimension makes use of a survey on birth control policies sent out by the New York Times to all candidates; the survey included questions asking whether they would codify 'Roe v. Wade,' make hormonal birth control available without a prescription, and preserve Planned Parenthood funding, among others ($\beta_1 = 0.999, t(4) = 1.152, p = 0.314$) (Astor 2019). The trade dimension is validated against analysis done by the Peterson Institute for International Economics that groups candidates by their opinions of various free-trade agreements, steel and aluminium tariffs, and "stances on China" ($\beta_1 = 2.683, t(4) = 0.472, p = 0.662$) (Hufbauer and Jung 2019). For the age and gender dimension we simply used the age each candidate would be on inauguration ($\beta_1 = 3.776, t(4) = 0.688, p = 0.529$) and the candidates' self-reported gender ($\beta_1 = 3.044, z(4) = 0.014, p = 0.989$) as the candidate alignment measure.

Figure 6 visualizes the regressions when using the candidate's policies or attributes as the metric of alignment. As can be seen, the Reddit dimensions show little relation to the candidate alignment mechanisms. The regression coefficients' large standard errors and small, non-significant R^2 values indicate our embedding dimensions are not predictive of the candidate and policy alignment metrics. Thus, Reddit communities are not reflective of political figures themselves or their campaign's policy points, but their supporters. We

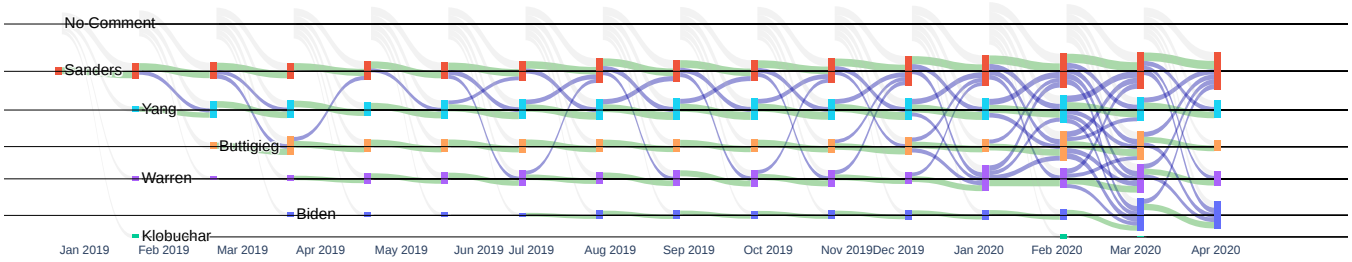

Figure 7: User self-selection into 'home' subreddits over time.

should stress that there is no a priori way of knowing this to be the case, and that this distinction has not been previously studied in this context. It may seem counterintuitive that the policy preferences of a candidate's supporters can act independently of a candidate's policy positions; however, policy preferences are only one consideration that contribute to a voter's candidate choice (Campbell et al. 1980). A multitude of factors including party affiliation, underlying attitudes, and superficial preferences can affect voter preferences independent of candidate policy positions (Campbell et al. 1980; Lewis-Beck et al. 2008). Thus, it is perhaps not surprising that a candidate's supporters, who presumably make up that candidate's subreddit in some proportion, are better reflected in our model than the candidate's policy positions.

Temporal Analysis

Finally, it is important to note that online communities are not static entities. Rather than each candidate subreddit being its own self-contained community, we note significant interplay between on Reddit. Figure 7 shows the flow between subreddits that users choose – with each path representing a user's 'home' subreddit, defined as the community that user commented in the most, for a particular month. As such, we apply our novel temporal community embedding, which is used to analyze candidate subreddits over time.

By including a temporal analysis we are able to analyze the stability of political communities and how they change relative to each other. This is especially pertinent within the literature of public sentiment analysis through social media, which emphasizes the benefits of being able to measure sentiment across time at a more granular level then administering surveys allows.

Subreddit Drift. Kim et al. demonstrate in their temporal word embedding, that by analyzing the cosine similarity of a word's vector representation from an initial start point, t_0, they are able to observe semantic and syntactic drift in language. We extend this to our community embedding model in Figure 8, in which t_0 = March 1, 2019. Earlier months were not included due to a lack of activity on *r/JoeBiden* and *r/BaemyKlobaechar*. As a reference, we also included two very large subreddits which we would expect to remain relatively stable due to their size and active community: *r/comicbooks* and *r/hiphopheads*. While all communities see some amount of churn, regular (*r/fitness* sees a spike of activity at the beginning of 2019) or otherwise, these two have particularly active communities — and thus we expect to see less churn relative to most. As can be seen, Joe Biden's subreddit saw the highest drift over time, presumably a result of having a small initial user base. Bernie Sander's subreddit had a similar stability to our baseline subreddits, which is congruent with past research noting Sander's sustained presence on Reddit (Mills 2018).

Centroid Similarity. An additional point of interest for this study concerns how candidate subreddits change in relation to each other over time. Previous research has found detrimental affects on information distribution as individuals self-select into political bubbles. Given the nature of our community embedding, this sort of self-selection would manifest in candidate subreddits becoming more orthogonal over time. In an effort to quantify self-selection we calculated the centroid of the candidate subreddits for each time step, c_t, and then for each candidate subreddit, s_t, calculate their cosine similarity: $cos(c_t, s_t)$. This is visualized in Figure 8. Given the variety of communities within the Reddit ecosystem it is not surprising that subreddits dedicated to Democratic political figures would have more similarities than differences – but in relative terms, the average centroid similarity is remarkably constant. While each of the candidate subreddit displayed some amount of drift throughout 2019, they all retained high levels of alignment with each other. This provides a way of quantifying the interplay between communities seen in Figure 7.

Discussion

Our results contribute a novel approach to the study of political organization online. While previous work has relied on qualitative, informal judgements of the ideological alignment between a political campaign and a community, we introduce a method to quantitatively measure the ideological association of candidate communities using political dimensions created using reference communities. We demonstrate that the political associations of these subreddits are only loosely associated with their candidate's explicit political views while having a much stronger association with the

views of their supporters. Given that these subreddit associations represent the ideological associations of their members, this suggests that Reddit candidate communities are more akin to self-organized grassroots groups than vehicles for candidate-aligned amplification of campaign messages as defined by Penney (2017). Rather than acting merely as a "distribution network" for the campaign's chosen messages and priorities, the content of these subreddits is produced by supporters with distinct political views and priorities. This demonstrates that a unidirectional view of the relationship between candidates and their online campaign sites is too simplistic; not only are these communities more than simply an "amplifier" for candidate messaging, they contain distinct ideological associations that could potentially exert pressure on the candidates themselves. Even if this distinct ideology does not affect the candidate, it could affect broader public perceptions of them; for example, the "Bernie Bro" stereotype (Penney 2017). However, the relative alignment of communities with the views of their respective candidate's supporters suggests that the level of reputational risk may not be as high as with BSDMS. Regardless, these results firmly demonstrate that online political communities cannot be viewed simply as an extension of a candidate, but as their own novel political entity, a finding with particular significance given that the continual shift towards online campaign methods (exacerbated by the COVID-19 pandemic) suggests that grassroots groups will only increase in prevalence and importance to political organization.

Our temporal analysis demonstrates that these dynamics exhibit a high degree of stability over time. We observe no major changes in between-candidate association, or candidate drift over a one-year time period. These findings suggest that the grassroots collective nature of these political communities is consistent. If candidate subreddits were acting as official amplifiers, one might expect to observe either consistent alignment with official messages, or an increase in alignment over time; we observe neither, instead observing that the distinct character of these communities remains over this time period. However, we do not intend to suggest that this implies that candidate and supporter ideologies never change. Our choice of a relatively short time period (the period surrounding the 2020 Democratic primaries) likely contributes to the stability we observe. Nevertheless, this analysis reinforces our findings on the grassroots nature of candidate subreddits and validates the promise of this method for estimating public ideological associations as an alternative to public opinion polling or informal subjective social media observation that political campaigns often must resort to.

Our work has limitations. Coordinated inauthentic activity (i.e. bot activity) has been observed on many social platforms, especially around elections and political campaigning, and it is unlikely that Reddit is removed from this phenomenon. Substantial amounts of bot activity could distort our estimates of supporters' ideological affiliations. The pseudonymous nature of Reddit also introduces potential complications; it is not possible to be certain whether accounts are not associated with the campaign, meaning official accounts could unintentionally be included in our

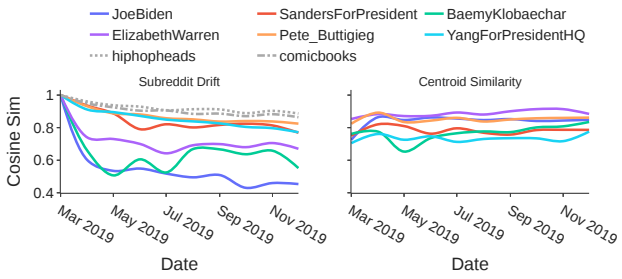

Figure 8: Subreddit drift and centroid similarity.

analysis, and users may create multiple accounts per person, reducing the amount of co-occurrence data used to calculate the embedding and dimension scores. Additionally, in our analysis of subreddit drift we note that candidates like Elizabeth Warren, Joe Biden and Amy Klobuchar saw significant variance in their embedding representation. However, it is unclear whether this is a function of user churn, or changes in preferences in the existing memberbase (Danescu-Niculescu-Mizil et al. 2013). While this paper focuses on the community as the unit of analysis, it is important to note that subreddits are by no means monoliths. Thus, this study provides the basis for more granular, intra-community research that extends our behavioural approach. Finally, we are missing complete data for 2020, an omission required due to technical restraints.

Our work opens up multiple other opportunities for future extensions. While we examine eight important issue- and ideology-based dimensions here, our method could be applied to any number of alternate dimensions. For example, future studies could investigate LGBT identity, more complex notions of left- and right- wing ideology, or distinct religious associations. A more complete collection of dimensions could provide a more complete analysis of supporter ideology. While we only examine one year's worth of data in our temporal analysis, there is no inherent limit to the time frame of such an analysis. As previously mentioned, this method acts as a promising alternative to public opinion polling in the context of online supporters. Temporal analysis of activity along issue- and ideology-based dimensions can give an accurate estimate of how supporter views change over time with a far lower cost and no possibility for self-reporting bias. Future work could apply this method to related studies on grassroots political organization. For example, a similar analysis could be performed for the 2016 Republican primaries to observe whether the ideological affiliation of Donald Trump supporters changed during the primary process. The coarse taxonomy presented by Penney and others could be expanded using our computational method. Application of our method to other political communities on Reddit could reveal alternate forms of organization somewhere between the grassroots and official campaign messaging. Finally, the large amount of linguistic data available in this dataset (i.e. the textual contents of comments) could be incorporated into this analysis in a hybrid approach.

Conclusion

Through a behavioural approach, this study has demonstrated a flexible methodology for the analysis of online political communities. When applied to the 2020 U.S. presidential primaries, we are able to use the community as a unit of analysis to measure candidate-focused communities along a variety of political, cultural and demographic axes. This approach is especially useful within the study of grassroots movements and public sentiment analysis. As noted earlier, political campaigns face a tension with respect to third-party actors: balancing the benefits of added energy and novel political content at the potential expense of a loss of campaign cohesion. Our embedding approach applied to Reddit's candidate-focused communities shows alignment with a candidate's supporters, but little alignment with the campaign's policies or the candidate's personal attributes. Within Penney's framing, this poses more reputational risk than campaign controlled platforms, but potentially less risk than communities with more problematic norms (Penney 2017). Within the realm of public sentiment analysis, we measure change in these grassroots communities along these relevant dimensions, but also change among the communities themselves. This structural approach does not rely on an understanding of language, or the varying norms of each community, but instead harnesses aggregate Reddit activity to position various communities in relation to each other.

It is clear that online mediums have changed the way politics are conducted, especially with respect to grassroots organizations. Our study interrogates grassroots communities in relation to the political campaigns they aim to support along important dimensions. Through this work we provide insights into an interesting and relevant microcosm of political life.

References

Abbott, J. 2012. Democracy@ internet. org revisited: analysing the socio-political impact of the internet and new social media in east asia. *Third World Quarterly* 33(2):333–357.

ActBlue. 2020a. Reddit for joe actblue donation portal. [Online; accessed 25-August-2020].

ActBlue. 2020b. Reddit for warren actblue donation portal. [Online; accessed 25-August-2020].

Astor, M. 2019. How the 2020 democrats responded to an abortion survey. *The New York Times*.

Bakshy, E.; Messing, S.; and Adamic, L. A. 2015. Exposure to ideologically diverse news and opinion on facebook. *Science* 348(6239):1130–1132.

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The pushshift reddit dataset.

Bennett, W. L., and Segerberg, A. 2013. *The logic of connective action: Digital media and the personalization of contentious politics*. Cambridge University Press.

Bennett, W. L. 2012. The personalization of politics: Political identity, social media, and changing patterns of participation. *The annals of the American academy of political and social science* 644(1):20–39.

Bossetta, M. 2018. The digital architectures of social media: Comparing political campaigning on facebook, twitter, instagram, and snapchat in the 2016 us election. *Journalism & mass communication quarterly* 95(2):471–496.

Campbell, A.; Converse, P. E.; Miller, W. E.; and Stokes, D. E. 1980. *The american voter*. University of Chicago Press.

Chadwick, A., and Stromer-Galley, J. 2016. Digital media, power, and democracy in parties and election campaigns: Party decline or party renewal?

Chadwick, A. 2007. Digital network repertoires and organizational hybridity. *Political Communication* 24(3):283–301.

Danescu-Niculescu-Mizil, C.; West, R.; Jurafsky, D.; Leskovec, J.; and Potts, C. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web*, 307–318.

Dommett, K., and Temple, L. 2018. Digital campaigning: The rise of facebook and satellite campaigns. *Parliamentary Affairs* 71(suppl_1):189–202.

Edwards, A. 2006. Ict strategies of democratic intermediaries: A view on the political system in the digital age. *Information Polity* 11(2):163–176.

Enli, G. S., and Skogerbø, E. 2013. Personalized campaigns in party-centred politics: Twitter and facebook as arenas for political communication. *Information, communication & society* 16(5):757–774.

Forgey, Q. 2019. Candidates views on the issues: Gun control. *Politico*.

Gibson, R. K., and Ward, S. 2012. Political organizations and campaigning online. *The SAGE handbook of political communication* 62–74.

Gibson, R. K. 2015. Party change, social media and the rise of 'citizen-initiated'campaigning. *Party politics* 21(2):183–197.

Hufbauer, G., and Jung, E. 2019. Where do democratic presidential candidates stand on trade? *The Peterson Institute for International Economics*.

Kim, Y.; Chiu, Y.-I.; Hanaki, K.; Hegde, D.; and Petrov, S. 2014. Temporal analysis of language through neural language models. *arXiv preprint arXiv:1405.3515*.

Kozlowski, A. C.; Taddy, M.; and Evans, J. A. 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review* 84(5):905–949.

Lewis-Beck, M. S.; Norpoth, H.; Jacoby, W.; and Weisberg, H. F. 2008. *The American voter revisited*. University of Michigan Press.

Margolis, M.; Resnick, D.; and Levy, J. 2003. Major parties dominate, minor parties struggle. In *Political parties and the Internet: Net gain?* Routledge. 53–69.

McGregor, S. C. 2020. "taking the temperature of the room" how political campaigns use social media to understand and represent public opinion. *Public Opinion Quarterly* 84(S1):236–256.

Mills, R. A. 2018. Pop-up political advocacy communities on reddit. com: Sandersforpresident and the donald. *AI & SOCIETY* 33(1):39–54.

Penney, J. 2017. Social media and citizen participation in "official" and "unofficial" electoral promotion: A structural analysis of the 2016 bernie sanders digital campaign. *Journal of communication* 67(3):402–423.

Pew Research Center. 2016. Voters' perceptions of the candidates: Traits, ideology and impact on issues. Online; accessed 07 September 2020.

Pew Research Center. 2018. News use across social media platforms 2018. Online; accessed 08 September 2020.

Pew Research Center. 2020. As voting begins, democrats are upbeat about the 2020 field, divided in their preferences. Online; accessed 08 September 2020.

Rajadesingan, A.; Resnick, P.; and Budak, C. 2020. Quick, community-specific learning: How distinctive toxicity norms are maintained in political subreddits. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 557–568.

Reddit. 2020a. Grassroots subreddit for 2020 democratic candidate for president, andrew yang. [Online; accessed 25-August-2020].

Reddit. 2020b. r/pete_buttigieg 2019 year in review. [Online; accessed 25-August-2020].

Reddit. 2020c. Sandersforpresident has raised $1 million! [Online; accessed 25-August-2020].

Scott, J. K., and Johnson, T. G. 2005. Bowling alone but online together: Social capital in e-communities. *Community Development* 36(1):9–27.

Soliman, A.; Hafer, J.; and Lemmerich, F. 2019. A characterization of political communities on reddit. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*, 259–263.

Stier, S.; Bleier, A.; Lietz, H.; and Strohmaier, M. 2018. Election campaigning on social media: Politicians, audiences, and the mediation of political communication on facebook and twitter. *Political communication* 35(1):50–74.

Sudeep Reddy, L. N., and Bland, S. 2019. Candidates views on the issues: Military. *Politico*.

Van de Mortel, T. F., et al. 2008. Faking it: social desirability response bias in self-report research. *Australian Journal of Advanced Nursing, The* 25(4):40.

Waller, I., and Anderson, A. 2019. Generalists and specialists: Using community embeddings to quantify activity diversity in online platforms. In *The World Wide Web Conference*, 1954–1964.

Waller, I., and Anderson, A. 2020. Community embeddings reveal large-scale cultural organization of online platforms. *arXiv preprint arXiv:2010.00590*.

XXXXX. 2020. Github repository. https://github.com/XXXXXX.