Artificial Communities and Bait and Switch Messaging of Russian Trolls on Twitter during the 2016 Election

Alexandra Cirone* William Hobbs†

July 11, 2019

Abstract

This paper studies network dynamics and coordinated messaging across Russian troll Twitter accounts during the 2016 US Presidential Election. Using Russian Internet Research Agency social media account data released by Twitter in late 2018, we apply a recently developed text scaling method, pivoted text scaling, to construct a simple timeline of text-based strategies used over the course of the campaign. We have three key results, in addition to replications of prior work using automated methods and publicly available data. First, we identify discontinuities in the use of network clusters of trolls, suggesting a directed strategy to amplify in-group messages and to build artificial communities for politicized content. Second, we show that from 2015 to 2016 the trolls switched from accounts focused on non-political content to political content. Third, we show a reversal on the left close to election time: after constructing a political community, the trolls imitating the Black Lives Matter activists increased seemingly *non*-political content relative to political content, suggesting efforts to increase distraction and reduce mobilization. Overall, the findings suggest that state-sponsored trolls' strategies to influence public opinion in 2016 went beyond simply divisive messages – they also used non-political content to selectively manipulate levels of attention to politics.

*Department of Government, Cornell University. E-mail: aec287@cornell.edu.

†Departments of Human Development and Government, Cornell University. E-mail: hobbs@cornell.edu.

The rapid rise of social media has highlighted the ability of authoritarian governments to influence public opinion – both domestically and abroad. These efforts came to prominence with Russian interference in the US Presidential Election in 2016 (Mueller 2019). Since that time, in late 2018, Twitter's Elections Integrity Initiative released a public data set of more than 10 million tweets sent by approximately 3,841 accounts affiliated with the Internet Research Agency (IRA), a Kremlin-based Russian troll farm. These accounts represent the efforts of human-controlled Russian operators, or "trolls," as opposed to computer-controlled accounts (or "bots"), to influence the 2016 U.S. presidential campaign, and provide a window into the strategic manipulation of content on social media. Academic studies have now established that IRA accounts overwhelmingly supported the Trump campaign over the Clinton campaign, and posted content to amplify existing social divisions between liberals and conservatives.

However, it is becoming increasingly clear that Russian IRA accounts adopted more subtle and sophisticated strategies. In this paper, we use a combination of text scaling and network analysis to trace the development of Russian foreign influence efforts on Twitter from 2015 to 2016. The results suggest that activity occurred in three stages: 1) contact and recruitment, 2) politicization, 3) selective mobilization and de-mobilization. The first and second of these have been shown or conjectured in prior studies. Our methods here enable us to identify selective de-mobilization among Black Lives Matter activists close to the election, and, to a lesser extent, selective mobilization of conservative leaning moderates.

Importantly, we argue these de-mobilization activities appear to resemble certain techniques used by authoritarian regimes to subvert collective action among their own citizenry. Demobilization is a strategy that involves messaging that discourages or deflects collective action (Østbø 2017). It could involve posting to actively suppress voter turnout or protest coordination, or "flooding" social media with social or innocuous content, in order to distract (Roberts 2018; Sanovich, Stukal and Tucker 2018). This technique is an increasingly

important tool in an autocrat's playbook, and is a strategy that Russia has invoked before to discourage domestic coordination among its citizens.[1] It has also been used in China, where government-sponsored troll factories filter collective action and flood social media with apolitical content (King, Pan and Roberts 2017*a*). While studies have documented apolitical content from the Russian accounts, none have tied this explicitly to authoritarian de-mbolization of a foreign population, as in the case of the US election.

The paper will proceed by replicating and unifying prior work on troll networks and their messaging, and then will focus on a) the gradual politicization of the conservative community and b) the creation and selective de-mobilization of a decoy Black Lives Matter activist community.

Data and Methods

Studies conducted on this public data and related non-public data have already classified the IRA accounts using account information and tweeted content. IRA accounts constructed "sock puppet" identities that mimicked legitimate users (Howard et al. 2018), and trolls often adopted either liberal or conservative identities; in all, studies have offered several typologies of the different accounts (Linvill and Warren 2018; Yin et al. 2018). Qualitative and quantitative coding of the content tweeted has further measured the extent to which these fake accounts were spreading misinformation compared to established and legitimate news sources (Linvill et al. 2019; Bastos and Farkas 2019; Yin et al. 2018; Broniatowski et al. 2018; Schafer 2018).

[1]Historically, Russia has long used disinformation as a strategy of political manipulation, but its organized factory efforts using trolls and bots began in 2012 to support Medvedev's presidency, and continued to influence domestic public opinion during the political crises over Ukraine in 2014 (Sanovich, Stukal and Tucker 2018; Sanovich 2017).

The Twitter IRA data set allows for studying somewhat more inclusive and dynamic text-based strategies – over time, across IRA account networks, and covering both political and non-political content. However, this data consists of millions of tweets and thousands of accounts; systematic classification and coding of troll activity represents a significant challenge. Scholars have relied on unrestricted open coding by experts (Linvill and Warren 2018; Bastos and Farkas 2019; Llewellyn et al. 2018), which is manually intensive and potentially subjective. Other scholars use automated machine learning methods, such as label propagation, sentiment analysis, or Latent Dirichlet Allocation (LDA) topic models, to classify tweet content but rely on subsets of accounts or snapshots from a particular point in time (e.g. Jensen 2018; Badawy et al. 2018; Zannettou et al. 2018; Dawson and Innes 2019).

Here, we use a recently developed text scaling method, pivoted text scaling (Hobbs 2017), in combination with network community detection, to score messages by trolls over the course of the campaign and to study messaging within the artificial communities constructed by trolls. This text scaling method is closely related to existing unsupervised scaling methods used in political science (Slapin and Proksch 2008; Bond and Messing 2015),[2] but produces a multi-dimensional scaling that estimates the high-level "gist" of a message, and the extent that it conforms to the most commonly used messages on a latent ideological dimension.[3]

[2]These techniques estimate low-dimensional structure in covariance matrices. Pivoted text scaling is intended for particularly high-dimensional covariance matrices, where a large number of variables are unreliable and complicate estimation of population covariance from sample covariance matrices (Johnstone and Paul 2018). In essence, it compresses variance in only words that tend to appear more often than their accompanying words, and then maps the remaining, rare words to the same low-dimensional space. Because a very small number of repeated words account for a large fraction of the reliable variation in text, and because the pivoted text scaling method specifically compresses variation in these common words, we can summarize variation in large volumes of content with a relatively small number of keywords.

[3]This high-level conformity is especially useful when analyzing a social network such as Twitter because, in a sense, it estimates the extent to which an individual belongs to a large group that uses a coordinated

Account Classification: network community detection

Similar to past work (Stewart, Arif and Starbird 2018), we use network community detection on the troll retweets and mentions to identify clusters of social activity. We replicate these analysis so that we can analyze messaging strategies specific to either the left or right-leaning filter bubbles created by the trolls – we expect to have very limited overlap across communities, which would justify cluster-by-cluster analyses. We include details for the process in the appendix.

Labeling identified communities: user profile descriptions

Most of our analyses below will focus on text scaling of tweet content. However, we also scale the text of the account descriptions alone so that we can more easily label the clusters identified in the network community detection.

In the account descriptions, we identify outliers of around 50 accounts tweeting local news, especially in 2015. Excluding outlier local news accounts from the subsequent text scaling focuses our scaling on ideological polarity. These accounts correspond to local news accounts identified in prior work (Yin et al. 2018). They can be identified using keywords that are variants of news – "local news", "breaking news", "top news" – often in combination with a specific locale (e.g. "San Jose's breaking news").

In the scaling of the account descriptions, the first dimension of that text scaling estimates conservatism, while the second estimates liberalism. There is very little overlap in vocabulary. We show this in Figure 5 in the appendix. Combining these two dimensions gives us a single left-right dimension.

---

vocabulary.

Troll Messaging: text scaling on tweets across and within clusters

To analyze Russian trolls' relevant activity during the 2016 election, we analyze tweets posted by the accounts before or on the election on November 8, 2016, and on or after June 1, 2016. We further restrict analyses to tweets coming from accounts using English.[4]

We run the text scaling on two sets of tweets: 1) all June 2015 to November 2016 tweets, and 2) June 2015 to November 2016 tweets within the liberal and conservative clusters. Using the output of this scaling, we identify three theoretically relevant latent variables: 1) ideological imitation, or the partisan dimension, 2) politics vs news, or the politicization dimension, and 3) social de-mobilization. Not all of the keywords are included in the text below, so the appendix provides full details on the dimension numbers and keywords for these dimensions.

In the overall text scaling, dimension 2 measures conservatism, while dimension 3 measures liberalism. We combine these dimensions to create a single partisan dimension. The news (crime and terror attacks) vs. politics dimension is the second dimension of the conservative cluster's text scaling, but also includes some political content (e.g. "teaparty"). Because of this, we also report in the appendix the same result for a combination of the 2nd and 3rd dimensions, where the poles of the combined news vs. politics dimension distinguish crime and terror attacks from the 2016 election.

The top dimension of the liberal cluster's text scaling appears to measure entertainment and de-mobilization, but could also measure informal conversation (e.g. "y'all", "shit", "lol") vs. retweeting of news. Because of this, we report the same substantive finding for a combination of the 2nd and 3rd dimensions, which distinguishes entertainment from Black Lives Matter mobilization, in the main text and the first dimension in the appendix.

[4]We identify English language accounts using Twitter's account language information, and also by using account descriptions, where descriptions using non-emoji Unicode in the supplementary multilingual plane are counted as non-English.

Results: Filter Bubbles

We first document that much of the Russian activity occurred within distinct filter bubbles. Journalistic investigations into the operations of the IRA Troll Farm, as well as the Mueller Report, suggest that Russian operators created politically neutral accounts to gain credibility, and cooperated with each other in teams to amplify messages.[5] Academic studies have also documented network effects among IRA Twitter accounts, and have found high levels of clustering. Howard et al. (2018) find that of 3,841 accounts in the IRA dataset, 2,648 are connected to at least one other IRA account. Further, accounts work in teams to mention each other, forming unique communities of interaction; both partisan and non-political clusters. Dawson and Innes (2019) identify 119 separate clusters on a wide range of topics. Other research focuses on communities centered on Black Lives Matter. Stewart, Arif and Starbird (2018) found activity was primarily located within cluster (as opposed to across the clusters), retweeting and mentioning accounts within the same group to appear authentic.

In order to compare our analysis to existing work, we replicate these results using network community detection here. We visualize this network in the right panel Figure 1. Colors correspond to the identified communities. The left panel corresponds to left-right assignments from the text scaling method, and we use the keywords from the account description scaling to label the clusters left or right. The left panel of Figure 1 shows the correspondence between these scores and their network clusters after we separate the scores at their mean. In this panel, green is assigned to accounts using only symbols or using vocabulary without any overlap in the sample.[6]

Similar to past work, we find very high clustering (i.e. there is little communication

[5]For example, see Radio Free Europe's investigation, *One professional Russian troll tells all* (2015).

[6]Note that this figure does not show accounts that were relatively inactive, and there are additional communities not connected to the liberal and conservative clusters (and that were inactive in 2016). These are also coded "green" in later analysis.

across clusters), and so use these clusters as groups in our later text analyses. Figure 1 shows clear clustering of IRA accounts into partisan filter bubbles. Clustering is even more apparent in the 2016 activity (Figure 4).

Results: Recruitment and Politicization

One puzzling finding from the Twitter IRA data is that much of the content posted by Russian trolls was seemingly apolitical. For example, Linvill et al. (2019) find that 52.5% of tweets in their data set were "camouflage" tweets with no clear connection to an IRA agenda. Schafer (2018) document that some trolls posted purely social content, such as recipes and celebrity gossip. Further, despite popular narrative about trolls and misinformation, only 6 percent of all URLs shared by IRA accounts led to junk news,[7] and almost half of all IRA-run accounts only spread reliable information (typically local news).

One plausible hypothesis for the content is that the high proportion of legitimate and non-political content served to attract followers and gain legitimacy. This form of narrative switching has been conjectured by studies before (Tucker 2018; Dawson and Innes 2019; Linvill et al. 2019), but scholars have only just began to study the tactical choices made by IRA account operators.

We construct a timeline of text-based strategies used over the course of the campaign to demonstrate the strategic use of non-political content, and to assess the timing and direction of both coordinated amplification strategies and "bait-and-switch" messaging. With the network clusters identified above, we combine the IRA clusters into two main categories: polarized accounts (either liberal or conservative) and ambiguous accounts (no clear ideological messaging), in addition to the local news accounts that were not part of the co-mention network because they tweeted links instead.

[7]However, that is twice the rate of a comparison, non-IRA group of Twitter users.

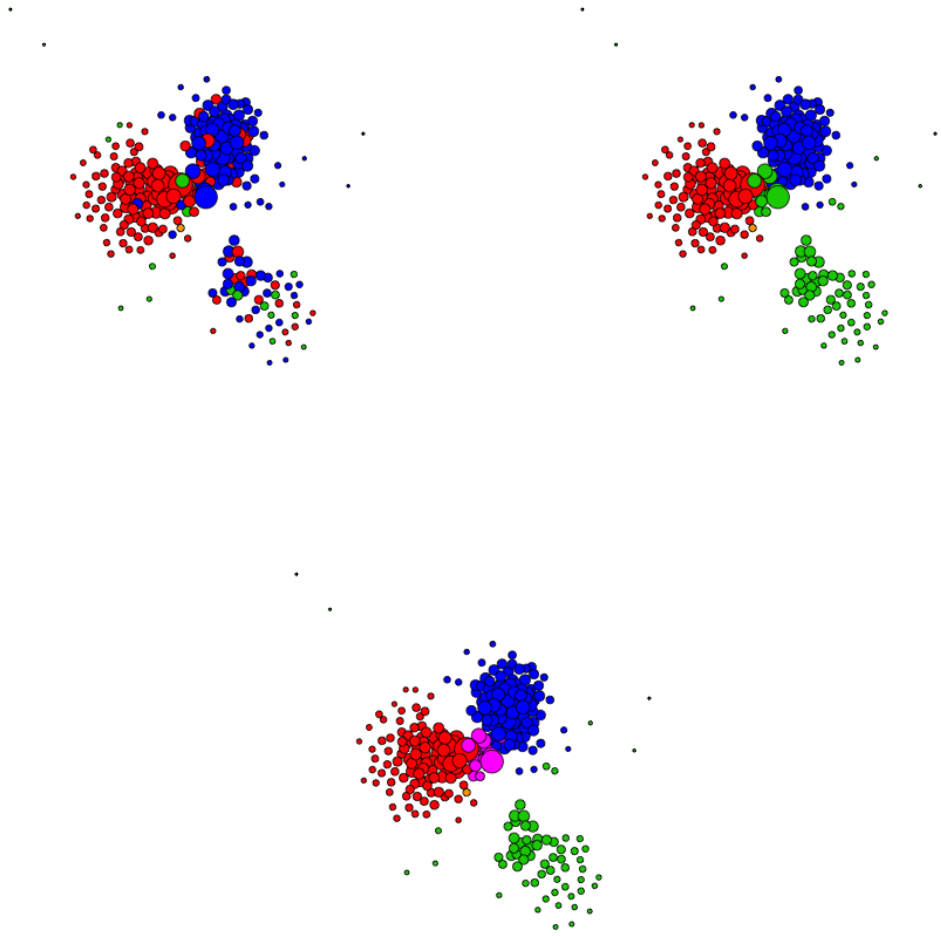

Figure 1: Accounts were classified by network cluster (right). Text scaling of the account descriptions provided the labels for the clusters (left). News accounts were identified by account descriptions alone. Less active accounts are omitted from this visualization. Figure 4 in the appendix shows the clusters for 2016 activity alone. Consistent with prior work, these clusters can be easily separated using community detection algorithms. We rely on the extremely limited overlap in clusters, especially in 2016, to justify the cluster-specific text analyses below, in addition to the combined text analysis, since we expect vocabulary to be distinct across clusters as well. The magenta colored nodes in the final plot were feeder/recruitment accounts that mentioned troll and non-troll accounts many times in 2015. They were the 15 most active accounts in the network in 2015, and account for the large number of mentions in Figure 2.

Figure 2 plots the activity of these accounts from June 2015. The y-axis in this figure is the number of tweets posted by each cluster. Over time, we see a reliance on local news and ambiguous accounts until fall of 2016, at which point there is a significant increase in activity of polarized accounts. The lower panel of this figure shows that ambiguous accounts mentioned non-trolls at extreme rates in 2015, suggesting a massive effort to contact and/or recruit Americans to follow the troll accounts. The most active of these recruitment accounts are visible in Figure 1 where they are the large green nodes situated between the red and blue clusters (colored magenta in the bottom panel of that figure).

Figure 3 similarly documents a shift in attention to non-troll accounts. In other words, this figure shows trends in amplification by different clusters of trolls, where amplification is the retweeting or mentioning of non-trolls. In 2015, less than 10% of amplification is coming from the polarized trolls accounts and 90% of amplification is coming from polarized trolls just before the election. Between these time points, there are several *abrupt* increases in the proportion of amplified messages coming from these accounts, suggesting a coordinated change in strategy.

Past studies have generally concluded that IRA troll accounts posted propaganda designed to divide, incite, and agitate viewers on both side of the political spectrum (Bastos and Farkas 2019). For example, this explains the sharp increases in the tweeting of conservative content in September 2016 (Tucker 2018), under the assumption that trolls' main goal was to support the Trump campaign. The red line in Table 1 documents this pattern in our data. The green line adds a further finding: that ambiguously political accounts suddenly shifted content to mobilize in support of Donald Trump.

Figures 7 and 9 analyze the news to politics shift on the right within the conservative cluster alone. While much of the shift from news to politics occurred across account within the same cluster, bait-and-switch changes in activity in summer 2015 also occurred within the same accounts.

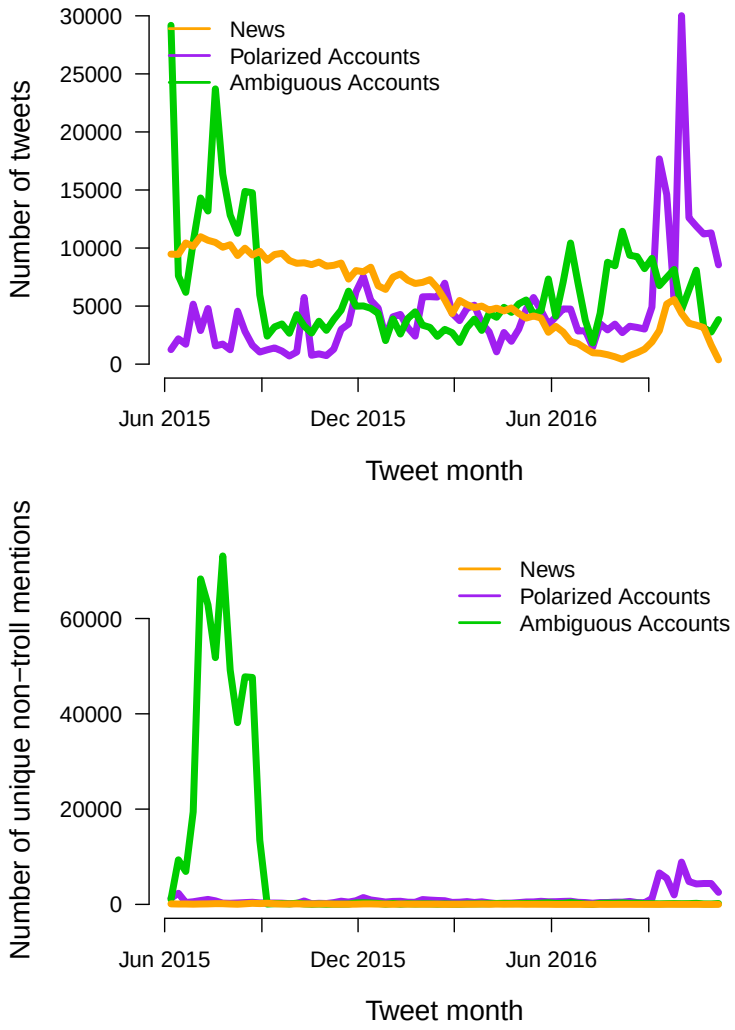

Figure 2: *Changes in use of different types of troll accounts.* This figure shows that accounts focused on tweeting local news gradually reduced their activity from 2015 into 2016, while accounts using polarized, partisan identities increased activity near the beginning of 2016 and dramatically increased activity close to the 2016 election. Ambiguous accounts, accounts not linked to the left-right filter bubbles, became less active at the same time that filter bubbles increased their activity in late 2015. The vast majority of troll interactions occurred during 2015 when news and ambiguously political accounts were most active. Around 500k non-trolls were mentioned by the troll accounts during 2015, including many messages with "follow me" requests, and 40k were mentioned in 2016. Most of the mention tweets in 2015 contained many mentions per tweet. The most active of these recruitment accounts are visible in Figure 1 where they are the large green nodes situated between the red and blue clusters.

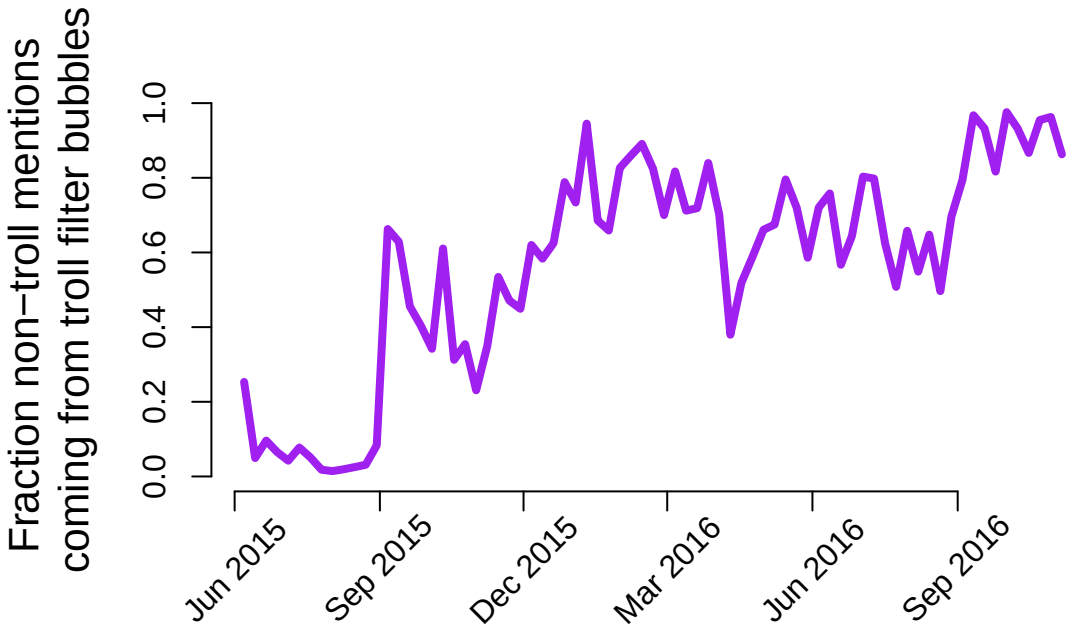

Figure 3: *Shifting amplification of non-trolls.* This figure shows trends in amplification by different clusters of trolls, where amplification is the retweeting or mentioning of non-trolls. In 2015, less than 10% of amplification is coming from the polarized trolls accounts and 90% of amplification is coming from polarized trolls just before the election.

Results: Selective De-mobilization

We have shown that Russian IRA Twitter accounts utilized non-political messaging in early stages of their efforts, and then activated political, right-leaning content right before the 2016 election. In this section, we consider a second purpose for seemingly apolitical content: selective de-mobilization.

To our knowledge, the idea that Russia would adopt a distraction and demobilization strategy on Twitter has not been systematically analyzed. Others have noted attempts to keep African-Americans from voting (MacFarquhar 2018) and more generally voter suppression tweets (Kim 2018). Only one study (Howard et al. 2018) documents that while conservative voters were targeted with Facebook ads encouraging them to actively support the Trump campaign, other segments of the population were targeted with ads against electoral

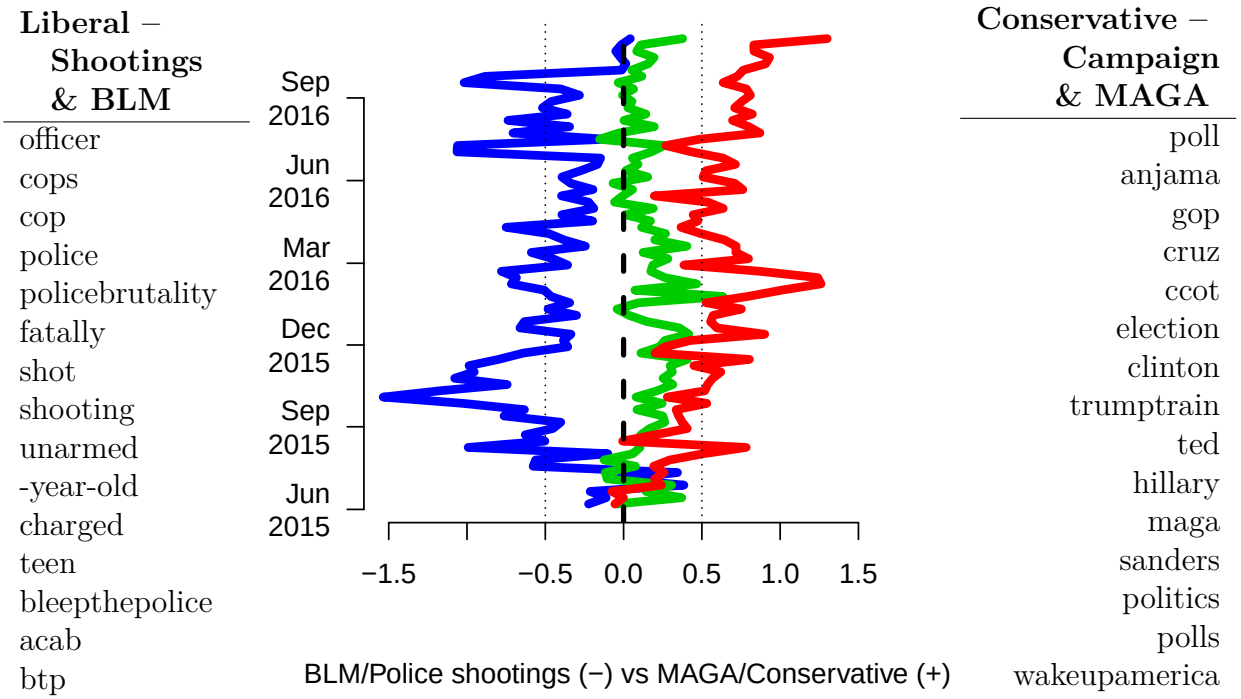

Table 1: *Conservative Mobilization.* This figure shows the average text scores on the partisan dimension of the text scaling. Accounts imitating Black Lives Matter activists tweeted less content about the movement near the end of the campaign 2016, and accounts imitating moderates and conservatives tweeted more right-leaning content in the last week of the election. Note that the mobilization of conservatives and moderates was larger within accounts – that is, the accounts that were not previously political contributed most to the spike (see appendix).

participation.

Below, we consider evidence for a demobilization strategy in the 2016 Election, focusing on the #BlackLivesMatter campaign. This was a salient political issue during the election, originating from the acquittal of American policeman George Zimmerman in the unprovoked death of teenage boy Trayvon Martin, and evolving to represent the systematic targeting of African-Americans by police officers. The BLM movement became associated with the left, and in response a counter-narrative (#AllLivesMatter) that also defended the police

(#BlueLives Matter) emerged on social media. The IRA attempted to capitalize on these racial and partisan divides by posting BLM content on Twitter. Facebook, Youtube, and Instagram, among others (Howard et al. 2018).

The blue line in Table 1, however, demonstrated that while these accounts were actively tweeting about BLM content in the months before the election, they were less likely to tweet such content near the end of the campaign. Table 2 explores this in further detail by examining BLM IRA accounts only. Here, it's clear these accounts switched to using apolitical content (talking about music, videos, the weather) instead of divisive content (such as the police, racism, or "stopthegop"). Linvill et al. (2019) have already shown that left-leaning trolls were more likely tweet apolitical ("camouflage") content, but don't discuss why. Here, we show that this activity is consistent with strategic demobilization.

Our findings from the 2016 election are supported by other studies of authoritarian regimes. For example, Østbø (2017) details that Russian agents used social media to "de-mobilize" the population in 2014, after the Crimea annexation. Pro-regime social media accounts deliberately switched from aggressive hate speech against the opposition to posting sad and empathetic content, in order to promote Kremlin events designed to distract the population. Similarly, China is notable for hiring thousands of government employees to fabricate approximately 448 million social media posts a year while pretending to be Chinese citizens (King, Pan and Roberts 2017*b*), and their censorship apparatus appears to filter collective action content rather than direct criticism of the government (King, Pan and Roberts 2013). Fabricated posts attempt to shift online discussions away from controversial issues by posting Chinese history and inspirational quotes. These show autocrats have used selective mobilization and de-mobilization on their own populations, however, we provide evidence from the 2016 that this strategy could be utilized to influence public opinion internationally.

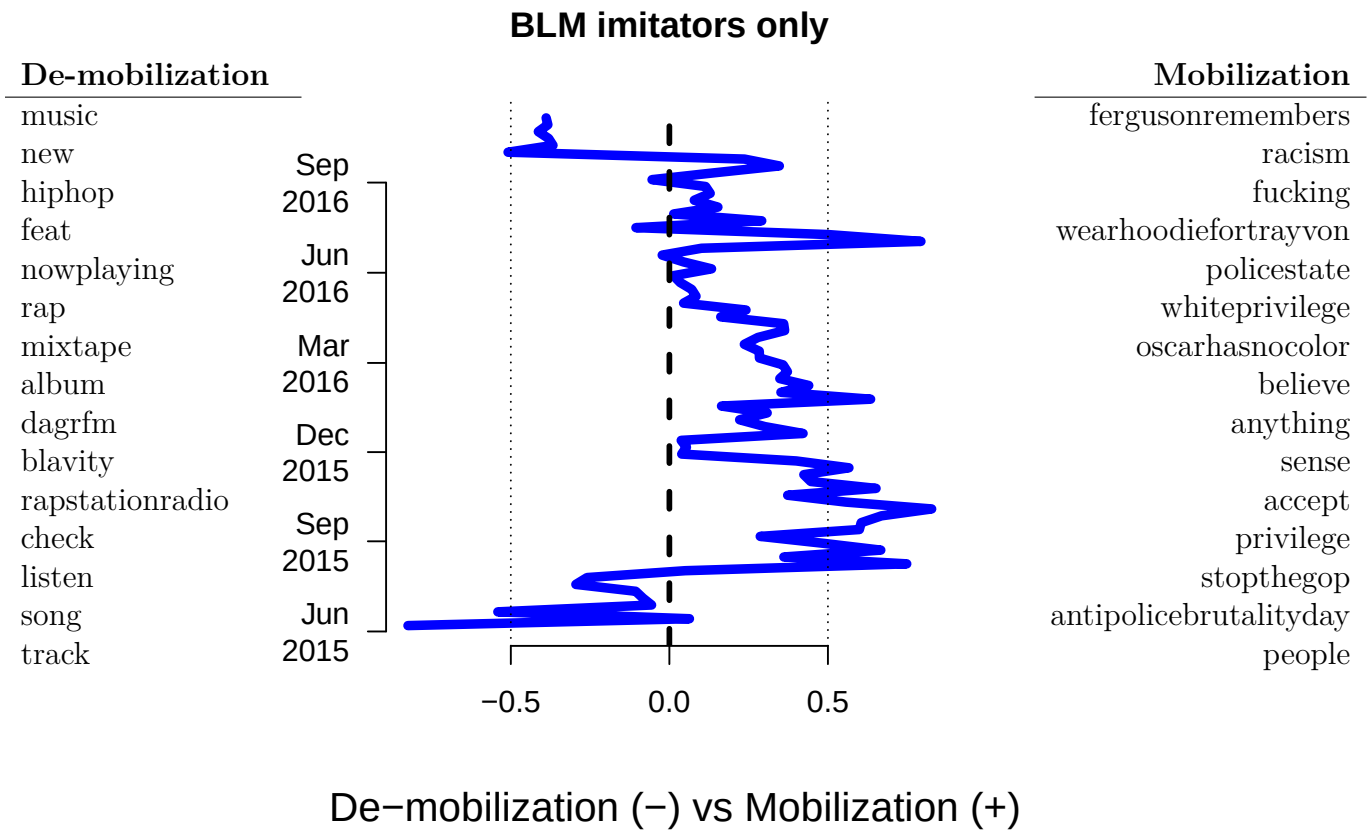

Table 2: *BLM Distraction.* Figure 1 showed a decrease in Black Lives Matter activity near the 2016 election. Given that the dimension opposite BLM in that dimension was focused on the election, it was possible that the accounts increased political attention. Here, we show that the BLM imitators instead increased content related to entertainment. The text scaling dimensions in this figure use data from the "blue" liberal cluster of accounts only (visible in Figure 1).

Conclusion

Past research has characterized the identities and tweets of IRA trolls using a wide range of methods and hand coding. These methods reflect significant expertise, but often rely on a small number of example posts and accounts, or exclude seemingly apolitical content from consideration. Here, in addition to replicating past findings, we provide two main contributions: 1) we show that the Russian Internet Research Agency attempted to influence the 2016 election through activity that was not clearly divisive or even political; and 2) we provide an analysis that can be easily replicated using publicly available data and to-be-released R code (i.e. no account-by-account hand coding or reliance on anecdote).

Specifically, we used automated network analysis and text scaling to replicate several existing findings, and then provide a timeline of messaging content within left and right-leaning troll communities from 2015 to 2016. In this, we consider messages on news and politics, as well as entertainment and other seemingly apolitical content. In the replications, we show trolls interacted within tight clusters of accounts and each account was largely consistent in its messaging, although trolls overall switched from non-political, local news content to political and divisive content over the course of the campaign.

We then proceed to the message timeline within clusters. We find that while right-leaning and moderate trolls mobilized followers in support of Donald Trump, left-leaning trolls used distraction and apolitical messaging to demobilize liberal constituents close to the election. This switch from political to distracting messaging in the Black Lives Matter community was a notable exception to the otherwise consistent messaging with accounts. In contrast with past work, our results suggest that direct efforts to demobilize, such as mentions of difficulty voting or opposition to Hillary Clinton, might have been secondary to indirect efforts to distract and dissolve decoy activist networks.

The accounts represented in the Elections Integrity data set have been deleted, but more

will inevitably take their place. Going forward, our results demonstrate the need for scholars and policymakers to not only focus not only on active, divisive messaging, but to consider the broader array of techniques in the autocrat's toolkit.

References

Badawy, Adam, Aseel Addawood, Kristina Lerman and Emilio Ferrara. 2018. "Characterizing the 2016 Russian IRA Influence Campaign." *CoRR* abs/1812.01997.

Bastos, M. T. and M.J. Farkas. 2019. ""Donald Trump is my President!" The Internet Research Agency Propaganda Machine." *Social Media and Society* .

Bond, Robert and Solomon Messing. 2015. "Quantifying Social Media's Political Space: Estimating Ideology from Publicly Revealed Preferences on Facebook." *American Political Science Review* 109(01):62–78.

Broniatowski, David A., Amelia M. Jamison, SiHua Qi, Lulwah AlKulaib, Tao Chen, Adrian Benton, Sandra C. Quinn and Mark Dredze. 2018. "Weaponized Health Communication: Twitter Bots and Russian Trolls Amplify the Vaccine Debate." *American Journal of Public Health* 108(10):1378–1384.

Clauset, Aaron, MEJ Newman and Cristopher Moore. 2004. "Finding community structure in very large networks." *Physical Review E* 70(6):066111.

Dawson, Andrew and Martin Innes. 2019. "How Russia's Internet Research Agency Built its Disinformation Campaign." *The Political Quarterly* 90(2):245–256.

Hobbs, William R. 2017. "Text Scaling for Open-Ended Survey Responses and Social Media Posts.".
URL: *https://ssrn.com/abstract=3044864*

Howard, Philip N., Bharath Ganesh, Dimitra Liotsiou, John Kelly and Camille François. 2018. "The IRA, Social Media and Political Polarization in the United States, 2012-2018." *Oxford Project on Computational Propaganda Working Paper* .

Jensen, Michael. 2018. "RUSSIAN TROLLS AND FAKE NEWS: INFORMATION OR IDENTITY LOGICS?" *Journal of International Affairs* 71(1.5):115–124.

Johnstone, Iain M and Debashis Paul. 2018. "PCA in High Dimensions: An Orientation." *Proceedings of the IEEE* 106(8):1277–1292.

Kim, Young Mie. 2018. "Uncover: Strategies and Tactics of Russian Interference in US Elections.".

King, Gary, Jennifer Pan and Margaret E. Roberts. 2013. "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review* 107(2):326–343.

King, Gary, Jennifer Pan and Margaret E Roberts. 2017*a*. "How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument." *The American Political Science Review* 111(3):484–501.

King, Gary, Jennifer Pan and Margaret E. Roberts. 2017*b*. "How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, not Engaged Argument." *American Political Science Review* 111(3):484–501.

Linvill, Darren L., Brandon C. Boatwright, Will J. Grant and Patrick L. Warren. 2019. ""THE RUSSIANS ARE HACKING MY BRAIN!" Investigating Russia's internet research agency twitter tactics during the 2016 United States presidential campaign." *Computers in Human Behavior* 99:292 – 300.

Linvill, Darren L. and Patrick L. Warren. 2018. "Troll Factories: The Internet Research Agency and State-Sponsored Agenda Building." *Working Paper* .

Llewellyn, Clare, Laura Cram, Adrian Favero and Robin L. Hill. 2018. "For Whom the Bell Trolls: Troll Behaviour in the Twitter Brexit Debate." *CoRR* abs/1801.08754.
URL: *http://arxiv.org/abs/1801.08754*

MacFarquhar, N. 2018. "Yevgeny Prigozhin, Russian Oligarch Indicted by U.S., Is Known as "Putin"s Cook"." *New York Times* .
URL: *https://www.nytimes.com/2018/02/16/world/europe/prigozhin-russia-indictment-mueller.html*

Mueller, R.S. 2019. *Report on the investigation into Russian interference in the 2016 presidential election*. Washington, D.C.: U.S. Department of Justice.
URL: *Volume I at https://www.justice.gov/storage/report*$_v$*olume*1.*pdf*, *accessed*18*June*2019; *VolumeIIath*
//www.justice.gov/storage/report$_v$*olume*2.*pdf*

One professional Russian troll tells all. 2015. *Radio Free Europe* .
URL: *https://www.rferl.org/a/how-to-guide-russian-trolling-trolls/26919999.html*

Østbø. 2017. "Demonstrations against demonstrations: the dispiriting emotions of the Kremlin's social media 'mobilization'." *Social Movement Studies* 16.

Roberts, Margaret. 2018. *Censored: Distraction and Diversion Inside China's Great Firewall*. Princeton: Princeton University Press.

Sanovich, Sergey. 2017. "Computational Propaganda in Russia: The Origins of Digital Disinformation, Working Paper." *Oxford Project on Computational Propaganda Working Paper* .

Sanovich, Sergey, Denis Stukal and Joshua A. Tucker. 2018. "Turning the Virtual Tables: Government Strategies for Addressing Online Opposition with an Application to Russia." *Comparative Politics* 50(3):435–482.

Schafer, Bret. 2018. "A View From the Digital Trenches: Lessons from Year One of Hamilton 68." *German Marshall Fund Report: Alliance for Securing Democracy* 3.

Slapin, Jonathan B and Sven-Oliver Proksch. 2008. "A Scaling Model for Estimating Time-Series Party Positions from Texts." *American Journal of Political Science* 52(3):705–722.

Stewart, Leo Graiden, Ahmer Arif and Kate Starbird. 2018. Examining Trolls and Polarization with a Retweet Network.

Tucker, Joshua A . 2018. "Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature." forthcoming.

Yin, Leon, Franziska Roscher, Richard Bonneau, Jonathan Nagler and Joshua A. Tucker. 2018. "Your Friendly Neighborhood Troll: The Internet Research Agency's Use of Local and Fake News in the 2016 US Presidential Campaign." *SMAPP Data Report* .

Zannettou, Savvas, Tristan Caulfield, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini and Jeremy Blackburn. 2018. "Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web." *CoRR* abs/1801.09288.

Supplemental Information

Network Community Detection

In the pre-processing for the network community detection, the user-mention data are represented using rows for the mentioned users and columns for the tweeting users, with the number of user-target mentions as elements. We standardize the rows of this matrix by dividing their row-wise sum and then use the cross-product of that matrix as the graph for the network detection algorithm. This graph represents the co-mentions of trolls conditioning on the overall mentions of the targeted accounts. We use the fast greedy algorithm (Clauset, Newman and Moore 2004) implemented in igraph to maximize modularity in the troll-to-troll graph. Modulary maximization algorithms select a number of clusters and cluster assignments that maximizes the number of within cluster connections and minimizes the number of across cluster connections.

All trolls who receive a community assignment by the network clustering algorithm are assigned that cluster in later analyses. Trolls without network information are assigned using the text scaling on their account descriptions. There are few enough of these accounts that including or excluding them has no noticeable effect on our analyses.

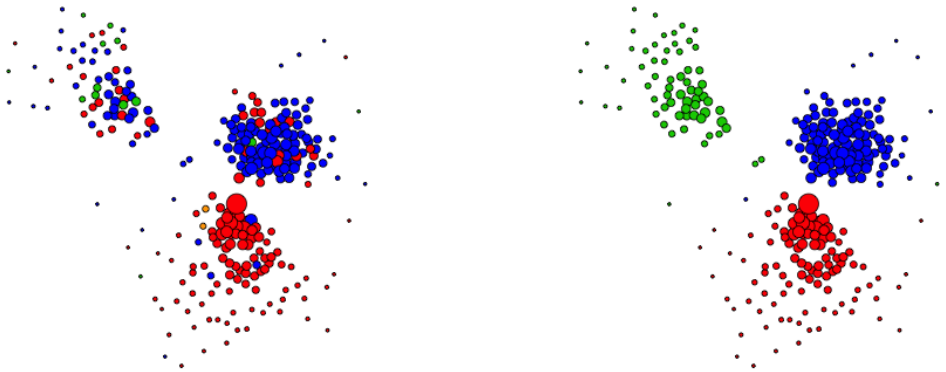

Figure 4: These networks replicate Figure 1 without tweets from 2015. Consistent with prior work, these clusters can be easily separated using community detection algorithms. The limited overlap justifies cluster-specific text analyses, since we expect vocabulary to be distinct as well.

Community Labels: text scaling

Account description keywords

The text scaling of user profile descriptions is weighted by the number of times an account posted in June 2015 through November 8, 2016.

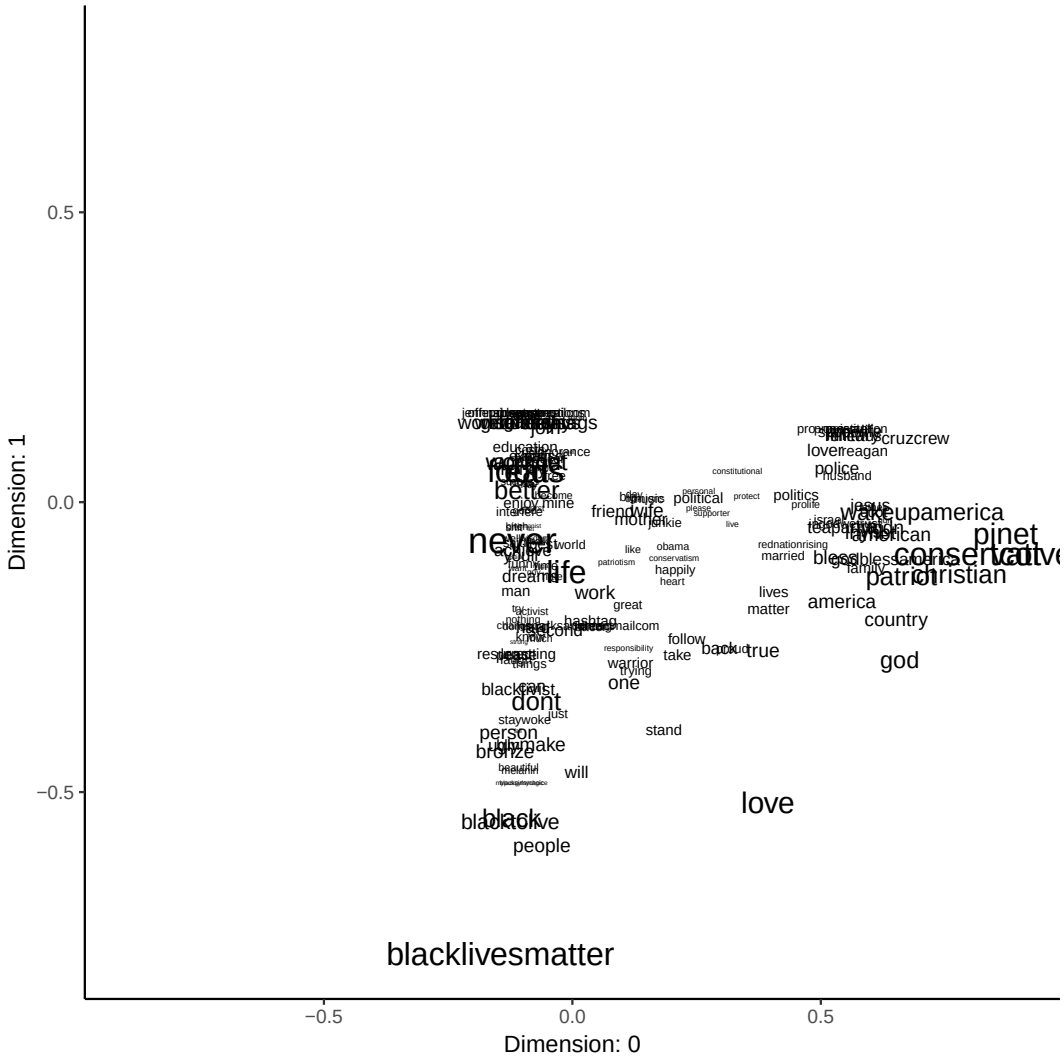

Figure 5: This figure shows the keywords for the conservative (dimension 0) and liberal (dimension 1) dimensions.

Mobilization of moderates

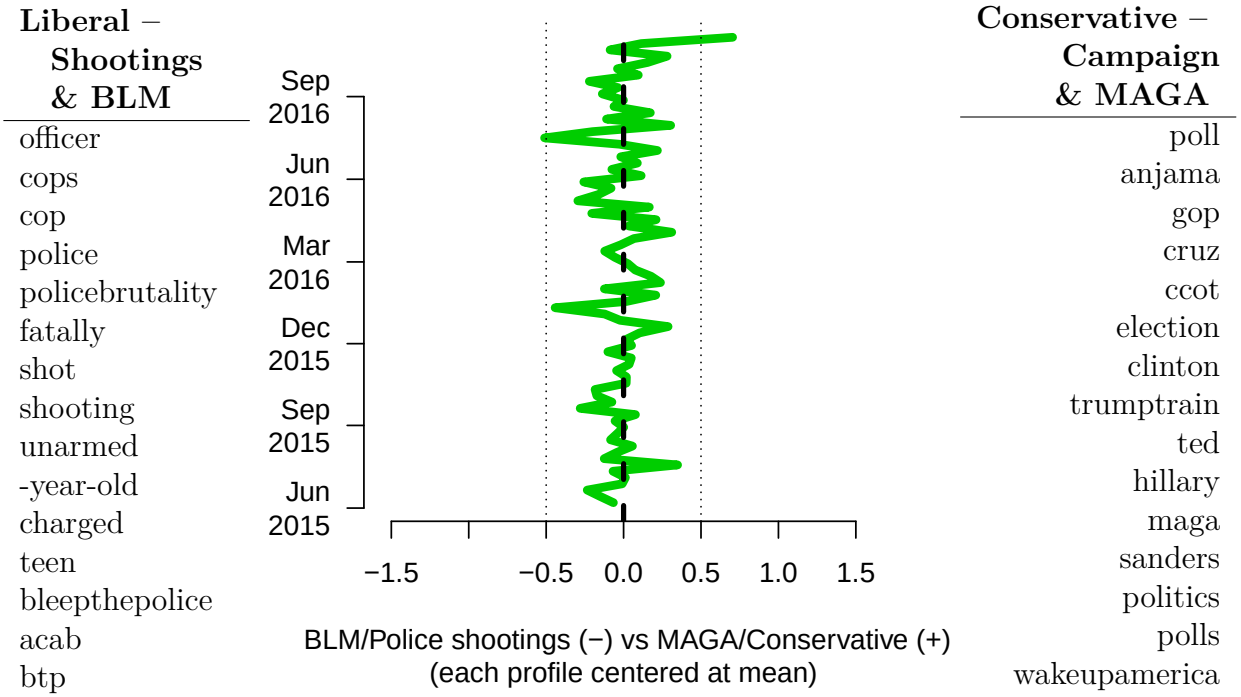

Table 3: *Moderate Mobilization.* This figure shows within account changes in messaging for politically ambiguous account cluster. This is the same data as Figure 1, but with each account centered at its mean.

Within Cluster Text Scaling

De-mobilization/mobilization time series, demeaned

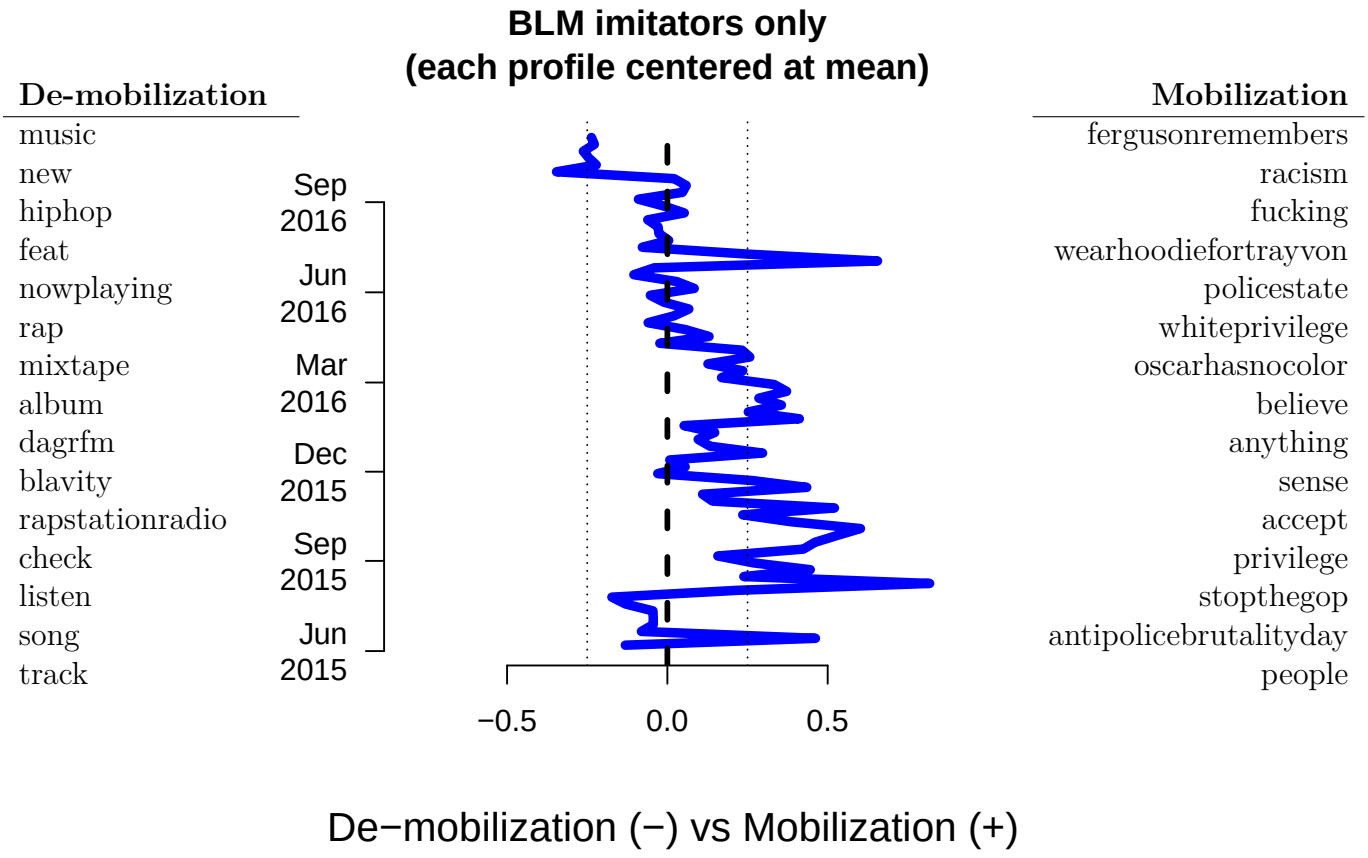

Table 4: *BLM Distraction.* Figure 2 showed a decrease in Black Lives Matter activity near the 2016 election. Here, we show the same result with each account centered at its mean. This shows bait-and-switch messaging within accounts over time. The text scaling dimensions in this figure use data from the "blue" liberal cluster only.

Dimension 1 of liberal cluster

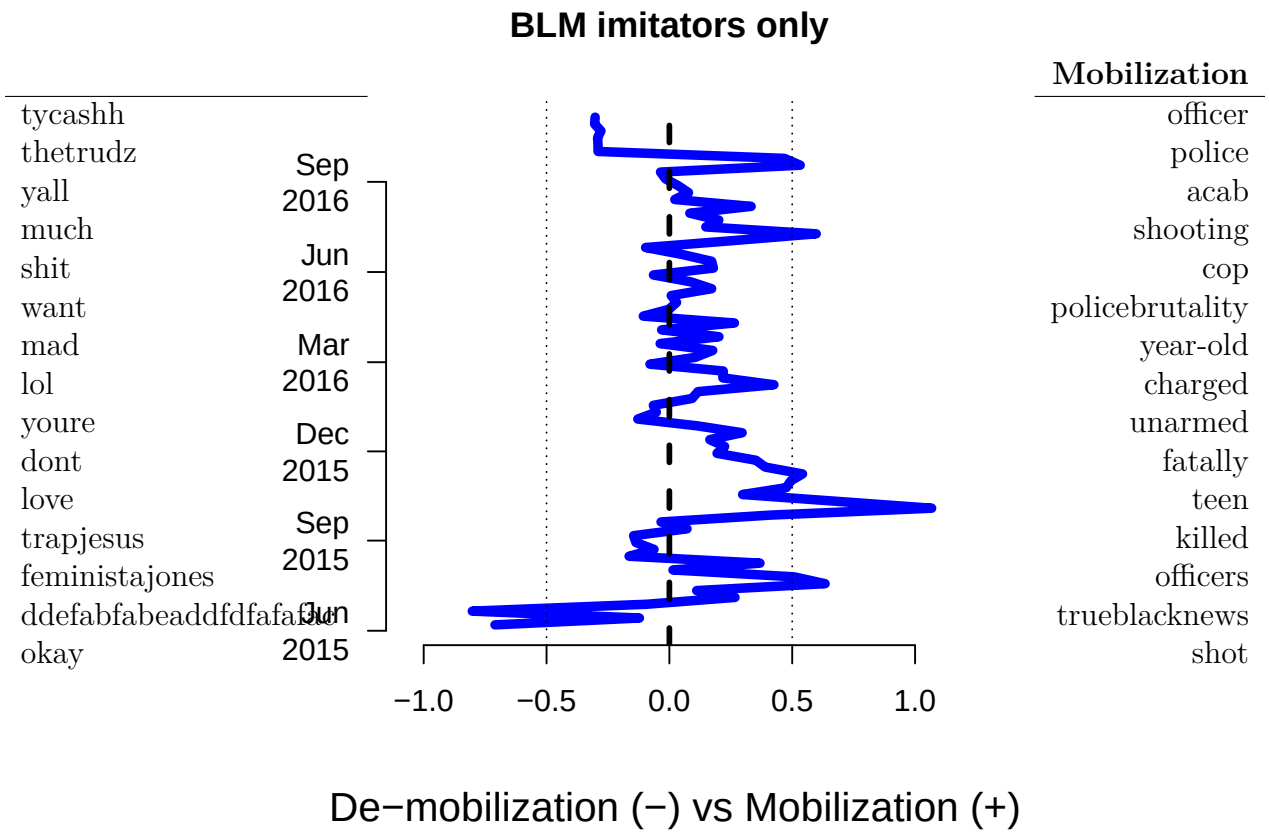

Table 5: *BLM Distraction in first dimension of text scaling.* Figure 2 showed a decrease in Black Lives Matter activity near the 2016 election. Here, we show the first dimension of the same text scaling. This dimension could have separated both informal language vs. news and demobilization vs. mobilization.

Table 6: fig:blm$_d$*istraction*$_d$1

Conservative news/politics time series

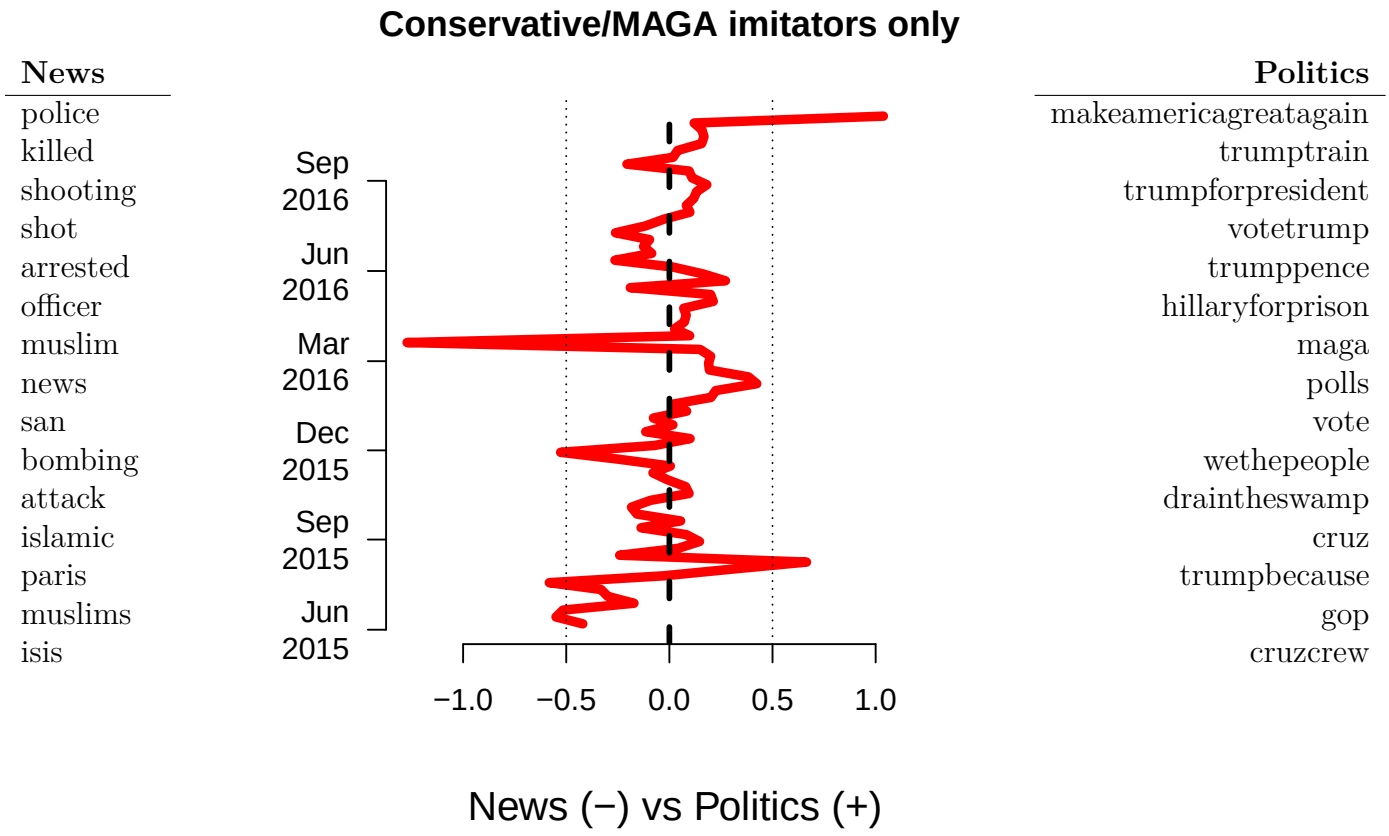

Table 7: *MAGA politicization.* This figure shows changes in how the conservative/MAGA cluster talked over time. The text scaling dimensions in this figure use data from the "red" conservative cluster only.

Dimension 2 of conservative cluster

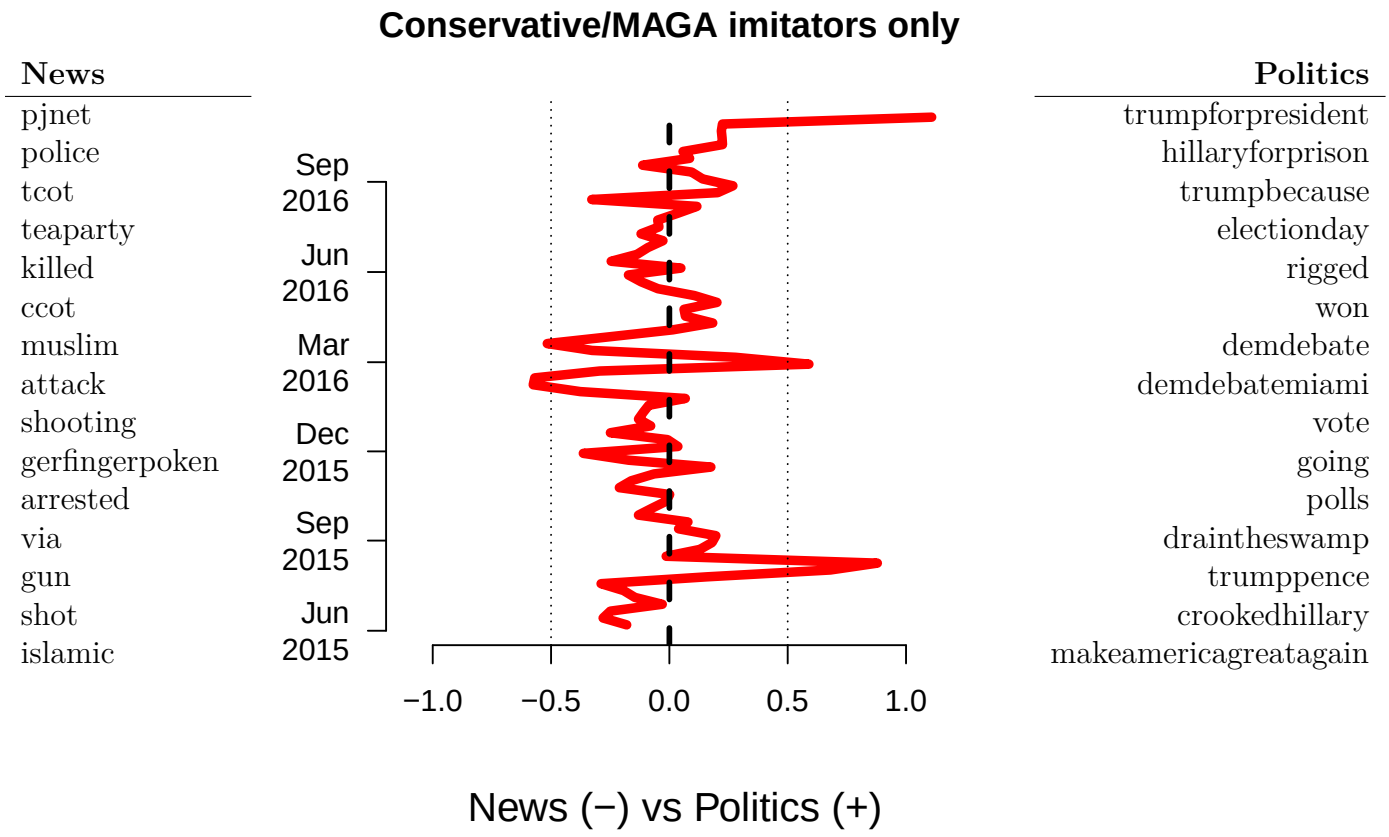

Table 8: *MAGA politicization.* Figure 7 show politicization of the conservative cluster over time. This figure shows changes in how the conservative/MAGA cluster talked over time using the first dimension. The text scaling dimensions in this figure use data from the "red" conservative cluster only. This dimension could have separated both international vs. domestic and news vs. politics. However, the timeline for the news dimension appears to be insensitive to this domestic vs. international distinction, however.

Conservative news/politics time series, demeaned

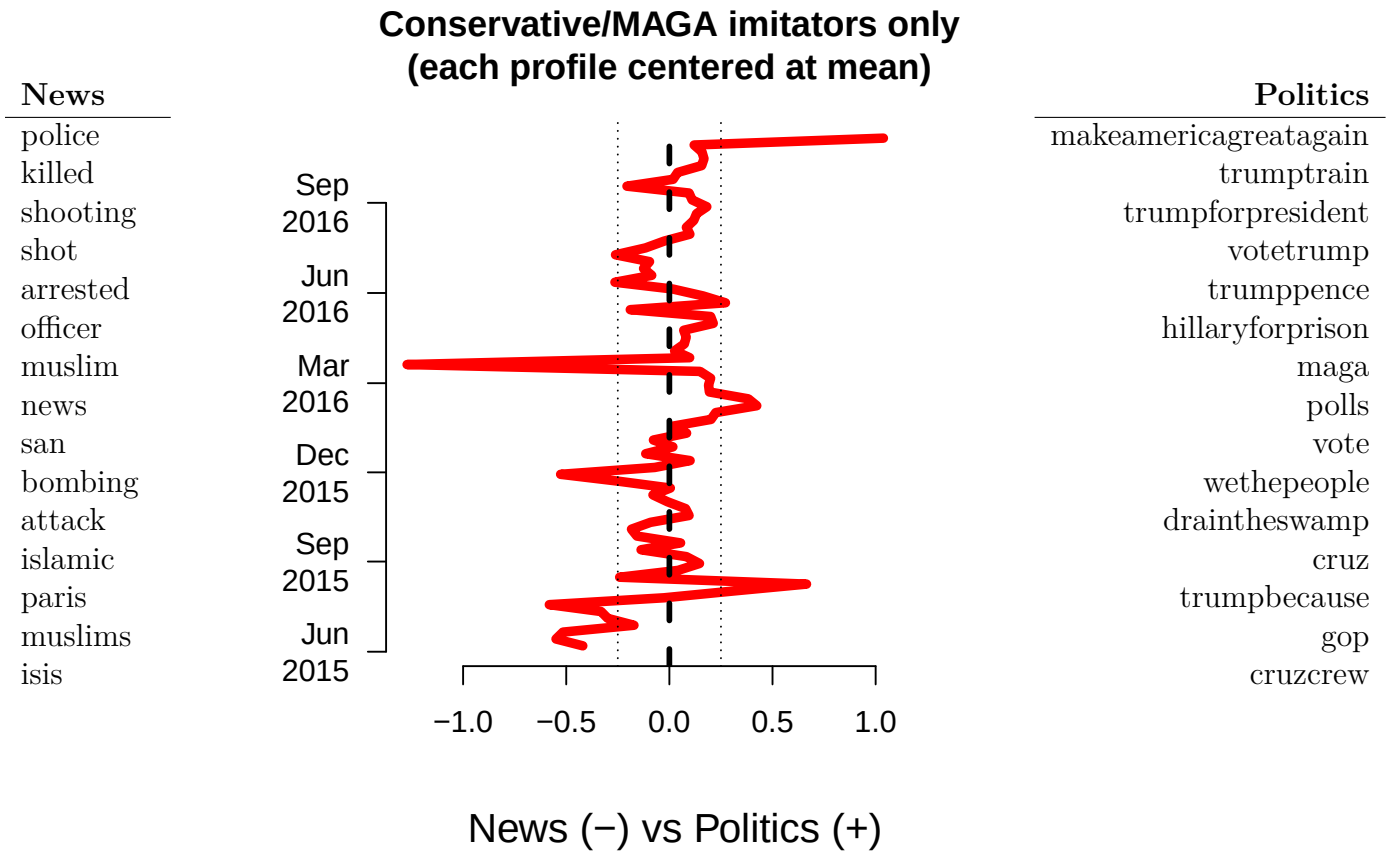

Table 9: *MAGA politicization.* Figure 7 showed changes in the conservative/MAGA cluster's messaging over time. This figure shows changes in how the conservative/MAGA accounts talked over time. It shows bait-and-switch messaging within accounts. The text scaling dimensions in this figure use data from the "red" conservative cluster only.

Text Scaling Keywords

Keywords: all tweets June 2015 to November 2016 data

Dim 1		Dim 2		Dim 3	
workout	news	ccot	workout	china	officer
lol	militants	gop	car	russia	cops
really	world	tcot	eat	chinas	unarmed
wanna	wounded	maga	exercise	talks	shooting
eat	turkish	pjnet	fatally	greece	cop
shit	screamymonkey	clinton	crash	britain	policebrutality
gonna	china	lindasuhler	sleep	anjama	police
dont	seeks	trumptrain	gym	deal	bleepthepolice
ass	south	wakeupamerica	girl	brexit	acab
like	dozens	hillary	injured	workout	-year-old
yall	rejects	cruz	weight	korea	shot
youre	officials	poll	killed	world	nypd
fat	kills	teaparty	fat	greek	btp
ihatepokemongobecause	rebels	trump	ass	minister	blacklivesmatter
weight	islamic	votetrump	shot	kerry	blm

Table 10: Top 3 dimensions' keywords: all tweets June 2015 to November 2016 data.

Keywords: MAGA/conservative cluster tweets June 2015 to November 2016 data

Dim 1		Dim 2		Dim 3	
clinton	prayerscalifornia	pjnet	trumpforpresident	ccot	islamkills
emails	islamkills	police	hillaryforprison	pjnet	police
email	stopislam	tcot	trumpbecause	tgdn	bombing
foundation	god	teaparty	electionday	cruzcrew	brussels
campaign	need	killed	rigged	tcot	shot
clintons	beingpatriotic	ccot	won	nra	shooting
wikileaks	guns	muslim	demdebate	teaparty	bomb
breaking	brussels	attack	demdebatemiami	wethepeople	officer
release	patriots	shooting	vote	trumptrain	found
poll	gunsny	gerfingerpoken	going	jstines	cop
via	cosproject	arrested	polls	tlot	arrested
investigation	constitution	via	draintheswamp	makeamericagreatagain	killed
haiti	country	gun	trumppence	wakeupamerica	prayforbrussels
hillary	guncontrol	shot	crookedhillary	cruz	san
fbi	shooterswife	islamic	makeamericagreatagain	stonewall	bombs

Table 11: Top 3 dimensions' keywords: MAGA/conservative cluster tweets June 2015 to November 2016 data.

Keywords: BLM/liberal cluster tweets June 2015 to November 2016 data

Dim 1		Dim 2		Dim 3	
tycashh	officer	blackpeopletwitter	clinton	trump	music
thetrudz	police	policestate	new	thehill	hiphop
yall	acab	policeviolence	thehill	gop	feat
much	shooting	policebrutality	campaign	donald	nowplaying
shit	cop	fucking	via	clinton	rap
want	policebrutality	acab	trump	hillary	rapstationradio
mad	-year-old	ebbdfcfdeaaedadadfefbeafdce	gop	obama	mixtape
lol	charged	blackpower	donald	sanders	listen
youre	unarmed	fergusonremembers	blicqer	republican	track
dont	fatally	wearhoodiefortrayvon	hillary	supporters	dagrfm
love	teen	whiteprivilege	news	vote	album
trapjesus	killed	alllivesmatter	politics	trumps	download
feministajones	officers	blacktolive	election	president	check
ddefabfabeaddfdfafafac	trueblacknews	blackskinisnotacrime	access	election	nineoh
okay	shot	cops	trumps	stopthegop	free

Table 12: Top 3 dimensions' keywords: BLM/liberal cluster tweets June 2015 to November 2016 data.