The Effect of Communication Networks on Cultural Change in Organizations: Evidence from Alt-Right Echo Chambers

Matthew Yeaton[*]

HEC Paris, 1 Rue de la Libération, 78350 Jouy-en-Josas, France

January 25, 2022

Click here to download the latest version

Abstract

Organizations seeking to change their cultures face a problem: changing culture directly is difficult. However, culture is deeply intertwined with other organizational features, especially informal organizational social networks. I theorize that changes to network "echo" (reflective within-group interaction) affect culture through the channel of language and communication. Since language is both an aspect of culture and the medium through which broader cultural priorities are communicated, changing the extent of network echo affects which cultural priorities organizational members observe around them and in turn what language these organizational members use themselves. I test this theory in the setting of online communities. Inducing cultural change in online communities and digital organizations is a key problem for platforms and in digitization policy as concerns mount about echo chambers that foment toxic cultures. I leverage a large-scale natural experiment on the platform Reddit to test the theory. The natural experiment shocked the networks of an Alt-Right community, transforming it from a high-echo network to a low-echo network. Across several ways of measuring language use, including measures derived from natural language processing techniques, I find that this shock causes a decrease in Alt-Right language imported to other communities by Alt-Right members (including a decrease in hate speech). This result overturns the conventional wisdom that breaking up echo chambers will simply displace problematic conversations.

I. Introduction

Disentangling the role of social networks on cultural change in organizations is a thorny problem because social networks and organizational culture are deeply and dynamically intertwined. Organizational culture is the systems and sets of meaning shared by organizational members, but the extent to which this meaning may be shared (or even observed) across an organization depends on the properties of the organization's informal social networks. Networks vary in

[*] yeaton@hec.fr

the extent that they exhibit social echo chambers, a situation where reflective, within-group interaction structurally insulates an organizational subgroup from the broader organization's culture. In a high-echo network, outside culture is unable to meaningfully penetrate the echo even if they greatly outnumber those in the echo chamber.

Network echo affects culture through the channel of language and communication. Since language is both an aspect of culture and the medium through which broader cultural priorities are communicated, changing the extent of network echo affects which cultural priorities organizational members observe around them and in turn what language these organizational members use themselves. Moreover, network echo is like a one-way mirror: outside culture is drowned out by the reflected culture inside the echo, but the those inside the echo bring their culture with them in their interactions with outsiders.

I test this theory in the setting of an online community on the platform Reddit, which is the fourth most visited website in the United States. I leverage a large-scale natural experiment that shocked the networks of an Alt-Right community, transforming it from a high-echo network to a low-echo network. I find that this network shock causes a decrease in Alt-Right language imported to other communities by Alt-Right members (including a decrease in hate speech). This finding is robust to a variety of methods for measuring language, including natural language processing measures related to document embeddings that are robust to spurious changes in the labels of words and the online problem of careful measurement of dog whistles.

This study makes several contributions. This study advances our understanding of the relationship between organizational culture and informal social networks. Despite the intuitive connection between them, it is nonetheless challenging to show strong empirical evidence of the impact of networks on culture. In traditional organizations, both culture and informal structure are hard to measure. Moreover, the tight link between culture and structure means that appropriately handling endogeneity is paramount to any meaningful inference. The possibility of measurement in the setting of online communities combined with the clean network shock partially ameliorate these empirical challenges.

Second, this study contributes to digitization policy and platform strategy. Inducing cultural change in online communities and digital organizations is a key problem for platforms and in digitization policy as concerns mount about echo chambers that foment toxic cultures. Additionally, online communities are a prime example of how organizations that leave toxic cultures to fester do so at their own peril. Online platforms and communities are one such highly visible setting facing this problem.

This study overturns the conventional wisdom that breaking up echo chambers will simply displace problematic conversations. If social networks affect linguistic culture, then network interventions become a reasonable strategy to combat toxic cultures in online communities. Finally, this study makes a methodological contribution through the development of a careful measurement of the significant problem of dog whistles and shibboleths that exists in all informal communication, but is especially present in (extreme political) online communities.

A. Organizational Culture as Language

This paper contributes to the growing view of organizational culture as accessible and encoded in language (Chen 2013; Crémer et al. 2007; Doyle et al. 2017; Srivastava and Goldberg 2017; Srivastava et al. 2018). The language and specfically the vocabulary used by members of the organization reveals the underlying cultural priorities of the organization.

First, the language facilitates the transfer of semantic content, the meaning itself. Language is the medium that some idea is communicated through. In the context of organizations, the focus has primarily been whether the language can efficiently communicate organizationally salient ideas (Weber and Camerer 2003; Wernerfelt 2004; Crémer et al. 2007; Koçak and Warglien 2020).

Given this premise of efficient language, what can the reader learn about the cultural priorities of the writer? One can learn about the cultural priorities of those in an organization by their use of specialized language or jargon. Consider the following example of frozen water in two languages: English and Inupiaq.

English has a handful of words for frozen water. Inupiaq, which is a language spoken in Alaska and northern Canada, has over fifty words for sea ice alone (Krupnik and Weyapuk 2010). The speakers of Inupiaq have many words for snow because their experience demands that kind of precision. While there is near infinite nuance in lived experience, part of what a language does is choose which of those things to describe precisely. Different groups have different priorities over ideas that are reflected in the language such that anyone looking at the language can then infer relative differences in priorities simply by looking at the language.

The relevance of language to an organization's culture goes beyond it being an aspect of that culture. This is because language encodes cultural information that is communicated alongside any substantive content being communicated with the language. That is, language is always communicating along at least two levels: the content itself, where language is simply the medium that this content is communicated through, and the encoded culture, that reflects the (possibly tacit) cultural priorities that went into the language's creation.

Language has several desirable properties with respect to measurement of organizational culture. I measure it directly in the same way it is observed by those in the organization. Additionally, I need not ex ante define the dimensions of culture before I measure it. This reduces the reliance on self-report or survey measures of culture, and reduces the possibility that the researcher may focus on aspects of culture that are of low relevance to the organization.

B. Cultural Evolution

Culture is dynamic rather than static, meaning culture changes through time and requires repeated observations to understand. This invites us to unpack the mechanisms involved in the evolution of culture. The dynamics of learning provide a useful tool to influence language and in turn culture. Social learning naturally invites the perspective of communication in networks and diffusion through networks.

However, this is an empirically challenging problem. Cross-sectional and time-series data on both the social network structure and culture of real-world organizations are scarce. Thus, strong causal evidence of a link between networks and culture is similarly scarce. The primary approaches thus far have been either theoretical or laboratory studies.

The theoretical side of the literature has focused on formal modeling especially using the DeGroot learning setup (DeGroot 1974; DeMarzo et al. 2003; Golub and Jackson 2010). Building off the results of the theoretical side of the literature, the laboratory studies have typically been parameter estimation tasks in order to map onto the spectral approach central to DeGroot learning (Centola and Baronchelli 2015; Becker et al. 2017; Centola et al. 2018).

The key result is that network structure can impact aggregate organization-level outcomes when people learn from each other. The network matters because it affects who learns from whom, and what the world looks like when you look around you.

However, we immediately run into a problem when we start trying to integrate this perspective with language, which is that language encourages us to think from the perspective of groups and subgroups, rather than from the perspective of individuals. I address this by generalizing the individual-oriented idea of spectral centrality to the group-oriented idea of conductance between groups.

C. Motivating the Model

The goal of our formal model of lanugage evolution in networks is to better understand the mechanisms involved in the evolution of language in social networks. The model has three interlocking components that are key to unpacking these mechanisms: network structure, language, and learning.

The primary result of our model of language evolution in social networks shows that priorities and language are endogenous to social network structure. This is the focus of our empirical predictions. The organization can shock or change the network to change users' language. Moreover, any shock that changes the language must also increase the conductance, or information flow, between individuals using idiosyncratic language and others in the community. This implies that a key mechanism in changing language is the increase in social learning across these groups.

The conductance between two groups in a network reflects the relative proportion of interaction with those from outside the group compared to inside the group. If the conductance between two groups is low, there is little interaction between the groups. Conversely, if the conductance between two groups is high, there is substantial interaction between the groups. This has lead to conductance being a popular measure in community detection algoritms (Yang and Leskovec 2015; Van Laarhoven and Marchiori 2016; Lu et al. 2018). In the context of a shock to the social network, changes in conductance reflect changes in the structural cohesion of interconnected groups. That is, increasing the conductance between two groups means we are making the groups more structurally integrated. Conductance can also be thought of as a spectrally-motivated version of the Krackhardt E/I Ratio (Krackhardt and Stern 1988). This structural integration, captured by increasing conductance, increases the amount of learning that takes place across group boundaries. This reduces the use of idiosyncratic vocabulary by the relatively marginal group.

To be clear, I am making a more specific hypothesis than simply that removing the forum removes some speech. Rather, changing one's neighbors is predicted to change one's language everywhere. Thus, while you might be more likely to reference dogs in a forum dedicated to dogs, you are predicted to write about them in a way that reflects how you write about dogs when the topic arises in another foum. Thus, should our hypothesis be true, individuals should use these words less in other forums compared to the pre-shock period. This is key for our empirical setting because it enables a much stronger test.

D. Online Communities and Echo Chambers

The share of all internet users who participate in online communites has continued to rise in recent years, from an already high 72% in 2017 to 76% in 2020. Amongst Gen Z and millenials, this share is close to 90% (Beer 2020). Thus, participants in online communities represent a growing share of a growing pie, as the number of global internet users continues to grow, rising to 4.13 billion in 2019 from 3.92 billion in the prior year (Clement 2020). People use online communites to connect with other users, have conversations, and share content.

The growth of online communities has led to a proliferation of digital records of natural language conversations. One key empirical area of interest in these communities is the extent to which they form echo chambers of opinions, urging participants toward ever more extreme points of view. Echo chambers emerge from self-selection into segregated communities of opinions (Sunstein 2001) and from algorithmic amplification of opinion segregation (Flaxman et al. 2013). However, while evidence for the existence of echo chambers accumulates, evidence for solutions remain limited.

This problem is especially relevant for online communtities where the content on the platform is the central value proposition of the community to its users. The typical response by platforms to the problem of extreme communities is to (a) ban the users themselves, (b) ban the forum hosting their conversations, or (c) both (Biddle 2015; Chandrasekharan et al. 2017). However, there is virtually no evidence on the impact of such interventions in changing user behavior. Additionally, the emergence of extreme communities opens platforms up to media risk, public scrutiny, and calls for government regulation.

A number of papers have sought to understand the extent of online segregation into echo chambers due to self-selection in networks (Sunstein 2001; Van Alstyne and Brynjolfsson 2005; Quattrociocchi et al. 2016), in news sources (Lawrence et al. 2010; Larcinese et al. 2011; Shore et al. 2016), and due to algorithmic influence on what media one is exposed to (Jeppesen and Frederiksen 2006; Ma and Agarwal 2007; Pariser 2011; Flaxman et al. 2013; Xu and Zhang 2013; Qiu and Kumar 2017). Additionally, many of these studies have focused on language as a means to infer opinions. However, most of these have focused on political ideology and political language (Sunstein 2001; Gentzkow and Shapiro 2010; Greenstein and Zhu 2012; Durante and

Knight 2012; Greenstein and Zhu 2016; Boxell et al. 2017; Greenstein and Zhu 2018) following the literature on ideological bias and political language in non-digital media (Gentzkow and Shapiro 2004; DellaVigna and Kaplan 2007; Gentzkow and Shapiro 2010).

While political ideology is an important aspect of preferences, especially with respect to news consumption, it is far from the only salient aspect of preferences and priorities in organizations, especially online. The political content of the Alt-Right community's posts in our empirical setting cannot be easily collapsed along the traditional left-right yardstick. Thus, one advantage of the empirical approach in this paper is in conceiving of priorities in a more nuanced way than this yardstick.

E. Cultural Change as Platform Strategy

Driven by the explosion of echo chambers in online communities, the problem of idiosyncratic vocabularies and language in organizations is an area that has skyrocketed in importance over the past decades. Despite the newness of the problem, the implications of this wave of idiosyncratic language are already broad and far-reaching.

Idiosyncratic language in online communities frequently takes the shape of extremist or even hate speech. Online hate speech transforms into real-world violence against marginalized groups (Fink 2018). Idiosyncratic language paves the way for the spread of misinformation that transforms into negative health outcomes and electoral tampering (Grinberg et al. 2019).

Faced with this environment, online platforms face increasing pressure to take action. This pressure takes the form of media scrutiny, fleeing advertisers and thus declining advertising revenue, brand risk, and calls for regulatory intervention. With this threat to their bottom lines, platforms must quickly find a way to "do well while doing good" or risk the loss of their mainstream advertising base.

One approach commonly taken by platforms is to ban users who engage in this type of speech. While changes in the population of users can make meaningful changes in the language content on online platforms (Danescu-Niculescu-Mizil et al. 2013; Kovacs and Kleinbaum 2020; Greenstein et al. 2020), banning users has an (obvious) negative effect on the number of users and may accordingly be revenue-decreasing for the platforms. Similarly, locking down the platforms too much chases off users to freer platforms and/or prompt blowback from the existing userbase.

My results demonstrate the possibility of another type of intervention available to platforms that lets them traverse this tightrope between advertisers and users: rewiring the social networks of their users while leaving the pool of users intact. The natural experiment on Reddit allows us to evaluate this type of "soft" intervention. I show that it is possible to change the vocabulary or language of those in an online community by changing the tapestry of local neighborhoods.

II. Model

The model augments a traditional DeGroot learning network setup (DeGroot 1974; DeMarzo et al. 2003; Golub and Jackson 2010) (DeGroot 1974; DeMarzo et al. 2003; Golub and Jackson 2010) with language-based communication (Crémer et al. 2007). Cultural evolution is a phenomenon at the intersection of social network structure, language, and social learning. Thus, the model has three central interacting elements that reflect this. The result is a dynamic model of language evolution in social networks. The main result is that language is partially endogenous to network structure. I propose that online communities can leverage the structural endogeneity of language to construct network shock interventions.

All of the formal propositions and their proofs can be found in the appendix.

A. Network Structure

Individuals in this model are organized in a network structure. Individuals are nodes in that network, and edges between individuals represent that one of them listens to or interacts with the other. In this model, edges serve as pathways that allow communication between individuals.

Formally, the social network is represented by the graph $G(V, A)$ which is defined over set of nodes or vertices V, and an adjacency matrix A. If there are N individuals in the network, then the adjacency matrix A is a $N \times N$ row-stochastic matrix, possibly weighted and directed (not necessarily symmetric). Element a_{ij} of A represents the influence of j on individual i. This influence may be due to interaction, conversation, or reputation. If $a_{ij} = 0$, then j has no influence on i. Since A is row-stochastic, all of the influence on an individual i must sum to one. Because of this, influence should be thought of in terms of proportions.

Throughout this analysis I will always assume there is some group within this network using idiosyncratic language that we want to disrupt, and there is some other group whose language is less idiosyncratic. Thus, it will be helpful for us to have a clear way to refer to these groups within this network. Formally, I will treat this allocation into groups as a network cut. A cut of a network partitions the nodes V into two disjoint subsets. Throughout this paper I will denote a cut by $(S, \bar{S})$ where $S, \bar{S} \subset V$.

The network cut enables us to think about the properties of the relationship between the individuals in these two groups. The principal measure of the relationship between these two groups for our purposes is the conductance of the network cut. This measure reflects the cross-pollination between these two groups, or how much those in the two groups can be observed and possibly influence one another. Formally, the conductance of a network cut is defined as

Definition 1 (Conductance of a network cut)**.** *Suppose we have a social network represented by the graph $G(V, A)$ which is defined over a set of nodes, V, and adjacency matrix, A. A cut of this network if given by $(S, \bar{S})$ where $S, \bar{S} \subset V$. The conductance of a cut is defined as*

$$\phi(S) = \frac{\sum_{i \in S, j \in \bar{S}} a_{ij}}{\min(a(S), a(\bar{S}))}$$

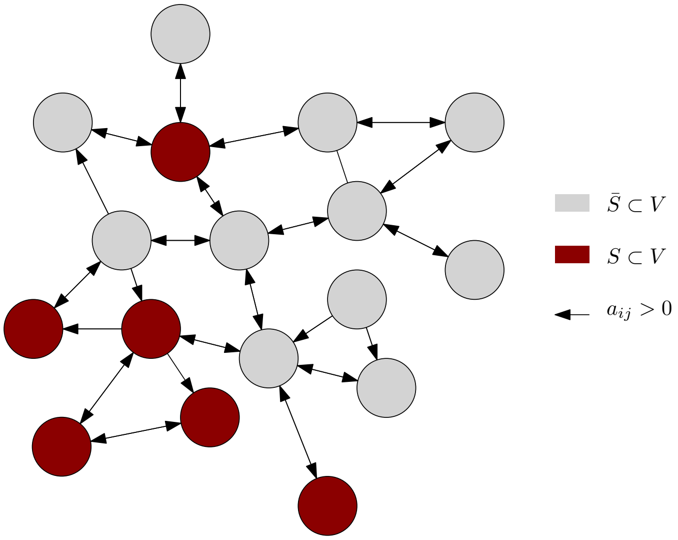


Figure 1. Directed Network with a cut. The nodes, V, in this example network are partitioned into two disjoint groups: S (red) and $\bar{S}$ (grey). This clarifies that a cut of a network is a property of the nodes, rather than the edges. The two partitions in the cut may be connected, as in this example, or comprise unconnected components.

where a_{ij} are the entries of the adjacency matrix A, so that

$$a(S) = \sum_{i \in S} \sum_{j \in V} a_{ij}$$

is the total weight of the edges incident with S.

The conductance measures the relative proportion of interaction and influence with those from outside the group ($\bar{S}$) compared to those from inside the group (S). If the conductance is low, then those in S are primarily interacting with others in S. This makes it difficult for the vocabulary of $\bar{S}$ to percolate through the individuals in S because it either does not reach them or is crowded out by other influence from those in S. Increasing the conductance allows greater interaction and learning to take place between the groups.

This measure clarifies one of the mechanisms of our proposed network shock interventions. Any shock that disrupts the idiosyncratic langauge of those in S will also increase the conductance between S and $\bar{S}$. This helps to emphasize that the social learning taking place is intrinsically tied to reducing the bottleneck of information between these two groups.

Thus, we can always increase the informational flow between two groups by focusing on the edges that cross over from one group to the other. If the network is serving as pipes for information to travel through, then increasing the conductance means that we are increasing the relative proportion of those pipes that cross between the groups so that we accelerate the flow of information between the groups.

B. Language

One source of novelty in this model, especially with respect to our setting, is clarity in what is transmitted in the pipes of the network: language. Especially in online settings (and certainly

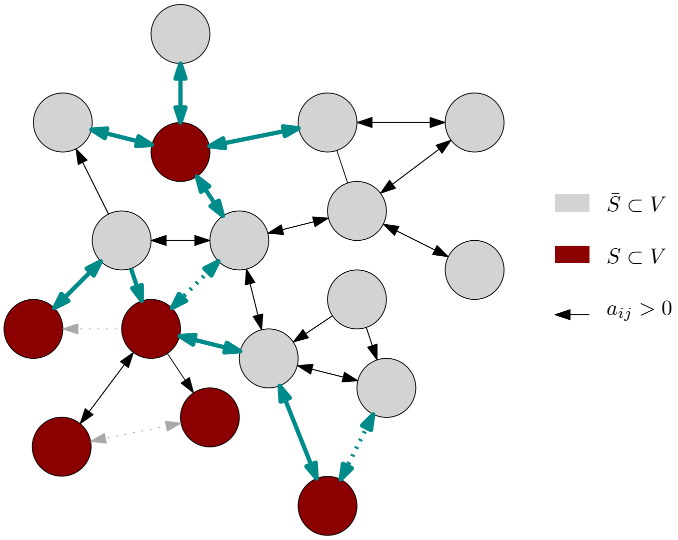


Figure 2. Directed network emphasizing the conductance-relevant edges. As in Figure 1, I create a cut of the network, i.e. I partition the nodes, V, into two disjoint groups: S (red) and $\bar{S}$ (grey). The conductance between these two groups captures the relative proportion of interaction with those from outside the group compared to inside the group. In particular, it measures the relative influence of the edges that 'cross the aisle,' or connect members of S to those in $\bar{S}$. I have colored these edges in teal. For the conductance between S and $\bar{S}$ to increase, we could increase the amount of influence from existing edges (making the solid teal lines larger), we could create new 'across the aisle' edges (the dotted teal lines), or, as is most relevant to our empirical context, we could *decrease* the influence of existing edges within a group (making the dotted light grey lines smaller).

in our empirical setting of Reddit), the primary way that individuals communicate with one another is by reading and responding to each other's posts and comments.

While one might wish that they could instantly and accurately communicate their rich inner state to those around them, we are constrained by the medium of communication, which in this case is the written word. While this may be a detriment to online communicators, it is a boon for us both theoretically and empirically. In a setting with no body language and no tone of voice, writers are constrained by what readers can observe. This match between observables for researcher and Redditor tightens the link between our modeling assumptions and the context we wish to understand.

Our model of language assumes that the set of priorities about which one would like to communicate is richer than the space of words one could construct. This induces the following tradeoff: if a word is precise, writing it should give the reader a very clear and immediate idea of what the writer intends. If a word is vague, it may take the reader longer to understand the meaning of the word, and possible confusion as to the writer's true underlying priority may ensue. Thus, if some priority is more important for the writer, it serves them to put that concept in a more specific word. Since the set of priorities is larger than the space of words, not every priority can be in a specific word. Thus, the choice of which concepts to assign to specific words encodes information about the writer's priorities.

This encoding of priorities via the relative specificity of words is precisely the aspect of language I am interested in for our setting of breaking up idiosyncratic speech in online communities. I suggest that idiosyncratic speech emerges in part because individuals in these communities place undue weight on concepts that are less important to the broader community. Idiosyncratic language is a natural result of our modeling assumptions if an individual places a high priority on the ideas underlying that language.

I will model this formally as a partition model of language in the style of Crémer et al. (2007). Briefly, in a partition language, individuals partition the set of priorities over what they would like to communicate into bundles, or words. They choose these bundles so that their most important priorities are in narrow bundles with little else, while less important priorities are in broader bundles with many other unimportant priorities.

Formally, every individual has priorities about what concepts are important to them from a finite set X. Each member's priorities are reflected by a probability distribution over this set. So for individual i, their priority for $x \in X,\ |X| = Z$ is denoted $p_{i,x} > 0$. $p_{i,x}$ reflects how important the concept or idea x is to individual i.

A language or vocabulary L is a partition $\{w_1, \dots, w_K\}$ of X. Saying word w_k indicates that one is trying to express some concept $x \in w_k$. I assume that the number of words $K < Z$ is fixed. This is a behavioral assumption that means that the language is coarser than the full breadth of possible concepts. The breadth of a word is given by $n_k = |w_k|$ and the frequency of a word is given by $p_k = \sum_{x \in w_k} p_x$. The breadth of a word reflects its specificity. Specific words are narrow, so that n is small, while vague words are broad, so that n is large. The frequency of a word reflects how often, in expectation, the ideas from that word will be used.

For a given word, the effort of ascertaining exactly which concept $x \in w_k$ is intended is given by a function c that is strictly increasing in n_k and convex so that vague words have higher "interpretation cost" than more precise words. The convexity assumption reflects the increasing ambiguity of parsing words with many possible meanings. Then the expected cost of a language L_i is

$$C(L_i; p_i) = \sum_{k=1}^{K} p_k^i c(n_k) \tag{1}$$

The optimal language for some individual i minimizes C for a given p_i. An example demonstrating the process of creating a partition langauge is shown in Figure 3.

The richness in this model and applicability to our context results from the dynamic interaction of the individuals in a social network. I am interested in the evolution of language as individuals converse and are influenced by one another. I now turn to learning.

C. Learning

The learning in this model is local social learning. It is local in the sense that individuals only observe their neighbors in the network. An individual i's neighborhood is comprised of those with whom i directly interacts, i.e. those j such that $a_{ij} > 0$. It is social in the sense that individuals observe other individuals in the network (rather than observing information shocks from outside the network, for example). Additionally, for individuals to learn, time must pass. I consider a two period model. Table I briefly describes the timing of the model.

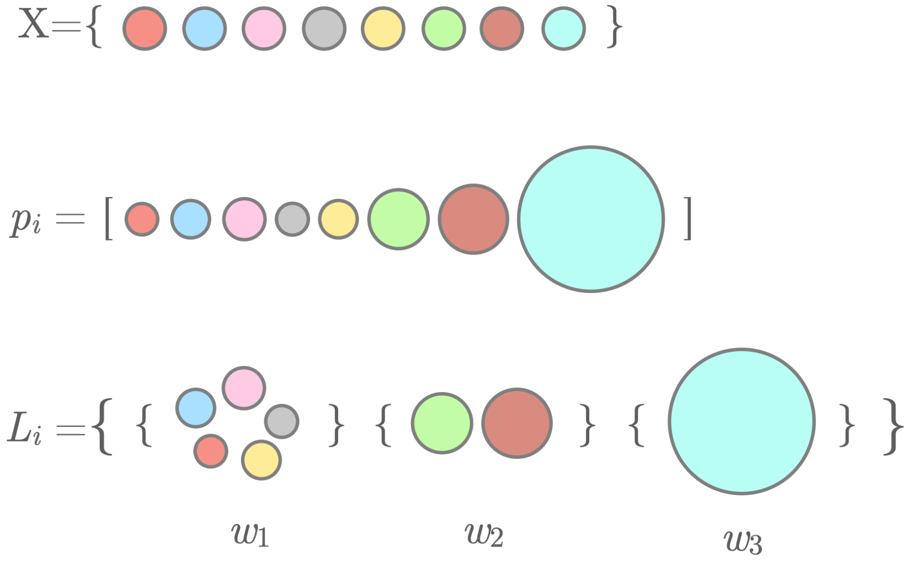


Figure 3. Partition Language. This figure gives an example of how to create a partition language. I begin with a finite set of concepts or ideas, X. This set of concepts is common to all individuals and has no inherent value or ordering. Individual i has priorities about which concepts are important to them, reflected by the probability distribution p_i. $p_{i,x}$ reflects the relative importance of the idea or concept x to the individual i. In the figure, this importance is reflected in the size of concept. The language takes the concepts, and partitions them into words. Since the space of possible ideas and concepts is richer than the language can allow, there are fewer words (in this case 3) than concepts in X (in this case 8). This implies that some words may need to contain more than one concept, making them less precise in conveying meaning. Additionally, it is costly for the listener to understand the meaning of a less precise word. The optimal language therefore grants the most important concepts the most precise words. In this example, this is reflected by the largest concept, light blue, to be granted its own word. The second largest set of concepts, red and green, share a relatively precise word, while the remaining concepts are grouped together into a quite imprecise word.

C.1. Learning Neighbors' Priorities

There are two types of learning in this model. First, individuals make an inference about their neighbors' priorities. Second, individuals update their own priorities by combining these estimates with their original priorities.

Individuals try to learn their neighbors' priorities by making inferences about the priorities underlying their neighbors' languages. They must do this because individuals observe the language used by their neighbors, but they do not observe the priorities that influenced the creation of that language. So they must make inferences about what those priorities are given the language that they observe. When they observe the language of a given neighbor, they observe two things: (1) the frequency of each of the words, which reflects how often in expectation each word is used, and (2) the set of possible priorities that could have been intended with each usage of the word.

Given the observed language, each individual makes a boundedly-rational inference about the priorities underlying that word. The boundedly-rational learning nests the Bayesian estimate as a special case. In other words, individuals are permitted to be fully rational, but they need not be for the results of the model to hold. Moreover, this may vary by individual. The

Table I. Model Timing

t_0: (a) We observe the initial network, including initial partition languages. We choose a cut of the network S such that those in S have a more specific word for some focal concept than those in $\bar{S}$.

(b) Network shock: We re-weight the network so that the conductance between S and $\bar{S}$ is now higher.

t_1: (a) Each individual outputs their language to each of their network neighbors.

(b) Learning Neighbors' Priorities: individuals make a boundedly rational inference about the underlying priorities of each of their neighbor's languages.

(c) Social Updating: individuals update their priorities by combining their estimates of their neighbors' priorities with their own.

(d) Each individual re-solves the partition language problem given their updated priorities.

full details of the boundedly rational learning are in the appendix, and can be best understood through their deviation from the fully rational case. The one-period Bayesian estimate of language under normally distributed priorities consists of taking an equally weighted estimate of each of the priorities in a word. If a word w has a frequency of p and n elements, the estimate for each p_j in w is $\hat{p}_j = p/n$. The intuition of the boundedly rational learning is that individuals may make any inference so long as they respect the ground truth of their observation of language so that their estimate can 'put the language back together again.'

There are two assumptions that go into this learning rule. Individuals observe the frequency of a word, and the possible priorities that resulted in that word's use. That one observes the frequency of a word is relatively innocuous. If two individuals are conversing they will eventually discuss enough to have a representative sample of each other's language use, and at minimum this will always hold in expectation. Indivudals also understand the possible priorities that resulted in a word's use. This is likely to hold more for some words than others. For example, when someone says precipitation everyone is likely to understand that the possible underlying meanings are one of rain, snow, sleet, or wintry mix. However, occasionally readers may be authentically confused as to the possible underlying meanings. To support this assumption, I lean on our empirical setting of online communication and suggest that if readers are truly confused they may perform time-consuming searches via asking in the thread, Googling for an answer, etc.

C.2. Social Learning

Individuals update their priorities in response to social information. Individuals now have an estimate of each of their neighbors' priorities that they combine with their own by weighted averaging. The weights have two components. The first is from the adjacency matrix A. The second component is a "stubbornness" coefficient that reflects that individuals are likely to overweight their own priorities and the priorities of those that share their views. These weights might be influenced by that individual's beliefs about the precision of their neighbors' information or by their neighbors' status or reputation.

Overall, this means that individuals are influenced by their neighbors in the network. One can take advantage of this via the network shock, which is formalized by the choice of the updated adjacency matrix A'. In particular, one chooses an A' that increases the conductance between S and $\bar{S}$. This increases the rate at which language of $\bar{S}$ can be observed by those in S by increasing the proportion of interactions that those in S are having with those in $\bar{S}$. Importantly, this shock to the network is the only tool at our disposal to change language in this model.

Effectively this stage of learning is a penalized version of DeGroot learning (DeGroot 1974) as individuals weight their updated priorities according to the adjacency matrix A' as a weighted sum. Additionally, weighted averaging of neighbors' estimates is rational under normally distributed priorities when the weights reflect the precisions of those priorities.

Overall, both stages of learning require a degree of bounded rationality. However, I believe that this model's case is helped that both stages of learning nest rational learning within a more general set of learning regimes. Thus, while I assume individuals make boundedly rational choices in learning from their neighbors, I neither rule out that they are rational, nor rely on knife-edge rationality assumptions.

D. Model Results and Discussion

These three interacting elements of network structure, language, and social learning are the building blocks that comprise our dynamic model of language evolution in social networks. There are two primary sets of results that are our focus. The first result demonstrates that very small changes in someone's priorities can cause them to change their language. In other words, language is not always robust to small changes in underlying priorities. This helps to reassure us that especially when individuals have extreme or marginal views, we may be able to change their language with even small shifts in their priorities.

The second set of results show that language is endogenous to social network structure. In particular, I show that we can shock the social network by "rewiring" it in order to disrupt and break apart idiosyncratic speech. Additionally, I show that any rewiring that accomplishes the task of disrupting idiosyncratic language must increase the conductance between those using this language and the rest of the community. Formal versions of these propositions and their proofs can be found in the appendix.

D.1. Clumpy Language Change

Individuals' language can be robust to changes in priorities, so that shifts in their priorities have no effect on their language. On the other hand, language can shift dramatically in response to extremely small changes in individuals' priorities. Both of these properties of the relationship between language and its underlying priorities can exist simultaneously because language is inherently coarser than the underlying distribution of priorities. This means that while priorities can change in a smooth or continuous way, the changes in language induced by those smooth changes in priorities will be ‘clumpy.’

The intuition for this idea is illustrated in Figure 4. Consider the priority weight p_n^*. Below p_n^*, one would create a word of size m. But as p_n^* grows, it eventually becomes worthwhile to reduce the size of the word so one can refer to n more precisely so that above p_n^* one would create a word of size $m-1$. Thus, for any possible (even infinitesimally small) neighborhood around p_n^*, there are two different languages. Since a language is a partition of discrete ideas or priorities into words, we are mapping a continuous object into a discrete object. This is the source of the ‘clumpiness.’ The takeaway is that we do not require large or dramatic shifts in priorities to result in a change in language.

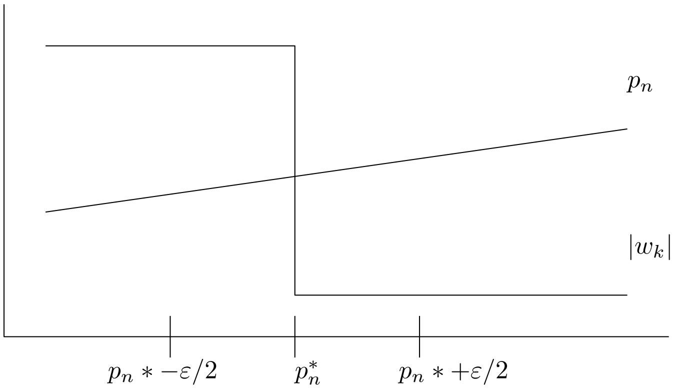


Figure 4. Clumpy Language Change. This illustrates how smooth changes in priorities can induce discrete or ‘clumpy’ changes in language. w_k is the word containing some idea or concept n. The smoothly moving line is the value of the priority on n, p_n. The other line is the precision of the word w_k, captured by the number of concepts in that word, $|w_k|$. Below p_n^*, one would create a word of size m. But as p_n^* grows, it eventually becomes worthwhile to reduce the size of the word so one can refer to n more precisely so that above p_n^* one would create a word of size $m-1$. Thus, for any possible $\varepsilon > 0$ (even infintessimally small), there will always be a neighborhood around p_n^*, $(p_n^* - \varepsilon/2, p_n^* + \varepsilon/2)$, that contains two different languages.

This property of clumpy language change is especially valuable given our empirical setting of echo chambers in Alt-Right communities. It is difficult to change people’s minds, especially in political contexts. However, this result demonstrates that changing someone’s mind a very small amount may still be enough to change the language they use. Idiosyncratic language emerges because individuals place greater priority on the ideas underlying those words than the broader community. However, idiosyncratic language is expensive to maintain, so that these priorities are more likely to be near the edge of their respective languages, ready to tip over into a more mainstream language with a small bump. One can therefore ‘unwind’ echo chambers by exploiting this property.

D.2. Change the Network to Change the Language

Language is endogenous to social network structure in the presence of social learning. In other words, one can shock, or ‘rewire,’ the social network in order to disrupt and break apart idiosyncratic language. Moreover, any such shock will increase the conductance between the group using the idiosyncratic language, S, and the rest of the community $\bar{S}$. This illustrates that a key source of the language change is the increase in social learning between S and $\bar{S}$.

This implies that language, and culture more broadly, is not a static construct. One need not consider it fixed in place, and need not take it as given. Instead, the language responds to the social structure of the community and the informational bottlenecks that the structure imposes (or removes). Network 'rewiring' is then a tool that an organization can use to change its language and culture. Moreover, this is a tool that is reasonably at the disposal of most organizations, whether online or not. This means that an organization could practically leverage this result to design an interventions to improve its culture.

The network shocks I focus on in this result are oriented around increasing the conductance between the group using the idiosyncratic language, S, and the rest of the community $\bar{S}$. Conductance is useful in illustrating the theoretical mechanism of the shock, but it is also useful as a simple scalar measure that captures changes in the mechanism.

The importance of conductance is also the importance of neighborhoods. Conductance captures the relative proportion of interaction or influence with those from outside the group compared to inside the group. With respect to one individual, to what extent does the composition of his or her neighborhood reflect their linguistic group membership? If conductance is low, neighborhood and group membership are strongly overlapping. One interacts primarily with those who share similar idiosyncratic language. This is an echo chamber. By increasing the conductance, we are decreasing the proportion of one's neighborhood that shares this idiosyncratic language.

While local neighborhoods convey information about conductance, conductance conveys information about neighborhoods and those neighbors neighborhoods, ad infinitum. Even if one's individual neighborhood remains constant, the growing conductance (a) shortens the travel time for outside information to reach that individual, and (b) reduces the amount of duplicated idiosyncratic language one will see. Thus, conductance is an excellent measure of structural insularity with respect to information flows.

Practically speaking, one can increase the conductance by increasing the weight of 'across the aisle' ties: edges that connect those in S to those in $\bar{S}$. One can do this by increasing the amount of interaction between these two groups by creating new edges or growing existing edges. However, one can also do this by decreasing the amount of interaction of individuals in S amongst themselves. Our empirical design is focused on the second case.

My formal modeling results lead directly to empirical hypotheses that I can test in my empirical setting. First, the ban of the Alt-Right subreddit will induce a network shock that increases the conductance between Alt-Right posters and others. Through this shock to insularity reflected in the conductance, those Alt-Right posters will decrease their use of idiosyncratic Alt-Right language. It is important to emphasize here that I am not making the trivial claim that deleting a forum removes language related to that forum. I make the much stronger claim that changing the network will change these individuals language use *wherever* they are posting.

III. Empirical Setting and Data

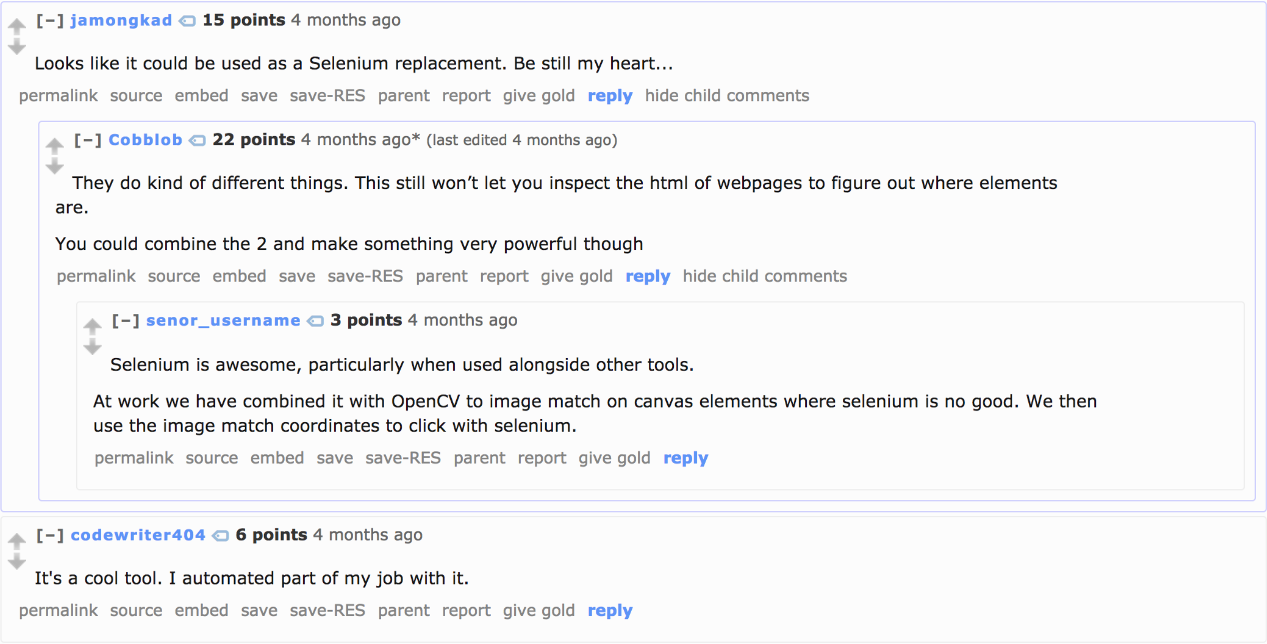
A. Empirical Setting

The empirical context for this study is an online community on the platform Reddit.com. Reddit bills itself as "the front page of the internet" and is the fourth most visited website in the United States and seventh most visited globally with 542 million monthly visitors according to the web analytics company Alexa.

The website is partitioned into communities called subreddits where users can post links and can comment and discuss with one another within the posts. Specifically, these comments are organized into a tree structure (so that users reply directly to each others' comments). An example of what comments on a Reddit post look like are shown in Figure 5.

Users have persistent accounts, so users can carry on persistent (if anonymous) conversations on the platform. The anonymity is advantageous given the questions we are interested in since users are more likely to speak truthfully about subjects that have associated stigma. The repeated patterns of interactions and conversations amongst users can be represented by a weighted, directed social network of users.

Consider two users as nodes of this network. If one writes a comment, and the other responds, then they now share an edge. The directed edge weights are the number of responses one makes to another specific user's comments and vice versa. Thus, if two users converse with each other frequently, they will have large edge weights, and on the other end of the spectrum if two users never respond to each other, then they will not share an edge. Since the network is weighted *and* directed, edges need not be symmetric.


[–] jamongkad 15 points 4 months ago

Looks like it could be used as a Selenium replacement. Be still my heart...

[–] Cobblob 22 points 4 months ago* (last edited 4 months ago)

They do kind of different things. This still won't let you inspect the html of webpages to figure out where elements are.

You could combine the 2 and make something very powerful though

[–] senor_username 3 points 4 months ago

Selenium is awesome, particularly when used alongside other tools.

At work we have combined it with OpenCV to image match on canvas elements where selenium is no good. We then use the image match coordinates to click with selenium.

[–] codewriter404 6 points 4 months ago

It's a cool tool. I automated part of my job with it.

permalink source embed save save-RES parent report give gold reply hide child comments
permalink source embed save save-RES parent report give gold reply hide child comments
permalink source embed save save-RES parent report give gold reply
permalink source embed save save-RES parent report give gold reply

Figure 5. Example of what comments on a Reddit post look like, showing (1) the tree structure, (2) the persistent user accounts (in blue).

B. Empirical Strategy

I exploit a natural experiment in order to evaluate the effect of informal organizational social networks on language. In January 2017, Reddit banned the subreddit r/Altright for violating the terms of service of the website by "doxxing," or posting on the subforum the real life identities and locations of people without their consent[1].

Although they banned the subreddit, they did not ban any individual users. Moreover, the users of r/Altright are participants on other subreddits both before and after the ban. The ban is therefore analogous to banning their meeting place, but allowing them to continue to meet in other places. By keeping the users (nodes) constant, but banning their meeting place, this is a shock to the edges of the social network. Specifically, the shock should dissolve or weaken the within-group bonds among r/Altright users.

This transforms the r/Altright network from a high-echo to a low-echo network. In the high-echo network, users are exposed to a consistent bundle of language, vocabulary, and viewpoints so that r/Altright users import the language and vocabulary from the Alt-Right community to discussions with other communities. In the low-echo network, these same users are have increased exposure to the non-Alt-Right bundle of language, and proportionally less exposure to the Alt-Right bundle of language. Consequently, the non-Alt-Right bundle appears to be a larger share of what the "world" looks like from the vantage point of their network neighborhood. This should decrease the amount of Alt-Right language that they import into other subreddits. Since decreasing their exposure to one community changes their language in other communities, their language is endogenous to their social network.

There are several aspects of the design of the empirical strategy that address a key alternative explanation to the causal mechanisms put forth in this study: self-censorship effects. Even if the ban cleanly shocks the social network by changing neighborhoods, it may have other effects on the language of users outside of the impact through the network change. In particular, the members of the community may react to the ban by changing their language in order to avoid getting their other communities banned. There are three main things I do to address this.

First is the choice of which community ban to examine. While content-based community bans are not uncommon on reddit, the r/Altright community is banned not for content-related reasons, but instead for violating other parts of the rules (in this case the rules against doxxing). This sets this ban apart from many others which explicitly claim content as the reason for the ban.

Second, the main tests are conducted in another Alt-Right-leaning subreddit: the_donald, which is a subreddit that describes itself as "a never-ending rally dedicated to the 45th President of the United States, Donald J. Trump." (Reddit 2019). It is a large subreddit with substantial overlap in membership with r/Altright. It is ostensibly ideologically and topi-

[1] Specifically, they shared identifying information about the man who punched Alt-Right figure Richard Spencer in the face after the January 20, 2017 inauguration of Donald Trump as President of the United States (Lancaster 2017).

cally similar to r/Altright, and is generally considered an Alt-Right-friendly space in media accounts. These factors combine to reduce concern over users censoring themselves for fear of social censure.

The third self-censorship concern is specific to the empirical setting. Alt-Right users make extensive use of dog whistles: coded language that communicate controversial ideas to ingroup members without the detection or scrutiny of outgroup members. While perhaps an interesting finding, it would be inaccurate to describe a change in the "label" of a dog whistle as a change in the semantic content of users' language. A change in language is more than a change in the label of a concept, and means a change in the underlying semantic meaning of users' conversations. I address this with a language measure that leverages document embeddings which help one to separate spurious changes in labels from changes in underlying meaning. This measure is elaborated on below.

Nevertheless, I cannot fully rule out the possibility of self-censorship effects. Self-censorship effects do not greatly affect the study's implications with respect to digitization policy and platform strategy. But to the extent that we can handle this alternative story it helps us to generalize these results to a more broad set of organizational contexts.

There are two steps required to test the effect. First, does the network shock induced by the subreddit ban transform the network from a high-echo network to a low-echo network with respect to Alt-Right users and others. Network echo is operationalized as network conductance between r/Altright users and non-r/Altright users. Figure 2 illustrates how rewiring network edges affects the conductance. The network shock primarily decreases the communication amongst Alt-Right posters (the dotted light-gray lines in the figure). This is a negative shock to the network that (proportionally) increases exposure to the rest of the network. This clarifies that we need not increase 'across the aisle' communication to increase the conductance. The shock affects the *relative* edge weights between Alt-Right users and others, thereby increasing the conductance between the groups.

Second, do the r/Altright users decrease the amount of Alt-Right language that they import into other subreddits, changing their language. The model plainly predicts that changing the network in a way that increases the conductance between these groups will reduce the frequency that Alt-Right users write with idiosyncratic language, in particular Alt-Right jargon. Continuing with the framing of our model, I will define group S as those who commented in r/Altright before the ban, and group $\bar{S}$ as those who did not comment in r/Altright before the ban but commented in other places that users in group S commented. The model predicts that since the conductance between S and $\bar{S}$ has gone up, the two groups should become closer in language after the ban than before the ban. This hypothesis is not trivial, as you could also imagine that users from S could simply go to another subreddit to continue their conversations. Specifically, the theory says that (a) if some commenters have more influence in the new network, one should see this influence in the new language, and (b) if the new network increases the conductance between these groups, then the language of S should reflect that.

To hone in on the effect of influence and to be as conservative as possible with respect to substitution/displacement effects, I compare Alt-Right poster comments *only in non-Alt-Right subreddits*. First, this ensures that our comparison in the two time periods is as close to apples-to-apples as possible, since the Alt-Right subreddit does not exist in the post-ban period by definition. Second, this makes our test conservative with respect to displacement effects of the ban. If all the ban does is shift conversations that were taking place in the Alt-Right subreddit to other subreddits, then one should see more Alt-Right jargon in formerly Alt-Right posters' comments in other subreddits after the ban. A displacement hypothesis predicts that Alt-Right users comments in these other subreddits in the post-ban period should be less similar to the average comment from these subreddits than before the ban. However, my model predicts that the shock should change Alt-Right users' language wherever they are posting, reducing the amount that Alt-Right users import idiosyncratic jargon into other subreddits.

Each test of this main prediction takes the same general form, but varies the dependent variable: the method of measuring idiosyncratic language. Specifically, I run the following set of regressions:

$$y_{ijt} = \beta_0 + \beta_1 postban_{ijt} + \gamma_j + \varepsilon_{ijt} \tag{2}$$

where y_{ijt} is the idiosyncratic language measure for comment i for individual j in time period t. $postban_{ijt}$ is an indicator that is 0 if the comment is made in the two weeks prior to the shock and 1 if it is made in the two weeks after the shock. γ_j is a set of user fixed effects and ε_{ijt} is an idiosyncratic error term. Standard errors are heteroskedasticity-robust and are clustered at the user level. β_1 identifies the effect of the network shock on idiosyncratic language use and is the main coefficient of interest.

C. Measuring Language

Idiosyncratic language is measured in a number of ways. Each measure helps to address different threats to the theory and mechanisms. I present these measures in an order that forms a spectrum from highly intuitive/less robust to highly robust/less intuitive.

C.1. Alt-Right Jargon Dictionary

The first of these is a dictionary of Alt-Right jargon sourced from media accounts which helps to ground-truth the study's findings. Specific, documented Alt-Right jargon gives us both some intuition for the later results, and gives us some confidence that the later results are also capturing the most troubling language that one might target. I construct a dictionary of Alt-Right jargon documented in media accounts and the Anti-Defamation League. The specific jargon and sources are in Table IX in the appendix. I include two versions of the measure: a simple count of the number of terms from this dictionary that are included in the comment, and a binary measure of whether the comment includes any terms from the dictionary.

C.2. Drift-Corrected Language Frequency

Next, I measure idiosyncratic Alt-Right language through words frequently used in the Alt-Right subreddit. The dependent variable in each of these tests is based on the top N most frequently used words from comments in the Alt-Right subreddit in the two weeks preceding the network shock. I omit stop words, which are commonly used words that convey no or minimal meaning, such as "the," "a," and"an." Given this top N words dictionary, I (1) count the usage and (2) inclusion of those frequent Alt-Right words in comments in the_donald subreddit. The simple frequency measures generalize the idea of the jargon measure to frequently used words from the Alt-Right subreddit.

Although the specific Alt-Right jargons in the original set of tests are unlikely to substantially change over the time period of our analysis, there is no guarantee that the underlying distribution of words that generated the frequency count measures ought to be stationary over this time period. In order to account for spurious changes in these measures resulting from possible non-stationarity, I construct a set of measures that corrects for the baseline drift in language over the time period. That is, even correcting for the baseline drift in language over this time period, how much more, if any, does Alt-Right language change? I do so by first constructing analogous frequency measures for the_donald subreddit over the same two week pre-shock period. The resulting corrected measure for each comment is

$$\text{Drift-Corrected}(N) = (\text{Top-}N\text{ Alt-Right}) - (\text{Top-}N\text{ TD})$$

where Top-N Alt-Right is the N most frequent words from r/altright pre-shock and Top-N TD is the N most frequent words from r/the_donald pre-shock. This measures how much *more* the Alt-Right language changes compared to the_donald language from that same time period. I include frequency cutoffs at $N = 50, 100, 250,$ and 1000 for both the simple frequency measures and the drift-corrected frequency measures.

C.3. Semantic Distance: Document Embeddings

The final set of measures address the possibility that the more intuitive language measures may capture spurious changes in vocabulary that are unbound from any shift in meaning. However, I hypothesize not just that vocabulary will change, but that this change should come bundled with changes in the meaning of what is being said. Specifically, I predict that the semantic meaning of Alt-Right users' comments should become more similar to the other r/the_donald users' comments after the network shock.

To improve the semantic robustness of our tests, I measure the semantic meaning of users' comments via document embeddings, which are an unsupervised method for measuring semantic meaning in bodies of texts (Le and Mikolov 2014). Document embeddings are an extension to word embeddings (Mikolov et al. 2013) that have shown superior performance in generating paragraph-level similarities. These techniques improve on bag-of-words based models by capturing similarities in meanings based on context. The approach uses shallow neural networks to predict a given word based on the surrounding words. Words are then mapped into a vector space such that semantically similar words are close in this vector space. Thus, if all one does is shift the label on a word, but leaves the meaning intact, they will be identical in embedding space. This is especially important in the empirical setting of Alt-Right echo chambers due to their extensive use of dog whistles. This measurement approach helps to address the concern of changing labels of dog whistles as a method of self-censorship in response to the subreddit ban.

The results presented in this paper are from a model trained with a vector size of 300, which is recommended for word embeddings (Yin and Shen 2018), but I see similar results for vector sizes of 200 and 400. For each comment, I measure the cosine distance in embedding space from each other comment. Specifically, the cosine distance is given by

$$\text{distance}(x,y) = 1 - \frac{x \cdot y}{\|x\|\|y\|}; \quad \text{distance}(x,y) : [0,1]^N \times [0,1]^N \to [0,1]$$

Cosine distance is a measure of the distance between any two vectors of an inner product space via the cosine of the angle between the two vectors. This measure captures the symmetric scalar distance between any two comments in embedding space and reflects the semantic distance between any two comments. The outcome variable for each comment is the mean distance from all other comments in r/the_donald, median distance from all other comments in r/the_donald, and the mean distance from all other posts in r/the_donald from users who did not post in r/Altright. Figure 6 summarizes the process used to create this measure.

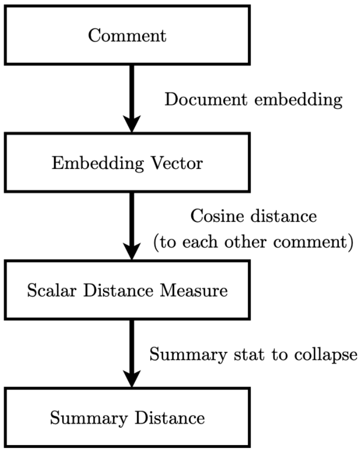


Figure 6. Document Embedding Pipeline. The corpus of M comments (unstructured text) are used to train a document embedding model, which maps comments into a vector space such that semantically similar comments are close together in that space ($[0,1]^{300}$). Cosine distance measures the pairwise semantic distance between comments ($[0,1]^M$). A summary statistic (e.g. mean, median) is then used to map the measure into a scalar distance measure ($[0,1]$).

D. Data & Descriptive Statistics

The data are comment text data and the associated social network from Reddit.com from the r/Altright subreddit and a large subreddit with substantial overlap in members between r/Altright: the_donald. The social network is constructed from interactions in r/Altright and r/the_donald from the two weeks before and the two weeks after the ban. This represents 81,000 unique users (nodes) across the entire sample period, with approximately 45,000 users in the pre-ban period and 65,000 users in the post-ban period. These users make approximately 1.6 million comments in these communities across the entire sample period. Table II presents summary statistics for the full sample.

The edges of the social network are constructed from the users' comments. For each user i, the influence $a_{i,j}$ of user j on i is represented by the proportion of i's comments that are responses that are to j. Denoting a comment k from i to j as $\text{comment}_{i,j,k}$, then

$$a_{i,j} = \frac{\sum_{k \in i} \text{comment}_{i,j,k}}{\sum_{k \in i, v \in V} \text{comment}_{i,v,k}}$$

If i never responds to j, then $a_{i,j} = 0$. This network construction requires no assumptions of symmetry. In practice, $a_{i,j} \neq a_{j,i}$ except in rare circumstances. Across all N users, this produces the weighted, directed, $N \times N$, row-stochastic adjacency matrix A. Considering an edge to exist if the adjacency weight is greater than zero, then the network has approximately 250,000 edges in the pre-ban period and 400,000 edges in the post-ban period.

In order to test changes in language use, I analyze the text content of users' comments. Accordingly, the unit of analysis for all tests presented below is the comment. In order to have an apples-to-apples comparison across periods, the main sample of comments reflects r/Altright users' comments *only* in r/the_donald. Out of 2,463 users who participate in r/Altright, 509 of them also participate in the_donald in both periods. In order to keep the language change tests as conservative as possible, I only include Alt-Right users who have posted in r/the_donald both before and after the r/altright ban. This leaves a sample of 198,698 comments in r/the_donald from 509 r/altright users in the four week period. Table III presents summary statistics for the apples-to-apples sample. Additionally, both the drift-correctly language frequency measure and the semantic distance measure integrate the full sample at various stages of their construction in order to make the relevant comparisons between sub-populations. This sample includes all users from both communities over the sample period and reflects 1.6 million comments from 81,000 users.

IV. Results

There are two sets of predictions that come out of the theoretical model. First, the ban of the Alt-Right subreddit will induce a network shock that transforms the network from a high-echo network into a low-echo network. We can robustly measure this change by the increase in the network conductance between Alt-Right posters and others. Through this shock to insularity

			Pre-ban	Post-ban	Full Sample
Users			44,740	64,735	81,187
	location	altright	2,463	-	2,463
		the_donald	43,022	64,735	79,659
Comments			601,405	1,018,627	1,620,032
	location	altright	35,061	-	35,061
		the_donald	566,344	1,018,627	1,584,971
	by	altright users	102,381	123,767	226,148
		the_donald (only) users	499,024	894,860	1,393,884
Comment Length			26.60	24.38	25.21
			(66.15)	(45.06)	(53.94)
	location	altright	34.25	-	34.25
			(60.59)	-	(60.59)
		the_donald	26.13	24.38	25.01
			(66.45)	(45.06)	(53.77)
	by	altright users	26.79	21.84	24.47
			(60.10)	(36.13)	(50.36)
		the_donald (only) users	26.20	24.44	25.07
			(66.58)	(45.29)	(53.88)

Table II. Summary statistics for full sample of users and comments from the reddit communities r/altright and r/the_donald. Summaries are provided for the pre-ban period, the post-ban period, and for the full sample. Further summary statistic breakdowns are provided for data partitioned by posting location (either altright or the_donald) and data partitioned by user (altright users or the_donald users who never post in altright). Mean comment lengths for each subcategory are shown. Standard deviations are in parentheses below the means.

			Pre-ban	Post-ban	Full Sample
Users			509	509	509
Comments			77,182	123,070	200,252
	location	altright	12,255	-	12,255
		the_donald	64,927	123,070	187,997
Comment Length			26.79	21.84	24.47
			(60.10)	(36.13)	(50.37)
	location	altright	31.86	-	31.86
			(48.99)	-	(48.99)
		the_donald	24.43	21.84	22.97
			(64.48)	(36.13)	(50.51)

Table III. Summary statistics for r/altright users who post in r/the_donald in both periods. Comments are from the reddit communities r/altright and r/the_donald. Summaries are provided for the pre-ban period, the post-ban period, and for the full sample. Further summary statistic breakdowns are provided for data partitioned by posting location (either altright or the_donald). Mean comment lengths for each subcategory are shown. Standard deviations are in parentheses below the means.

reflected in the conductance, those Alt-Right posters will decrease their use of idiosyncratic Alt-Right language wherever they post, including decreasing the amount that they import Alt-Right language into other subreddits.

A. High-Echo to Low-Echo Network

The conductance results compare the social networks of interactions for the two weeks before and after the ban of the Alt-Right subreddit. The network shock should increase the conductance between Alt-Right commenters and others. Indeed, the conductance between Alt-Right commenters and the remaining The_Donald commenters went from 0.48 before the ban to 0.88 after the ban. The shift in conductance seems to have affected this community fairly completely. Figure 7 shows a histogram of Alt-Right commenters' contributions to the group conductance with with non-Alt-Right commenters. Before the ban, the community appears to have been somewhat fragmented, with some users highly integrated with other Alt-Right users (the high-echo portion of the network), while others were highly integrated with the The_Donald community. However, after the ban, this fragmentation coalesced into a much more well-knit whole (only a low-echo network). The shock transforms the network from a high-echo to a low-echo network.

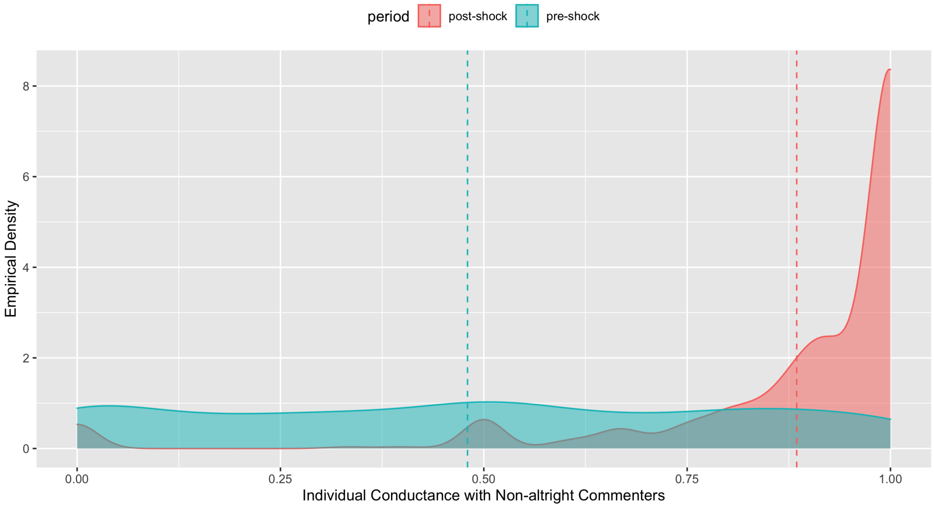


Figure 7. Conductance Contribution Distribution. Empirical distribution of individual Alt-Right users' conductance contribution with the The_Donald subreddit users. The blue distribution is the pre-ban distribution and the orange is the post ban distribution. Individuals are included in the pool if they have ever commented in the Alt-Right subreddit and they have commented in the The_Donald both before and after the Alt-Right ban (508 users). The dotted lines show the total conductance for each period, which is 0.48 for the pre-ban period and 0.88 for the post-ban period.

This can also be seen in the social network itself before and after the ban. Figures 9 and 10 show the reply networks for the two weeks before and after the ban, respectively. For ease of viewing, the networks for both periods show only the largest connected component, which in both cases accounts for over 99% of all users. Users who have ever posted in the Alt-Right subreddit have been colored red, while others have been colored grey. In the pre-shock period, there are clearly two main clusters. While clearly connected, there is a relatively insular cluster of Alt-Right posters who interact much more with each other than with the rest of the community. However, after the shock, the network is dominated by a single cluster that includes Alt-Right posters and non-Alt-Right posters alike. This decrease in insularity of the Alt-Right community helps to demonstrate what the increase in conductance is capturing, and gives face validity to the idea that the shock indeed transforms the network from a high-echo to a low-echo network.

B. The Effect of Social Network on Linguistic Change

The transformation from a high-echo to a low-echo network informs the prediction that the network shock will reduce the amount of idiosyncratic Alt-Right language these users import into other communities. Equation 2 is the regression specification for the following tests. Each test varies the dependent variable y_{ijt} which is the language measure.

B.1. Alt-Right Jargon Dictionary

The first language measure is a dictionary of Alt-Right jargon sourced from media accounts. Table IV presents regression results for both the inclusion and count of Alt-Right jargon in Alt-Right users' posts in r/the_donald. Both measures of specific jargon use fall after the network shock. Alt-Right subreddit commenters decrease their use of this idiosyncratic language in the the_donald subreddit approximately once per two-hundred comments in post-ban period as compared to the pre-ban period. While this may appear to be a small decrease in absolute terms, this reflects a decrease in usage of about 33% with respect to the pre-ban period. Given that the words in this dictionary are so frequently hate speech, this is a meaningful drop. Additionally, this measure helps to ground-truth the study's findings from the following langauge measures.

B.2. Drift-Corrected Language Frequency

The regression results for the Alt-Right language frequency measures for various word rank cutoffs is presented in Table V. All the tests show that use of these words in r/the_donald by former alright posters drops after the network shock. Alt-Right subreddit commenters decrease their use of frequently used language from the Alt-Right subreddit in the the_donald subreddit approximately once per twenty comments in post-ban period as compared to the pre-ban period.

The regression results for the drift-corrected langauge frequency measures for various word rank cutoffs are shown in Table VI. All the tests show that use of these words in r/the_donald by former alright posters drops after the network shock after accounting for the baseline drift in language. Alt-Right subreddit commenters decrease their use of frequently used language from the Alt-Right subreddit in the the_donald subreddit approximately once per twenty-five comments in post-ban period as compared to the pre-ban period beyond the amount that one

	Dependent variable:	
	Alt-Right Jargon (count)	Alt-Right Jargon (incl)
	(1)	(2)
Post-ban	-0.005^{***}	-0.004^{***}
	(0.001)	(0.001)
Constant	0.014^{**}	0.008^{***}
	(0.007)	(0.001)
User FE	Yes	Yes
Observations	198,698	198,698
R^2	0.077	0.076
Adjusted R^2	0.076	0.075
Note:		*p<0.1; **p<0.05; ***p<0.01

Table IV. Change in language use in r/the_donald for Alt-Right users who have posted in r/the_donald both before and after the r/altright ban. This reflects a sample of 198,698 comments in r/the_donald from 508 r/altright users in the four week period. The unit of analysis is the comment. All tests include user-level fixed effects. Standard errors are heteroskedasticity-robust and are clustered at the user level. The dependent variable in (1) is count of words from our Alt-Right jargon dictionary created from media accounts and the ADL. The dependent variable in (2) is a binary measure of whether the comment included any words from that same jargon dictionary. Both tests show a small but significant decrease in the use of these specific jargons. This represents a decrease of about 5 uses of these words per 1000 comments. Given the small size of the jargon dictionary and that it is composed of blatant hate speech, this is a meaningful decrease.

	Dependent variable: Includes Alt-Right Words			
	Top 50	Top 100	Top 250	Top 1000
	(1)	(2)	(3)	(4)
Post-ban	-0.036^{***}	-0.033^{***}	-0.033^{***}	-0.033^{***}
	(0.002)	(0.002)	(0.001)	(0.002)
Constant	0.154^{**}	0.106^{***}	0.060^{***}	0.047^{***}
	(0.064)	(0.035)	(0.018)	(0.007)
User FE	Yes	Yes	Yes	Yes
Observations	198,698	198,698	198,698	198,698
R^2	0.112	0.095	0.086	0.089
Adjusted R^2	0.110	0.093	0.084	0.087
Note:	*p<0.1; **p<0.05; ***p<0.01			

(a) Dependent variables for each regression is a binary measure of whether the comment (in r/the_donald) included a word from the top N most frequently used words in the Alt-Right subreddit.

	Dependent variable: Count of Alt-Right Words			
	Top 50	Top 100	Top 250	Top 1000
	(1)	(2)	(3)	(4)
Post-ban	-0.056^{***}	-0.053^{***}	-0.058^{***}	-0.068^{***}
	(0.003)	(0.003)	(0.002)	(0.003)
Constant	0.189^{**}	0.097^{***}	0.052^{***}	0.072^{***}
	(0.073)	(0.024)	(0.003)	(0.009)
User FE	Yes	Yes	Yes	Yes
Observations	198,698	198,698	198,698	198,698
R^2	0.089	0.078	0.068	0.061
Adjusted R^2	0.087	0.076	0.066	0.059
Note:	*p<0.1; **p<0.05; ***p<0.01			

(b) Dependent variables for each regression is the count for each comment (in r/the_donald) of words from the top N most frequently used words in the Alt-Right subreddit.

Table V. Change in language use in r/the_donald for Alt-Right users who have posted in r/the_donald both before and after the r/altright ban. This reflects a sample of 198,698 comments in r/the_donald from 508 r/altright users in the four week period. The unit of analysis is the comment. All tests include user-level fixed effects. Standard errors are heteroskedasticity-robust and are clustered at the user level. The dependent variables for each regression are (a) a binary measure of whether the comment (in r/the_donald) included a word from the top N most frequently used words in the Alt-Right subreddit, and (b) the count for each comment (in r/the_donald) of words from the top N most frequently used words in the Alt-Right subreddit. All tests show a significant decrease in the inclusion of these words. This represents a decrease of about (a) 3 comments including these words per 100 comments, and (b) 5 mentions of these words per 100 comments.

would expect their language to change compared to their use of frequent words from within the_donald subreddit.

	Dependent variable: Count of Alt-Right Words Over Baseline			
	Top 50	Top 100	Top 250	Top 1000
	(1)	(2)	(3)	(4)
Post-ban	-0.039^{***}	-0.031^{***}	-0.041^{***}	-0.041^{***}
	(0.003)	(0.002)	(0.003)	(0.002)
Constant	0.139^{**}	0.026^{***}	-0.387	0.019^{***}
	(0.057)	(0.003)	(0.240)	(0.001)
User FE	Yes	Yes	Yes	Yes
Observations	198,698	198,698	198,698	198,698
R^2	0.051	0.052	0.033	0.038
Adjusted R^2	0.049	0.050	0.031	0.036
Note:				*p<0.1; **p<0.05; ***p<0.01

Table VI. Change in language use in r/the_donald for Alt-Right users who have posted in r/the_donald both before and after the r/altright ban. This reflects a sample of 198,698 comments in r/the_donald from 508 r/altright users in the four week period. The unit of analysis is the comment. All tests include user-level fixed effects. Standard errors are heteroskedasticity-robust and are clustered at the user level. The dependent variables for each regression are the count for each comment (in r/the_donald) of words from the top N most frequently used words in the Alt-Right subreddit minus the count of words from the top N most frequently used words in r/the_donald. All tests show a significant decrease in the count of these words over the relevant baseline. This represents a decrease of about 5 mentions of these words per 100 comments. This is a sizable decrease.

B.3. Semantic Distance: Document Embeddings

The semantic distance measures allows us to directly test the prediction that the semantic meaning of Alt-Right users' comments should become more similar to the other r/the_donald users' comments after the network shock. Additionally, tests based on this measure speak to concerns over changing labels of dog whistles as a method of self-censorship in response to the subreddit ban. Outside of concerns over strategic label changes, handling dog whistles and shibboleths is important for any empirical study of online echo chambers. Table VII shows the results from these tests. They suggest that Alt-Right users' comments get closer in meaning to the rest of r/the_donald following the network shock.

These results permit us speak to the idea that language use is not fully contextual. In the pre-shock period, Alt-Right posters were frequently importing langauge from the Alt-Right subreddit into the_donald, which is one avenue by which extreme language is ported gradually into less extreme communities. By increasing the conductance between the Alt-Right commenters and non-Alt-Right commenters, the network shock has decreased the volume of jargon imported. The count variables are useful in part because they allow us to generalize this result beyond hate speech and the specific Alt-Right setting, instead suggesting that the effect of the shock on language change is a more general process.

Across measures of language use, these tests consistently show that the transformation of the network from a high-echo to a low-echo network decreases the amount of idiosyncratic Alt-Right language that Alt-Right users import into the_donald, a large and topically-related community. The result is robust to a wide range of idiosyncratic language measures and adjustments for overall changes in language over the time period.

	Dependent variable:		
	Mean Distance	Median Distance	Mean Distance (TD)
	(1)	(2)	(3)
Post-ban	-0.009^{***}	-0.007^{***}	-0.009^{***}
	(0.003)	(0.003)	(0.003)
Constant	0.891^{***}	0.893^{***}	0.892^{***}
	(0.002)	(0.002)	(0.002)
User FE	Yes	Yes	Yes
Observations	198,698	198,698	198,698
R^2	0.068	0.060	0.068
Adjusted R^2	0.066	0.057	0.066
Note:			*p<0.1; **p<0.05; ***p<0.01

Table VII. Change in language use in r/the_donald for Alt-Right users who have posted in r/the_donald both before and after the r/altright ban. This reflects a sample of 198,698 comments in r/the_donald from 508 r/altright users in the four week period. The unit of analysis is the comment. All tests include user-level fixed effects. Standard errors are heteroskedasticity-robust and are clustered at the user level. I train a document embedding model on our corpus of comments, which maps each comment into an embedding space such that semantically similar comments are close together in that space. For a given comment, I calculate the cosine distance between that comment and every other comment in embedding space. The dependent variables for each comment in regression (1) is the mean distance from all other comments in r/the_donald, (2) is the median distance from all other comments in r/the_donald, and (3) is the mean distance from all other all other posts in r/the_donald from users who did not post in r/altright. Each regression shows a significant decrease in the semantic distance between Alt-Right users' comments and the comments of others in the post-ban period.

V. Discussion

Theoretically, I show that one can generalize this idea of the importance of local neighborhoods on group-level language dynamics by focusing on the conductance between those using idiosyncratic vocabulary and others. Because conductance is a group-oriented measure, it captures the relationship of one group to another group that would be difficult to encapsulate through egocentric measures alone (and thus is more appropriate to capture social network echo). Ultimately, the phenomena of culture and echo chambers are theoretically oriented around group influence and relative insularity: to what extent does some group reflect its language back on itself.

The conductance is a spectral property of a network. Spectral measures are among the richest and most informative properties of a network with respect to the dynamics of information flows. This is because spectral measures capture rich *global* structural properties of networks in addition to *local* structural properties.

While this structural richness is not inherently valuable in understanding every phenomena, it is valuable in understanding the dynamics of echo chambers. Across fields, the literature on spectral graph measures have been most influential with respect to communication and insularity (DeMarzo et al. 2003; Golub and Jackson 2010; Becker et al. 2017). However, my theoretical findings suggest that these measures are insufficient on their own in helping us understand the diffusion of language as they omit an important level of aggregation: individuals in our model must communicate through language rather than directly transmitting their beliefs.

The problem of aggregation of human behavior from the micro scale to the macro scale remains one of the central motivating challenges in the social sciences. The extent to which one can abstract away from the medium of communication depends in part on the extent to which the medium of communication affects the social aggregation process. I show that it does indeed affect the aggregation process. Consequently, omitting the medium of communication omits a key step in the process of social diffusion. This means that one cannot directly generalize results around the diffusion of continuous phenomena to non-continuous phenomena like language without carefully thinking through the dynamics of the aggregation process.

While the phenomenon of idiosyncratic language in organizations most obviously applies to echo chambers in online communities, there is a clear link to how these processes relate to the problems faced by managers in other organizations, including firms. The ability to change language via a network rewiring intervention can be a valuable strategic lever for an organization. Language allows one to communicate on two levels. It is the medium that one must use to communicate, so it facilitates the transfer of semantic content. Additionally, it encodes information about the speaker's priorities. Precise words convey that their underlying concepts are more important than some concept in an imprecise word.

Thus, reducing the precision of an idiosyncratic word degrades a speaker's ability to communicate some idiosyncratic concept on both of these levels. Embedding a concept in a less precise word increases the likelihood that the intended semantic content of the speaker's message will be misconstrued, so that the listener understands some other underlying concept from the now more vague word. Conversely, by decreasing the frequency of the word that embeds that idiosyncratic concept (a consequence of reducing its precision), one decreases the likelihood that some other intended meaning will be misinterpreted by the listener as that idiosyncratic concept.

If one believes that accurately and precisely communicating this idiosyncratic concept is undesirable for the organization, for example in the case of hate speech, then decreasing the frequency and precision of that communication is a highly desirable property. Even if the communication is simply unimportant, as is more likely the case in the context of idiosyncratic speech in a firm, then one must recognize that the precision of this unimportant speech comes at the expense of some actually important concept made more ambiguous. By reducing the precision of the idiosyncratic word, one makes space in the language for the more important or relevant concept to take the forefront. Having the organizational language and culture better match the organization's actual problems and circumstances is valuable in enabling them to be accurately communicated. This improves organizational performance and may be a source of competitive advantage.

Some questions may remain about the causal relationships between the social network and the language. First, while the ban of the subreddit does indeed shock the social network of users, it may have other effects. Two effects in particular may affect the causal relationship between the network shock and the language change: self-censorship, and displacement.

The extent of self-censorship hinges on users beliefs about whether the ban is related to the content of their language. We focus on this ban because the subreddit is banned *not* because their conversation or content was against Reddit's rules, but by breaking a rule around revealing personally identifiable information online. This sets the community ban apart from many others online which explicitly claim content as the reason for the ban. This is important because content-related bans may be more likely to bias our results through self-censorship effects, while other rule bans should have minimal effect on self-censorship. Nonetheless, I cannot fully rule out this alternative hypothesis, and it may have a moderating effect on language change.

In order to address displacement effects, I look at this group's language use in another subreddit with substantial (but not complete) overlap in membership and compare it before and after the ban. I find that their language post-ban shifts to reflect that this other community constitutes a larger proportion of what they see when they browse Reddit. This result is in contrast to the prevailing wisdom that such a ban would instead simply displace these users to this other community to continue their conversations. Instead, what I find is the complete opposite. These users are *less* likely to import Alt-Right language into places outside the Alt-Right subreddit after the ban as compared to before.

There is a question of the ultimate equilibrium of such a process that is beyond the scope of this paper. That is, can the new network structure be sustained in the very long run (and how long should one think of that being in the fast moving digital world). This is a key scope condition of this paper. In other words, I only claim that language and culture are partially structurally dependent over the medium term.

However, the sustainability of the equilibrium is significantly influenced by the freedom individuals have in network formation. Thus, organizations with more formal structure, such as firms, may have more success in long-run sustainability through their fiat power over organizational structure. The upside is that the framework has implications beyond online communities to organizations of many forms. Additionally, this may be one of the first times one can measure and test such a model, since online communities allow us to observe both language use and social network structure. Thus, digitization here provides us not only an opportunity to address and possibly to fix the problems it has created, but also an opportunity to better understand our analog world.

These results permit a strong test of the view that individuals are highly social creatures whose beliefs and priorities evolve in the context of communities, including organizations. Moreover, I suggest that language is a tool for transmitting tacit knowledge about users' priorities. The knowledge theory of the firm suggests that the ability to efficiently transmit tacit knowledge is an important source of competitive advantage (Kogut and Zander 1992). Thus, I suggest that the suitability of an organization's language to its context may be one such mechanism by which some firms outperform others.

Knowledge development, and especially knowledge transfer, has a central linguistic basis. Both of language's dual communication channels of content and encoded cultural priorities are valuable to a firm. Language makes possibly tacit knowledge that is relevant to the firm easier and faster to refer to (Weber and Camerer 2003), and it encodes organizational priorities as it is used. In other words, language is not only communicating on two levels, but is communicating different types of tacit knowledge on each of those levels. Thus, a suitable language can be a source of significant competitive advantage for firms, especially those for whom tacit or difficult-to-encode knowledge is key to their business efforts.

Emphasizing the dynamic process of language evolution implies that organizations ought to be able to influence the language, and make strategic choices about their languages and cultures. I propose that organizations can leverage the endogeneity of language to organizational structure by making strategic changes to organizational design. If language is what flows through the pipes of a social network, and language is strategically valuable, changing the size and placement of those pipes is a strategic decision.

VI. Conclusion

The proliferation of online communities over the past several decades has created a novel set of both concerns and opportunities. Digitization has permanently shifted the landscape of information aggregation. In particular, many are concerned about the emergence of echo chambers. At best, echo chambers make information and news aggregation difficult for even concientious consumers, and at worst they lead to the radicalization of the members of these online communities.

On the other hand, the proliferation of online digital communities and communications also presents us with an opportunity to better understand the mechanisms and processes underlying this shift in information aggregation. At their core, echo chambers are a phenomonon at the intersection of organizational structure and language. Therefore, any attempt to evaluate solutions to the problem of echo chambers must take both langauge and social structure into account. I present a model that unites these perspectives and demonstrates how language can be endogenous to network structure.

By taking advantage of a natural experiment on one of the largest online communities and contemporary advances in natural language processing techniques, I am able to evaluate one solution to the problem of echo chambers via the partial endogeneity of language to network structure. The natural experiment shocks the network adjacency weights of the members of an Alt-Right community by shutting down the community, but not banning any of the members.

I demonstrate the shift in language use in several ways, first via Alt-Right jargon identified in the news media which helps to grant some face validity to our approach. Second, I demonstrate a reduction in the most frequently used terms from the Alt-Right subreddit. Finally, and most informatively, I show a decrease in semantic difference captured via word embeddings. This approach is the most robust because (a) it is the most wholistic, and (b) it will capture shifting dog-whistles which have the same underlying meaning. This type of robustness is especially important in our empirical setting. That the tests are unambiguous in this case is strong support for our hypothesis that these users have changed the semantic content of their comments.

This result suggests that intervention in online radicalized communities can prevent individuals in those communities from becoming further radicalized, and indeed can soften their radicalization. Instead, this suggests that they are at least partially a product of their community, and that changing their community encourages them to change their priorities as expressed in their language. This may be valuable in changing their minds, but perhaps even more valuable in affecting those who are newcomers to the community who now end up seeing fewer uses of extreme language. This is consistent with our modeling predictions that relatively small changes to the network's structure can have large effects on aggregate language use.

Unlike many models in the network diffusion literature, I incorporate the medium through which individuals communicate. This is valuable not only for versimilitude, but also because it allows us to directly hypothesize on that same observable medium of language. This allows us to test on observable language use directly, instead of having to make appeals to either political opinions as a yardstick for beliefs or by having to use stylized parameter estimation tasks instead of natural language use.

Language is a central aspect of culture in every organization, but especially online. In an online setting, language *is* the culture. In other types of organizations where culture may be broader than language, culture is nonetheless accessible in language. It is invaluable to show how culture evolves in an organization, and how it can be influenced by the structure of the organization. Advances in natural language processing—like those used in this paper—make culture measureable and enable the analysis of these processes.

References

Becker, J., D. Brackbill, and D. Centola (2017). Network dynamics of social influence in the wisdom of crowds. *Proceedings of the National Academy of Sciences 114*(26), E5070–E5076.

Beer, C. (2020). The rise of online communities. [GlobalWebIndex, Online].

Biddle, S. (2015, August). Reddit (finally) bans coontown. [Online].

Boxell, L., M. Gentzkow, and J. M. Shapiro (2017, March). Is the internet causing political polarization? evidence from demographics. Working Paper 23258, National Bureau of Economic Research.

Centola, D. and A. Baronchelli (2015). The spontaneous emergence of conventions: An experimental study of cultural evolution. *Proceedings of the National Academy of Sciences 112*(7), 1989–1994.

Centola, D., J. Becker, D. Brackbill, and A. Baronchelli (2018). Experimental evidence for tipping points in social convention. *Science 360*(6393), 1116–1119.

Chandrasekharan, E., U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, and E. Gilbert (2017). You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction 1*(CSCW), 31.

Chen, M. K. (2013, April). The effect of language on economic behavior: Evidence from savings rates, health behaviors, and retirement assets. *American Economic Review 103*(2), 690–731.

Clement, J. (2020). Number of internet users worldwide from 2005 to 2019. [Statista, Online].

Crémer, J., L. Garicano, and A. Prat (2007). Language and the theory of the firm*. *The Quarterly Journal of Economics 122*(1), 373.

Danescu-Niculescu-Mizil, C., R. West, D. Jurafsky, J. Leskovec, and C. Potts (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*, pp. 307–318.

DeGroot, M. H. (1974). Reaching a consensus. *Journal of the American Statistical Association 69*(345), 118–121.

DellaVigna, S. and E. Kaplan (2007, 08). The Fox News Effect: Media Bias and Voting*. *The Quarterly Journal of Economics 122*(3), 1187–1234.

DeMarzo, P. M., D. Vayanos, and J. Zwiebel (2003). Persuasion bias, social influence, and unidimensional opinions*. *The Quarterly Journal of Economics 118*(3), 909.

Doyle, G., A. Goldberg, S. Srivastava, and M. C. Frank (2017). Alignment at work: Using language to distinguish the internalization and self-regulation components of cultural fit in organizations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 603–612.

Durante, R. and B. Knight (2012, 06). Partisan Control, Media Bias, and Viewer Responses: Evidence from Berlusconi's Italy. *Journal of the European Economic Association 10*(3), 451–481.

Fink, C. (2018). Dangerous speech, anti-muslim violence, and facebook in myanmar. *Journal of International Affairs 71*(1.5), 43–52.

Flaxman, S., S. Goel, and J. M. Rao (2013). Ideological segregation and the effects of social media on news consumption. *Available at SSRN 2363701.*

Gentzkow, M. and J. M. Shapiro (2010). What drives media slant? evidence from u.s. daily newspapers. *Econometrica 78*(1), 35–71.

Gentzkow, M. A. and J. M. Shapiro (2004, September). Media, education and anti-americanism in the muslim world. *Journal of Economic Perspectives 18*(3), 117–133.

Golub, B. and M. O. Jackson (2010). Naive learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics 2*(1), 112–49.

Greenstein, S., G. Gu, and F. Zhu (2020). Ideology and composition among an online crowd: Evidence from wikipedians. *Management Science - forthcoming.*

Greenstein, S. and F. Zhu (2012, May). Is wikipedia biased? *American Economic Review 102*(3), 343–48.

Greenstein, S. and F. Zhu (2016). Open content, linus' law, and neutral point of view. *Information Systems Research 27*(3), 618–635.

Greenstein, S. and F. Zhu (2018). Do experts or crowd-based models produce more bias? evidence from encyclopedia britannica and wikipedia. *MIS Quarterly 42*(3).

Grinberg, N., K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer (2019). Fake news on twitter during the 2016 u.s. presidential election. *Science 363*(6425), 374–378.

Jeppesen, L. B. and L. Frederiksen (2006). Why do users contribute to firm-hosted user communities? the case of computer-controlled music instruments. *Organization Science 17*(1), 45–63.

Koçak, Ö. and M. Warglien (2020). When three'sa crowd: how relational structure and social history shape organizational codes in triads. *Journal of Organization Design 9*(1), 1–27.

Kogut, B. and U. Zander (1992). Knowledge of the firm, combinative capabilities, and the replication of technology. *Organization Science 3*(3), 383–397.

Kovacs, B. and A. M. Kleinbaum (2020). Language-style similarity and social networks. *Psychological science 31*(2), 202–213.

Krackhardt, D. and R. N. Stern (1988). Informal networks and organizational crises: An experimental simulation. *Social psychology quarterly*, 123–140.

Krupnik, I. and W. U. Weyapuk (2010). *Qanuq Ilitaavut: "How We Learned What We Know" (Wales Inupiaq Sea Ice Dictionary)*, pp. 321–354. Dordrecht: Springer Netherlands.

Lancaster, L. (2017, February). Reddit shuts down 'alt-right' subreddit. [Online; posted 1-February-2017].

Larcinese, V., R. Puglisi, and J. M. Snyder (2011). Partisan bias in economic news: Evidence on the agenda-setting behavior of u.s. newspapers. *Journal of Public Economics 95*(9), 1178 – 1189. Special Issue: The Role of Firms in Tax Systems.

Lawrence, E., J. Sides, and H. Farrell (2010). Self-segregation or deliberation? blog readership, participation, and polarization in american politics. *Perspectives on Politics 8*(1), 141–157.

Le, Q. and T. Mikolov (2014). Distributed representations of sentences and documents. In *International conference on machine learning*, pp. 1188–1196.

Lu, Z., J. Wahlström, and A. Nehorai (2018). Community detection in complex networks via clique conductance. *Scientific reports 8*(1), 1–16.

Ma, M. and R. Agarwal (2007). Through a glass darkly: Information technology design, identity verification, and knowledge contribution in online communities. *Information Systems Research 18*(1), 42–67.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119.

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you.* Penguin UK.

Qiu, L. and S. Kumar (2017). Understanding voluntary knowledge provision and content contribution through a social-media-based prediction market: A field experiment. *Information Systems Research 28*(3), 529–546.

Quattrociocchi, W., A. Scala, and C. R. Sunstein (2016). Echo chambers on facebook. *Available at SSRN 2795110*.

Reddit (2019, January). Community details for r/the_donald. [Online].

Shore, J., J. Baek, and C. Dellarocas (2016). Network structure and patterns of information diversity on twitter. *Boston University Questrom School of Business Research Paper* (2813342).

Srivastava, S. B. and A. Goldberg (2017). Language as a window into culture. *California Management Review 60*(1), 56–69.

Srivastava, S. B., A. Goldberg, V. G. Manian, and C. Potts (2018). Enculturation trajectories: Language, cultural adaptation, and individual outcomes in organizations. *Management Science 64*(3), 1348–1364.

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

Van Alstyne, M. and E. Brynjolfsson (2005). Global village or cyber-balkans? modeling and measuring the integration of electronic communities. *Management Science 51*(6), 851–868.

Van Laarhoven, T. and E. Marchiori (2016). Local network community detection with continuous optimization of conductance and weighted kernel k-means. *The Journal of Machine Learning Research 17*(1), 5148–5175.

Weber, R. A. and C. F. Camerer (2003, April). Cultural conflict and merger failure: An experimental approach. *Manage. Sci. 49*(4), 400–415.

Wernerfelt, B. (2004, September). Organizational Languages. *Journal of Economics & Management Strategy 13*(3), 461–472.

Xu, S. X. and X. Zhang (2013). Impact of wikipedia on market information environment: Evidence on management disclosure and investor reaction. *Mis Quarterly*, 1043–1068.

Yang, J. and J. Leskovec (2015). Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems 42*(1), 181–213.

Yin, Z. and Y. Shen (2018). On the dimensionality of word embedding. In *Advances in Neural Information Processing Systems*, pp. 887–898.

A Appendix

A. Proofs

This section contains all of the formal definitions and proofs referred to in the main body of the paper.

Lemma 1 (Conductance of a network cut with a row-stochastic network)**.** *Suppose we have a social network represented by the graph $G(V, E)$ which is defined over a set of vertices, V, and adjacency matrix, A. Next, suppose we have a fixed subset $S \subset V$. Then, if the adjacency matrix A is row-stochastic*

$$\phi(S, A') > \phi(S, A) \Leftrightarrow \sum_{i \in S, j \in \bar{S}} a'_{ij} > \sum_{i \in S, j \in \bar{S}} a_{ij}$$

Proof. Fix a cut $S = V \setminus \bar{S}$ of a social network (so that we are "holding the people constant"). However, we will not fix the adjacency matrix. Then the conductance of this cut is given by

$$\phi(S) = \frac{\sum_{i \in S, j \in \bar{S}} a_{ij}}{\min(a(S), a(\bar{S}))}$$

Since A is row stochastic, this implies that

$$a(S) = \sum_{i \in S} \sum_{j \in V} a_{ij} = \sum_{i \in S} \left[\sum_{j \in V} a_{ij} \right] = \sum_{i \in S} 1 = |S|$$

$$\Rightarrow \min\left(a(S), a\left(\bar{S}\right)\right) = \min\left(|S|, \left|\bar{S}\right|\right)$$

Thus, for a fixed cut,

$$\phi(S, A') > \phi(S, A) \Leftrightarrow \sum_{i \in S, j \in \bar{S}} a'_{ij} > \sum_{i \in S, j \in \bar{S}} a_{ij}$$

□

Definition 2 (Reconstitution-consistent priority estimate)**.** *Given any partition language L, we define the set $\mathcal{P}$ as*

$$\mathcal{P} = \left\{ p \middle| \min_{L} \sum_{k=1}^{K} p_k^i d(n_k) \right\}$$

which is the set of priorities that could yield the langugage L.

Then, an estimate $\hat{p}$ is reconsitution consistent if $\hat{p} \in \mathcal{P}$.

Figures **??** and 8 illustrate this property. This is a form of bounded rationality oriented around 'putting things back together again.' This means that individuals can choose any estimate of priorities given a neighbor's language as long as they can then use those priorities to generate the language that they observe. Effectively, the mapping from priorities to language is surjective, and we only require that individuals respect this surjectivity in their estimates. Note also that this nests the Bayesian estimate of language if we assume normally distributed priorities, which for one period is the mean of the set of possible priorities. Thus, the boundedly rational learning nests the fully rational case. Importantly, they can even choose an estimate of these underlying priorities that is most flattering or favorable to their own priorities, as long as it respects this reconstitution property.

Table VIII. Model Assumptions

1. Suppose we have a social network represented by the graph $G(V, A)$ which is defined over a set of vertices, V, and adjacency matrix, A. (Network structure)
2. Indiviudals construct and output optimal partition langauges based on their priorities. Word decoding costs, d, are strictly increasing and convex. (Partition language)
3. We can find a fixed subset $S \subset V$ with the property that there exists a focal word $w' \in C$ such that
$$p_x^i \in w', p_x^j \in w'' \quad \forall i \in S, \forall j \in \bar{S} = V \setminus S$$
and $\exists w_a, w_b \in C$
$$|w'| < |w_a| < |w_b| < |w''|$$
That is, some concept p_x is embedded in a more specific word for those in S than for those in $\bar{S}$ (and those words aren't "too close" in specificity).
4. Individuals make "very weakly rational" inferences about their neighbors' priorities given the observed language. These inferences must respect the reconstitution property, so that individuals may make any inferenece on the underlying priorities in a given word so long as that estimate of priorities yields the language that they observe. (Language inference learning)
5. A is row-stochastic. Individuals update their priorities via weighted averaging with their neighbors, and individuals do not fully ignore information from any neighbor. (Penalized DeGroot learning)
6. Timing as described in Table I.

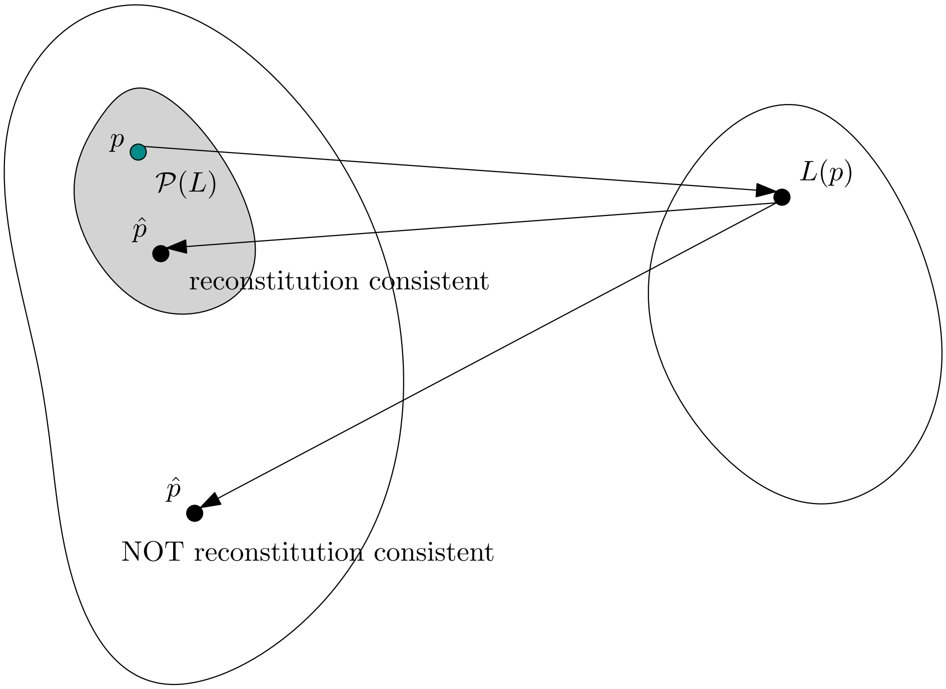


Figure 8. Reconstition consistent estimates

Proposition 1 (Change network, change priorities)**.** *Suppose we satisfy the assumptions described in Table VIII. Then there always exists A' such that $p_x^{i'} < p_x^i \; \forall i \in S$.*

That is, we can always find a network shock $A \to A'$ that will reduce a focal idiosyncratic priority that is embedded in a more specific word for S than for $\bar{S}$.

Moreover, any such A' must also increase the conductance between S and $\bar{S}$.

Proof. First, by assumption (1), individuals in G solve the optimal partition langauge problem given their priorities. This is a mapping from the priorities vector to a language partition that satisfies

$$\min_{L_i} D(L_i, p_i) = \min_{L_i} \sum_{k=1}^{K} p_k^i d(n_k)$$

where d is strictly increasing and convex in the breadth of a given word, and $p_k^i = \sum_{x \in k} p_x^i$.

Next, by assumption (2), we can find a fixed subset $S \subset V$ with the property that there exists a focal word $w' \in C$ such that

$$p_x^i \in w', p_x^j \in w'' \quad \forall i \in S, \forall j \in \bar{S} = V \setminus S$$

and $\exists w_a, w_b \in C$

$$|w'| < |w_a| < |w_b| < |w''|$$

That is, some concept p_x is embedded in a more specific word for those in S than for those in $\bar{S}$.

Next each individual outputs their language for the period to each of their neighbors. Specifically, some individual i in the network observes the language of some other person j whenever $a_{ij} > 0$ for a_{ij}th element of A.

By assumption (3), individuals may make any inference on the underlying priorities in a given word so long as they can construct some ordering of priorities that can be made to fit in the size of the words they observe from their neighbors. We denote this inference from language back to priorties by $g : L \to p$. Recalling from Prop 1 of CGP 2007 that in any optimal language

$$|w_i| > |w_j| \Rightarrow p_i \leq p_j \forall i \in w_i, j \in w_j$$

then if there are

$$|w'| < |w_a| < |w_b| < |w''|$$

then

$$\min_{i \in w'} p_i = \frac{\sum_{j \in w_a} p_j}{|w_a|}$$

and similarly

$$\max_{k \in w''} p_k = \frac{\sum_{m \in w_b} p_m}{|w_b|}$$

Since $|w_a| < |w_b|$ and $p_i \leq p_j \forall i \in w_b, j \in w_a$ then we can conclude that

$$\min_{i \in w'} p_i = \frac{\sum_{j \in w_a} p_j}{|w_a|} > \frac{\sum_{m \in w_b} p_m}{|w_b|} = \max_{k \in w''} p_k$$

Thus any $g : L \to p$ such that the inference does not violate the size of the words will find that $g(p_{x,S}) > g(p_{x,\bar{S}})$. Note that this nests the Bayesian estimate of p for one period under the assumption of Normally distributed signals (which is the mean).

Next, by assumption (4) A is row-stochastic so that for each individual $i \in V$ and $\forall p_x \in p$ their updated priorities $p'_{x,i}$ are given by

$$p'_{x,i} = g(p'_x) \cdot a'_i = g\left(p_{x,S}\right) \cdot a'_{i,S} + g\left(p_{x,\bar{S}}\right) \cdot a'_{i,\bar{S}}$$

For simplicity, let us denote

$$g_x = g\left(p_{x,S}\right), \bar{g}_x = g\left(p_{x,\bar{S}}\right)$$

so that rewriting we have

$$p'_{x,i} = g_x \cdot a'_{i,S} + \bar{g}_x \cdot a'_{i,\bar{S}} = \sum_{j \in S} g_{x,j} a'_{ij} + \sum_{j \in \bar{S}} \bar{g}_{x,j} a'_{ij} \tag{3}$$

We have just shown that $g(p_{x,S}) > g(p_{x,\bar{S}})$ whenever individuals are at least "very weakly rational." Then for every $i \in S$,

$$\sum_{j \in \bar{S}} a'_{ij} > \sum_{j \in \bar{S}} a_{ij} \Rightarrow p_x^{i'} < p_x^i$$

As an addendum (from assumption 5), this works even when we include a "stubbornness" penalty term as a weighted combination between the two groups. That is, even if $\exists \lambda \in (0, 2)$ such that

$$p'_{x,i} = \lambda g_x \cdot a'_{i,S} + (2 - \lambda) \bar{g}_x \cdot a'_{i,\bar{S}} = \lambda \sum_{j \in S} g_{x,j} a'_{ij} + (2 - \lambda) \sum_{j \in \bar{S}} \bar{g}_{x,j} a'_{ij}$$

then we can still choose an A' that satisfies the inequality, specifically whenever

$$(2 - \lambda) \sum_{j \in \bar{S}} a'_{ij} > \sum_{j \in \bar{S}} a_{ij} \Rightarrow p_x^{i'} < p_x^i$$

This completes the proof.

□

Proposition 2 (Change network, change langauge)**.** *Suppose we satisfy the assumptions described in Table VIII. Then there always exists A' such that*

$$w'(t_1) = \{w|p_x^i \in w, A\} \neq \{w|p_x^i \in w, A'\} = w'(t_2)\ \forall i \in S$$

That is, we can always find a network shock $A \to A'$ that will break up or reduce the specificity of a focal idiosyncratic word that is more specific for S than for $\bar{S}$.

Moreover, any such A' must also increase the conductance between S and $\bar{S}$.

Proof. First, by assumption (1), individuals in G solve the optimal partition langauge problem given their priorities. This is a mapping from the priorities vector to a language partition that satisfies

$$\min_{L_i} D(L_i, p_i) = \min_{L_i} \sum_{k=1}^{K} p_k^i d(n_k)$$

where d is strictly increasing and convex in the breadth of a given word, and $p_k^i = \sum_{x \in k} p_x^i$.

Next, by assumption (2), we can find a fixed subset $S \subset V$ with the property that there exists a focal word $w' \in C$ such that

$$p_x^i \in w', p_x^j \in w'' \quad \forall i \in S, \forall j \in \bar{S} = V \setminus S$$

and $\exists w_a, w_b \in C$

$$|w'| < |w_a| < |w_b| < |w''|$$

That is, some concept p_x is embedded in a more specific word for those in S than for those in $\bar{S}$.

Next each individual outputs their language for the period to each of their neighbors. Specifically, some individual i in the network observes the language of some other person j whenever $a_{ij} > 0$ for a_{ij}th element of A.

By assumption (3), individuals may make any inference on the underlying priorities in a given word so long as they can construct some ordering of priorities that can be made to fit in the size of the words they observe from their neighbors. We denote this inference from language back to priorties by $g : L \to p$. To avoid too much repition, we can follow a similar argument from Prop. 1 to find that any $g : L \to p$ such that the inference does not violate the size of the words will find that $g(p_{x,S}) > g(p_{x,\bar{S}})$. Note that this nests the Bayesian estimate of p for one period under the assumption of Normally distributed signals (which is the mean).

Next, (again by Prop. 1) we know from equation 3 that for individual $i \in V$ and for all x that estimate of p_x' is given by

$$p_{x,i}' = g_x \cdot a_{i,S}' + \bar{g}_x \cdot a_{i,\bar{S}}' = \sum_{j \in S} g_{x,j} a_{ij}' + \sum_{j \in \bar{S}} \bar{g}_{x,j} a_{ij}'$$

There are two cases we now need to consider to achieve $\{w|p_x^i \in w, A\} \neq \{w|p_x^i \in w, A'\}\ \forall i \in S$. The first is the case where we reconstitute w to change its meaning (swapping elements in w with elements from other words). The second is the case where we is the case where we increase the breadth of w (making it less specific).

1. For this first case, we can change the meaning of w' by changing its contents while maintaining the size of the word. For this to happen, at least two priorities must swap places. This happens as soon as the smallest element of w' swaps with largest element of some other w_z, here:

$$p_k d(n_k) + p_z d(n_z) < [p_k + \max_p p_z - \min_p p_k] d(n_k) + [p_z - \max_p p_z + \min_p p_k] d(n_z)$$

$$\Rightarrow [p_k d(n_k) + p_z d(n_z)] - [[p_k + \max_p p_z - \min_p p_k]d(n_k) + [p_z - \max_p p_z + \min_p p_k]d(n_z)] > 0$$

$$\Rightarrow -\max_p p_z d(n_k) + \min_p p_k d(n_k) + \max_p p_z d(n_z) - \min_p p_k d(n_z) > 0$$

$$\Rightarrow [\min_p p_k - \max_p p_z]d(n_k) > [\min_p p_k - \max_p p_z]d(n_z)$$

However, since we know that $d(n_z) > d(n_k)$, then

$$\Rightarrow [\min_p p_k - \max_p p_z]d(n_k) > [\min_p p_k - \max_p p_z]d(n_z) \Leftrightarrow \min_p p_k - \max_p p_z < 0$$

By assumption (3) we know that $g(p_{x,S}) > g(p_{x,\bar{S}})$ and by assumption (2) we know that

$$|w'| < |w_a| < |w_b| < |w''|$$

which implies that

$$g(p_{x,S}) \geq \max_{j \in w_a} p_j > g(p_{x,\bar{S}})$$

Then we can choose $w_z = w_a$. Then there always exists some A (trivially where $\sum_{j \in S} a'_{ij} = 0, \sum_{j \in \bar{S}} a'_{ij} = 1$) such that the inequality is satisfied, and the word is broken up.

2. For the second case, we can increase the breadth of w (making it less specific). We can do this by shrinking the elements of z and/or increasing the size of some smaller element in a broader word. We acheive this as soon as

$$p_k d(n_k) + p_z d(n_z) > [p_k + \max_p p_z]d(n_k + 1) + [p_z - \max_p p_z]d(n_z - 1)$$

$$\Rightarrow p_z[d(n_z) - d(n_z - 1)] + \max_p p_z[d(n_z - 1) - d(n_k + 1)] > p_k[d(n_k + 1) - d(n_k)]$$

Since d is convex, we know that

$$d(n_z) - d(n_z - 1) \geq d(n_k + 1) - d(n_k)$$

Additionally, since (by assumption 2) $n_k - n_z \geq 2 \Rightarrow d(n_z - 1) - d(n_k + 1) \geq 0$, so that all of the action in determining when it crosses over is in the relative size of the p.

By assumption (3) we know that $g(p_{x,S}) > g(p_{x,\bar{S}})$ and by assumption (2) we know that

$$|w'| < |w_a| < |w_b| < |w''|$$

which implies that

$$g(p_{x,S}) \geq \max_{j \in w_a} p_j > g(p_{x,\bar{S}})$$

Then we can choose $w_z = w_a$. Then there always exists some A (trivially where $\sum_{j \in S} a'_{ij} = 0, \sum_{j \in \bar{S}} a'_{ij} = 1$) such that the inequality is satisfied, and the word is broken up.

□

Proposition 3 (Clumpy/Discontinuous Language Change). *Given any partition language, L, let us define the set $\mathcal{P}$ as*

$$\mathcal{P} = \left\{ p \middle| \min_L \sum_{k=1}^{K} p_k^i d(n_k) \right\}$$

which is the set of priorities that could yield the langugage L.

Then there always exists some $p^* \in \mathcal{P}$ such that $\forall \varepsilon > 0$

$$(p^* - \varepsilon, p^* + \varepsilon) \cap \mathcal{P} \neq \emptyset \text{ and } (p^* - \varepsilon, p^* + \varepsilon) \cap \mathcal{P}^c \neq \emptyset$$

so that every neighborhood of p^* contains at least one point in $\mathcal{P}$ and not in $\mathcal{P}$.

That is, a change in language need not be preceded by a large change in priorities.

Proof. First, note that $p \in \mathcal{P} \Leftrightarrow \forall k, z \in L, \forall x \in k$

$$d(n_k)p_k + d(n_z)p_z \leq d(n_k - 1)(p_k - p_x) + d(n_z + 1)(p_z + p_x) \tag{4}$$

since by definition of $\mathcal{P}$ the language L solves $\min_L \sum_{k=1}^{K} p_k^i d(n_k)$ given p.

Now consider the mapping

$$G(p) = [d(n_k)p_k + d(n_z)p_z] - [d(n_k - 1)(p_k - p_x) + d(n_z + 1)(p_z + p_x)]$$

G is linear and hence continuous in p. Now choose p^* such that $G(p^*) = 0$. Then $\forall \varepsilon > 0$ we know that

$$(p^* - \varepsilon, p^* + \varepsilon) \cap \mathcal{P} \neq \emptyset$$

since by construction $p^* \in (p^* - \varepsilon, p^* + \varepsilon)$ and $G(p^*) = 0$ satisfies the inquality in equation 4 so that $p^* \in \mathcal{P}$.

It will be easiest to consider the intersections with $\mathcal{P}^c$ in cases.

1. First, consider the case (1) where $n_k = n_z = n$, so that the width of these two words is the same. This implies that for p^*

$$d(n)p_k - d(n-1)(p_k - p_x) = d(n+1)(p_z + p_x) - d(n)p_z$$

$$\Rightarrow p_k[d(n) - d(n-1)] = p_z[d(n+1) - d(n)] + p_x[d(n+1) - d(n-1)]$$

By the convexity of d we know that

$$d(n+1) - d(n-1) > d(n+1) - d(n) > d(n) - d(n-1)$$

Recall that by definition

$$p_k = \sum_{j \in w_k} p_j = \left[\sum_{j \in w_k \setminus \{x\}} p_j \right] + p_x$$

Within the neighborhood $(p^* - \varepsilon, p^* + \varepsilon)$ select $p * -\varepsilon_x/2$, so that we travel from p^* a distance of $\varepsilon/2$ along the dimension of p_x. This means that $p_z[d(n+1) - d(n)]$ will be unaffected, while p_k and p_x both increase by the same amount, $\varepsilon/2$. However, since

$$d(n+1) - d(n-1) > d(n) - d(n-1)$$

we can conclude that $G(p^* - \varepsilon_x/2) > 0$. Thus, for this case

$$(p^* - \varepsilon, p^* + \varepsilon) \cap \mathcal{P}^c \neq \emptyset$$

2. Now consider the case (2) where $n_k > n_z$. By convexity of d we know that

$$d(n_k) - d(n_k - 1) \geq d(n_z + 1) - d(n_z)$$

Then, for p^*, we know that

$$p_k[d(n_k) - d(n_k - 1)] + d(n_k - 1)p_x = p_z[d(n_z + 1) - d(n_z)] + d(n_z + 1)p_x \tag{5}$$

(a) If (case 2.a) $n_k = n_z + 1$ then

$$p_k[d(n_k) - d(n_k - 1)] + d(n_k - 1)p_x = p_z[d(n_k) - d(n_k - 1)] + d(n_k)p_x$$

$$\Rightarrow p_k[d(n_k) - d(n_k - 1)] = p_z[d(n_k) - d(n_k - 1)] + p_x[d(n_k) - d(n_k - 1)]$$

since $d(n_k) - d(n_k - 1)$

$$\Rightarrow p_k = p_z + p_x$$

Recall that by definition

$$p_z = \sum_{j \in w_z} p_j = \left[\sum_{j \in w_z \setminus \{y\}} p_j \right] + p_y$$

where withouut loss of generality p_y chosen among those in w_z. Within the neighborhood $(p^* - \varepsilon, p^* + \varepsilon)$ select $p * -\varepsilon_y/2$, so that we travel from p^* a distance of $\varepsilon/2$ along the dimension of p_y. Clearly then $G(p^* - \varepsilon_y/2) > 0$. Thus, for this case

$$(p^* - \varepsilon, p^* + \varepsilon) \cap \mathcal{P}^c \neq \emptyset$$

(b) For case 2.b consider $n_k \geq n_z + 2$. In this case we know by the convexity of d that

$$d(n_k - 1) \geq d(n_z + 1)$$

then

$$p_k[d(n_k) - d(n_k - 1)] + p_x[d(n_k - 1) - d(n_z + 1)] = p_z[d(n_z + 1) - d(n_z)]$$

Choose some element $p_y \neq p_x \in w_k$, then $G(p^* - \varepsilon_y/2) > 0$ following the same logic as above.

3. Next, consider the case where $n_z > n_k$. Then by convexity of d

$$d(n_z + 1) - d(n_z) > d(n_k) - d(n_k - 1)$$

and

$$d(n_z + 1) > d(n_k - 1)$$

Returning again to equation 5 we see that for p^*

$$p_k[d(n_k) - d(n_k - 1)] + d(n_k - 1)p_x = p_z[d(n_z + 1) - d(n_z)] + d(n_z + 1)p_x$$

$$\Rightarrow p_k[d(n_k) - d(n_k - 1)] = p_z[d(n_z + 1) - d(n_z)] + p_x[d(n_z + 1) - d(n_k - 1)]$$

Again choose some element $p_y \neq p_x \in w_k$, then $G(p^* - \varepsilon_y/2) > 0$ following the same logic as above.

Thus, for every case of the relationship of n_k and n_z we can find an element of the neighborhood $(p^* - \varepsilon, p^* + \varepsilon)$ such that

$$(p^* - \varepsilon, p^* + \varepsilon) \cap \mathcal{P}^c \neq \emptyset$$

This completes the proof.

□

B. Additional Figures & Tables

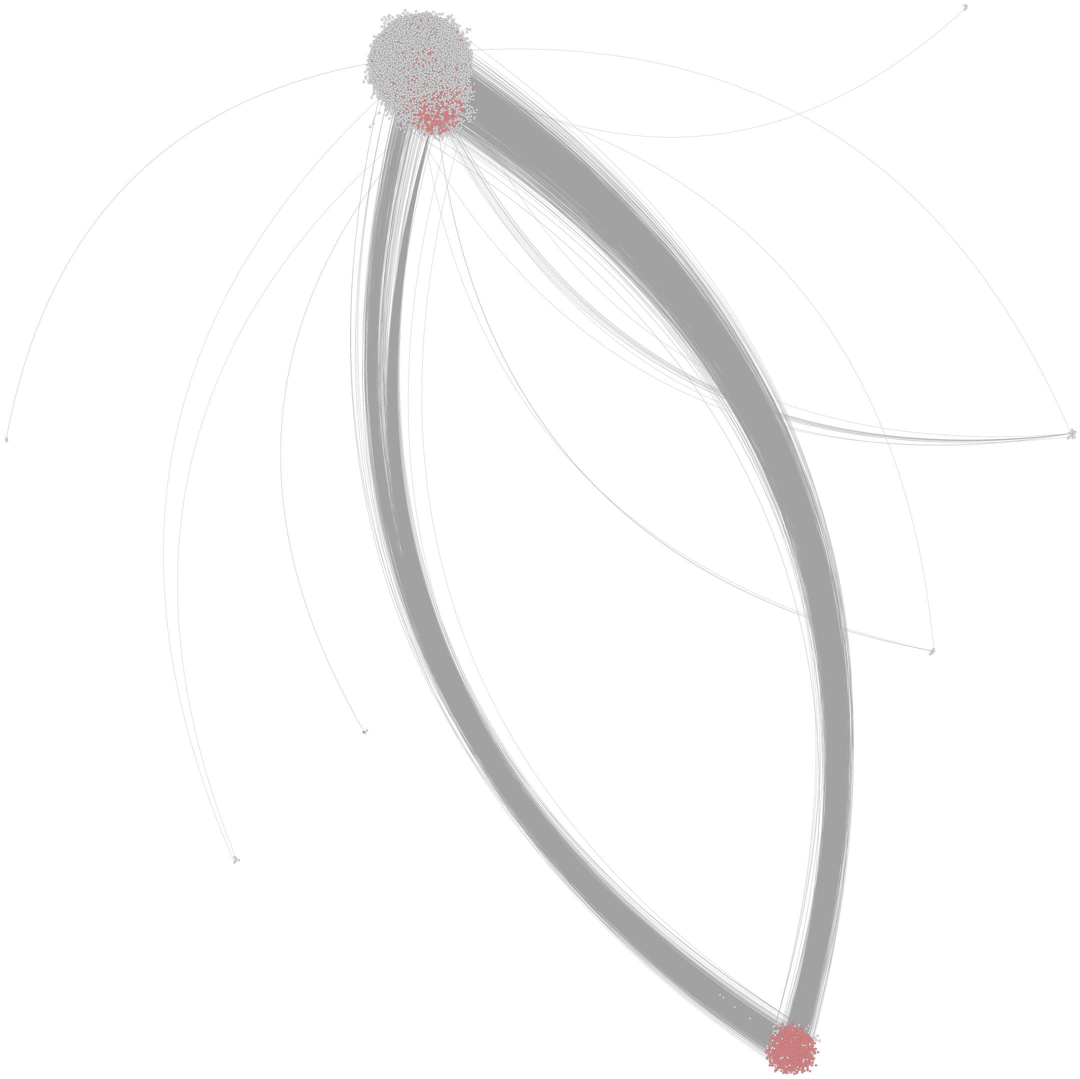


Figure 9. Pre-shock reply network. The reply network for the two weeks prior to the Alt-Right ban consists of approximately 35,000 nodes (users) and 250,000 edges. This is the largest connected component of the complete reply network. It reflects over 99% of users for that period. Users who have ever posted in the Alt-Right subreddit are represented by red nodes, while others are represented by grey nodes. While some of the red nodes are densely connecected with the largest cluster in the network, there is a clear secondary cluster comprising a high proportion of red nodes. This relative insularity of the Alt-Right commenters is reflected in a lower conductance in the pre-ban period as compared to the post ban period.

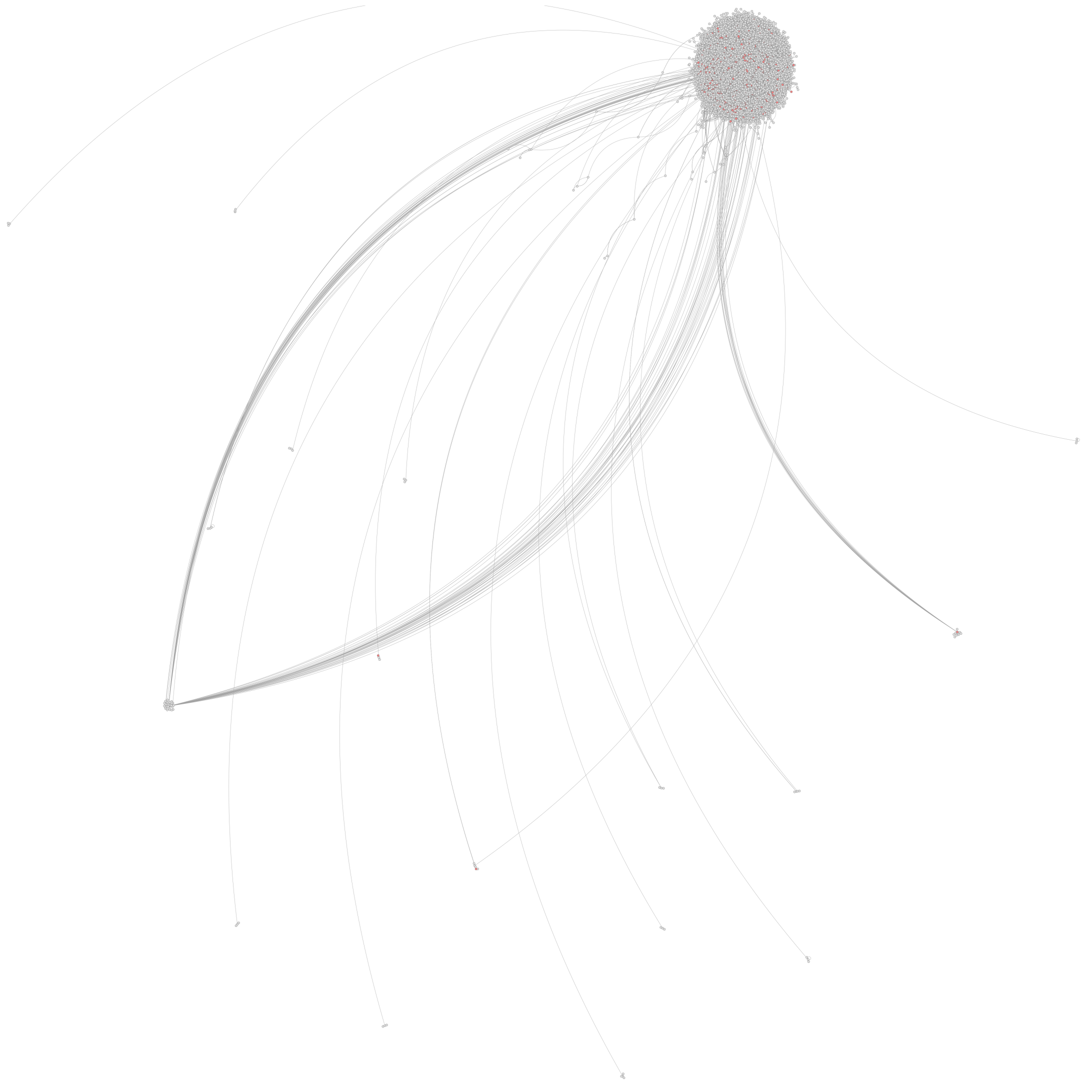


Figure 10. Post-shock reply network. The reply network for the two weeks after the Alt-Right ban consists of approximately 50,000 nodes (users) and 400,000 edges. This is the largest connected component of the complete reply network. It reflects over 99% of users for that period. Users who have ever posted in the Alt-Right subreddit are represented by red nodes, while others are represented by grey nodes. After the ban, the network has only one main cluster, suggesting that the Alt-Right commenters have become more integrated with the rest of the community after the ban. This is reflected in a higher conductance in the post-ban period as compared to the pre-ban period.

Table IX. Alt-Right jargon dictionary and sources.

Term	Source	Reference
Antifa	NYT	https://www.nytimes.com/article/what-antifa-trump.html
Asatru	SPLC	https://www.splcenter.org/fighting-hate/extremist-files/ideology/neo-volkisch
Blackpill/black pill	Quartz	https://qz.com/1092037/the-alt-right-is-creating-its-own-dialect-heres-a-complete-gu
Chad	Quartz	https://qz.com/1092037/the-alt-right-is-creating-its-own-dialect-heres-a-complete-gu
Deus Vult	Vice	https://www.vice.com/en_us/article/ezagwm/get-to-know-the-memes-of-the-alt-right-and
Dindu Nuffin	SPLC	https://www.splcenter.org/hatewatch/2018/04/19/day-trope-white-nationalist-memes-thr
Echoes (triple parentheses)	ADL	https://www.adl.org/education/references/hate-symbols/echo
Femoid	Quartz	https://qz.com/1092037/the-alt-right-is-creating-its-own-dialect-heres-a-complete-gu
Gibs	SPLC	https://www.splcenter.org/hatewatch/2018/04/19/day-trope-white-nationalist-memes-thr
Goolag	Quartz	https://qz.com/1092037/the-alt-right-is-creating-its-own-dialect-heres-a-complete-gu
Hypergamy	Quartz	https://qz.com/1092037/the-alt-right-is-creating-its-own-dialect-heres-a-complete-gu
Jewish question/JQ	SPLC	https://www.splcenter.org/hatewatch/2018/04/19/day-trope-white-nationalist-memes-thr
Mudshark	SPLC	https://www.splcenter.org/hatewatch/2018/04/19/day-trope-white-nationalist-memes-thr
Odin	SPLC	https://www.splcenter.org/fighting-hate/extremist-files/ideology/neo-volkisch
Rapefugees	SPLC	https://www.splcenter.org/hatewatch/2018/04/19/day-trope-white-nationalist-memes-thr
Social justice warrior/SJW	Quartz	https://qz.com/1092037/the-alt-right-is-creating-its-own-dialect-heres-a-complete-gu
Skittles	Quartz	https://qz.com/798305/alt-right-trolls-are-using-googles-yahoos-skittles-and-skypes-
Six Gorillion	ADL	https://www.adl.org/education/references/hate-symbols/six-gorillion
Snowflake	LA Times	https://www.latimes.com/nation/la-na-pol-alt-right-terminology-20161115-story.html
TPTB	Quartz	https://qz.com/1092037/the-alt-right-is-creating-its-own-dialect-heres-a-complete-gu
Transtrender	Quartz	https://qz.com/1092037/the-alt-right-is-creating-its-own-dialect-heres-a-complete-gu
We Wuz Kangz	SPLC	https://www.splcenter.org/hatewatch/2018/04/19/day-trope-white-nationalist-memes-thr
White genocide	ADL	https://www.adl.org/resources/glossary-terms/white-genocide
Muh holocaust	ADL	https://www.adl.org/education/references/hate-symbols/muh-holocaust
Wrongthink	Quartz	https://qz.com/1092037/the-alt-right-is-creating-its-own-dialect-heres-a-complete-gu