Huyen Nguyen, Charles Huyi, Paul Warren*
CS224W
12/10/2017

The Propagation of Lies: Impeding the Spread of Misinformation by Identifying and Invading Echo Chambers in Networks

Introduction and objectives

The World Economic Forum recently declared the volume and rapid spread of misinformation online as one of the top threats to society today [1]. Biased — and, at its limit, fake — news swayed recent elections in the US and Thailand [2,3], confused health workers during the Ebola crisis [6], and has been an enabler of violent extremism [4].

Echo chambers are among the largest known vectors for the spread of misinformation [6], in part because they are the manifestation of natural human tendencies. Indeed, in 2016, the Pew Research Center reported that 83% of users simply ignore content that conflicts with their existing views, 31% tailor their social news feeds to see less disagreeable content, and 27% block or unfriend the source of that information [7]. And as social media companies increasingly personalize their platforms, echo chambers will only become more prevalent over time.

Existing research has identified echo chambers among policymakers, researchers, social networks, and other communities. Invading or breaking down these echo chambers, however, remains a largely unsolved problem.

Our project has two main objectives. The first is to identify echo chambers in a real-world dataset using both supervised and unsupervised methods. The second is to prototype and test methods for invading echo chambers by exposing their members to nonnative content and diversifying the content they see. We hope this work helps start a discussion on exact methods that can be used to combat the spread of misinformation in social networks.

Related work

Prior research has identified and quantified tightly knit communities within social networks. Colleoni et al. [8] identified and characterized groups of like-minded users within Lang and Leskovec's sample of 467 million tweets in 2009 (representative of the entire population of Twitter users at that time). They first classified tweets and users as political or nonpolitical, then political users as either Democrats or Republicans (ignoring moderate users). The authors found the number of users who follow the official Republican account to be eight times higher than the number of users following the official Democratic account, but that Democrats exhibit significantly higher homophily than Republicans (88% of Democrats' Twitter interactions are with other Democrats, compared with 24% for Republicans). Their methods are heavily supervised, however, limiting the adaptability of their work to other datasets.

Researchers have also begun to quantify the spread of misinformation within social networks. Howard et al. [2] used a manually curated list of hashtags and keywords to query Twitter's public streaming API and retrieve roughly 22 million politically-related tweets between November 1st and November 11th, 2016. The authors manually sorted the URLs in these tweets into 16 different categories like "Professional Political Content - Experts" and "Other - Junk News" so that they could measure the amount of each type of news over time. Among their findings: a dramatic increase in the spread of "fake news" in the days leading up to the election.

We've read a significant amount of research over the course of working on our project, but the above research represents the primary basis from which we build. Our work shares many of the same goals as the authors above — for example, finding and quantifying biased echo chambers — and we replicate some of their methodology. That said, we aim to differentiate ourselves by reducing the amount of supervision required in our approach. Though we use Howard et al.'s political dataset in order to be able to easily validate our results, we develop methods that can be applied to any social network containing insular communities. Finally, our project has the additional objective of invading echo chambers by suggesting nonnative content from other communities, using unsupervised methods.

Data

We use the Michigan Fake News dataset recently published by the Computational Propaganda Project at the Oxford Internet Institute [19]. This dataset contains 63,277 tweets by Michigan-based users tweeted from Nov 1th to Nov 11th, 2016, in the lead up to the 2016 US Presidential Election. The Oxford group manually curated a list of political hashtags and key terms, searched for tweets that directly contained at least one of those tokens or referenced a URL or another tweet that contained at least one of those tokens, and selected only the tweets by users who claimed to be based in Michigan via their user-provided "city" and "state" account tags. Michigan was chosen because it was a key battleground

* Paul Warren is not in 224W.

state where public support was evenly split between both candidates right up to Election Day. Michigan voters were 47.6% for Donald Trump, 47.3% for Hillary Clinton [12].

Our dataset contains 38,936 unique users, 63,277 tweets, 10,318 unique hashtags, and 25,920 unique URLs. The top 5 most frequently occurring hashtags are:

Hashtag	Count
#MAGA (short for Make America Great Again)	6714
#Election2016	4979
#Trump	4683
#ImWithHer	3563
#DrainTheSwamp	2451

The top ten most frequently occuring URL domains are:

Domain	Count
youtube.com	1142
wikileaks.org	864
instagram.com	599
facebook.com	506
truthfeed.com	218
infowars.com	201
foxnews.com	189
briebart.com	154
politicususa.com	117
dailycaller.com	96

Some sample tweets from the Michigan dataset:

User	Tweet
lastara36	@shondarhimes I swear this election is a few pages out of a scandal season! Do you write this stuff? #Election2016
davidreisig1	RT @wikileaks: Hillary Clinton campaign's Pied Piper strategy (see attachments) #PodestaEmails https://t.co/DAmWNq9K0f
FramingNebula	RT @mcspocky: Reminder: Donald Trump due in court after Election Day on child rape and racketeering charges https://t.co/Pl5nUo0GC8 #ctl #p…

Methodology and results to date

We tackle our objectives in four steps:

1. **Graph preprocessing:** Convert tweets and their metadata into a graph.
2. **Community identification:** Identify clusters of users, or communities, based on interactions.
3. **Echo chamber diagnosis:** Identify which communities are echo chambers, based on content and characteristics.
4. **Content suggestion / invading echo chambers:** Prototype and test methods for invading echo chambers by exposing their members to nonnative content and diversifying the content they see.

1: Graph preprocessing

First, we transform the Michigan dataset into a graph where nodes are users and edges are interactions between users. When creating edges, we consider the following directed interactions:

- If user A retweets a tweet of user B, there is an edge from A to B
- If user A mentions user B in one of their tweet, there is an edge from A to B
- If user A replies to a tweet of user B, there is an edge from A to B

Types of graphs

We experiment with three different types of graphs.

- Single directed graph (snap.py's PNGraph): there is at most one directed edge from one source node to a destination node. There is a directed edge from user A to user B if A interacts with B, and there is a directed edge from user B to user A if B interacts with A. There can be up to two directed edges between A and B.
- Single undirected graph (snap.py's PUNGraph): There is at most one undirected edge between a pair of nodes. There is an edge between user A and user B if either user interacts with the other. There can be at most one undirected edge between A and B.
- Multi directed graph (snap.py's TNEANet): There can be more than one directed edge from one source node to a destination node. There is a directed edge from user A to user B for each time A interacts with B, and vice versa. There can be multiple directed edges in either direction between the two users.

With 38,936 unique users and 63,277 tweets, there are, on average, fewer than 2 tweets per user. All graphs have the same number of nodes (38,936), but the number of edges differ. The fact that the multi-directed graph has nearly twice the number of edges as the single-directed graph suggests that repeat interactions between users are not uncommon.

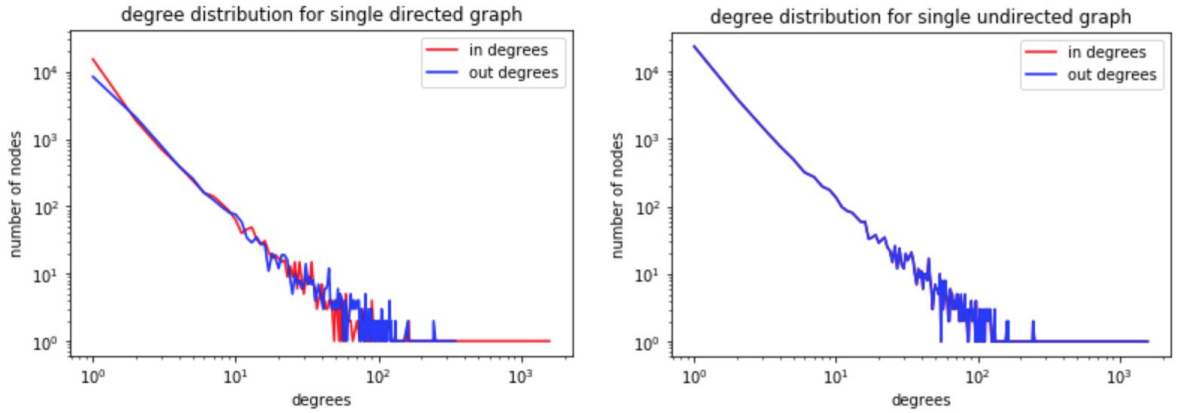
Graph	Edges
Single directed	51,732
Single undirected	51,702
Multi directed	109,185

Power-law degree distributions

The in- and out-degree distributions of all three graphs confirm that there are indeed many repeated interactions between users.

Graph	5 max in degrees	5 max out degrees
Single directed	277, 300, 417, 991, 1575	289, 328, 328, 333, 345
Single undirected	333, 345, 417, 991, 1575	333, 345, 417, 991, 1575
Multi directed	1409, 1426, 1991, 2124, 6030	732, 833, 834, 841, 935

Moreover, the degree distributions of our graphs follow a power law distribution like many real world graphs such as the web graph, actor-collaborations, citations to papers, and other online social networks.



Degree distributions of two of our networks.

2: Community identification

To identify groups of users who interact with each other significantly more than they do with users outside of the group, we use the notion of the modularity score.

Modularity score

Developed by Newman and Girvan in their classic work, “Finding and evaluating community structure in networks,” modularity scores allow us to quantify the strength of a community [16]. The modularity score of a community is calculated as:

$$Q = \sum_i (e_{ii} - a_i^2) = \mathrm{Tr}\ \mathbf{e} - \|\mathbf{e}^2\|,$$

where e_{ij} is the fraction of all edges in the network that link vertices in community i to vertices in community j, and $a_i = \sum_j e_{ij}$. In a network where edges fall between vertices irrespective of communities, $e_{ij} = a_i a_j$, so this quantity measures how much the fraction of edges in the network that connect vertices within a community exceeds a baseline. Modularity scores fall between 0 and 1, where higher numbers suggest higher levels of community.

Understanding our data

First, we want to understand the size of groups who use or interact with the same “anchors,” where an anchor is a hashtag or a well-known account (e.g., @wikileaks). That is, for each hashtag, we consider the set of users who use this hashtag to be a potential community. Similarly, for each well-known account (those with over a tunable threshold of in-links), we consider the set of users who interact with this account to be a potential community.

While modularity scores between 0.3 and 0.7 indicate significant community structures, the modularity scores of most of our same-hashtag groups are negative.

Hashtag	# users using hashtag	Modularity
MAGA	1468	-0.077375
MakeAmericaGreatAgain	757	-0.031860
ImWithHer	2256	-0.064901
NeverTrump	533	-0.015996
Any of pro-Trump hashtags	3453	-0.108515
Any of pro-Hillary hashtags	3342	-0.097418

Similarly, when we looked at groups of users who interact with one of the top 100 accounts based on in-degree, such as famous political figures @realDonaldTrump, @HillaryClinton, @wikileaks, @DonaldJTrumpJr, etc., we find negative modularity scores.

Popular account	# users interacting with account	Modularity
realDonaldTrump	1575	-0.07983
HillaryClinton	991	-0.061574
wikileaks	417	-0.026257
DonaldJTrumpJr	300	-0.016676
FoxNews	277	-0.01553
DanScavino	273	-0.02192
bfraser747	265	-0.02244
mike_pence	260	-0.0178
LindaSuhler	245	-0.02229
CNN	235	-0.01432

This suggests that the communities within our network may be smaller than these groups, and that a group of users who share the same viewpoint don't necessary know each other or have interactions with each other. The sparsity of our graph also makes it less likely to have an edge between two arbitrary nodes.

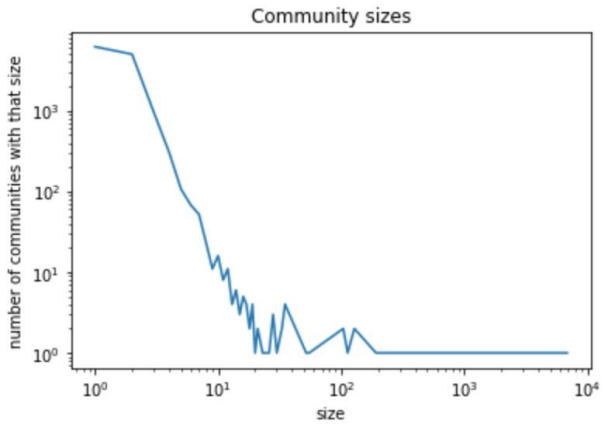
Communities based on network structure

We then leverage the network structure itself to optimize more directly for the modularity score and use the Clauset-Newman-Moore community detection method to detect communities [17]. The algorithm has been proven to be considerably faster than most previous general algorithms such as Girvan-Newman [16] or the algorithm developed by Radicchi et al. in [18]. At every step of the algorithm, two communities that contribute maximum positive value to global modularity are merged.

We enforce that communities should consist of at least 20 nodes, since from the perspective of a platform like Twitter or Facebook, smaller communities — even extremely insular ones — are unlikely to pose much of a treat to the spread of misinformation. This parameter can be easily adjusted for other applications.

With this method, we find 12,847 communities, 6,209 of them consisting of only one user and 5,012 with only two users. The predominance of communities with only one user makes sense given the sparsity of our graph. There are many users that aren't connected to any other user (see figure below).

Still, there are 32 communities with at least 20 users. The three largest communities have 6,847, 4,881, and 1,156 users, respectively.



Frequency of communities of various sizes.

3: Echo chamber diagnosis

For our project, we define echo chambers as groups that, in addition to being insular (i.e., a community), exhibit strong single-mindedness when it comes to content. As such, echo chambers cannot be identified based on network structure alone and require us also to look at the types of interactions that connect users within the group. In this step of our methods, we aim to determine which communities we identify in the previous step are indeed echo chambers.

Establishing a baseline truth using prior knowledge ("the hashtag method")

Given the political nature of our dataset, the easiest way to make this determination is to look at what hashtags are used in each community. There are a total of 10,318 hashtags in our dataset, and 19,595 out of 38,936 users include at least one hashtag in their tweets.

Hashtag	Used by # users	# occurrences
#election2016	3611	5726
#imwithher	2256	4945
#trump	2209	5994
#maga	1468	7325
#draintheswamp	966	3510

Top 5 hashtags in our dataset.

Using hashtag content is our baseline echo-chamber identification method. Like Colleoni et al. [8], we begin by manually curating our own list of pro-Hillary (448) and pro-Trump (711) hashtags. Consistent with Colleoni's findings, we find more pro-Trump hashtags, and that Trump supporters tweeted more during our dataset's timeframe than Hillary supporters [2].

We then classify each user as pro-Trump if s/he uses at least twice as many pro-Trump hashtags as pro-Hillary hashtags; and vice versa. We label all users not using one type hashtag twice as much as the other as neutral.

	# pro-Trump	# pro-Hillary	# neutral	Trump:Hillary
Hashtags	708	444	9166	1.595
Users	3073	2530	205	1.215

After classifying our users, we move on to communities. We consider a community to be pro-Trump if two conditions are met:

1) The ratio of pro-Trump users / pro-Hillary users is at least twice the ratio of pro-Trump users / pro-Hillary users over the entire network, or the number of pro-Hillary users is 0.
2) There are at least 5 more pro-Trump users than pro-Hillary users.

Pro-Hillary communities are defined symmetrically, and communities not meeting either set of requirements are considered neutral. We find that 8 (25%) of the communities we identified in the previous step are pro-Hillary, another 8 (25%) are pro-Trump, and the remaining 16 (50%) communities are neutral. There are strong signs of homophily within the non-neutral communities: among pro-Hillary communities, for example, the ratio of pro-Hillary to pro-Trump users is extremely high, often exceeding 10. Among pro-Trump communities, the ratio of pro-Trump to pro-Hillary users is lower, at 3-6. (Hillary supporters exhibit strongly homophily, another consistent finding to Colleoni [2].)

Community	# tags	# T-tags	# H-tags	T/H tags	# users	# T-users	# H-users	T/H users	Affiliation
Entire graph	10318	708	444	1.594595	38936	3073	2530	1.214625	N
0	4227	669	145	4.613793	6847	1101	163	6.754601	T
25	296	56	13	4.307692	322	86	21	4.095238	T
1566	33	0	7	0	33	0	11	0	H
31	341	115	13	8.846154	341	35	5	7	T
49	61	15	4	3.75	35	11	2	5.5	T
41	158	28	15	1.866667	192	11	36	0.305556	H
685	28	2	5	0.4	128	4	29	0.137931	H
693	19	0	3	0	33	0	3	0	N
56	232	24	13	1.846154	486	24	26	0.923077	N
66	185	32	12	2.666667	133	17	4	4.25	T
1231	8	2	0	-1	26	19	0	-1	T
312	97	13	17	0.764706	349	7	79	0.088608	H
1237	14	2	4	0.5	35	1	24	0.041667	H
18	51	6	8	0.75	52	1	13	0.076923	H
368	50	16	2	8	55	18	1	18	T
368	64	5	10	0.5	104	0	18	0	H
634	47	5	8	0.625	249	2	44	0.045455	H

In two of our communities, there are only one or two pro-Trump users among many pro-Hillary users, suggesting that a couple of Trump supporters might have tried to provoke or interact with Hillary supporters.
We label these pro-Trump and pro-Hillary communities, which are both insular (i.e., a network community) and single-minded in terms of content, as evidenced by the strong homophily in hashtag usage, as echo chambers. The remaining neutral groups are network communities but not echo chambers.

Using our understanding of the political leaning of each hashtag to identify echo chambers gives us a baseline truth for this dataset with which to compare the results of other approaches. Now, we see how we can improve these methods.

Generalizing our approach for identifying polarized hashtags

The non-generalizable step in the above approach (and the prior research we read) for identifying echo chambers from communities is the curation of two large lists of polarized hashtags (one Democratic and one Republican). This step is also laborious and takes time. To improve the generalizability of our approach to other, not necessarily political, datasets, we experiment with a semi-supervised way of growing lists of polarized hashtags given two small lists of opposing hashtags. Here, we use "pro-Trump" and "pro-Hillary" hashtags, but the approach can be applied to "pro-X" and "anti-X" in another dataset just as easily.

We start with a small list of 20 manually curated pro-Trump hashtags and a list of 20 pro-Hillary hashtags and run the following algorithm:

1. Use these two small lists to identify pro-Trump users and pro-Hillary users.
2. Get a list of all the hashtags used by pro-Trump users we identify, call them T-tags. Similarly, get a list of all the hashtags used by pro-Hillary users, call them H-tags.
3. Eliminate the tags shared by both list T-tags and H-tags from the two lists.
4. From the T-tags list, keep the tags that are used by at least K pro-Trump users. Similarly, from the H-tags list, keep the tags that are used by at least K pro-Hillary users.

We evaluate the results of this algorithm, for various values of K, against the curated lists of pro-Trump and pro-Hillary hashtags we built previously. Here are the results:

K	T-tags				H-tags				F1 score
	Total	*Trump*	*Hillary*	*Neutral*	*Total*	*Trump*	*Hillary*	*Neutral*	*T / H*
5	524	206	9	309	102	1	54	47	0.56 / 0.65
4	669	243	13	413	145	1	75	69	0.52 / 0.64
3	917	276	16	625	211	1	95	115	0.46 / 0.59

Higher values of K result in better performance as measured by F1. Our algorithm can misclassify neutral hashtags as pro-Hillary or pro-Trump, but it rarely confuses a pro-Trump hashtag for pro-Hillary, and vice versa. The results suggest that a semi-supervised method as the above can be effective for growing lists of polarized hashtags, which in turn can be used in an approach like the one above to identify echo chambers.

T-tags output samples	H-tags output samples
#trumpvoters	*#drumpf*
#backtheblue	*#flipcongress*
#unionstrong	*#blacklivesmatters*
#lancastervotes	*#shareblue*
#hillaryforprison	*#sexualpredator*
#evangelical	*#andalsowithher*
#wjc	*#dirtydonald*
#trumpforceone	*#sociopath*
#broward	*#trumpvsclinton*
#corruptdnc	*#expectus*
#davechappelle	*#mo*
#5days	*#sowithher*
#neverhillary	*#trumpuniversity*
#benghazibutcher	*#hillarysarmy*

Unsupervised identification of echo chambers

The most basic unsupervised approach for selecting echo chambers from a list of communities is to randomly select communities and label them as echo chambers. Given a sufficiently polarized dataset and a high modularity-score bar for identifying communities, this may work well in practice, especially for very large datasets where more involved methods require too much time or computing resources. Applying this approach to our dataset would generate a 50% accuracy rate relative to our baseline truth (as 50% of our identified communities were categorized as echo chambers using the hashtag method above).

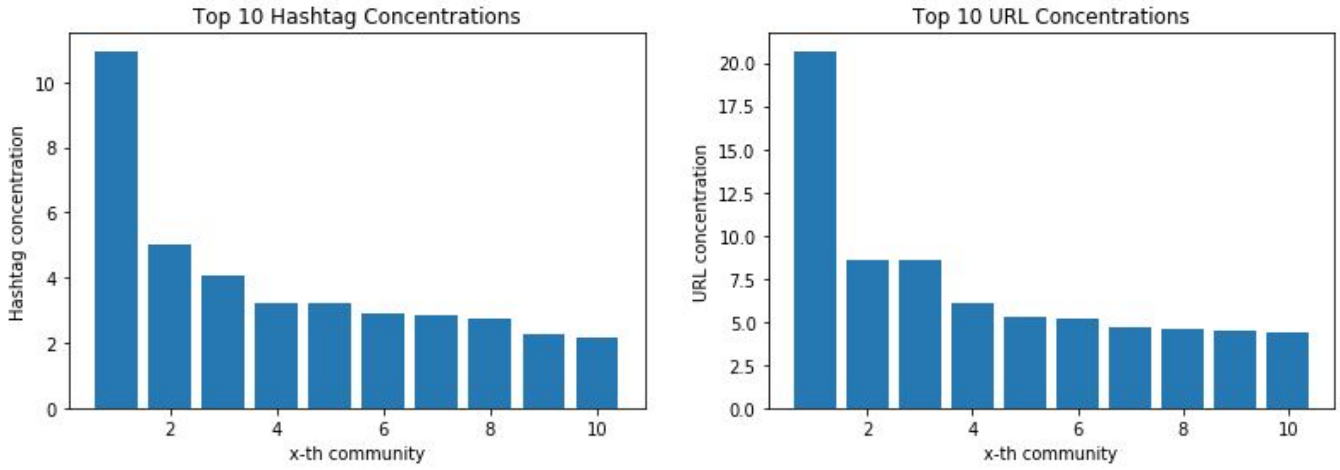
A more intentional unsupervised approach involves looking at hashtags and URLs, but, unlike before, ignoring their content. Instead, we define concentration, a function of a community C, and a set of hashtags H or urls U, as:

$$concentration(C, H) = \frac{\#\ tweets\ with\ hashtags\ by\ users\ in\ C}{\#\ distinct\ hashtags\ in\ H\ used\ by\ users\ in\ C} \qquad concentration(C, U) = \frac{\#\ tweets\ with\ URLs\ by\ users\ in\ C}{\#\ distinct\ urls\ in\ U\ used\ by\ users\ in\ C}$$

The idea is that users of a community that is not an echo chamber are expected to tweet about a variety of topics and represent a normal range in terms of points of view. We use hashtags and URLs as a proxy for both topics and points of view. Therefore, in a regular community, we expect users' tweets to contain a range of hashtags and a range of URLs. On the other hand, users in an echo chamber are more likely to tweet about a limited range of topics and represent a narrower range of points of view, relying on a smaller set of hashtags and URLs. This is one property we want to capture about echo chambers. The other, with a view toward biased content, is that the echo chambers most worth disrupting are those that generate the greatest amount of biased content.

The concentration score measures the degree to which a community both: (a) generates a significant amount of content and (b) relies on a limited set of hashtags or URLs.

The set of hashtags H or urls U can be limited to the output of our polarized lists, generated by our semisupervised method above. However, we find that simply letting H be the set of all hashtags used in the community and U be the set of all URLs yields meaningful results. For each community identified by the Clauset-Newman-Moore method in step 2, we compute this concentration score.



Distribution of hashtag / URL concentration scores in our dataset.

Ordering the communities by concentration scores (for both hashtags and URLs) from high to low, we see that there is a steep drop-off in scores after the top few communities. For example, with URL concentration, the top community has four-times the concentration of the 6th highest community (20 to 5). There is no definitive threshold separating echo chambers from communities, but for the purposes of our dataset, we consider the top community by hashtags and the top three communities by URLs to be echo chambers.

4: Content suggestion / invading echo chambers and frameworks for measurement

Echo chambers are among the largest known vectors for the spread of misinformation [6], and, having identified echo chambers, we turn our attention to breaking them down. In practice, a social media company (e.g., Twitter) can take various heavy-handed approaches such as automatically blocking communication between members of established echo chambers or limiting the number of tweets a user can post using a single hashtag. In this paper, we explore a less authoritarian solution: suggesting relevant content to a member of an echo chamber that presents a different point of view than that of the echo chamber.

We also propose a few simple frameworks for evaluating how well a particular approach achieves its goal of breaking down echo chambers, given that there is no formal definition of an echo chamber. One can imagine these (or similar frameworks) informing key-performance metrics for social media companies resisting the formation of echo chambers — and, consequently, the spread of misinformation — on their platforms.

Generating diversity of content through recommendations

Content recommendation is a large field that typically deals with giving users more of what they want. To disrupt echo chambers, we deviate somewhat from this path and recommend content with the intent of bringing diverse viewpoints and information to a member of an echo chamber. Recommended content can take many forms: tweets to read, users to follow, or websites to visit. The recommendation should also be made to maximize user engagement with the content. Populating a Trump supporter's feed with inflammatory pro-Hillary tweets is unlikely to inspire open-mindedness (and might even have a counterproductive effect).

We explore only tweet recommendations in this paper, though similar approaches can be taken to recommend users to follow or URLs to visit. At a high level, our approach consists of three steps: (1) arbitrarily choosing a user to target, (2) identifying relevant tweets for this user, and (3) selecting which tweet (from the relevant tweets) to show the user.

In step 1, we arbitrarily choose a user from one of our identified echo chambers, along with all of that user's tweets. The user is an active Trump supporter from whom we have 368 tweets. For step 2, in order to identify relevant tweets for this user, we first need to group all tweets in our dataset based on semantic content. We do this by using word2vec to convert each word in every tweet to a vector, then averaging the word-vectors in each tweet to encode a tweet-vector. We calculate the cosine similarity between all pairs of tweet-vectors and build a large graph whose nodes are tweet-vectors. An edge exists between two nodes if the cosine similarity exceeds a tunable threshold. We choose our threshold to be 1, the highest possible cosine similarity score, to see if the theoretically highest-quality results are good enough for recommendations. Future work may involve looking into other thresholds.

We run the Clauset-Newman-Moore community detection algorithm [17] on this tweet graph. Unlike the user communities we identify earlier in this paper, these clusters represent tweets with similar vector representations.

- We identify 98 clusters of various sizes.
- The average number of tweets in a cluster is 97. The median number of tweets in a cluster is 3.
- The smallest cluster contains 2 tweets, and the largest cluster contains 3,127 tweets.
- Three clusters contain more than 2,700 tweets. The rest of the clusters contain fewer than 120 tweets.
- The 368 tweets by our chosen user, a Trump supporter, fall into 8 clusters. Three of these clusters have size 2,700; the others have size 118, 33, 7, 7, and 5.

Now we are ready to identify relevant tweets for our user (step 2 above). For each tweet by our selected Trump supporter, we look at the cluster into which it falls and can either define relevant tweets as either (a) all pro-Clinton tweets in the cluster (supervised) or (b) all tweets from other echo chambers in the cluster (unsupervised). An unsupervised approach can be applied more readily to datasets in general, though when the polarities in a network are known, a supervised approach can be taken at this step in combination with unsupervised techniques for the other steps.

Regardless of how relevant tweets are determined, we then need to select which relevant tweets to show to the user (step 3). One option is to randomly select among relevant tweets. Another is to order relevant tweets by when they were posted, and recommend the most recently posted content. We find that the best results come from using a third approach: recommending tweets based on their posters' popularity.

We determine a user's popularity by generating a directed graph of user interactions and calculating a user's popularity score as the in-degree count divided by the out-degree count. We use this ratio instead of using simple in-degree in order to have a sense of proportion. Intuitively, a user with 2 in-degree connections and 100,000 out-degree edges is "less popular" than a user with 2 in-degree connections and 10 out-degree edges. We sort the relevant tweets to our selected Trump supporter by the popularity of the tweet's poster, and suggest the top relevant tweets from that list. This process is repeated for each tweet by our selected Trump supporter. That is, for each of his tweets, we generate a corpus of associated relevant tweets and sort these tweets by their posters' popularity to select which relevant tweets to show. User popularity only needs to be computed once. Also, when multiple relevant tweets' posters' have equal popularity, we randomly select one of the tweets to show.

Tweet by selected Trump supporter	**Number of relevant tweets**	**Recommended tweet from relevant tweets based on popularity of poster**	**Handpicked "good" recommendation**	**Handpicked "poor" recommendation**
RT @realDonaldTrump: Just landed in North Carolina-heading to the J.S. Dorton Arena. See you all soon!	3,684	@wellnesscoachgc We're going to win because of volunteers like you! Thank you! #ImWithHer -Lauren	Know your rights: Check out each state's law on employee time off to vote:	@krystalball @KatrinaPierson 48 more hours of this utter B.S. It cannot pass quick enough. #ImWithHer #LetsDoThis

Lets #MakeAmericaGreatAgain! https:...			https://t.co/1EhTPiUHcy #1uvote #imwithher	
RT @TrussElise: Bill O'Reilly Exposes George Soros! #Hillary Clinton & Democrat Funding 8... https://t.co/2yHCUcOyVF via @YouTube #TrumpPen	2,016	RT @smileydevil: Yep. Make America Great Again. Complete lunatics. #ImWithHer #VoteBlue #NeverTrump https://t.co/UsjEB32ELg	#Hillary maintains lead in new poll after #FBI letter"	@Plaid_Underdog #MARA MAKE AMERICA RUSSIAN AGAIN #VoteTrump #IfYoureStupid #racist #sexist #misogynist #Trumprussia #trumpserver #DumpTrump
RT @RobertPercin: Sick Hillary Caught Lying Again (Or Fading Memory)? #MAGA3X\n\nSays She Was in NYC on 9/11 (Wasn't) https://t.co/pcPbQ5N2en...	116	RT @UniteWomenOrg: #Trump has a #rape case pending, so why am I only hearing about #Hillary's emails? https://t.co/dK2G8Vlvon	RT @IMPL0RABLE: Trump 'Special Session' to Replace Obamacare Is Imaginary, lacks understanding of Govt. https://t.co/kjixHUPBcy #MAGA #Trum...	Trump's closing argument: A woman cannot be president https://t.co/O3jlk9Xit4 #maga… https://t.co/LLpwG1CEPw #politics #MSNBC #CNN #p2...
RT @PhxKen: Father of soldier killed in #Benghazi endorses @realDonaldTrump @HillaryClinton responsible for the death of my son https://t.c\u2026	32	RT @tesskrasne: 10,000s of phone calls, 1,000s of doors, 100s of voters registered, 141 days, and 1 vote. I'm yours, #HillaryClinton .	🇺🇸How many dishonorable discharges? Serve. Protect and defend the Constitution of the United States. #NeverTrump… https://t.co/K07nzVhAFM	RT @kharyp: 28% of GOP early voters in Florida voted for #HillaryClinton & she has an 8 point lead in early voting!
RT @wikileaks: Arianna Huffington, co-founder of Huffinton Post, prefers covert influence #PodestaEmails\n https://t.co/zn0NhFxBwA	2,016	RT @smileydevil: Yep. Make America Great Again. Complete lunatics. #ImWithHer #VoteBlue #NeverTrump https://t.co/UsjEB32ELg	#FBI Horrified As Spy Says #Russia Has Been Supporting And Cultivating #Trump For Years @politicususa https://t.co/E7JBrFAnI5 #NeverTrump	@Plaid_Underdog #MARA MAKE AMERICA RUSSIAN AGAIN #VoteTrump #IfYoureStupid #racist #sexist #misogynist #Trumprussia #trumpserver #DumpTrump

Tweets by our selected Trump support, the associated recommendations for each, and hand-picked relevant tweets to show the range in the quality of relevant tweets.

The results in the table above — to be read horizontally — show 5 representative tweets by our selected Trump supporter (in the leftmost column), along with the tweet we recommend (middle column) based on the procedure described above. In the two rightmost columns, we include tweets we pick by hand from the list of relevant tweets; this is in order to demonstrate that some relevant tweets are more suited to being the recommended tweet than other relevant tweets. The tweets in the "poor" recommended tweets column, for example, are unlikely to be effective at encouraging our selected Trump supporter to be more open-minded. Some of them are not entirely relevant to the Trump supporter's tweet, while others are inflammatory or divisive. Our takeaway is that the process of selecting which relevant tweet to show the user (step 3 above) is indeed an important component to any methodology.

In examples (rows) 2 and 5 above, the recommended tweet based on popularity is also inflammatory / ad-hominem in nature. Better options do exist in the relevant tweets, however, as seen in the "good" recommendation column. Using popularity as the basis for the recommendation is effective insofar as both sides of the polarized topic take a high road. In situations where opposing parties have devolved into mud-slinging, other recommendation approaches might be more appropriate. In general, we are quite impressed with the recommendations generated by our approach.

Framework for evaluating a method's success

The results above give rise to an important question, which is how to measure the quality of any approach taken to reduce the prevalence of echo chambers on a platform. This question is challenging in part because there is no formal definition of an echo chamber. Without being able to evaluate the impact of a given approach, however, a social media company would have a hard time justifying its efforts to combat echo chambers, since it wouldn't know if its efforts were having a positive impact.

A manual way to evaluate the quality of a recommendation-based approach is to have manual reviewers look through recommendations suggested for a particular user's tweet. Given data in the format presented above, with the user's original tweet matched with a recommended tweet, a group of reviewers can review recommendations and score the overall quality of recommendations. Another option is to track how often users engage with recommended tweets — for example, by navigating to a recommended tweet's user's Twitter page or clicking on a URL in a recommended tweet. The

manual approach is cumbersome, and both approaches evaluate the recommendation algorithm more than it measures whether the algorithm has an effect on disrupting echo chambers.

An echo chamber's culpability with regard to the spread of misinformation lies not in its underlying community structure but in the biased content that can circulate, unchecked, within it. (A silent echo chamber poses limited risk.) Therefore, we propose that the quantity to be minimized within a network is the average size of the largest set of semantically similar tweets within identified echo chambers. This quantity can be derived from a combination of the approaches described above. First, communities must be identified based on network structure, and some content-based criteria (e.g., URL concentration) can used to tease out which communities are echo chambers. Secondly, in parallel, all tweets can be clustered based on their content to identify semantically similar tweets. Combining these two results, one can compute the number of tweets within each echo chamber that belong to the same tweet cluster. If a single echo chamber consists of tweets only from one tweet cluster, that would suggest that the echo chamber is biased. If the number of tweets in this echo chamber is, in addition, large, then this bias becomes problematic. Across the network, one hopes that the average size of the largest set of same-cluster tweets within echo chambers is low.

An alternative formulation is using the notion of a PageRank-like random walker who traverses the network of users and, at each jump, stops to read a random tweet by that user. This random walker follows directed interactions between users with some probability p and jumps randomly to another user with (smaller) probability $1-p$. Moreover, for each user he comes across, the random walker stops and picks a random tweet by this user to read. If for some subset of starting nodes for the random walker, the likelihood of reading tweets of similar semantic content after multiple steps is high, then there is likely to be an echo chamber in the network. Put in practice, a social media platform could run multiple trials with a random walker initialized at different starting nodes and compute the fraction of time that the first piece of content and the last piece of content read by the walker are semantically similar. This would be done before and after some sort of intervention (e.g., the content recommendation methods described above) to evaluate the effectiveness of the intervention. A good outcome would be that the fraction of time that content remains similar after multiple iterations, drops.

One challenge with this approach is whether it is able to distinguish between echo chambers that spread misinformation from more benign, siloed communities that share, say, content related to cute pets. In some ways, this distinction is somewhat fuzzy, as any insular community lacks healthy checks and balances and therefore is more likely to propagate biased information. However, heuristics such as (a) the presence of language whose sentiment is at the extremes (either very negative or very positive) or (b) offensive language may be a good ways to easily distinguish traditional "undesirable" echo chambers from more benign insular communities. Allowing for more supervised approaches, platforms could also look for the presence of keywords related to violence or strong social / political affiliations.

Conclusions

The spread of misinformation is a prevalent and growing threat globally. On social media platforms, echo chambers are among the largest known vectors for the spread of misinformation [6], in part because they are the manifestation of natural human tendencies. Indeed, in 2016, the Pew Research Center reported that 83% of users simply ignore content that conflicts with their existing views and 31% tailor their social news feeds to see less disagreeable content [7].

Our work addresses two objectives. The first is identifying echo chambers among Twitter members using both supervised and unsupervised methods. Modeling users as a network enables a wide range of analyses, and, for example, we find that the pattern of user interactions in real data follows a power-law distribution. Furthermore, unsupervised methods based on network structure and modularity score are able to tease out strongly knit communities, and measures such as the frequency of hashtag repeat-use can then be used to distinguish between communities that are traditional "undesirable" echo chambers from more-benign, siloed user groups.

Our second objective is to prototype and test methods for disrupting echo chambers by exposing their members to non-native content and diversifying the content they see. Using natural language processing and modeling tweets in a network, we are able to identify relevant tweets for any given tweet by a user in an echo chamber. Choosing which relevant tweet to show a particular user deserves further attention. However, selecting recommendations from relevant tweets based on the popularity of users yields sound results. Finally, the question of how to evaluate the effectiveness of any method for disrupting echo chambers is a tricky one. We propose that the objective to be minimized is the average size of the largest set of semantically similar tweets within identified echo chambers across the network.

We hope this work helps start a discussion around additional approaches for combating the formation of echo chambers and the spread of misinformation in social networks, and we thank Anunay Kulshrestha for his guidance throughout this work.

References

[1] World Economic Forum. 10. The Rapid Spread of Misinformation Online. Outlook on the Global Agenda 2014 (2014).
[2] Howard, Philip N., et al. Junk news and bots during the US election: What were Michigan voters sharing over Twitter. Data Memo 2017.1. Oxford, UK: Project on Computational Propaganda. [link]
[3] Grömping, Max. "'Echo Chambers' Partisan Facebook Groups during the 2014 Thai Election." Asia Pacific Media Educator 24.1 (2014): 39-59. - link
[4] O'Hara, Kieron, and David Stevens. "Echo chambers and online radicalism: Assessing the Internet's complicity in violent extremism." Policy & Internet 7.4 (2015): 401-422. link
[5] Your Filter Bubble is Destroying Democracy, WIRED (2016) - link
[6] The inevitable rise of ebola conspiracy theories, March 2014.
[7] The Political Environment on Social Media, Pew Research 2016 - link
[8] Colleoni, Elanor, Alessandro Rozza, and Adam Arvidsson. "Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data." Journal of Communication 64.2 (2014): 317-332. - link
[9] Menczer, Filippo. Fake Online News Spreads Through Social Echo Chambers, Scientific American - link
[10] Garrett, R. Kelly. "Echo chambers online?: Politically motivated selective exposure among Internet news users." Journal of Computer-Mediated Communication 14.2 (2009): 265-285. - link
[11] Morstatter, Fred, et al. "Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose." *ICWSM*. 2013. - link
[12] https://www.politico.com/2016-election/results/map/president/michigan/
[13] Broder, Andrei, et al. "Graph structure in the web." Computer networks 33.1 (2000): 309-320.
[14] Leskovec, Jure, and Eric Horvitz. "Planetary-scale views on a large instant-messaging network." *Proceedings of the 17th international conference on World Wide Web*. ACM, 2008.
[15] Leskovec, Jure, and Julian J. Mcauley. "Learning to discover social circles in ego networks." *Advances in neural information processing systems*. 2012.
[16] Newman, Mark EJ, and Michelle Girvan. "Finding and evaluating community structure in networks." *Physical review E* 69.2 (2004): 026113.
[17] Clauset, Aaron, Mark EJ Newman, and Cristopher Moore. "Finding community structure in very large networks." *Physical review E* 70.6 (2004): 066111.
[18] Radicchi, Filippo, et al. "Defining and identifying communities in networks." *Proceedings of the National Academy of Sciences of the United States of America* 101.9 (2004): 2658-2663.
[19] Philip N. Howard, Gillian Bolsover, Bence Kollanyi, Samantha Bradshaw, Lisa-Maria Neudert. "Junk News and Bots during the U.S. Election: What Were Michigan Voters Sharing Over Twitter?" Data Memo 2017.1. Oxford, UK: Project on Computational Propaganda. Download Data.
[20] Morstatter, F., Pfeffer, J., Liu, H. & Carley, K. M. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. ArXiv13065204 Phys. (2013).

Contributions

Huyen Nguyen and Charles Huyi contributed equally to this project. Paul Warren is not in the class. He downloaded and characterized several Twitter datasets and worked on the natural language processing-based content recommendation system with Huyen and Charles.

We presented this paper at a NIPS workshop, and on Dec 9, 2017, we found out that our paper was selected as a regional winner of the Ericsson Innovation Awards, and we're excited about that!