Community Interaction and Conflict on the Web

Srijan Kumar
Stanford University, USA
srijan@cs.stanford.edu

William L. Hamilton
Stanford University, USA
wleif@stanford.edu

Jure Leskovec
Stanford University, USA
jure@cs.stanford.edu

Dan Jurafsky
Stanford University, USA
jurafsky@stanford.edu

ABSTRACT

Users organize themselves into communities on web platforms. These communities can interact with one another, often leading to conflicts and toxic interactions. However, little is known about the mechanisms of interactions between communities and how they impact users.

Here we study intercommunity interactions across 36,000 communities on Reddit, examining cases where users of one community are mobilized by negative sentiment to comment in another community. We show that such conflicts tend to be initiated by a handful of communities—less than 1% of communities start 74% of conflicts. While conflicts tend to be initiated by highly active community members, they are carried out by significantly less active members. We find that conflicts are marked by formation of echo chambers, where users primarily talk to other users from their own community. In the long-term, conflicts have adverse effects and reduce the overall activity of users in the targeted communities.

Our analysis of user interactions also suggests strategies for mitigating the negative impact of conflicts—such as increasing direct engagement between attackers and defenders. Further, we accurately predict whether a conflict will occur by creating a novel LSTM model that combines graph embeddings, user, community, and text features. This model can be used to create early-warning systems for community moderators to prevent conflicts. Altogether, this work presents a data-driven view of community interactions and conflict, and paves the way towards healthier online communities.

ACM Reference Format:
Srijan Kumar, William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2018. Community Interaction and Conflict on the Web. In *WWW 2018: The 2018 Web Conference, April 23–27, 2018, Lyon, France.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3178876.3186141

1 INTRODUCTION

User-defined communities are an essential component of many web platforms, where users express their ideas, opinions, and share information. These communities not only provide a gathering place for *intracommunity* interactions between members of the same community, they also facilitate *intercommunity* interactions, where
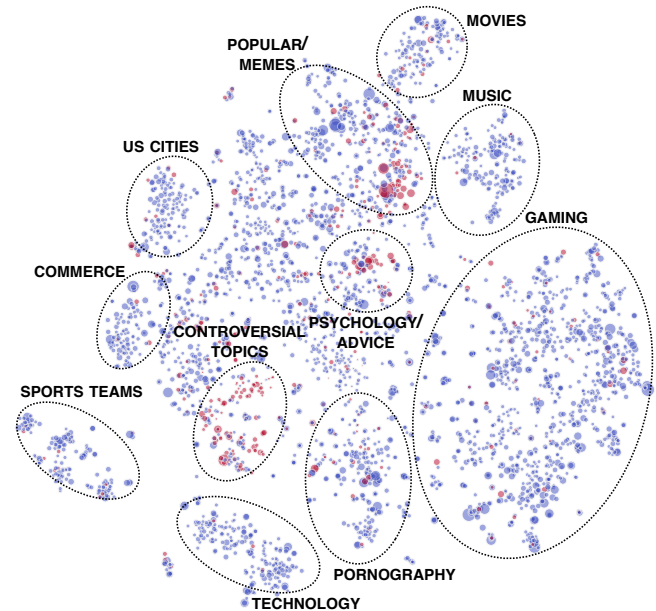
Figure 1: Communities in Reddit: each node represents a community. Red nodes initiate more conflicts, while blue nodes do not. Communities are embedded using user-community information, as described in Section 6. Figure best viewed in color.

members of one community engage with members of another. Studies of intercommunity dynamics in the offline setting have shown that intercommunity interactions can be positive—leading to the exchange of information and ideas [6, 22, 65, 70]—or they can take a negative turn, leading to overt conflicts between community members [31, 58, 75]. Due to the prevalence of community-level interactions on web platforms, understanding their mechanisms and impact on users is important to foster positive engagement between communities, and to reduce the adverse effects of intercommunity conflicts on users.

However, analyses of intercommunity interaction and conflict are largely absent in previous work on web communities, which tend to focus on community detection [18, 80], interactions within individual communities [38, 47, 49], or how users spread their time between multiple communities [26, 76, 83]. Moreover, existing research in social psychology on conflicts between small groups in lab settings [31, 70, 75] does not give us a window into the fine-grained details of how individual conflicts start on the web, the details of

user behavior during these interactions, and their impact on the individuals involved.

A consequence of this methodological limitation is that we lack the tools to predict and mitigate intercommunity conflicts in large-scale web environments. Instances of one online community harassing, "brigading", or "trolling" another will inevitably occur [23, 28, 51], but how can communities defend against these attacks? Studies on "bridging echo chambers" [21] and the positive effects of intergroup dialogue [5, 64] suggest that direct engagement could be effective for mitigating such conflicts, but there is also ample evidence suggesting that the best defense is simply to ignore the anti-social behavior of the attacking community [7]. Therefore, in web community conflicts, is it more effective to ignore and isolate these attacking users, or is it better to directly engage with them? We answer these questions in this work.

Present work. We provide a large-scale view of intercommunity interactions and conflict through the lens of Reddit, a popular website where users create and participate in interest-based communities called subreddits. We analyze 40 months of data, containing 1.8 billion comments made by over 100 million users across 36,000 communities [3].

Naturally, there are no explicit labels of when communities interact with and attack each other. A key methodological innovation in our work is that we identify and analyze concrete instances of intercommunity interaction and conflict: we identify cases where a post in one "source" community hyperlinks to a post in another "target" community, and we create a null model of user activity to detect instances where these hyperlinks mobilize a significant number of users to comment in the target post. We further use crowd-sourced labels to classify the sentiment of these *cross-linking posts* to identify cases of *negative mobilization*, where users were explicitly mobilized by posts with negative sentiment.

To give an example of such negative mobilization, the following post was made on the 'r/conspiracy' community, which criticizes and links to a post in 'r/Documentaries' community:

> **Come look at all the brainwashed idiots in r/Documentaries**
> Seriously, none of those people are willing to even CON-SIDER that our own country orchestrated the 9/11 attacks. They are all 100% certain the "turrists" were behind it all, and all of the smart people who argue it are getting downvoted to the depths of hell. Damn shame. Wish people would do their research. The link is [CROSS-LINK].

This post led to several members of r/conspiracy posting angry and uncivil comments on the cross-linked r/Documentaries' post.

Here we provide a data-driven view of these negative mobilizations, which can broadly be divided into three phases (Figure 2). The first phase is *initiation*: a user in the source community makes a cross-link to the target community that (potentially) mobilizes a subset of users (Figure 2, left). The second phase is *interaction*: once the cross-linking post has mobilized a subset of users, these attackers begin to comment on the comment thread of the target post and interact with members of the target community (*i.e.*, the defenders; Figure 2, middle). The final phase is *impact*: even after the negative mobilization is over, the event may have long-term impacts on the behavior of the attacking and defending users (Figure 2, right). By analyzing how thousands of negative mobilizations
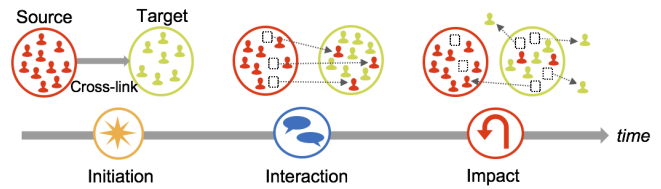

Figure 2: Timeline of interactions: community interactions can be divided into three phases: (i) initiation, where a cross-link is created from the source to the target community, (ii) interaction, where members of the source interact with those in the target, and (iii) long-term impact, where some source members may stay in the target and some target members may leave.

proceed through these three phases, we characterize the types of users and communities who initiate and participate in intercommunity conflicts, and highlight possible mitigation strategies that defending communities could use to minimize their adverse effects.

Initiation of mobilizations. Our findings show that a small set of communities is responsible for the majority of negative mobilizations: 74% of negative mobilizations are initiated by 1% of source communities. Moreover, we find that these interactions generally occur between highly similar communities. Figure 1 shows a learned, two-dimensional social map of the Reddit communities (see Section 6 for details), where each node is a community and its redness indicates the proportion of cross-links it generates with negative sentiment. The relatively few conflict-initiating (*i.e.*, red) communities are concentrated in three dense regions of this Reddit social map. Surprisingly, at the user level, we find that while negative mobilizations are initiated by highly active members of the source community, the actual perpetrators of negative mobilizations are users that are significantly less active.

User interactions during mobilizations. By analyzing the user-user replies on the target threads during negative mobilizations, we gain fine-grained insight into dynamics of user behavior in intercommunity conflicts. We create two variants of the PageRank algorithm, which measure the flow of information from the perspective of the attackers and defenders, respectively. Using these measures, we find that an important marker of negative mobilizations is an echo chamber effect, where participants preferentially interact with members of their "home" community, *i.e.*, attackers talk to other attackers, and defenders to other defenders. Furthermore, we find that attackers tend to single out and collectively "gang-up" on a small set of defenders.

Impact of mobilizations. Negative mobilizations have a long-term adverse impact on the involved users and communities. We find that negative mobilizations lead to a "colonization" effect, where defenders reduce their participation in the target community while attackers become more active. However, building off our analysis of user-level interactions, our results suggests a possible route for mitigating this adverse outcome: we show that a reduction in the echo-chamber effect—marked by an increased interaction between attackers and defenders—and an increase in defenders' use of anger words towards attackers, are associated with decreased rates of colonization and increased future participation rates for the defenders. Thus, it appears that increased engagement and a

增加参与，增加防御，相比忽略或孤立攻击方的策略更好

<mark>more fierce defense may be a more effective mitigation strategy, compared to ignoring or isolating the attacking users.</mark>

Predicting mobilizations. Lastly, we develop models to predict whether a cross-link will lead to a mobilization. Our model combines recent advancements in deep learning on social networks with a novel variant of recurrent neural network LSTM model, which we call "socially-primed" LSTM model. Our model achieves an AUC of 0.76 in predicting if a cross-link will lead to a mobilization, significantly outperforming a strong baseline based on expert-crafted features, which achieves an AUC of 0.67. This model could be used in practice to create early warning systems for community moderators to predict potential negative mobilizations as soon as the cross-linking post is created, and help moderators to curb the adverse effects of intercommunity conflicts.

Altogether, our results shed light on how intercommunity interaction and conflicts occur on the web, and pave the way towards the development and maintenance of healthier online platforms.

2 DATA AND DEFINITIONS

Users on Reddit form and join interest-based communities called subreddits (*e.g.*, 'r/Documentaries' or 'r/StarWars'), and within these communities they post and comment on content (*e.g.*, images, videos, links to articles, etc.). We use 40 months of Reddit post and comment data, from January 2014 to April 2017 [3], to analyze the interactions between these subreddit communities. All our analysis was conducted with publicly available data [3], and relevant codes and data are available on our project website [1].

As there are no explicit labels of intercommunity interaction on Reddit, we leverage hyperlinks where a post in one community (which we call the *source*) links to a post in another community (which we call the *target*). These *cross-links* initiate a (possible) interaction between the two communities (*i.e.*, the first step *'initiation'* in the timeline of Figure 2). By following the cross-link, users may be *mobilized* from the source to comment on the linked post in target community, thus leading to intercommunity *'interactions'* (step two in Figure 2). After removing overlapping cross-links, we obtain 137,113 cross-links made between 36,000 communities, which is the set we analyze.

User activity on Reddit primarily occurs in discussion threads associated with posts. In order to identify the explicit effect of cross-links and to control for community-level differences and temporal confounds in our analysis of these threads, we use a matching process to provide a reference point for the posts that are involved in a cross-link (*i.e.*, both the source post that contains the cross-link and the target post that is linked to). In particular, for each post p that we analyze, we select a *matched post* from the same community that was created closest in time to p and that does not have any outgoing or incoming cross-links. The temporal restriction ensures that hourly, daily, or weekly changes in user behavior are controlled for, and selecting matched comparison posts from within the respective communities ensures that we are not confounded by community-level differences.

For the purpose of our analysis, we also define the *members* of the source, S, and target, T, communities at the time of each cross-link. In particular, for a cross-link that is made on day d, we define members of community S (resp., T) as users who have made at least one comment in S (resp., T) in the 30 days prior to d, but who did
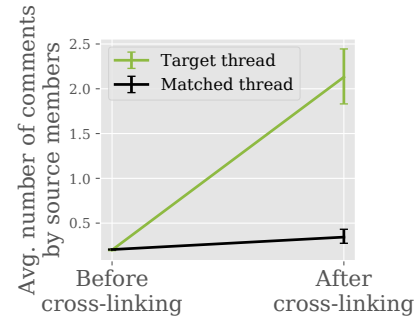

Figure 3: Cross-links lead to an increase in number of comments by source members in the target thread.

not comment in T (resp., S) during this time period.[1] In other words, members are users who participated in community S (resp., T) but not in community T (resp., S) before the cross-link was made.[2] To provide baseline comparisons for these users, we again use a matching process. For each member of S who comments on the target thread due to mobilization, we sample a random comparison member of community S who did not do so. This *matched user* is selected such that it made similar number comments in the past 30 days as the source member. An analogous process is used to match members of community T.

2.1 Defining mobilizations

We start by identifying cases of mobilization. We define mobilizations as cases where a cross-link leads to an increase in the number of comments by current source members on the discussion thread of the target post. Identifying such mobilizations is challenging because simply counting the number of source community members who comment in the target thread ignores that current source members may comment on the target community at random, instead of as an effect of the cross-link. Therefore, to identify cases where source members are mobilized *due to the cross-link*, we compare against a null model that measures the expected rate of comments of source members on the target thread. To control for the initial popularity of the posts, the null model is created using comments made on the matched target posts, restricting to cases where the target and matched posts have a near-equal number of comments before the cross-link is made (in particular, where the difference in their comment counts is less than 5). We compare the number of comments made by source members on the two threads, within a 12 hour window before and after the cross-link is created.

Figure 3 shows the change in average number of comments made by source members in the cross-linked target thread compared to a matched thread, restricting to those pairs whose pre-cross-link comment counts are equal. However, after cross-linking, there is an increase in both the threads: on average, the matched threads exhibit a 1.6× relative after-to-before increase, while the target

---

[1] But note that for analyses of user history, we always ignore comments made within ±3 days of the cross-link to remove possible confounds.
[2] We focus on exclusive members of the two interacting communities, following research conducted in the social psychology literature [31, 70, 75] where participants are always exclusively members of either communities. Other types of users, such as common members, conflate the observations of the two communities, while new users or non-members add little additional information. Their effects can be studied in future research.

threads exhibit an 8.8× increase. The baseline 1.6× increase for the matched threads illustrates the expected increase in comments by source members under our null model (*i.e.*, irrespective of cross-links). Based on these observations, we define *mobilizations* as cases where a cross-link leads to an increase in the number of comments by source community members in the target thread that is more than the baseline rate (*i.e.*, a >1.6× after-to-before increase). This gives a total of 22,075 mobilizations, which is about 16% of the total set of cross-link posts we study.

2.2 Classifying the sentiment of mobilizations

To further categorize interactions, we classify cross-links based on the sentiment of the source post. Overtly negative intent of the source post towards the post in the target community may signal outgroup derogation [30], which is a fundamental component of intergroup conflict [73, 75].

We collected labels from Mechanical Turk crowdworkers, who were shown source and target posts pairs, and were asked to label the sentiment of the source post towards the target as either negative, neutral, or positive. The workers reported very few instances of positive intent, so we merged the positive and neutral classes. We obtained two labels each (negative vs. neutral) for a randomly sampled set of 1020 pairs, with inter-rater agreement of over 0.95. We then converted the text of the source post[3] into a large set of features: lexical features from the LIWC [77] and VADER [36] lexicons, as well as stylistic linguistic features, such as average word length, readability scores, and punctuation counts.[4] Using a Random Forest classifier, our classification model achieves an accuracy of 0.80 using 10-fold cross validation.[5] We used this classifier to categorize the remaining source posts, and it labeled 8% of them as negative.

Coupling this sentiment score with mobilization, we call mobilizations that are initiated with negative sentiment as *negative mobilizations* (a total of 1809 cases), and those that start with neutral sentiment as *neutral mobilizations* (a total of 20266 cases).

Here, we also define two terms: *attackers* are the members of the source community, and *defenders* are the members of the target community, who comment in the (cross-linked) target thread during a negative mobilization. Note that we use these terms as they are clear and intuitively understandable. However, not all "attackers" are necessarily negative while commenting (though we do find that they most often are (*c.f.* Section 4)). Similarly, "defenders" may not be defending during a negative mobilization as such, but the terminology is used for ease of explanation.

3 INITIATION OF MOBILIZATIONS

We start our observations with the first phase, the *initiation* of intercommunity interactions, as shown in Figure 2. We aim to characterize the properties of the key entities involved in these interactions: the two communities and their participating members.
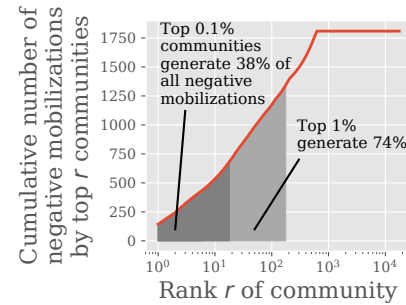

Figure 4: A small number of communities initiate the majority of conflicts.

3.1 Which communities initiate negative mobilizations?

Contrary to in-lab studies which bring together two groups to elicit interactions and intergroup conflict [31, 70, 75], web-based platforms have thousands of communities that could potentially interact. So, among these communities, which tend to initiate negative mobilizations, and which communities do they target?

We start by looking at the properties of the initiating community. We find that most negative mobilizations are initiated by a small number of communities (see Figure 4)—less than 0.1% and 1% of source communities are responsible for 38% and 74% of the negative mobilizations, respectively. This means that these handful of communities are hubs of negatively mobilizing users. This finding echoes those on anti-social behavior, which show that troll-like behaviors are concentrated in small number of communities [12], and that taking precautionary measures, such as banning particularly egregious communities, can be effective in curbing this behavior [10].

Next, we look at the similarity between the source and target communities involved in mobilizations. Previous work has suggested that conflicting communities are likely to focus on topically similar areas [14, 25], so we evaluate this hypothesis here. Looking at the pair of interacting communities, we quantify the similarity of two communities as their tf-idf post similarities.[6] We find that negative mobilizations (and mobilizations in general) tend to occur between highly similar communities—the average tf-idf content similarity between the source and target communities in mobilizations is 1.5× higher than the average value computed between random community pairs (0.51 *vs.* 0.34; $p < 0.001$).[7] Manually examining these community pairs reveals that they tend to focus on similar topics, but have different views on the subject matter (*e.g.*, r/conspiracy *vs.* r/worldnews, or r/mensrights *vs.* r/againstmensrights), which fits with previous discussions of this phenomenon [14, 25].

3.2 Which users are involved in negative mobilizations?

In web-based communities comprised of thousands or even millions of users—with varying levels of participation—only a small fraction

---

[3]We removed words common in the source and target posts as source post may quote (part of) the target post.
[4]The full feature list is available in the online appendix [1].
[5]We used the implementation in scikit-learn package [60] with forests of 400 trees.

[6]The tf-idf similarity value [69] is computed over all communities posts using the top-10,000 words in terms of overall frequency-rank on Reddit.
[7]All reported p-values are based on the Wilcoxon signed-rank test for paired comparisons and the Mann Whitney U-test otherwise [71].
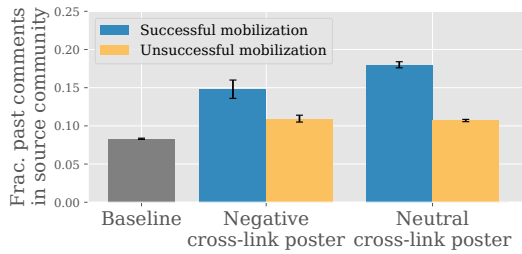
Figure 5: Successfully negative mobilizing posts are created by highly active 'core' members of the source community.
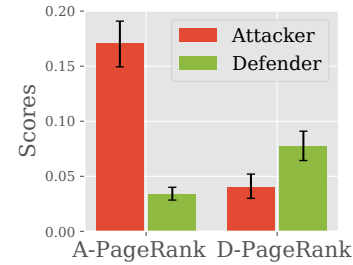

Figure 6: Echo chambers are formed during negative mobilization: attackers have higher A-PageRank scores than defenders, and defenders have higher D-PageRanks.

of users will actually be involved in any individual negative mobilization. So, what are the properties of the users that get involved as opposed to the ones that do not? We answer this question by looking at the user who creates the cross-link post that initiates the negative mobilization, as well as the users that get mobilized and participate in the resulting discussion.

Examining the cross-link creator, we see that negative posts with cross-links are created by users that are 10% more active in the source community compared to its random matched user ($p < 0.001$, Figure 5(a); see Section 2 for details on matching). As not all cross-links lead to mobilization, we further see that users that are successful in mobilizing others are significantly more active than the ones who are not ($p < 0.001$). Thus, we see that highly active members of the source community are responsible for initiating negative mobilizations. Similar observations are made in the case of neutral mobilizations.

But who are the actual perpetrators of negative mobilizations? Are they also highly active members of the source community? To answer this, we compare the mobilized attacking source members (the attackers) to their matched counterparts. We find that the attackers are, in fact, significantly less active in the source community (fraction of past comments in source community is 0.30 vs. 0.17, $p < 0.001$). We also find that the attackers expressed more 'anger' in their past comments;[8] attackers use anger words (from the LIWC lexicon [77]) 1.2× more than their matched counterparts (0.31 vs. 0.26, $p < 0.001$).

Similarly, examining the members of the target community who are mobilized to defend during these attacks, we find analogous results—these defenders are less active in the target community (fraction of past comments in target community is 0.27 vs. 0.19, $p < 0.001$) and use 2.2× more anger words than their matched counterparts (0.32 vs. 0.145, $p < 0.001$). Thus, we see that the defending and attacking users are similar: they tend to be less active in their home communities and have used more anger words in the past.

Overall, we find that negative mobilizations are initiated by a handful of communities that attack highly similar communities. While these interactions are initiated by the highly active users of the source community, the attackers and defenders who actually get mobilized to participate in the negative mobilization are much less active than them—a finding that dovetails with psychological studies showing that peripheral group members are more likely to

engage in public displays of outgroup derogation [57]. Prior work has argued that this phenomenon is due to the desire of peripheral members to increase their ingroup status (*ibid.*), and interestingly we see this effect even on Reddit, which is pseudonymous.

4 USER INTERACTION DURING MOBILIZATIONS

Now that we have characterized the entities involved in negative mobilizations—both in terms of the communities and users involved—we turn to the task of characterizing the dynamics of user interactions during these events. Do attackers and defenders talk to each other, or do they create separate bubbles where they talk to other similar users, creating an echo chamber-like effect?

We answer these questions by analyzing the user-to-user reply network of discussions on the target post, in the second phase of the mobilization, *i.e.*, the *interaction* phase (see Figure 2). This gives us the unique ability to study thousands of highly rich user-user interactions during negative mobilizations, providing a new degree of both scale and granularity to the study of intercommunity conflict [74].

Negative mobilizations lead to negative interactions. We find that negative mobilizations have a negative impact on the sentiment and civility of the target discussion thread. We compare the comments made on the target threads of negative mobilizations and their matched threads (see Section 2 for matching details). The target threads are more likely to contain anger words (44% increase; $p < 0.001$) and are 25× more likely to have a comment removed by a moderator (deletion rates of 0.205 *vs.* 0.008; $p < 0.001$).[9] For comparison, these adverse effects are drastically reduced in neutral mobilizations (*i.e.*, no significant change in deletion rate, $p > 0.05$).

Quantifying echo chambers in negative mobilizations.
Do attackers and defenders talk to each other during negative mobilizations? Or do they occupy separate worlds, thus exhibiting homophily and creating echo chambers? We answer this question by constructing user-user reply networks from comments made on the target thread. The network is directed and weighted, where the edge weight from user i to user j corresponds to the number of times i directly replied to one of j's comments.

Our analysis suggests a homophily and echo chamber effect, where attackers preferentially interact with other attackers and

---

[8]This is calculated using comments made in past 30 days.

[9]Note that this result is also statistically significant after macro-averaging deletion rates across communities, indicating that it is not simply due to the target communities have higher baseline deletion rates.

defenders with other defenders. For example, examining the direct interactions between users (*i.e.*, cases where *i* replied to *j* or vice-versa), we find that attackers interact 2× more with other attackers compared to their interactions with defenders, while defenders interact 20× more with other defenders compared to attackers ($p < 0.001$). We quantify this echo chamber effect using two variants of the PageRank score [59]: the Defender PageRank (D-PageRank) and Attacker PageRank (A-PageRank), which quantify the centrality of a user in the reply network from the perspective of defenders and attackers, respectively. Like the classic PageRank statistic [59], the D- and A-PageRank values for a user, *i*, correspond to the probability of a random walk visiting the node corresponding to *i*, where this random walk follows directed edges with transition probabilities proportional to edge weights and "teleports" to a new random start point with a probability of 0.25 at each step. However, unlike the standard PageRank, we compute the D-PageRank and A-PageRank by restricting the teleport set to defender nodes and attacker nodes respectively, similar to the approach used by Garimella et al. to quantify controversy [20].

Figure 6 shows the values of these scores. We see the attackers have significantly higher A-PageRanks—meaning that they are more likely to be visited on random walks starting from other attackers—while defenders have significantly higher D-PageRanks. Thus, we see that the discussion threads during negative mobilizations seem to be marked by a bifurcation between attackers and defenders, where members primarily talk to other members from their "home" community. Similar echo chamber effects have been found during controversial discussions in social media platforms [13, 16, 19, 67].

Attackers "gang-up" on defenders. Of course—despite the echo chamber effect elucidated above—attackers and defenders will inevitably have some direct interactions. But what do these interactions look like? Examining the defenders' A-PageRank scores, we find that while 83% defenders have exactly zero score, a very small fraction of defenders (1.14%) have an A-PageRank score which is at least ten times the mean value among all attackers and defenders. This highly skewed distribution of the defenders' A-PageRank scores indicates that a small set of defenders are involved in most of the interactions with the attackers. Additionally, we find that the comments made between attackers and defenders employ very angry language—attackers use anger words more frequently when talking to defenders than when talking to fellow attackers (0.015 *vs.* 0.011; $p < 0.001$), and vice-versa for defenders (0.017 *vs.* 0.014; $p < 0.001$). Together, these results suggests that while attackers do not communicate with most of the defenders, they tend to single out and "gang-up" on a handful of defenders.

Overall, we find that negative mobilizations lead to the creation of echo chambers in the target discussion, where attackers primarily talk to other attackers and defenders talk to other defenders. But whenever they do talk to each other, they exchange highly angry comments. Moreover, our analysis shows that attackers tend to "gang-up" to attack some defending users, while other defending users do not engage with the attackers.
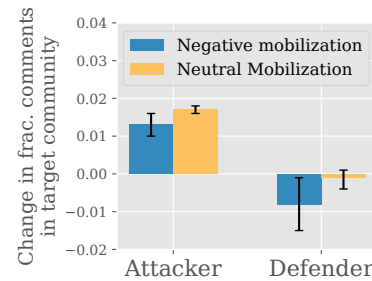

Figure 7: Impact of negative mobilization: defenders become less active in target community, while attackers become more active.

5 IMPACT OF MOBILIZATIONS

As seen in the previous section, attackers "gang-up" on defending users during negative mobilizations. But in the long-term, do these attacks have any impact on the involved users? And what should the defenders do to prevent the negative effects of these events—should they ignore the attackers during discussions, or should they actively engage with them? In this section, we answer these questions by studying the third phase of mobilization (Figure 2), *i.e.*, the *impact* phase, where we measure changes in user activity after the conflict is over.

5.1 Quantifying impact

We measure the impact of mobilization after it is over as the change in attackers' and defenders' participation in the target community. All changes are measured as differences between average activity levels of the users in the 30 days before mobilization starts and in the 30 days after the mobilization ends (measured as 3 days after the initial cross-link post was created). As a baseline, we again compare with randomly matched comparison users (see Section 2 for details).

In Figure 7, we measure the change (after minus before) in the fraction of comments that attackers and defenders make in the target community after negative mobilizations end (with analogous values for neutral mobilizations shown). A positive (negative, *resp.*) change indicates an increased (decreased, *resp.*) engagement in the target community. We see that after negative mobilizations, attackers post more frequently in the target community while defenders post less frequently ($p < 0.001$), indicating that negative mobilizations often lead to "colonization", where members of the attacking community become regular members of the target community. This colonization may lead to further adverse effects for the target community, since these incoming users are generally angrier than average and tend to violate community norms (*c.f.*, Section 4).

In contrast, neutral mobilizations lead to "immigration", where the mobilized users become more active in the target community, with no significant change for the 'defenders' (Figure 5). Moreover, as mobilized users in neutral mobilizations are more civil (*c.f.*, Section 4), this effect could lead to an increase in positive cross-community exchanges, which is an open area for future research.
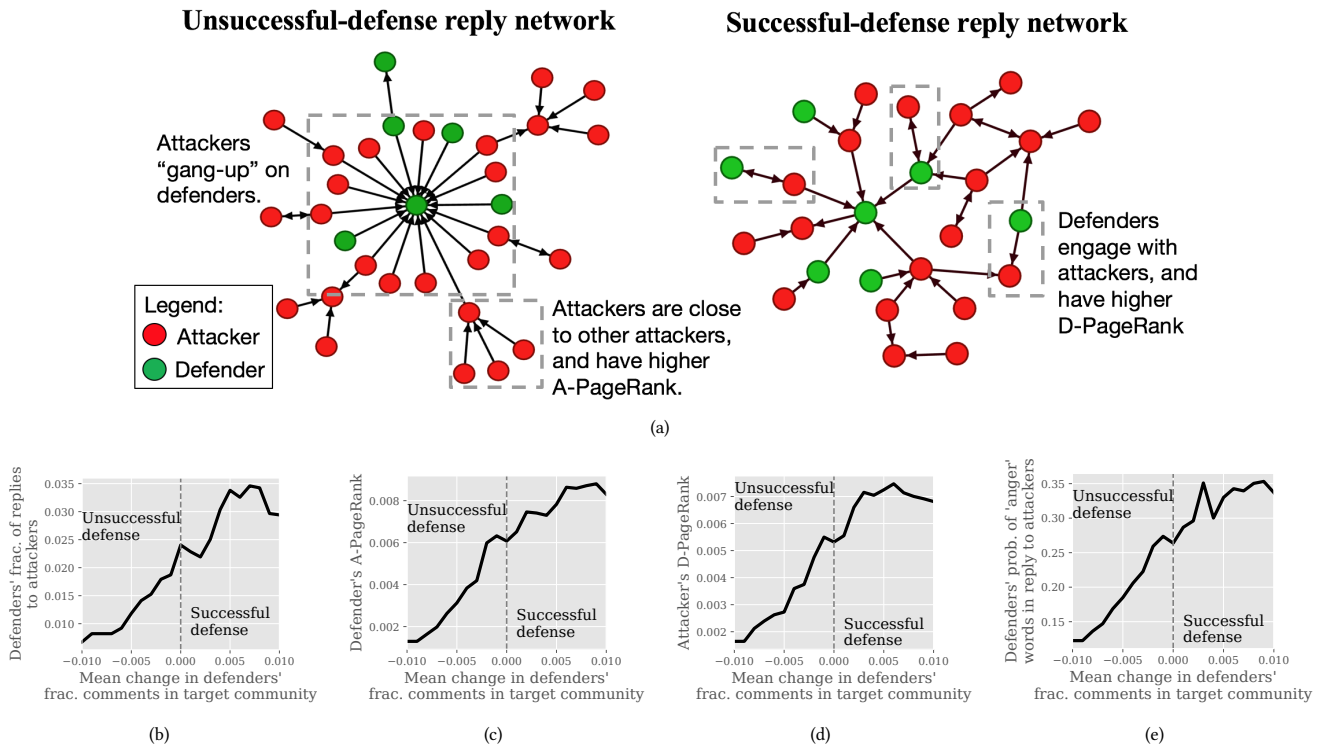
Figure 8: (a) Two reply networks on two real target threads from our data showing key characteristic differences between successful and unsuccessful defense networks. In an unsuccessful defense, attackers "gang-up" on defenders, while defenders engage directly with attackers in successful defenses. In successful defenses: (b) defenders reply more to attackers, (c) defender's A-PageRank is higher, (d) attacker's D-PageRank is higher, and (e) defenders use more angry words in replies to attackers.

5.2 What makes a defense successful?

As negative mobilizations adversely affect the target community, are there any possible ways of preventing this from happening? Is ignoring attackers the optimal way, as in the case of anti-social behavior such as trolling [7], or is direct engagement more effective, as suggested in hostility reduction theories [5, 63, 64]?

In order to understand strategies that may be useful for mitigating these adverse impacts of negative mobilizations, we compare cases of successful and unsuccessful defense. We define a defense to be 'successful' when defenders become more active in their community after the negative mobilization ends, while an 'unsuccessful' defense is when the defenders become less active. [10] We compare the 10% most and 10% least successful defenses and report differences in mean statistics between the two classes.

Figure 8(a) illustrates the largest connected components of two real reply networks from our data—one for an unsuccessful defense and another for a successful defense, with several observations. The accompanying Figures 8(b–e) show characteristic differences between successful and unsuccessful defenses overall. The x-axis in these figures varies the mean change (after minus before) in defender's proportion of comments in the target community after the negative mobilization ends—positive values indicate increased

proportion (i.e., successful defense) and negative values indicate a decrease (i.e., unsuccessful defense), relative to the expected baseline change (calculated from the matched users defined in Section 2). The plots are smoothened as moving average (window size of ±5) to remove minor variations. Together, these figures show that a defense is successful when defenders tend to engage in a direct fierce dialogue with the attackers, thus preventing them from "ganging-up" on defenders, and breaking echo chambers that generally form during negative mobilization (see Section 4). We explain the figures in detail in the next few paragraphs.

While there is no statistically significant increase in the number of defenders ($p = 0.05$), we find that successful defenses are marked by a strong increase in fraction of defenders' replies towards attackers (Figure 8(b); correlation coefficient = 0.97). Further, we use the two novel PageRank variants, A-PageRank and D-PageRank, developed in Section 4. Successful defenses have defenders with higher A-PageRank (0.036 *vs.* 0.031, $p < 0.001$; Figure 8(c)) and attackers with higher D-PageRank (0.052 *vs.* 0.028; $p < 0.001$; Figure 8(d)). Thus, defenders facilitate more discussions with attackers in successful defense and there is significantly more cross-communication between them, preventing the formation of echo chambers.

But does this increased cross-communication entail a civil discussion or a fierce debate? We answer this by calculating the probability of users using 'anger' words (from LIWC dictionary [77]) in direct replies by attackers to defenders, and vice-versa. We find that in

---

[10]Alternate definitions of 'success', *e.g.*, those involving quality of future comments, can be explored in future work.
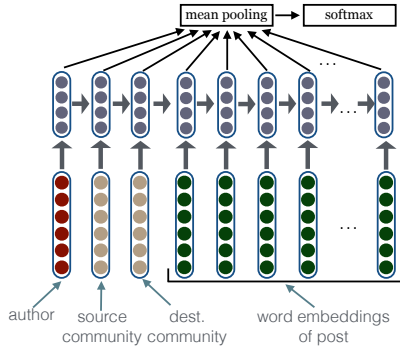
Figure 9: Socially-primed LSTM architecture.

successful defenses, defenders use more anger words towards attackers than attackers do to defenders (0.017 *vs.* 0.015; $p < 0.05$), while in unsuccessful defenses, both are equally angry towards each other ($p = 0.37$). This shows that defenders become more aggressive towards attackers in cases of successful defense.

Overall, we find that negative mobilizations often lead to "colonization", where after the conflict is over, attackers become regular members of the target community, while defenders become less active there. However, we find that this negative outcome is prevented when defenders break echo-chambers and engage in direct heated conversations with attackers.

6 PREDICTION OF MOBILIZATIONS

With the analysis in the previous sections, we have shown that negative mobilizations can have long-term adverse impact. So, can we predict whether a post will lead to a negative mobilization as soon as it is created? An efficient model can then be used to create an early warning system for community moderators to take appropriate precautions. In this section, we focus on this prediction task.

We create a novel "socially-primed" LSTM model along with a feature-based baseline for comparison. Our deep learning model leverages both the text of the post as well as the user-to-community interaction network for predictions. We show that our model significantly outperforms the feature-based baseline for the task of predicting whether a cross-linking post will mobilize users.

6.1 Feature-based baseline model

As a baseline, we create a traditional feature based model, with a large set of 263 hand-crafted features, including[11]

- Post features: Lexical count statistics from the post (*e.g.*, from LIWC [77] and from VADER [36]) and tf-idf similarity values between the post content of the source/destination communities.
- User features: Activity levels of the post creator (*e.g.*, fraction of posts in target) and averaged lexical features of the user's previous posts.
- Community features: tf-idf similarities between source and target content.

These features are used both as a baseline in a Random Forest model [60] and combined together with the deep model as an ensemble, as explained later.

---

[11]The full feature list is available in the project website [1].

6.2 Text, user, and community embeddings

While it is common for deep learning models to use textual information, our data also contains rich social information about users and communities, which we can leverage in our predictions. Therefore, we convert text, user, and communities into 300-dimensional embeddings, which will be used as input to our deep learning model. To generate text embeddings, we use standard techniques.[12] To generate user and community embeddings, we innovatively use the 'which user posts in which community' network to convert users and communities into embedding vectors with the following intuition: two community embeddings will be close together if similar users post on them, and two user embeddings will be similar if they tend to post on the same communities.

More formally, we train user and community embeddings based on the bipartite multigraph between users and communities, where one user-community edge, (u_i, c_j), exists in the graph for each time a user, u_i, posts on community, c_j. Adapting the approach used in a number of recent successful works [27, 53, 62], we learn user and community embeddings, $\mathbf{u}_i \in \mathbb{R}^d$, $\mathbf{c}_j \in \mathbb{R}^d$, from this graph using a negative-sampling optimization algorithm with the following loss:

$$\mathcal{L} = \sum_{(u_i, c_j) \in \mathcal{E}} -\log(\sigma(\mathbf{u}_i^\top \mathbf{c}_j)) - K \cdot \mathbb{E}_{c_n \sim P_n} \log(-\sigma(\mathbf{u}_i^\top \mathbf{c}_n)), \quad (1)$$

where $\mathcal{E}$ is the set of edges in the bipartite multigraph, P_n is a uniform distribution over all communities (used to generate negative samples), and $K = 5$ is the number of negative samples used.

We trained the embeddings using 100 passes over the full 40 months of Reddit posts, with stochastic gradient descent code adapted from Levy et al. [48]. Figure 1 shows the resulting community embeddings plotted in a two-dimensional visualization using t-SNE [50], with communities sized according to the number of posts that are made in it, and colored as the proportion of its outgoing cross-links that are negative (red indicates high negativity, and blue indicates low negativity). Communities were grouped manually into high-level categories, such as sports teams, technology, etc. We observe that most community clusters are mostly blue, while controversial topics (*e.g.*, cluster with r/conspiracy and r/theredpill) and advice communities generate a high number of negative cross-links.

These text, user, and community embeddings will be used as input to our socially-primed LSTM model explained below.

6.3 Socially-primed LSTM model

To combine social and textual data in a structured manner for predictions, we create a novel "socially primed" LSTM model. Using the word, user, and community embeddings as inputs, we train a recurrent LSTM model, as shown in Figure 9. LSTMs are a popular variant of recurrent neural networks (RNNs), *i.e.*, neural networks that operate on sequential input sequences, $[\mathbf{x}_1, ..., \mathbf{x}_T]$ [34]. The basic idea behind LSTMs is that they use a combination of dense neural networks to map the current input value, $\mathbf{x}_t \in \mathbb{R}^d$, along with a hidden-state vector, $\mathbf{h}_t \in \mathbb{R}^h$, to a new hidden state, $\mathbf{h}_{t+1} \in \mathbb{R}^h$.

In our socially-primed LSTM model, the input sequences are the concatenation of (i) the user embedding of the post author, $\mathbf{u}_i$, (ii) the community embeddings of the source and target communities, $\mathbf{c}_s$ and $\mathbf{c}_t$, and (iii) the word embeddings of the post text,

---

[12]Off-the-shelf GloVe word embeddings trained on 840 billion words [61].

$[\mathbf{w}_0, ..., \mathbf{w}_L]$ (where L denotes the length of the text).[13]Traditional LSTMs work with text data only, while 'social-priming' signifies that user and community information are additionally used for training. To predict mobilization, we take mean of the hidden LSTM states from each time step and then feed the resulting vector through a softmax (i.e., logistic) layer:

$$[\mathbf{h}_1, ..., \mathbf{h}_{L+3}] = \text{LSTM}([\mathbf{u}_i, \mathbf{c}_s, \mathbf{c}_t, \mathbf{w}_1, ..., \mathbf{w}_L]) \quad (2)$$

$$y = \frac{1}{1 + \exp\left(\boldsymbol{\theta}^\top \left(\frac{1}{L+3} \sum_{t=1}^{L+3} \mathbf{h}_t\right)\right)}, \quad (3)$$

where y denotes the probability of the post leading to mobilization, and $\mathbf{h}_t$ are the LSTM hidden states.[14]

6.4 Prediction results

The prediction task we consider is: whether a post will mobilize source members, using the full set of cross-links from Section 2. We randomly use 80% of the data for training, 10% for validation, and the remaining 10% as held-out test set.[15]

In this task, the Random Forest feature-based model (size 500) achieves an AUC of 0.67. For reference, we include the traditional LSTM model that only uses text information (*i.e.*, no user and community embeddings), which achieves an AUC of 0.66, almost equal to the random forest model. Our socially-primed LSTM model significantly outperforms both these models, with an AUC of 0.72. We also create a Random Forest ensemble model that uses all the hand-engineered features, user embeddings, community embeddings, and the average hidden state of the LSTM, further improving performance to an AUC of 0.76. This shows that both features and the LSTM model contribute towards the improved performance of the ensemble.

Overall, in this section, we created several machine learning prediction models and achieve high AUC of 0.76 for predicting if a post will mobilize users. These models can be used in practice to create early warning systems for community moderators to predict potential negative mobilizations as soon as the cross-link post is created, and help them to curb the adverse effects of these events.

7 FURTHER RELATED WORK

In addition to the computational social science and social psychology literature previously mentioned, our work also draws on and contributes to a rich line of research on online hate speech and anti-social behavior. Research on negative interactions in online platforms has focused primarily on controversies [4, 13, 20, 21, 45, 52] and anti-social behavior, in the form of trolling [11, 12, 42], sockpuppetry [39], harassment and cyberbullying [8, 23, 32, 35, 55, 66, 79], vandalism [43], hate speech [9, 15, 51, 56, 72], and others [17, 24, 40, 41, 44, 46, 54, 68, 78, 81]. While these studies cover a broad spectrum of anti-social behavior, they focus on user-level analyses. In contrast, our study focuses on interactions at the community-level. Recent research has also shown that organized negative behavior can spread across platforms [33, 82], and that banning communities that foster hate-speech was effective in decreasing hate speech on

Reddit [10]. Unlike these studies, which give important insights into the dangers of online negative behavior and effectiveness of correction measures, our research focuses on (negative) interactions between community-community dyads and their impact.

8 DISCUSSION AND CONCLUSIONS

Our findings provide a novel, large-scale view of intercommunity interaction and conflict on the web. By leveraging cross-links between posts made in communities and carefully controlling for baseline rates of user activity, we were able to identify explicit cases of intercommunity mobilization.

Our analysis shows that negative mobilizations have important long-term adverse effects, leading to processes of "colonization", where ill-behaved users come to dominate the target community. However, we also uncovered mitigating factors that correlate with decreased rates of these adverse outcomes—most prominently, that increased engagement in discussions between members of the interacting communities leads to improved outcomes. Further, we designed a novel socially-primed LSTM model, which combines user, community, and text embeddings, to predict whether mobilizations will occur with an AUC of 0.76.

Nonetheless, our analysis has limitations that point to interesting directions for future work. First, we study interactions between pairs of communities, but understanding interactions between more than two communities could reveal larger-scale dynamics of intercommunity interactions. Our analysis is also based on a single platform where users are pseudonymous, while the nature of community interactions and conflicts may be different where people use their 'real' identities (*e.g.*, Facebook or Twitter). Moreover, while we study the impact of mobilization on the involved users, an analysis of conversations in the rest of the community could further our understanding of the impact of intercommunity interactions. Additionally, the relation between intercommunity interactions and anti-social behavior, such as trolling [12] and sockpuppetry [39], is yet unexplored. Lastly, our analysis of the sentiment of intercommunity mobilizations is limited to instances of *explicit sentiment*, where the negative (or neutral) sentiment is clear to a relatively uninformed crowdworker. However, community restrictions may prevent creation of explicit derogatory posts and lead to instances of *implicit negative sentiment*. Detecting implicit sentiment and using it to study intercommunity interaction is an important open direction for future work.

By identifying, analyzing, and predicting intercommunity mobilizations, our work outlines a data-driven methodology for quantifying the impact of these interactions and predicting when they will occur. The management of highly negative and disruptive communities in particular is becoming increasingly important for multi-community platforms [29], and our analysis helps to pave the way towards the development of policies, strategies, and tools for promoting positive interactions on these platforms.

---

[13]We use a maximum length of $L = 50$, discarding words that occur after this point.
[14]We trained the model using PyTorch [2], the Adam optimizer [37], and a standard cross-entropy loss. The project website [1] contains data and code for replication.
[15]We do not restrict to negative mobilizations, since the sentiment labels themselves were generated by a machine learning classifier. Predicting negative mobilization can be done by coupling this classifier with the sentiment classifier developed in Section 2.

REFERENCES

[1] Online appendix. http://snap.stanford.edu/conflict.
[2] Pytorch v0.2. http://pytorch.org/.
[3] Reddit data dump. http://files.pushshift.io/reddit/. Accessed: 2017-10-27.
[4] A. Addawood, R. Rezapour, O. Abdar, and J. Diesner. Telling apart tweets associated with controversial versus non-controversial topics. In *Proceedings of the 2nd Workshop on NLP and Computational Social Science*, 2017.
[5] G. W. Allport. *The nature of prejudice*. Basic Books, 1979.
[6] V. Belák, S. Lam, and C. Hayes. Cross-community influence in discussion fora. *ICWSM*, 12:34–41, 2012.
[7] A. Binns. Don't feed the trolls! managing troublemakers in magazines' online communities. *Journalism Practice*, 6(4):547–562, 2012.
[8] J. Blackburn and H. Kwak. Stfu noob!: predicting crowdsourced decisions on toxic behavior in online games. In *Proceedings of the 23rd international conference on World wide web*, pages 877–888. ACM, 2014.
[9] P. Burnap and M. L. Williams. Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data Science*, 5(1):11, 2016.
[10] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, and E. Gilbert. You canâĂŹt stay here: The efficacy of redditâĂŹs 2015 ban examined through hate speech. In *Proceedings of the ACM Human-Computer Interaction*, 2017.
[11] J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil, and J. Leskovec. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 20th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, 2017.
[12] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec. Antisocial behavior in online discussion communities. In *Proceedings of the The International AAAI Conference on Web and Social Media (ICWSM)*, 2015.
[13] M. Conover, J. Ratkiewicz, M. R. Francisco, B. Gonçalves, F. Menczer, and A. Flammini. Political polarization on twitter. *Proceedings of the 5th International AAAI Conference on Web and Social Media (ICWSM)*, 2011.
[14] S. Datta, C. Phelan, and E. Adar. Identifying misaligned inter-group links and communities. *Proceedings of the ACM Human-Computer Interaction*, 2017.
[15] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati. Hate speech detection with comment embeddings. In *Proceedings of the 24th International Conference on World Wide Web (WWW)*. ACM, 2015.
[16] R. Faris, H. Roberts, B. Etling, N. Bourassa, E. Zuckerman, and Y. Benkler. Partisanship, propaganda, and disinformation: Online media and the 2016 us presidential election. *Berkman Klein Center for Internet & Society Research Paper*, 2017.
[17] E. Ferrara. Contagion dynamics of extremist propaganda in social networks. *Information Sciences*, 2017.
[18] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.
[19] D. Garcia, F. Mendez, U. Serdült, and F. Schweitzer. Political polarization and popularity in online participatory media: an integrated approach. In *Proceedings of the 1st Workshop on Politics, Elections and Data*. ACM, 2012.
[20] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis. Quantifying controversy in social media. In *Proceedings of the 9th ACM International Conference on Web Search and Data Mining (WSDM)*. ACM, 2016.
[21] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis. Reducing controversy by connecting opposing views. In *Proceedings of the 10th ACM International Conference on Web Search and Data Mining (WSDM)*. ACM, 2017.
[22] H. Giles. *Intergroup communication: Multiple perspectives*, volume 2. Peter Lang, 2005.
[23] J. Golbeck, Z. Ashktorab, R. O. Banjo, A. Berlinger, S. Bhagwan, C. Buntain, P. Cheakalos, A. A. Geller, Q. Gergory, R. K. Gnanasekaran, et al. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on Web Science Conference (WebSci)*. ACM, 2017.
[24] S. C. Guntuku, D. B. Yaden, M. L. Kern, L. H. Ungar, and J. C. Eichstaedt. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49, 2017.
[25] W. Hamilton, K. Clark, J. Leskovec, and D. Jurafsky. Inducing domain-specific sentiment lexicons from unlabeled corpora. *Proceedings of the 2016 Conference on Empirical Methods on Natural Language Processing (EMNLP)*, 2016.
[26] W. Hamilton, J. Zhang, C. Danescu-Niculescu-Mizil, D. Jurafsky, and J. Leskovec. Loyalty in online communities. In *Proceedings of 2017 The International AAAI Conference on Web and Social Media (ICWSM)*, 2017.
[27] W. L. Hamilton, R. Ying, and J. Leskovec. Representation learning on graphs: Methods and applications. *IEEE Data Engineering Bulletin*, 2017.
[28] C. Hardaker. Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. *Journal of Politeness Research*, 2010.
[29] C. Hauser. Reddit bans nazi groups and others in crackdown on violent content. *New York Times*, October 2017. [Online; posted 26-October-2017].
[30] M. Hewstone, M. Rubin, and H. Willis. Intergroup bias. *Annual Review of Psychology*, 53(1):575–604, 2002.
[31] M. E. Hewstone and R. E. Brown. *Contact and conflict in intergroup encounters*. Basil Blackwell, 1986.

[32] S. Hinduja and J. W. Patchin. *Bullying beyond the schoolyard: Preventing and responding to cyberbullying*. Corwin Press, 2014.
[33] G. E. Hine, J. Onaolapo, E. De Cristofaro, N. Kourtellis, I. Leontiadis, R. Samaras, G. Stringhini, and J. Blackburn. Kek, cucks, and god emperor trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. In *Proceedings of the The International AAAI Conference on Web and Social Media (ICWSM)*, 2017.
[34] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.
[35] H. Hosseinmardi, A. Ghasemianlangroodi, R. Han, Q. Lv, and S. Mishra. Towards understanding cyberbullying behavior in a semi-anonymous social network. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, pages 244–252. IEEE, 2014.
[36] C. J. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2014.
[37] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.
[38] R. E. Kraut, P. Resnick, S. Kiesler, M. Burke, Y. Chen, N. Kittur, J. Konstan, Y. Ren, and J. Riedl. *Building successful online communities: Evidence-based social design*. MIT Press, 2012.
[39] S. Kumar, J. Cheng, J. Leskovec, and V. Subrahmanian. An army of me: Sockpuppets in online discussion communities. In *Proceedings of the 26th International Conference on World Wide Web*, 2017.
[40] S. Kumar, B. Hooi, D. Makhija, M. Kumar, C. Faloutsos, and V. Subrahamanian. Rev2: Fraudulent user prediction in rating platforms. *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*, 2018.
[41] S. Kumar and N. Shah. False information on web and social media: A survey. In *Social Media Analytics: Advances and Applications*. CRC, 2018.
[42] S. Kumar, F. Spezzano, and V. Subrahmanian. Accurately detecting trolls in slashdot zoo via decluttering. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, pages 188–195. IEEE, 2014.
[43] S. Kumar, F. Spezzano, and V. Subrahmanian. Vews: A wikipedia vandal early warning system. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015.
[44] S. Kumar, R. West, and J. Leskovec. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th International Conference on World Wide Web*, 2016.
[45] H. Lamba, M. M. Malik, and J. Pfeffer. A tempest in a teacup? analyzing firestorms on twitter. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2015.
[46] K. Lee, P. Tamilarasan, and J. Caverlee. Crowdturfers, campaigns, and social media: Tracking and revealing crowdsourced manipulation of social media. In *ICWSM*, 2013.
[47] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in online social networks. In *Proceedings of the 19th international conference on World Wide Web (WWW)*. ACM, 2010.
[48] O. Levy and Y. Goldberg. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2014.
[49] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the Association for Information Science and Technology (JASIST)*, 58(7):1019–1031, 2007.
[50] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research (JMLR)*, 9(Nov):2579–2605, 2008.
[51] J. N. Matias, A. Johnson, W. E. Boesel, B. Keegan, J. Friedman, and C. DeTar. Reporting, reviewing, and responding to harassment on twitter. *arXiv:1505.03359*, 2015.
[52] Y. Mejova, A. X. Zhang, N. Diakopoulos, and C. Castillo. Controversy and sentiment in online news. *Computation and Journalism Symposium*, 2014.
[53] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 2013.
[54] T. Mitra and E. Gilbert. Credbank: A large-scale social media corpus with associated credibility annotations. In *ICWSM*, pages 258–267, 2015.
[55] K. Munger. Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, 39(3):629–649, 2017.
[56] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web (WWW)*, 2016.
[57] J. G. Noel, D. L. Wann, and N. R. Branscombe. Peripheral ingroup membership status and public negativity toward outgroups. *Journal of Personality and Social Psychology*, 68:127–127, 1995.
[58] U. of Oklahoma. Institute of Group Relations and M. Sherif. *Intergroup conflict and cooperation: The Robbers Cave experiment*, volume 10. University Book Exchange Norman, 1961.

[59] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[60] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research (JMLR)*, 2011.

[61] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[62] B. Perozzi, R. Al-Rfou, and S. Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 701–710. ACM, 2014.

[63] T. F. Pettigrew. Generalized intergroup contact effects on prejudice. *Personality and Social Psychology Bulletin*, 23(2):173–185, 1997.

[64] T. F. Pettigrew and L. R. Tropp. Does intergroup contact reduce prejudice? recent meta-analytic findings. *Reducing Prejudice and Discrimination*, 93:114, 2000.

[65] M. A. Rahim. *Managing conflict in organizations*. Transaction Publishers, 2010.

[66] J. Ratkiewicz, M. Conover, M. R. Meiss, B. Gonçalves, A. Flammini, and F. Menczer. Detecting and tracking political abuse in social media. *Proceedings of the 5th International AAAI Conference on Web and Social Media (ICWSM)*, 2011.

[67] M. H. Ribeiro, P. H. Calais, V. A. Almeida, and W. Meira Jr. "everything i disagree with is# fakenews": Correlating political polarization and spread of misinformation. *arXiv:1706.05924*, 2017.

[68] H. Saif, M. Fernandez, M. Rowe, and H. Alani. On the role of semantics for detecting pro-isis stances on social media. In *Proceedings of the CEUR Workshop*, volume 1690, 2016.

[69] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988.

[70] M. Sherif. *Group conflict and co-operation: Their social psychology*, volume 29. Psychology Press, 2015.

[71] S. Siegal. *Nonparametric statistics for the behavioral sciences*. McGraw-hill, 1956.

[72] L. A. Silva, M. Mondal, D. Correa, F. Benevenuto, and I. Weber. Analyzing the targets of hate in online social media. In *Proceedings of the 2016 The International AAAI Conference on Web and Social Media (ICWSM)*, 2016.

[73] H. Tajfel. Social psychology of intergroup relations. *Annual Review of Psychology*, 33(1):1–39, 1982.

[74] H. Tajfel. *Social identity and intergroup relations*. Cambridge University Press, 2010.

[75] H. Tajfel and J. C. Turner. An integrative theory of intergroup conflict. *The Social Psychology of Intergroup Relations*, 1979.

[76] C. Tan and L. Lee. All who wander: On the prevalence and characteristics of multi-community engagement. In *Proceedings of International World Wide Web Conference (WWW)*, 2015.

[77] Y. R. Tausczik and J. W. Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, Mar. 2010.

[78] W. Wang, L. Chen, K. Thirunarayan, and A. P. Sheth. Cursing in english on twitter. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing (CSCW)*. ACM, 2014.

[79] E. Wulczyn, N. Thain, and L. Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399. International World Wide Web Conferences Steering Committee, 2017.

[80] J. Xie, S. Kelley, and B. K. Szymanski. Overlapping community detection in networks: The state-of-the-art and comparative study. *ACM Computing Surveys (CSUR)*, 45(4):43, 2013.

[81] T. Yasseri, R. Sumi, A. Rung, A. Kornai, and J. Kertész. Dynamics of conflicts in wikipedia. *PloS One*, 7(6):e38869, 2012.

[82] S. Zannettou, T. Caulfield, E. De Cristofaro, N. Kourtelris, I. Leontiadis, M. Sirivianos, G. Stringhini, and J. Blackburn. The web centipede: understanding how web communities influence each other through the lens of mainstream and alternative news sources. In *Proceedings of the 2017 Internet Measurement Conference*, pages 405–417. ACM, 2017.

[83] J. Zhang, W. Hamilton, C. Danescu-Niculescu-Mizil, D. Jurafsky, and J. Leskovec. Community identity and user engagement in a multi-community landscape. In *Proceedings of 2017 The International AAAI Conference on Web and Social Media (ICWSM)*, 2017.