Extracting Inter-Community Conflicts in Reddit

Srayan Datta
University of Michigan
srayand@umich.edu

Eytan Adar
University of Michigan
eadar@umich.edu

Abstract

Anti-social behaviors in social media can happen both at user and community levels. While a great deal of attention is on the individual as an 'aggressor,' the banning of entire Reddit subcommunities (i.e., subreddits) demonstrates that this is a multi-layer concern. Existing research on inter-community conflict has largely focused on specific subcommunities or ideological opponents. However, antagonistic behaviors may be more pervasive and integrate into the broader network. In this work, we study the landscape of conflicts among subreddits by deriving higher-level (community) behaviors from the way individuals are sanctioned and rewarded. By constructing a *conflict network*, we characterize different patterns in subreddit-to-subreddit conflicts as well as communities of *'co-targeted'* subreddits .The dynamics of these interactions also reveals a shift in *conflict focus* over time. [1]

Introduction

Anti-social behavior in social media is not solely an individual process. Communities can, and do, antagonize other groups with collective anti-social behaviors. Similarly, both individuals and communities can be sanctioned in reaction to this behavior. On Reddit, for example, individuals can be banned or otherwise penalized (e.g., have their posts downvoted). Likewise, entire subreddits can also be sanctioned when multiple individuals violate general Reddit norms and use the community as a platform for generating conflict.

Critically, the form of anti-social behavior at the community level can be quite varied. The ability to identify and coordinate with others means that actions considered anti-social for the individual can be expanded to group settings. A group can thus act anti-socially— producing mass spamming, trolling, flame wars, griefing, baiting, brigading, fisking, crapflooding, shitposting, and trash talking—against both individuals and other subreddits (Kumar et al. 2018). On Reddit, as in other discussion boards, the ability to create (multiple) accounts under any pseudonym can further exacerbate such behaviors. Although a vast majority of users are generally norm-compliant, anonymity can lead to less inhibited behavior from users (Suler 2004). In aggregate, the re-

sult is an entire embedded network of subreddit-to-subreddit conflicts inside of the Reddit ecosystem. Research has found specific instances of these conflicts. Our goal is to inferentially identify the structure and dynamics of this *community-to-community* conflict network at scale.

To achieve this, we address a number of challenges. First among them is the lack of explicit group membership. Group 'membership' in Reddit, and systems like it, can be vague. While subscriptions are possible, individuals can display member-like behaviors by posting to subreddits they are not part of. Such behaviors—subscription and posting—are not, however, a clear indication of the individual's 'social homes.' An individual can display both social and anti-social behaviors within the community via posting. What we seek is not simply identifying an individual's 'home' but to also discriminate between *social homes* and *anti-social homes*.

To achieve this separation, we apply a definition that extends Brunton's construct of spam (Brunton 2012): a community defines spam as (messaging) behavior that is not consistent with its rules and norms. That is, we seek to separate *norm-compliant* behaviors that indicate social membership and those that are *norm-violating* (indicating an anti-social home). Rather than relying on a global definition of norms, we utilize the sanctioning and rewarding behavior of individual subreddits in response to norm violation and compliance respectively. An explicit measure we leverage is up- and down-voting on posted comments. While these are not the only sanctions and rewards, they are (a) consistently used, and (b) can be aggregated both at the individual and community levels. As we demonstrate below, inference based on these lower level signals can help identify broader conflicts.

A further appeal of the bottom-up approach is that the converse, top-down identification of sanctions at the subreddit level, does not provide a clear indication of conflict. First, this signal is sparse as the banning of subreddits remains rare. Except for explicit *brigading*, which are (hard to detect) coordinated attacks on another subreddit, community-based anti-social behaviors may not result in a community being visibly sanctioned. Second, even when a sanction is employed it may be due to other reasons than community-on-community attacks. For example, subreddits such as *fatpeoplehate* (a fat-shaming subreddit) and *europeannationalism* (a racist subreddit) have been banned, not necessarily due to any specific 'attack,' but rather non-compliance with

[1]The network datasets from our study are publicly available at https://github.com/srayandatta/Reddit_conflictgraph.

Reddit-wide norms on hate speech.

Our bottom-up inference is different in that we can identify pairs of social and anti-social homes and aggregate these to find conflicts. Specifically, we can find authors implicated in conflicts—which we call *controversial authors*—by identifying those that have both social *and* anti-social homes. From this, we can say that if multiple authors have a social-home in subreddit A *and* an anti-social home in subreddit B, then there is a directed conflict between A and B. By finding aggregate patterns using all Reddit comments from 2016 (9.75 million unique users and 743 million comments), we can construct the subreddit *conflict graph* at scale. Furthermore, we demonstrate how directed edges in our graph can be weighted as a measure of *conflict intensity*. The process of identifying conflict edges and their associated weights is complicated by inherent noise in behaviors and high-variance of community sizes. A specific contribution of our work is the use of different aggregation and normalization techniques to more clearly identify the conflict graph.

Using this graph, we can determine not only the broad landscape of community-to-community conflicts but to answer specific questions as well: Which subreddits are most often instigators of conflict (versus targets)? Are conflicts reciprocal and are they proportional in intensity? Does 'attacking' multiple subreddits imply broad member misbehavior or the work of just a few individuals? Are certain subreddits targeted 'together?' Do conflicts shift over time?

Briefly, we find that subreddit conflicts are often reciprocal, but the conflict intensity is weakly negatively correlated with the intensity of the 'response.' We also find the larger subreddits are more likely to be involved in a large number of subreddit conflicts due to their size. However, our analysis of the fraction of users involved can isolate situations where both relative and absolute counts of involved authors are high. Additionally, we find different patterns of conflict based on intensity. For example, a single subreddit targeting many others may divide its attention, resulting in decreased intensity across the targets. On the other hand, we find anecdotal evidence that subreddits which act as social homes to many controversial authors and have high average conflict intensity against other subreddits often display communal misbehavior. Because of the longitudinal nature of our data, we are also able to perform a dynamic analysis to isolate temporal patterns in the conflict graph. We find, for example, that subreddits that conflict with multiple other subreddits change their main 'targets' over time.

Our specific contributions are mapping the static and dynamic subreddit conflict networks across Reddit. We identify group membership and define the concept of social and anti-social homes as a way of defining conflicts. By analyzing the different static and temporal patterns in subreddit conflicts, we provide evidence for mechanisms that can identify communal misbehavior. We provide a baseline for quantifying conflicts in Reddit and other social networks with 'noisy' community structure and where individuals can (mis)behave in a communal fashion. Our work has implications in identifying community features which can be used to automatically monitor community (mis)behavior in such social networks as an early warning system.

Related Work

Conflicts in Social Media

Undesirable behavior in online communities are widely studied in social media research. Qualitative analysis often focuses on identifying and characterizing different types of inappropriate online behaviour or provides case studies in different forums. Analysis of Usenet news, for example, helped explore identity and deception (Donath 1999). General anti-social behaviors (Hardaker 2010) can also manifest in 'site-specific' ways as in the trolling and vandals on Wikipedia (Shachaf and Hara 2010), or griefing and combative strategies in Second Life (Chesney et al. 2009).

Predicting trolls and other anti-social behavior is another well explored research area. For example, researchers have identified a connection between trolling and negative mood which provided evidence for a 'feedback' loop that contributes to further trolling (Cheng et al. 2017). In the context of *prediction*, several studies focused on detecting certain anti-social behaviors on specific sites (e.g., vandalism in Wikipedia (Adler et al. 2011; Kumar, Spezzano, and Subrahmanian 2015)). Others have attempted to predict both anti-social behaviors or sanctions. Examples of the former include finding sockpuppets (same user with multiple accounts) on discussion sites (Kumar et al. 2017) and Twitter (Galán-García et al. 2013). Within Reddit, Kumar et al. (Kumar et al. 2018) studied controversial hyperlink cross-postings between subreddits to identify community conflict. Examples of the latter (sanction prediction) include banning prediction based on comments (Cheng, Danescu-Niculescu-Mizil, and Leskovec 2015) and using abusive content on one forum to predict abuse on others (Chandrasekharan et al. 2017b). The bulk of this research emphasized individual behavior rather than inter-community anti-social behavior (rare exceptions emphasized specific types of anti-social behavior). While we draw upon this literature to understand individual behaviors, we focus on a broad definition of higher-order inter-community conflicts. That is, our aim is to identify inter-community conflict (rather than individual-on-individual or individual-on-community conflicts) by developing behavioral mapping mechanisms in the context of the full network.

Trolling in Reddit

Individual trolling in Reddit is predominantly studied through content analysis (e.g., (Merritt 2012)). A key focus for Reddit research has been comparing the differences in trolling behavior between a smaller number of communities. For example, Metzger performed a contrastive study on intercultural variation of trolling in *ShitRedditSays* and *MensRights* (Metzger 2013). Most related to our work is the study by Kumar et al. (Kumar et al. 2018) which found that very few subreddits are responsible for the majority of conflicts. This has implications to our conflict graph constructions in that we anticipate finding key conflict 'nodes.'

More recently, there has been research on interventions (e.g., banning) to combat anti-social behavior. For example, Chandrasekharan et al. (Chandrasekharan et al. 2017a) studied the effect of banning two particular subreddits, *fatpeo-*

plehate and *CoonTown*, to combat hate-speech. However, this work does not elaborate on subreddit-to-subreddit relations before or after the ban. Subreddit relations are discussed from an ideological frame by identifying subreddits which discuss the same topic from different point of views (Datta, Phelan, and Adar 2017). However, this approach does not capture conflict explicitly.

Signed Social Networks

We analyze subreddit conflicts by creating a subreddit conflict graph, which can be viewed as a *signed graph* (where all the edges are marked negative). Use of signed graphs for trolling detection is uncommon but has been explored in past research. Kunegis et al. (Kunegis, Lommatzsch, and Bauckhage 2009) predicted trolls and negative links in Slashdot (a technological news website and forum where users are able to tag other users as 'friend' or 'foe'). Multiple studies (Wu, Aggarwal, and Sun 2016; Shahriari and Jalili 2014) proposed models to rank nodes in signed social networks. Signed networks incorporate both positive and negative edges. In our case, it is difficult to make claims about positive relations in the conflict graph. Because most individuals are norm-compliant, edges constructed between two *social* homes may be an artifact of authors being largely norm-compliant and simply reflect correlated interests. In contrast, an author that displays both norm-compliant and norm-violating behaviors provides a better indication of likely conflict.

Dataset

In our research, we focus on Reddit (www.reddit.com) both due to its similarity (in features) to many other discussion boards and its vast scale. Reddit is a social aggregator and discussion forum for millions of individuals who regularly post news, video, images or text and discuss them in different comment threads. Reddit divides itself into focused subreddits which discuss particular topics or perform specific kinds of social aggregation (e.g., image or video sharing). In each subreddit, user-submitted content is referred to as a Reddit *post*, and a post is discussed via different dedicated *comment threads*. Each post and comment can be rewarded or sanctioned by other Reddit users using upvotes and downvotes. Within each subreddit, comments can also be flagged or moderated by subreddit-specific moderators. Individuals who behave inappropriately can also be banned from specific subreddits.

For the analysis presented here, we used all publicly available Reddit comments from 2016. This was a subset of the multi-year Reddit data (posts, authors, comments, etc.) compiled by Baumgartner (Baumgartner 2017). We specifically mined commenting behavior (rather than posting) for building conflict graphs. Comments are much more prevalent than posts, and anti-social behavior in Reddit often involves inflammatory comments rather than posts. For each comment, we make use of the following metadata: author of the comment, which subreddit the comment was posted on and how many upvotes and downvotes the comment received. We found that there are 9,752,017 unique authors who commented at least once in Reddit in our sample. Though largely

a 'human population,' bots can also be programmed to generate comments. Of the 9.7M authors, 1,166,315 were 'highly active,' posting more than 100 comments throughout the year. On average, a Reddit user posted in 7.2 subreddits and commented 76.2 times in 2016. As may be expected, most Reddit users are 'pro-social.' In 2016, we find that 79.2% of authors (across all of Reddit) have at least 90% of their comments upvoted.

Identifying Inter-community Conflicts

To define the *conflict graph* between subreddits we first identify edges that capture community-on-community 'attacks.' We would like these edges to be directed (as not all conflict is reciprocated) and weighted (to indicate the strength of the conflict). Our goal is not to only identify 'passive' ideological opposition but also behaviors where one subreddit actively engages with the other.

This distinction is important as there are instances where two subreddits are discussing the same topic through different ideologies (as determined through text analysis), but have very low author overlap. For example, the *askscience* (discussion forum for science-related topics) and *theworldisflat* (forum for scientific evidence that the world is flat) could be considered to be ideologically opposed (Datta, Phelan, and Adar 2017). However, there are very few authors who post in both subreddits, meaning there is no engagement and no 'conflict' by our definition. These individuals do not agree but largely leave each other alone.

Instead, we focus on identifying individuals that post to multiple subreddits and behave *differently* depending on the subreddit. In our model, behaviors, such as commenting, can be norm-compliant or norm-violating. Norm-compliant are those behaviors that the community finds agreeable in that they are consistent both with the way behaviors (e.g., message posting) should be done and/or the content of the message itself. Norm-violating are those behaviors that are disagreeable in the same way (how they are posted or what is in them). Norm-violating behavior can include traditionally anti-social behaviors: flame wars, griefing, spamming, trolling, brigading, baiting, fisking, crapflooding, shitposting, and trash talking. This, again, is consistent with Brunton's spam definition (Brunton 2012). The appeal of this localized definition of spam is that each community can assert what they consider social or anti-social behavior (i.e., norm-compliant and norm-violating) and can make local decisions to reward or sanction such behaviors respectively.

Our inferential goal is to operationalize social and anti-social behavior by leveraging reward and sanction behaviors as indicators. For this purpose, we use up- and downvotes. Obviously, not all compliant behaviors are rewarded through up-votes, nor are all norm-violating sanctioned through down-votes (banning being a notable alternative). Other metrics for norm-violation may include identifying banned users or users whose comments are regularly removed by moderators. Unfortunately, such data is not readily available (removed comments and authors will be missing from the dataset). Posts can also be marked as 'controversial' to signal undesirable behavior, but these are not always anti-social per se. Additionally, both banning and con-

troversial post 'tagging' may not be reliably imposed. Up-voting and downvoting, however, are part Reddit's incentive structure and are uniformly applied and ubiquitous.

An individual who reliably produces enough *measurable norm-compliant behavior* (e.g., many upvoted messages) can be said to have a *social home* in that community. Likewise, an individual that produces a substantial amount of *measurable norm-violating behavior* (i.e., many downvotes) is said to have an *anti-social home* in that community. An individual can have multiple social and multiple anti-social homes. Because our goal is to find conflict *edges* we do not consider authors that are *only* social or *only* anti-social. Those who are globally norm-violating (e.g., spammers, malicious bots, etc.) and are negatively treated in all subreddits in which they post are removed from consideration. Figure 1 (left) illustrates this idea.

A second key aspect in building the conflict graph is in aggregation. One *particular* individual may have a social home and an anti-social home. However, from the single example, we can not infer that the other members of that person's social home would endorse the messages the person is posting to the other subreddit. Instead, we look for signals in the aggregate. If there are many individuals, who cross-post to two subreddits— where one subreddit is clearly a social home, and the other is clearly an anti-social home—we infer that a conflict exists. This conflict need not be reciprocated, but as we show below, it often is.

We can roughly quantify the anti-social behavior of a user within a subreddit if he/she has more downvoted comments compared to upvoted comments. Note that a single comment can have multiple upvotes and downvotes and Reddit automatically upvotes a user's own comment (all comments in Reddit start with one upvote). We consider the user's upvote as a 'baseline' (i.e. we consider user's own upvote as one upvote when counting all upvotes) as we assume the author views his or her own comment positively[2]. We say a comment is *downvoted* (in aggregate) if the total number of downvotes for the comments exceeds upvotes, and *upvoted* when upvotes exceed downvotes. Similarly, we say that a user has shown social behavior if they have more upvoted comments (rewarded, norm-compliant) compared to downvoted ones (sanctioned, norm-violating) within a subreddit.

To distinguish between an author's 'home' and simply a 'drive-by' comment, we enforce a threshold (we call this *significant presence*) of more than ten comments in the subreddits over the course of the year (2016). This threshold also ensures that we can observe enough up and down votes for any particular author. Additionally, new authors in a subreddit might break some unfamiliar rules, and receive downvotes initially. Our threshold gives sufficient data to determine if they 'learned.' We also enforce that the user has more than 100 total posts in 2016, which ensures that they have an overall significant presence on Reddit.

As authors were automatically assigned to *default subreddits* (*AskReddit*, *news*, *worldnews*, *pics*, *videos*), many Red-

dit authors began by posting in these groups[3]. Norms (and norm-compliance) in these subreddits may be significantly different from rest of the subreddits. Using our definition of social homes, a large number of users have at least some default subreddit as their social home or anti-social home just because they started by posting in these forums. For this reason, we exclude default subreddits from our analysis.

Controversial Authors

We denote an author with at least one social and one anti-social home as a *controversial author* (the purple figures in Figure 1). In 2016, 1,166,315 authors had more than total 100 comments over the year. After filtering for significant presence in subreddits, we found 23,409 controversial authors. This indicates that only about 2% of the more prolific Reddit users fall into this category. Among the controversial authors, 82% have only a *single* anti-social home. The vast majority (92.5%) of controversial authors have *more* social than anti-social homes. This indicates that these authors differ from the conventional idea of a "troll" who misbehaves in every forum they participate in. This also means that a typical controversial author focuses his/her 'misbehavior' on a small number (usually 1) subreddits. This result is consistent with Reddit users being loyal (in posting) to a small set of subreddits (Hamilton et al. 2017).

In aggregate, if there are many controversial authors that have a social home in subreddit A and anti-social home in subreddit B, we view this to be a directed conflict from A against B. We call this a *conflict edge*. The sum of all these edges, after some additional filtering, captures the *conflict graph* (Figure 1, right).

Our approach has the benefit that aggregation can eliminate various types of noise. While upvoting/downvoting is noisy at the level of any particular message, aggregation at the author level allows us to look for *consistent* behaviors (i.e., are messages from an author always rewarded in one place and sanctioned in another?). Noise at the level of a *particular* controversial author is similarly mitigated by aggregation (are there *multiple* individuals being rewarded in one place and sanctioned in the other?).

The Subreddit Conflict Graph
Constructing the Conflict Graph

To construct the conflict graph, we apply the following strategy. If k authors have a social-home in subreddit A, and an anti-social home in subreddit B, we can create a weighted directed *conflict edge* from A to B. If we create these edges for all subreddit pairs, we have a graph of antagonistic subreddit relations. Weights for these edges must be normalized as different subreddits have a different number of users. Thus, a raw author count (i.e., common authors with a social home in A and anti-social home in B) is misleading. Larger subreddits would dominate in weights as more authors often means more controversial authors. For convenience, we

---

[2]The algorithms described in this paper can be applied with or without a individuals' personal upvotes. We find the variation to be small in this instance.

[3]Though this does not impact our analysis (for 2016 data), we note that default subscription was replaced in 2017 with a dynamic popular subreddit homepage.
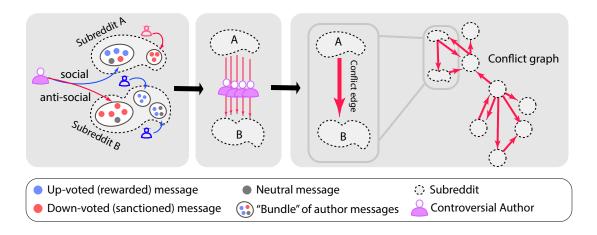
Figure 1: Methodology for identifying conflicts and creating the conflict graph.

refer to the 'source' of the edge as the *instigating* subreddit and the 'destination' as the *targeted* subreddit.

We normalize the raw controversial author counts by the number of common authors in both subreddits. Furthermore, for each subreddit pair, we require that there are at least five controversial authors between them to ensure that we are not misidentifying a conflict due to very few controversial authors (i.e. if there is k_1 authors with social home in subreddit A and anti-social home in subreddit B, and k_2 authors with social home in subreddit B and anti-social home in subreddit A, $k_1 + k_2$ must be at least five). We emphasize that the weight, direction, or even existence, of an edge from subreddit A to B, is different from an edge from B to A.

Eliminating Edges Present due to Chance

While defining conflict between a pair of subreddits, we need to make sure that users are not perceived negatively in the attacked subreddit by chance. For two subreddits A and B with n_{common} common users and n_{actual} users perceived positively in subreddit A but negatively in subreddit B (we only consider users who posted more than 10 times in both subreddits), we calculate the number of users who can be perceived negatively in subreddit B by chance. First, we define an empirical multinomial distribution of comment types for subreddit B, i.e., we calculate the probabilities of a random comment in subreddit B being positive (upvoted), negative (downvoted) or neutral. To create this multinomial distribution, we only use comments from users who posted more than ten times in subreddit B as these are the users we consider when declaring controversial authors. For a common user i, if i posted n_i times in subreddit B, we sample n_i comments from the probability distribution and calculate if he/she is perceived negatively in the sample. We sample all common users and count the total number of users perceived negatively in subreddit B. We repeat this experiment 30 times to create a sampling distribution of the expected number of negatively perceived users and calculate the z-score of n_{actual} using this sampling distribution. We only retain conflicts from A to B, where this z-score is greater

than 3, i.e., the number of users perceived negatively in the attacked subreddit is significantly higher than the number expected from random chance. The final set of subreddits (nodes) and associated edges are the *conflict graph*.

Conflict Graph Properties

The final subreddit conflict graph for 2016 consists of 746 nodes and 11,768 edges. This is a small fraction of active subreddits in 2016 (around 76,000) which is, in part, due to the low amount of 'multi-community posting' on Reddit (Hamilton et al. 2017) (i.e., very few authors regularly post to more than one 'home' community). As we require multi-community posts to create an edge, many subreddits are 'free floating' and are removed from consideration. Of the 746, nine were banned sometime between the end of 2016 and April of 2018: *PublicHealthWatch* (a subreddit dedicated to documenting the 'health hazards' of, among others, LGBTQ groups), *altright*, *Incels* (involuntary celibate), *WhiteRights*, *european*, *uncensorednews*, *european-nationalism*, *DarkNetMarkets* and *SanctionedSuicide*. An additional six became 'private' (requiring moderator approval to join and post), which include known controversial subreddits (*Mr_Trump* and *ForeverUnwanted*).

The conflict graph consists of five components, with the giant component containing 734 nodes. The next largest component consists of only six nodes representing different sports streaming subreddits (*nflstreams*, *nbastreams*, *soccer-streams* etc.). Through manual coding of subreddits we identify the following high-level categories: political subreddits (e.g. *politics*, *The_Donald*, *svenskpolitik*) discussion subreddits, video game subreddits (e.g. *Overwatch*, *pokemongo*), sports fan clubs, location-focused subreddits (e.g. *canada*, *Seattle*, *Michigan*, *Atlanta*), subreddits for marginalized groups (e.g. *atheism*, *DebateReligion*, *TrollXChromosomes*, *lgbt*, *BlackPeopleTwitter*), *porn subreddits (these are image sharing subreddits with their name ending with porn, they are not pornographic in nature – e.g. *MapPorn*, *HistoryPorn*) and NSFW subreddits (e.g. *nsfw*, *NSFW_GIF*). Because of our use of 2016 data and the associated (and contentious)

election, political subreddits are heavily represented in the conflict graph. Figure 2 shows the ego network for the subreddit *Liberal* in the conflict graph.
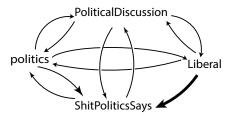

Figure 2: Ego network for the subreddit *Liberal*. Thicker edges denote higher conflict intensity.

Edge weights in the conflict graph are often low. On average, only 3.57% (median is 1.70%) of authors in the 'conflict source' subreddit (i.e., their social home) post to the target subreddit (i.e., their anti-social home). There are, however, edges with extremely high weights. The highest edge weight in our data is 85.71% from *The_Donald* to *PanicHistory*. However, in this case, this is due to the disproportionate difference in size of the two (they share only seven common authors). Thus, a high conflict intensity does not necessarily mean that a large fraction of originating subreddit users are antagonistic to the target subreddit. Nonetheless, it does point to the fact that larger subreddits with many controversial authors can overwhelm a smaller subreddit. The high edge-weight here indicates the degree to which this happens. Using the subreddit conflict graph, we can isolate the main source and targets of conflicts and understand where conflicts are one-sided or mutual.

Are conflicts reciprocal? We find that if a conflict edge exists between subreddit A and B, in 77.2% cases the inverse edge will *exist*. Calculating the Spearman correlation between conflict intensities of pairs of reciprocated edges, $\rho(5126) = -0.111, p < 0.0001$, we observe a weak (but significant) negative relationship. Figure 3 depicts the outgoing conflict (source) intensity versus incoming conflict (the conflict target) intensity. This indicates that a targeted subreddit will likely reciprocate, but the intensity is usually not proportional.

Which subreddits are most targeted in 2016? The indegree of a subreddit roughly indicates the number of other subreddits targeting it. The weighted sum of these edges (weighted indegree) corresponds to the intensity. The top 10 most targeted subreddits by indegree are *politics*, *SubredditDrama*, *AdviceAnimals*, *EnoughTrumpSpam*, *atheism*, *SandersForPresident*, *The_Donald*, *PoliticalDiscussion*, *technology* and *KotakuInAction* respectively. However, when we order subreddits by *total incoming conflict intensity* (see Table 1) the list is somewhat different. In both lists, we observe that the most targeted subreddits are social and political discussion forums as well as forums that discuss Reddit itself. The heavy presence of political forums can be attributed to the 2016 US presidential election. We try to deduce if, in general, the most targeted subreddits
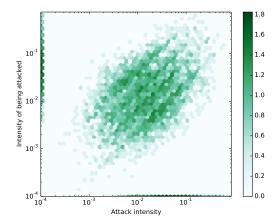

Figure 3: Conflict intensity vs intensity of reciprocation in the subreddit conflict graph (log-scale). Un-reciprocated edges appear at the bottom and left edge.

by degree are also the most targeted subreddits by average incoming intensity (total intensity/number of sources) and vice versa. When contrasting indegree to average intensity for subreddits that are targeted by at least one subreddit (we have 673 such subreddits), we find a weak positive correlation with Spearman $\rho(673) = 0.242, p < 0.0001$. A subreddit targeted by many subreddits is not necessarily targeted with high intensity. Conversely, subreddits targeted by only a few others can nonetheless be targeted with high intensity.

Subreddit	Indegree	Weighted indegree
SubredditDrama	272	19.51
EnoughTrumpSpam	217	13.25
BestOfOutrageCulture	46	10.59
ShitPoliticsSays	48	10.29
Enough_Sanders_Spam	48	9.80
sweden	81	9.62
KotakuInAction	168	9.19
ShitAmericansSay	94	8.83
PoliticalDiscussion	185	8.08
vegan	71	7.26

Table 1: Top-10 targeted subreddits ranked by total incoming intensity.

Which are the most conflict 'instigating' subreddits in 2016? By using the conflict graphs outdegree (weighted or not) we can similarly find the largest conflict sources. The top-10 subreddits ranked by outdegree are *politics*, *AdviceAnimals*, *The_Donald*, *SandersForPresident*, *WTF*, *technology*, *atheism*, *SubredditDrama*, *EnoughTrumpSpam* and *PoliticalDiscussion*.

When ordered by total intensity, the top-10 list changes to include more news, politics, and controversy focused subreddits (Table 2). This list also includes a now banned subreddit (*uncensorednews*). However, we observe that most of these subreddits have low average conflict intensity (i.e., intensity per edge is low). If we order by average intensity

(Table 3), we find that subreddits targeting very few others (usually 1 or 2 subreddits) show up at top spots. However, we find that the subreddits at the first, third and ninth position of this list (*europeannationalism*, a Nazi subreddit, *PublicHealthWatch*, an anti-LGBT subreddit and *WhiteRights*) are banned by Reddit. A controversial now private subreddit (*ForeverUnwanted*) also appears in this list. This may have implications for identifying problematic subreddits.

As before, we can check if the subreddits most often at the source of a conflict (by outdegree) are also the most instigating (by average conflict intensity). Using 719 'source' subreddits in our conflict graph, we find a weak positive correlation between the number of targeted subreddits and the average outgoing conflict intensity (Spearman $\rho(719) = 0.189, p < 0.0001$), in line with our previous discussion.

Subreddit	Outdegree	Weighted outdegree
The_Donald	260	17.75
politics	542	10.15
conspiracy	141	7.39
KotakuInAction	152	7.35
uncensorednews	113	7.14
AdviceAnimals	268	6.76
SandersForPresident	211	6.62
CringeAnarchy	148	6.12
ImGoingToHellForThis	114	5.99
Libertarian	92	5.86

Table 2: Top 10 subreddits (conflict source) ranked by total conflict intensity.

Subreddit	Outdegree	Average outdegree
europeannationalism	1	0.75
OffensiveSpeech	1	0.62
PublicHealthWatch	1	0.62
askMRP	1	0.57
ForeverUnwanted	2	0.50
theworldisflat	1	0.43
FULLCOMMUNISM	2	0.42
marriedredpill	1	0.38
WhiteRights	2	0.34
SargonofAkkad	2	0.33

Table 3: Top 10 subreddits (conflict source) ranked by average conflict intensity.

Do larger subreddits get involved in more conflicts due to their size? We find that larger subreddits are more likely to get involved in both incoming and outgoing conflicts. Using number of unique authors who posted more than 10 times in 2016 in the subreddit as a measure of subreddit size, we find moderate positive correlation between both size and number of incoming conflicts (Spearman $\rho(673) = 0.403, p < 0.0001$), and size and outgoing conflicts (Spearman $\rho(719) = 0.457, p < 0.0001$). However, taking conflict intensities into account, we find size and average incoming conflict intensity is moderately negatively correlated (Spearman $\rho(673) = -0.594, p < 0.0001$). Similarly, size and average outgoing conflict intensity is also weakly negatively correlated (Spearman $\rho(719) = -0.222, p < 0.0001$). This tells us that subreddits with larger size are more likely to be involved in conflicts just because there are more authors commenting in them, but average conflict intensity is not indicative of subreddit size.

Node properties Edge weights alone do not tell us if controversial authors are particularly prevalent in a specific subreddit. Rather, it only indicates the fraction of common users who are sanctioned (norm-violating) in the target subreddits. However, these common users might represent only a small fraction of users of a subreddit. This is especially possible for the larger subreddits. To determine which subreddits are the social home for *many* controversial authors, we use three additional metrics: *con_author_percent* is the percentage of controversial authors who make their social home in a subreddit relative to the number of authors who posted in that subreddit (more than 10 times in the year); *avg_subs_targeted* and *median_subs_targeted* are the average and median of number of subreddits that these controversial authors 'target.' These numbers can tell us (a) what fraction of a subreddit are engaged in conflict, and (b) are they engaging in broad (across many subreddits) or focused conflicts. We limit our study to subreddits with at least 20 controversial authors who have a social home on that subreddit (overall, we find 698 subreddits meet this criterion). Removing smaller subreddits minimally affects the top-10 subreddits (see Table 4) by con_author_percent (only *theworldisflat*, with 11 controversial authors, is removed from the list). We refrain from listing one pornographic subreddit in the table at rank 8.

Subreddit	Con_author_%	Average	Median
PublicHealthWatch	35.25	2.44	2.0
OffensiveSpeech	34.78	2.86	2.0
WhiteRights	32.60	3.19	2.0
ThanksObama	32.59	2.34	1.5
europeannationalism	32.43	2.77	2.0
subredditcancer	27.51	2.33	2.0
subredditoftheday	24.90	1.94	1.0
POLITIC	23.94	1.91	1.0
undelete	23.04	2.13	1.0
SRSsucks	22.18	2.07	1.0

Table 4: Top 10 subreddits with highest percentage of positively perceived controversial authors (with at least 20). The average and median columns correspond to *avg_subs_targeted* and *median_subs_targeted* respectively.

Most subreddits in top 10 list are either political forums or somewhat controversial in nature. To lend further credence to this measure, *PublicHealthWatch* (an anti-LGBT subreddit, rank 1), *WhiteRights* (rank 3) and *europeannationalism* (a Nazi subreddit, rank 5) score highly with our metric and were recently banned by Reddit. It is also important to note that most controversial authors have only one anti-social home. Thus in almost all cases, median_subs_targeted is 1. The only exceptions are the first six subreddits in Table 4, *The_Farage* (median is 2) and *sjwhate* (median is 2). Note that all banned subreddits shown in this table have a median

of 2. The median con_author_percent for all 698 subreddits is 4.09%, and the lowest is 0.36%. It is worth noting that, the three banned subreddits in this list targeted only one or two subreddit each but with very high conflict intensity (e.g., *europeannationalism* attacked *AgainstHateSubreddits* with conflict intensity of 0.75). All three subreddits also show up in the list of most conflict-source subreddits by average conflict intensity. This shows that a subreddit does not have to target multiple other subreddits to be problematic.

Subreddit	Con_author_% rank (value)	Average conflict intensity rank (value)
PublicHealthWatch	1 (35.25)	2 (0.62)
europeannationalism	5 (32.43)	1 (0.75)
WhiteRights	3 (32.60)	9 (0.34)
altright	23 (18.18)	19 (0.24)
european	31 (17.41)	24 (0.20)
Incels	183 (7.87)	57 (0.11)
uncensorednews	13 (19.94)	123 (0.06)
DarkNetMarkets	635 (1.59)	494 (0.02)
SanctionedSuicide	475 (2.94)	199 (0.04)

Table 5: Banned subreddits and their ranks and values by average conflict intensity and con_author_percent.

Compared to the top-10 subreddits by con_author_percent, large political subreddits in the most instigating subreddit list (conflict source) had a lower percentage of controversial authors who engage in conflict with other subreddits (e.g., *The_Donald* (8.09%), *SandersForPresident* (6.75%), *politics* (5.71%)). However, in many cases, these values are higher than the median.

Banned subreddits Three out of nine banned subreddits in the conflict graph rank within the top 10 when ranked by con_author_percent and average conflict intensity. Table 5 shows rank (and value) by con_author_percent and average conflict intensity for all nine banned subreddits (lower ranks means higher con_author_percent and higher average intensity respectively). We observed that moderately low ranks in both measures for three other banned subreddits. High con_author_percent means that a large fraction of the corresponding subreddit is participating in norm-violating behavior and high average intensity means that a large fraction of common authors between the source and target subreddits are norm-violating. Low ranks by both these measures should indicate that the corresponding subreddit is misbehaving as a community. This is supported by the fact that six out of nine banned subreddits and two controversial subreddits (both set to private by moderators of the respective subreddits) display this behavior. We emphasize again that subreddits can be banned due to their content and not due to the conflict they caused. Such subreddits will not rank low in these two measures.

Co-Conflict Communities

Creating the Co-conflict Graph

Although most controversial authors have only one anti-social home, there are nonetheless patterns of conflict di-

rected from one subreddit against multiple others. Subreddits targeted by same set of authors gives us further insight about these authors and the subreddits they call home. Using all subreddits from the conflict graph, we can create graphs that map the subreddits that are co-targeted. In the co-conflict graph, nodes are still subreddits. Edges are determined by generating a weighted edge between two subreddits A and B if there are at least 2 common negatively perceived controversial authors between subreddits A and B, so that we do not misidentify an edge due to one single author. The weight of the edge is determined by the Jaccard coefficient, which denotes how many of such authors A and B have in common compared to distinct negatively perceived controversial authors in both subreddits. If X and Y denotes the set of such authors in subreddit A and B respectively, the Jaccard coefficient between X and Y is defined as: $Jaccard(X, Y) = \frac{X \cap Y}{X \cup Y}$.

Co-conflict Graph Properties

As the majority of controversial authors misbehave in only one subreddit, the co-conflict graph has many disconnected components. We only focus on the largest connected component (i.e. the giant component) which consists of 237 nodes and 780 edges. Unlike the subreddit conflict graph, the co-conflict graph is undirected. Furthermore, edge semantics are different as edges denotes the similarity between two subreddits. Common network analysis algorithms can be applied to this graph more intuitively. Use of community detection, for example, can help us determine which groups of subreddits (rather than pairs) are 'co-targeted.' There are multiple algorithms for community detection in undirected networks (Fortunato 2010) (e.g., FastGreedy, Infomap, Louvain, etc.). The algorithms have different trade-offs (Aldecoa and Marín 2013; Lancichinetti and Fortunato 2009; Prat-Pérez, Dominguez-Sal, and Larriba-Pey 2014), though generally both Louvain and Infomap are shown to perform well. The Louvain or multilevel algorithm (Blondel et al. 2008; Meo et al. 2011) is based on modularity maximization, where modularity is a measure of cohesiveness of a network. An attractive property of Louvain is that it follows a hierarchical approach by first finding small, cohesive communities and then iteratively collapsing them in a hierarchical fashion. This approach on the co-attacked graph produced reasonably sized communities and the results of the community detection algorithm were very stable (i.e., do not change much on different runs). Note that, we use the weighted Louvain algorithm for this purpose.

Community Detection Results

We evaluate the communities using μ-score and clustering coefficient (CC). μ-score is defined as fraction of edges from within the community to outside the community compared to all edges originating from the community. The clustering coefficient of a node is the fraction of connected neighbor pairs compared to all neighbor pairs. For a community, the CC is the average of clustering coefficients of all nodes in the community. In general, low μ-score and high CC denotes a 'good' community. Using the weighted multilevel algorithm on the co-attacked graph we find 15 distinct communities.

Table 6 shows exemplar subreddits per community, size of the community, μ-score and clustering coefficient for subreddits with at least 10 nodes in them.

Figure 4 shows the co-conflict graph and its communities. In general, most communities show low μ-score and low clustering coefficient due to presence of star-like structures (i.e. a large number of nodes are connected to one single node). For example, *politics* is connected to 103 other subreddits. Smaller subreddit communities are topically more cohesive compared to larger communities. For example, community 6 (video game subreddits) and 7 (gun-related subreddits) in table 6 are both topically very cohesive.
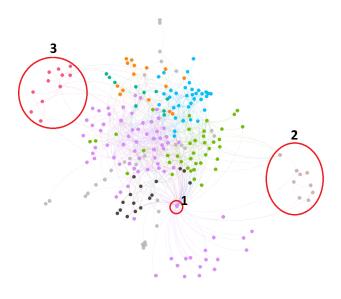

Figure 4: The Co-conflict Graph (each node is a subreddit). Different communities are shown in different colors. Circle 1 includes *politics*, a subreddit demonstrating a star pattern. Circles 2 and 3 are communities of gun-related and video game subreddits respectively.

It is worth re-emphasizing that the co-conflict graph does not necessarily mean that a pair of subreddits in the same community are 'friendly' and do not have a conflict with each other. For example, *Christianity* and *atheism* belong to same community and there exists a conflict edge from *Christianity* to *atheism*. Similarly, *SandersForPresident* and *Enough_Sanders_Spam* are in the same community and are very much "at war". This is mostly due to presence of aforementioned star-like structures. For example, *Republican* and *democrats* both are only connected to *politics* and thus belong in the community containing *politics*. This does not mean that *Republican* and *democrats* have a common group of people perceived negatively.

Conflict Dynamics

One interesting question for our conflict graphs is how they change over time? It is possible that controversial authors maintain the same social and anti-social homes over time. Conversely, a subreddit with controversial authors may 'shift' its negative behaviors to different subreddits over time. To better understand these dynamics, we study this in both an aggregate manner (i.e. does the most targeted and most instigating subreddits vary each month or do they remain mostly static?), and from the perspective of a few individual subreddits (how does rank of a particular subreddit among most targeted and most instigating subreddits vary over time?). To do so, we created conflict graphs for each month in 2016. These monthly graphs use the same set of subreddits and the same set of controversial authors used in constructing the yearly conflict graph.

We focus this preliminary analysis on subreddits that targeted five or more other subreddits over the year and model how their 'conflict focus' varies. That is, do they specifically focus on a single subreddit over all months, or does their most targeted subreddit in a specific month vary from month to month? To determine this, we count the number of times the most targeted subreddit for each conflict source subreddits change from one month to the next. We call this the *change_count* for the attacking subreddit. By definition, change_count can vary from 0 (most targeted subreddit did not change in all 12 months) to 11 (most targeted subreddit changed every month). If a subreddit did not target any other in a particular month, but targeted some subreddit in the next (or vice versa), we count that as a change. Figure 5 shows the distribution of change_count for source subreddits.
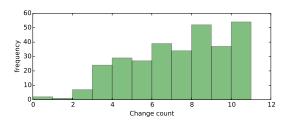

Figure 5: Change count for source subreddits who targeted at least 5 subreddits

On average, we find that change_count is 6.91 (median of 7), indicating that most subreddits shifts their primary focus over time. We find only 2 subreddits did not change their target at all in 12 months. One example of this is *CCW* (concealed carry weapons subreddit) targeting *GunsAreCool* (a subreddit advocating for gun control in USA).

Because of the 2016 US election, the monthly 'most targeted' and 'most instigating' subreddits are still predominantly political. However, some subreddits only appear in the beginning of the year (e.g. *SandersForPresident* is in the list of top 3 most instigating subreddits for the first four months, *The_Donald* is the top 3 most targeted subreddit list for the first 3 months) or end of the year (e.g. *EnoughTrumpSpam* is in the list of top 10 most targeted subreddits for the last 7 months and during that time, it is the most targeted subreddit). On the other hand, some subreddits show remarkable consistency—*The_Donald* is always the most instigating subreddit (for all 12 months) and *politics*, *SubredditDrama* are always in the top-10 most targeted.

No	Example subreddits	Size	μ-score	cc	description
1	*politics*, *PoliticalDiscussion*, *hillaryclinton*, *SandersForPresident*, *EnoughTrumpSpam*, *Enough_Sanders_Spam*, *AskTrumpSupporters*, *SubredditDrama*	74	0.33	0.41	mostly politics and political discussion subreddits
2	*KotakuInAction*, *conspiracy*, *undelete*, *MensRights*, , *PublicFreakout*, *WikiLeaks*, *worldpolitics*, *Political_Revolution*, *europe*, *The_Donald*	39	0.20	0.29	Political subreddits, controversial subreddits
3	*nsfw*, *NSFW_GIF*, *woahdude*, *cringepics*, *trashy*, *WatchItForThePlot*	34	0.23	0.38	mostly NSFW subreddits, subreddits making fun of others
4	*nba*, *nfl*, *baseball*, *Patriots*, *canada*, *toronto*, *ontario*	16	0.17	0.15	sports subreddits, Canada related subreddits
5	*TopMindsofReddit*, *AgainstHateSubreddits*, *SRSsucks*, *worstof* ,*ShitAmericansSay*, *TrollXChromosomes*	15	0.19	0.24	Subreddits focusing on other subreddits
6	*Overwatch*, *DotA2*, *GlobalOffensive*, *NoMansSkyTheGame*, *leagueoflegends*	11	0.06	0.00	video game related subreddits
7	*guns*, *progun*, *Firearms*, *gunpolitics*, *shitguncontrollerssay*	10	0.06	0.23	gun-related subreddits
8	*relationships*, *OkCupid*, *AskMen*, *AskWomen*, *niceguys*, *instant_regret*, *sadcringe*, *TheBluePill*	10	0.39	0.51	relationship subreddits, satirical subreddits

Table 6: Communities in co-conflict network with at least 10 nodes. For each community, exemplar subreddits, size of the community, μ-score and clustering coefficient(cc) is shown

Figure 6 illustrate the rank of four political subreddits related to the US presidential election (*The_Donald*, *EnoughTrumpSpam*, *SandersForPresident* and *Enough_Sanders_Spam*). The figures capture the rank of these in the most targeted (upper plot, largest weighted indegree in the conflict graph) and most instigating lists (lower plot, largest weighted outdegree) respectively. These demonstrate both the pattern of stable conflict as well as varying ones.
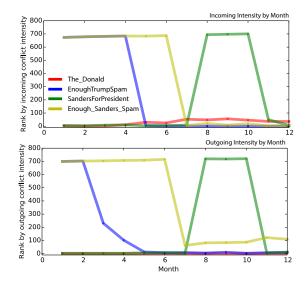

Figure 6: Rank by intensity of being targeted (top) and conflict intensity (bottom) for four political subreddits in 2016.

Perhaps the most important observation from these plots is how similar they are. *The_Donald* is always the most instigating subreddit and it is consistently targeted back. *EnoughTrumpSpam* gained popularity during March 2016 and gradually became more instigating in the next two months. For the last seven months of 2016, *EnoughTrumpSpam* is the most targeted subreddit. *SandersForPresident* is near the top in both most targeted and most instigating list until the end of July 2016 and from November 2016. However, in three months between July and November, this subreddit did not have any antagonistic relation with any other subreddit as it was shutdown after US political conventions in July and subsequently brought back after in November. *Enough_Sanders_Spam* was formed in July 2016 and instantly became highly targeted due to its content. This shows that, a subreddit instigating/targeted can be highly dependant on external events.

Discussion

In this paper, we demonstrate a quantitative method for identifying community-to-community conflicts by aggregating users who behave differently depending on the community they interact with. We define social and anti-social homes of a user based on a local perception of norm-compliance and norm-violation (which we measure by reward and sanction through voting). This method allows us to find conflict in any social network with 'noisy' community structure. Though we focus on Reddit in a specific year (2016), we believe the work is more broadly usable both across time and other social media sites.

Before discussing in which situations our approach may or may not be usable, we briefly summarize our key findings. We find that community-to-community conflicts are usually reciprocal but mutual conflict intensities usually do not match up. We identify which subreddits generated most conflict and which subreddits were most targeted. By analyzing subreddits banned by Reddit in relation to our measures (e.g., average conflict intensity, a high percentage of positively perceived controversial authors, etc.) we illustrate how our technique may be useful for identifying problematic subreddits. Co-conflict subreddit communities show that subreddit conflicts are not random in nature, as we observe topically similar subreddits usually belong to the same co-conflict community. We perform a preliminary analysis of temporal patterns in subreddit conflicts and find that the conflict focus usually shifts over time.

Downvotes for determining community conflicts

A downvoted comment in a particular subreddit may be a reaction to a number of factors ranging from innocuous norm-violation, being off-topic, presenting a non-conforming viewpoint, low-effort posts (e.g. memes), reposts, and truly malicious behavior. Furthermore, a user in a subreddit might have downvotes (and a few downvoted comments) due to being new in the subreddit (i.e., not knowing all the rules) or brigading, where users from antagonistic subreddits downvote random or targeted comments as a 'downvote brigade'. Due to these factors, we rely on aggregate signals in our analysis instead of individual comments.

A downvote does not provide a *global* quality assessment of a comment. Rather, a downvoted comment within a subreddit signifies that this particular subreddit perceives the comment as low quality. This is a localized definition of quality defined by the subreddit and it is consistent with Brunton's model of spam (Brunton 2012). Globally, these comments might not be seen as norm-violating or low-quality. We acknowledge the fact that users may receive negative feedback not for their own antisocial behavior, but for the antagonistic stance of the receiving community. We do not assume that, for a conflict edge, the instigating community is a 'community of aggressors'. In fact, depending on the viewpoint, it might be viewed as a 'refuge for social outcasts.' New users in a subreddit are more susceptible to innocuous norm-violation due to them not knowing all rules of a new subreddit, but with time they tend to learn. To eliminate these users from the list of controversial users, we enforce a minimum threshold of comments in a subreddit. Excluding these users, we use downvote within a subreddit to determine subreddit conflicts and not as an indicator of the global quality of the comment.

Limitations

Using controversial authors to find subreddit conflicts has some limitations. First, this method does not take into account comments deleted by users or moderators (this data is not available for collection). Some subreddits are especially aggressive in deleting downvoted or moderated comments. In some cases, misbehaving authors in a subreddit are banned from further posts. As with comments, we do not have records of this type of moderation. When a subreddit aggressively bans many people, it can change the conflict graph from a static and dynamic perspective.

A clear example of this is *The_Donald*, which banned thousands of individuals over its lifetime (these banned individuals formed a subreddit *BannedFromThe_Donald*, with a subscriber count of 2,209 in November of 2016 and over 27,000 in July of 2018). These individuals do not show up as controversial authors as their comments are gone. We also do not account for sockpuppetry, i.e., having multiple accounts, one for normal posting behavior on Reddit and others for misbehaving. Presence of many users with sockpuppets can skew the estimation of controversial authors in different subreddits.

One final limitation of our model is that *correlated* multi-community posting may appear as a conflict edge. For example, members of community A (a subreddit for a specific computer game) are found to conflict with community B (a feminist subreddit). However, it may not be appropriate to say that A conflicts with B. The topics of the two communities are completely orthogonal. In this situation, it might be due to the presence of a third subreddit, C (e.g., an anti-feminist subreddit) that conflicts with B. It simply happens that many members of A (the game) also have a social-home on C (the anti-feminist subreddit). It would thus be more accurate to say that C conflicts with B. One approach for handling this is to ensure that there is some topical correspondence between the communities we are considering (based on text). This eliminates the $A - B$ edge but retains $C - B$. It is nonetheless possible that we may want to know that the $A - B$ link exists. Moderators of subreddit A might want to be made aware of this correlation and take action.

Implications

Although we perform our analysis on Reddit, our method is equally applicable in other social media with inherent or inferred community structure with associated community feedback. For example, we can perform a similar analysis on Facebook pages and groups, online news communities and Twitter hashtag communities (people who tweeted a particular hashtag are part of that hashtag community). We quantify user behavior based on upvotes and downvotes in a particular community, and this data is more easily available for many social media websites compared to a list of banned or otherwise sanctioned users from a particular community. Our approach is highly adaptable and can incorporate new information (e.g., lists of banned and sockpuppet accounts). The analysis is also fully automated and highly parallelizable making it viable for analysis of large datasets.

In addition to providing insight into communities, we also believe that our work can be used for moderation purposes. We observe that several banned subreddits rank very high on particular metrics for measuring conflict. We can calculate these measures for monthly (or otherwise temporal) subreddit conflict graphs and see how different subreddits rank in these measures over time. This can be used to monitor problematic subreddit behavior as a whole or create an early-warning system based on machine learning by treating currently banned subreddits as positive examples of communal misbehavior and use the metrics above as features.

Conclusion

In this work, we study community-on-community conflict. We describe a mechanism for determining the social and anti-social homes for authors based on commenting behavior. From these, we construct 'conflict edges' to map the conflicts on Reddit. Using our approach, we allow for a contextual definition of anti-social behavior based on local subreddit behavior. This provides a different perspective than studying global-norm violating behaviors.

We find that most conflicts (77.2%) are reciprocated, but the intensities from both sides do not necessarily match. Larger subreddits are more likely to be involved in more subreddit conflicts due to their large user-base, but most of these conflicts are minor, and this does not imply large-scale com-

munal misbehavior. On the other hand, we find that high average conflict intensity and a large fraction of subreddit users perceived negatively in other subreddits may have implications for communal misbehavior. Finally, we explore temporal patterns in conflicts and find that subreddits that target multiple others, will shift their main conflict focus over time. We believe this analysis can be applied to other social media sites which display community structure, create early warning systems for norm-violating communities and help encourage discussion about community-wide misbehavior in social media.

References

Adler, B. T.; De Alfaro, L.; Mola-Velasco, S. M.; Rosso, P.; and West, A. G. 2011. Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. In *Proc. of CICLing – Volume II*, 277–288.

Aldecoa, R., and Marín, I. 2013. Exploring the limits of community detection strategies in complex networks. *CoRR* abs/1306.4149.

Baumgartner, J. 2017. Reddit data. files.pushshift.io/reddit/.

Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; and Lefebvre, E. 2008. Fast unfolding of communities in large networks. *J. of Statistical Mechanics: Theory and Experiment* 2008(10):P10008.

Brunton, F. 2012. Constitutive interference: Spam and online communities. *Representations* 117(1):30–58.

Chandrasekharan, E.; Pavalanathan, U.; Srinivasan, A.; Glynn, A.; Eisenstein, J.; and Gilbert, E. 2017a. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proc. ACM Hum.-Comput. Interact.* 1(CSCW):31:1–31:22.

Chandrasekharan, E.; Samory, M.; Srinivasan, A.; and Gilbert, E. 2017b. The bag of communities: Identifying abusive behavior online with preexisting internet data. In *Proc. CHI'17*, 3175–3187.

Cheng, J.; Bernstein, M. S.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. *CoRR* abs/1702.01119.

Cheng, J.; Danescu-Niculescu-Mizil, C.; and Leskovec, J. 2015. Antisocial behavior in online discussion communities. In *Proc. ICWSM'15*.

Chesney, T.; Coyne, I.; Logan, B.; and Madden, N. 2009. Griefing in virtual worlds: causes, casualties and coping strategies. *Information Systems J.* 19(6):525–548.

Datta, S.; Phelan, C.; and Adar, E. 2017. Identifying misaligned inter-group links and communities. *Proc. ACM Hum.-Comput. Interact.* 1(CSCW):37:1–37:23.

Donath, J. S. 1999. Identity and deception in the virtual community. *Communities in cyberspace* 1996:29–59.

Fortunato, S. 2010. Community detection in graphs. *Physics Reports* 486(3):75 – 174.

Galán-García, P.; de la Puerta, J. G.; Laorden, C.; Santos, I.; and Bringas, P. G. 2013. Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying. In *SOCO-CISIS-ICEUTE*.

Hamilton, W. L.; Zhang, J.; Danescu-Niculescu-Mizil, C.; Jurafsky, D.; and Leskovec, J. 2017. Loyalty in Online Communities. *ArXiv e-prints*.

Hardaker, C. 2010. Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. *J. of Politeness Research* 6(2):215–242.

Kumar, S.; Cheng, J.; Leskovec, J.; and Subrahmanian, V. 2017. An army of me: Sockpuppets in online discussion communities. In *Proc. WWW'17*, 857–866.

Kumar, S.; Hamilton, W. L.; Leskovec, J.; and Jurafsky, D. 2018. Community interaction and conflict on the web. In *Proc. WWW'18*.

Kumar, S.; Spezzano, F.; and Subrahmanian, V. S. 2015. VEWS: A wikipedia vandal early warning system. *CoRR* abs/1507.01272.

Kunegis, J.; Lommatzsch, A.; and Bauckhage, C. 2009. The slashdot zoo: Mining a social network with negative edges. In *Proc. WWW'09*, 741–750.

Lancichinetti, A., and Fortunato, S. 2009. Community detection algorithms: A comparative analysis. *Phys. Rev. E* 80:056117.

Meo, P. D.; Ferrara, E.; Fiumara, G.; and Provetti, A. 2011. Generalized Louvain method for community detection in large networks. In *ISDA*, 88–93. IEEE.

Merritt, E. 2012. An analysis of the discourse of internet trolling: A case study of reddit. com.

Metzger, F. 2013. Impoliteness in CMC: Differences in trolling of men's rights activists vs. trolling of feminists.

Prat-Pérez, A.; Dominguez-Sal, D.; and Larriba-Pey, J.-L. 2014. High quality, scalable and parallel community detection for large real graphs. In *WWW'14*, 225–236.

Shachaf, P., and Hara, N. 2010. Beyond vandalism: Wikipedia trolls. *J. of Information Science* 36(3):357–370.

Shahriari, M., and Jalili, M. 2014. Ranking nodes in signed social networks. *Social Network Analysis and Mining* 4(1):172.

Suler, J. 2004. The online disinhibition effect. *Cyberpsychology & behavior* 7:321–6.

Wu, Z.; Aggarwal, C. C.; and Sun, J. 2016. The troll-trust model for ranking in signed networks. In *Proc. WSDM'16*, 447–456.