Community Interaction and Conflict on the Web

Srijan Kumar @srijankr
William Hamilton @williamleif
Jure Leskovec @jure
Dan Jurafsky @jurafsky
Why study inter-community interactions?

• Users form communities
• Communities interact with one another
• Little is known about how community interaction occurs
• So, we study inter-community interactions between 20,000+ communities on Reddit
Conflict across communities

“Come look at all the brainwashed idiots in Documentaries....”

- Can disrupt communities
- Can decrease long-term engagement

Members go and post negative/hateful comments

Conspiracy

Documentaries

Understanding how communities fight and how to prevent conflicts is important to foster a healthy online environment.
Reddit Dataset

We use public Reddit data for this study
• 40 months (2014—2017)
• 1.8+ billion comments
• 100+ million users
• 20,000+ communities

But…
There are no labels of community interactions and conflicts. How to define these?
"Come look at all the brainwashed idiots in Documentaries...."

Source community links to a post in target community

Members of source may be mobilized to comment in the linked target post

Inter-community interaction happens if a hyperlink mobilizes users from the source to the target community
How does the left (source) post refer to the right (target) post?

A. With neutral or no opinion
B. With a negative opinion

Defining conflicts using crowdsourcing

- Turkers labeled 1000 pairs of source-to-target posts
- We developed text classifier (0.80 AUC) to label remaining pairs
- We define conflicts as interactions that are initiated with negative sentiment.
- Identified 1800 conflicts

Conflicts = Interactions initiated by negative-sentiment source post
Our model: Three phases of conflict

Which communities engage in conflicts?

Question: Are all communities prone to conflict, or is it restricted to a few bad apples?

Our solution:
• Create who-posts-where network
• Generate embedding vector for each user and community, similar to word2vec
• Vectors learned to maximize probability of a user posting in a community

Dot = community

Blue dot = community that initiates fewer conflicts

Red dot = community that initiates more conflict

1% of communities start 74% of all conflicts

Conflicts are concentrated in some areas

Who do communities attack?

Question: Do communities attack other random communities, or is there a relation between the source and target community?

Our solution:

• TF-IDF similarity between communities:
  • Create word vector for each community from its posts
  • Calculate cosine similarity between source and target community

• TF-IDF similarity is 1.5x expected value

Phases of conflict

- Initiated by handful of communities
- Attack similar, but opposing, communities

Source \rightarrow Target

Initiation \rightarrow Interaction \rightarrow Impact

**Hypothesis 1**: attackers and defenders reply significantly to one another.

**Hypothesis 2**: attackers and defenders primarily reply to users of the same type.

Legend:  
- **Attacker node**
- **Defender node**
A-PageRank: Run PageRank but restrict the teleport set to just attackers.
• Quantifies node centrality with respect to all attackers.
Echo-chambers form during conflicts

Attackers have higher average A-PageRank scores than defenders.

Defenders have higher average D-PageRank scores than attackers.

So, attackers are closer to other attackers.

So, defenders are closer to other defenders.
Echo-chambers form during conflicts

**Hypothesis 1**: attackers and defenders reply significantly to one another

**Hypothesis 2**: attackers and defenders primarily reply to other users of the same type

Ganging-up effect during conflicts

• Some defenders are very close to attackers: 10x average A-PageRank score
• Most defenders are unreachable: zero A-PageRank score
• Linguistic analysis shows attackers swear more in replies to defenders

Phases of conflict

- Initiated by handful of communities
- Attack similar, but opposing, communities

- Conflicts create echo-chambers
- Attackers gang-up on defenders
Do conflicts change future engagement?

If activity increases, then conflicts make users more loyal and active

OR

If activity decreases, then conflicts drive users away

Attacksers “colonize” the target community and defenders leave.

What prevents colonization?

Future activity - previous activity in target community

More active

Less active

How to defend against attacks?

Successful-attack reply network
(Defenders become less active)

Successful-defense reply network
(Defenders become more active)

Legend:
- Red: Attacker
- Green: Defender

Attacker “gang-up” on defenders.

Attackers are close to other attackers, and have higher A-PageRank.

Defenders engage with attackers, and have higher D-PageRank.

Successful vs unsuccessful defense properties

When defense is successful:

- Defenders reply directly more to attackers
- Attackers and defenders are closer to each other in the reply network
- Defenders tend to use more ‘anger’ words

Direct and angry replies to attackers (“fighting-back”) marks a successful defense.
Phases of conflict

- Initiated by handful of communities
- Attack similar, but opposing, communities

- Conflicts create echo-chambers
- Attackers gang-up on defenders

- Conflicts lead to colonization
  - Successful defense: direct heated engagement with attackers
Can we predict conflicts before they happen?

Task: Given a post from source to target community, will it lead to a conflict?

Mobilization of attackers

No mobilization

Predicting conflicts

• We create a “socially-primed” LSTM structure.
• Takes user, community, and word embeddings as input for the prediction.

• A strong feature baseline gets 0.67 AUC
• Socially-primed LSTM gets 0.72 AUC
• Combination of both gets 0.76 AUC

Community Interaction and Conflict on the Web

- Initiated by handful of communities
- Attack similar, but opposing, communities

- Conflicts create echo-chambers
- Attackers gang-up on defenders

- Conflicts lead to colonization
- Successful defense by direct heated engagement with attackers

- Conflicts predicted with 0.76 AUC
- More results on positive inter-community interactions in the paper

Data and code: snap.stanford.edu/conflict