Anti-social behavior in online spaces

Bad things happen on the web…
Anti-social behavior in online spaces

Bad things happen on the web…

Misinformation
Anti-social behavior in online spaces

Bad things happen on the web…

Misinformation

Radicalization

Anti-social behavior in online spaces

Bad things happen on the web…

- Misinformation
- Radicalization
- Hate-speech
Anti-social behavior in online spaces

Bad things happen on the web…

Misinformation

Radicalization

Hate-speech

…
Intercommunity conflict

What leads to intercommunity conflict on the web?

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

long-term decrease in engagement (and presumably well-being)

post negative/hateful content

vote manipulation, etc....

r/conspiracy

r/Documentaries
Reddit dataset

40 months (2014-2017)
1.8 billion comments
100+ million users
20,000+ communities
Defining conflict: intercommunity mobilizations

*Source* community links to *target* community.
Defining conflict: intercommunity mobilizations

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

Source community links to target community.
Defining conflict: intercommunity mobilizations

Source community links to target community.

Users from source mobilize to target (with potential negative consequences)

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

*conflict/interaction

sentiment analysis
Conflict: the three stages

1. Initiation

2. Interaction

*conflict/interaction

3. Impact

users driven away!

~20,000 detected cases of intercommunity conflict in our data.
Initiation: Which communities start conflicts?

blue = “initiates less conflict”
red = “initiates more conflict”

74% of conflicts are started by 1% of communities
Initiation: Which users start conflicts?

Highly active “core” members start conflicts, but “peripheral” members carry them out.

Users with more within-community posts are better at mobilizing others.
Interaction: What happens during conflicts?

Conflicts tend to involve lots of anger words, etc… …but who are the central users in these interactions?

Idea: PageRank on who-replies-to-who networks

Finding: “Attackers” tend to be more central
Interaction: “Echochambers” during conflicts

Idea: “Attacker” and “Defender” PageRanks

Run PageRank but restrict the teleport set to just attackers/defenders:

Quantifies centrality from perspective of only defenders/attackers.
Interaction: “Echochambers” during conflicts

Idea: “Attacker” and “Defender” PageRanks

Run PageRank but restrict the teleport set to just attackers/defenders:

Quantifies centrality from perspective of only defenders/attackers.
Interaction: “Echochambers” during conflicts

Idea: “Attacker” and “Defender” PageRanks

Run PageRank but restrict the teleport set to just attackers/defenders:

Quantifies centrality from perspective of only defenders/attackers.
Interaction: “Echochambers” during conflicts

Idea: “Attacker” and “Defender” PageRanks

Run PageRank but restrict the teleport set to just attackers/defenders:

Quantifies centrality from perspective of only defenders/attackers.
Interaction: “Echochambers” during conflicts

Idea: “Attacker” and “Defender” PageRanks

Run PageRank but restrict the teleport set to just attackers/defenders:

Quantifies centrality from perspective of only defenders/attackers.
Interaction: “Echochambers” during conflicts

Idea: “Attacker” and “Defender” PageRanks

Run PageRank but restrict the teleport set to just attackers/defenders:

Quantifies centrality from perspective of only defenders/attackers.
Interaction: “Echochambers” during conflicts

Idea: “Attacker” and “Defender” PageRanks

Conflicts form echochambers!!

E.g., attackers are only central to other attackers.
Impact: How do conflicts impact the target users?

Attackers “colonize” the target community and defenders leave.
Impact: How do conflicts impact the target users?

Attackers “colonize” the target community and defenders leave.

What prevents this?
Impact: What makes a “successful” defense?

Unsuccessful-defense reply network

Successful-defense reply network

Legend:
- Red: Attacker
- Green: Defender

Attackers “gang-up” on defenders.

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

Defenders engage with attackers, and have higher D-PageRank.
... conclusions and other fun stuff

NLP and network analysis can give us social insights!!

We also built a deep learning model to predict conflict.

Collaborators (besides me and Dan):

Srijan Kumar
Postdoc, Stanford CS

Jure Leskovec
Associate Prof., Stanford CS

Discussion happening on Reddit right now!!
https://www.reddit.com/r/MachineLearning/comments/841lhj/d_community_interaction_and_conflict_on_the_web/