Sockpuppets and conflicts on social media

Srijan Kumar

http://stanford.edu/~srijan

Postdoctoral researcher, Computer Science
Stanford University
Web: A platform for everyone
Web allows...

- Web enables social interaction
- Web is not a static library that people passively browse
- Web is a place where people:
  - Act as prosumers, i.e., content producers and content consumers
  - Interact with other people:
    - Internet forums, Blogs, Social networks, Twitter, Wikis, Podcasts, Slide sharing, Bookmark sharing, Product reviews, Comments, …
Why study interactions on the web?

• Opinions are formed and shaped online via social interactions
  – Web is a good proxy for people’s opinions
• People gather majority of their information online via interactions with friends and family

![Facebook news driven by friends and family](image)

*Graphic Source: Pew Research Center’s Project for Excellence in Journalism / 2012 State of the News Media*
...but the web has increasingly become polarized.

This leads to deception and conflicts

Reference: Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Balancing opposing views to reduce controversy. WSDM 2017
Deception via multiple account use in social media

• Anyone can create more than one account in social media
• This is especially true to win arguments and infiltrate opposing groups
• We study million of comments made on online discussion platforms to analyse how sockpuppets are used for deception
Conflict on the web

• Online groups interact with one another
  – Prior research has shown that positive interactions leads to exchange of information and ideas [1, 2]
  – …but communities can fight and attack one another to win arguments, infiltrate, and destabilize them [2]
• Conflict between groups can potentially lower user engagement on the platform
• Therefore, we answer how inter-group conflict occurs in and impacts social media

Inter-group conflict is a by-product of formation of polarized groups

References:
Sockpuppets in online discussions

Sockpuppets in Online Discussions

To spread false information and push one’s own agenda, malicious entities use more than one account simultaneously.
Sockpuppets in Online Discussions

Why DC is better than Marvel

April 28, 2013 by Eric_17

bdiaz209
April 28 2013, 11PM
Possibly the best blog I’ve ever read major props to you

Eric_17
April 28 2013, 12AM
Thanks. I knew Marvel fans would try to flame me, but they have nothing other than “oh that’s your opinion” instead of coming up with their own argument

Fellstrike
April 29 2013, 6PM
Quit talking to yourself, *******. Get back on your meds if you’re going to do that

bdiaz209 posts only on this discussion to support and defend Eric_17
Wikipedia blocks hundreds of 'scam' sock puppet accounts

2 September 2015 | Technology

Sock puppetry and fake reviews: publish and be damned

The Hand That Controls the Sock Puppet Could Get Slapped

By BRAD STONE and MATT RICHTEL. JULY 16, 2007
Sockpuppets vs bots

• Bots are semi or fully automated, while sockpuppets are completely operated by humans.

• Bots do simple tasks like retweet and tweet and simple replies, while sockpuppets engage in actual conversation and arguments.

• Bots operate in 1000s, while sockpuppets operate in smaller number (typically 2-10).

Sockpuppets give an early look into the future of interactive malicious bots
Why are sockpuppets created?

- Diversify identity
- Privacy concern
- Anonymize identity
- Multiply identity

References: Gilbert, et al. (CHB, 2014), Caspi et al. (C&B, 2006)
Data: Online discussion platforms

Disqus is a plugin that enables commenting on any website.

We have all comments made on:

2.9M Users
2.1M Articles
62M Posts
2 Years
Defining sockpuppets

No ground-truth sockpuppet labels.

So, we adopt currently used definition from Wikipedia, after statistical validation for our task, as follows:

Sockpuppets are users that post from the same IP address in the same discussion very close in time (15 min), in at least 3 different instances.

3,656 Sockpuppets
1,623 Puppetmasters (not a real account)

Note: we use the IP addresses only for defining sockpuppets, but not in the detection tasks later.
Characteristics of sockpuppets
How to compare sockpuppets and ordinary users?

We have to match.

For each sockpuppet, match an ordinary user that makes similar number of posts on similar discussions.
Where do sockpuppets post?

Most sockpuppets are active in controversial/opinionated topics.
Relation between pair of sockpuppets

Interact more with each other

\[ p < 10^{-3} \]

Falcon-X32
Feb 5 2013, 3PM
I agree. You are absolutely right!

Smoothzilla
Feb 5 2013, 3PM
Thanks for your support!!!!

jakey008
Feb 5 2013, 2PM
should have read the reviews first :(.

ricobean27
Feb 5 2013, 3PM
Couldn’t agree more.

Upvote each other more

\[ p < 10^{-3} \]
Do puppetmasters lead double lives?

Double life hypothesis:
Puppetmaster maintains distinct personality for the two sockpuppets

Similarity is measured as cosine similarity between user posts’ features: LIWC, sentiment, number of words, etc.
Do puppetmasters lead double lives?

Alternate hypothesis:
Puppetmaster operates both sockpuppets similarly

Ordinary ——— Sockpuppet 1 ——— Sockpuppet 2

Less similar ——— More similar

Similarity is measured as cosine similarity between user posts’ features: LIWC, sentiment, number of words, etc.
Do puppetmasters lead double lives?

Both sockpuppets are more similar to each other

\[ p < 10^{-3} \]

“Good sock/Bad sock” strategy is not common
Why are sockpuppets created? Only for deception?
Hypothesis: Deceptive sockpuppets of the same master have very different usernames.

Majority of sockpuppets are deceptive.
Pretender vs Non-pretender Sockpuppets

srijan
Feb 5 2013, 2PM
best article i have read!!!

ricobinson
Feb 5 2013, 3PM
But this article doesn't make any sense

More opinionated $p < 10^{-3}$

theRealBatman
Feb 5 2013, 3PM
YOU ARE STUPID AND A ****

Srijan Kumar. Sockpuppets and conflicts on the web.
How are sockpuppets used?
Do sockpuppets always support one another?
Neutral sockpuppets

We quantify the amount of support by counting assenting, negation and dissenting words from LIWC

srijan       Nov 3, 2018, 3PM
I agree with you

theRealBatman Nov 3, 2018, 3:05 PM
why is that?

60%
Neutral
Supporter sockpuppets

We quantify the amount of support by counting assenting, negation and dissenting words from LIWC.

srijan
Feb 5 2013, 3PM
I agree with you

theRealBatman
Feb 5 2013, 3PM
I second that!

60% Neutral
30% Supporter
Dissenter sockpuppets

We quantify the amount of support by counting assenting, negation and dissenting words from LIWC.

srijan  Feb 5, 2013, 3PM
I agree with you

theRealBatman  Feb 5, 2013, 3PM
Huh, you are stupid

60% Neutral
30% Supporter
10% Dissenter
Supportiveness and Deceptiveness

Deception is important to create an illusion of public consensus.

High support is marked by high pretention.
Detecting sockpuppets
Features

Activity
Number of posts, number of replies, reciprocity of posts, age of account, ...

Post
Number of words, characters, etc., LIWC counts, Readability, Sentiment, ...

Community
Number of upvotes and downvotes, Fraction of reported posts, Is account reported, ...

Note: we are not using the IP based features
Is an account a sockpuppet?
Is an account a sockpuppet?

- **Post**: 0.57
- **Community**: 0.54
- **Activity**: 0.59
- **All**: 0.68

AUC
Do two accounts belong to the same person?
Do two accounts belong to the same person?

![Bar chart showing AUC values for different categories: Post, Community, Activity, and All. The AUC values are: Post 0.80, Community 0.56, Activity 0.86, and All 0.91.](image-url)
Conclusion

Spectrum of sockpuppetry

Benign usage by non-pretender sockpuppets: Primarily created to separate interest, are respectful, operate similarly, and are neutral towards each other

Malicious usage by pretender sockpuppets: Primarily created to create an illusion of consensus, are abusive, and support & defend each other
Conflicts between online communities

Why study inter-community interactions?

• Users organize themselves into groups, 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
• Why study Reddit? #5 site in USA, #19 in the world.
Conflict between communities: An example

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

Members go and post negative/hateful comments

• Can disrupt communities
• Can decrease long-term engagement

Conspiracy

Documentaries

Understanding how communities fight and how to prevent conflicts is important to reduce polarization and 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?
Defining inter-community interactions

“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

Attackers

Defenders

Inter-community interaction happens if a hyperlink mobilizes users from the source to the target community
Defining conflicts using crowdsourcing

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

- Amazon Mechanical 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

Highly similar communities attack each other
Phases of conflict

• Initiated by handful of communities
• Attack similar, but opposing, communities
Attacker-Defender Interactions

**Network is created as follows:** nodes are users and edge \((i, j)\) means node \(i\) replies to node \(j\).

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

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

**Legend:**
- Red: Attacker node
- Brown: 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. ⇒ Attackers are closer to other attackers.

Defenders have higher average D-PageRank scores than attackers. ⇒ 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

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

Legend:
- Attacker node
- Defender node
Phases of conflict

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

Future - previous activity in target community

Attackers “colonize” the target community and defenders leave.

What prevents colonization?
How to defend against attacks?

Successful-attack reply network
(Defenders become less active)

Successful-defense reply network
(Defenders become more active)

Attackers “gang-up” on defenders.

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

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

Legend:
- Red: Attacker
- Green: Defender
Successful vs unsuccessful defense

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
- Attackers gang-up on defenders
- Conflicts lead to colonization
- Successful defense: direct heated engagement with attackers
Predicting 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
Conflicts on the web

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

- 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
Conclusions

• **Sockpuppets**: Used primarily for malicious purposes. Work by creating an illusion of consensus in discussion.

• **Conflicts**: Communities attack one another online, that has reduces the user-base of the attacked community

• **Detection** works well with feature-based methods

• More at [http://stanford.edu/~srijan](http://stanford.edu/~srijan)

Thanks!
I’m in the job market!