Sockpuppet Detection in the English Wikipedia

Srijan Kumar, Leila Zia, Jure Leskovec

Sockpuppets

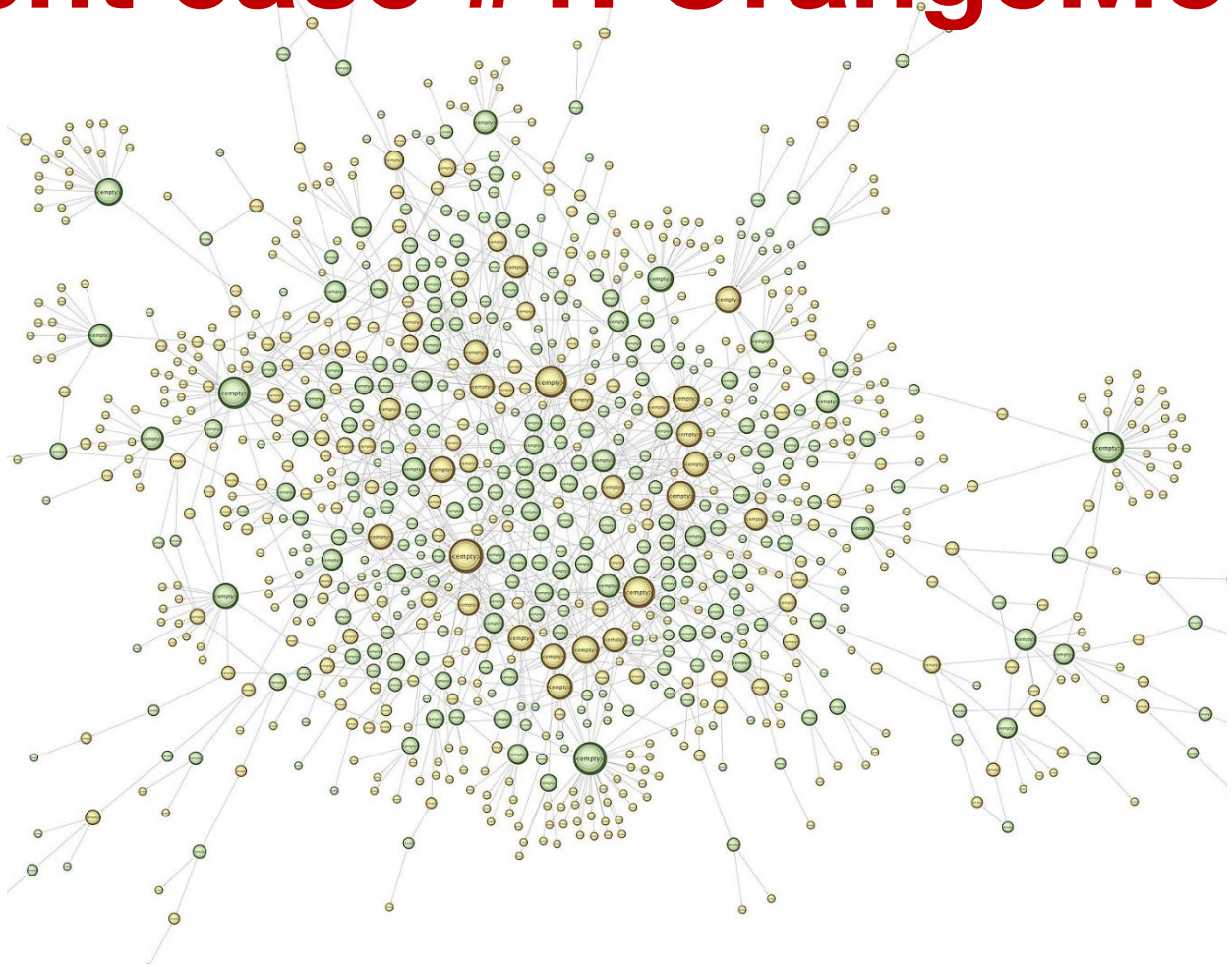
- **Definition:** Use of more than one account
- **Reasons of sockpuppetry:**
  - Benign, e.g., work vs personal account
  - Malicious, e.g., point of view editing
- **Malicious sockpuppet abuse is harmful and continues for long-term**

Recent case #1: OrangeMoody

- **381 socks used for paid/promotional editing**
- Edits from April, 2015 to August, 2015
- **Well-planned editing strategy:**
  - **"Article creation" socks:** created promotional articles
  - **"Helper" socks:** added content, promo links
- Lot of hard and investigative work by the **Checkuser team** to find these accounts
- Still in "**active**" status

Reference: https://en.wikipedia.org/wiki/Wikipedia:Long-term_abuse/Orangemoody

Recent case #1: OrangeMoody



OrangeMoody accounts: Yellow bubbles represent IP addresses, and green bubbles represent registered accounts. Edges mean co-editing a page.

Reference: https://en.wikipedia.org/wiki/Wikipedia:Long-term_abuse/Orangemoody

Recent case #2: Morning277

- **323 socks** from August, 2010 (still "**active**")
- **How they got caught:** multiple accounts were editing without ever using talk pages
- **Complex editing strategy:**
  1. Create new account. Auto-confirm by making several trivial edits
  2. Create article page in sandbox
  3. Add significant content and images to the article. Look credible: cite links from external websites
  4. Use another account to remove sandbox and move to main article space
  5. Abandon these accounts and repeat from step 1

Reference: https://en.wikipedia.org/wiki/Wikipedia:Long-term_abuse/Morning277

Challenges in finding sockpuppets

- Sockpuppets use **complex editing strategies**
- They split suspicious edits across **multiple accounts**
- Unlike vandals, they work **slowly and cautiously** to avoid getting caught

- Thus, it requires a lot of **hard work by Checkusers to detect them**

Main question:

How can we help Checkusers to find sockpuppets before they do harm?

Our solution:

Use public data and surface suspicious accounts to Checkusers for verification as soon as possible

Our Task

- **Input:** All edits made by all accounts (IP addresses and registered users)

- **Output:** A ranking of all accounts based on their probability of being a sockpuppet

Dataset

- We create a big dataset from **all edits made in one entire month on the English Wikipedia**
  - Number of **users**: 446,075
  - **Sockpuppet users: 1,601** (0.35% of all users!)
  - Total number of **pages**: 1,349,918
  - Number of **edits**: 4,418,932
- **Setting:** we use first 27 days of edits for model training and last 3 days of edits for model evaluation
- **Performance metric:** AUROC (max value is 1, higher is better)

Our Approach

- We create **two machine learning** models:
  1. **Feature engineering solution:** more interpretable models
  2. **Deep learning solution:** higher performance expectation, as it can capture non-trivial user-user interactions
- We use **confirmed sockpuppet accounts** to train our models

Model #1: Feature engineering

- For each user, we extract **>800 features** across multiple attributes:
  - **Text:** number of characters, words, punctuations, pronouns, edit sentiment, readability, psycholinguistic attributes
  - **Activity:** number of edits, fraction of edits on the same page
  - **Article:** fraction of edits made on namespace = 0
  - **Time:** time since previous edit, distribution of time difference between edits
  - **Network:** embedding vectors from who-edits-what network

Model #1: Feature engineering

- Once edit features are extracted, we train a **logistic regression classifier** on the first 27 days of edits
- The model is trained to predict a:
  - 0 if the user is only using a single account
  - 1 if the user is using another account simultaneously
- **Prediction result:** 0.7941 AUROC

Model #2: Deep learning

- Uses the **network of who-edits-what** to learn a representation of each user and each article
- Representations capture complex relations:
  - **Two users** have a similar representation if they edit similar articles in similar times
  - **Two articles** have a similar representation if they are edited by similar users in similar times
- These representations are **trained to predict if user is sockpuppet**. More details on the project page
- **Prediction result:** 0.821 AUROC

Example Result #1: 128.61.83.176

- We identified this account to be suspicious **3 days before** it was banned by moderators
- The account seems normal at a high-level from its activity
- **Looking closer at its edits** indicates sockpuppetry, vandalism, and bad edits

Example Result #1: 128.61.83.176

Georgia Bulldogs football: Difference between revisions

From Wikipedia, the free encyclopedia

Browse history interactively

Revision as of 05:24, 20 January 2016 (view source) Jim1138 (talk | contribs) m *(Reverted edits by 128.61.83.176 (talk) (HG) (3.1.18))* ← Previous edit	Revision as of 05:28, 20 January 2016 (view source) 128.61.83.176 (talk) Next edit →
Line 1:	**Line 1:**
− {{Use mdy dates|date=**September** 2011}}	+ {{Use mdy dates|date=**Septem er** 2011}}
− {{**Infobox** NCAA **football** school	+ {{**Info ox** NCAA **foot all** school
− | CurrentSeason = 2015 Georgia **Bulldogs football** team	+ | CurrentSeason = 2015 Georgia **ulldogs foot all** team
− | TeamName = Georgia **Bulldogs football**	+ | TeamName = Georgia **ulldogs foot all**
| Image = UGA logo.svg	| Image = UGA logo.svg
| ImageSize = 150	| ImageSize = 150
| Helmet =	| Helmet =
| ImageSize2 =	| ImageSize2 =
− | HeadCoachDisplay = [[**Kirby** Smart]]	+ | HeadCoachDisplay = [[**Kir y** Smart]]

Removing the character “b” from a page

Example Result #2: CCL-DTL

- Blatant vandalism and harassment of other users using more than one account
- Active from 2015

Example Result #2: CCL-DTL

WIKIPEDIA
The Free Encyclopedia

Not logged in Talk Contributions Create account Log in

User page | Talk | Read | Edit | More | Search Wikipe

User:CCL-DTL

From Wikipedia, the free encyclopedia

leftists are brainlets that think a system that flopped think "omfg the real shiEt hasn't been tried yet!!11!!" should explain why the dprk literally begged ebil imperialist for food and oil and only showed a small portion of area that been "fl00Ded"

Main page
Contents
Featured content
Current events
Random article

User page today

Example Result #2: CCL-DTL

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction

Help
About Wikipedia
Community portal
Recent changes
Contact page

User:CCL-DTL

From Wikipedia, the free encyclopedia

This is an old revision of this page, as edited by CCL-DTL (talk | contribs) at 07:37, 23 February 2015. The present address (URL) is a permanent link to this revision, which may differ significantly from the current revision.

(diff) ← Previous revision | Latest revision (diff) | Newer revision → (diff)

Hi

User_talk:Coryentymatototok is stupid

Can someone try to pacify this User:Seidico?

He did a very nonsensical edit on Portstation MRT Station.

I reverted it. Thanks!

Mayan Mayo Lover:

User_talk:2401:7400:C800:37DA:6C2F:93B4:E86:97E5

and a false info editor:

User_talk:175.156.203.70

Same user in February, 2015

Discussion and Next Steps

- We will present the result of this research to checkusers to get their feedback. If you're a checkuser, we want to talk with you in Wikimania.
- Here are some questions for you:
  - What are your thoughts about the trade-off between receiving the possible sockpuppet information early vs. accepting that sockpuppets can go undetected and they can create damage?
  - Any other thoughts you want to share?

Sockpuppet Detection in the English Wikipedia

Leila Zia, Srijan Kumar, Jure Leskovec

More details at: https://meta.wikimedia.org/wiki/Research:Sockpuppet_detection_in_Wikimedia_projects

Credits

- Page 1: Note that the logos used on the first slide belong to the corresponding institutions.