KDD 2017
Halifax, Nova Scotia - Canada
August 13 - 17, 2017

Tutorial: Data-Driven Approaches towards Malicious Behavior Modeling

Meng Jiang
University of Notre Dame

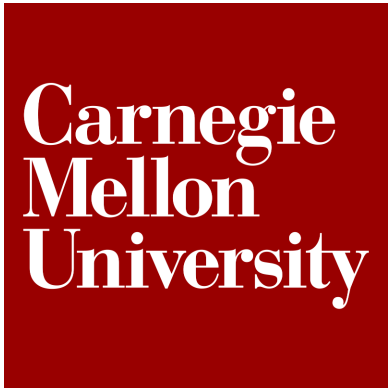
Srijan Kumar
Stanford University

Christos Faloutsos
Carnegie Mellon University

V.S. Subrahmanian
University of Maryland, College Park

http://www.meng-jiang.com/tutorial-kdd17.html

Acknowledgement



Outline

Introduction

Feature-based algorithms

- Bots
- Sockpuppets
- Vandals
- Hoaxes

Spectral-based algorithms

- Visualization: "spokes", "blocks", "staircases"
- Camouflage
- Theoretical guarantee

Density-based algorithms

- Ill-gotten Likes
- Synchronized Behaviors
- Advertising campaigns
- Social spam

Conclusions and future directions

Outline

Introduction

Feature-based algorithms

- Bots
- Sockpuppets
- Vandals
- Hoaxes

Spectral-based algorithms

- Visualization: "spokes", "blocks", "staircases"
- Camouflage
- Theoretical guarantee

Density-based algorithms

- Ill-gotten Likes
- Synchronized Behaviors
- Advertising campaigns
- Social spam

Conclusions and future directions

Social Honeypots for Spam Detection

K. Lee et al. SIGIR 2010

- MySpace: 51 honeypots over 3 months
- Twitter: Unknown number of honeypots over 2 months.
- Two step process:
  - Identify accounts that friend/follow the honeypots.
  - Use an SVM classifier to distinguish between spammers and benign accounts.

K. Lee, J. Caverlee, S. Webb. Uncovering Social Spammers: Social Honeypots + Machine Learning, *Proc. SIGIR 2010*.

MySpace Spam Profiles

- Click Traps: Users clicking on objects on the profile page are redirected to another webpage.
- Infiltrators: Spams friends of those who accept a friend request.
- Pornography: "About Me" section of the profile shows porn stories and links to porn sites
- Dubious Pills: Similar to the above
- Winnies: All these profiles have the headline "Hey its winnie" even though the rest of the profile is different. Links lead to porn sites.

Social Spammer Detection with Sentiment Information (X. Hu et al. ICDM 2014)

- Used 3 datasets
  - TAMU Honeypot data 30K users (7 months) with about a 50/50 split into benign vs. spammers
  - Twitter Suspended Spammers data. ~2 mths , ~20K users with ~4K spammers
  - Stanford Twitter Sentiment. 40K tweets over 2.5 months with labeled sentiment.

X. Hu, J. Tang, H. Gao, H. Liu. Social Spammer Detection with Sentiment Information, ICDM 2014.

1) Associate sentiment vector *s(u)* with each user *u*. *s(u)* is the vector of sentiment for ALL tweets in the data set.
2) Defined distance between two users' sentiment vectors.
3) Defined Sentiment Graph – users are nodes, edges are weighted by sentiment correlation between the two users
4) Set up the problem of finding spammers as non-convex optimization problem
5) Develop a novel algorithm to solve this problem.

Achieve high precision and recall (over 0.9 for both) on both test datasets.

Detecting Bots/Cyborgs on Twitter

(Z. Chu et al. IEEE TDSC 2012)

- Introduces cyborgs – bot-assisted human accts or human-assisted bot accts
- Developed a training set with about 2K accounts per category (human, bot, cyborg)
- Studied the main differences between these categories.

Do bots have more friends than followers? NO

- $Reputation(u) = \frac{\#followers}{\#followers+\#friends}$
- Reputation is ~1 for humans
- Cyborgs are not far behind
- Bots have a reputation score closer to 0.5

Z. Chu, S. Gianvecchio, H. Wang and S. Jajodia. Detecting Automation of Twitter Accounts: Are you a Human, Bot, or Cyborg? IEEE Transactions on Dependable & Secure Computing, Vol 9, Nr. 6, pages 811-824, 2012

Detecting Bots/Cyborgs on Twitter

(Z. Chu et al. IEEE TDSC 2012)

Does automation generate more tweets?

- Cyborgs post the most tweets
- They are followed by humans
- Then bots

Does automation yield higher tweet frequency?

- Bots are the most frequent posters
- Followed by cyborgs
- Followed by humans

Detecting Bots/Cyborgs on Twitter (Z. Chu et al. IEEE TDSC 2012)

Are bots posts more regular ?

- Based on entropy
- Let $X = \{X_i\}$ be a sequence of random vars.
- Use inter-arrival times, i.e. time since last post
- Let $\boldsymbol{P}(x_i) = \boldsymbol{P}(X_i = x_i)$.
- Entropy of sequence $H(X_1,..,X_n) = \sum_{i=1}^{m} \boldsymbol{P}(x_i) * \log(\boldsymbol{P}(x_i))$
- Conditional entropy $H(X_m|X_1,\dots,X_{\{m-1\}}) = H(X_1,..,X_m) - H(X_1,\dots,X_{\{m-1\}})$
- Entropy rate $\lim_{m\to\infty} H(X_m|X_1,\dots,X_{\{m-1\}})$.
- Bot posts have the lowest entropy, cyborgs are next, and humans have the highest entropy w.r.t. interarrival time.

How do bots post vs. humans?

- > 50% of human posts are from the Twitter website
- 42.39% of tweets by bots are from unregistered API tools.
- Tools used by bots are automatic, with no human intervention.

Detecting Bots/Cyborgs on Twitter (Z. Chu et al. IEEE TDSC 2012)

Do bots include more links in their tweets than humans?

- Average number of URLs in bot tweets is the highest
- Followed closely by cyborgs
- Followed by humans

Classification Task

- Use entropy-based features.
- Use Random Forest classifier.
- Show confusion matrix with very high accuracy in the three way classification.

CASE STUDY 1: IDENTIFYING BOTS ON TWITTER

Using Sentiment to Detect Bots on Twitter: Are Humans more Opinionated than Bots?
J. Dickerson, V. Kagan, and V.S. Subrahmanian.
ASONAM 2014

Our Approach to Training Set Creation

- Associate with each user *u*, a set of variables learned from past data.
- Data from July 15 2013 to May 15 2014 associated with bots in the 2014 Indian election
  - 25M+ tweets
  - 17M+ users
  - 45M+ edges
- 2014 Indian Election
  - Largest democratic election in history
  - Social media played huge role
- Defined set of topics of interest (TOI):
  - Political parties: Shiv Sena, BJP, …
  - Politicians: Rajnath Singh, Nitish Kumar, …

V. Kagan, A. Stevens, and V.S. Subrahmanian. Using Twitter Sentiment to Forecast the 2013 Pakistani Election and the 2014 Indian Election. *IEEE Intelligent Systems*, pp. 2-5, Jan-Feb 2015.

Sentiment Extraction

- For each user *u*, day *d*, and topic *t*:

SS(*d*,*u*,*t*): sentiment score in [-1,+1] for topic *t* averaged across all *u*'s tweets on *t* for day *d*

- Past work did not look at *topic-specific* sentiment for detecting malicious actors
- Used SentiMetrix's commercially-available:
  - SS(*d*,*u*,*t*) = -1 → "maximally negative"
  - SS(*d*,*u*,*t*) = +1 → "maximally positive"
- Could use other methods as long as they assign a sentiment score to a topic

Network Extraction

- Given a set of users *U* who tweeted about TOI
  - Collected followers of each *u* for two hops
  - Collected accounts *u* follows for two hops
- Local structure: about 45 million edges
- Allows commonly-used features like:
  - # followers
  - # friends
  - # friends / # followers

Features

- Tweet Syntax
  - E.g. #hashtags, #mentions, #links, etc
- Tweet Semantics
  - Lots of sentiment related features for user
- User Behavior
  - Tweet spread/frequency/repeats/geo
  - Tweet volume histograms by topic
  - Sentiment: normalized flip flops(t), variance(t), monthly variance(t)
- User Neighborhood (and behavior)
  - Multiple measures looking at agreement/disagreement between user sentiments and those of people in his neighborhood

Using Sentiment to Detect Bots on Twitter: Are Humans more Opinionated than Bots?,
J. Dickerson, V. Kagan, and V.S. Subrahmanian.
ASONAM 2014

Network Features

Contradiction Rank

- $CR(u,t) = x_t^+ y_t^- + x_t^- y_t^+$ where
  - x_t^+ is the fraction of *u*'s tweets with sentiment that are positive w.r.t. *t*
  - y_t^+ is the fraction of all tweets [not just *u*'s] with sentiment that are positive w.r.t. *t*
  - x_t^-, y_t^- defined similarly
- High contradiction rank => most users disagree with *u* on *t*
- Low contradiction rank => most users agree with *u* on *t*

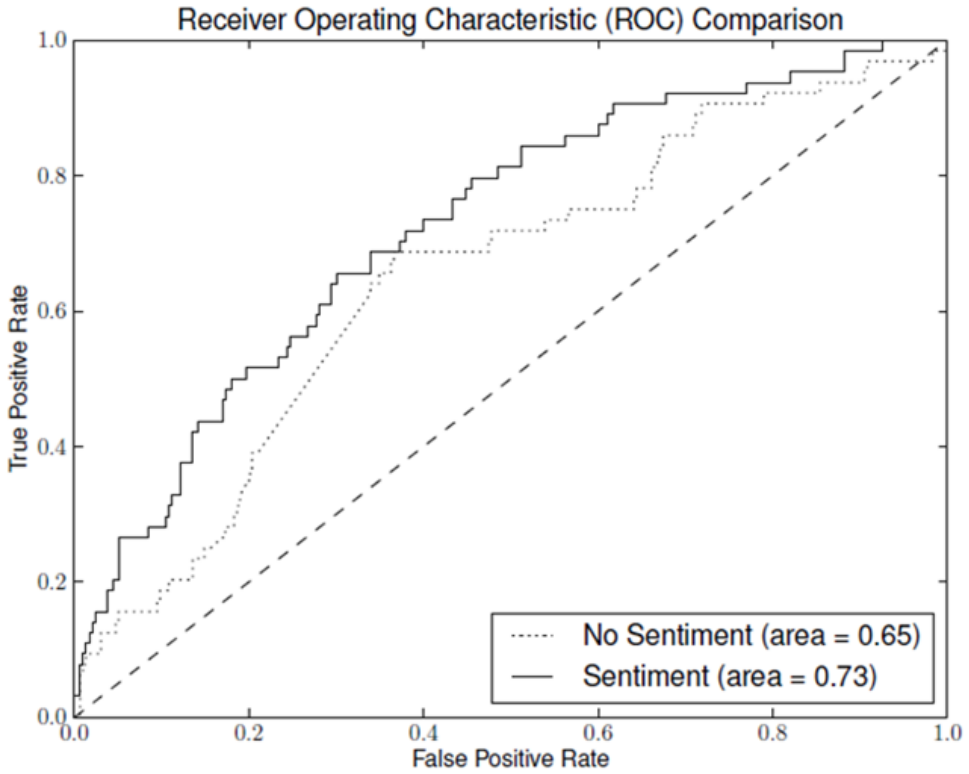
- Agreement Rank: $AR(u,t) = x_t^+ y_t^+ + x_t^- y_t^-$
- Dissonance rank of user
- $DR(u) = \sum_{t \in TOI} \frac{CR(u,t)}{AR(u,t)}$
- Positive Sentiment Strength
  - Average sentiment score (for *t*) from *u*'s tweets that are positive about *t*
- +/- Sentiment Polarity Fraction
  - Percentage of *u*'s tweets on *t* that are positive/negative

Network Features

- Neighborhood Contradiction Rank
  - Similar to contradiction rank: but y_t^+, y_t^- are computed by just considering *u*'s neighbors' tweets.
- Intuition:
  - *u*'s (global) contradiction rank could be high because *u*'s opinions on *t* are inconsistent with the majority view
  - But may be consistent with *u's* immediate neighborhood.

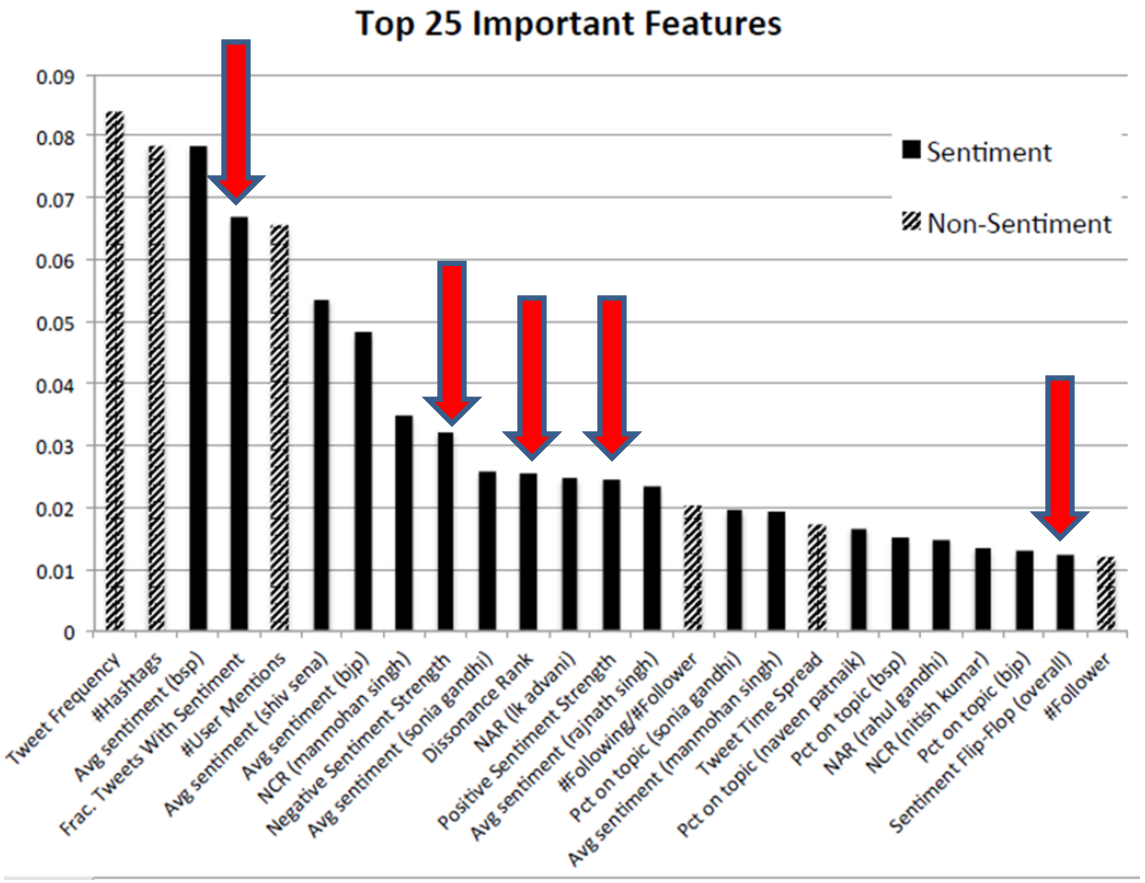
Can extend agreement rank and dissonance rank similarly

Predictive Accuracy



Which of the features do you think are the most important?

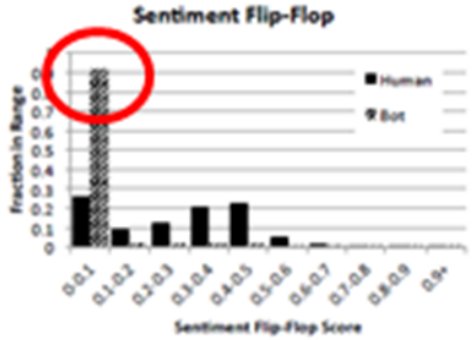
Most Important Features



Question: Humans vs. Bots

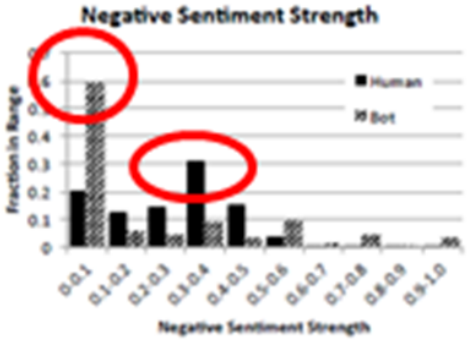
1. Do bots or humans flip flop more?
2. Whose positive opinions are stronger?
3. Whose negative opinions are stronger?
4. Who tend to write more tweets with sentiment?
5. Who tend to disagree more?

Question: Humans vs. Bots

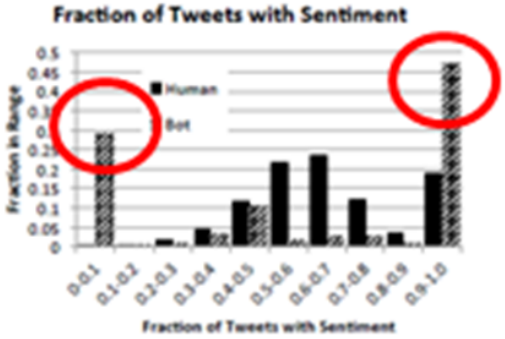


(a) Sentiment flip-flop score.

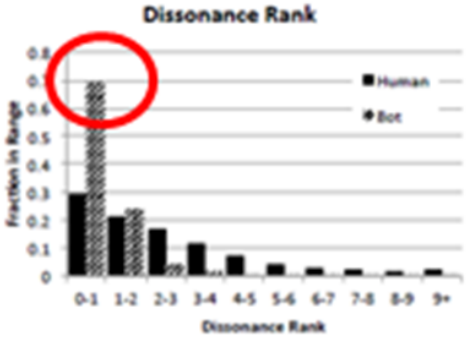
(b) Positive sentiment strength.



(c) Negative sentiment strength.



(d) Negative sentiment strength.



(e) Dissonance rank.

CASE STUDY 2: THE DARPA TWITTER BOT CHALLENGE

The DARPA Twitter Bot Challenge
V.S. Subrahmanian et al.
IEEE Computer, June 2016, pages 38-46

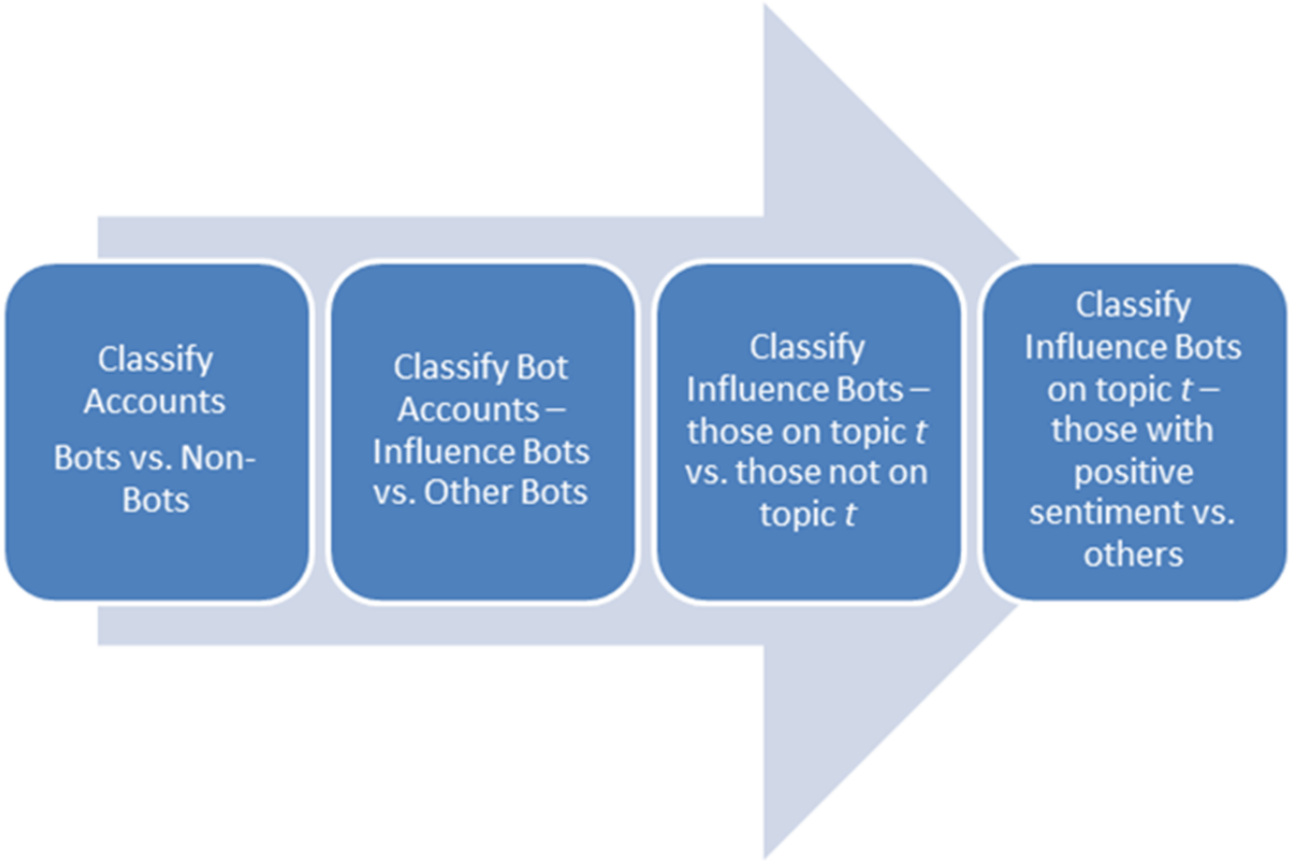
The DARPA Twitter Bot Challenge

- Run over a 28-day period in Feb/March 2015.
- One day 1, DARPA provided 4 weeks of data.
- Another 4 weeks played out in real-time.
- Goal: Identify all bots in DARPA-provided data.
- Scoring. All guesses about bots confirmed in real-time
  - 1 point for each correct guess
  - -1/4 point for each incorrect guess
- Bonus: If all bots are guessed and there are still *d* days left in the competition, you get *d* bonus points

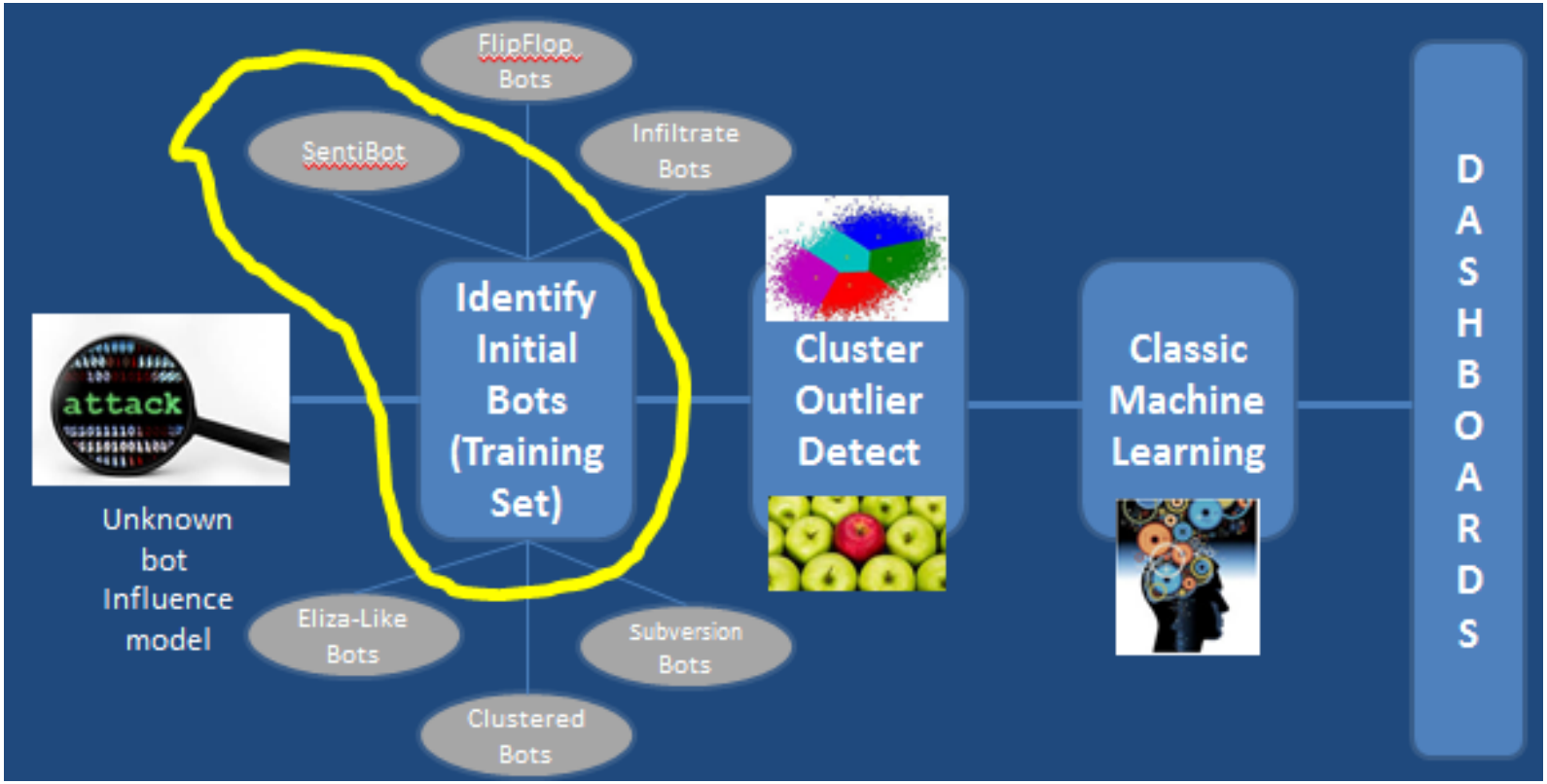
DARPA Twitter Bot Challenge Results

	Misses	Hits	Guesses	Accuracy	Speed	Final Score
Sentimetrix	1	39	40	38.75	12	50.75
USC	0	39	39	39	6	45
DESPIC	7	39	46	37.25	6	43.25
IBM	4	39	43	38	5	43
B. Fusion	9	39	48	36.75	5	41.75
G. Tech	56	38	94	24	0	24

Challenges


Classify Accounts
Bots vs. Non-Bots
Classify Bot Accounts – Influence Bots vs. Other Bots
Classify Influence Bots – those on topic *t* vs. those not on topic *t*
Classify Influence Bots on topic *t* – those with positive sentiment vs. others

Heterogeneity of Methods Used

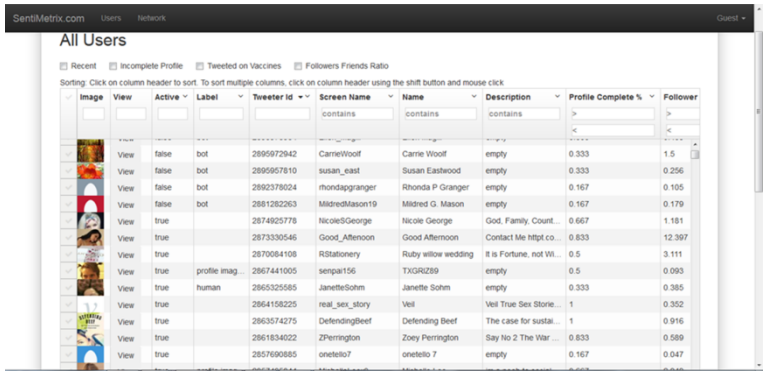


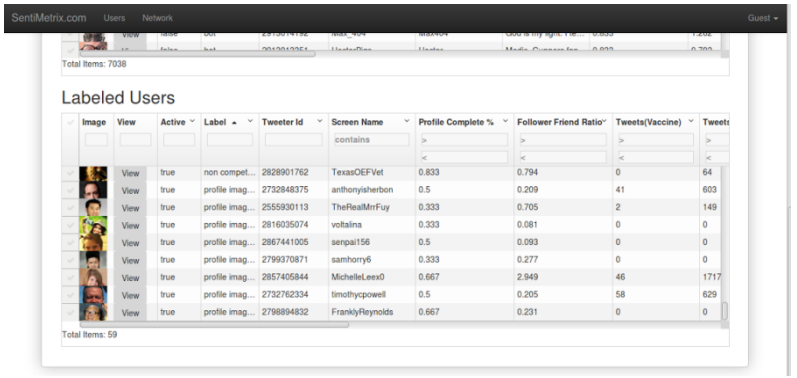
Human in the loop process used to identify bots used in new social media influence campaigns including adversary strategies never seen before.

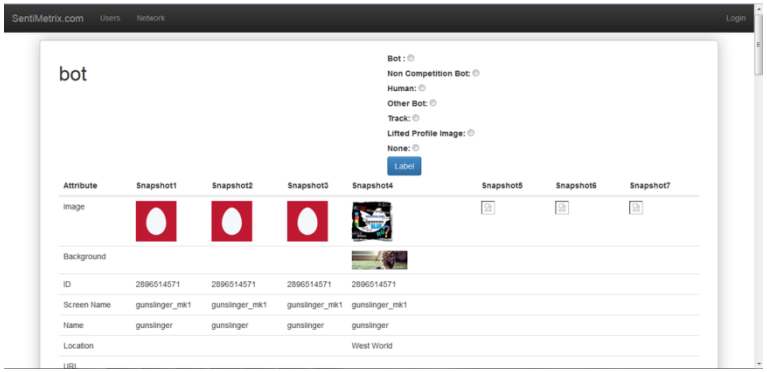
Goal 1: Find a few Initial Bots

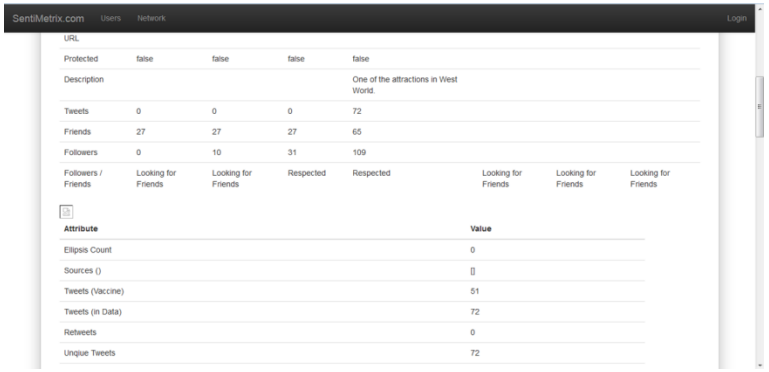
- Missing or "Stock Image" profile images
  - Landscapes/Nature
  - Middle-aged mothers
    - Used a human feature recognizer that would extract the expected age, sex, and number of humans in a profile image
      - Was not actually very useful during the competition
- Common naming patterns, e.g.
  - firstname_lastname_number
- Bots would follow other bots to bolster their # of followers and retweets ("botnet")
  - Did not actually happen as much as expected
- Similarities amongst bots. During the competition, we noticed many users were following 38-42 users
- Screwed-up Profiles. Any bots that were initially setup with incomplete profiles

Human-in-the-loop is Key









Goal 1: Flip-Flopping

- We expected bots to be firmly "pro-vax" by the end of the competition
- In SentiBot 1.0, very few Indian Election bots flipped sentiment
- In this competition, however, bots are attempting to change influence
- Hypothesize that pro-vax bots should always remain pro-vax
- Infiltration bots will remain pro-vax once they begin to "whistleblow"
- Define "positive" users as either anti-anti-vax or pro-vax
- Positive hashtags found during the competition:
  - #VaccinesWork
  - #MMRisSafe
  - #GetaFluVax

Goal 1 Hypothesis: Infiltration

- Immediately after creation, bots would begin to tweet at leaders of the anti-vax movement, such as @TannersDad, @ceestave, and @Wonderwon
- Tweets would mostly be anti-vax or neutral in sentiment, in an attempt to get the victim to retweet the bot. If the victim retweeted, then it was possible that the victim's social followers would begin to follow the bot.
- After "trapping" the anti-vax users with sweet words, would begin tweeting pro-vax resources
- In the competition, many bots attempted to Infiltrate, which we did not expect. Initially, we suspected most bots would immediately be pro-vax

Goal 1 Hypothesis: Eliza-Bots

- Eliza-bots are a well-established way to create chat bots
- http://en.wikipedia.org/wiki/ELIZA
- Often have large amounts of common subsequences. Use a DNA subsequence algorithm (Smith-Waterman) to detect
- After identifying that some of the competition bots were indeed displayed Eliza behavior, we learned a partial phrase list of 53 phrases from identified bots – suspicion of other account exhibiting such tweets went up:
  - "haha… love your opinons"
  - "Really?!"
  - "where is the evidence?"

Goal 1 Hypothesis: Clustered Bots

- We believed that bot creators would not devote a significant amount of resources to generate a bot with its own unique behavior.
- Instead, bots would come in behavioral groups of 5 or more
- Run DBScan on our extracted features to generate clusters
- Analyze social network for significant overlap in friends or followers
- Detect "same-origin" by doing the Jaccard similarity of other users compared to confirmed bots:
  - Let B be the set of unique tweets made by a confirmed bot
  - Let U be the set of unique tweets made by a user
  - Avg. Jaccard = mean(| B ^ U | / | B v U |)

Goal 1 Hypothesis: Subversion Bots

- Bots would substitute links in anti-vax or neutral-vax tweets with links to informative, pro-vax resources
- May also include memes or content intended to confuse and annoy anti-vaxxers.
- Unlike our other hypotheses, this behavior started occuring two weeks into the competition, rather than immediately

- *They lied, we knew 10 years ago, we saw the truth. #CDCwhistleblower #BREAKaBillion for truth in #autism http://bit.ly/16bBiEc*

Chronology

- *Week 1: No guesses*
- *Day 8: Guessed two bots that used very short adverb-adjective combinations.*
- *Day 9: Used similarity metrics to guess two more bots.*

AVA: Adjective Verb Adverb Combinations for Sentiment Analysis
D. Reforgiato and V.S. Subrahmanian
IEEE Intelligent Systems, Vol. 23, 4, pp. 43-50, July/Aug 2008.

Chronology II

- *Day 10: Found 4 clusters (nature, Robo_, Lowercase, NurseMama) of similar bots.*
  - Performed DBScan on their friends/followers social network to see if we could find similarily named users with similar friends
  - Look for friends/followers that followed confirmed bots and other users
  - Jaccard similarity of user tweets versus confirmed bots

Chronology III

- *Days 10-12: Found 4 clusters (nature, Robo_,Lowercase, NurseMama) of similar bots.*
  - Perform DBScan on their friends/followers social network and see if we could find similarily named users with similar friends
  - Look for friends/followers that followed confirmed bots and other users
  - Jaccard similarity of user tweets versus confirmed bots
- By the end of day 12, had correctly guessed 29 bots

Chronology IV

- *Days 10-12: Applied classical ML algorithms*
  - Small training set with the 29 discovered bots + 79 very obvious human accounts.
  - Trained SVM and Random Forest Classifiers with another 75 new features that we added.
  - Discovered all remaining 10 bots with classical ML.

Conclusions of the Case Studies

- New subversive influence campaigns will exhibit new techniques which we cannot fully anticipate.
- Need an architecture that can quickly and dynamically adapt to new social media attacks
- Proven in a competitive setting, winning DARPA Twitter Bot Challenge, beating mega-corporations like IBM

Effective in real world!

Outline

Introduction

Feature-based algorithms

- Bots
- Sockpuppets
- Vandals
- Hoaxes

Spectral-based algorithms

- Visualization: "spokes", "blocks", "staircases"
- Camouflage
- Theoretical guarantee

Density-based algorithms

- Ill-gotten Likes
- Synchronized Behaviors
- Advertising campaigns
- Social spam

Conclusions and future directions

Example

Why DC is better than Marvel

April 28, 2013 by Eric_17

bdiaz209

Possibly

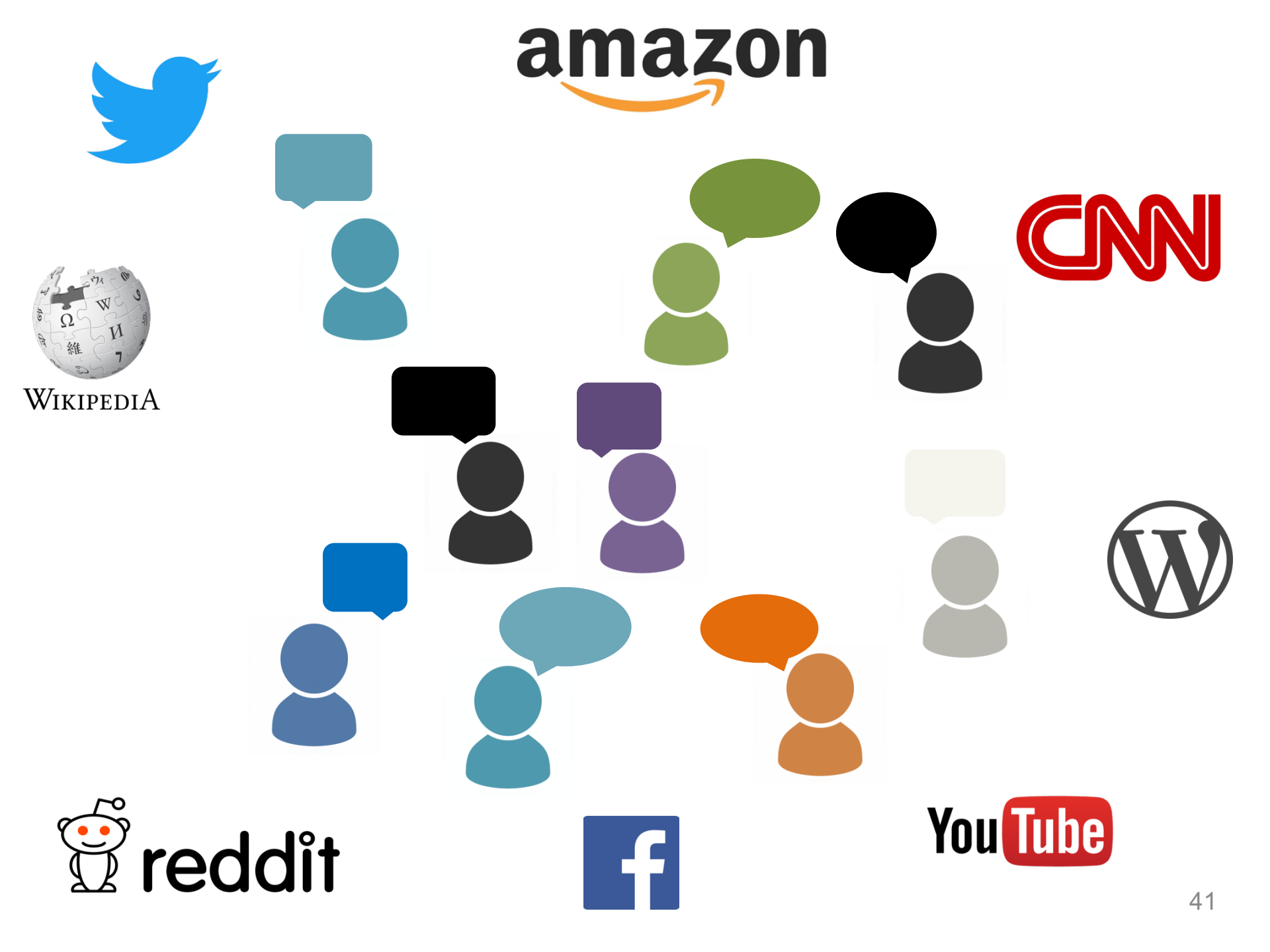
bdiaz209 posts only on this discussion to support and defend Eric_17

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


amazon
CNN
WIKIPEDIA
reddit
YouTube

Blog Home »

Why DC is better than Marvel

April 28, 2013 by Eric_17

I know I am a complete minority when it comes to this, but DC is simply better than Marvel. I'm not being bias in this because overall I love them both, but DC is much better. This is for several reasons:

Marvel is just to kid-friendly today. The newest X-Men movie, Ironman, Thor, Spider-Man, and of course the Avengers just were not serious enough. They cracked to many stupid jokes, and were not violent enough. Batman 89, Batman Returns, Watchmen and the newest Batman trilogy are how a super hero flick should be. DC movies like Batman, Superman and Green Lantern and WATCHMEN were darker with a more gruesome attitude and action. Marvel movies are just bright lights and stupid one-liners

Want to subscribe to this blog?

Grab the RSS Feed

Archives

- 2013

Categories

- uncategorized

Why DC is better than Marvel

April 28, 2013 by Eric_17

EuShock

MAY 25 2013 11AM

DC is better than Marvel but NOT because random heroes would beat other random heroes. It's because DC has a darker tone usually. It emphasizes on storyline a lot more, creates more immersive worlds and has a bunch of non-serial comics that are awesome (Watchmen, V for Vendetta, Neil Gaiman's Sandman). Marvel is probably visually more impressive with great use of colours and awesome character designs. But I still prefer substance over looks.

Eric_17

SEP 10 2013 9PM

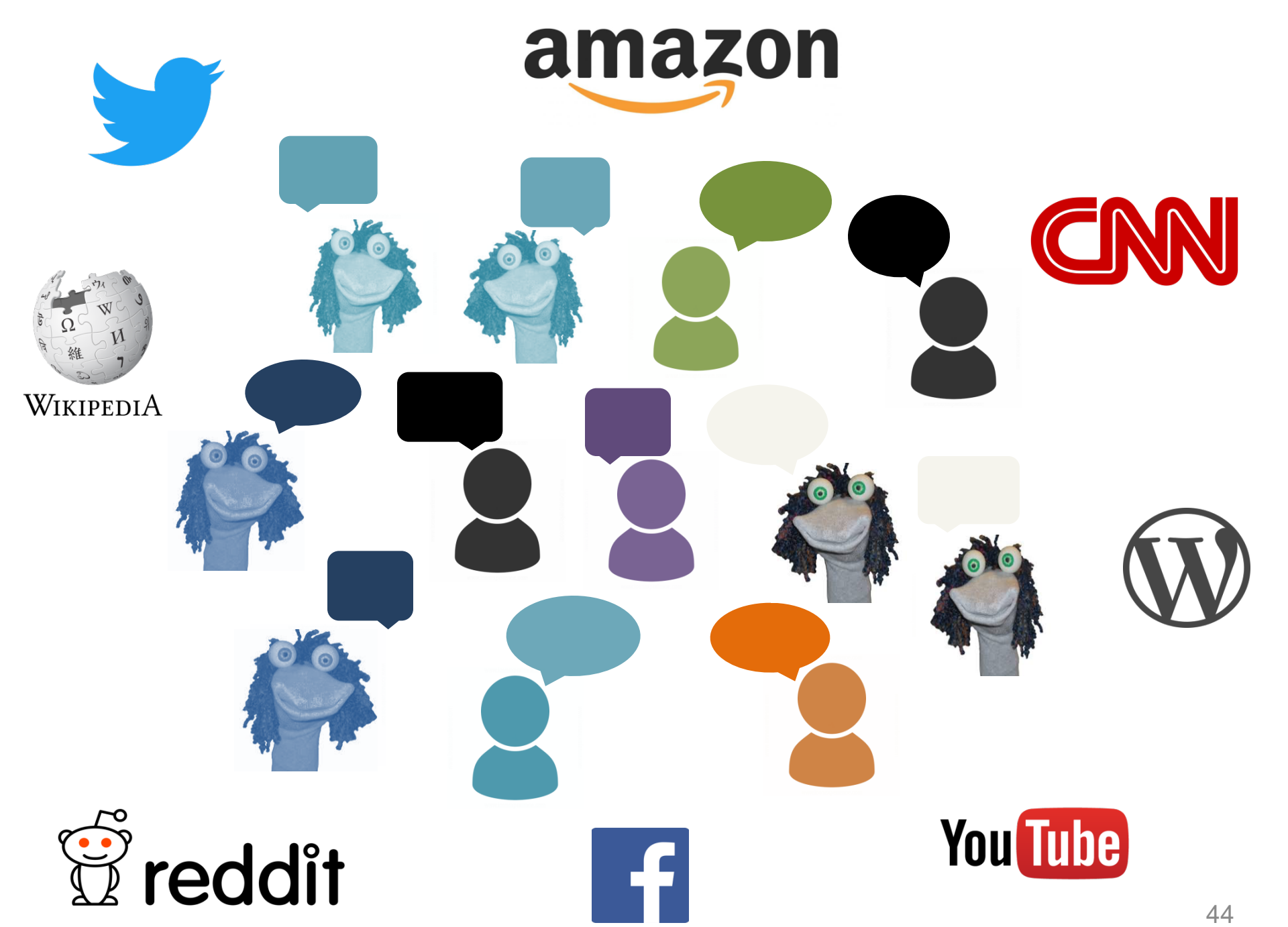
Exactly.

daddyboomboom

MAY 23 2013 8PM

I still don't see why dc is better than marvel. I personally like dc better and I like their superheroes more than marvels, but I do not think that this article tells exactly why.

(Superman and Batman are by far more popular than the avengers combined and lets face it superman has like all the avengers abilities combined.)


amazon
CNN
WIKIPEDIA
reddit
YouTube

amazon

Sock puppetry and fake reviews: publish and be damned

YAHOO!

The Hand That Controls the Sock Puppet Could Get Slapped

By **BRAD STONE** and **MATT RICHTEL** JULY 16, 2007

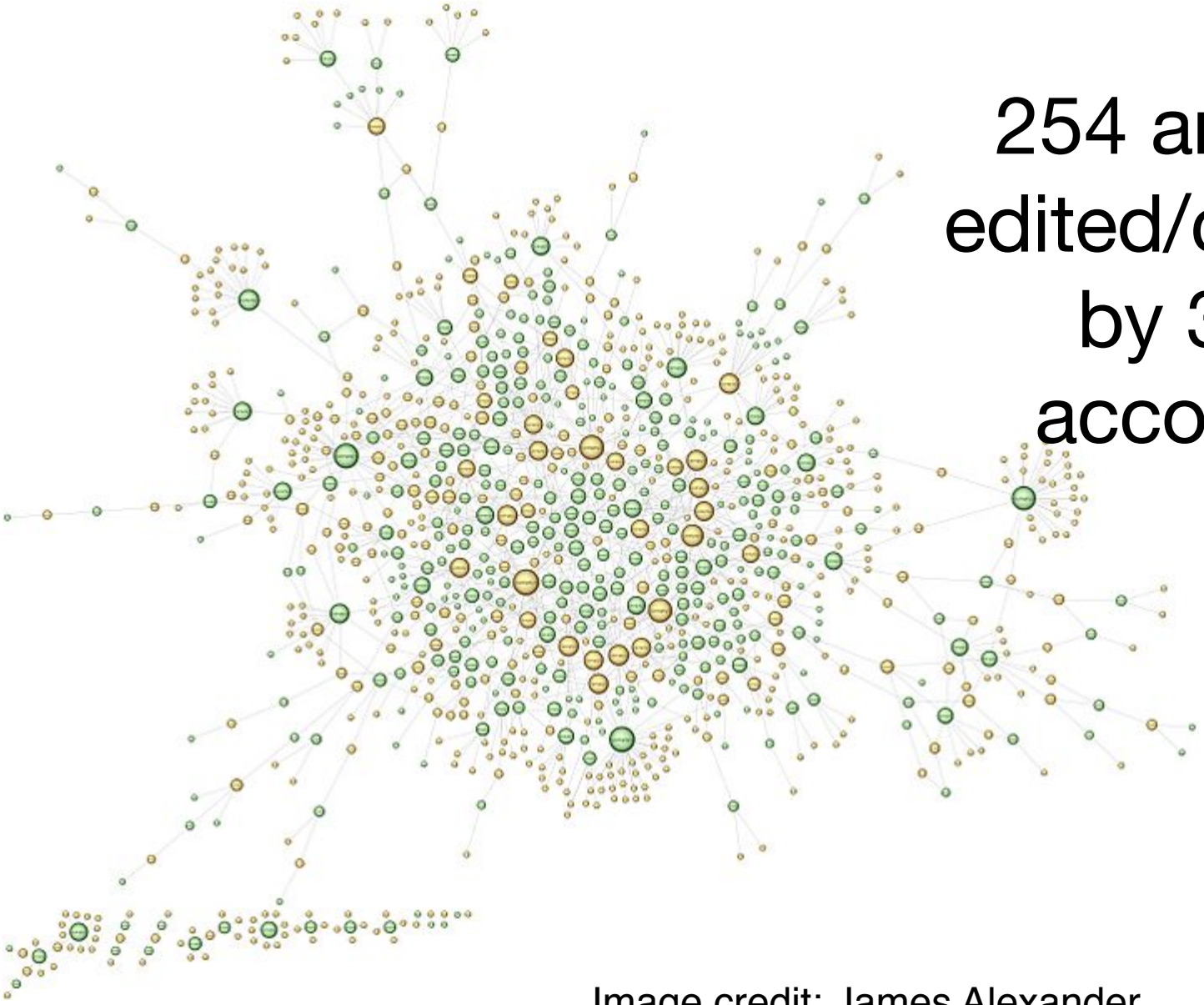
WIKIPEDIA

Wikipedia blocks hundreds of 'scam' sock puppet accounts

2 September 2015 | Technology

Share

Wikipedia: Orangemoody sockpuppet case



254 articles edited/created by 381 accounts

Image credit: James Alexander

Online sockpuppetry – why?

Diversify identity

Anonymize identity

Multiply identity

Privacy concern

Gilbert, et al. (CHB, 2014), Caspi et al. (C&B, 2006)

Sockpuppets in Wikipedia

Inappropriate uses of alternative accounts

Editors must not use alternative accounts to mislead, deceive, disrupt, or undermine consensus. This includes, but is not limited to:

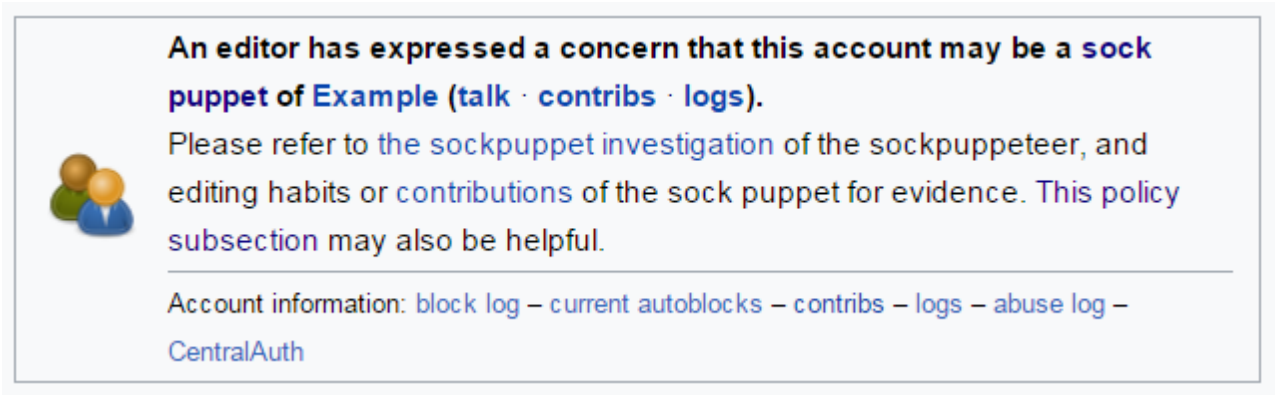
- **Creating an illusion of support**
- **Strawman socks**
- **Editing project space**
- **Circumventing policies or sanctions**
- **Contributing to the same page or discussion with multiple accounts**
- **Avoiding scrutiny**
- **Editing logged out to mislead**
- **Misusing a clean start.**
- **"Good hand" and "bad hand" accounts**

Legitimate uses

Valid reasons for an alternative account include:

- **Security**
- **Privacy**
- **Doppelgänger accounts**
- **Clean start under a new name**
- **Username violations**
- **Compromised accounts**
- **Humor accounts**

Wikipedia Sockpuppets



Suspected by volunteers,
confirmed and deleted by administrators

Tsikerdekis et al. (ISFP, 2013), Yamak et al. (WWW 2016), Solorio et al (LASM 2013)

Sockpuppets in social media

- Similar login time
- Similar login IP address
- Similar usernames

Liu et al. (FCS 2016)

- Write similar to each other
- Similar point of view

Bu et al. (KBS 2013)

- Support one another

Zheng et al. (IIH-MSP 2011)

Sockpuppets in online discussions

Data: Sockpuppets

2.9M Users

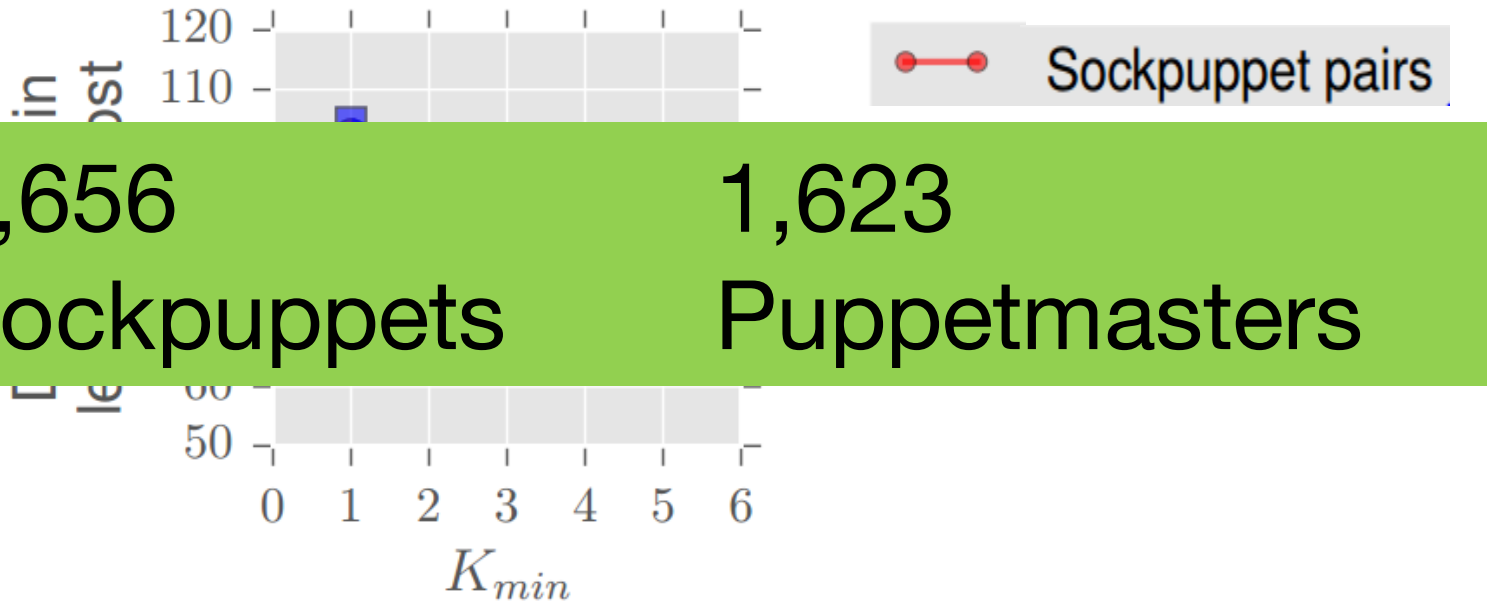
2.1M Articles

62M Posts

Defining sockpuppets

We define sockpuppets as:

Sockpuppets are accounts 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

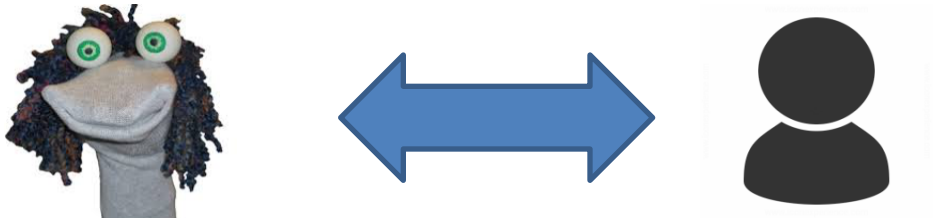
1,623 Puppetmasters

Note: we do not use the IP addresses for detection

Characteristics of sockpuppets

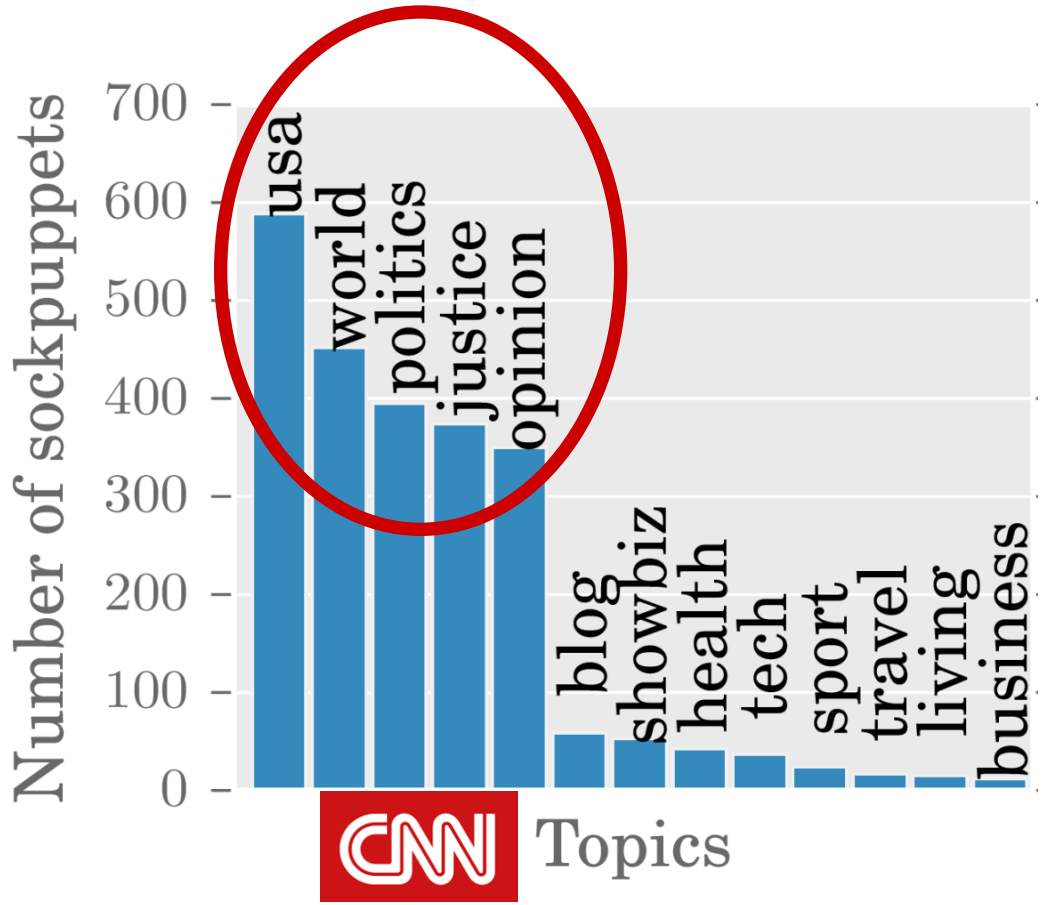
How do we compare sockpuppets and ordinary users?

We have to match!



For each sockpuppet, match an ordinary user that makes
similar number of posts
on
similar discussions

Where do sockpuppets post?



How do sockpuppets write?

jakey008 Feb 5 2013, 2PM

should have read the reviews first :(

ricobeans27

Couldn't agree more!!

Start fewer discussions
$p < 10^{-3}$

Agree more
$p < 10^{-3}$

Falcon-X32 Feb 5 2013, 3PM

I agree. You are absolutely right!

Write shorter sentences
$p < 10^{-3}$

Address others directly
$p < 10^{-3}$

Downvoted more
$p < 10^{-3}$

Write more self-centered posts
$p < 10^{-3}$

Relation between pair of sockpuppets

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

ricobeans27 Feb 5 2013
Couldn't agree more.

Upvote each other more
$p < 10^{-3}$

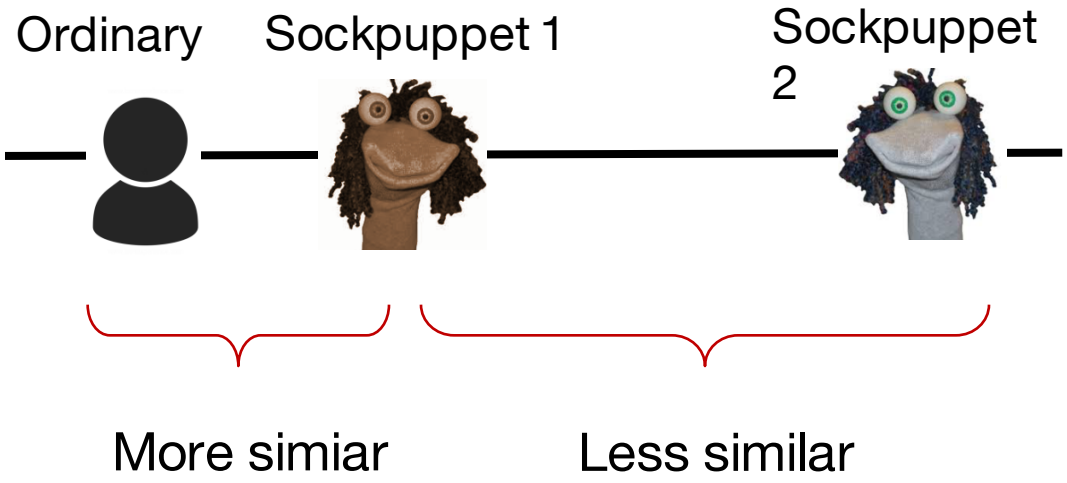
Falcon-X32 Feb 5 2013, 3PM
I agree. You are absolutely right!

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

Interact more with each other
$p < 10^{-3}$

Do puppetmasters lead double lives?

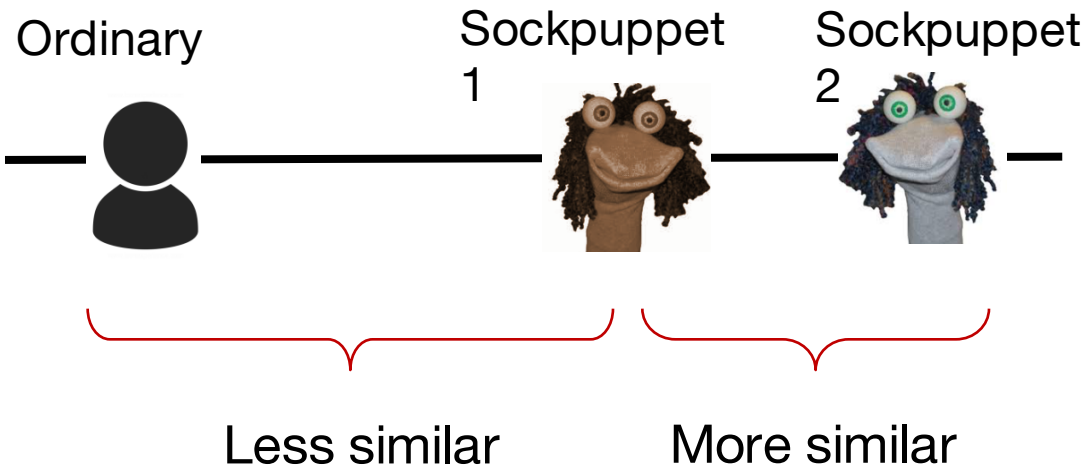
Double life hypothesis:
Puppetmaster maintains distinct personality for the two sockpuppets



Similarity is measured as cosine similarity between user posts' features: LIWC, sentiment, number of words, etc.

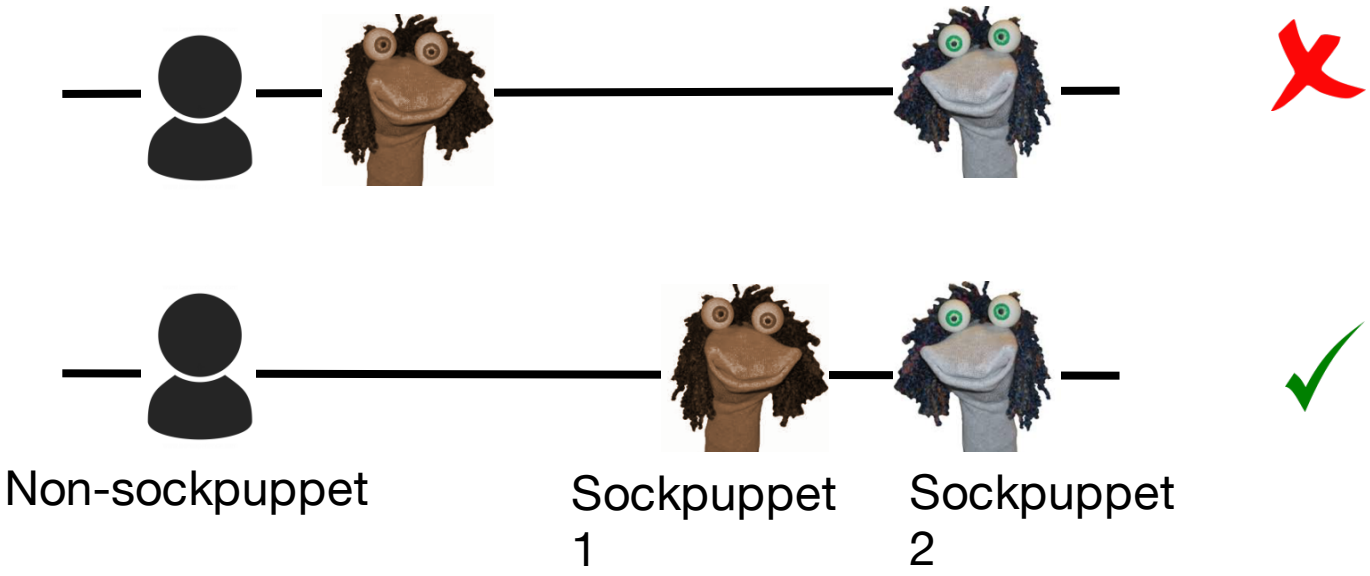
Do puppetmasters lead double lives?

Alternate hypothesis:
Puppetmaster operates both sockpuppets similarly



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

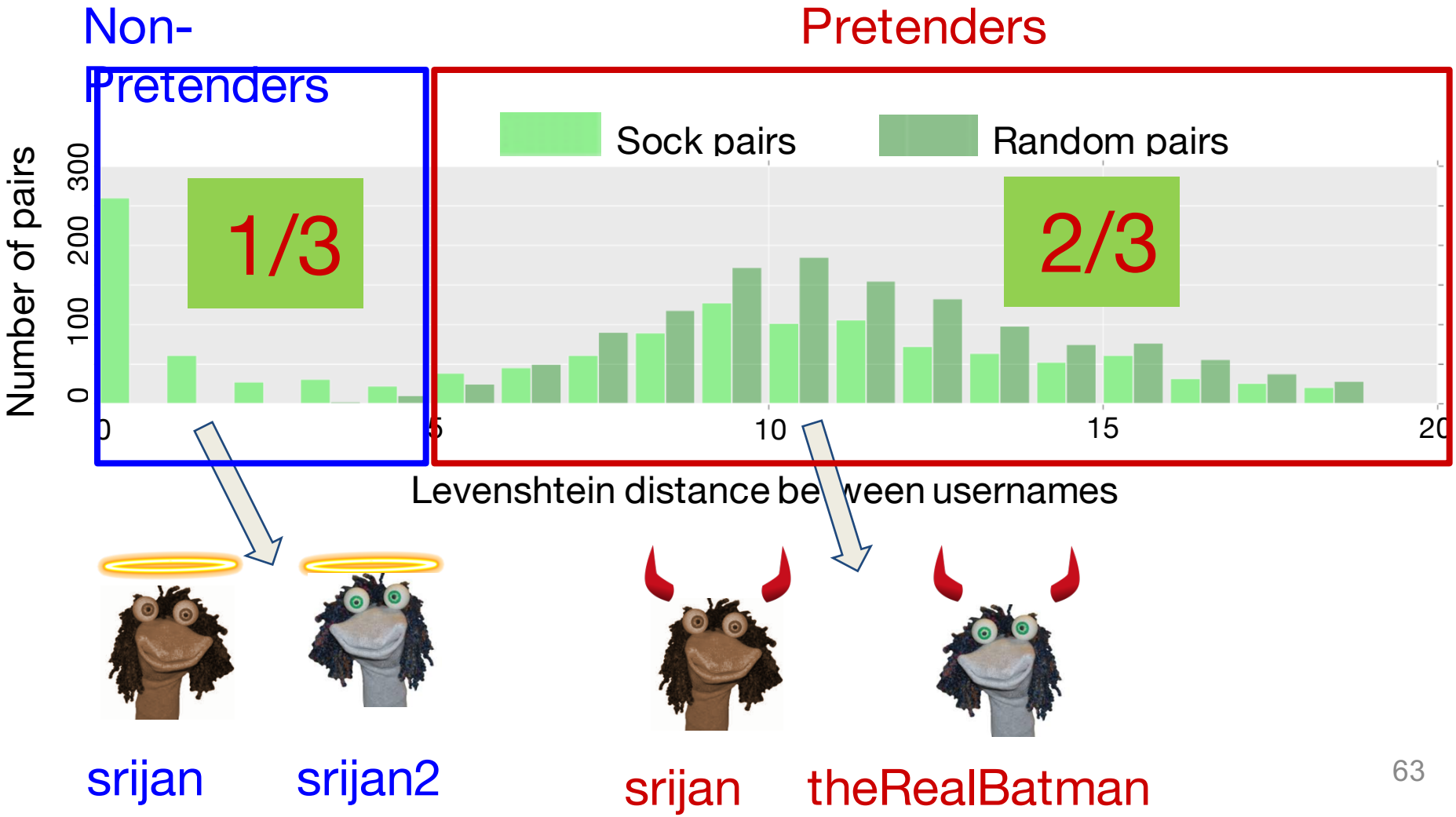
$p < 10^{-3}$

"Good sock/Bad sock" not common

How are sockpuppets used?
Are they used for deception?

Deceptiveness

Hypothesis: Deceptive sockpuppets of the same master have very different usernames.



Pretender vs non-pretender sockpuppets

srijan Feb 5 2013, 2PM

best article i have read!!!

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

ricob 3PM

But this article doesn't make any sense

theRealBatman Feb 5 2013, 3PM

YOU ARE STUPID AND A *****

srijan Feb 5 2013, 3PM

i agree.. these morons dont know a thing

Swear more
$p < 10^{-3}$

Downvoted and reported more
$p < 10^{-3}$

How are sockpuppets used?
Do sockpuppets support one another?

Neutral sockpuppets

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

srijan Feb 5 2013, 3PM

best article ever!

theRealBatman Feb 5 2013, 3PM

why so?

60%

Neutral

Supporter sockpuppets

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

srijan Feb 5 2013, 3PM

best article ever!

theRealBatman Feb 5 2013, 3PM

Totally agree!!

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

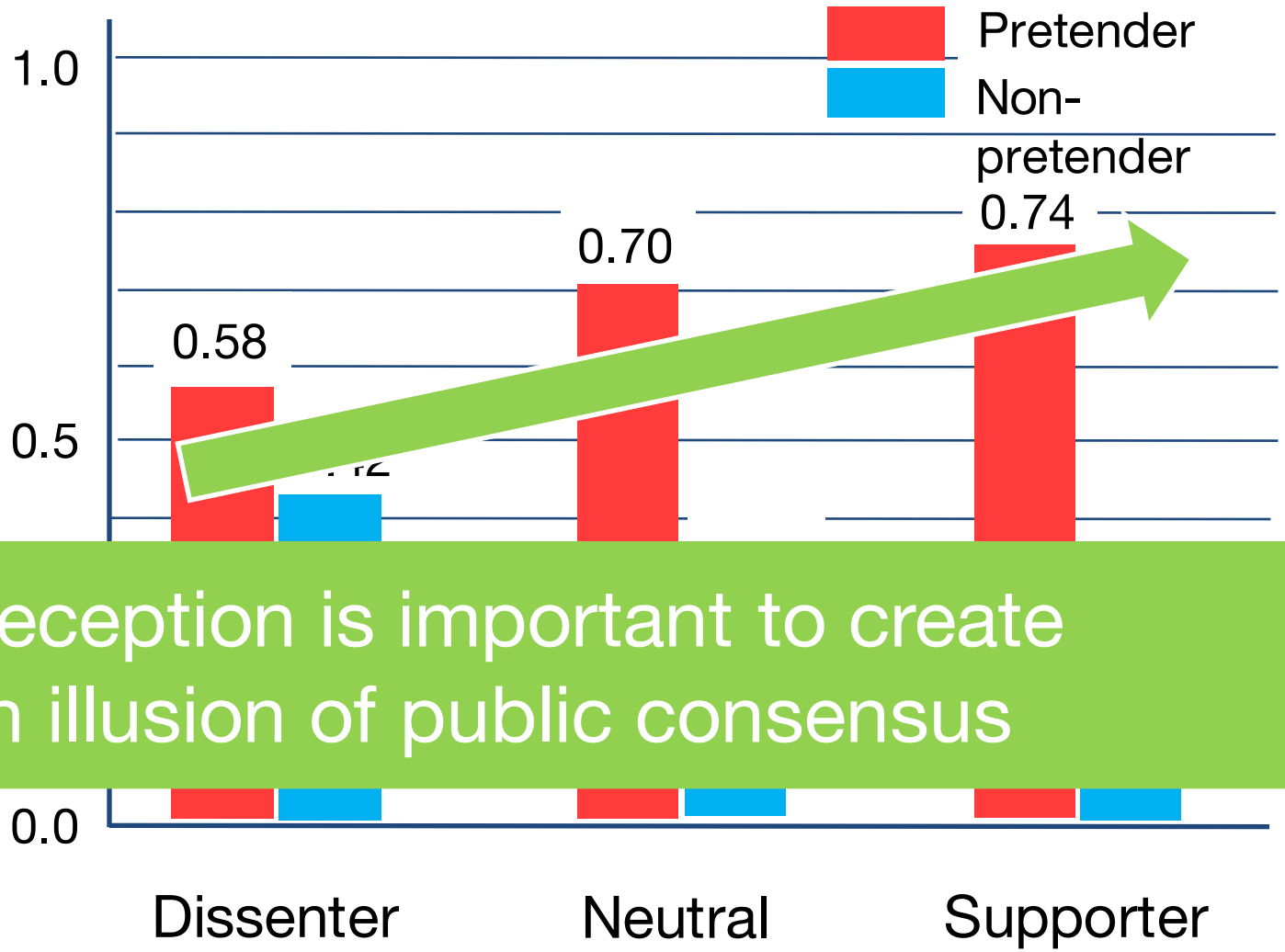
best article ever!

theRealBatman Feb 5 2013, 3PM

I don't think so

60%	30%	10%
Neutral	Supporter	Dissenter

Supportiveness and Deceptiveness



Deception is important to create an illusion of public consensus

Detecting sockpuppets

Features

Activity

Number of posts,
number of replies,
reciprocity of posts,
age of account,
…

Post

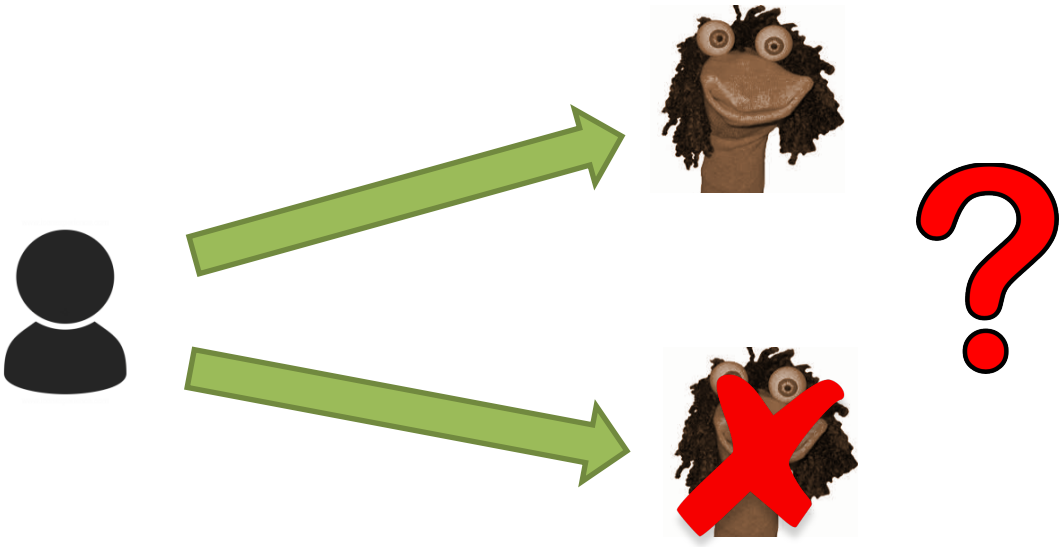
Number of words,
characters, etc.,
LIWC counts,
Readability,
Sentiment,
…

Community

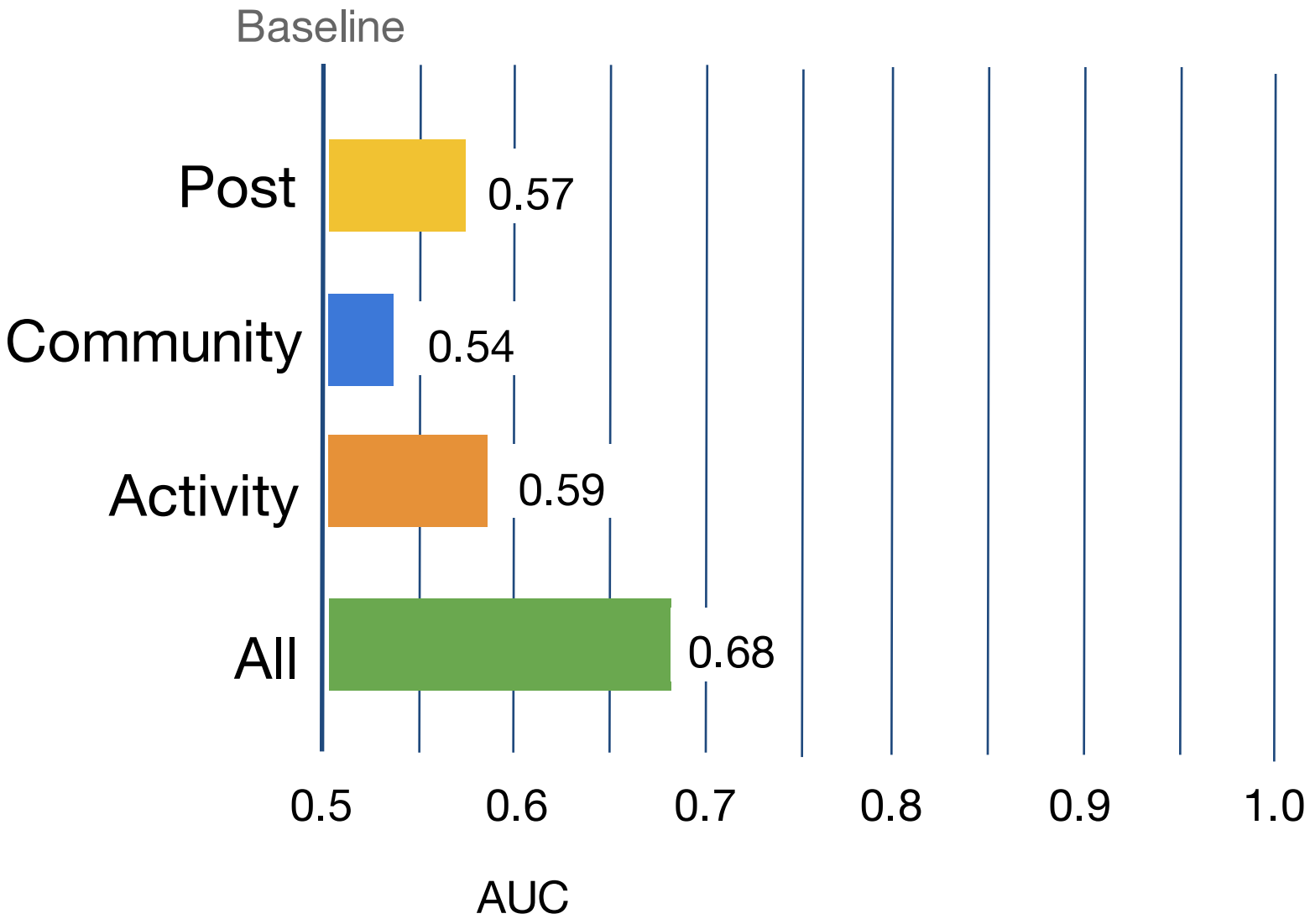
Number of upvotes and
downvotes,
Fraction of reported posts,
Is account reported,
…

Note: we are not using the IP based features

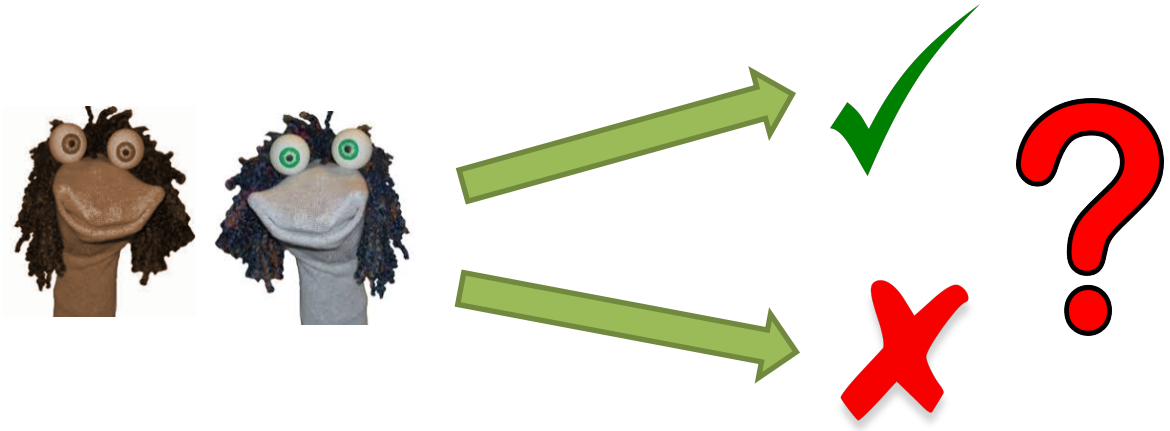
Is an account a sockpuppet?



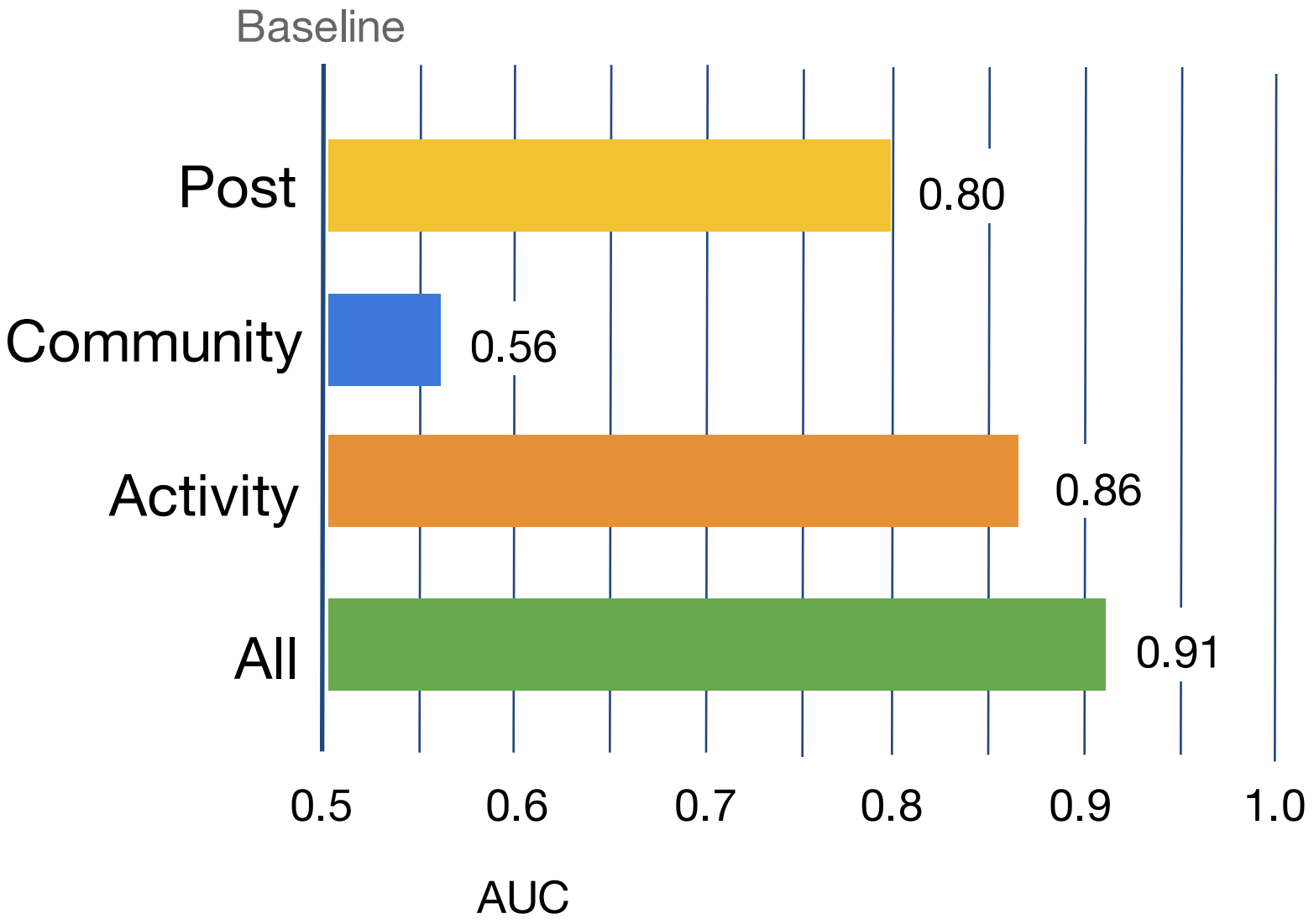
Is an account a sockpuppet?



Do two accounts belong to the same person?



Do two accounts belong to the same person?



Outline

Introduction

Feature-based algorithms

- Bots
- Sockpuppets
- Vandals
- Hoaxes

Spectral-based algorithms

- Visualization: "spokes", "blocks", "staircases"
- Camouflage
- Theoretical guarantee

Density-based algorithms

- Ill-gotten Likes
- Synchronized Behaviors
- Advertising campaigns
- Social spam

Conclusions and future directions

Vandalism

Vandalism is
"an action involving deliberate destruction of or damage to public or private property."

Vandalism is common on Wikipedia

- Freely accessible
- Large reach
- Major source of information for many

WIKIPEDIA

Easy to add content

The free encyclopedia that anyone can edit

Vandalism: An edit that is:

- Non-value adding
- Offensive
- Destructive in removal

Vandalism

Charlie Sheen

From Wikipedia, the free encyclopedia

Charlie Sheen (born September 3, 1965) is half man, half cocaine.

Contents [hide]

1 Early life
2 Career
3 Political views and activities
3.1 Charitable activities
3.2 September 11 attacks
4 Personal life
5 Awards and honors
6 Filmography
6.1 Films
6.2 Short films
6.3 Television
7 References
8 External links

Charlie Sheen

Sheen in March 2009

Born Carlos Irwin Estevez
September 3, 1965 (age 45)
New York City, New York, U.S.

Occupation Actor

More images

Emma Stone

Actress

Emily Jean "Emma" Stone is a hot American actress with a beautiful smile. In 1987, she fell out of the sky as an angel. Wikipedia

Born: November 6, 1988 (age 24), Scottsdale, Arizona, United States
Height: 1.68 m
Siblings: Spencer Stone
Parents: Krista Stone, Jeff Stone
Upcoming movie: The Amazing Spider-Man 2

Movies

The Amazing Spider-Man 2012
The Croods 2013
Gangster Squad 2013
Easy A 2010
Zombieland 2009

~ 7% edits are vandalism
~ 3-4 % editors are vandals

Tools to detect vandalism on Wikipedia

STiki: Metadata

EDITOR

registered?, account-age, geographical location, edit quantity, revert history, block history, is bot?, quantity of warnings on talk page

ARTICLE

age, popularity, length, size change, revert history

REVISION COMMENT

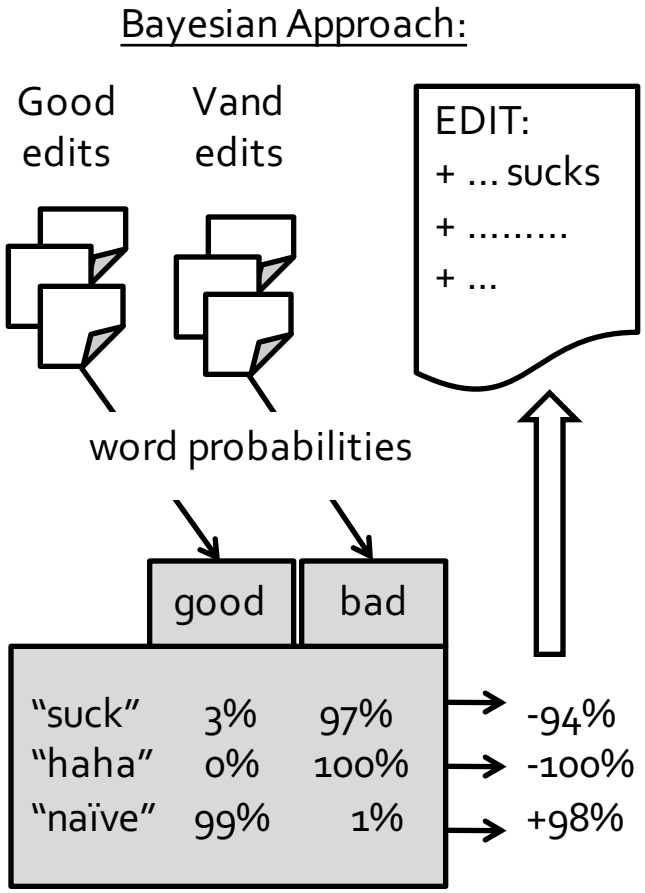
length, section-edit?

TIMESTAMP

time-of-day, day-of-week

West et al. (EuroSec, 2010)

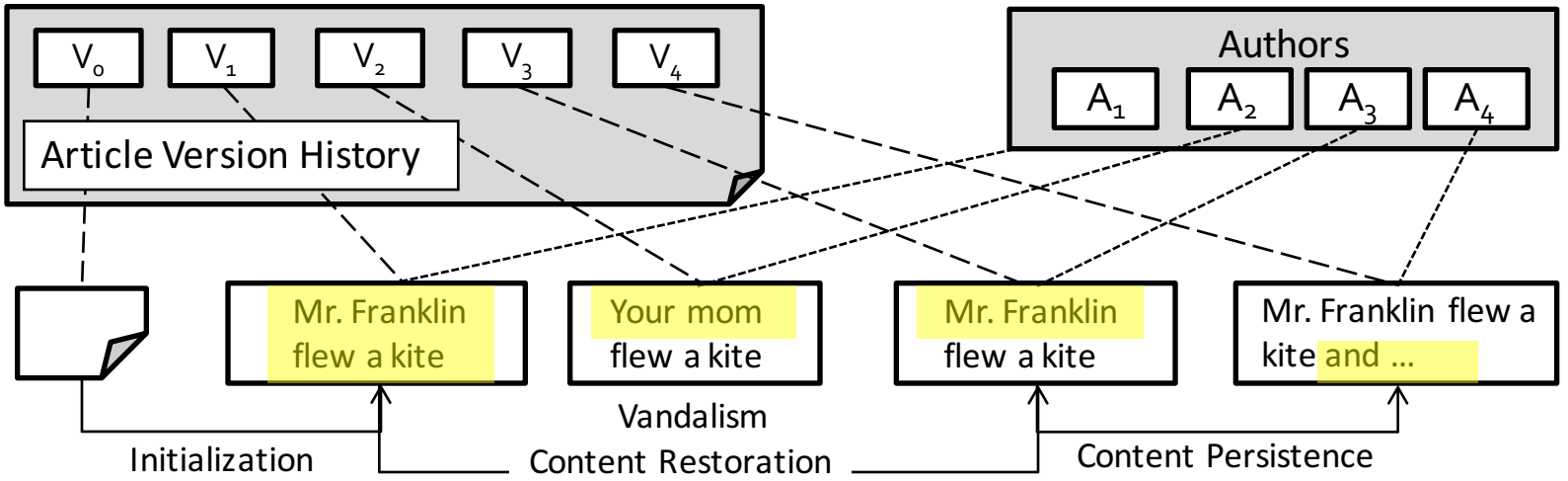
ClueBot NG: Textual



- Vocabularies differ between vandalism and innocent edits
- Automatically assess individual word "goodness" probability

Valesco et al. (CLEF 2010)

WikiTrust: Content driven



- Content that survives is good content
  - Good content builds reputation for its author

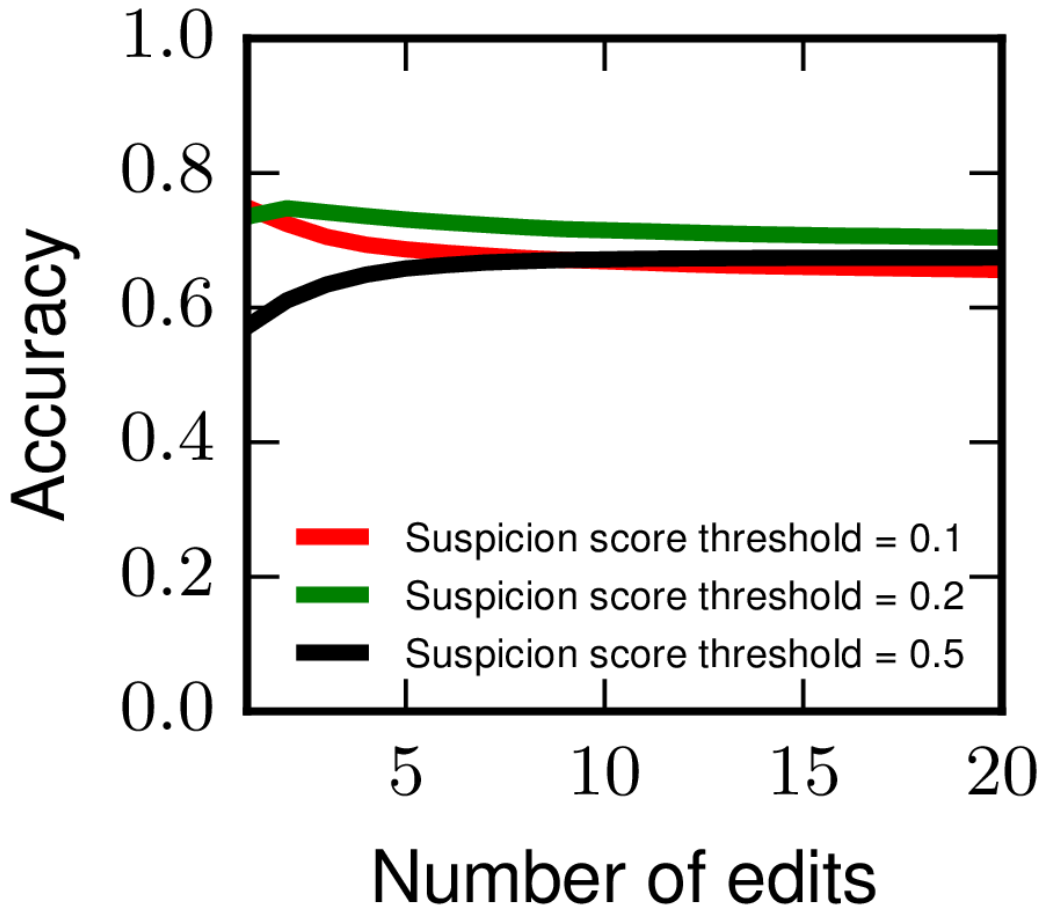
Adler et al. (WWW, 2007)

Detection of vandals

Vandalism detection

Vandal detection

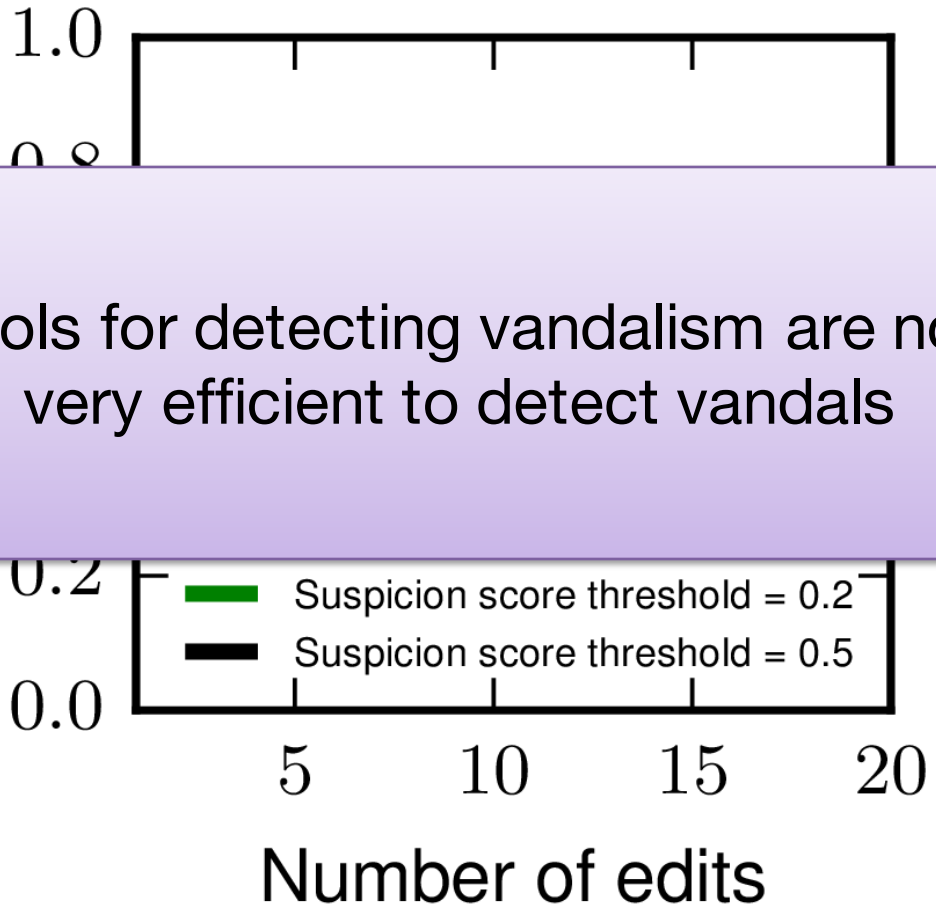
Using STiki to detect vandals



Stiki rule: Editor is a vandal if any edit's suspicion score exceeds threshold

Kumar et al. (KDD, 2015)

Using STiki to detect vandals

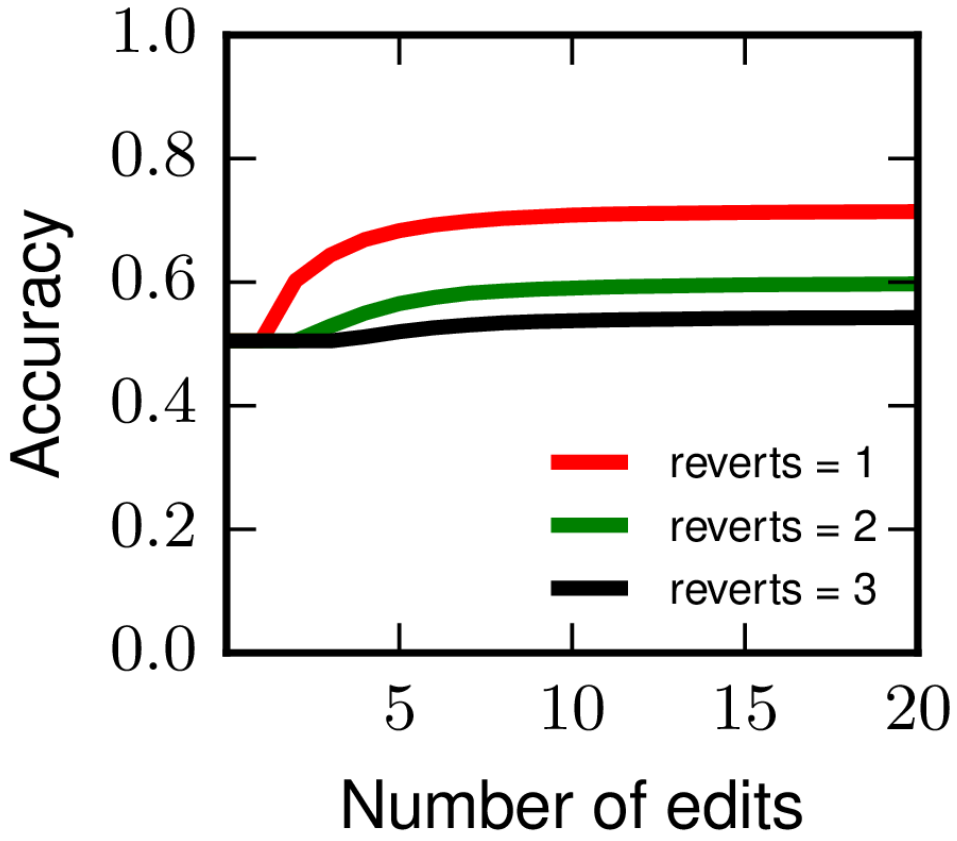


Tools for detecting vandalism are not very efficient to detect vandals

Stiki rule: Editor is a vandal if any edit's suspicion score exceeds threshold

Kumar et al. (KDD, 2015)

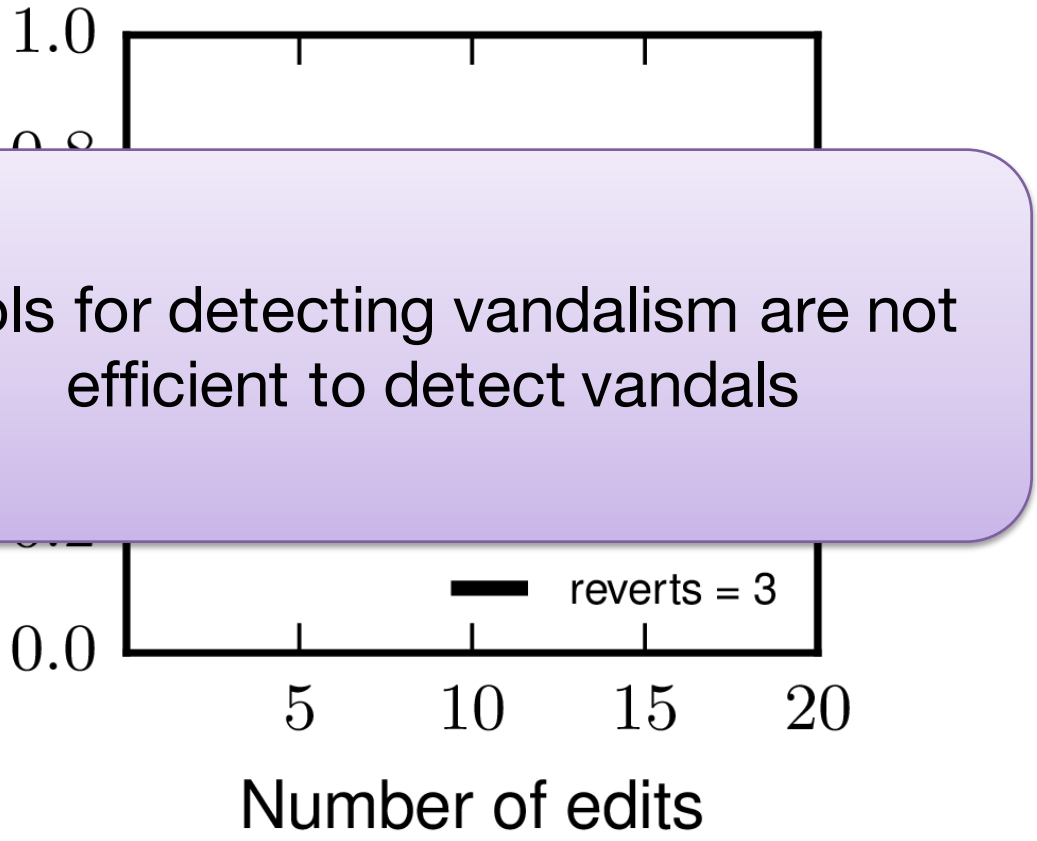
Using ClueBot NG to detect vandals



ClueBot rule: Editor is a vandal if it reverts at least N edits

Kumar et al. (KDD, 2015)

Using ClueBot NG to detect vandals



Tools for detecting vandalism are not efficient to detect vandals

ClueBot rule: Editor is a vandal if it reverts at least N edits

Kumar et al. (KDD, 2015)

Objective:
Detect vandals in as few edits as possible

Data: Wikipedia Vandals

34,000 Editors Half are vandals

770,000 Edits 160,000 edits by vandals

Time: Jan 2013 - July 2014

Kumar et al. (KDD, 2015)

Characteristics of vandals

WIKIPEDIA
The Free Encyclopedia

Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history Search Wikipedia

Perth

From Wikipedia, the free encyclopedia

Coordinates: 31°57′8″S 115°51′32″E

This article is about the capital of Western Australia. For the city in Scotland, see Perth, Scotland. For other uses, see Perth (disambiguation).

Perth (/ˈpɜːrθ/) is the capital and largest city of the Australian state of Western Australia. It is the fourth-most populous city in Australia, with an estimated population of 2.06 million (as of 30 June 2016) living in

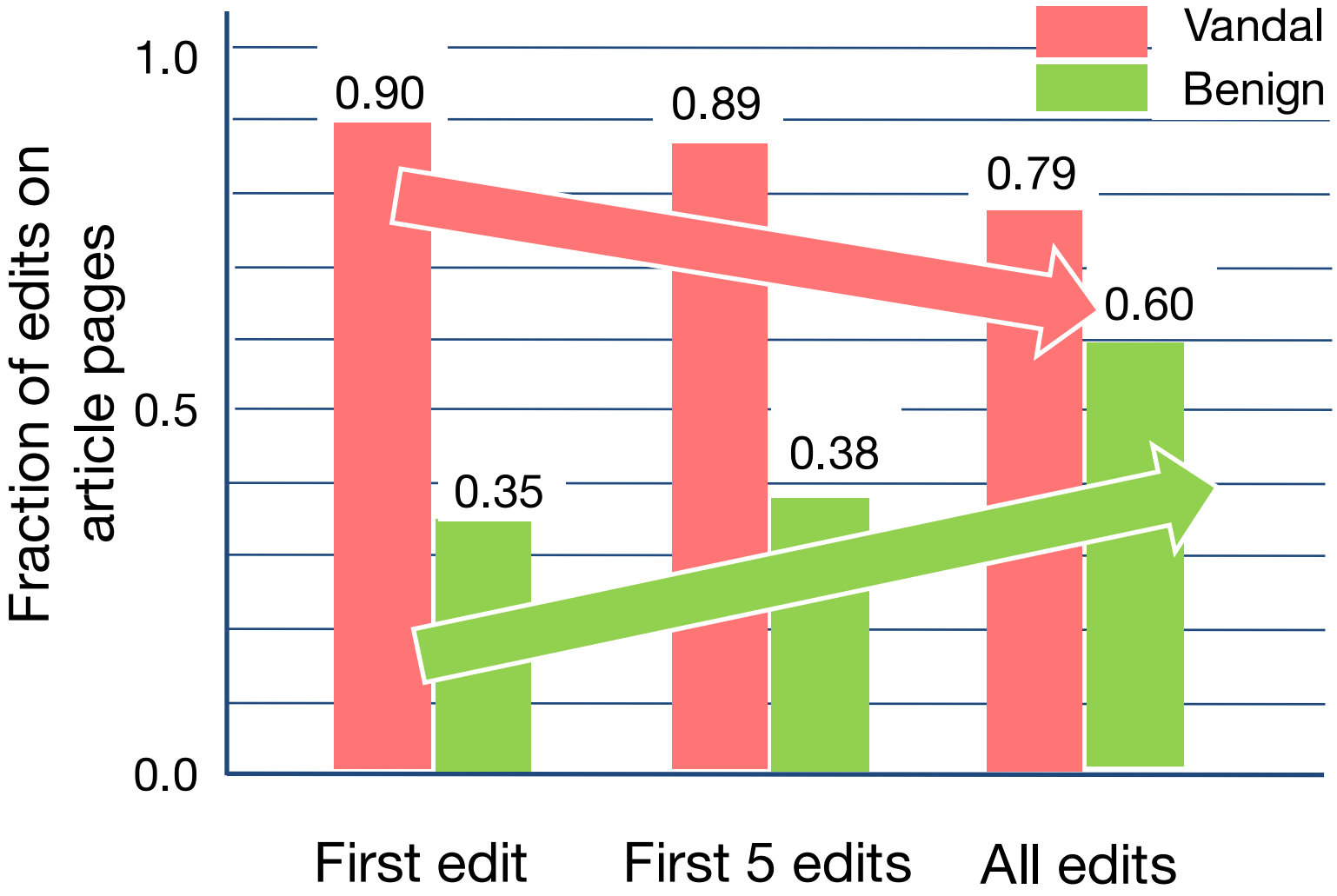
Perth
Western Australia

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

Interaction

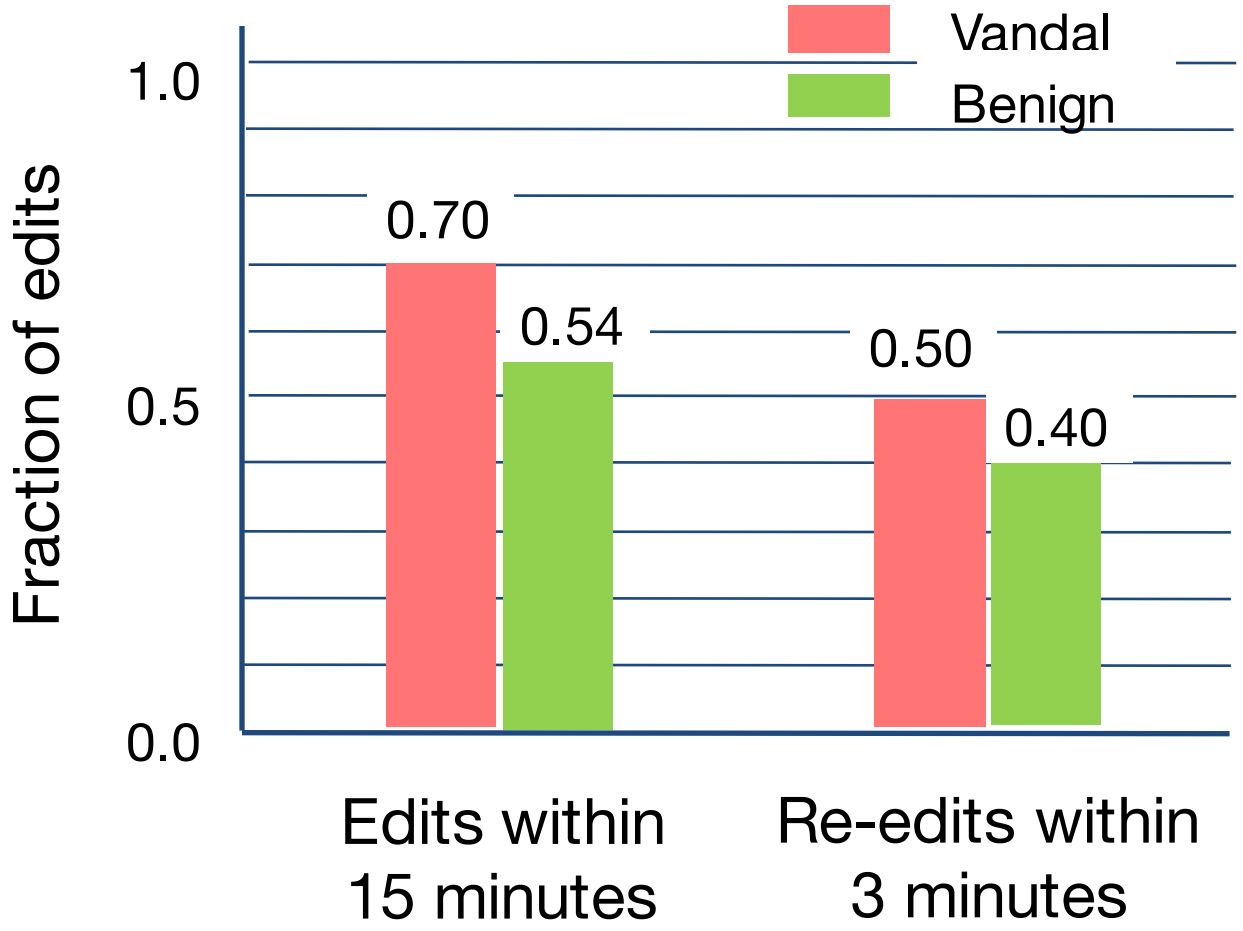
Editors can edit article pages and talk pages

Vandals make visible edits



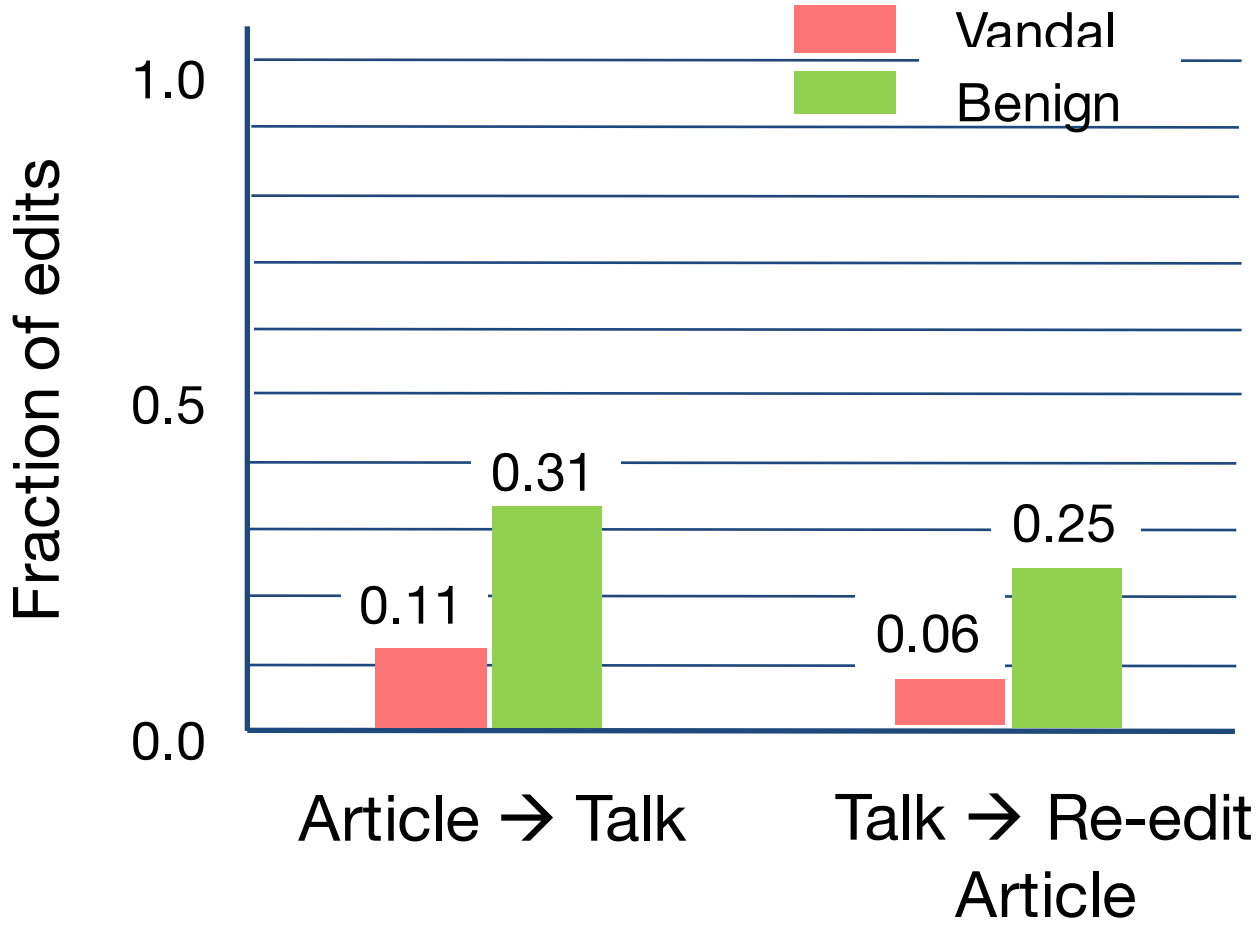
Kumar et al. (KDD, 2015)

Vandals are quicker



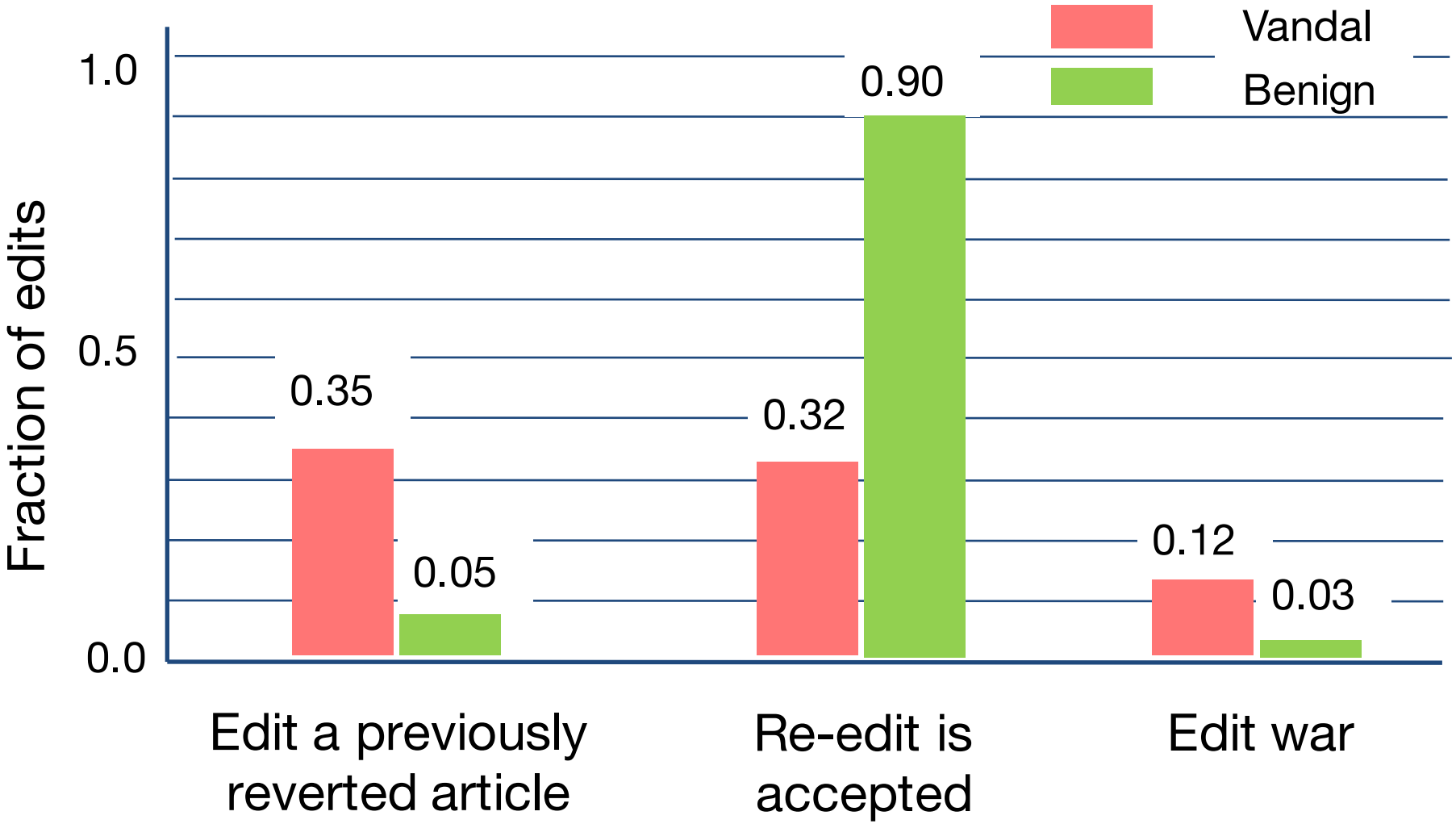
Kumar et al. (KDD, 2015)

Vandals do not discuss



Kumar et al. (KDD, 2015)

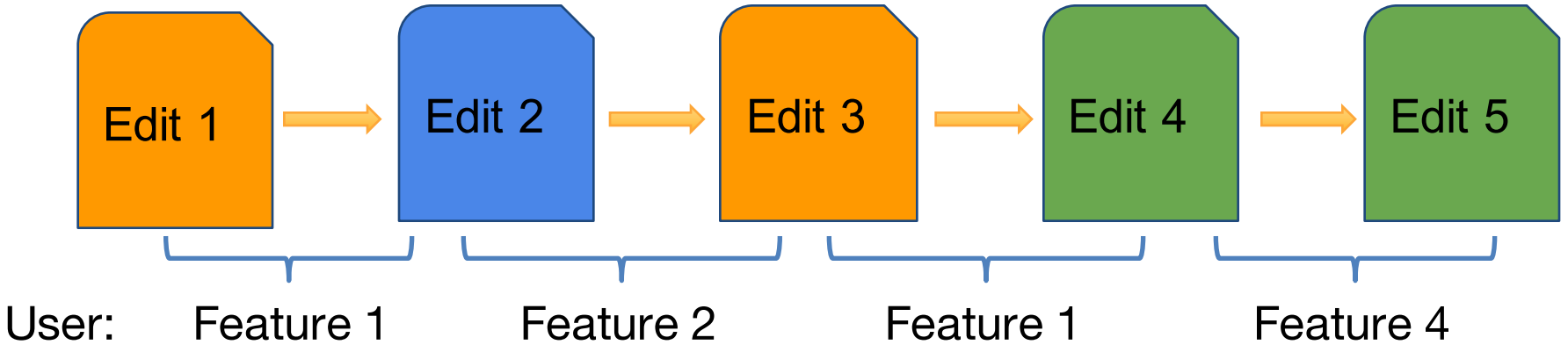
Vandals make reversion driven edits



Kumar et al. (KDD, 2015)

Detecting vandals

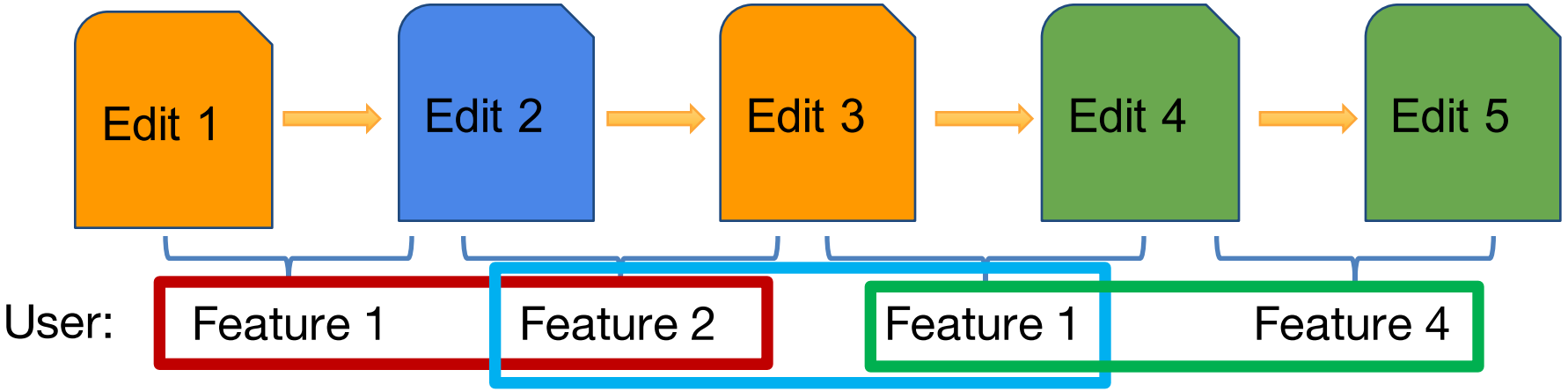
Pairwise Edit Features



Time x Type of page x First edit x Distance x Similarity x Reverted or not

Kumar et al. (KDD, 2015)

Meta-Features: Transitions

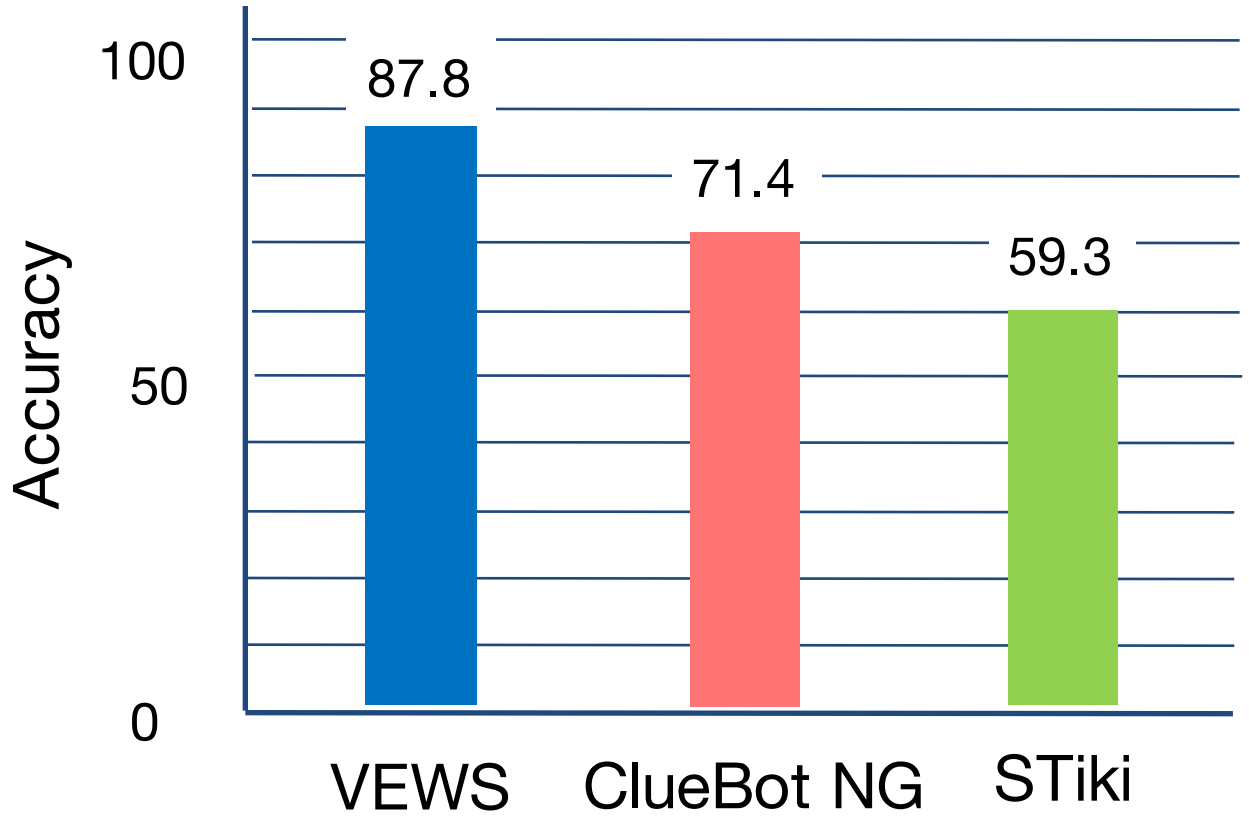


	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5
Feature 1		1		1	
Feature 2	1				
Feature 3					
Feature 4					
Feature 5					

N x N

Kumar et al. (KDD, 2015)

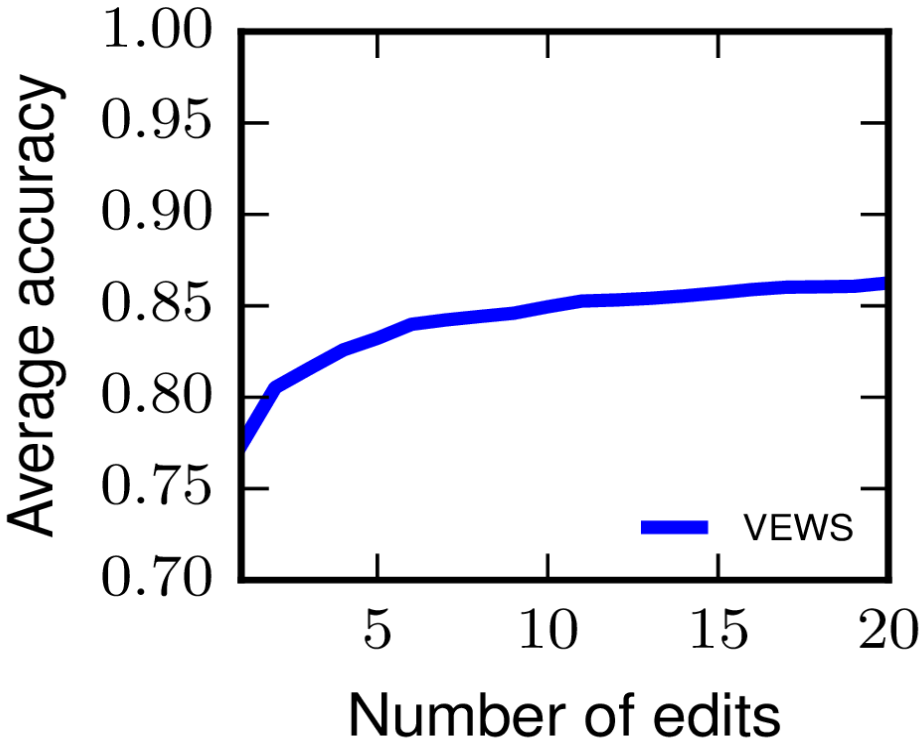
Vandal Detection



VEWS identifies 87% vandals on or before first reversion.
44% vandal are identified before first reversion.

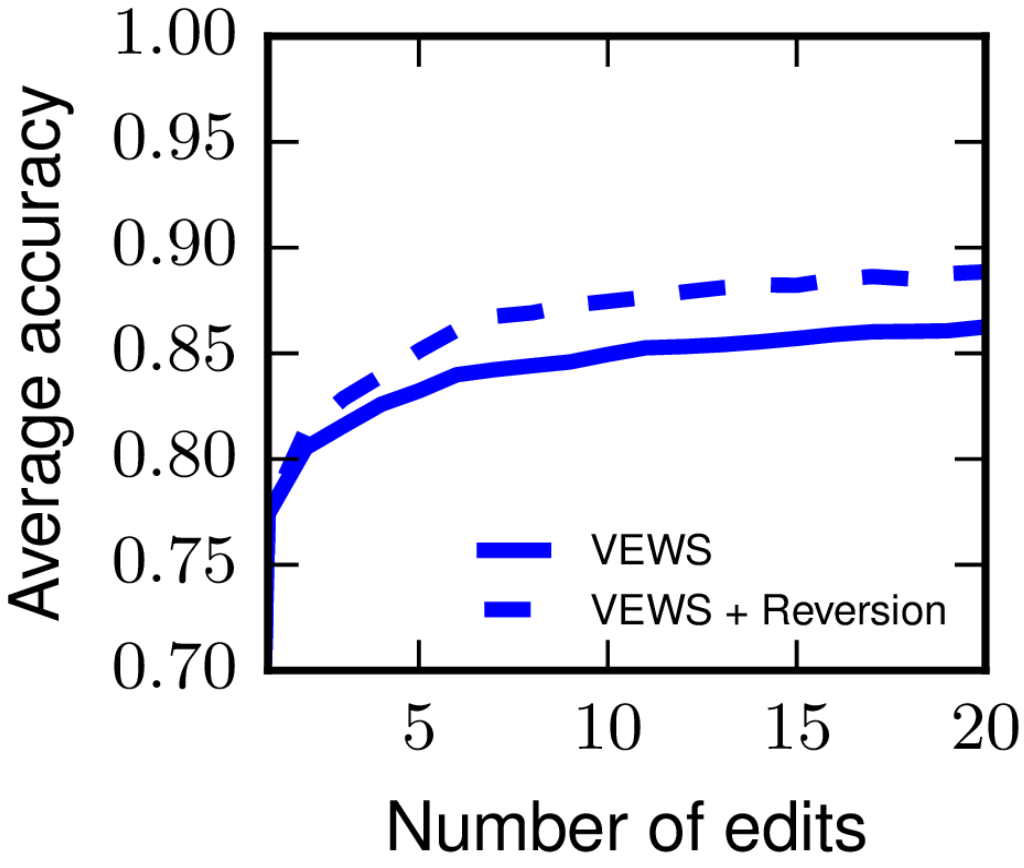
Kumar et al. (KDD, 2015)

Early Warning System

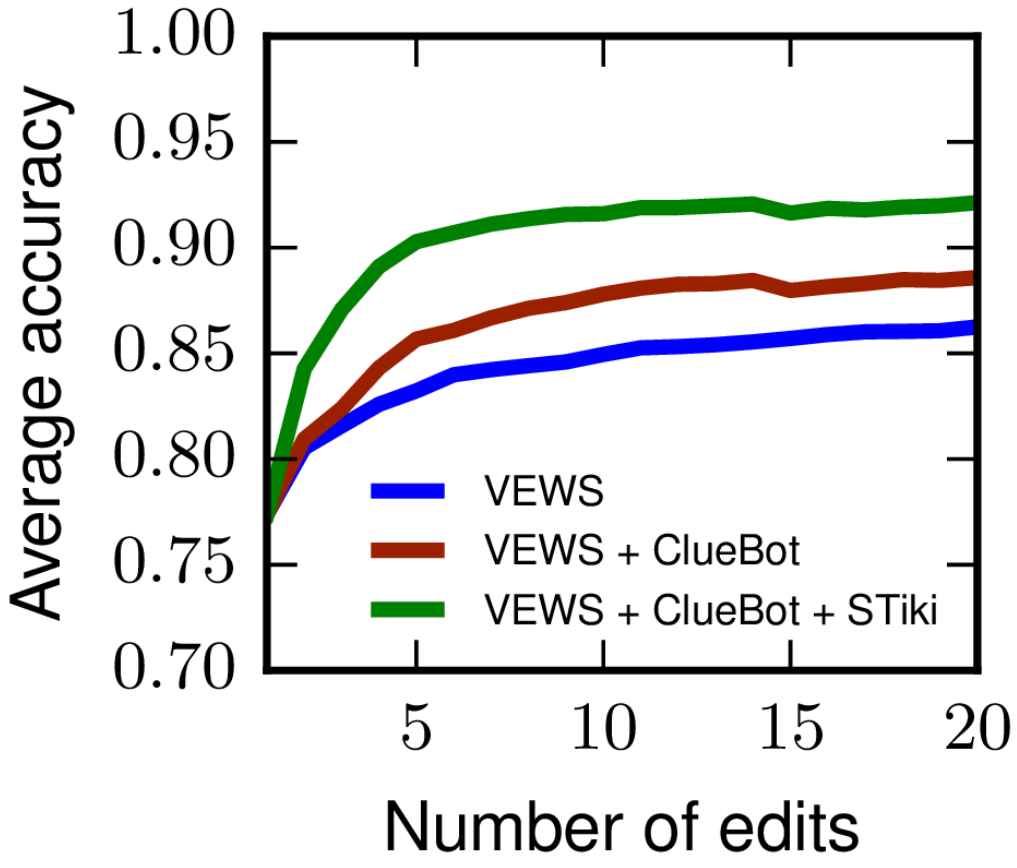


VEWS identifies vandals in 2.13 edits on average

Does reversion information help?



Combining Multiple Systems



Summary: Vandals

- **Vandals:** Users that make non-constructive contribution
- Vandals are aggressive: they make visible edits without discussing and edit war
- Vandals can be detected early by using temporal features and relation between edited pages
- Combination of metadata, text and human feedback is the best in detecting vandals

Outline

Introduction

Feature-based algorithms

Bots | Sockpuppets

Vandals | Hoaxes

Spectral-based algorithms

Visualization: "spokes", "blocks", "staircases"

Camouflage | Theoretical guarantee

Density-based algorithms

Ill-gotten Likes | Synchronized Behaviors

Advertising campaigns | Social spam

Conclusions and future directions

Types of false information

Misinformation
honest mistake

Disinformation
deliberate lie to mislead

Wikipedia defines "hoax" as "deliberately fabricated falsehood made to masquerade as truth"

http://abcnews.com.co/obama-signs-executive-order-banning-national-anthem/

abc NEWS

Obama Signs Executive Order Banning The National Anthem At All Sporting Events Nationwide

By ***Jimmy Rustling, ABC News*** - November 11, 2016 1421 4

Snopes.com
Rumor Has It

Articles tagged: Fake News

158 Total

Fact Check › Politics

U.S. Attorney General Jeff Sessions Disbarred for Misconduct?

Mar 29th, 2017 - Hyperpartisan web sites spread the false claim that Attorney General Jeff Sessions will be disbarred thanks to a letter of complaint signed by 2,000 U.S. lawyers.

News › Political News

IJR Staffers Suspended for Promoting Obama Conspiracy Theory

Mar 21st, 2017 - Independent Journal Review suspended three staff members, including chief content officer Benny Johnson, for suggesting that Obama may have interfered in a Hawaii judge's ruling on the Trump travel ban.

Fact Check › Fake News

Nancy Pelosi Was Just Taken from Her Office in Handcuffs?

Mar 11th, 2017 - Reports that the House Minority Leader was taken from her office in handcuffs for plotting to overthrow the president are fake news.

Fact Check › Fake News

Shepard Smith Fired from Fox News?

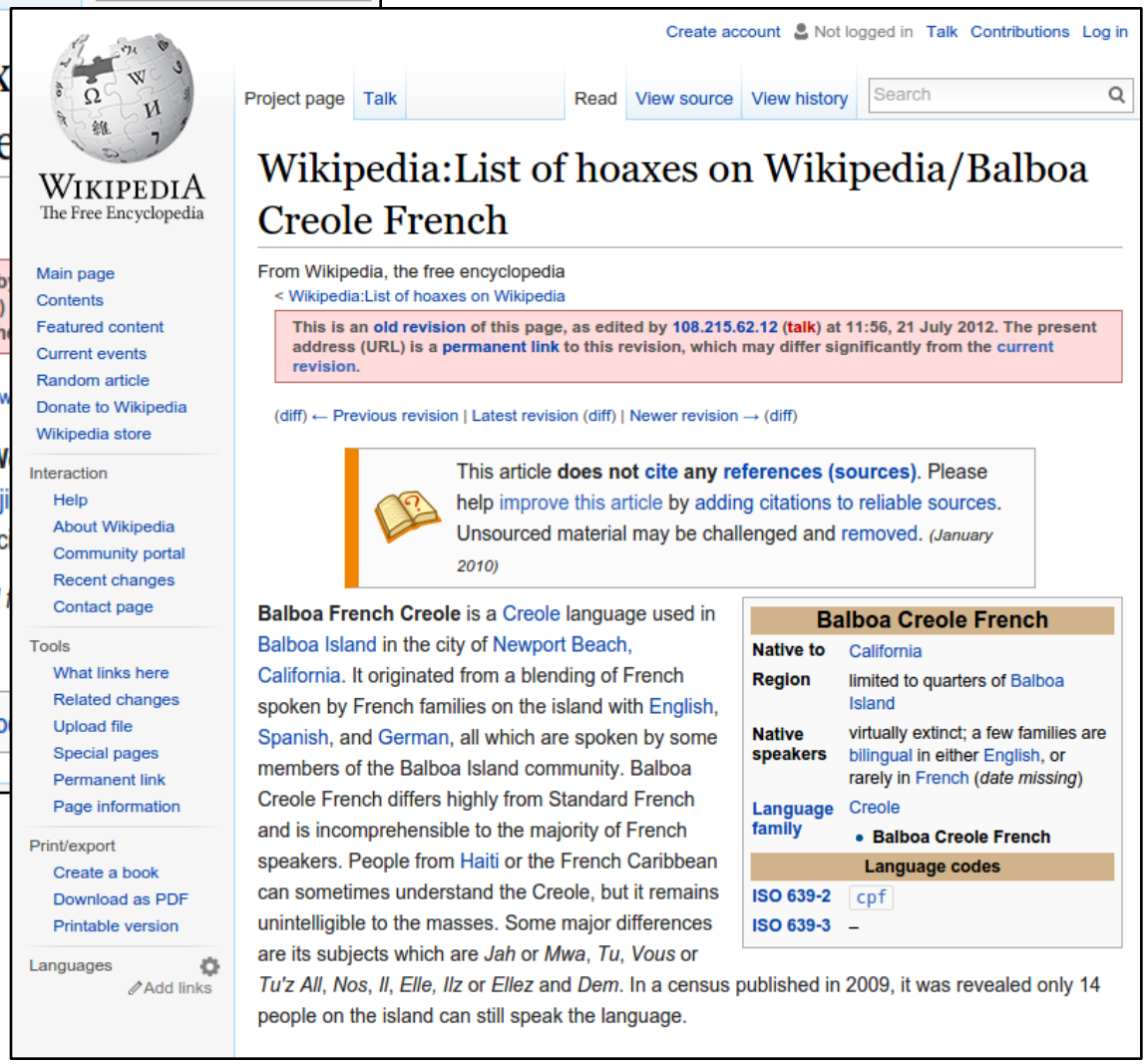
Mar 9th, 2017 - Hoax outlets reported that Fox News anchor Shep Smith has been fired by chairman Rupert Murdoch for being "too controversial."

Fact Check › Fake News

Did Betsy DeVos Say History Textbooks Should Be Based on the Bible?

Mar 8th, 2017 - A report stating that Trump's Secretary of Education wants to exclude all information not found in the Bible from history textbooks is satire, not fact.

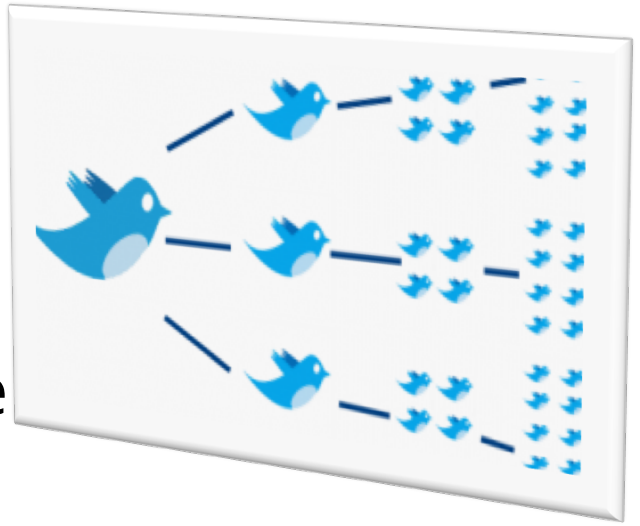
Hoaxes on Wikipedia



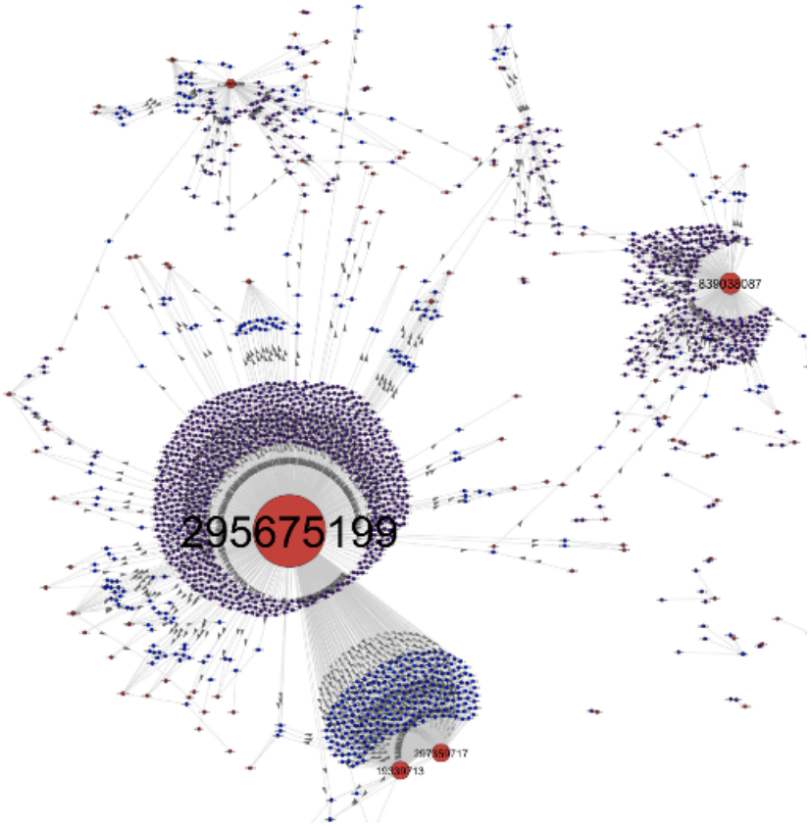
Properties of disinformation

False Information Goes "Viral" Online

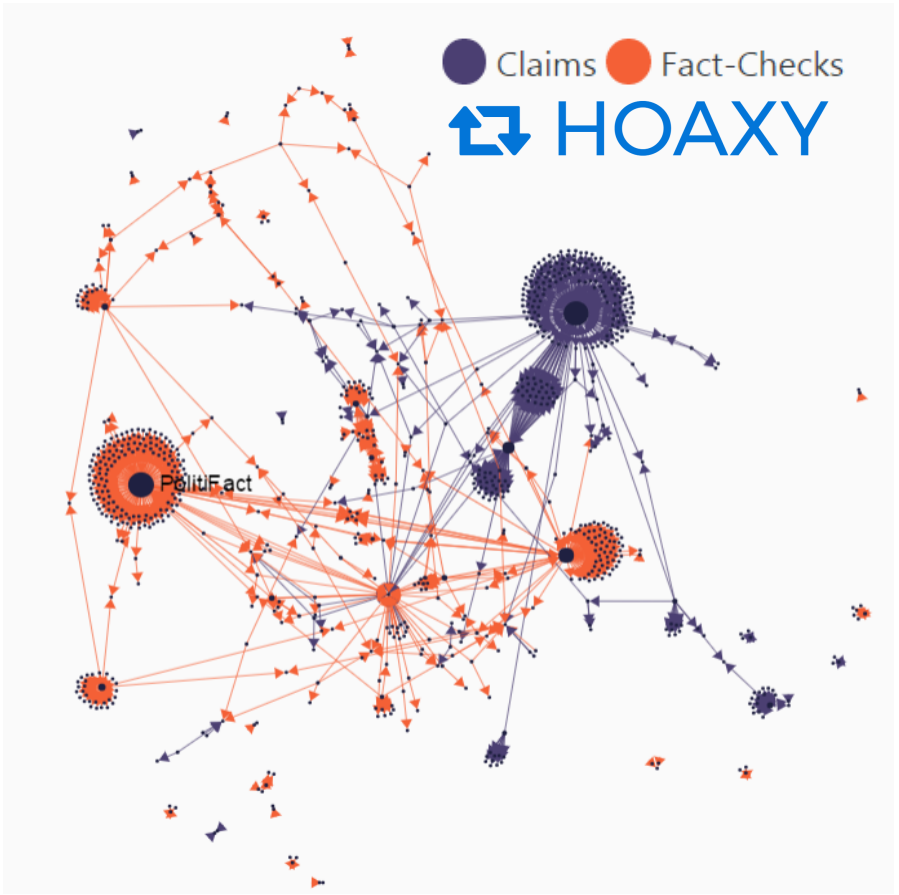
- Many social media users "retweet", "share", and "like" these erroneous reports.
- These users include average citizens who don't fact-check before spreading the news.
- Examples about how hoaxes spread.



False information spreads quickly



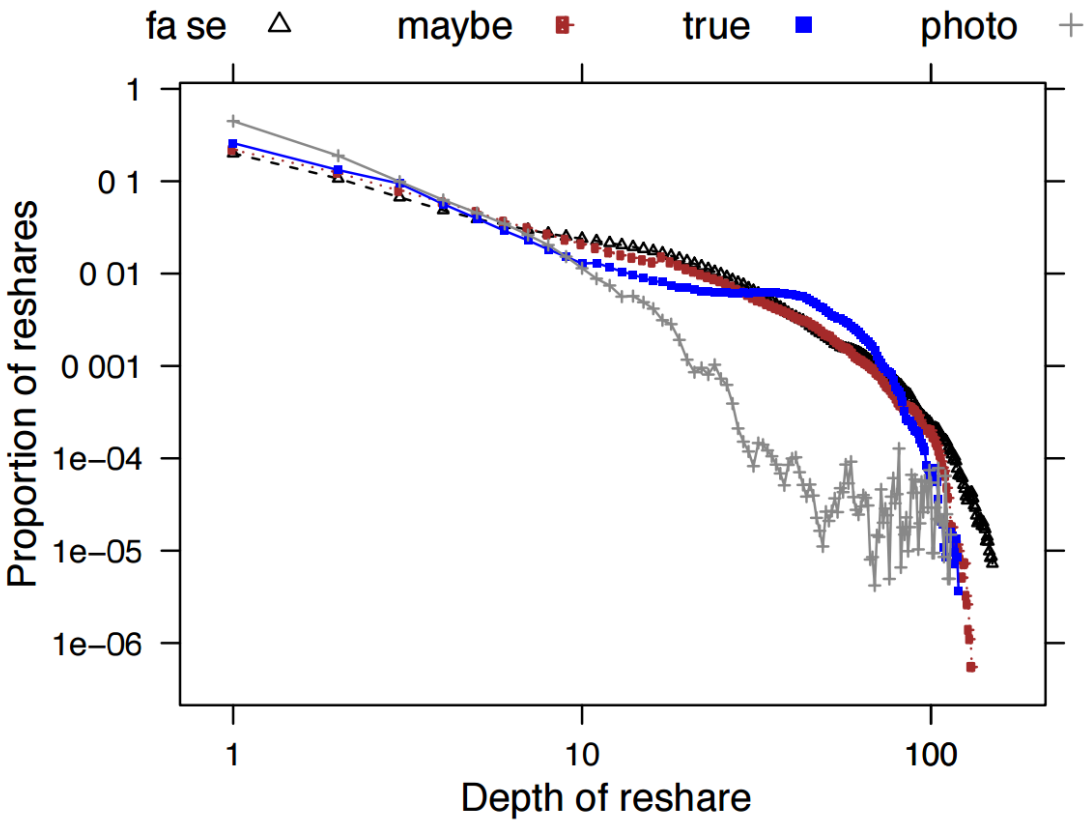
Tweets and retweets for spread of a fake image during first 2 hours



Tweets and retweets by users on claims and fact checks on a topic

Gupta, et al. (WWW 2013), Shao, et al. (WWW 2016)

False information cascades deep



Rumor cascades tend to be deeper, in that more reshares are at greater depths, than the reference cascades.

Frigerri, et al. (ICWSM 2014)

Which of these news is false?

BREAKING BOMBSHELL: NYPD Blows Whistle on New Hillary Emails: Money Laundering, Sex Crimes with Children, Child Exploitation, Pay to Play, Perjury

Preexisting Conditions and Republican Plans to Replace Obamacare

Horne, et al. (ICWSM 2017)

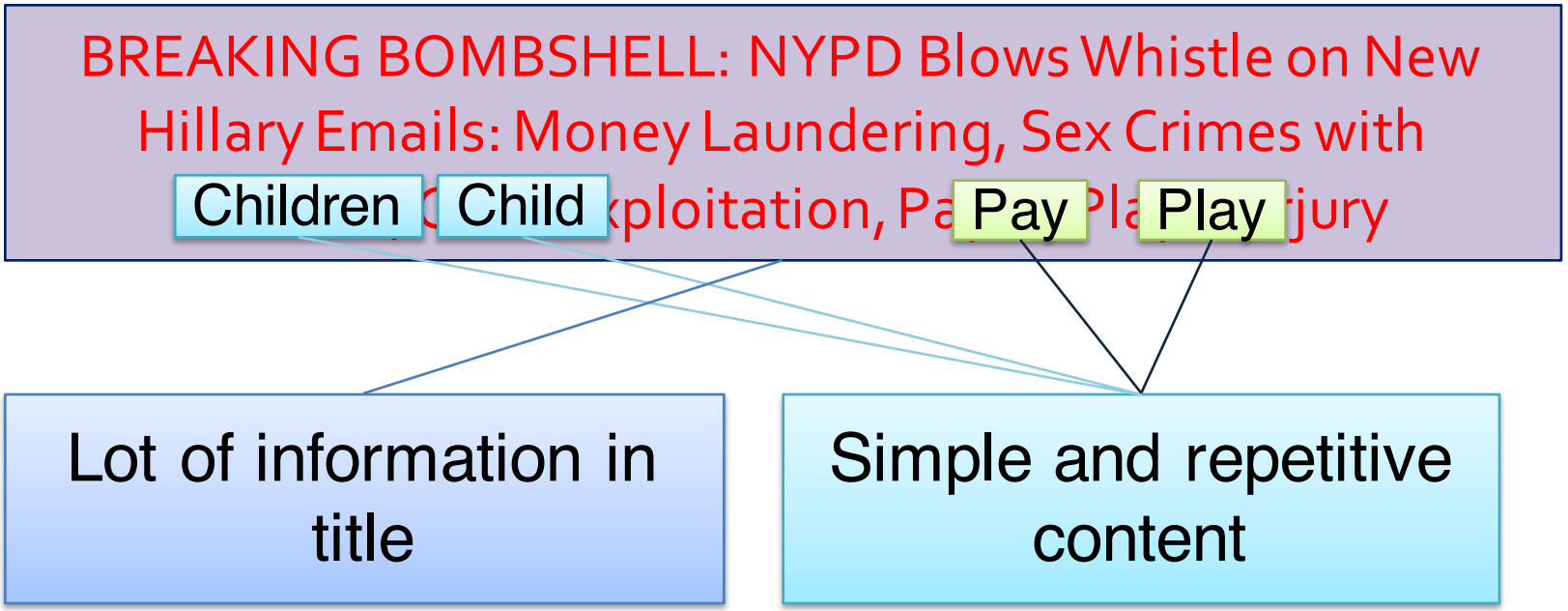
Which of these news is false?

BREAKING BOMBSHELL: NYPD Blows Whistle on New Hillary Emails: Money Laundering, Sex Crimes with Children, Child Exploitation, Pay to Play, Perjury

Preexisting Conditions and Republican Plans to Replace Obamacare

Horne, et al. (ICWSM 2017)

How is fake news written?



Horne, et al. (ICWSM 2017)

Case study:
Disinformation on Wikipedia

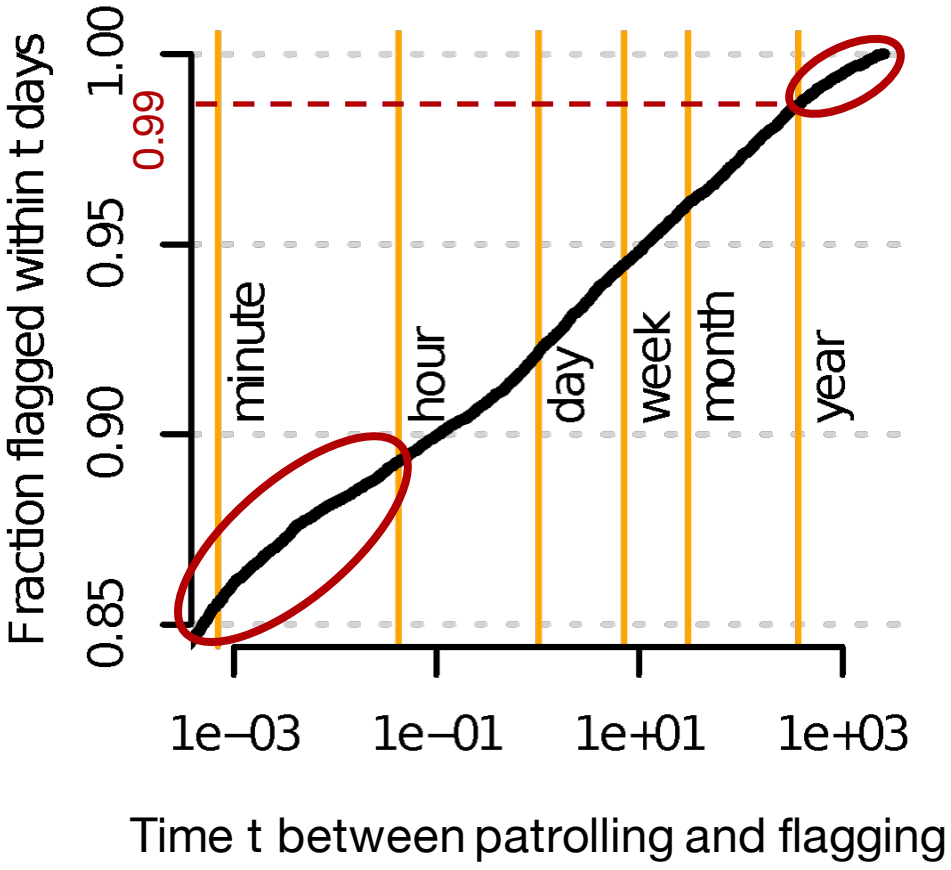
Impact of Wikipedia hoaxes

"The worst hoaxes are those which
(a) last for a long time,
(b) receive significant traffic,
(c) are relied upon by credible news media."
Jimmy Wales on Quora

Kumar, et al. (WWW 2016)

Impact of Wikipedia hoaxes

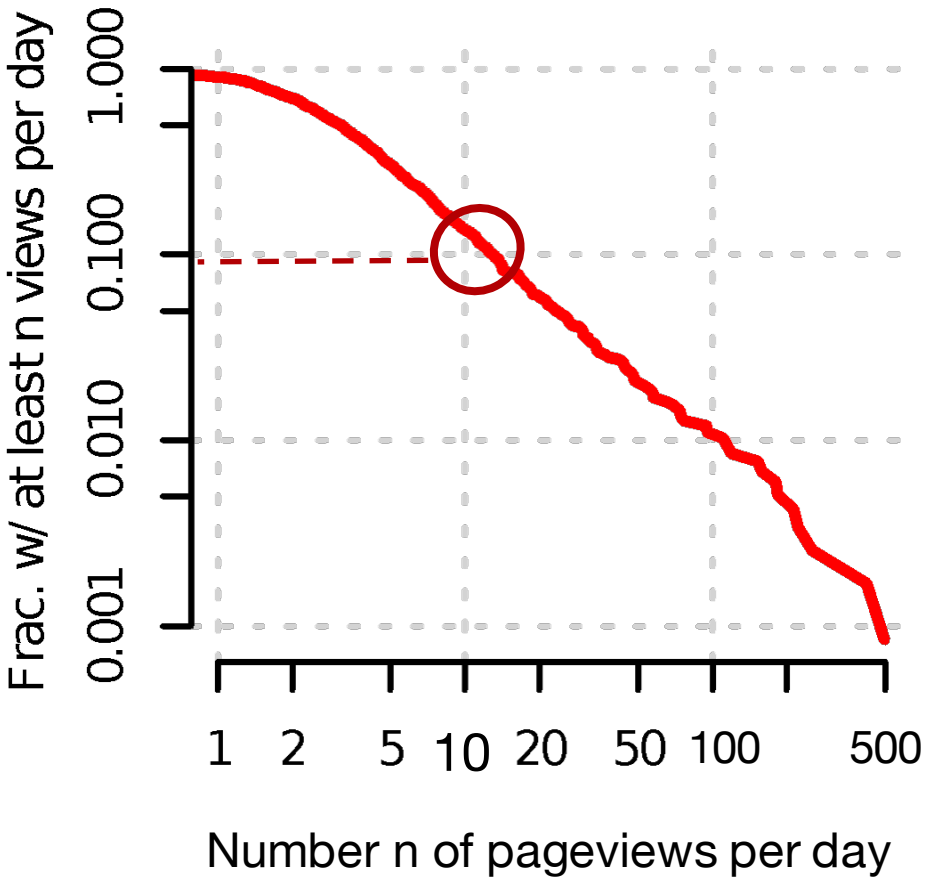
"The worst hoaxes are those which
(a) last for a long time"



Kumar, et al. (WWW 2016)

Impact of Wikipedia hoaxes

"The worst hoaxes are those which
(b) receive significant traffic"



Kumar, et al. (WWW 2016)

Impact of Wikipedia hoaxes

"The worst hoaxes are those which
(c) are relied upon by credible news media"

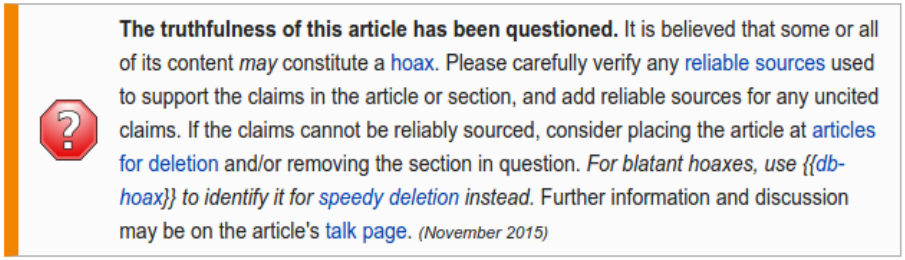
1.08
active inlinks
from entire web

Kumar, et al. (WWW 2016)

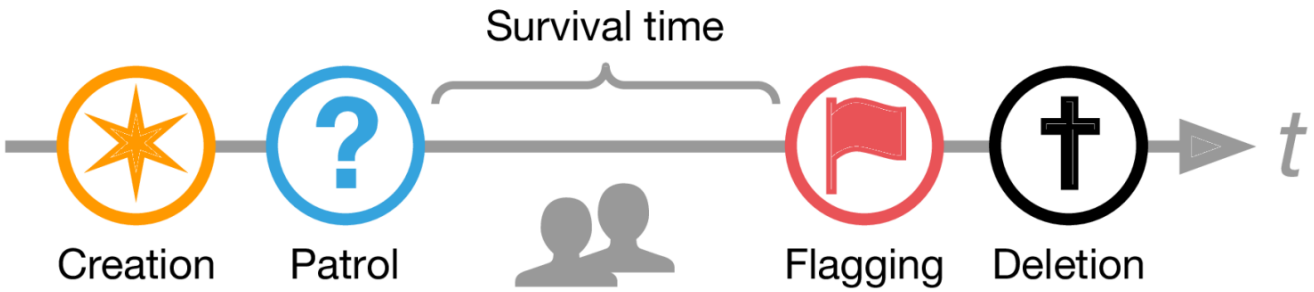
Wikipedia Hoaxes

Hoax article vs hoax facts

21,218 hoax articles


The truthfulness of this article has been questioned. It is believed that some or all of its content *may* constitute a hoax. Please carefully verify any reliable sources used to support the claims in the article or section, and add reliable sources for any uncited claims. If the claims cannot be reliably sourced, consider placing the article at articles for deletion and/or removing the section in question. *For blatant hoaxes, use {{db-hoax}} to identify it for speedy deletion instead.* Further information and discussion may be on the article's talk page. *(November 2015)*

Hoax lifecycle:



Data: http://cs.umd.edu/~srijan/hoax/

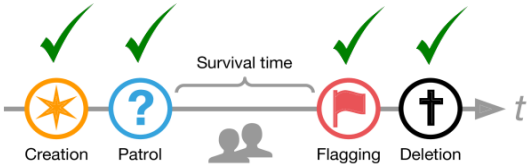
Kumar, et al. (WWW 2016)

What are Wikipedia hoaxes like?

Hoax

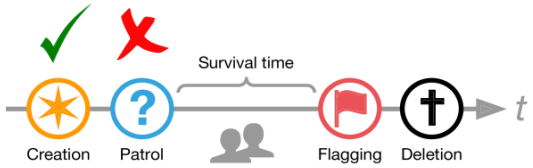
Successful hoax

pass patrol
survive for a month
viewed frequently



Failed hoax

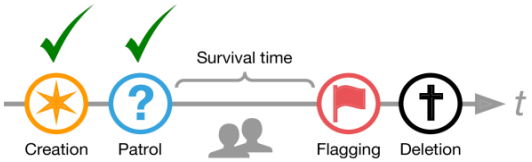
flagged and deleted during patrol



Non-hoax

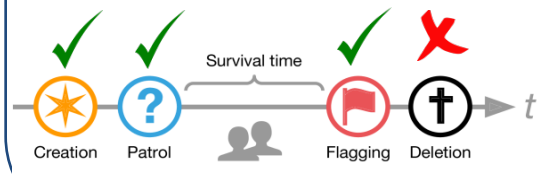
Legitimate articles

never flagged



Wrongly flagged

temporarily flagged



Kumar, et al. (WWW 2016)

Characteristics of Wikipedia hoaxes

Appearance:
how the article looks

Link-network:
how the article connects

Support:
how other articles refer to it

Editor:
how the article creator looks

Kumar, et al. (WWW 2016)

Characteristics of Wikipedia hoaxes

Appearance:
how the article looks

Link-network:
how the article connects

Support:
how other articles refer to it

Editor:
how the article creator looks

Features:

- Plain-text length
- Plain-text-to-markup ratio
- Wiki-link density
- Web-link density

Hoax articles are longer, but they mostly have plain text and have lesser web and wiki links.

Kumar, et al. (WWW 2016)

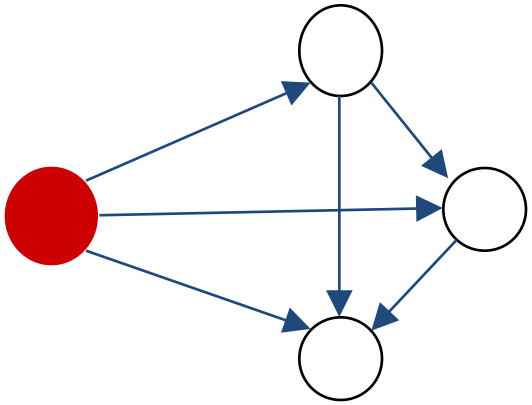
Characteristics of Wikipedia hoaxes

Appearance:
hoaxes mostly have text and few references.

Link-network:
how the article connects

Support:
how other articles refer to it

Editor:
how the article creator looks



CC = 0
incoherent article

CC > 0
coherent article

Legitimate articles are more coherent than successful hoaxes

Kumar, et al. (WWW 2016)

Characteristics of Wikipedia hoaxes

Appearance:
hoaxes mostly have text and few references.

Link-network:
hoaxes have incoherent wikilinks.

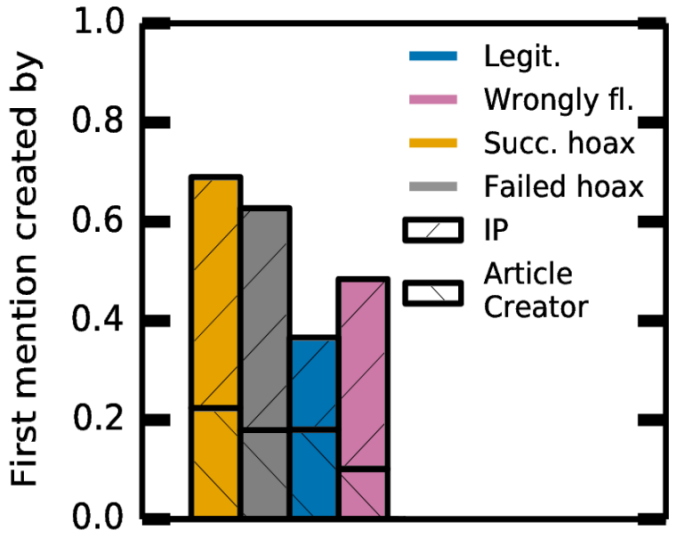
Support:
how other articles refer to it

Editor:
how the article creator looks

Features:

- Number of prior mentions
- Time since first mention
- Creator of first mention

Hoax mentions are less in number, more recently created, and mostly created by IP addresses or article creator



Kumar, et al. (WWW 2016)

Characteristics of Wikipedia hoaxes

Appearance:
hoaxes mostly have text and few references.

Link-network:
hoaxes have incoherent wikilinks.

Support:
hoaxes have few, recent, suspicious mentions.

Editor:
how the article creator looks

Features:

- Creator's age
- Creator's experience

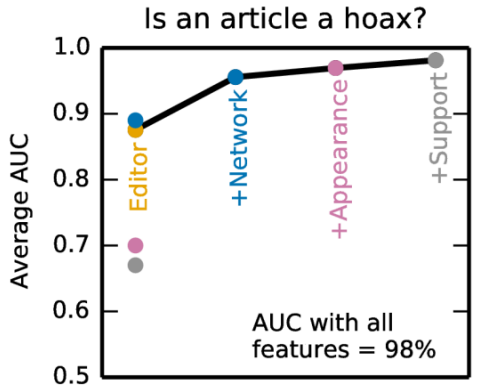
Hoax creators are more recently registered, and
have lesser editing experience.

Kumar, et al. (WWW 2016)

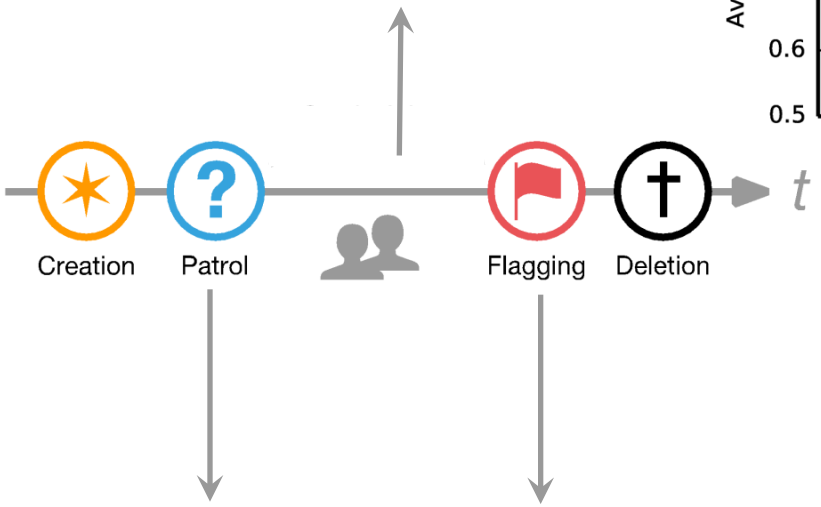
Detection of disinformation

Detecting Wikipedia hoaxes

AUC = 98%
Editor and
Network features



Is an article a hoax?



AUC = 71%
Appearance
features

Will a hoax get past patrol?

Is an article flagged as hoax really one?

AUC = 86%
Editor and
support features

Kumar, et al. (WWW 2016)

Identifying real Wikipedia hoaxes

Flagged by us and deleted by Wikipedia administrators

Steve Moertel

American popcorn entrepreneur

Survived for 6 years 11 months!

Kumar, et al. (WWW 2016)

Detecting False Tweets

75% NDCG score of prediction

Linguistic: swear words, emotion words, "I", "my", pronouns, etc.

Author: number of followers, friends

Tweet network: number of retweets, mentions, reply? retweet?

Time: time since author registration, time since tweet

Gupta, et al. (SocInfo 2014)

Can humans identify fake information?

HOW TO SPOT FAKE NEWS

CONSIDER THE SOURCE

Click away from the story to investigate the site, its mission and its contact info.

READ BEYOND

Headlines can be outrageous in an effort to get clicks. What's the whole story?

CHECK THE AUTHOR

Do a quick search on the author. Are they credible? Are they real?

SUPPORTING SOURCES?

Click on those links. Determine if the info given actually supports the story.

CHECK THE DATE

Reposting old news stories doesn't mean they're relevant to current events.

IS IT A JOKE?

If it is too outlandish, it might be satire. Research the site and author to be sure.

CHECK YOUR BIASES

Consider if your own beliefs could affect your judgement.

ASK THE EXPERTS

Ask a librarian, or consult a fact-checking site.

IFLA
International Federation of Library Associations and Institutions

Can readers identify Wikipedia hoaxes?

320 random hoax and non-hoax pairs
10 raters on Amazon Mechanical Turk rated each pair

Results

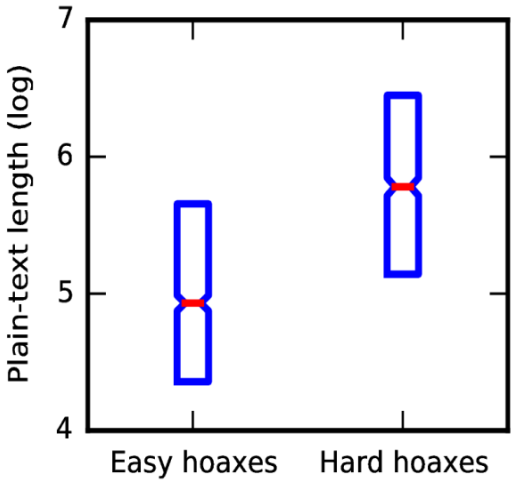
50%	66%	86%
Random	Human	Classifier

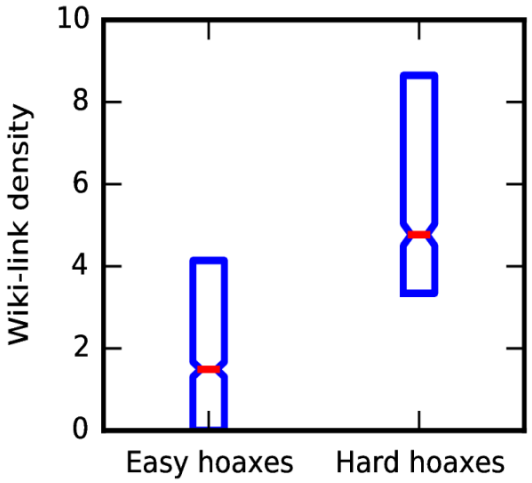
Casual readers are gullible to hoaxes.
Accurate detection needs non-appearance features.

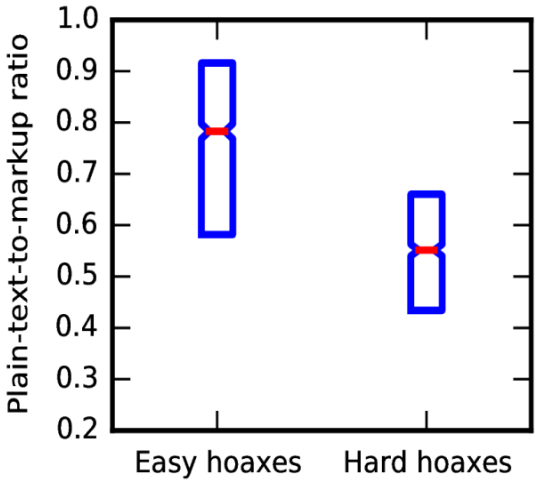
Kumar, et al. (WWW 2016)

What fools humans?

Comparing easy- vs hard-to-identify hoaxes



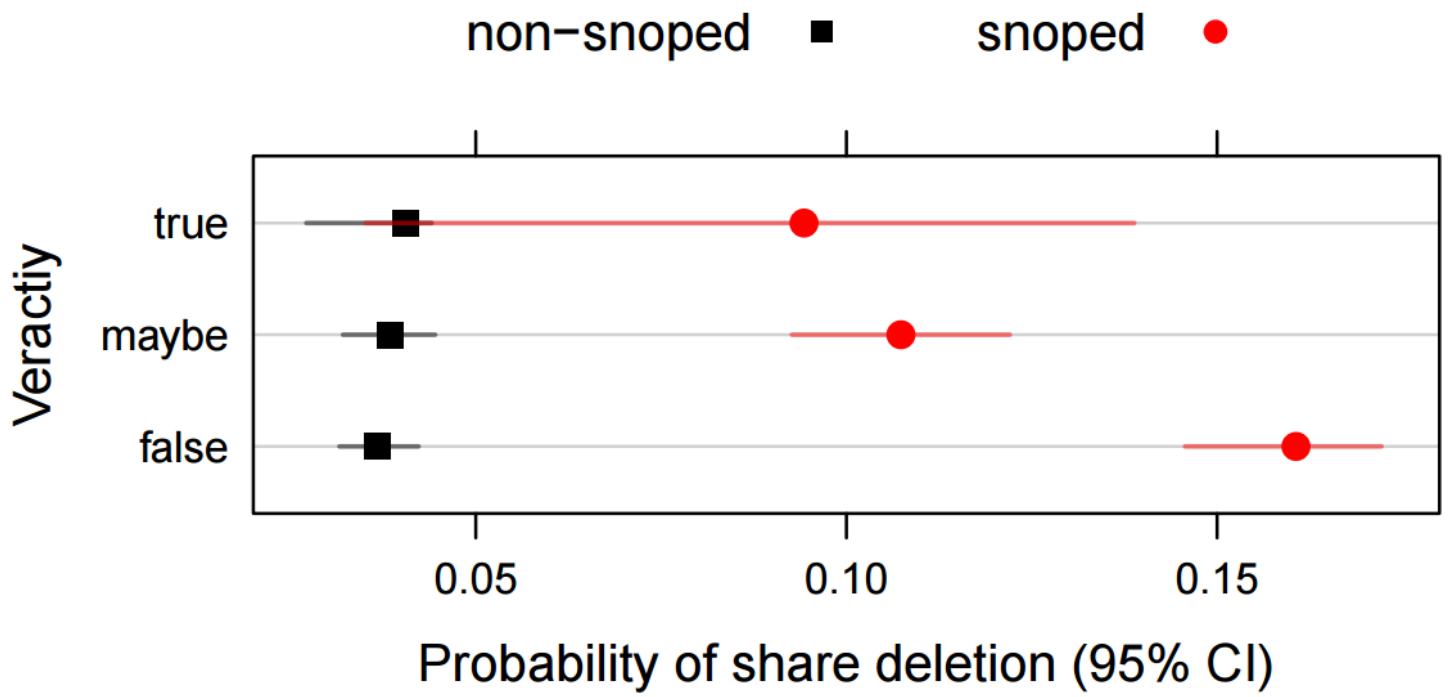




Humans get fooled when article looks more "genuine", and it is assumed to be credible.

Kumar, et al. (WWW 2016)

What happens when false information is pointed out?



Pointing out false information leads to its deletion, as observed in case study of Facebook

Frigerri, et al. (ICWSM 2014)

Summary: Hoaxes

- **Hoaxes:** False information pretending to masquerade as genuine information
- Disinformation spreads wide and fast, can survive for a long time, are viewed frequently and cited from across the web
- Wikipedia hoaxes are longer, but lack references, and are created by newer editors
- Hoaxes can be detected efficiently using non-superficial features
- Humans get fooled into believing hoaxes are genuine if it looks genuine
- But pointing out false information leads to its deletion

References

S. Kumar, R. West and J. Leskovec. Disinformation on the Web: Impact, Characteristics and Detection of Wikipedia hoaxes. WWW 2016.

A. Friggeri, L. A. Adamic, D. Eckles, and J. Cheng. Rumor cascades. In ICWSM, 2014

TweetCred: Real-Time Credibility Assessment of Content on Twitter. A. Gupta, P. Kumaraguru, C. Castillo, P. Meier. International Conference on Social Informatics. Springer 2014

B. Horne and S. Adali. This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News. The 2nd International Workshop on News and Public Opinion at ICWSM 2017.

A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi. Faking Sandy: Characterizing and identifying fake images on Twitter during hurricane Sandy. In WWW Companion, 2013.

References

S. Kumar, F. Spezzano and V.S. Subrahmanian. VEWS: A Wikipedia Vandal Early Warning System. SIGKDD 2015.

B. T. Adler, L. de Alfaro, and I. Pye. Detecting wikipedia vandalism using wikitrust - lab report for PAN at CLEF 2010. CLEF, 2010

B. T. Adler, L. de Alfaro, S. M. Mola-Velasco, P. Rosso, and A. G. West. Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. CICLing, 2011.

A. G. West, S. Kannan, and I. Lee. Detecting wikipedia vandalism via spatio-temporal analysis of revision metadata? in EUROSEC, 2010.

M. Potthast, B. Stein, and R. Gerling. Automatic vandalism detection in Wikipedia. Advances in Information Retrieval, ser. Lecture Notes in Computer Science, C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, and R. White, Eds. Springer Berlin Heidelberg, 2008.

S. Mola-Valesco. Wikipedia vandalism detection. WWW 2011.

References

Kumar, S., Cheng, J., Leskovec, J. & Subrahmanian, V.S. (2017). An Army of Me: Sockpuppets in Online Discussion Communities. Proceedings of the 26th International Conference Companion on World Wide Web. WWW 2017. http://bit.ly/sockpaper

Solorio, T., Hasan, R. & Mizan, M. (2013). A case study of sockpuppet detection in Wikipedia. LASM 2013.

Zheng, X., Lai, Y.M., Chow, K., CK, Lucas & Yiu, S.M. (2011). 2011 Seventh International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP)

Bu, Z., Xia, Z. & Wang, J. (2013). A sock puppet detection algorithm on virtual spaces. Knowledge-Based Systems 2013

Liu, D., Wu, Q., Han, W. & Zhou, B. Sockpuppet gang detection on social media sites. Frontiers of Computer Science

Gilbert, R., Thadani, V., Handy, C., Andrews, H., Sguigna, T., Sasso, A. and Payne, S.. The psychological functions of avatars and alt (s): A qualitative study. Computers in Human Behavior, 2014

References

Gilbert, R.L., Foss, J.A., and Murphy, N.A.. Multiple personality order: Physical and personality characteristics of the self, primary avatar and alt. In Reinventing ourselves: Contemporary concepts of identity in virtual worlds, Springer, 2011.

A. Caspi and P. Gorsky. Online deception: Prevalence, motivation, and emotion. CyberPsychology & Behavior, 2006.

M. Tsikerdekis and S. Zeadally. Multiple account identity deception detection in social media using nonverbal behavior. IEEE Transactions on Information Forensics and Security, 9(8):1311–1321, 2014.

M. Tsikerdekis and S. Zeadally. Online deception in social media. Communications of the ACM, 57(9):72–80, 2014.

Z. Yamak, J. Saunier, and L. Vercouter. Detection of multiple identity manipulation in collaborative projects. In Proceedings of the 25th International Conference Companion on World Wide Web. 2016.

KDD 2017
Halifax, Nova Scotia - Canada
August 13 - 17, 2017

Tutorial: Data-Driven Approaches towards Malicious Behavior Modeling

Meng Jiang
University of Notre Dame

Srijan Kumar
Stanford University

Christos Faloutsos
Carnegie Mellon University

V.S. Subrahmanian
University of Maryland, College Park

http://www.meng-jiang.com/tutorial-kdd17.html