



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

Tutorial link: <http://bit.ly/kdd2017>

Outline

Introduction

Feature-based
algorithms

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

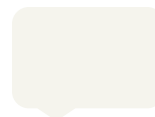
Conclusions and future directions

Tutorial link: <http://bit.ly/kdd2017>

amazon



WIKIPEDIA



reddit



You**Tube**



US

BIG STORY

Nintendo Switch
Pro Controller
Video Review



Bungie Announces
Destiny 2 Reveal
Trailer Launch Date

4 Reasons Legends
of Tomorrow
Became Amazing

Daily Deals: PS Plus
1yr Membership for
\$48

The Best Action
Movies to Stream
on Netflix

Browse

Search

[Shows](#) [Store](#) [Reviews](#) [PS4](#) [Xbox](#) [PC](#) [Nintendo](#) [Movies](#) [TV](#) [VR](#)

3
following

3
followers

33
games

Add New Game ID

Level 7

[Activity](#)[Blog](#)[Games](#)[People](#)[View My IGN »](#) [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

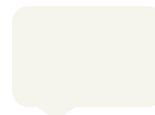
I still don't see why dc is better than marvel. I personally like dc better and I like their superheroes more than marvels, but I do not think that this article tells exactly why.

(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



WIKIPEDIA



reddit



You**Tube**

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



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



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

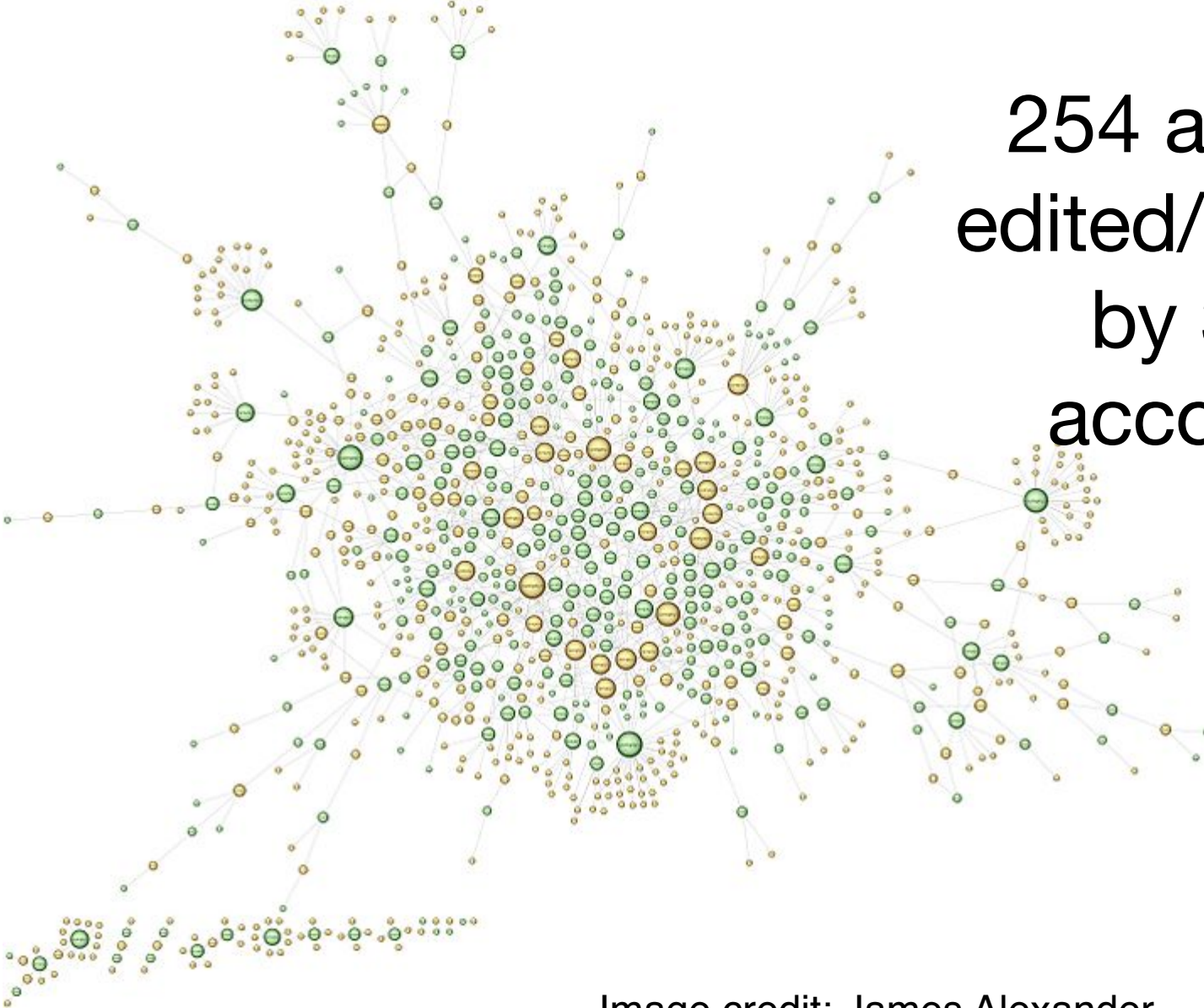


Image credit: James Alexander

Online sockpuppetry – why?



Diversify identity



Anonymize identity



Multiply identity



Privacy concern

Sockpuppets in Wikipedia

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**

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**

Wikipedia Sockpuppets



An editor has expressed a concern that this account may be a **sock puppet** of [Example](#) ([talk](#) · [contribs](#) · [logs](#)).

Please refer to [the sockpuppet investigation](#) of the sockpuppeteer, and editing habits or [contributions](#) of the sock puppet for evidence. [This policy subsection](#) may also be helpful.

Account information: [block log](#) – [current autoblocks](#) – [contribs](#) – [logs](#) – [abuse log](#) – [CentralAuth](#)

Suspected by volunteers,
confirmed and deleted by administrators

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

DISQUS



BREITBART



A.V. CLUB



2.9M
Users

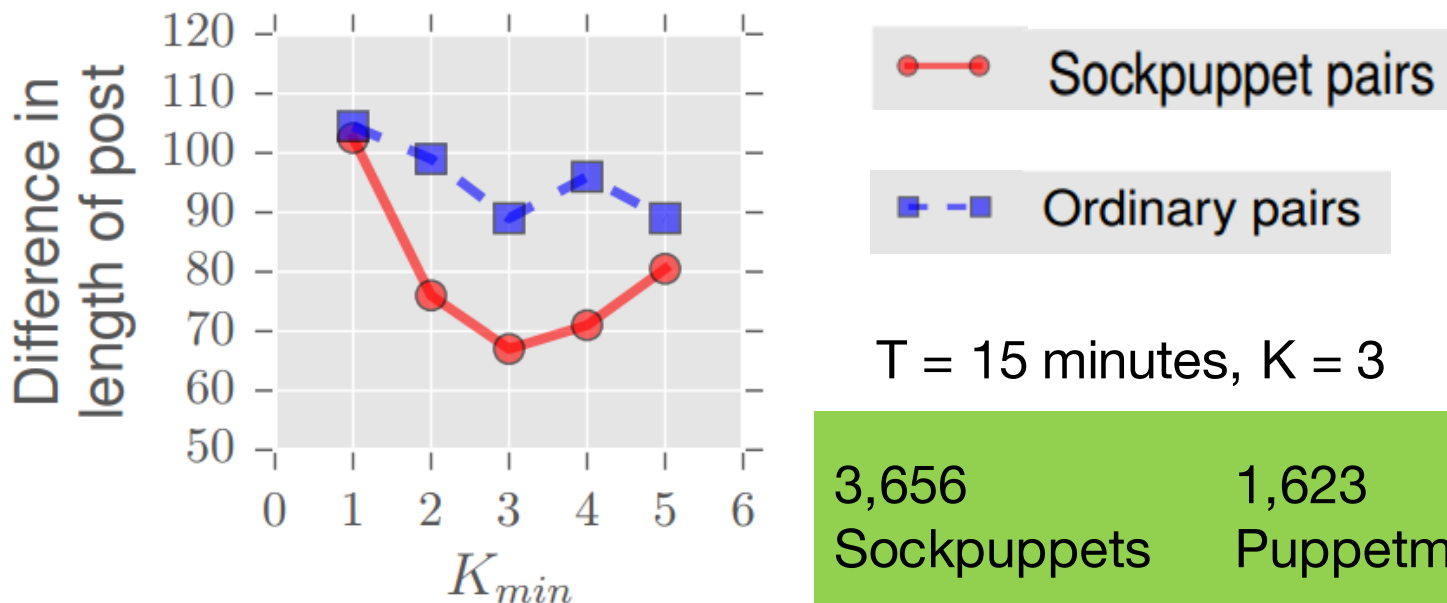
2.1M
Articles

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**.

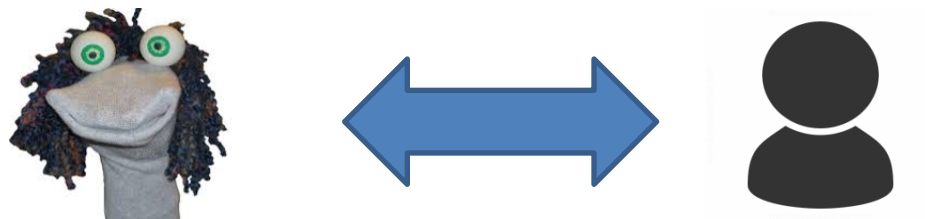


Note: we do not use the IP addresses for detection

Characteristics of sockpuppets

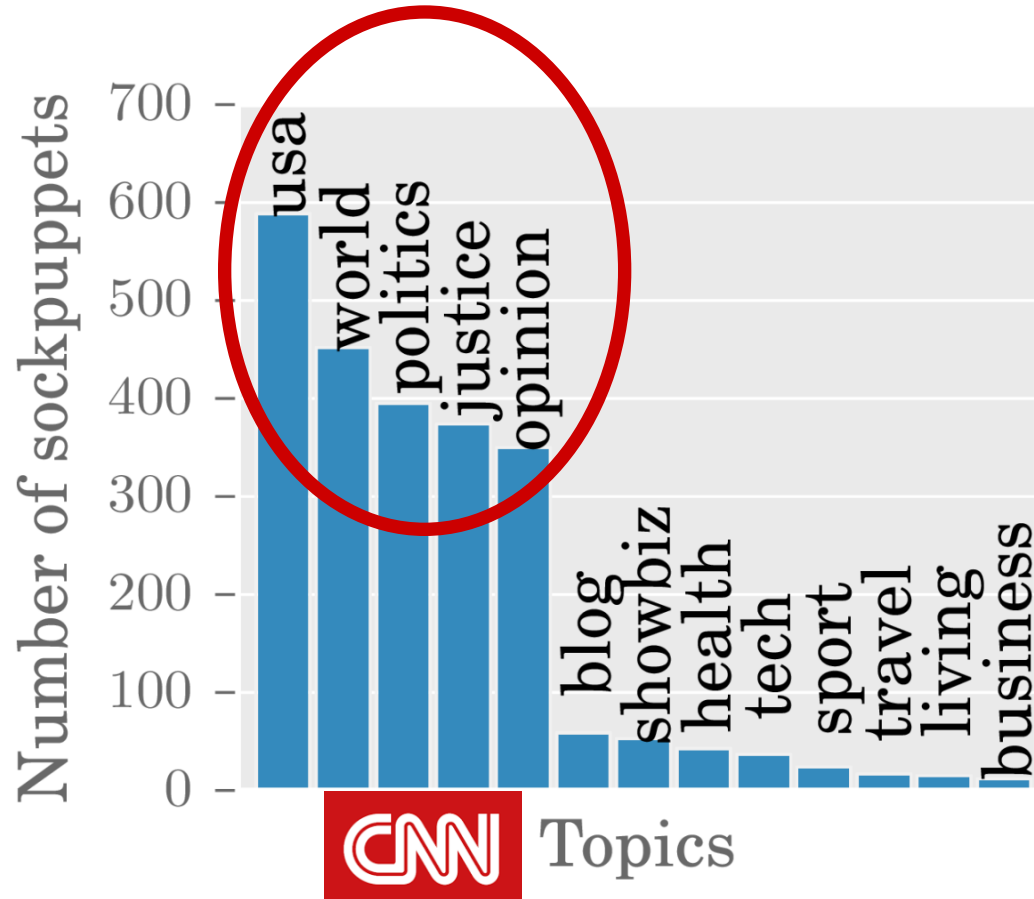
How do we compare sockpuppets and ordinary users?

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!!!!

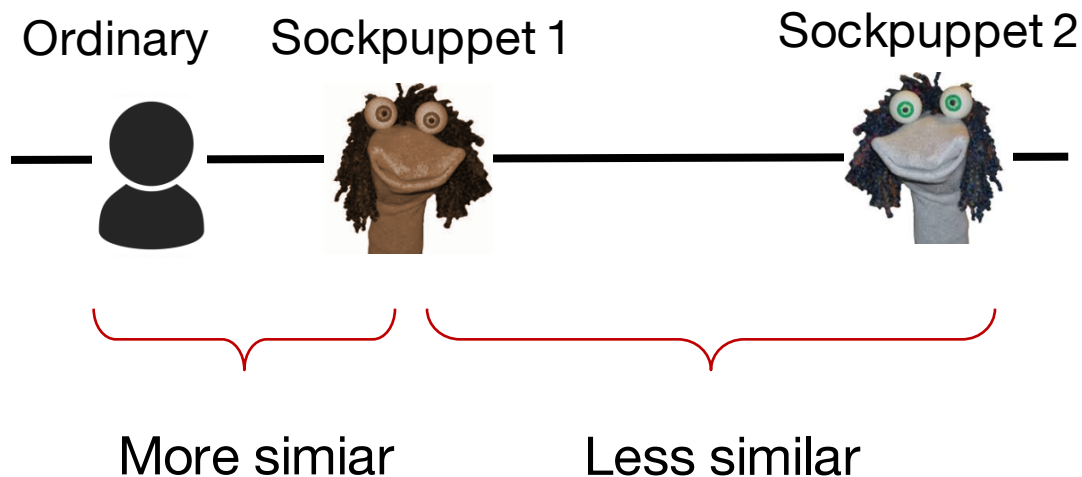


Interact more with each other

$p < 10^{-3}$

Do puppetmasters lead double lives?

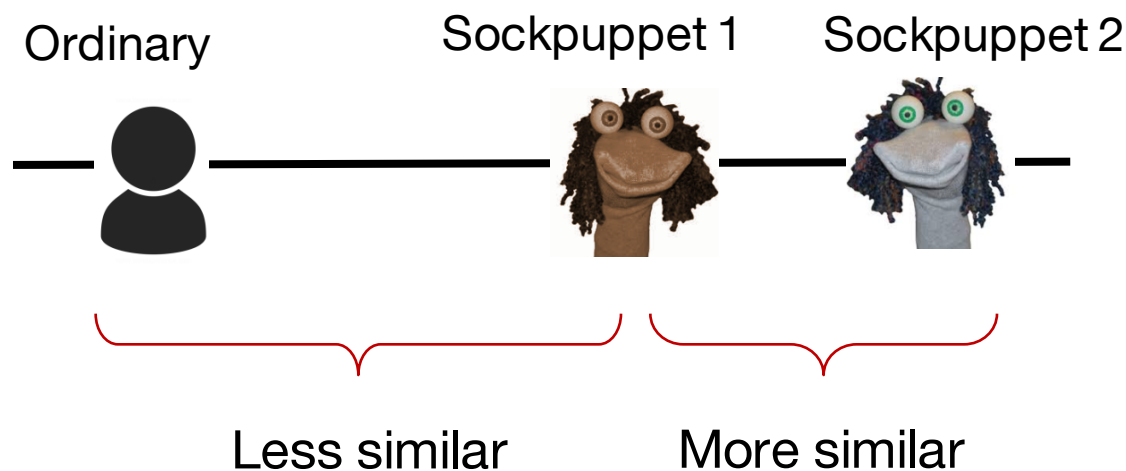
Double life hypothesis:
Puppetmaster maintains distinct personality for
the two sockpuppets



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

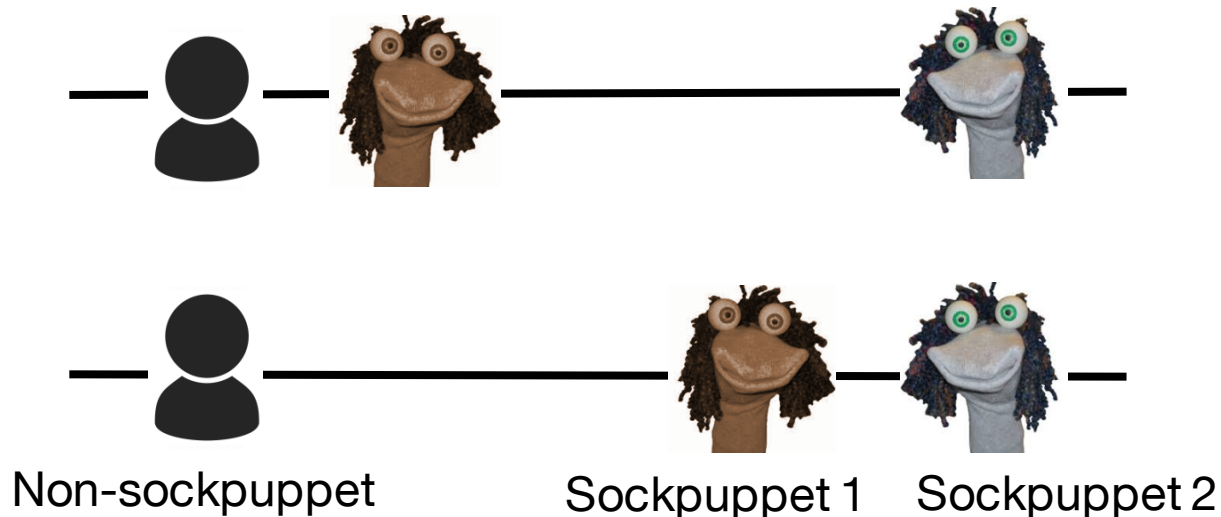
Do puppetmasters lead double lives?

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

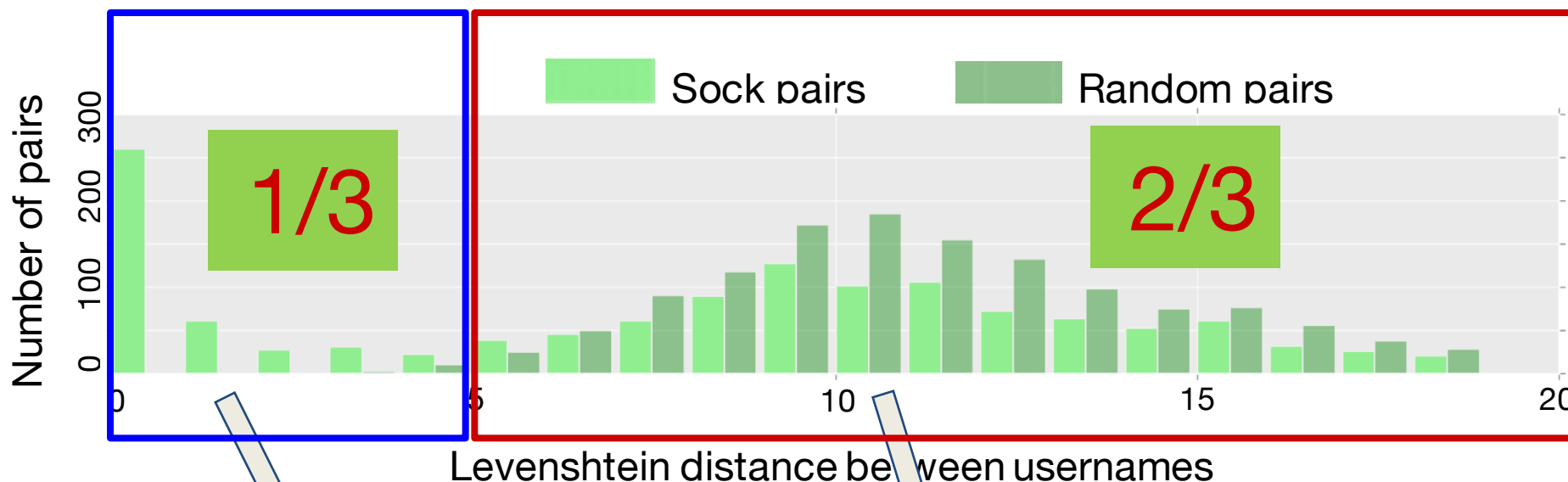
How are sockpuppets used?
Are they used for deception?

Deceptiveness

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

Non-Pretenders

Pretenders



srijan

srijan2



srijan

theRealBatman

Pretender vs non-pretender sockpuppets



srijan

Feb 5 2013, 2PM

best article i have read!!!



ricob

More opinionated

$p < 10^{-3}$

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

Kumar, et al. (WWW 2017)

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

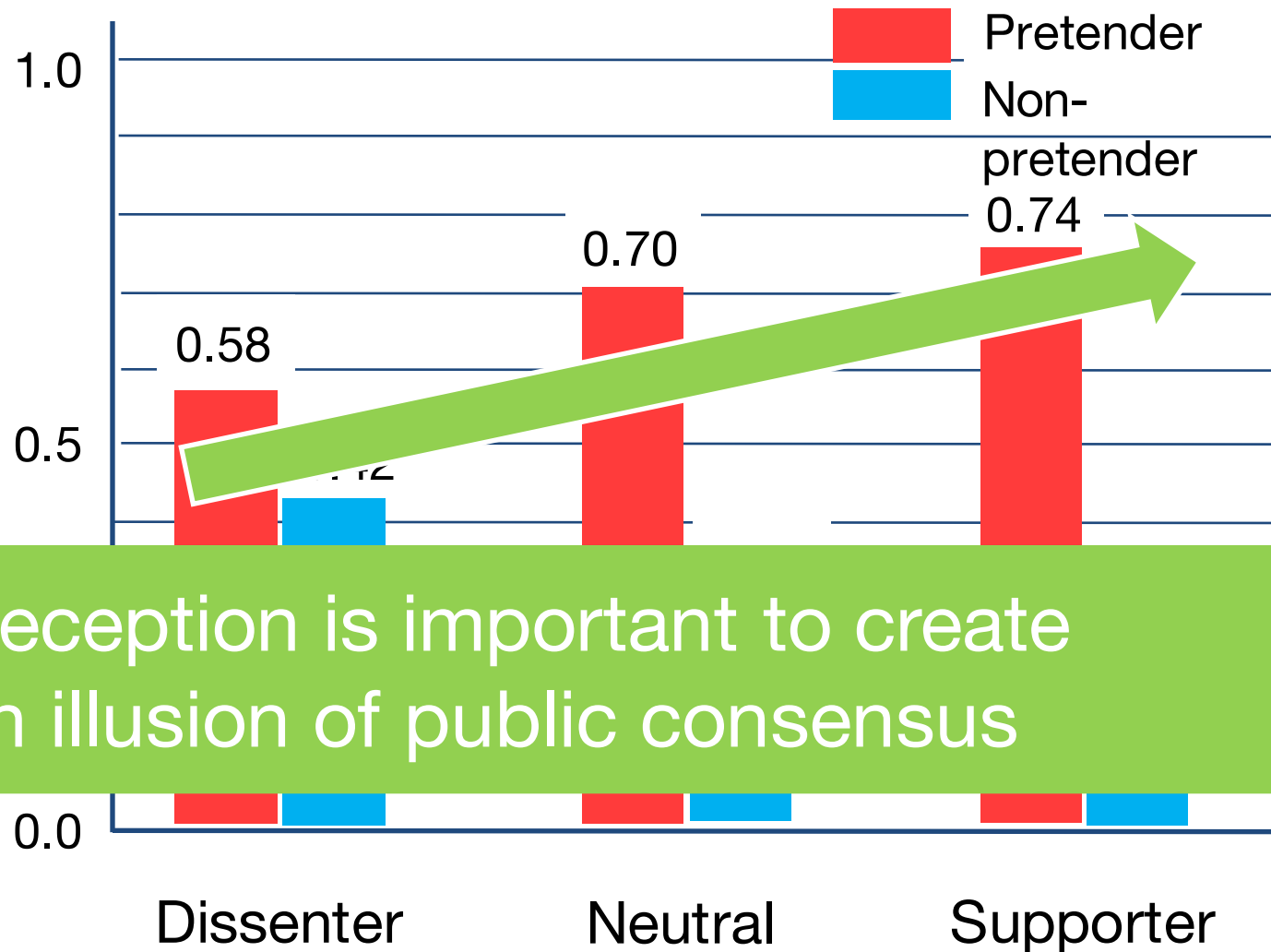
I don't think so

60%
Neutral

30%
Supporter

10%
Dissenter

Supportiveness and Deceptiveness



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,
...

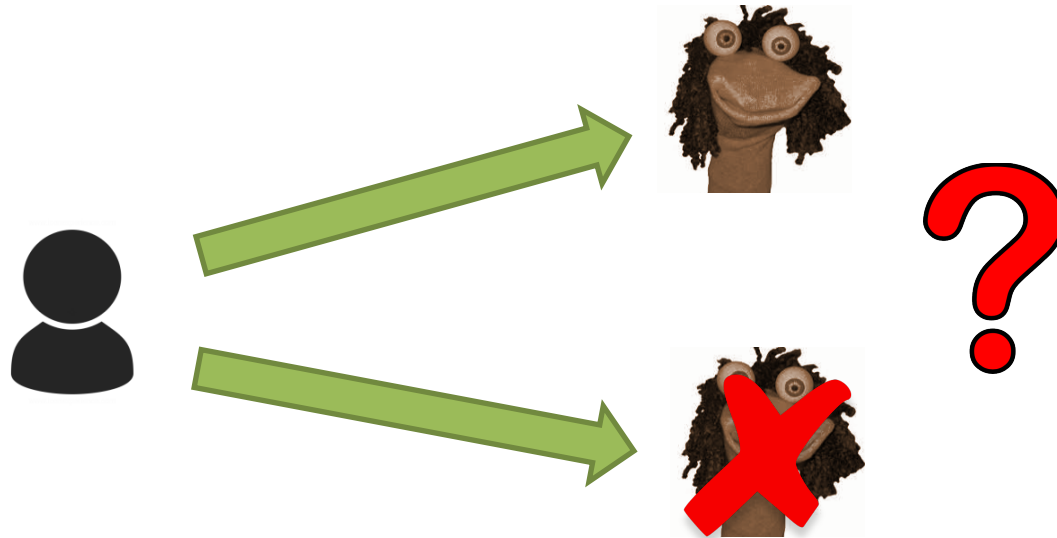


Community

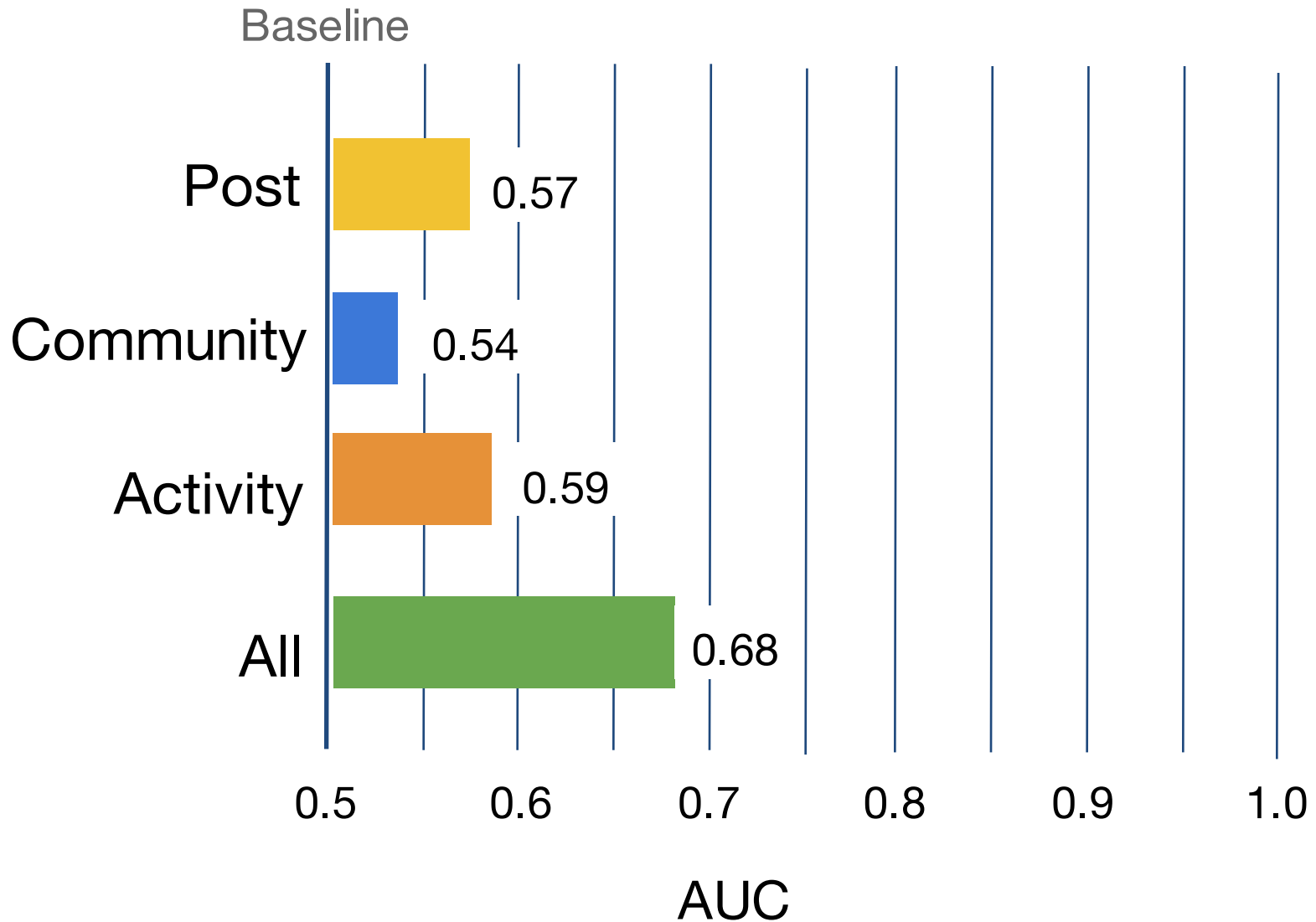
Number of upvotes and
downvotes,
Fraction of reported posts,
Is account reported,
...

Note: we are not using the IP based features

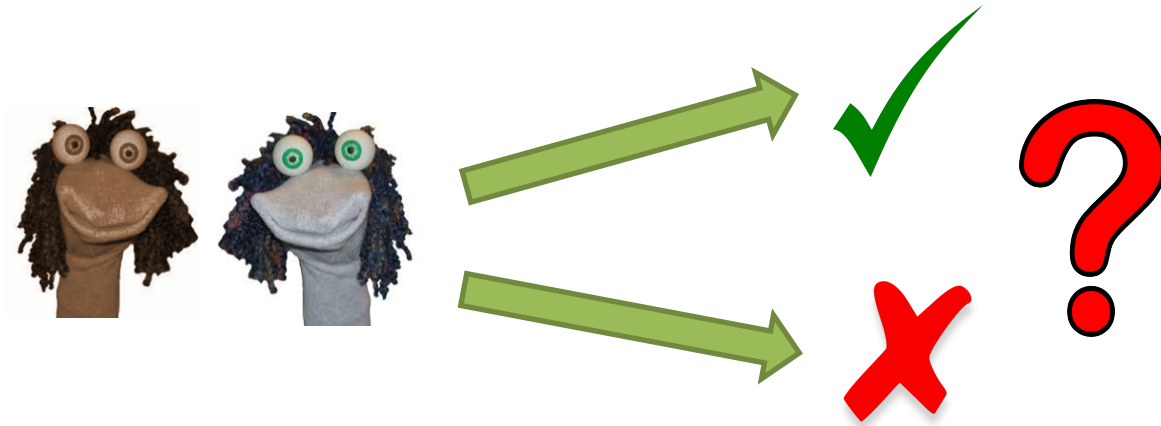
Is an account a sockpuppet?



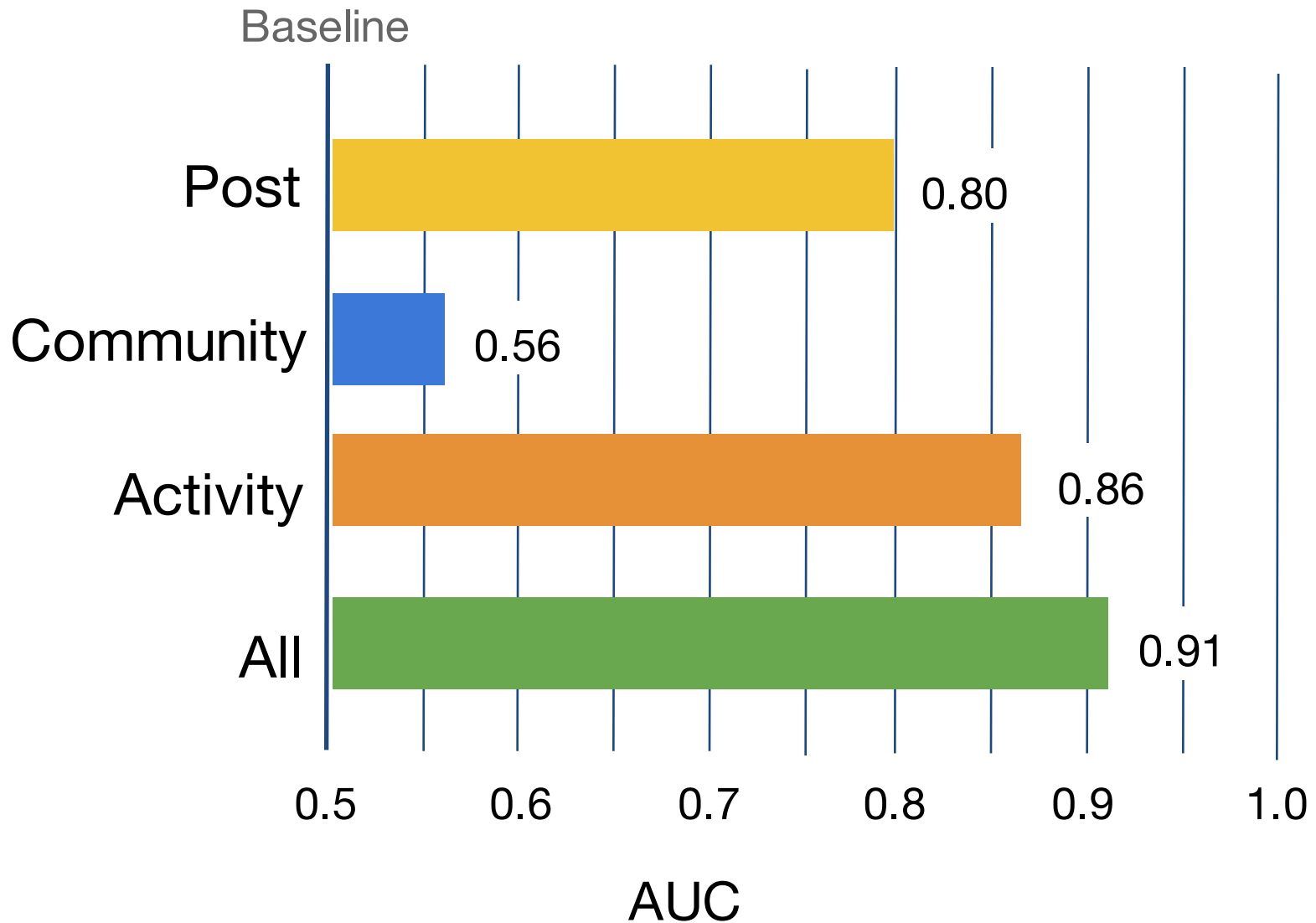
Is an account a sockpuppet?



Do two accounts belong to the same person?



Do two accounts belong to the same person?



Outline

Introduction

Feature-based
algorithms

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

Tutorial link: <http://bit.ly/kdd2017>

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



Easy to add content

WIKIPEDIA

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



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

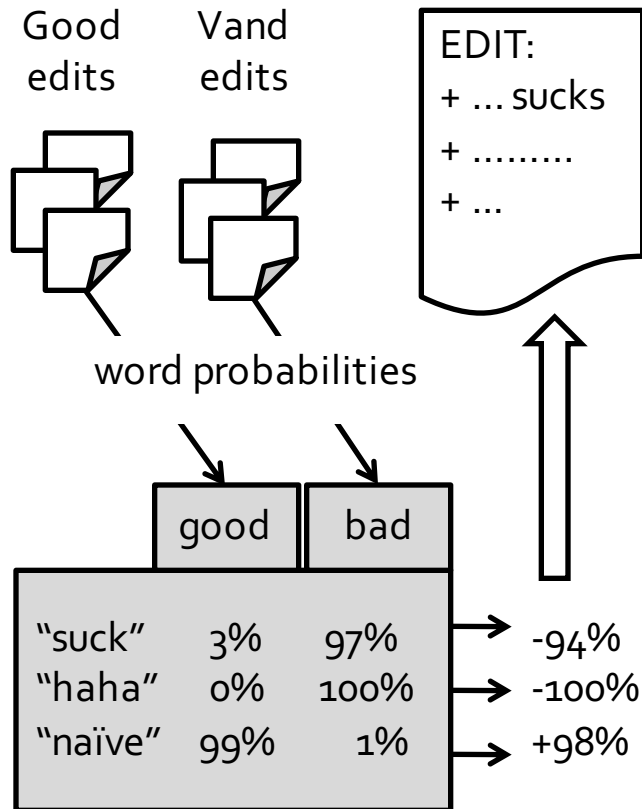
length, section-edit?

TIMESTAMP

time-of-day, day-of-week

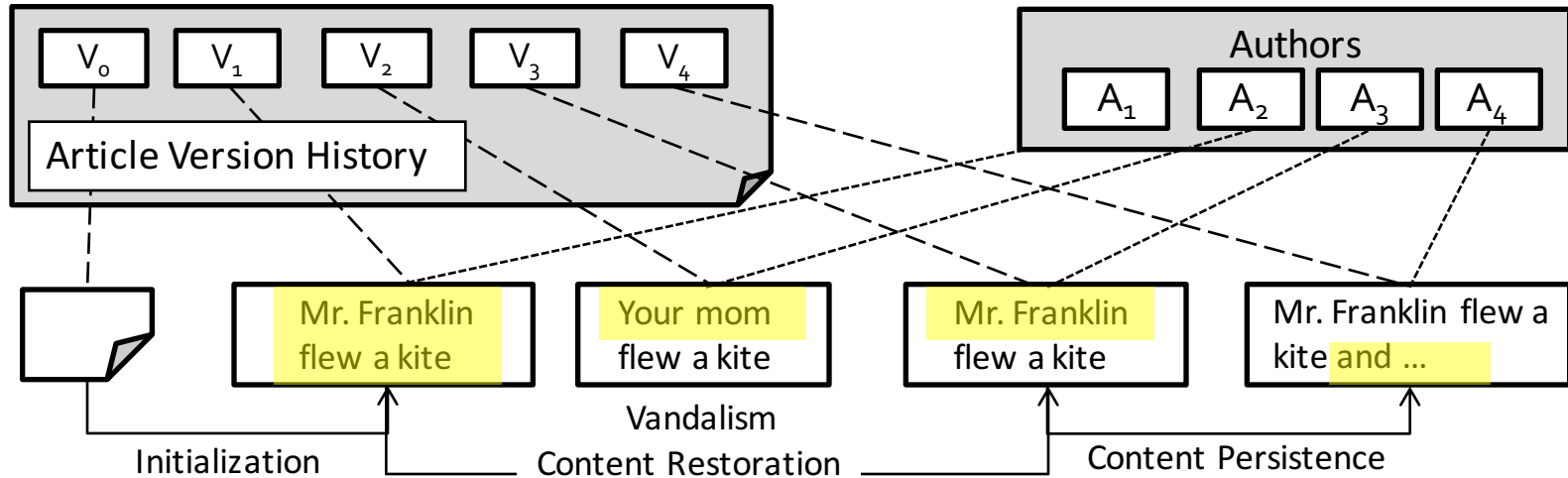
ClueBot NG: Textual

Bayesian Approach:



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

WikiTrust: Content driven



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

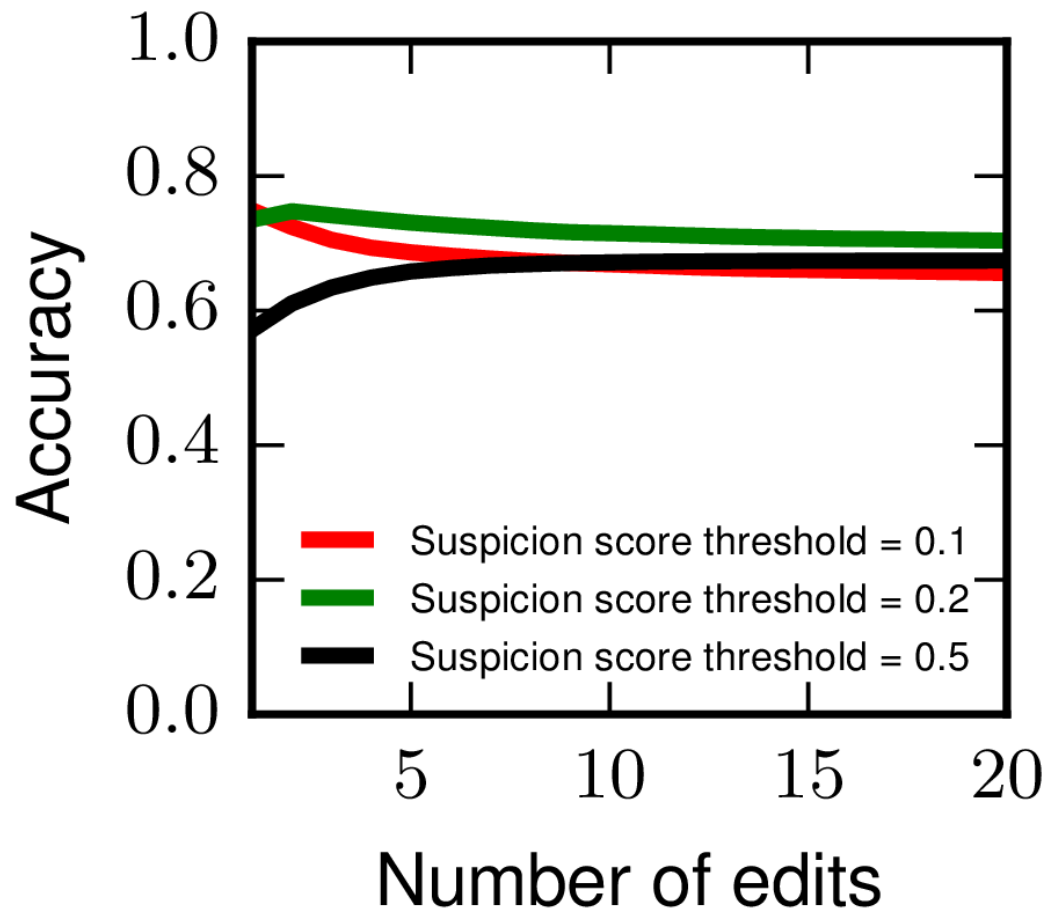
Detection of vandals

Vandalism
detection



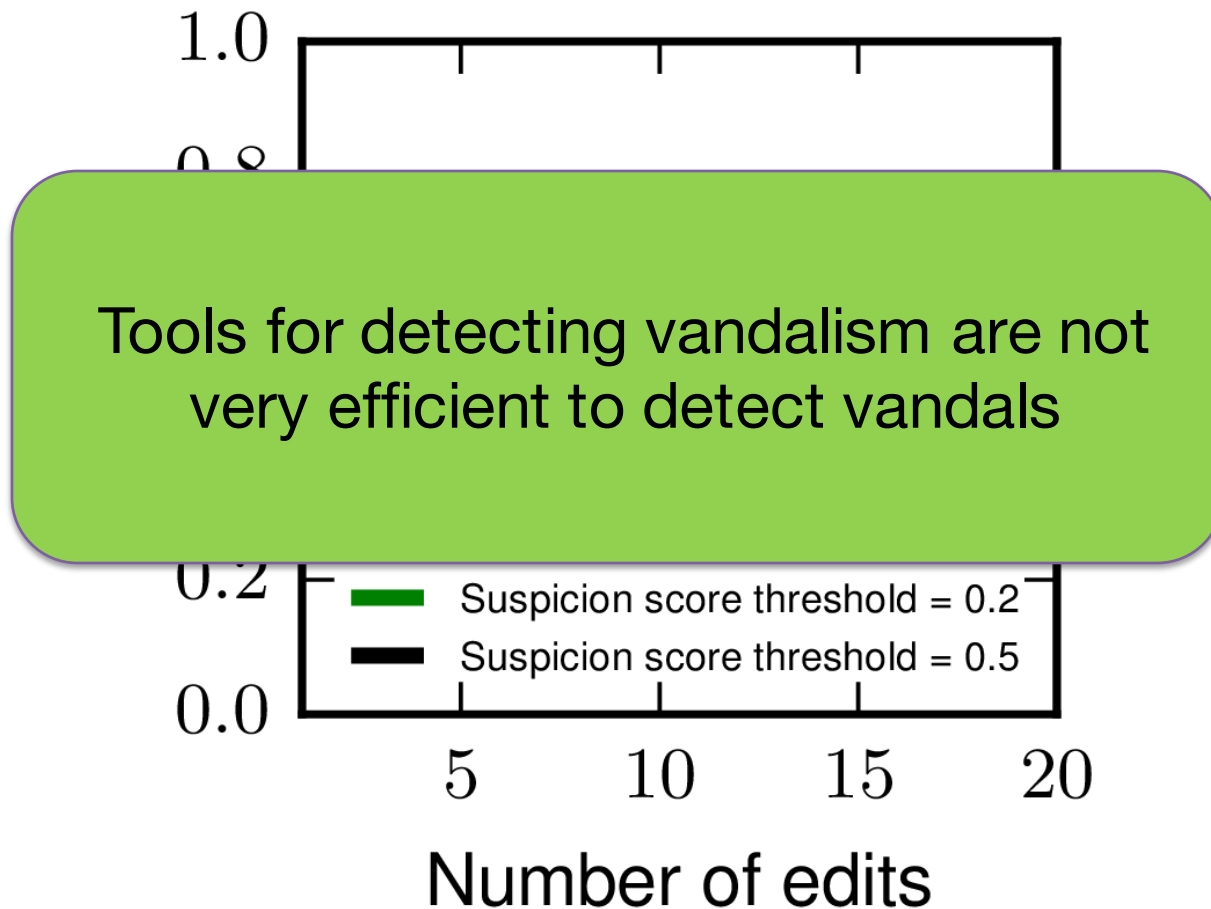
Vandal
detection

Using STiki to detect vandals



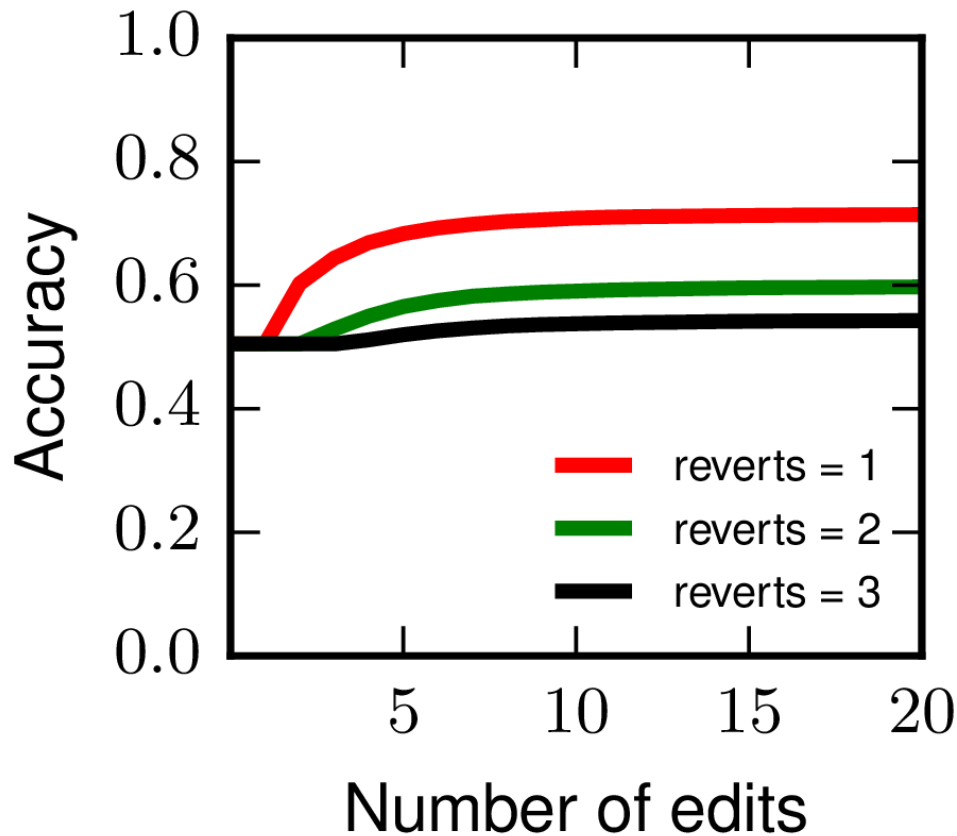
Stiki rule: Editor is a vandal if any edit's suspicion score exceeds threshold

Using STiki to detect vandals



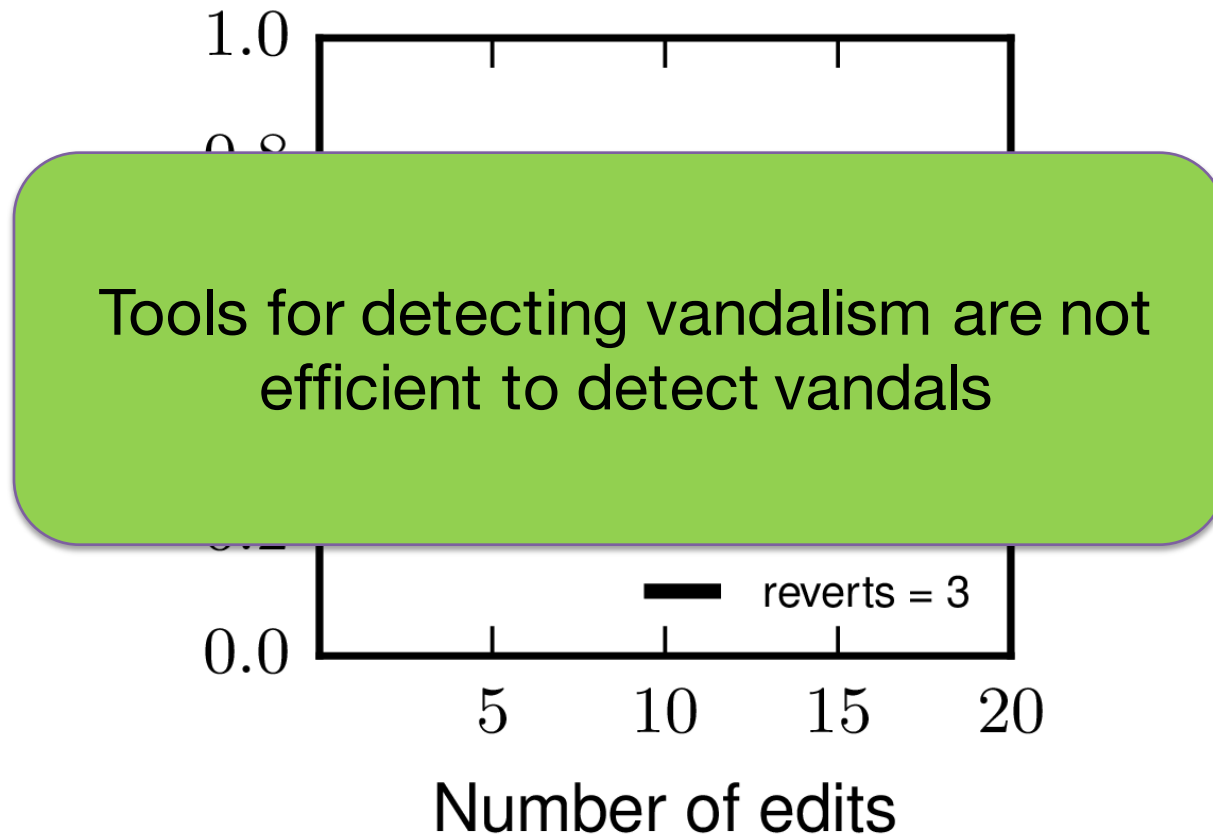
Stiki rule: Editor is a vandal if any edit's suspicion score exceeds threshold

Using ClueBot NG to detect vandals



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

Using ClueBot NG to detect vandals



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

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

Characteristics of vandals

Visit the main page



WIKIPEDIA

The Free Encyclopedia

Main page

Contents

Featured content

Current events

Random article

Donate to Wikipedia

Wikipedia store

Not logged in

Talk

Contributions

Create account

Log in

Article

Talk

Read

Edit

View history

Search Wikipedia

Halifax, Nova Scotia

From Wikipedia, the free encyclopedia

Coordinates:  44°38′52″N 63°34′17″W

Halifax (/ˈhælɪfæks/, locally /ˈhɛlɪfæks/), legally known as the **Halifax Regional Municipality (HRM)**, is the capital of the province of Nova Scotia, Canada. The municipality had a population of 403,131 in 2016, with 316,701 in the urban area centred on Halifax Harbour.^{[3][4]} The regional municipality consists of four former municipalities that were amalgamated in 1996: Halifax, Dartmouth, Bedford, and the Municipality of Halifax County.

Halifax

K'jipuktuk

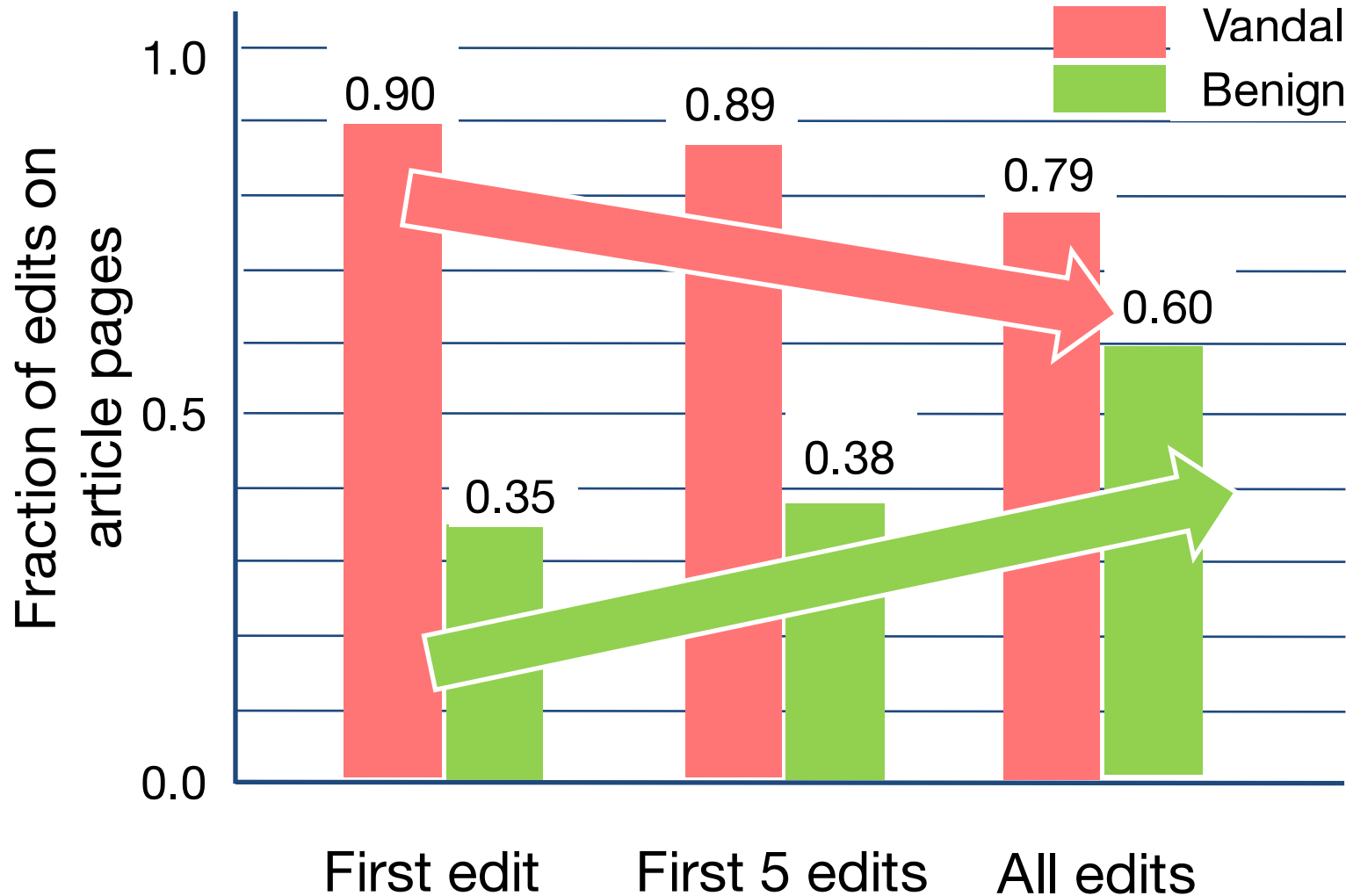
Regional municipality

Halifax Regional Municipality

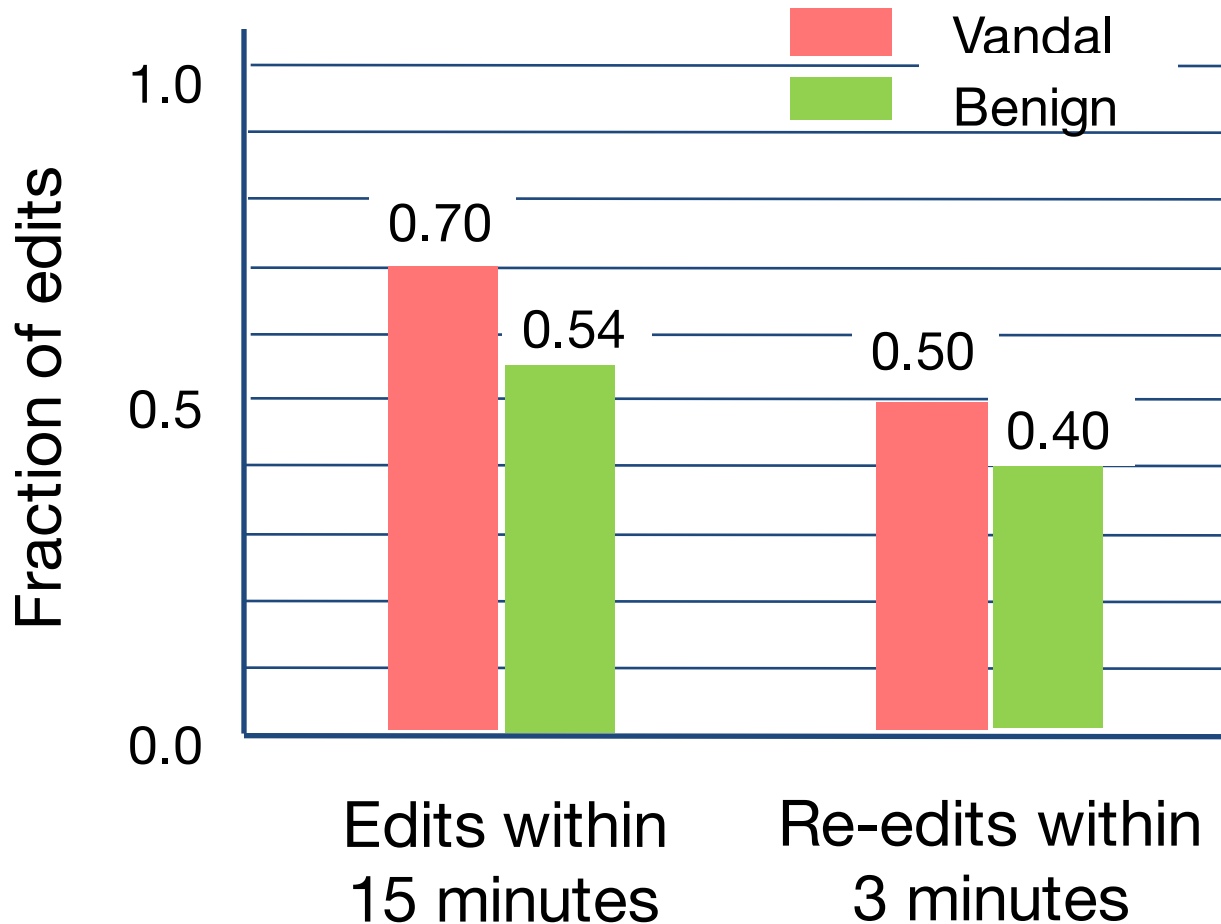


Editors can edit article pages and talk (discussion) pages

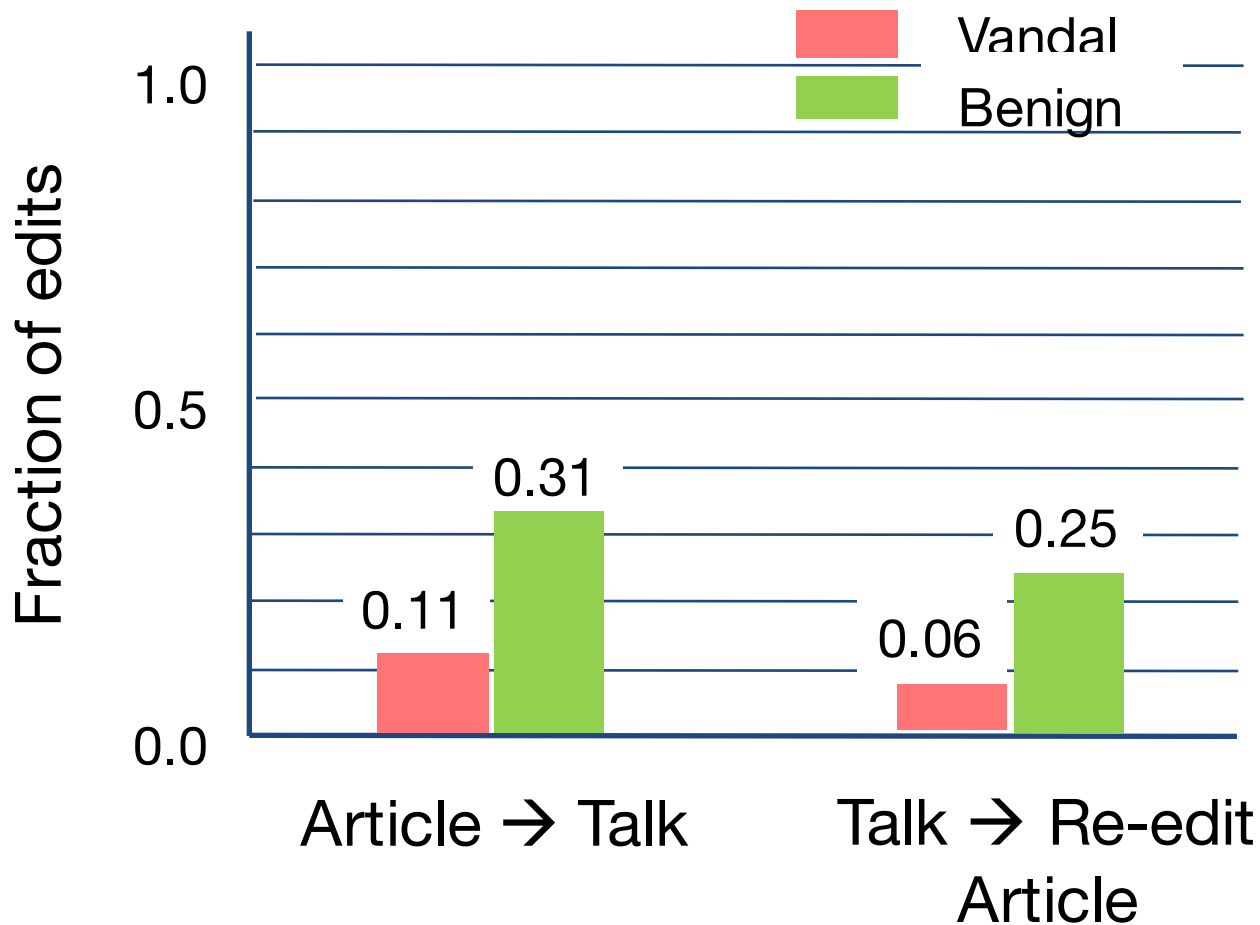
Vandals make visible edits



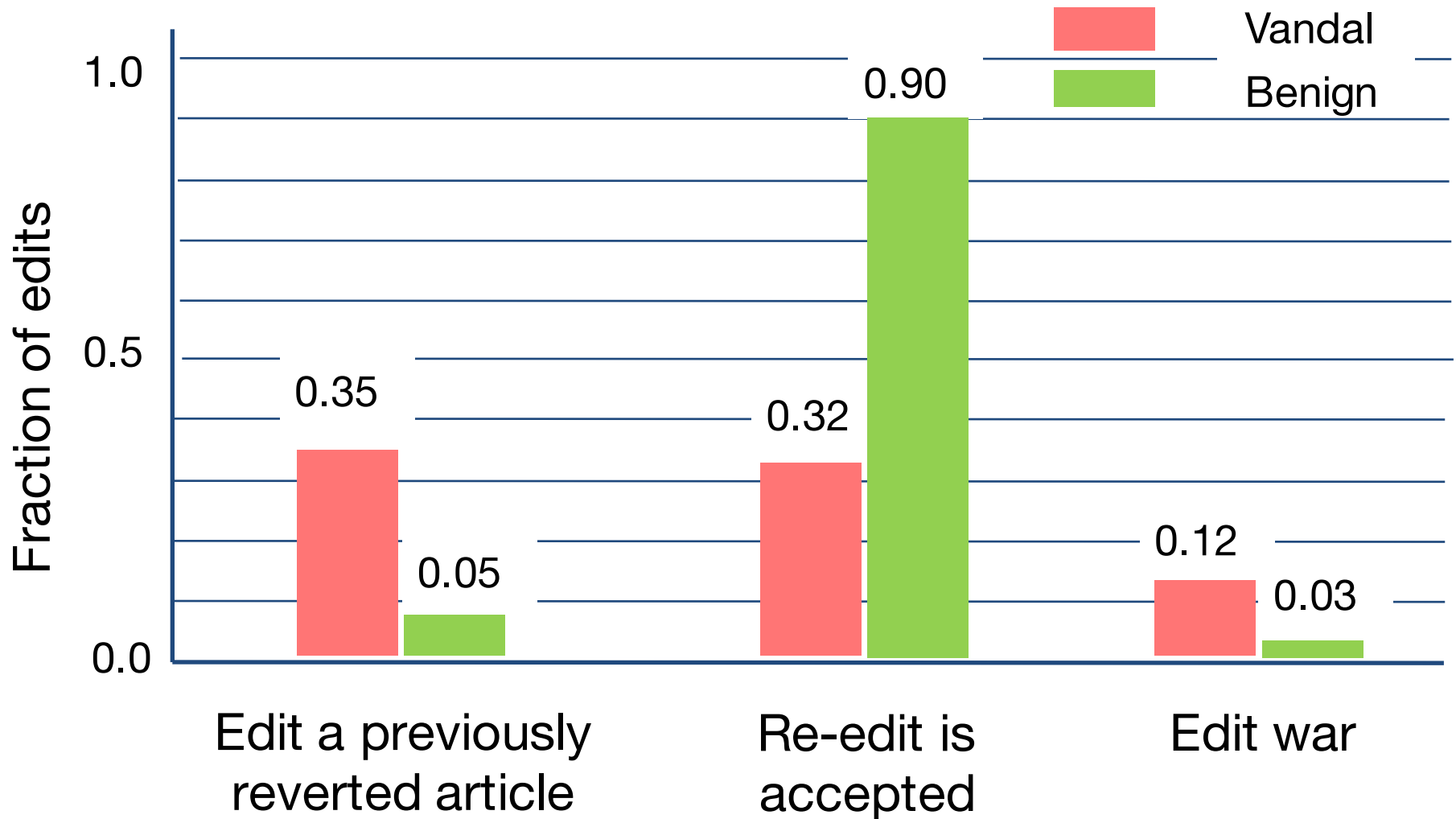
Vandals are quicker



Vandals do not discuss

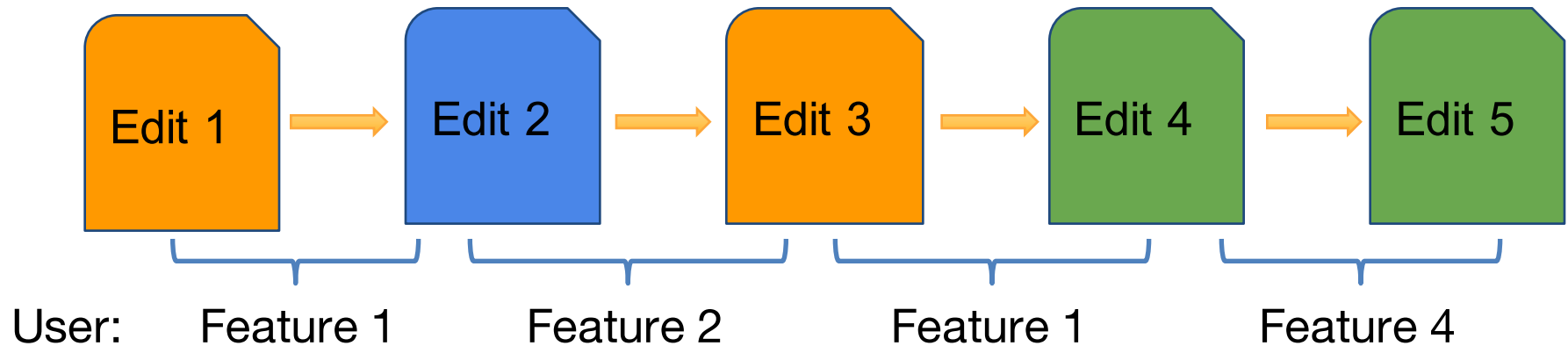


Vandals make reversion driven edits



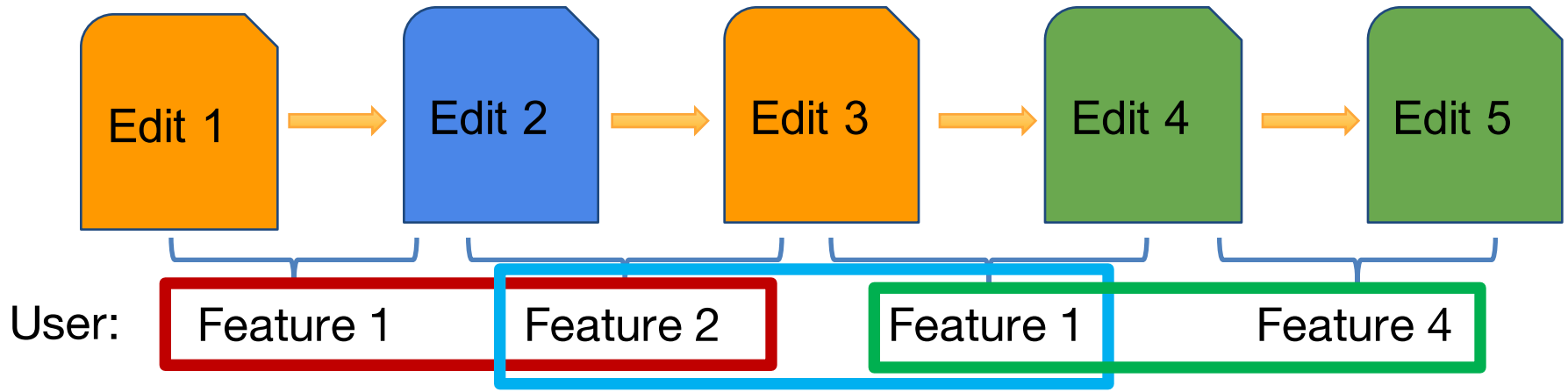
Detecting vandals

Pairwise Edit Features



Time \times Type of page \times First edit \times Distance \times Similarity
 \times Reverted or not

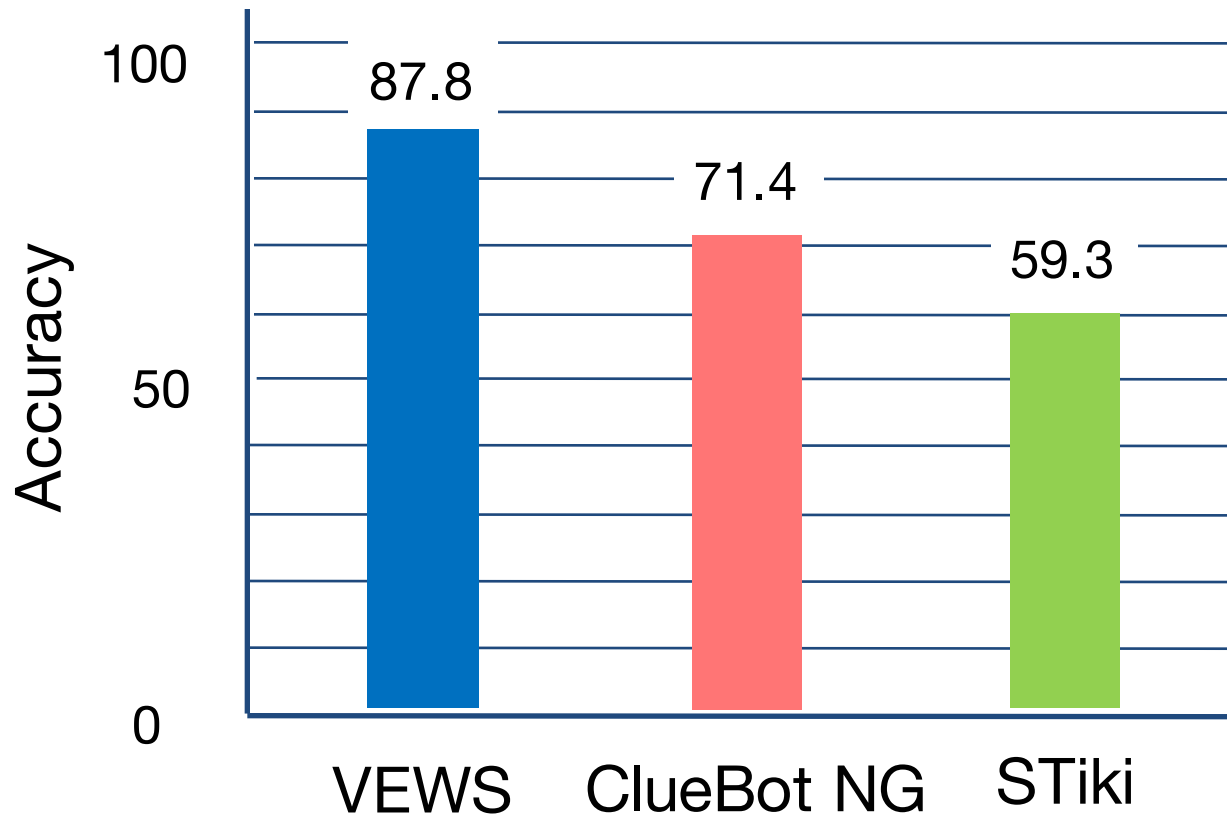
Meta-Features: Transitions



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

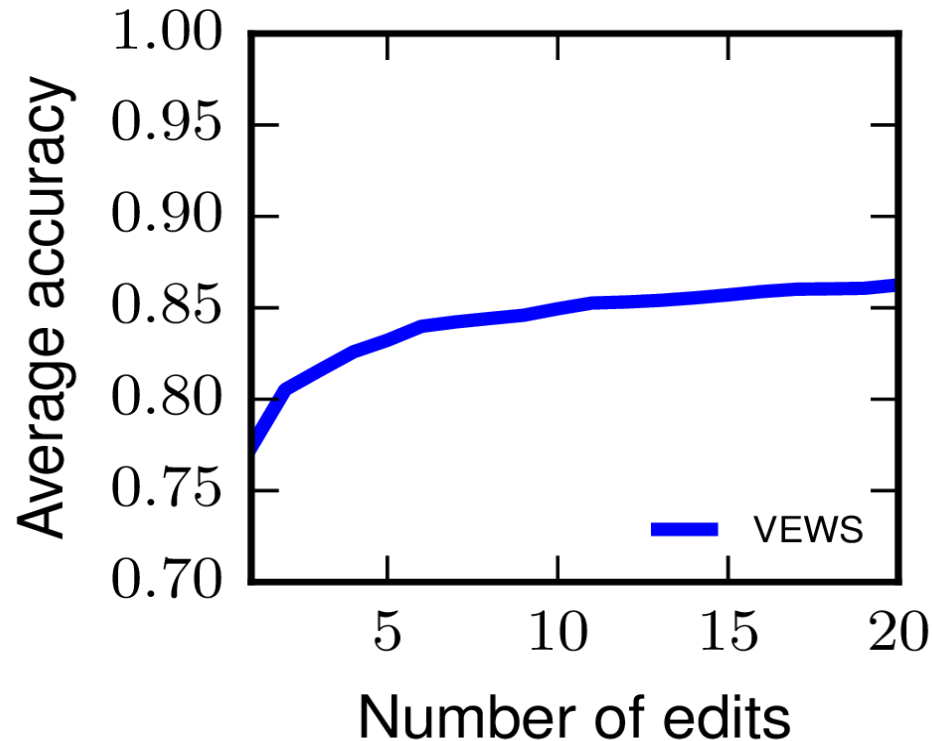
N x N

Vandal Detection



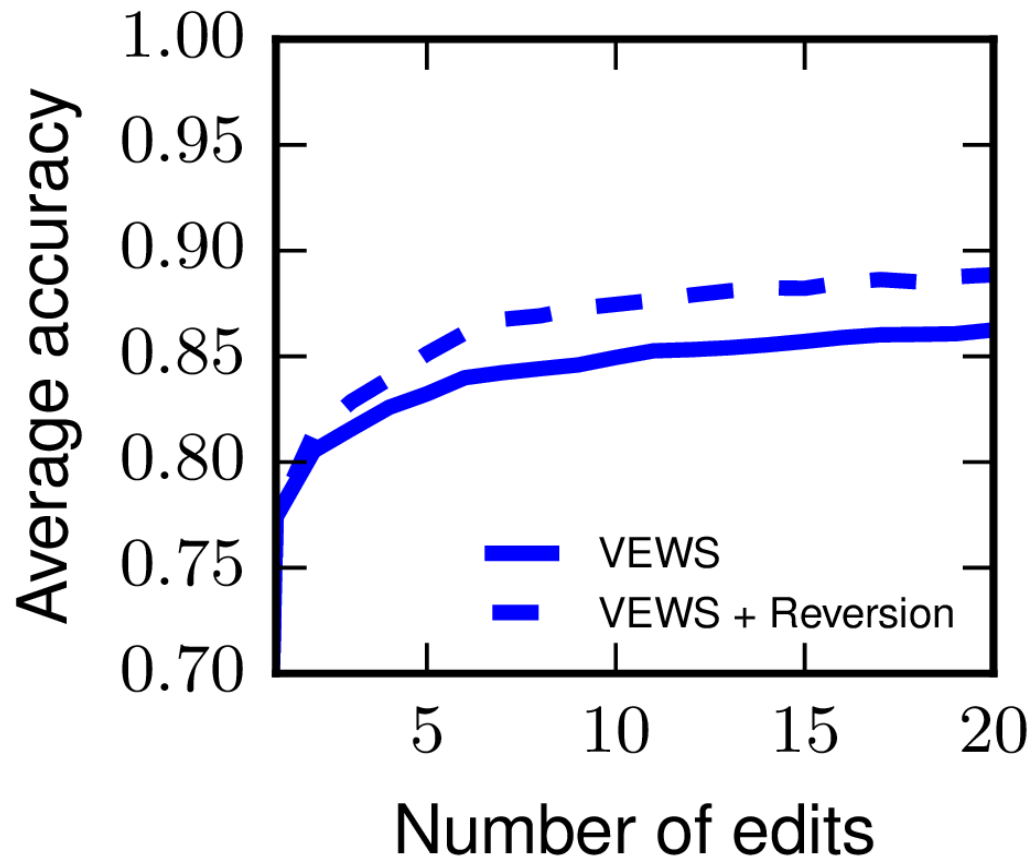
VEWS identifies 87% vandals on or before first reversion.
44% vandal are identified before first reversion.

Early Warning System

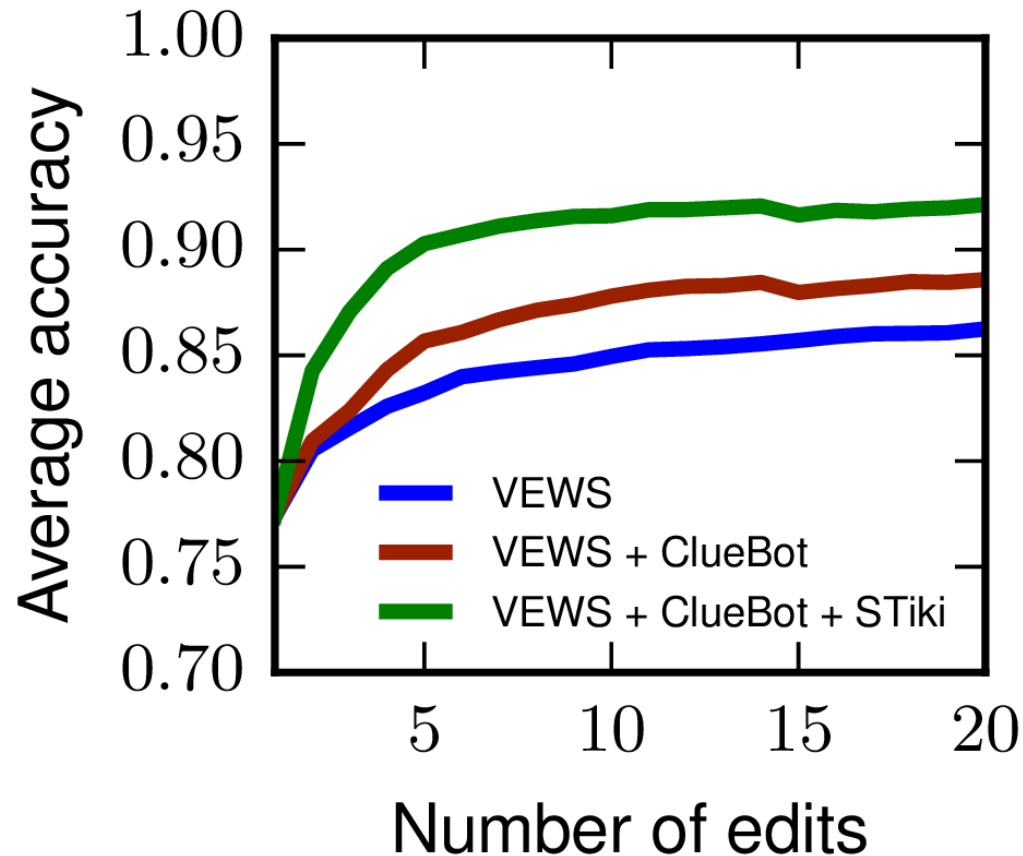


VEWS identifies vandals in
2.13 edits on average

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

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

Tutorial link: <http://bit.ly/kdd2017>

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/>

NEWS FASHION ▼ TECH ▼ VIDEO ▼ WORLD ▼



Home > News > Obama Signs Executive Order Banning The National Anthem At All Sporting Events...

NEWS

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

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

SHARE



Facebook



Twitter



Recent Comments

DOCS on *Gay Wedding Mobile Vans Cashing In On The Legalization Of Gay Marriage*

eric turner on *Obama Signs Executive Order Declaring Investigation Into Election Results; Revote Planned For Dec. 19th*

friols on *Obama Signs Executive Order Declaring Investigation Into Election Results; Revote Planned For Dec. 19th*

Brian on *Obama Signs Executive Order Declaring Investigation Into Election Results; Revote Planned For Dec. 19th*



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



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](#)

[Tools](#)
[What links here](#)
[Related changes](#)

[Create account](#) [Not logged in](#) [Talk](#) [Contributions](#) [Log in](#)

[Project page](#) [Talk](#) [Read](#) [View source](#) [View history](#)

Wikipedia:List of hoaxes on Wikipedia/Jar'Edo We

From Wikipedia, the free encyclopedia
< [Wikipedia:List of hoaxes on Wikipedia](#)

This is an **old revision** of this page, as edited by [108.215.62.12](#) (talk) at 11:56, 21 July 2012. 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)

In [Australian aboriginal mythology](#), **Jar'Edo We** is a [mythological figure](#) of knowledge and physical might, created by [Altji](#). He is associated with getting too arrogant or self-conceited. He is associated with the [Warramiri](#) people.


This article relating to a myth or legend needs to be improved. You can help Wikipedia by expanding it.

Categories: [Aboriginal gods](#) | [Knowledge gods](#)

[Tools](#)
[What links here](#)
[Related changes](#)
[Upload file](#)
[Special pages](#)
[Permanent link](#)
[Page information](#)

[Print/export](#)
[Create a book](#)
[Download as PDF](#)
[Printable version](#)

[Languages](#) [Add links](#)



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](#)

[Tools](#)
[What links here](#)
[Related changes](#)
[Upload file](#)
[Special pages](#)
[Permanent link](#)
[Page information](#)

[Print/export](#)
[Create a book](#)
[Download as PDF](#)
[Printable version](#)

[Languages](#) [Add links](#)

[Create account](#) [Not logged in](#) [Talk](#) [Contributions](#) [Log in](#)


[Project page](#) [Talk](#) [Read](#) [View source](#) [View history](#)

Wikipedia:List of hoaxes on Wikipedia/Balboa Creole French

From Wikipedia, the free encyclopedia
< [Wikipedia:List of hoaxes on Wikipedia](#)

This is an **old revision** of this page, as edited by [108.215.62.12](#) (talk) at 11:56, 21 July 2012. 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)



This article **does not cite any references (sources)**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. *(January 2010)*

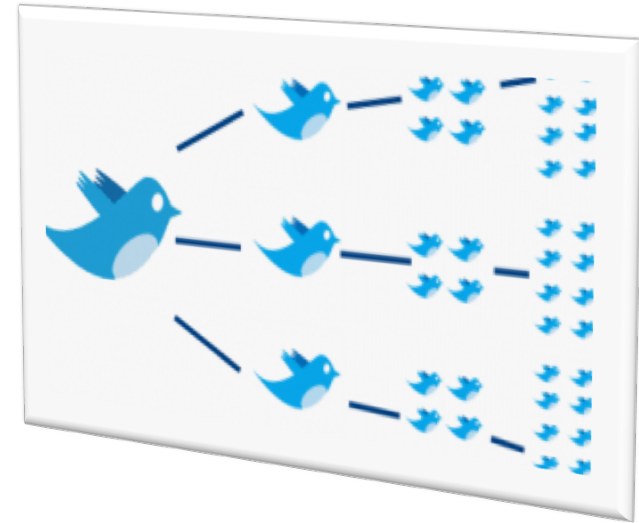
Balboa French Creole is a [Creole language](#) used in [Balboa Island](#) in the city of [Newport Beach, California](#). It originated from a blending of French spoken by French families on the island with [English](#), [Spanish](#), and [German](#), all which are spoken by some members of the Balboa Island community. Balboa Creole French differs highly from Standard French and is incomprehensible to the majority of French speakers. People from [Haiti](#) or the French Caribbean can sometimes understand the Creole, but it remains unintelligible to the masses. Some major differences are its subjects which are *Jah* or *Mwa*, *Tu*, *Vous* or *Tu'z All*, *Nos*, *Il*, *Elle*, *Ilz* or *Ellez* and *Dem*. In a census published in 2009, it was revealed only 14 people on the island can still speak the language.

Balboa Creole French	
Native to	California
Region	limited to quarters of Balboa Island
Native speakers	virtually extinct; a few families are bilingual in either English , or rarely in French (<i>date missing</i>)
Language family	<div>Creole<ul style="list-style-type: none">Balboa Creole French</div>
Language codes	
ISO 639-2	cpf
ISO 639-3	–

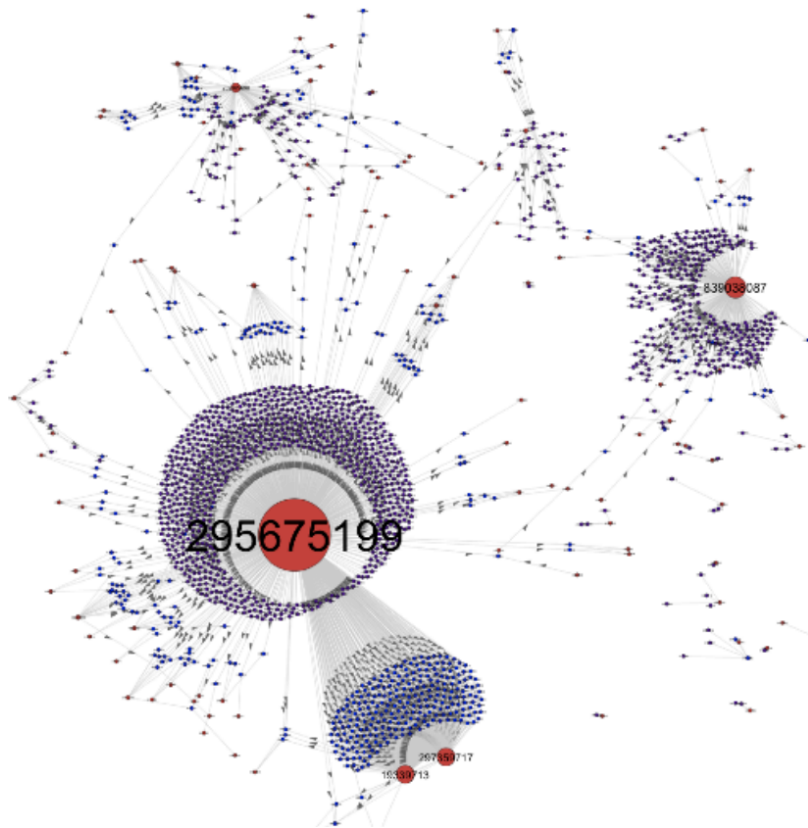
Properties of disinformation

False Information Goes “Viral” Online

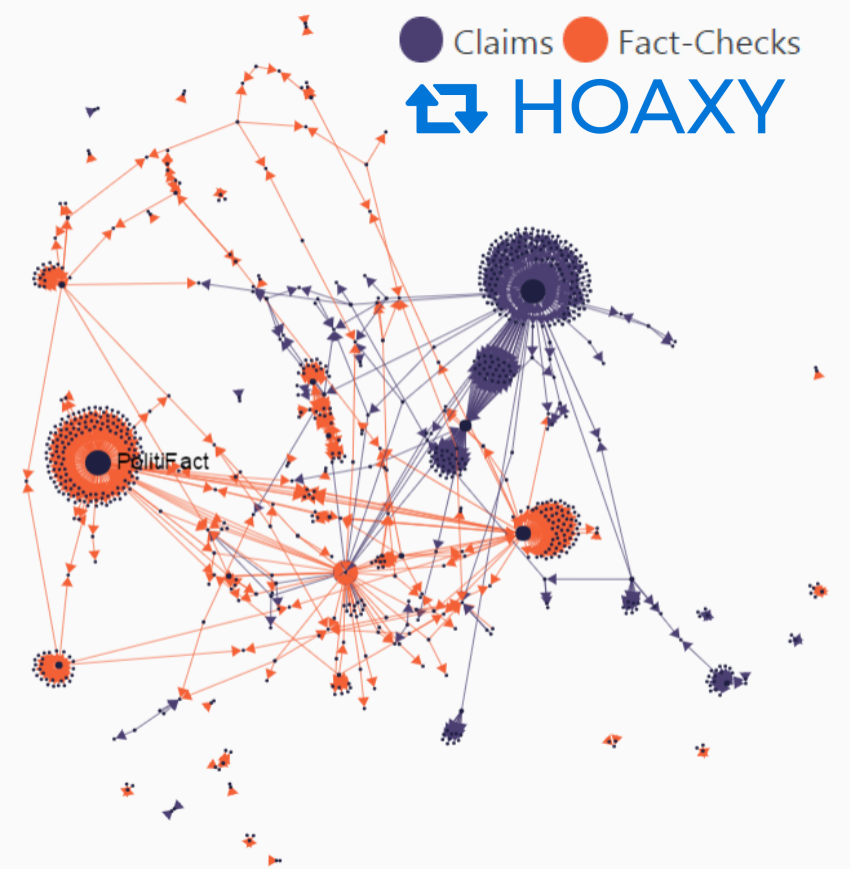
- Many social media users “retweet”, “share”, and “like” these erroneous reports.
- These users include average citizens who don’t fact-check before spreading the news.
- Examples about how hoaxes spread.



False information spreads quickly

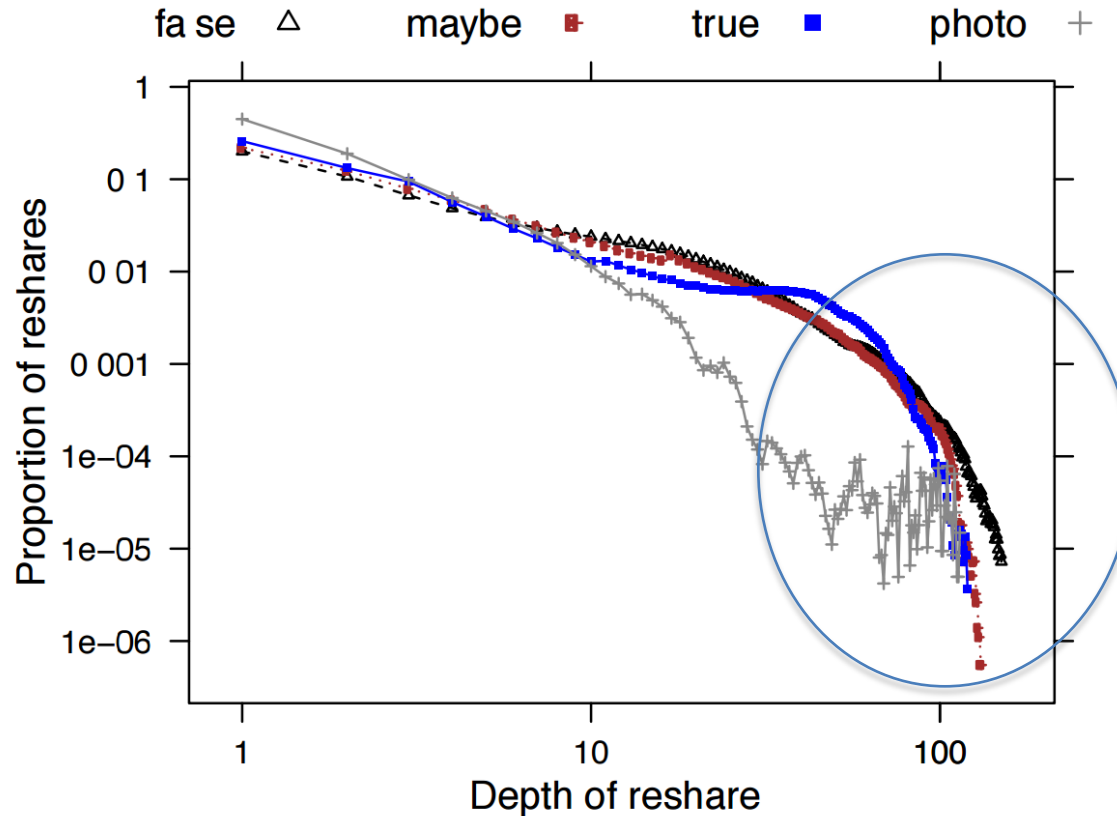


Tweets and retweets for spread of a fake image during first 2 hours



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

False information cascades deep



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

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

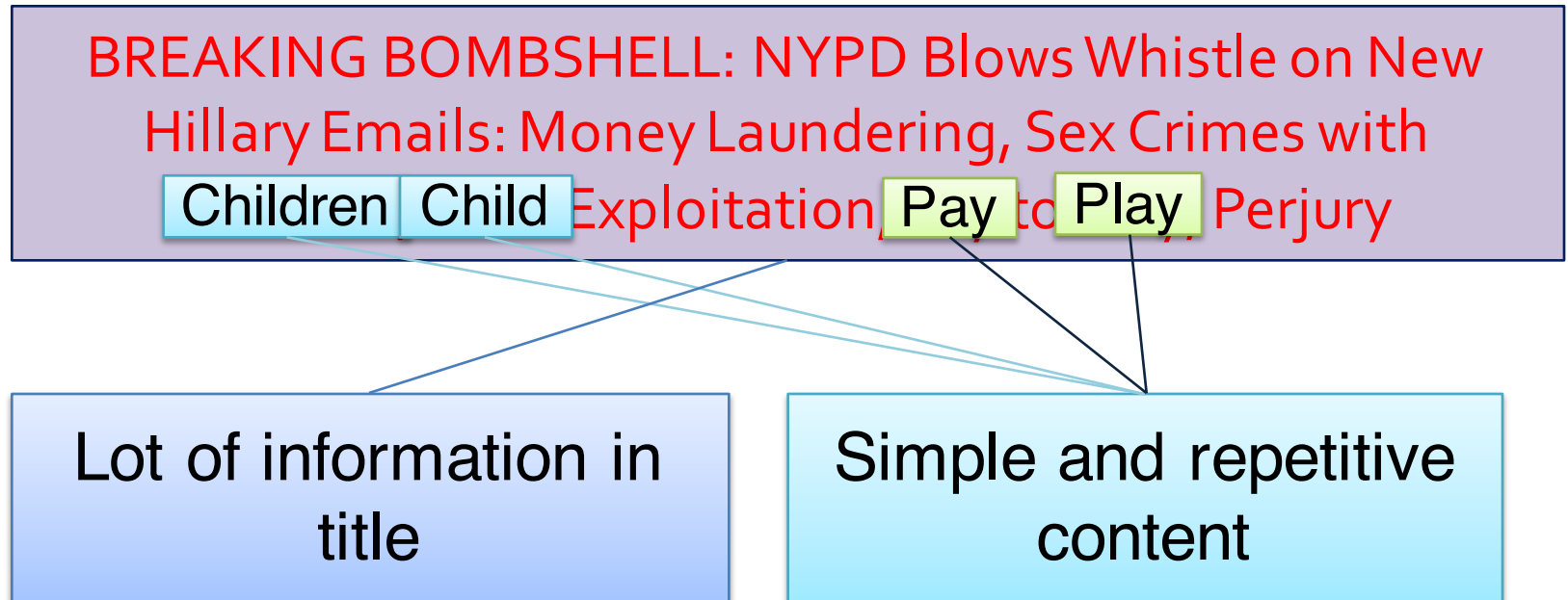
Preexisting Conditions and Republican Plans to Replace Obamacare

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

How is fake news written?



Case study: Disinformation on Wikipedia

Impact of Wikipedia hoaxes

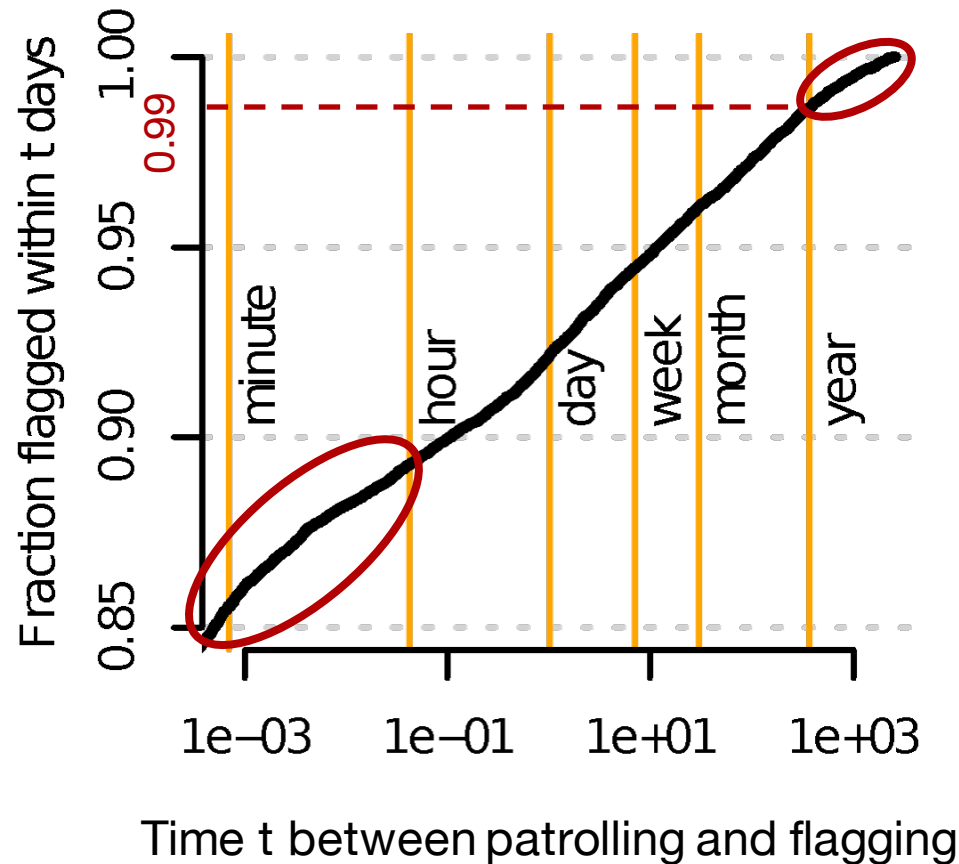
“The worst hoaxes are those which

- (a) last for a long time,
- (b) receive significant traffic,
- (c) are relied upon by credible news media.”

Jimmy Wales on Quora

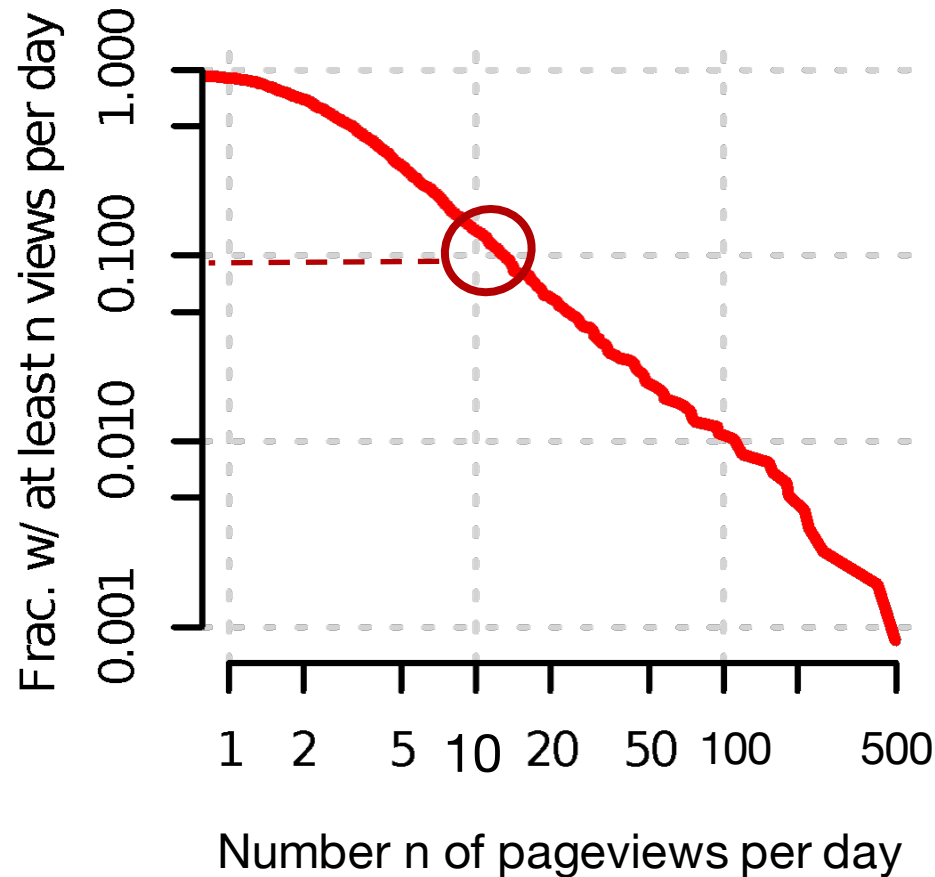
Impact of Wikipedia hoaxes

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



Impact of Wikipedia hoaxes

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



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

Wikipedia Hoaxes

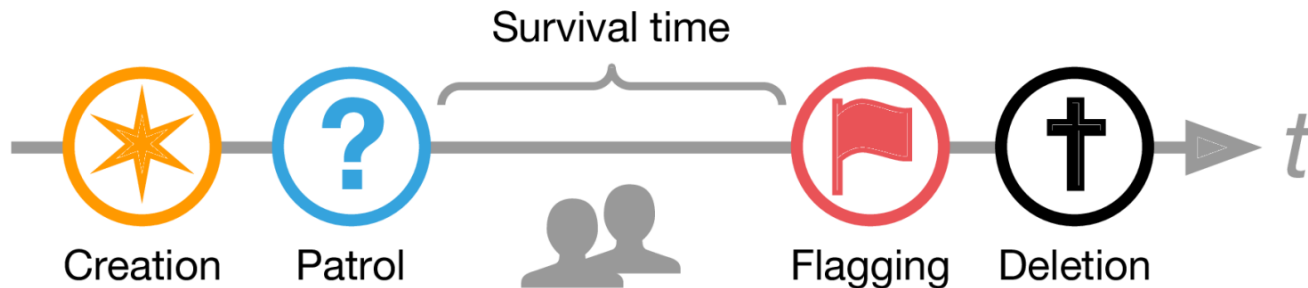
Hoax article vs hoax facts

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:



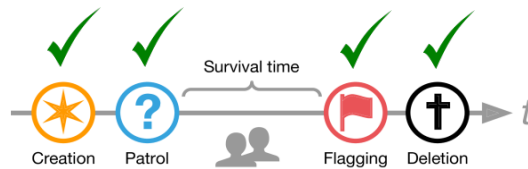
Data: <http://cs.umd.edu/~srijan/hoax/>

Kumar, et al. (WWW 2016)

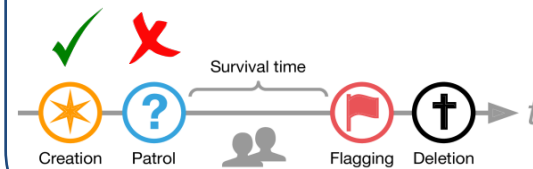
What are Wikipedia hoaxes like?

Hoax

Successful hoax
pass patrol
survive for a month
viewed frequently

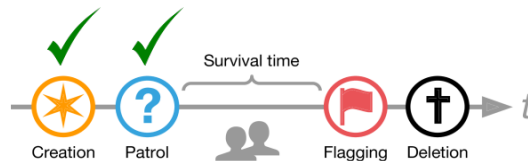


Failed hoax
flagged and
deleted during
patrol



Non-hoax

Legitimate
articles
never flagged



Wrongly flagged
temporarily flagged



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

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.

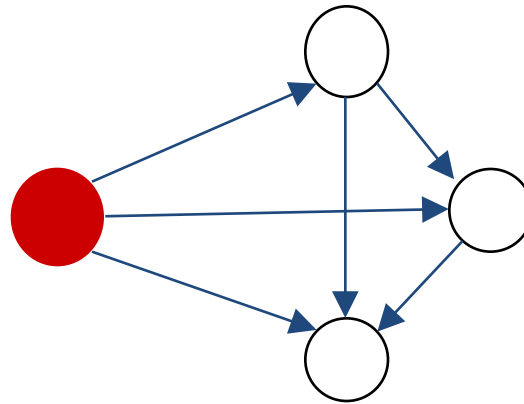
Characteristics of Wikipedia hoaxes

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

Characteristics of Wikipedia hoaxes

Appearance:
hoaxes mostly
have text and
few references.

Link-network:
hoaxes have
incoherent
wikilinks.

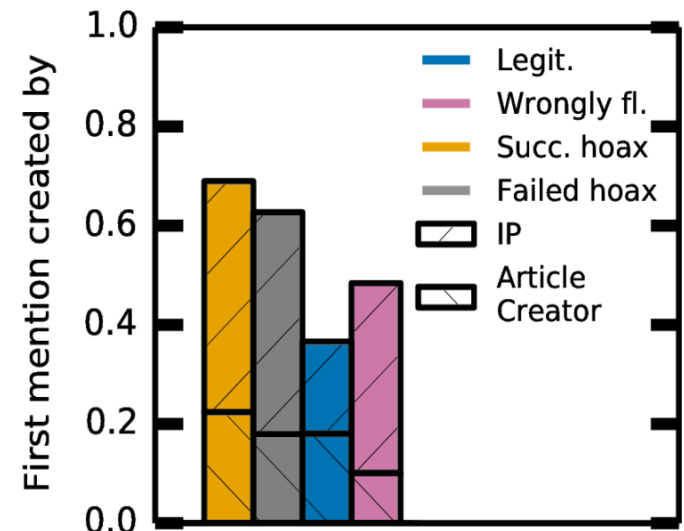
Support:
how other articles
refer to it

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



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

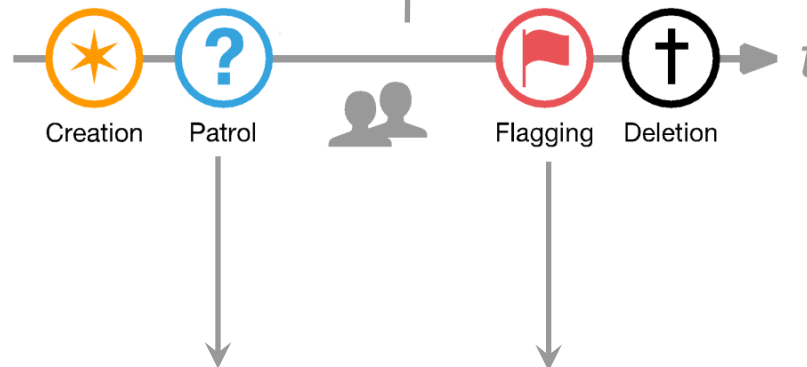
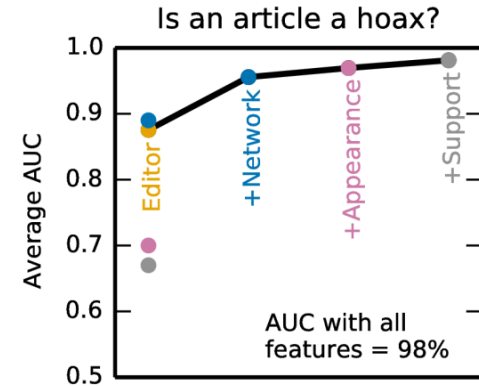
Hoax creators are more recently
registered, and
have lesser editing experience.

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

Identifying real Wikipedia hoaxes

Flagged by us and deleted by Wikipedia administrators

Steve Moertel

American popcorn
entrepreneur

Survived for
6 years 11 months!

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

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.

Can readers identify Wikipedia hoaxes?

320 random hoax and non-hoax pairs

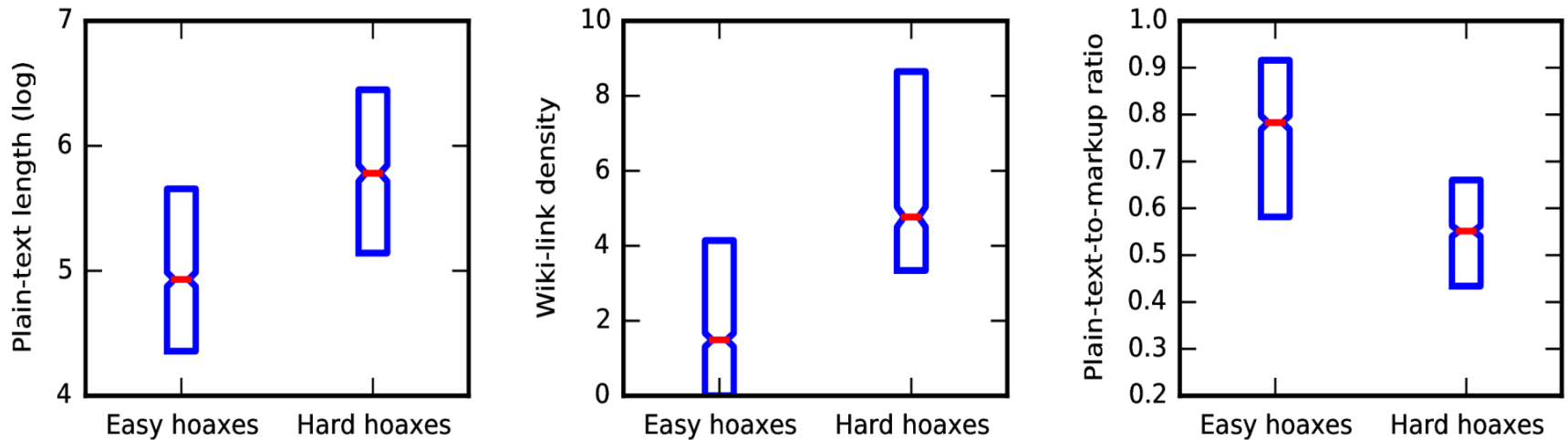
10 raters on Amazon Mechanical Turk rated each pair

Results		
50%	66%	86%
Random	Human	Classifier

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

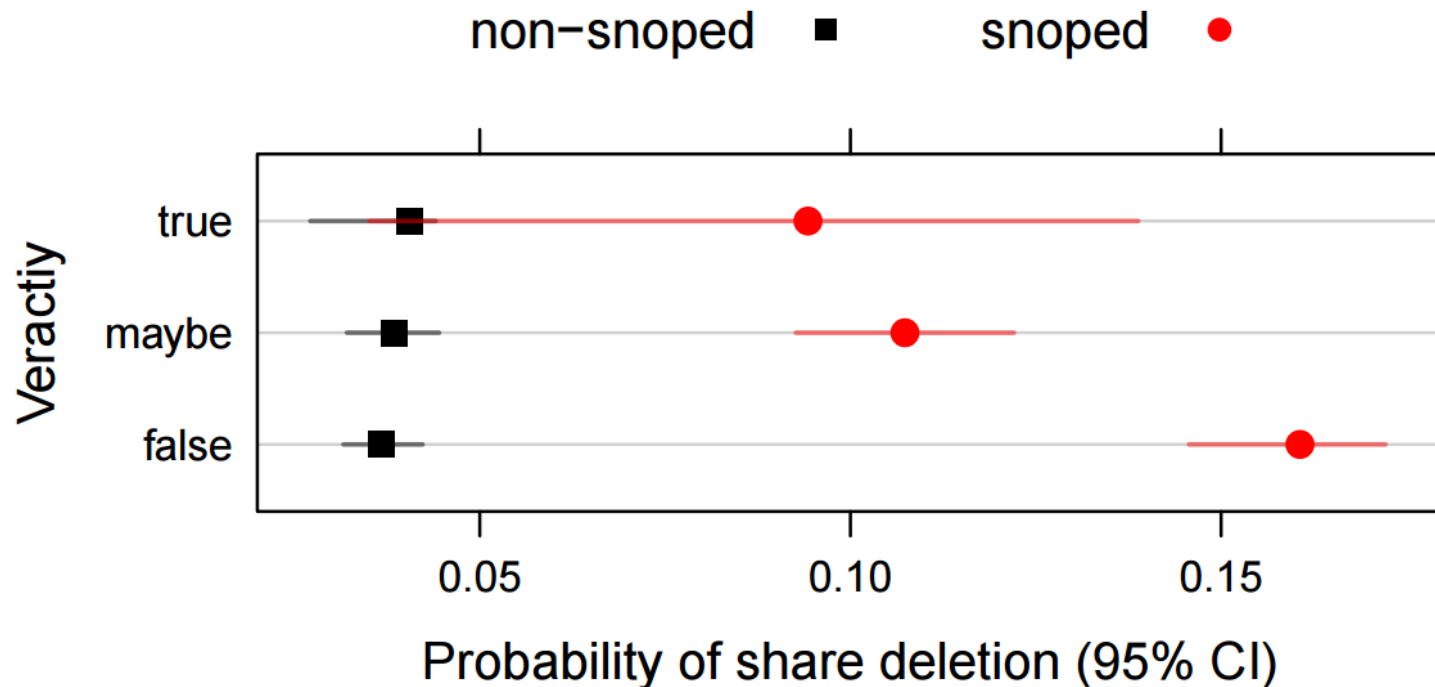
What fools humans?

Comparing easy- vs hard-to-identify hoaxes



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

What happens when false information is pointed out?



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

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

Tutorial link: <http://bit.ly/kdd2017>