



# Malicious Behavior on the Web: Characterization and Detection

Srijan Kumar (@srijankr)

Justin Cheng (@jcccf)

Jure Leskovec (@jure)

Slides are available at <http://snap.stanford.edu/www2017tutorial/>

# Tutorial Outline

## Malicious users

Trolling

Sockpuppets

Vandals

## Misinformation

Fake reviews

Hoaxes

<http://snap.stanford.edu/www2017tutorial>

# Types of false information



Misinformation  
honest mistake

Disinformation  
deliberate lie to mislead

Wikipedia defines “hoax” as  
“deliberately fabricated  
falsehood made to  
masquerade as truth”



NEWS FASHION ▼ TECH ▼ VIDEO ▼ WORLD ▼



Home &gt; News &gt; 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](#) 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 [Jingarr](#) people.

 *This article relating to a myth or legend is a stub. You can help Wikipedia by expanding it.*

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



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](#) (*date missing*)

**Language family**  
[Creole](#)  
• [Balboa Creole French](#)

#### Language codes

**ISO 639-2** [cpf](#)  
**ISO 639-3** –

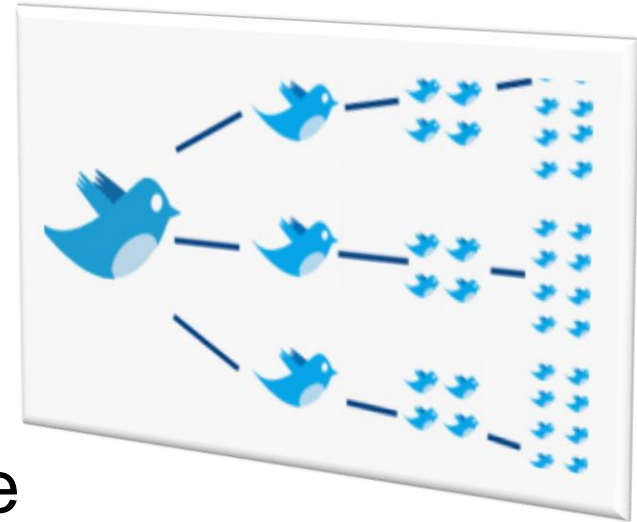
# Properties of disinformation

# False Information Goes “Viral” Online

Many social media users “retweet”, “share”, and “like” these erroneous reports.

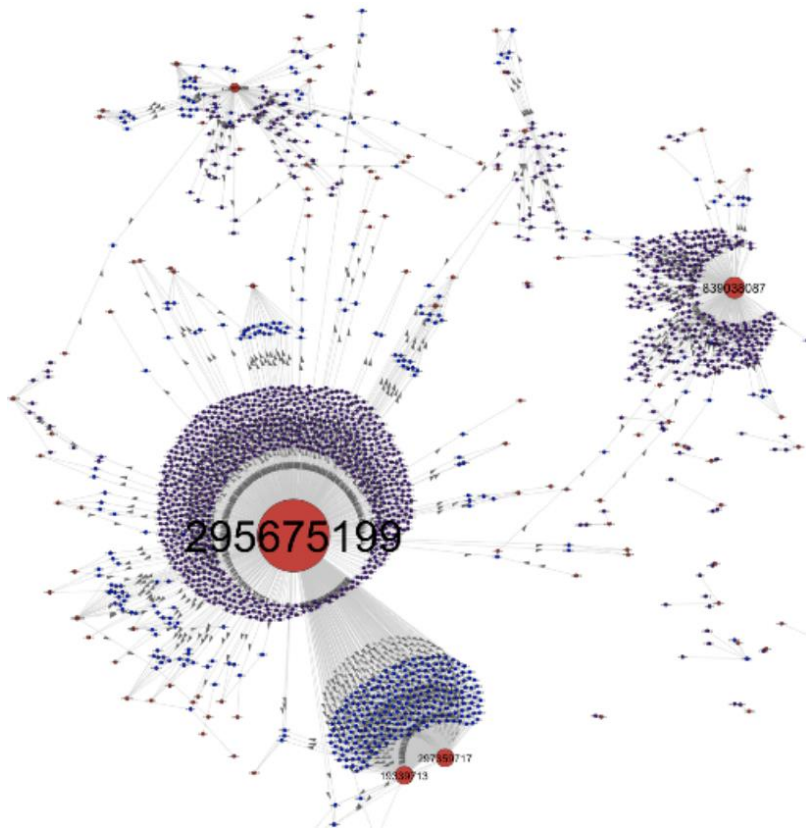
These users include average citizens who don’t fact-check before spreading the news.

Examples about how hoaxes spread.

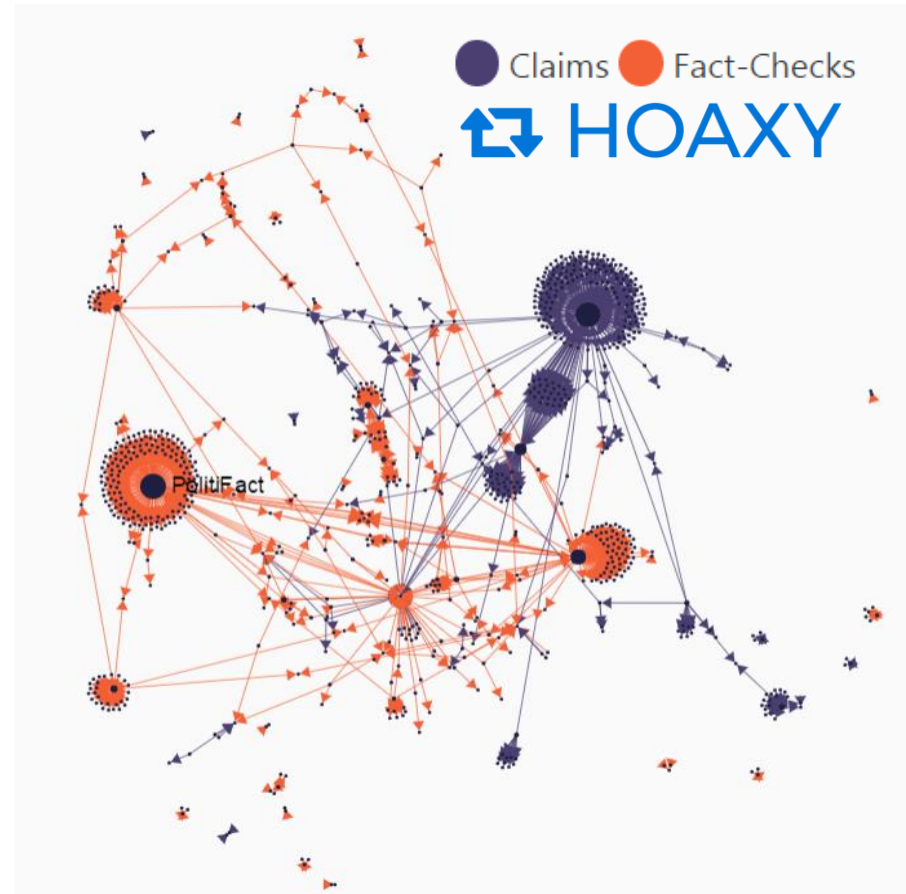




# False information spreads quickly

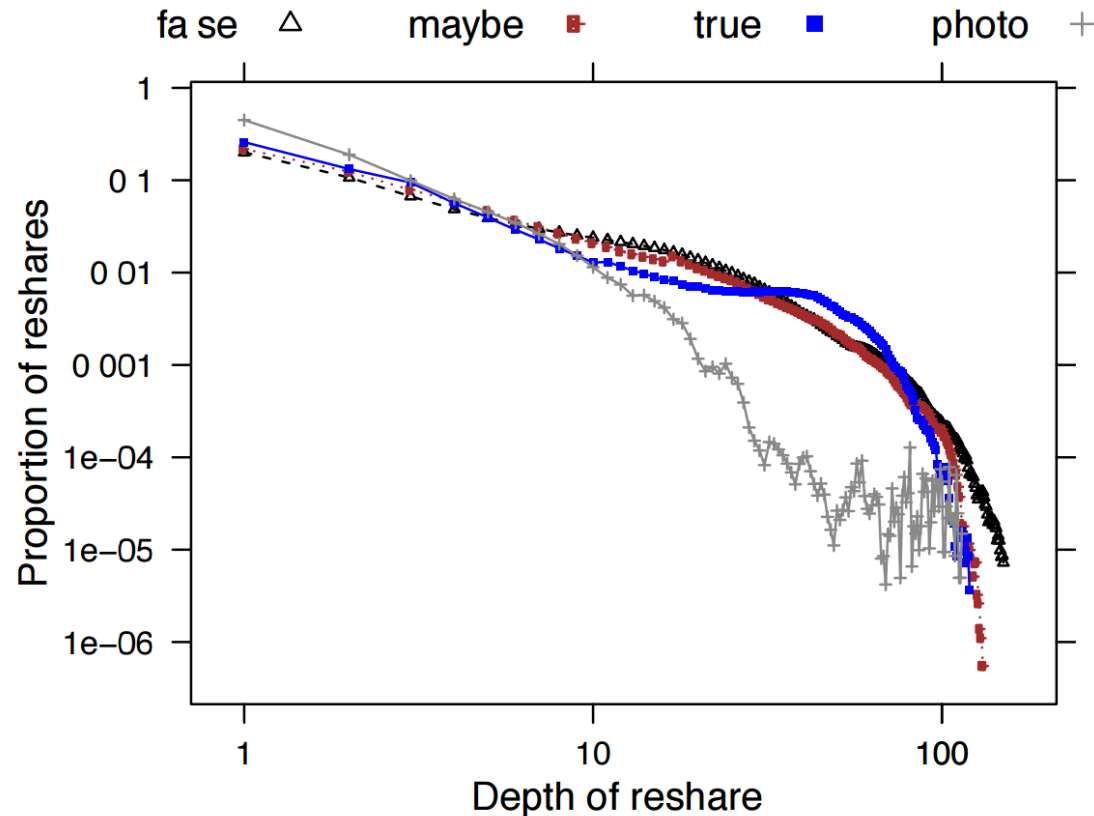


Tweets and retweets for spread of a fake image during first 2 hours



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?

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

The diagram shows a title box at the top with several words highlighted in colored boxes: 'Children' (light blue), 'Child' (light blue), 'Exploitation,' (light blue), 'Pay' (light blue), 'to' (light blue), 'Play' (light blue), and 'Perjury' (light blue). Lines connect these highlighted words to two analysis boxes below. An orange line connects 'Children' to the 'Lot of information in title' box. A light blue line connects 'Child' to the 'Simple and repetitive content' box. A light blue line connects 'Exploitation,' to the 'Simple and repetitive content' box. A light blue line connects 'Pay' to the 'Simple and repetitive content' box. A light blue line connects 'to' to the 'Simple and repetitive content' box. A light blue line connects 'Play' to the 'Simple and repetitive content' box. A light blue line connects 'Perjury' to the 'Simple and repetitive content' box.

Lot of information in  
title

Simple and repetitive  
content



# Case study: Disinformation on Wikipedia

# Impact of Wikipedia hoaxes

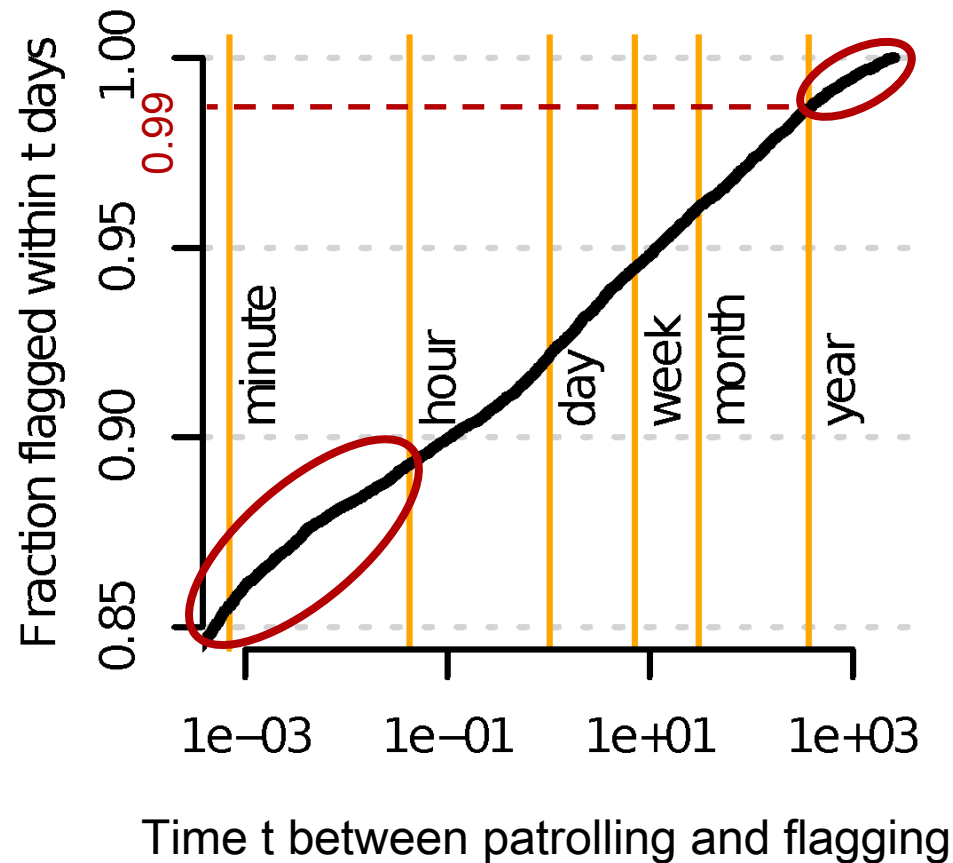
“The worst hoaxes are those which

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

Jimmy Wales on Quora

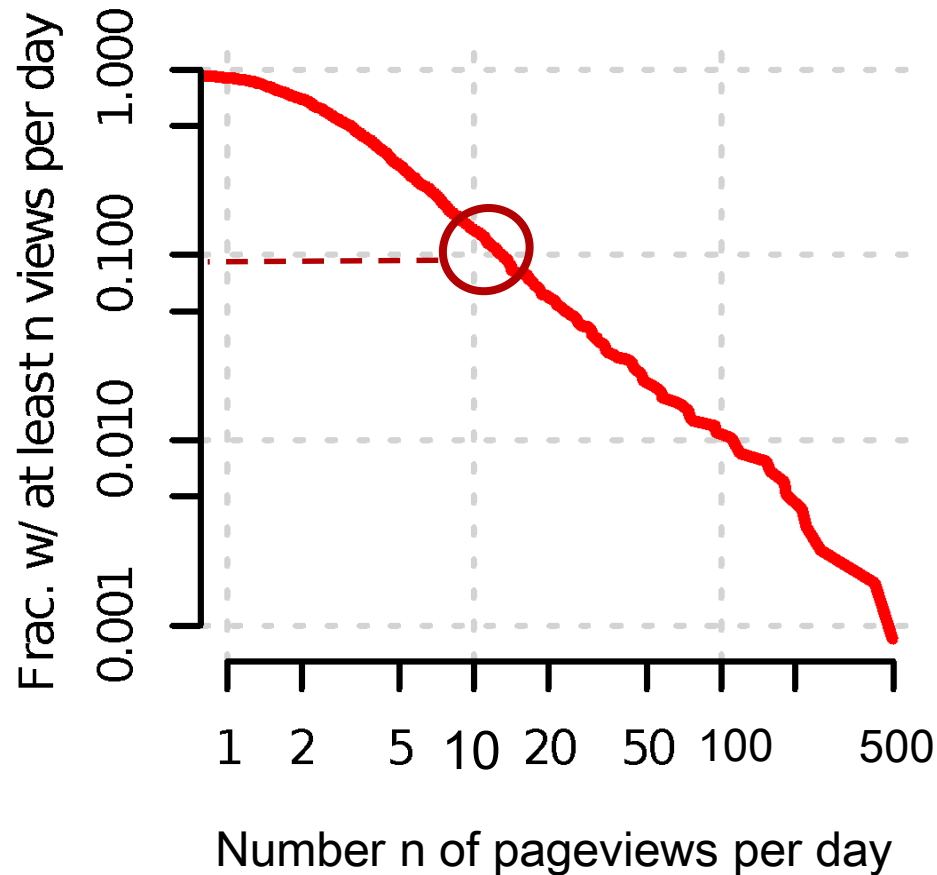
# Impact of Wikipedia hoaxes

“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

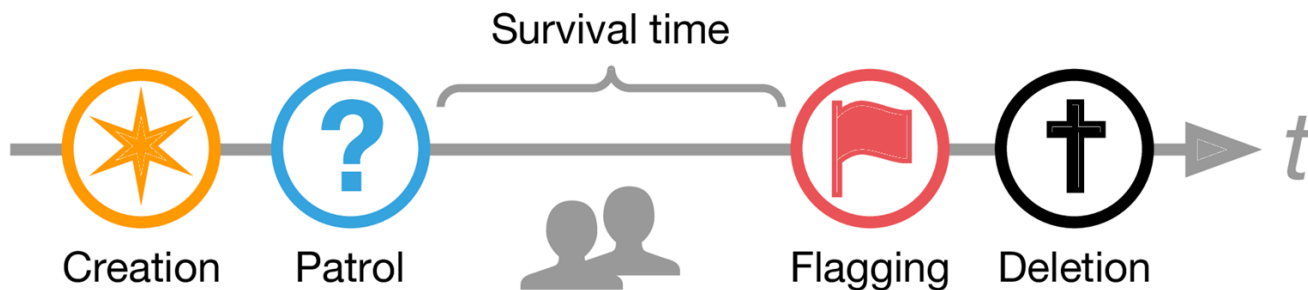
Hoax article vs hoax facts

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:



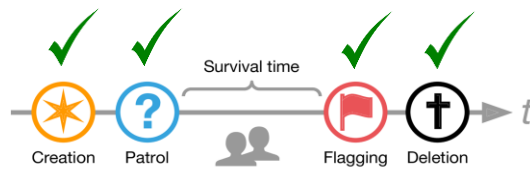
Data: <http://cs.umd.edu/~srijan/hoax/>

Kumar, et al. (WWW 2016)

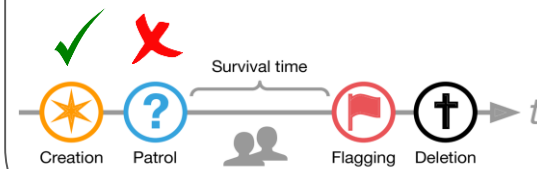
# What are Wikipedia hoaxes like?

Hoax

Successful hoax  
pass patrol  
survive for a month  
viewed frequently



Failed hoax  
flagged and deleted  
during patrol

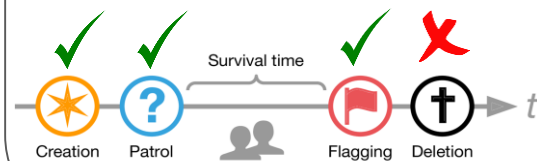


Non-hoax

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.

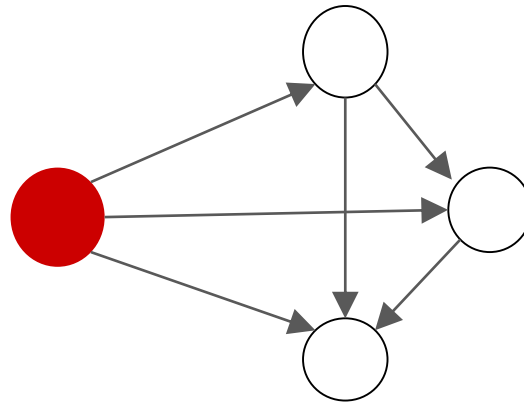
# Characteristics of Wikipedia hoaxes

**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.

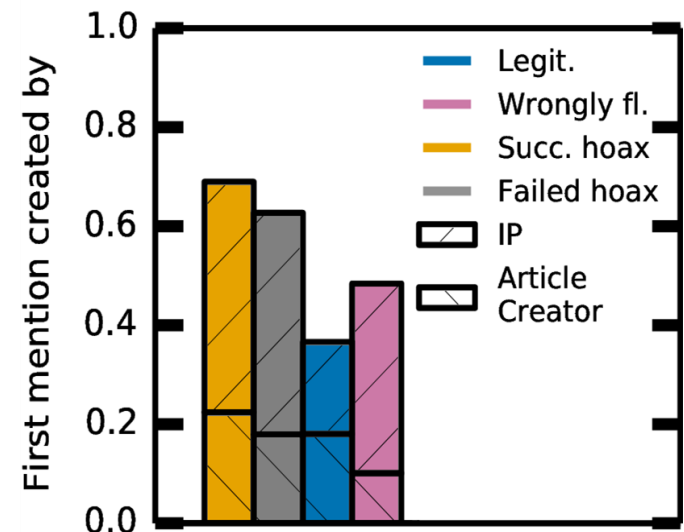
Support:  
how other articles  
refer to it

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

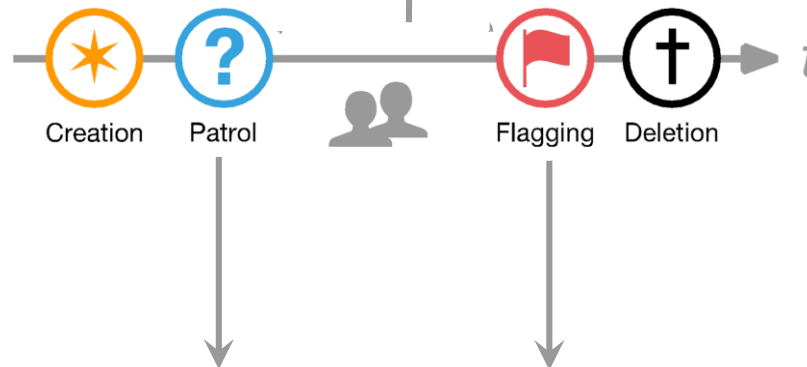
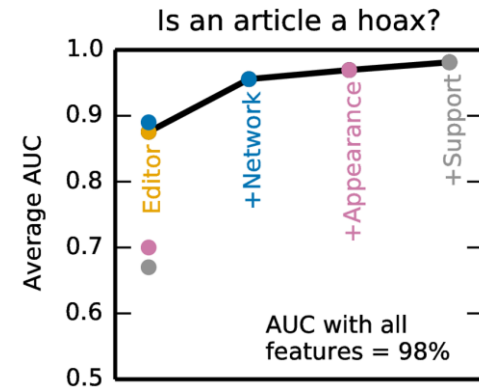
Hoax creators are more recently  
registered, and  
have lesser editing experience.

# 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

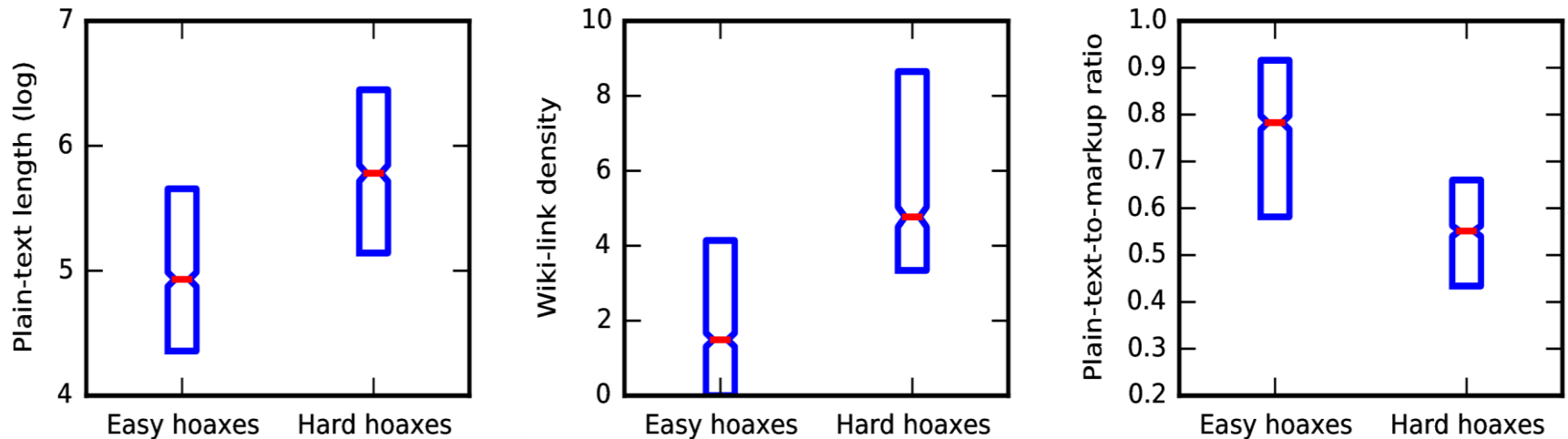
10 raters on Amazon Mechanical Turk rated each pair

Results		
50%	66%	86%
Random	Human	Classifier

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

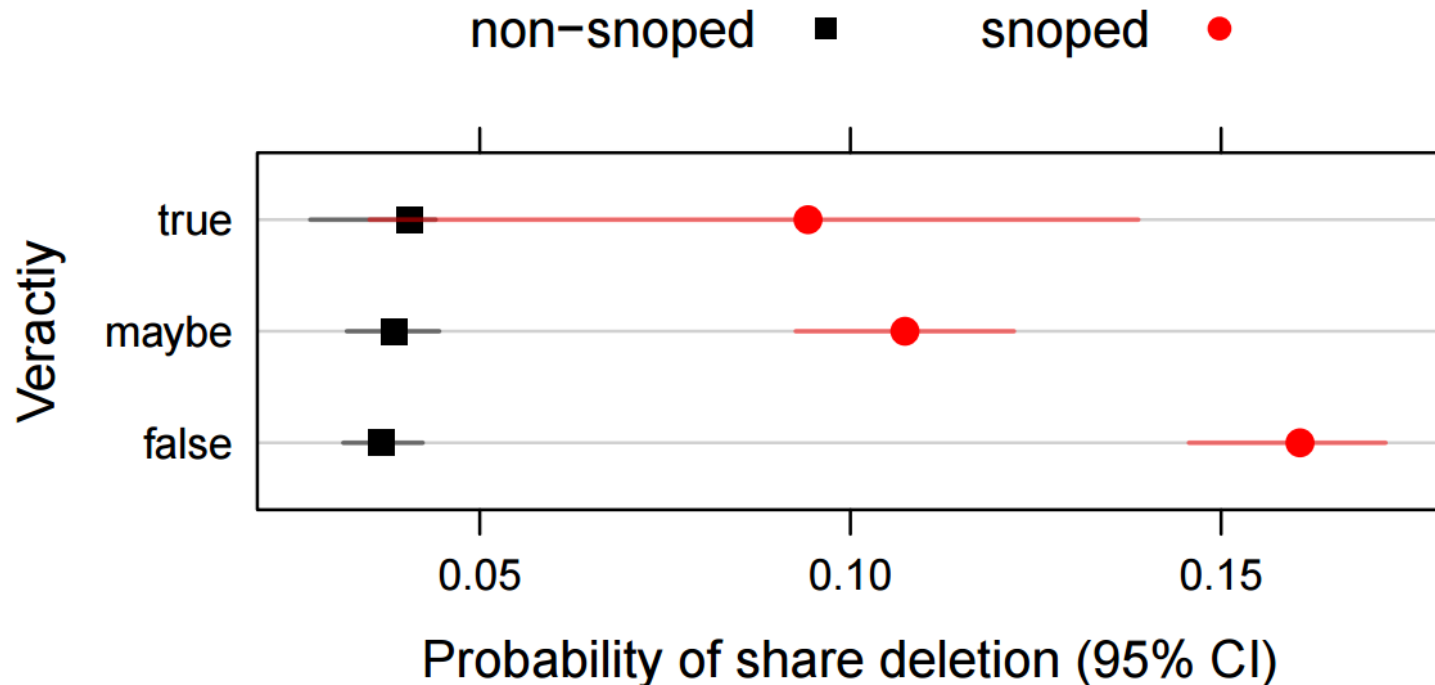
# What fools humans?

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.