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/
Trolling
Sockpuppets
Vandals

Fake reviews
Hoaxes

Types of false information

Misinformation
honest mistake

Disinformation
deliberate lie to mislead

Wikipedia defines “hoax” as “deliberately fabricated falsehood made to masquerade as truth”
Obama Signs Executive Order Banning The National Anthem At All Sporting Events Nationwide

By Jimmy Rustling, ABC News - November 11, 2016

Articles tagged: Fake News

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.

Fact Check > Political News

**IUR 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: List of hoaxes on Wikipedia

In Australian aboriginal mythology, Jar'Edo Wi is a figure with knowledge and physical might, created by Altij landowners to ensure that those who attempt to steal the land get too arrogant or self-conceited. He is associated with the totemic animal the black swan (Cygnus atratus). This article relating to a myth or legend.

Categories: Aboriginal gods | Knowledge god | Myths | Myths of Australia | Traditional Aboriginal mythology | Aborigines | Legends

Balboa French Creole

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 of which are spoken by some members of the Balboa Island community. Balboa French Creole 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, Vouls or Tu'z All, Nos, Il, Elfe, liz or Eilez and Dem. In a census published in 2009, it was revealed only 14 people on the island can still speak the language.
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 tend to be deeper, in that more reshares are at greater depths, than the reference cascades.

Frig erri, 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
How is fake news written?

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

Lot of information in title

Simple and repetitive content

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
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

**Hoax lifecycle:**

Data: [http://cs.umd.edu/~srijan/hoax/](http://cs.umd.edu/~srijan/hoax/)

Kumar, et al. (WWW 2016)
What are Wikipedia hoaxes like?

Successful hoax:
- pass patrol
- survive for a month
- viewed frequently

Failed hoax:
- flagged and deleted during patrol

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

AUC = 71%
Appearance features

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

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)

75% NDCG score of prediction
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% Random
- 66% Human
- 86% 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


