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/

Conclusions and Open Challenges

On the internet…

Anyone can use multiple identities to put something on the internet.

From anywhere in the world.

They can say anything they like.

Leave it there as long as they like.

Change it whenever they feel like it.

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

Open Challenges

How do we design a healthier, more welcoming web?

What tools do we need to build?
What are appropriate user interaction techniques?
How do we change incentives?

Tutorial Summary

Malicious users

- Trolling
- Sockpuppets
- Vandals

Misinformation

- Fake reviews
- Hoaxes

http://snap.stanford.edu/www2017tutorial

Summary: Trolling

- **Trolling:** behavior that does not adhere to community norms
- Trolls are not sociopathic individuals
- Trolling can be induced by bad mood
- Trolling is contagious: discussion context leads to increased bad behavior
- Voting and feedback creates social feedback loops, which lead to downward spirals

Summary: Sockpuppets

- **Sockpuppets:** Usage of multiple accounts, both for benign and malicious intent
- Sockpuppets write worse than non-sockpuppets
- Sockpuppet accounts help each other
- Sockpuppets can vary in how deceptive they are and how supportive they are
- Sockpuppets can be detected from what they post and how they post, but not efficiently from community feedback

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

Summary: Fake Reviewers

- **Fake Reviewers:** Users who write non-truthful reviews for products
- Fake reviews are worse: shorter, more positive, use more “I”s and more verbs and adverbs
- Fake reviewers are deceptive: they collude among themselves and are faster
- Textual, behavioral and network based algorithms can detect fake reviewers
- Combination of several components performs the best

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

Open Challenges

P1. Anonymity

What is the role of anonymity and the lack of single verified identify in antisocial behavior on the internet?

Open Challenges

P2. Early detection

How can antisocial behavior and disinformation be detected as early as possible?

What features can we use?
Can we skip semantic analysis and fact checking?

Open Challenges

P3. Adversarial setting

Bad users can actively change behavior in presence of new detection measures to avoid detection.
How do we deal with this?

Open Challenges

P4. Organized adversaries

How do we detect coordinated attacks on social media?

Open Challenges

P5. Multi-platform malicious behavior

How do antisocial entities behave across several platforms?

Signs of malicious behavior to look out for

- Activity: malicious behavior is often done with “throwaway” and recent accounts
- Temporal: malicious users are often faster
- Linguistic: malicious users are often abusive and more opinionated
- Network: malicious users often collude and are densely connected to each other
- Community feedback: malicious users are harshly treated by other users, but regular negative feedback can be harmful

Datasets

- Wikipedia hoax dataset: www.cs.umd.edu/~srijan/hoax
- Wikipedia personal attack dataset: https://figshare.com/projects/Wikipedia_Talk/16731
- Wikipedia vandals: www.cs.umd.edu/~srijan/vews/
- Wikipedia vandalism: http://wikipapers.referata.com/wiki/List_of_vandalism_datasets
- TAMU Twitter honeypot dataset: http://infolab.tamu.edu/data/
- Twitter synchronized malicious behavior data: http://www.meng-jiang.com/pubs/catchsync-kdd14/catchsync-kdd14-code-and-data.gz
- Amazon, Yelp, TripAdvisor review datasets:
- http://shebuti.com/collective-opinion-spam-detection/
- http://cs.unm.edu/~aminnich/trueview/
- https://www.cs.uic.edu/~liub/FBS/fake-reviews.html
- http://snap.stanford.edu/data/#reviews

End of Part 2

Malicious Behavior on the Web: Characterization and Detection

Srijan Kumar (@srijankr)
Justin Cheng (@jcccf)
Jure Leskovec (@jure)

http://snap.stanford.edu/www2017tutorial