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

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
- **Lockstep behavior**:

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?

Subrahmanian et al. (Science 2017)

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?

Subrahmanian et al. (Science 2017)

Open Challenges

P4. Organized adversaries

How do we detect coordinated attacks on social media, as opposed to lone wolf attacks?

Subrahmanian et al. (Science 2017)

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

Let us know if you have one too!

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