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

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

WIKIPEDIA

Easy to add content

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

More images

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

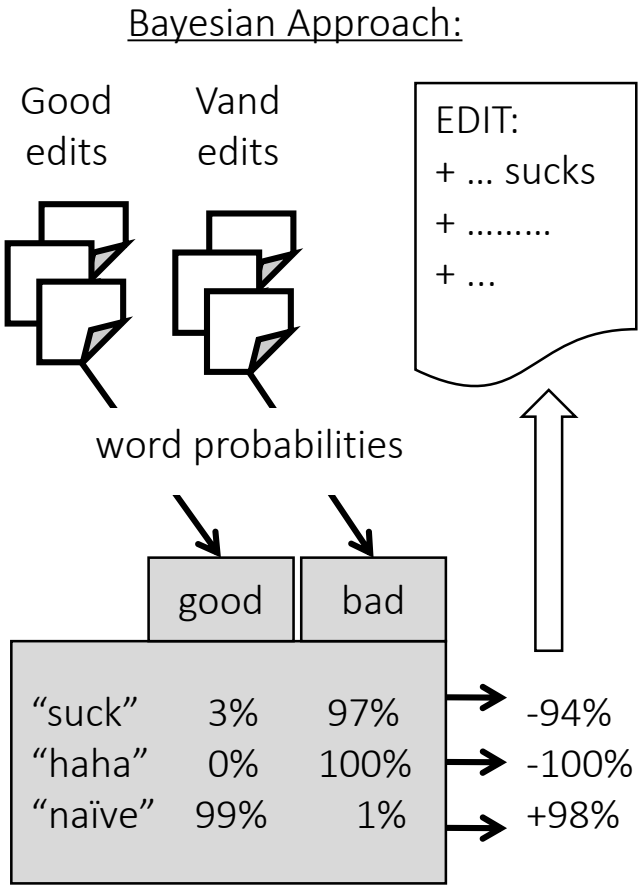
REVISION COMMENT

length, section-edit?

TIMESTAMP

time-of-day, day-of-week

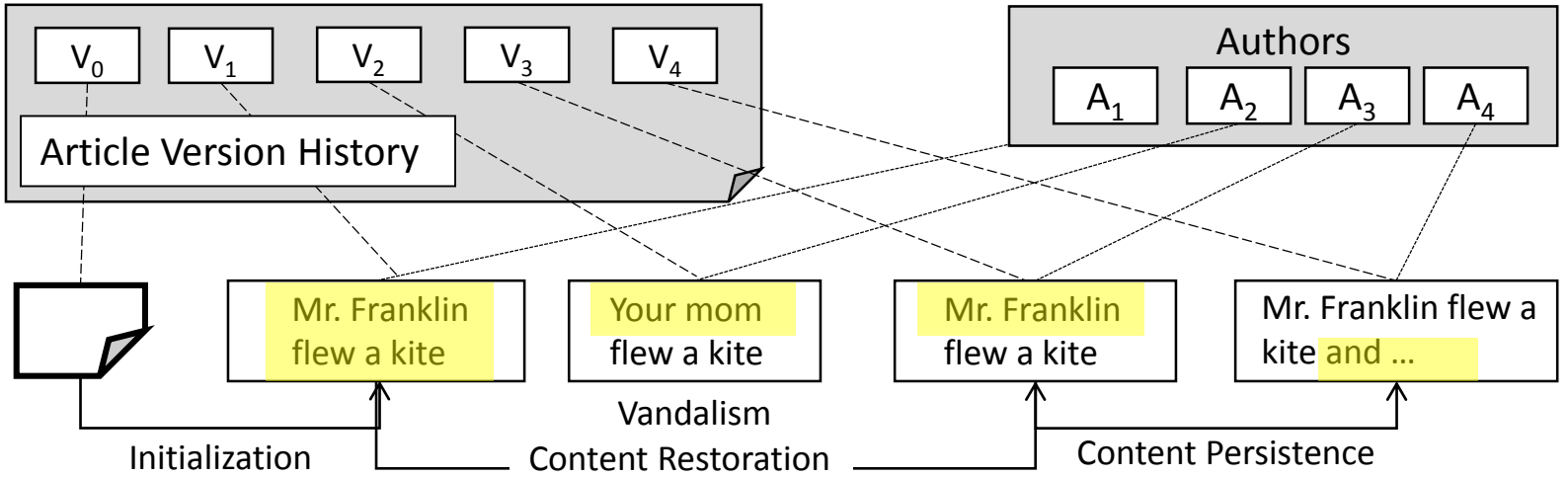
ClueBot NG: Textual



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

Valesco et al. (CLEF 2010)

WikiTrust: Content driven



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

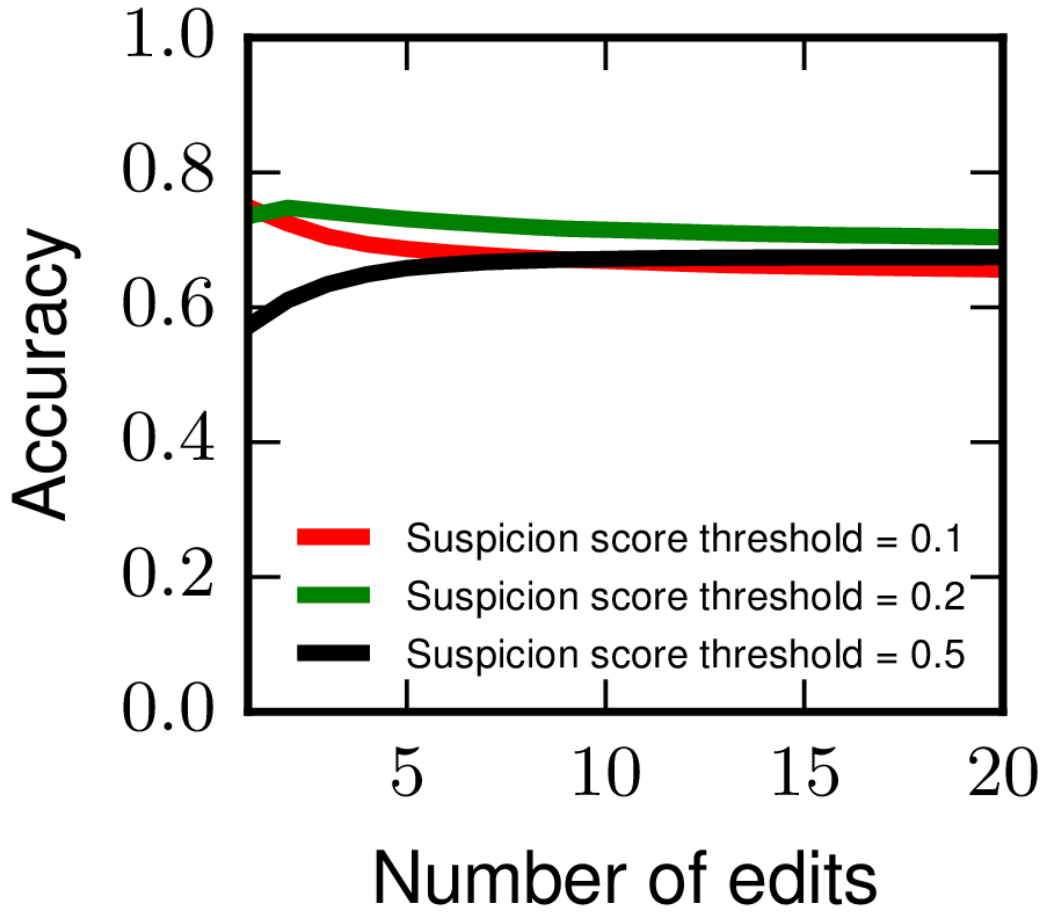
Adler et al. (WWW, 2007)

Detection of vandals

Vandalism detection

Vandal detection

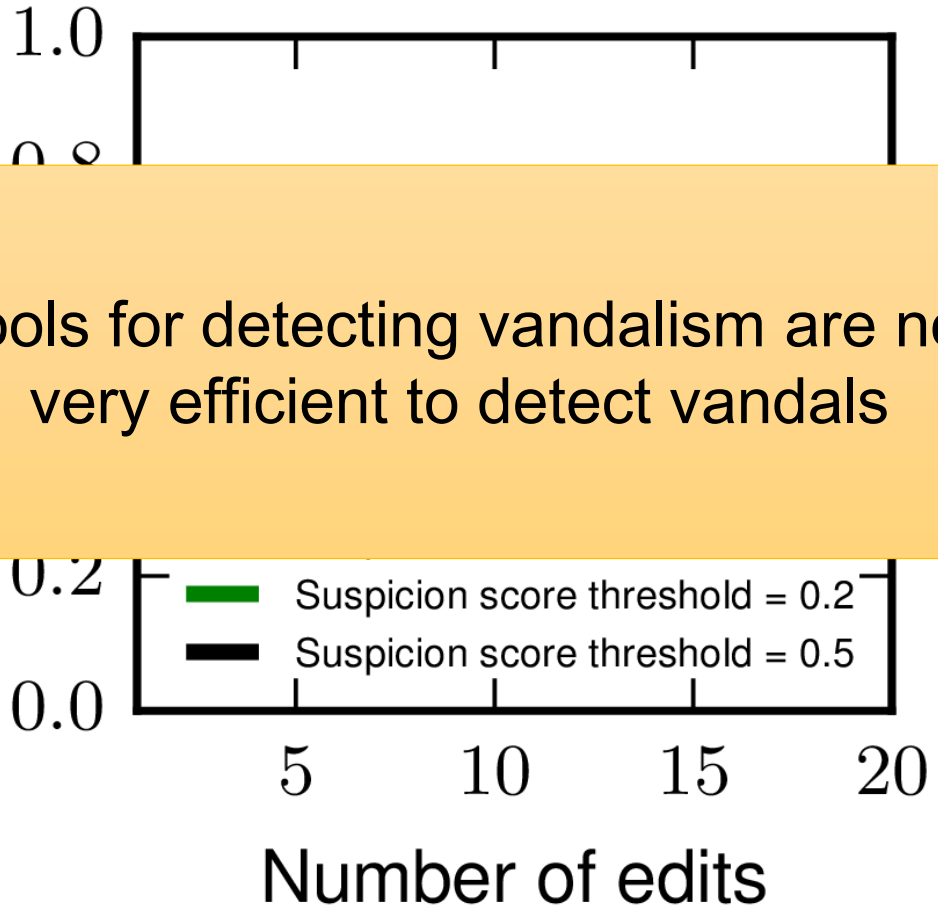
Using STiki to detect vandals



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

Kumar et al. (KDD, 2015)

Using STiki to detect vandals

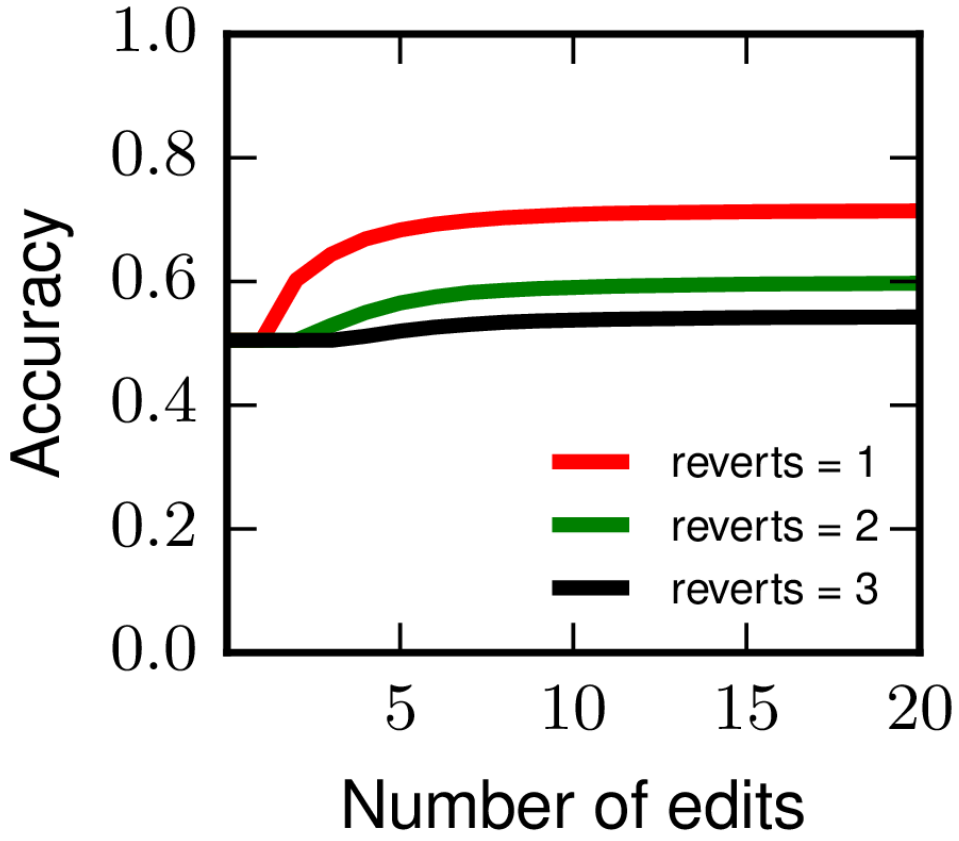


Tools for detecting vandalism are not very efficient to detect vandals

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

Kumar et al. (KDD, 2015)

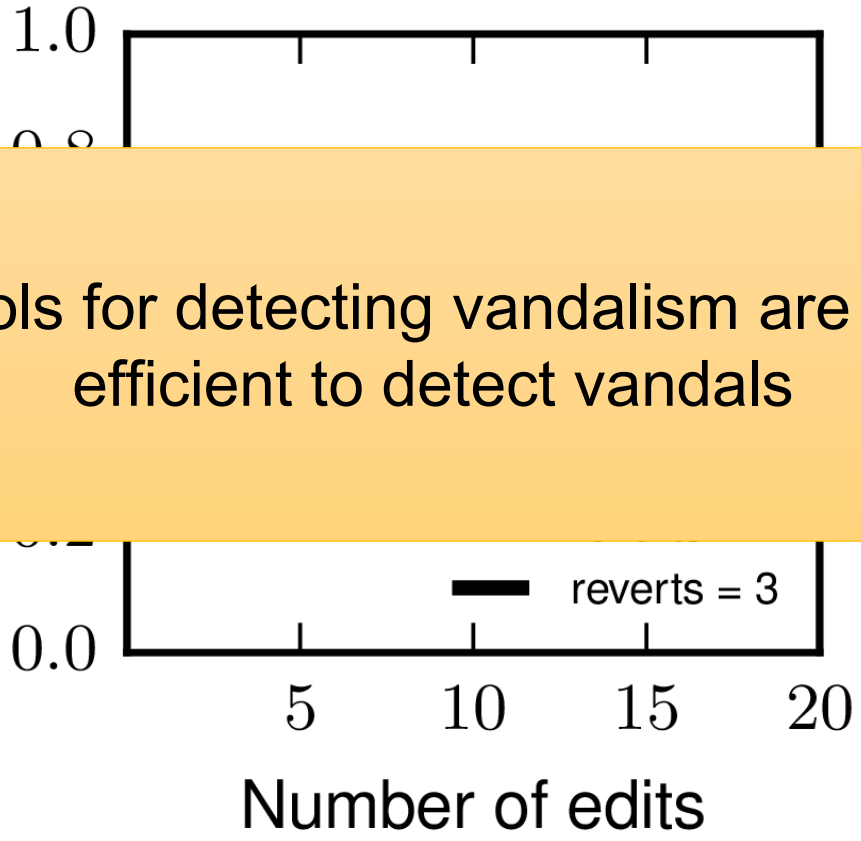
Using ClueBot NG to detect vandals



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

Kumar et al. (KDD, 2015)

Using ClueBot NG to detect vandals



Tools for detecting vandalism are not efficient to detect vandals

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

Kumar et al. (KDD, 2015)

Objective:
Detect vandals in as few edits as possible

Data: Wikipedia Vandals

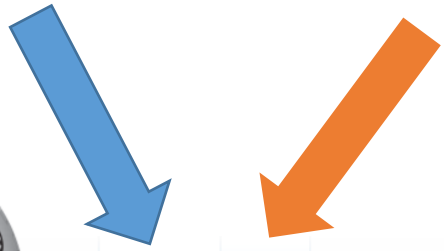
34,000 Editors Half are vandals

770,000 Edits 160,000 edits by vandals

Time: Jan 2013 - July 2014

Kumar et al. (KDD, 2015)

Characteristics of vandals



Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history Search Wikipedia

WIKIPEDIA
The Free Encyclopedia

- Main page
- Contents
- Featured content
- Current events
- Random article
- Donate to Wikipedia
- Wikipedia store

Interaction

Perth

From Wikipedia, the free encyclopedia

Coordinates: 31°57′8″S 115°51′32″E

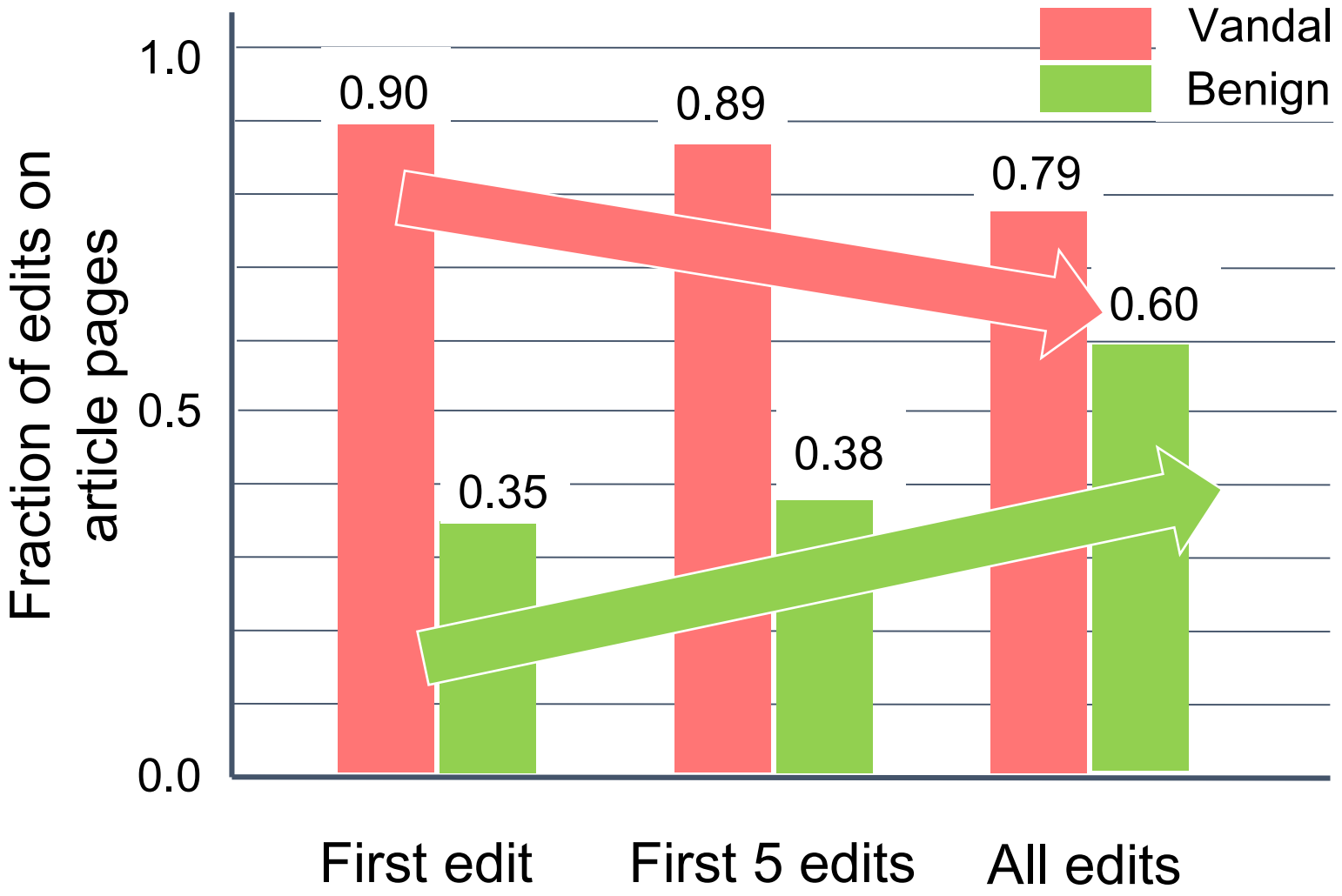
This article is about the capital of Western Australia. For the city in Scotland, see Perth, Scotland. For other uses, see Perth (disambiguation).

Perth (/ˈpɜːrθ/) is the capital and largest city of the Australian state of Western Australia. It is the fourth-most populous city in Australia, with an estimated population of 2.06 million (as of 30 June 2016) living in

Perth
Western Australia

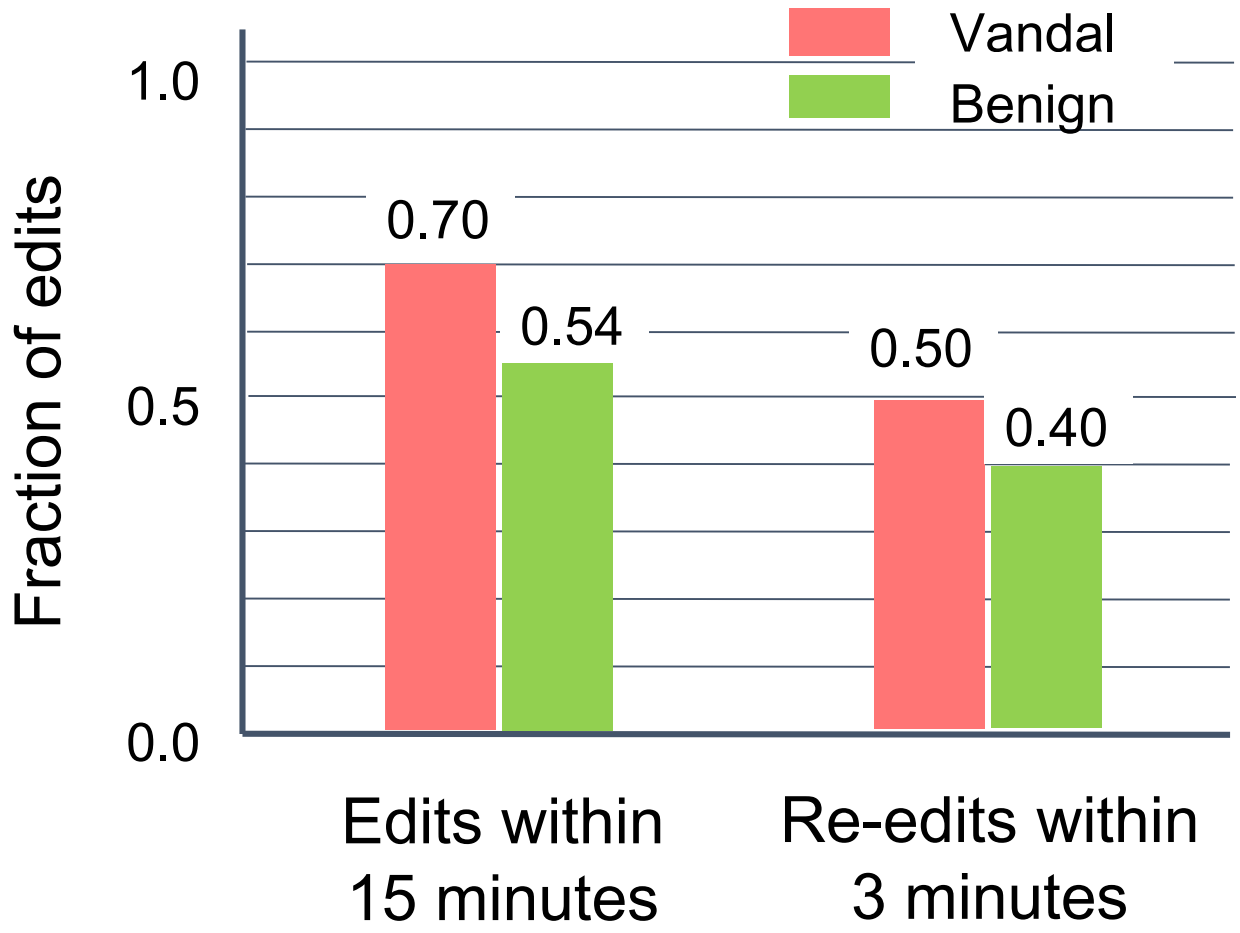
Editors can edit article pages and talk pages

Vandals make visible edits



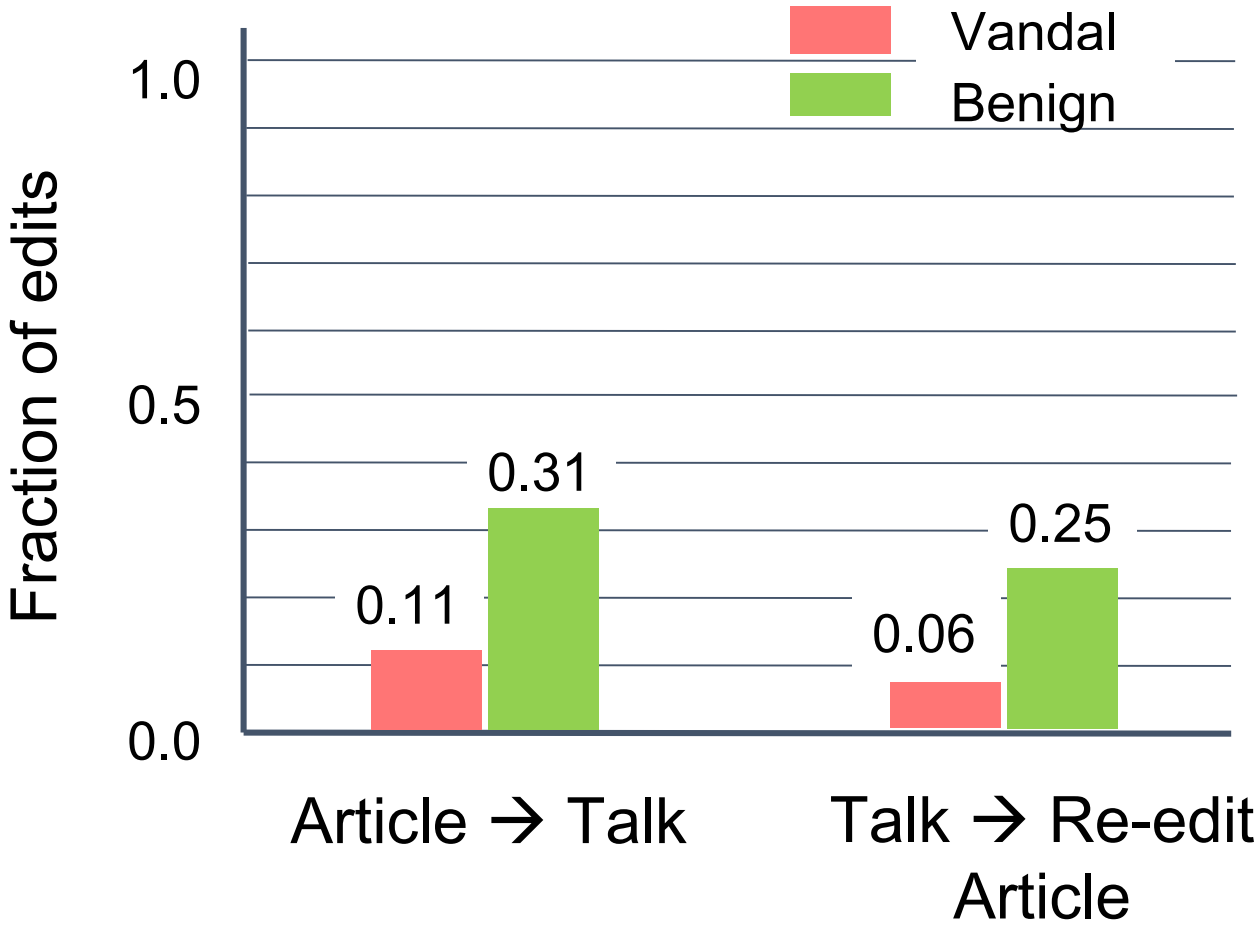
Kumar et al. (KDD, 2015)

Vandals are quicker



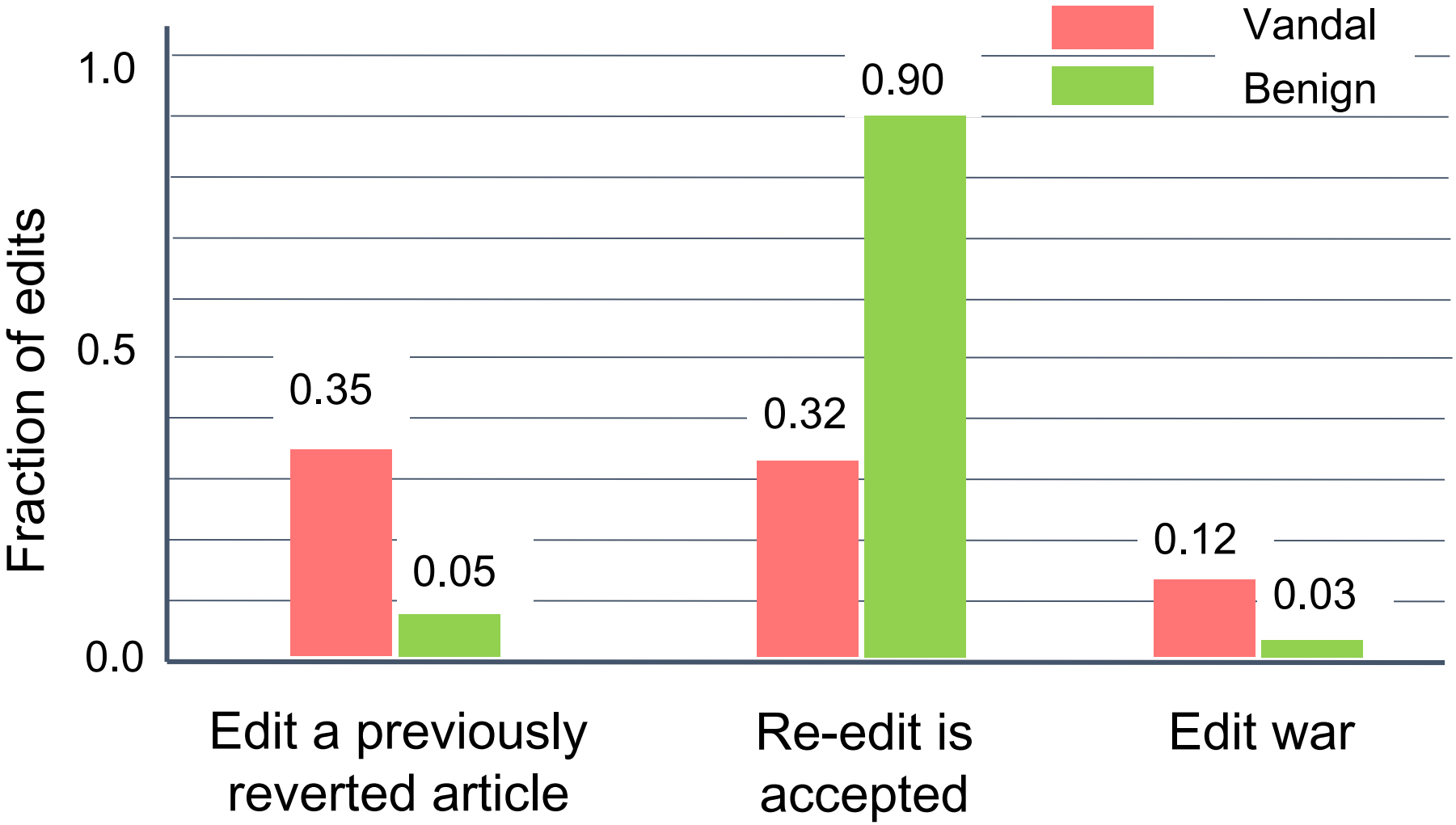
Kumar et al. (KDD, 2015)

Vandals do not discuss



Kumar et al. (KDD, 2015)

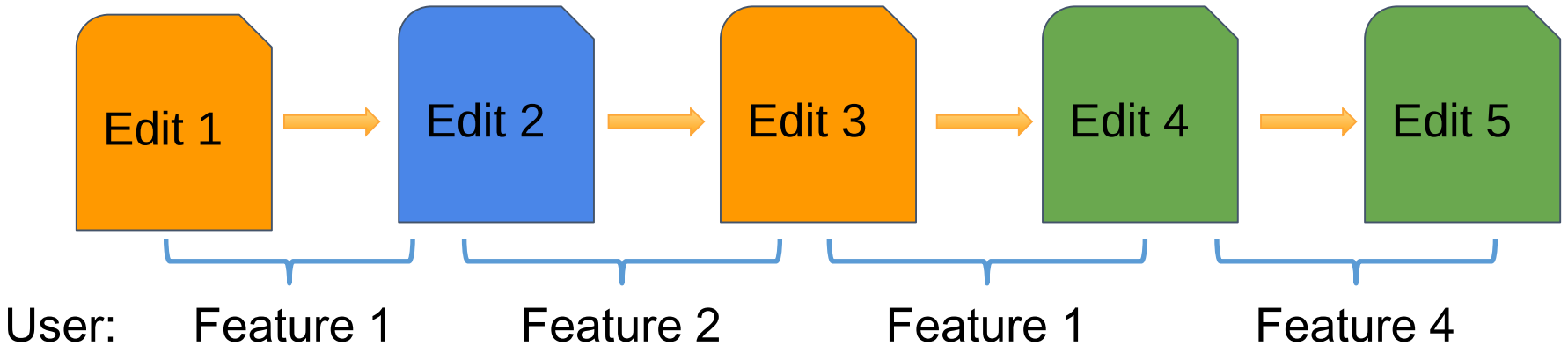
Vandals make reversion driven edits



Kumar et al. (KDD, 2015)

Detecting vandals

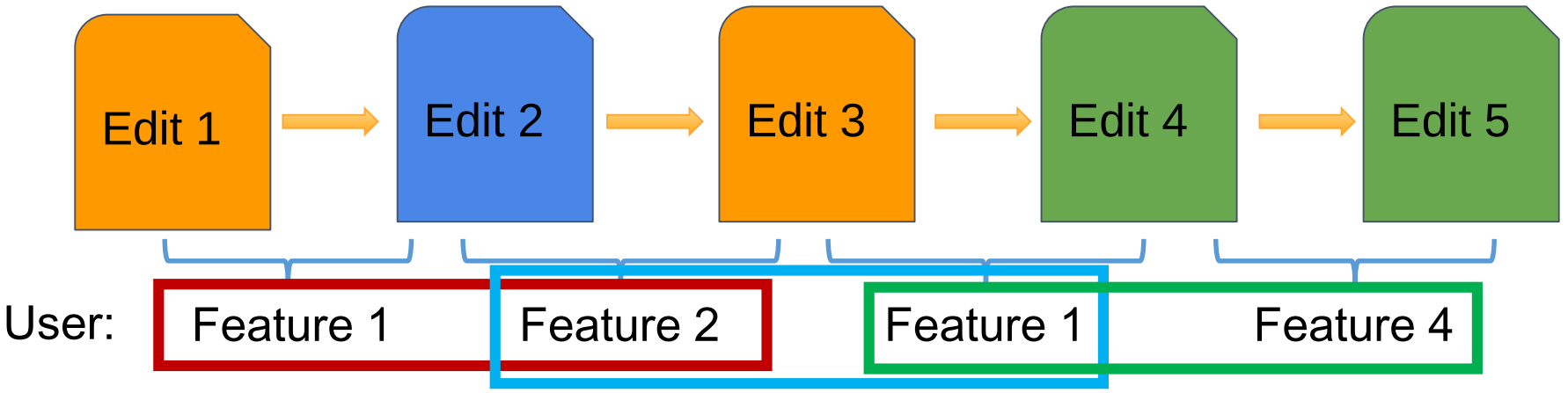
Pairwise Edit Features



Time x Type of page x First edit x Distance x Similarity x Reverted or not

Kumar et al. (KDD, 2015)

Meta-Features: Transitions

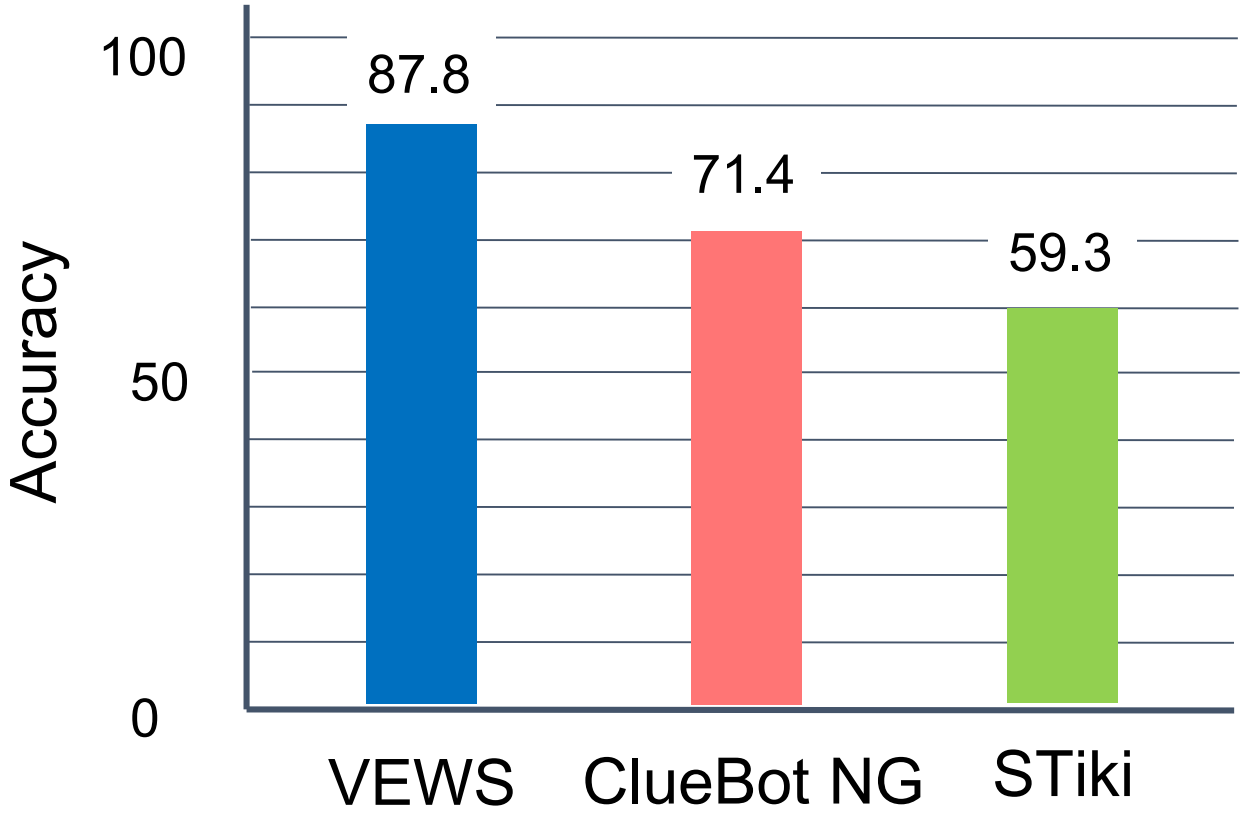


	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5
Feature 1		1		1	
Feature 2	1				
Feature 3					
Feature 4					
Feature 5					

N x N

Kumar et al. (KDD, 2015)

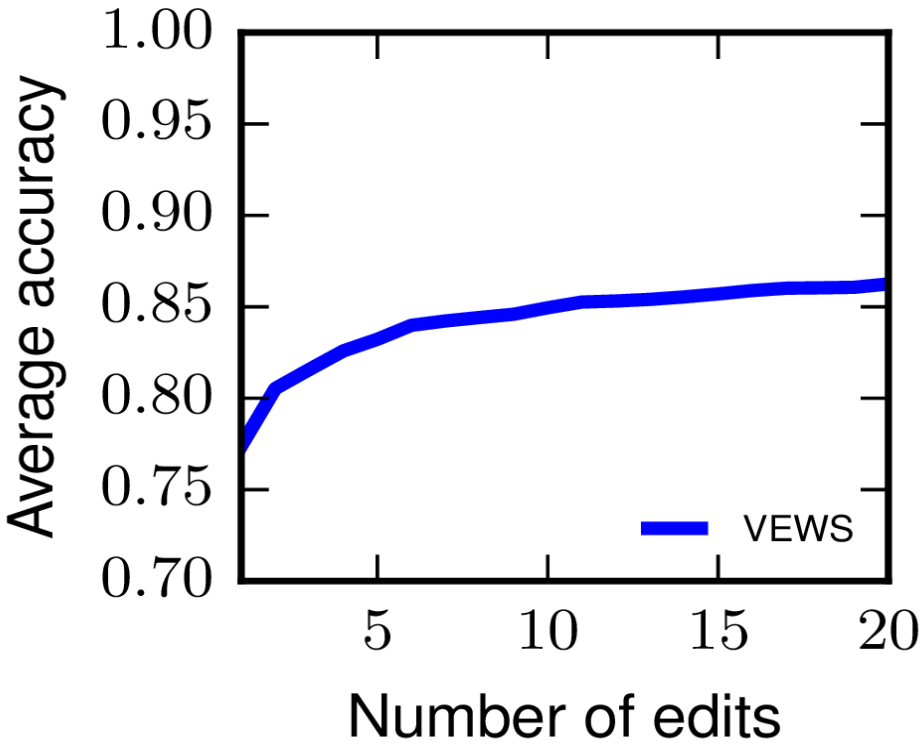
Vandal Detection



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

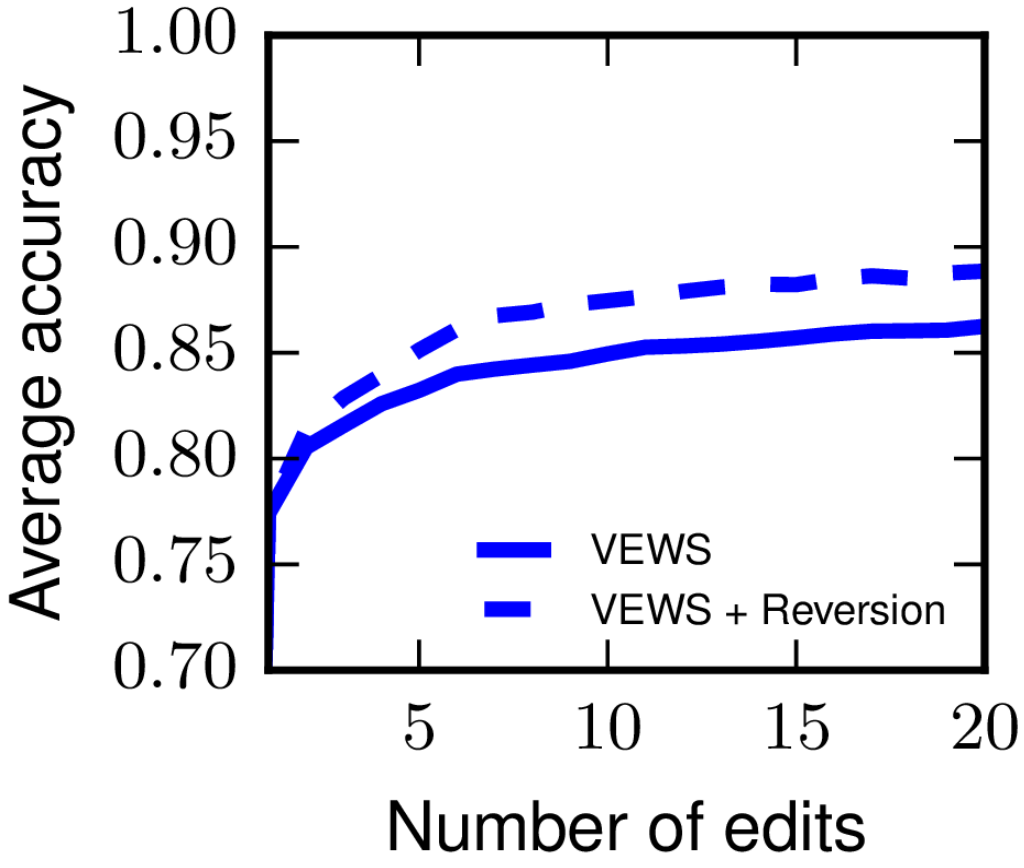
Kumar et al. (KDD, 2015)

Early Warning System

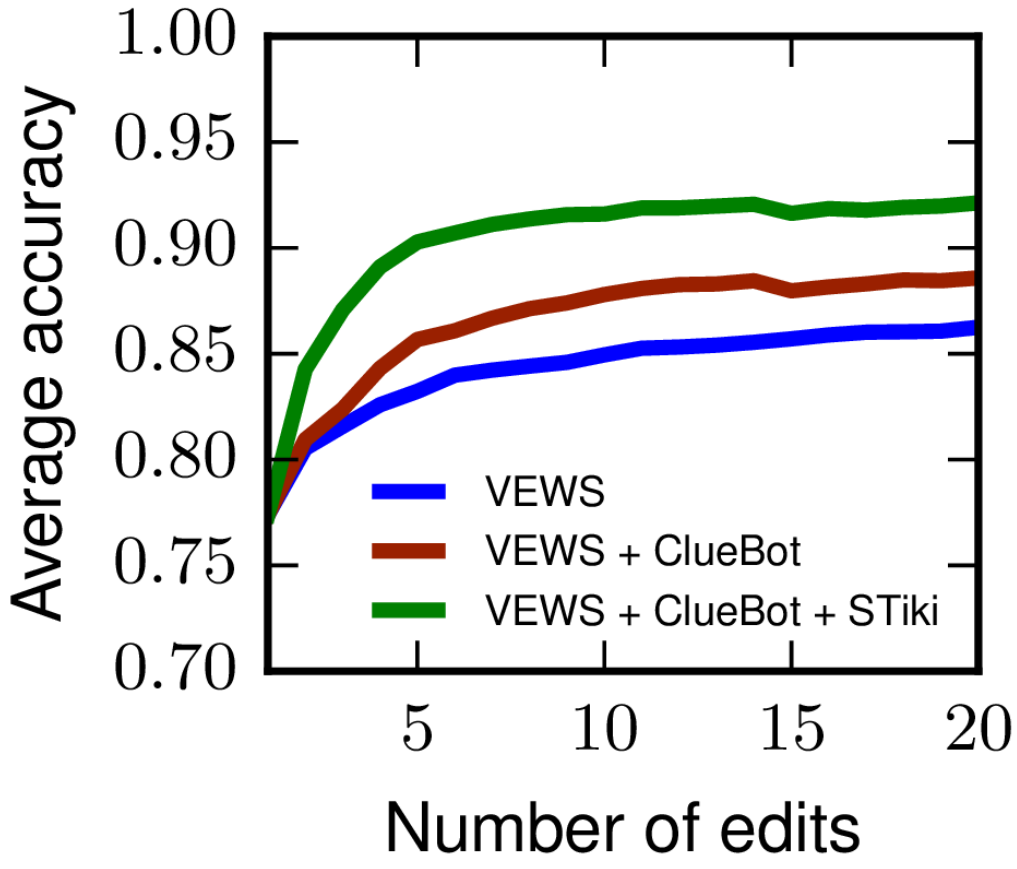


VEWS identifies vandals in 2.13 edits on average

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

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.