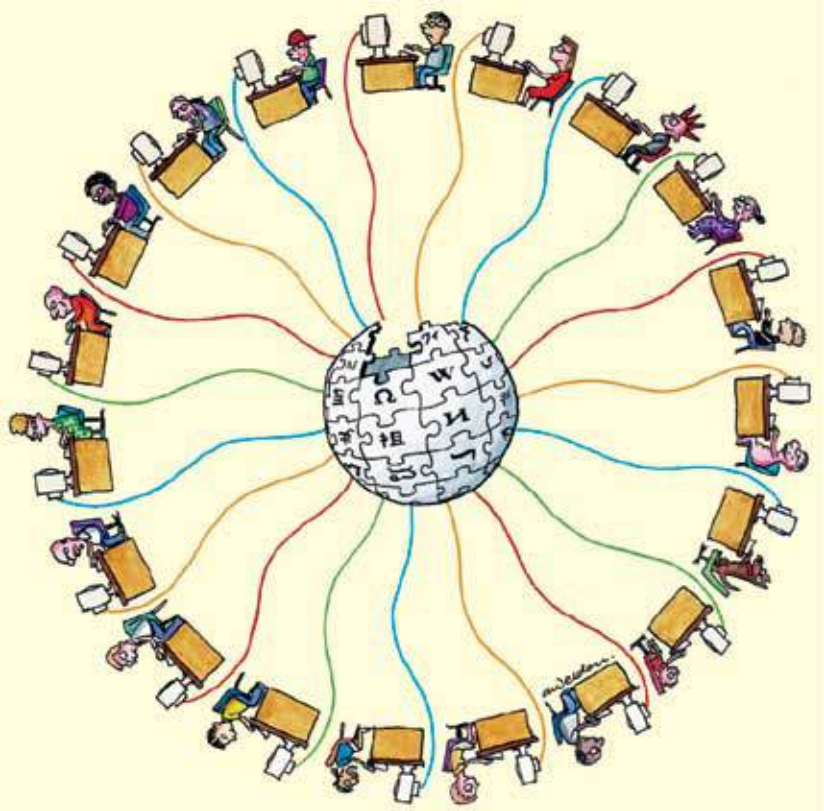
VEWS: A Wikipedia Vandal Early Warning System

Srijan Kumar, Francesca Spezzano, V.S. Subrahmanian
University of Maryland, College Park

LCCD Laboratory for Computational Cultural Dynamics

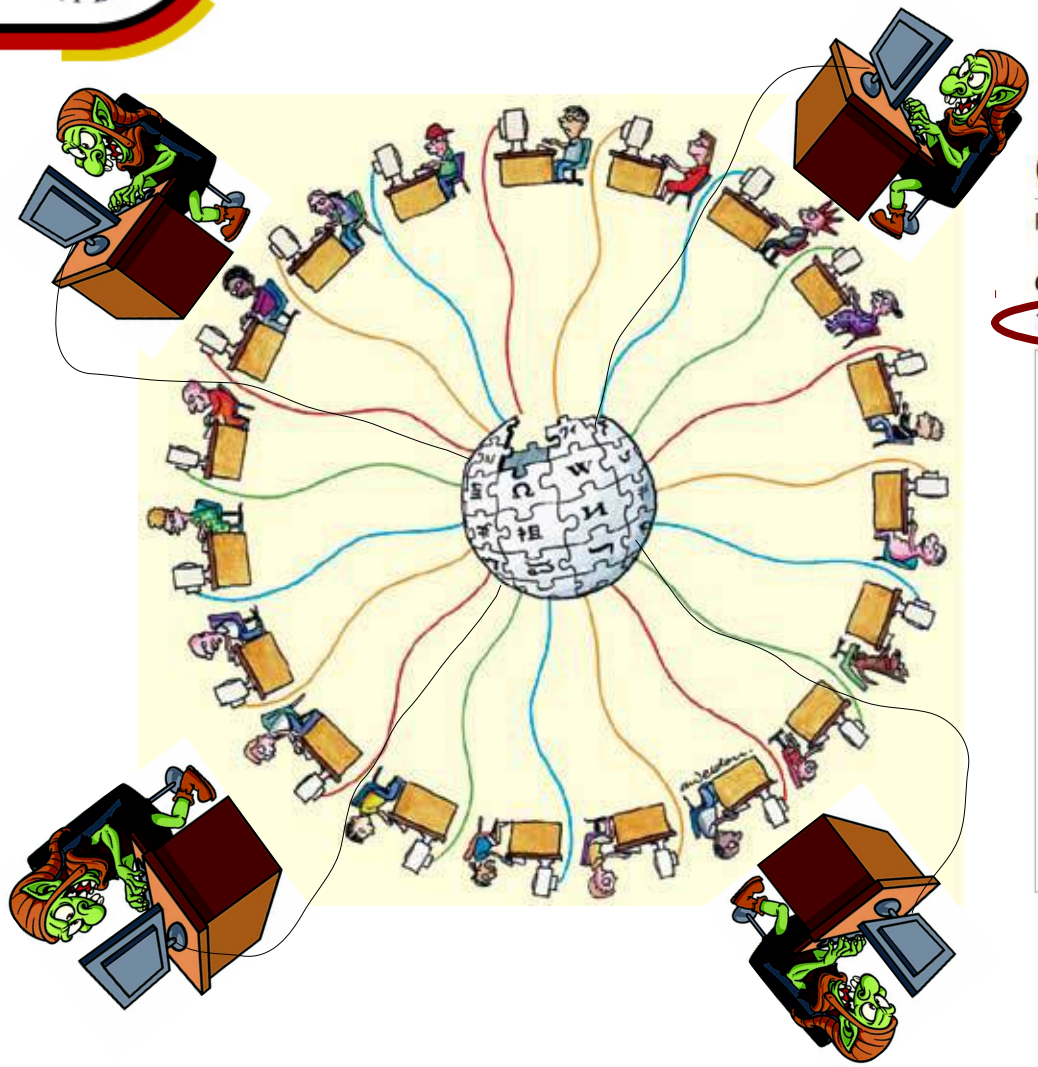
Wikipedia



The largest encyclopedia

- 35M articles over 290 languages
- 73K+ active editors

Vandals on Wikipedia



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

Current Tools: Detect Vandalism

ClueBot NG
Bot to detect vandalism edits

STiki
Tool available to trusted users for reverting vandalism

Current Tools: Detect 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

Our Goal

DETECT VANDALS BEFORE THEY ARE DETECTED BY OTHER WIKIPEDIA USERS/BOTS

Our Goal

DETECT VANDALS BEFORE THEY ARE DETECTED BY OTHER WIKIPEDIA USERS/BOTS

VEWS

USES EDITING BEHAVIOR DIFFERENCE TO CATCH VANDALS

Outline

- UMDWikipedia Dataset
- User Editing Behavior: Similarities and Differences
- VEWS: Vandal Early Warning System
- Efficiency of VEWS
- Comparison with State of the Art Tools

UMDWikipedia Dataset

- 1 M users who registered between Jan 2013 – July 2014:
  - 17k vandals - All users blocked by administrators for vandalism
  - 17k benign users

- 770k edits: 160k vandals, 610k benign users

  edit(ID, user, page, timestamp, reverted)

- Information about each page:

  page(ID, title, article/meta, categories)

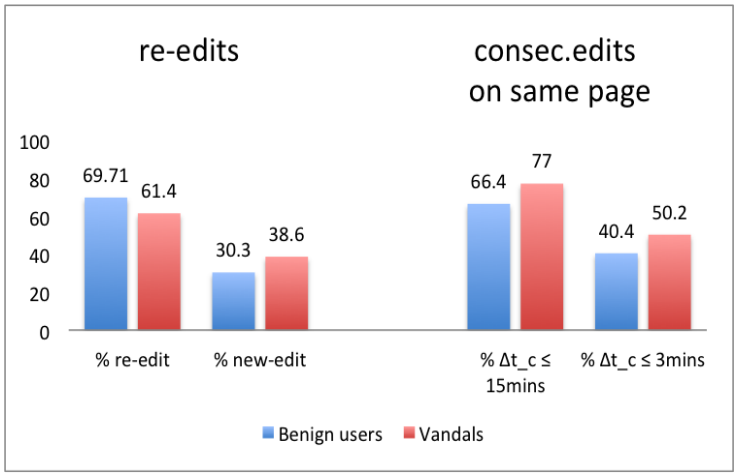
  page_distance(p1, p2)
- STiki and ClueBot NG data

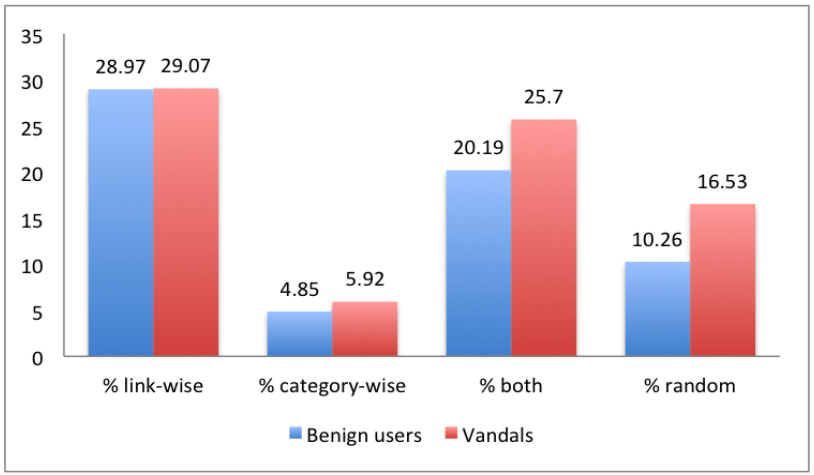
The dataset is available at *http://cs.umd.edu/~vs/vews/*

Outline

- UMDWikipedia Dataset
- User Editing Behavior: Similarities and Differences
- VEWS: Vandal Early Warning System
- Efficiency of VEWS
- Comparison with State of the Art Tools

Editing Behavior: Similarities

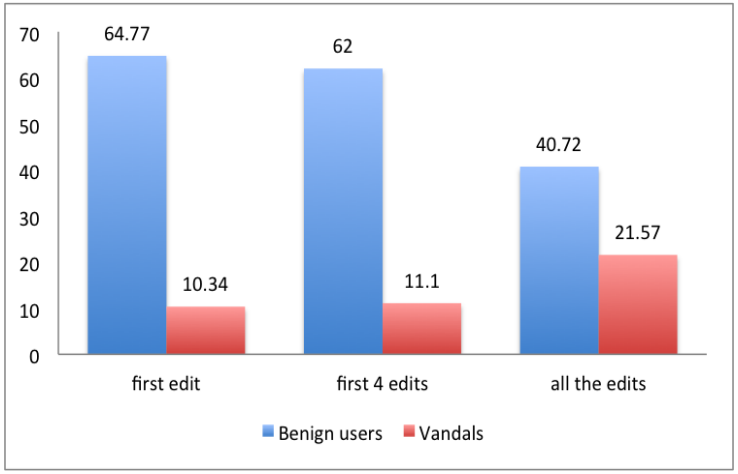


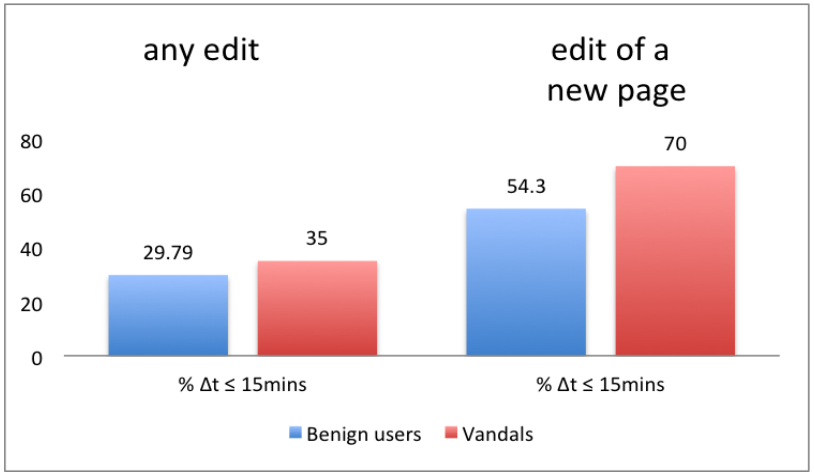


- Wikipedia users are more likely to re-edit a page than edit a new page
- Wikipedia users re-edit the same page quickly

- Both vandals and benign users exhibit similar navigation patterns:
  - Link-wise: Close but low semantic similarity
  - Category-wise: High semantic similarity, but not close

Editing Behavior: Differences





- Benign users edit far more meta-pages than vandals (even in the first edit)

- Vandals make faster edits than benign users
- Vandals spend less time editing a new page

Outline

- UMDWikipedia Dataset
- User Editing Behavior: Similarities and Differences
- VEWS: Vandal Early Warning System
- Efficiency of VEWS
- Comparison with State of the Art Tools

VEWS Features

VEWS

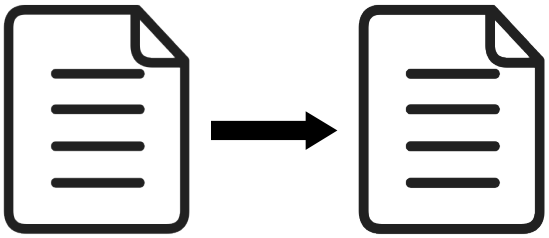
Behavioral Features

Features inspired from the editing behavior analysis

State Transition Features

Patterns in sequence of edits made by users

Behavioral Features



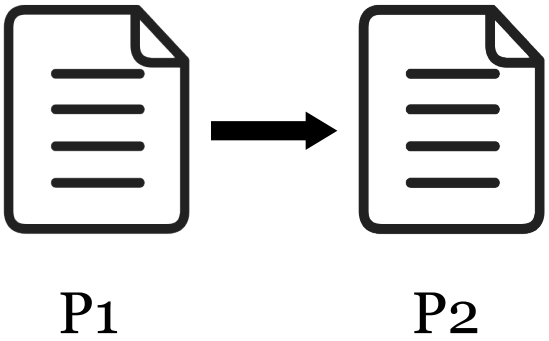
User U

Use consecutive edits made by user

Requirement:

- Unified representation for consecutive edits

Behavioral Features

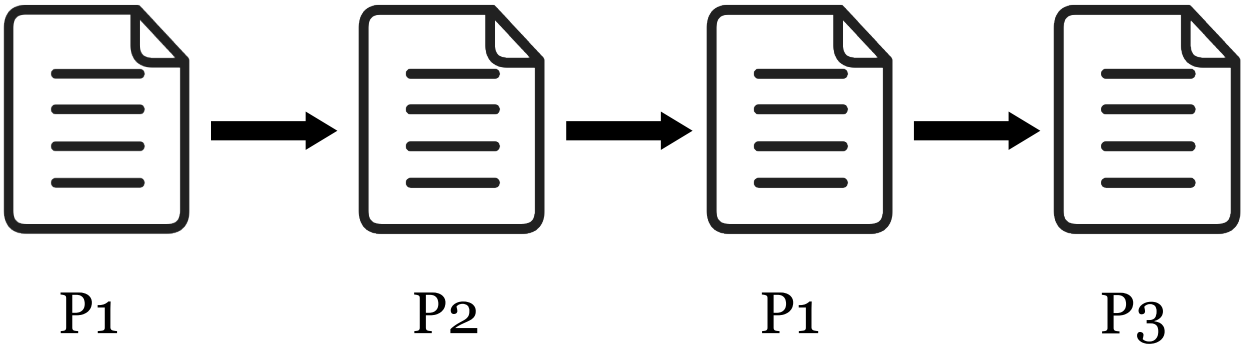


User U: f(U, P1,P2)

Reduces any pair of edits *(U, P1, P2)* into a feature vector:

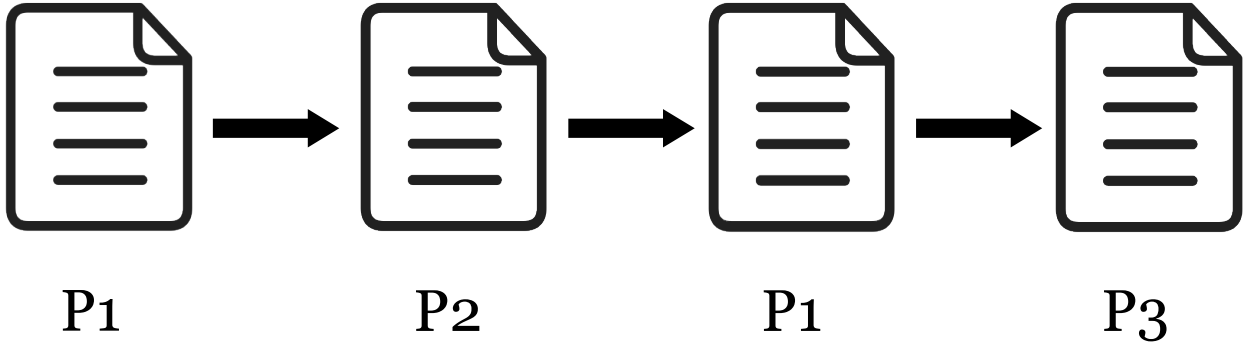
- Is P2 an article page?
- Time difference:
  very fast/fast/slow
- Is P2 U's First edit?
- P1 = P2?
- First time edit on P2?
- Hops from P1 to P2:
  <= 3, > 3 or null
- Number of common categories

Behavioral Features



User U: f(U, P1,P2) f(U, P2,P1) f(U, P1,P3)

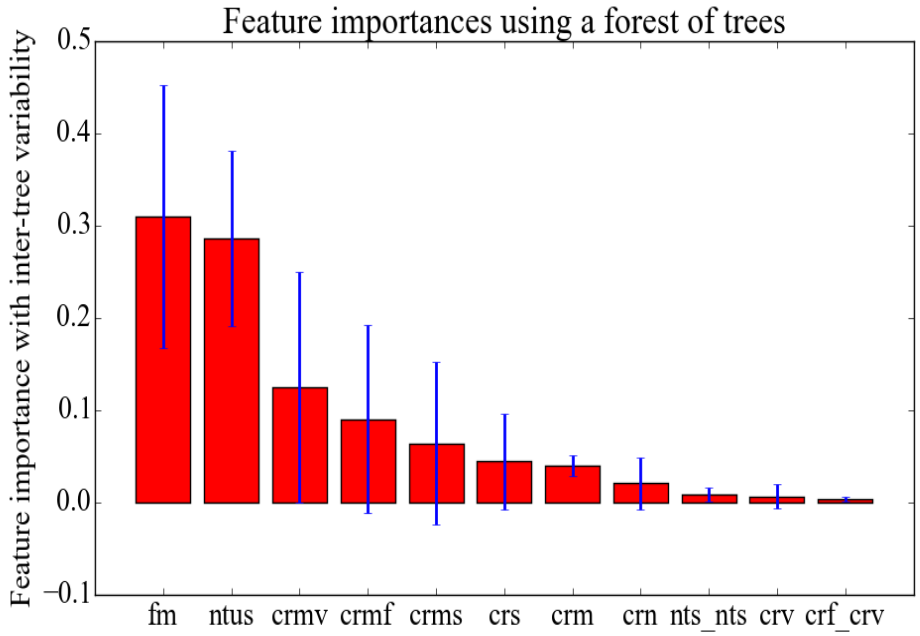
Behavioral Features



User U: f(U, P1,P2) f(U, P2,P1) f(U, P1,P3)

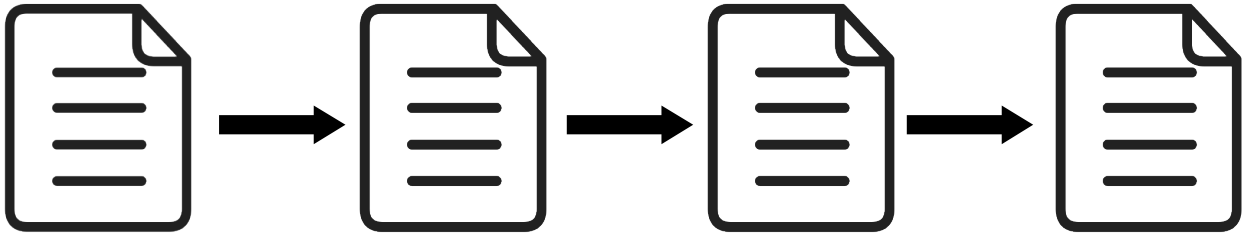
Frequent Pattern Mining on vandal and benign users
11 features with large difference

Behavioral Features



- ***fm***: If the first edit is on a meta page, then user is more likely to be benign.
- ***ntus***: Benign users are likely to take longer to edit a new page than a vandal.

State Transition Features



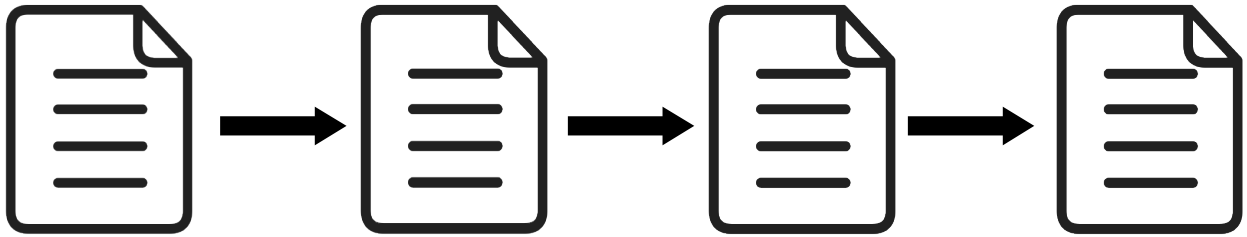
P1 P2 P1 P3

User U: f(U, P1,P2) f(U, P2,P1) f(U, P1,P3)

Use the entire sequence of edits of the user

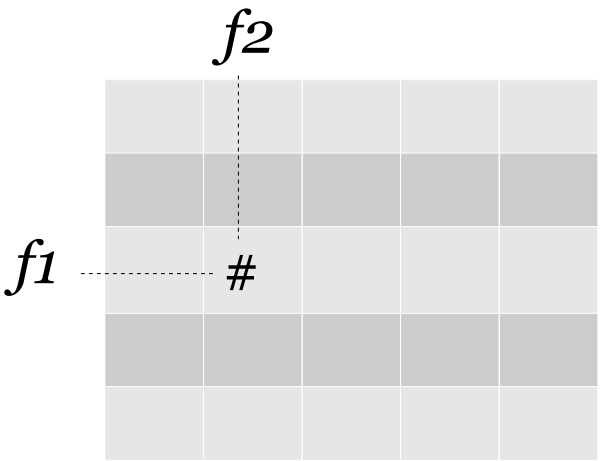
- But each user can have variable number of edits!

State Transition Features



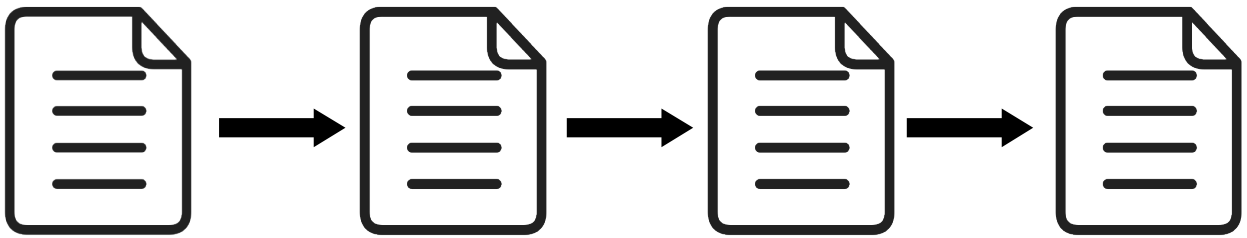
P1 P2 P1 P3

User U: f(U, P1,P2) f(U, P2,P1) f(U, P1,P3)



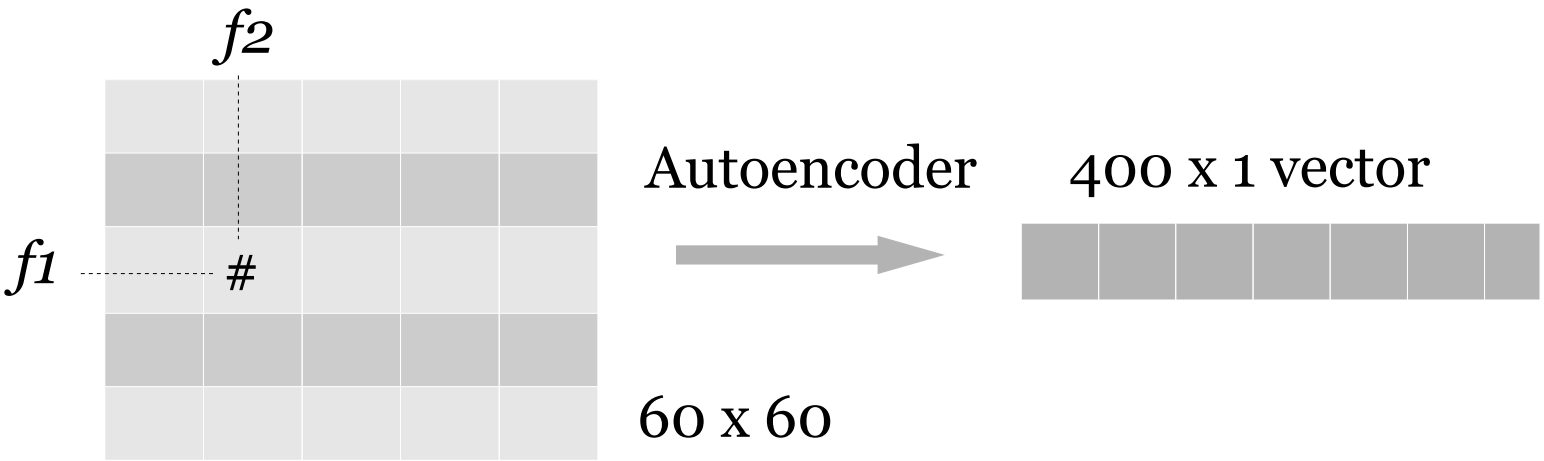
60 x 60

State Transition Features



P1 P2 P1 P3

User U: f(U, P1,P2) f(U, P2,P1) f(U, P1,P3)



Outline

- UMDWikipedia Dataset
- User Editing Behavior: Similarities and Differences
- VEWS: Vandal Early Warning System
- Efficiency of VEWS
- Comparison with State of the Art Tools

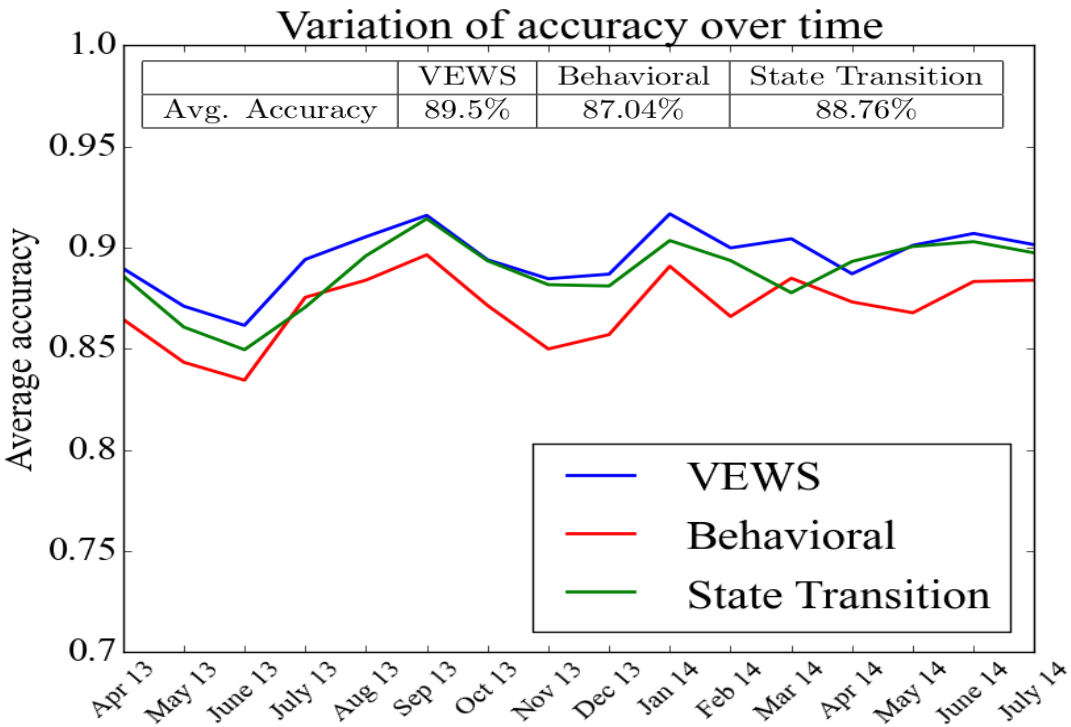
Efficiency of VEWS

	Accuracy	TPR	TNR	FPR	FNR
Behavioral	86.5%	0.85	0.89	0.11	0.15
State Transition	87.39%	0.88	0.90	0.10	0.12
VEWS	87.82%	0.87	0.92	0.08	0.13

- 10 fold cross validation using Support Vector Machine
- VEWS has the highest accuracy.

Temporal Advantage

Using recent data improves accuracy!

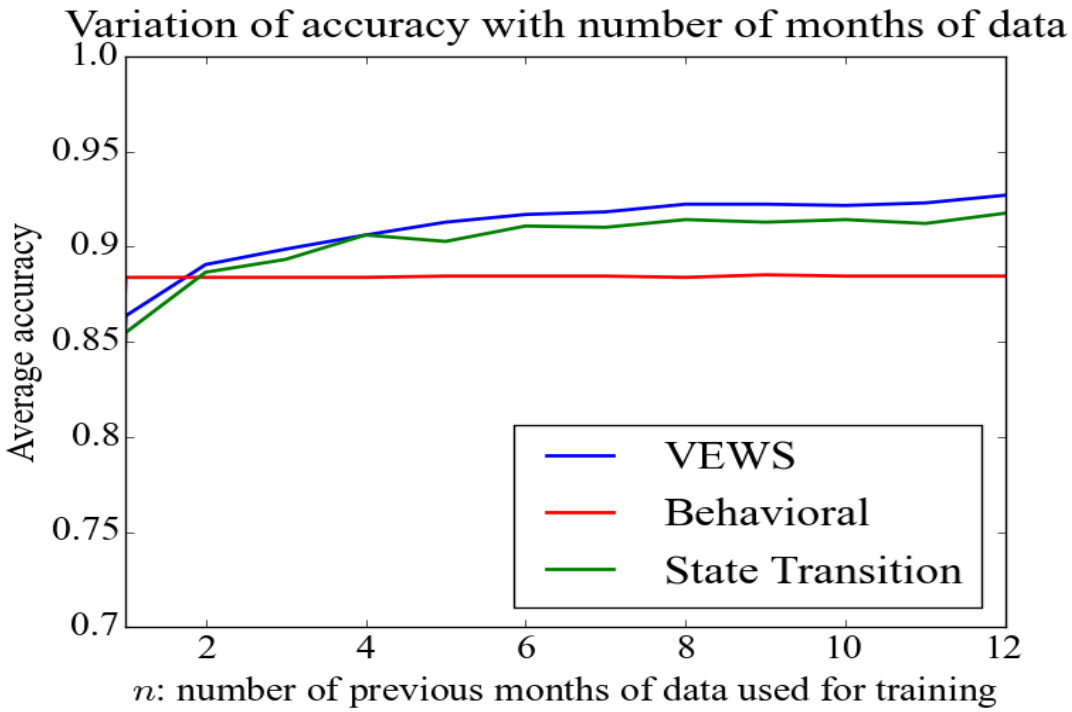


	VEWS	Behavioral	State Transition
Avg. Accuracy	89.5%	87.04%	88.76%

Training using previous 3 months of data

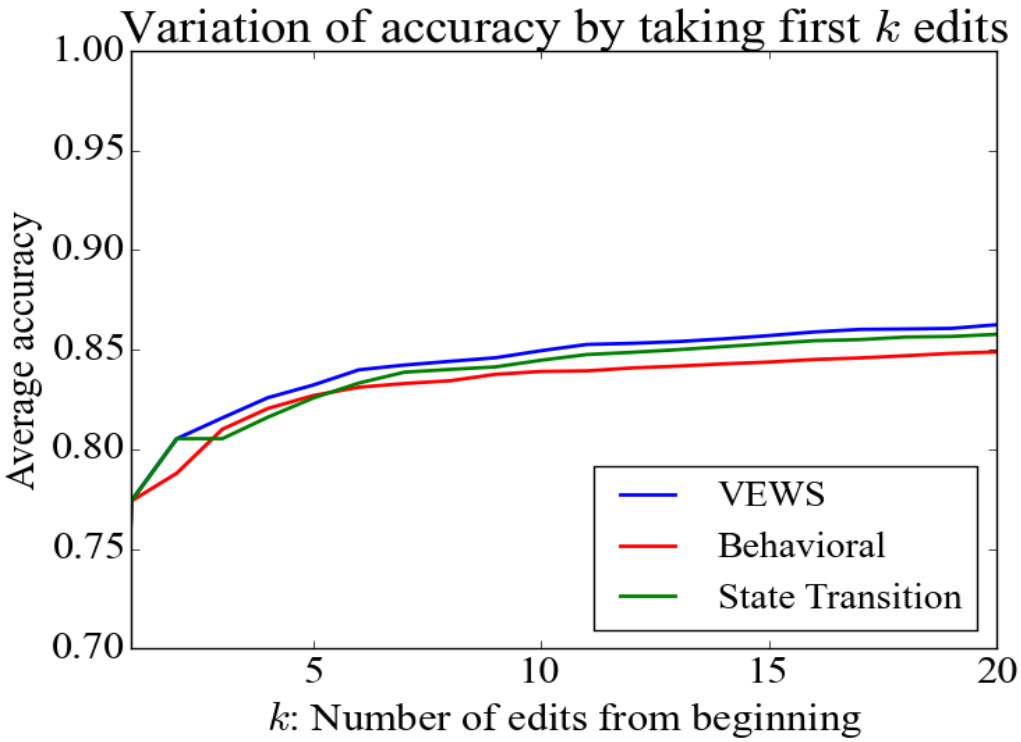
Temporal Advantage

More temporal data, more accuracy!



Previous n months of data for training and testing on July 2014

Early warning system



- Only first few edits needed to accurately identify the vandal
- With only the first edit, VEWS has 78% accuracy
- In 43.68% cases, vandal is identified before its first reversion.

Outline

- UMDWikipedia Dataset
- User Editing Behavior: Similarities and Differences
- VEWS: Vandal Early Warning System
- Efficiency of VEWS
- Comparison with State of the Art Tools

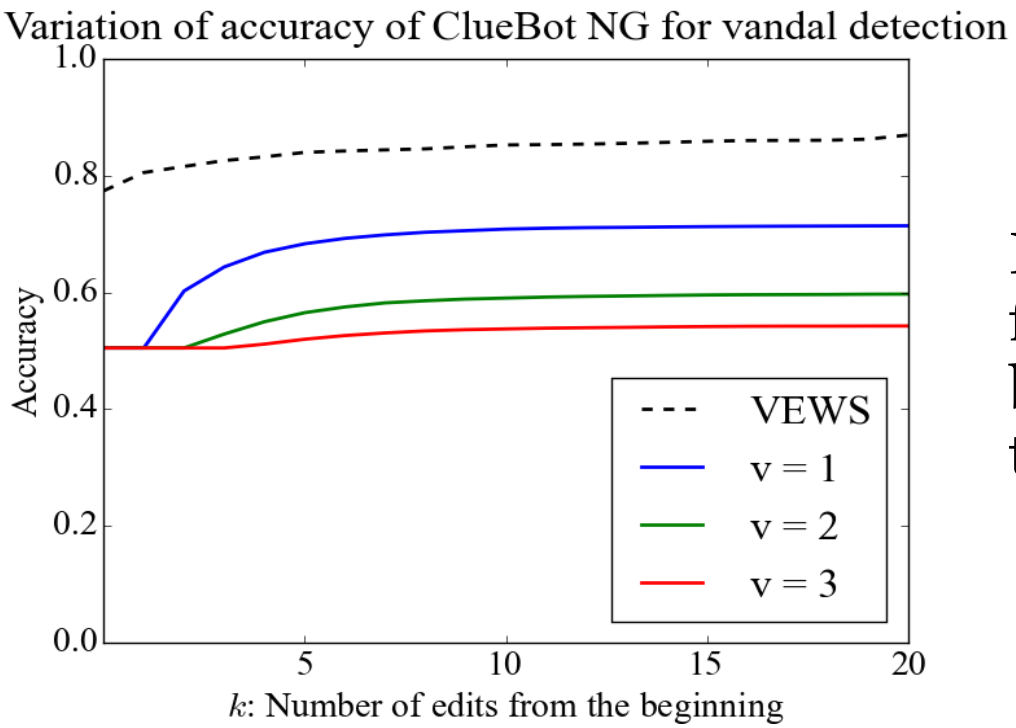
State of the Art tools

ClueBot NG

STiki

Detect vandalism, not vandals!

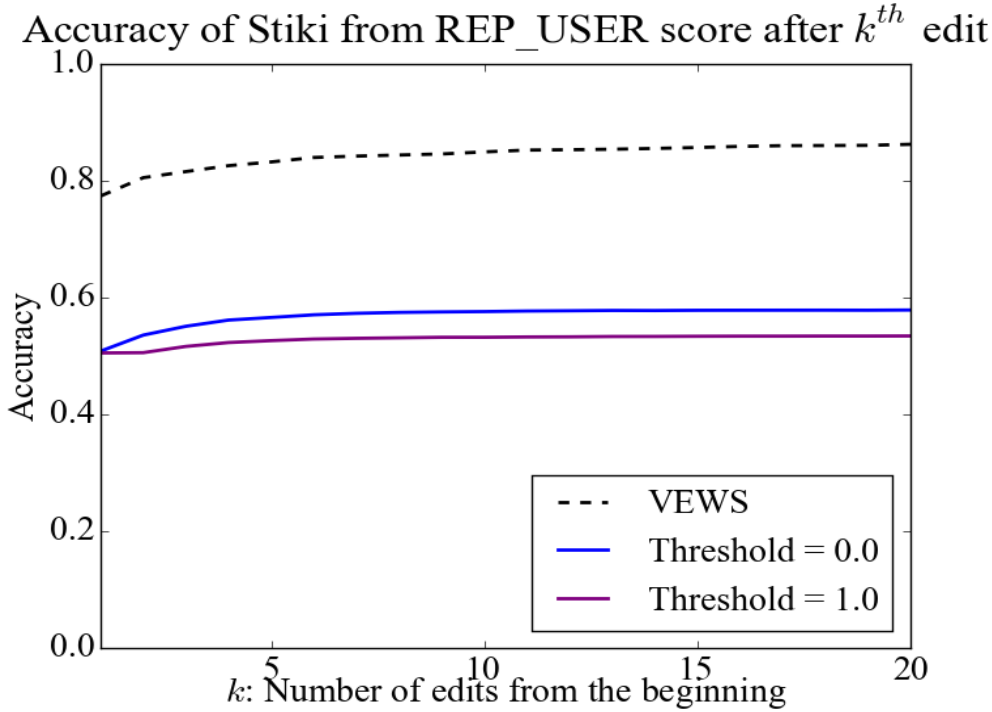
Comparison with ClueBot NG



Rule:
If v edits of a user's first k edits are caught by ClueBot NG, then the user is a vandal.

- ClueBot NG has accuracy of 70%
- When both VEWS and ClueBot NG detect the vandal, VEWS does it 2.39 edits earlier.

Comparison with STiki



Rule:
If the STiki user reputation score after the kth edit is above the threshold, then the user is a vandal.

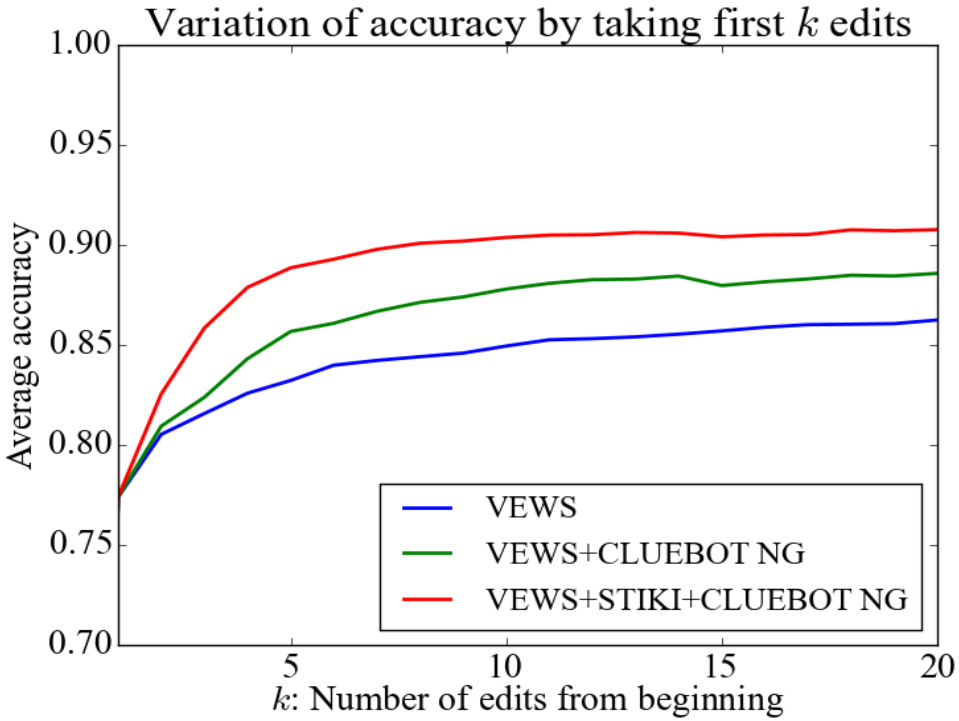
- STiki has an accuracy of 60%
- VEWS consistently finds vandals much more effectively than STiki's user reputation score and vandalism probability score.

State of the Art tools

ClueBot NG and STiki: Revert edit based on content

ClueBot NG STiki	VEWS
Natural Language Processing features	Editing Behavior features

Combination is the best!



- Adding features from ClueBot NG and STiki improves VEWS !
- VEWS + ClueBot NG: A fully automated system with 88.6% accuracy.

Summary

VEWS:

- Is a fully automated vandal early warning system
- Detects vandals by exploiting editing behavior differences
- Detects vandals within a few edits with high accuracy
- Improves by combining with state-of-the-art tools

Related Work

- Vandalism detection algorithms and PAN-WVC-2010 dataset:
  [Potthast et al. 2008], [Mola-Velasco 2010], [West et al. 2010], [Adler et al. 2010], [Adler et al. 2011], [Sumbana et al. 2012], [Ferschke et al. 2013]

- Automated/Semi-automated tools for vandalism detection:
  ClueBot NG, Stiki [West et al. 2010], WikiTrust (now defunct) [Adler et al. 2010]

- User navigation behavior:
  [Catledge et al. 1995], [Cockburn et al. 2001], [Adar et al. 2008], [West et al. 2012]

Thank you!

UMDWikipedia dataset is available at
http://cs.umd.edu/~vs/vews

Please drop by our poster!

Questions?