

Detecting Wikipedia Vandalism via Spatio-Temporal Analysis of Revision Metadata

Andrew G. West

June 10, 2010

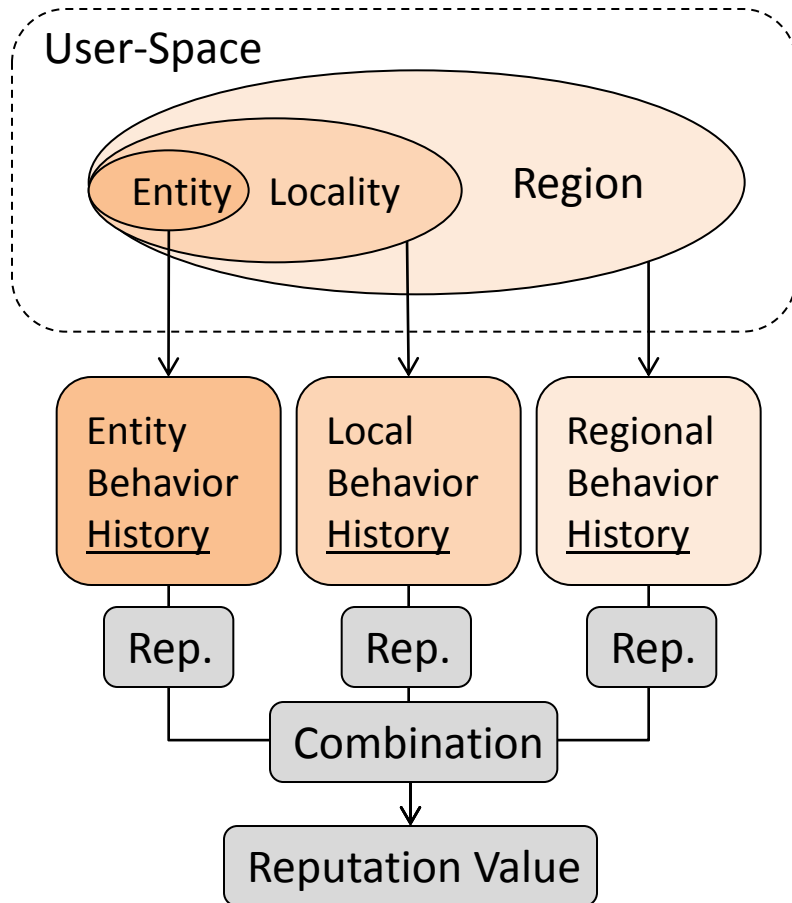
ONR-MURI Presentation

Report Documentation Page				Form Approved OMB No. 0704-0188	
Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.					
1. REPORT DATE 10 JUN 2010		2. REPORT TYPE		3. DATES COVERED 00-00-2010 to 00-00-2010	
4. TITLE AND SUBTITLE Detecting Wikipedia Vandalism via Spatio-Temporal Analysis of Revision Metadata				5a. CONTRACT NUMBER	
				5b. GRANT NUMBER	
				5c. PROGRAM ELEMENT NUMBER	
6. AUTHOR(S)				5d. PROJECT NUMBER	
				5e. TASK NUMBER	
				5f. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) University of Pennsylvania, Computer and Information Science, 3451 Walnut St, Philadelphia, PA, 19104				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)				10. SPONSOR/MONITOR'S ACRONYM(S)	
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution unlimited					
13. SUPPLEMENTARY NOTES MURI Review, June 2010. U.S. Government or Federal Rights License					
14. ABSTRACT					
15. SUBJECT TERMS					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT Same as Report (SAR)	18. NUMBER OF PAGES 40	19a. NAME OF RESPONSIBLE PERSON
a. REPORT unclassified	b. ABSTRACT unclassified	c. THIS PAGE unclassified			

Where we left off....

FROM THE LAST MURI REVIEW

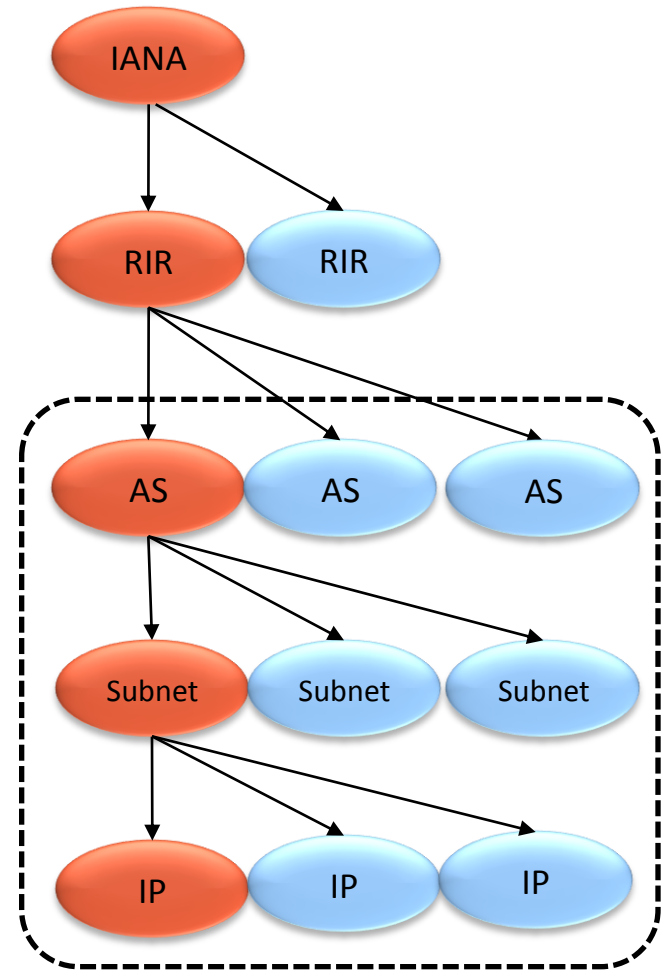
Spatio-Temporal Reputation



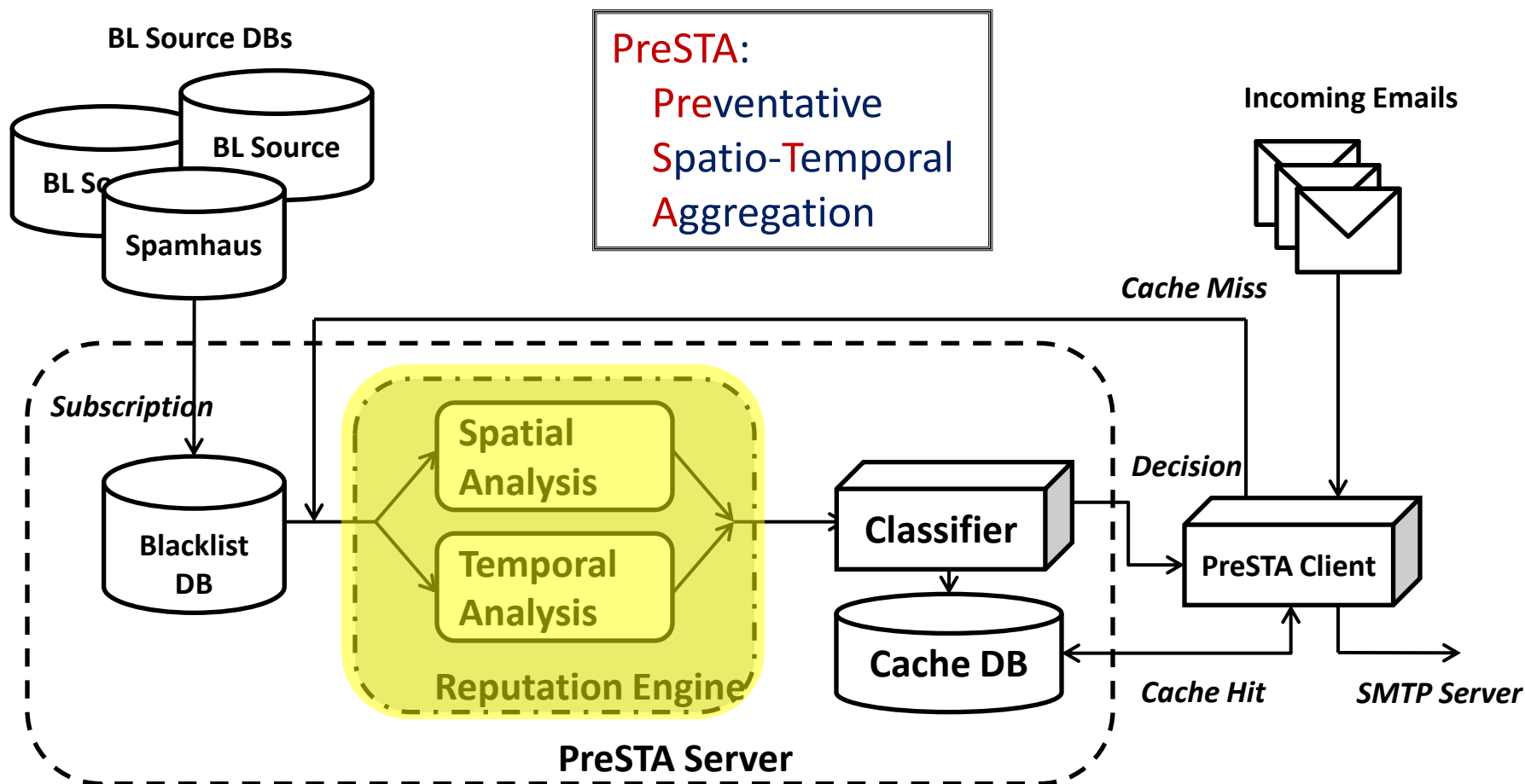
- **Single-entity** reputation values are the status quo
 - Issue: Sybil attacks (*e.g.*, spam botnets)
- **Spatial** reputation:
 - No entity-specific data? Use broader groupings
 - Exploit **homophily**
 - Clarity in borderline classification cases

Hierarchical Groupings = TDG = QTM

- Spatial groupings for spam detection leverage the IP assignment hierarchy
 - Entities are IP addresses
 - {AS, Subnet, IP} groups used
- TDGs are hierarchies, thus spatio-(temporal) techniques may fulfill the reputation component of QTM/QuanTM



PreSTA for Spam Detection



APPLYING SPATIO- TEMPORAL PROPERTIES TO WIKIPEDIA

Vandalism

Barack Hussein Obama II (
🔊 /bəˈrɑːk huːˈseɪn ouˈbɑːmə/; born August 4, 1961) is !!! THE WORSTEST PRESIDENT EVER. PLEASE RESIGN IMMEDIATELY!!!
the 44th and current President of the United States. He is the first African American to hold the office. Obama previously served as the junior United States Senator from Illinois, from January 2005 until he resigned after his election to the presidency in November 2008.

Originally from Hawaii, Obama is a graduate of Columbia University and Harvard Law School, where he was the president of the Harvard Law Review. He was a community organizer in Chicago before earning his law degree. He worked as a civil rights attorney in Chicago and taught constitutional law at

Barack Obama



- Serious problem. One source [3] estimates hundreds of millions of `damaged page views`
- NLP effective for blatant instances. Subtle ones (e.g., insertion of `not`, name replacement) – much harder to find
- Our method: Alternative means of detection, complementing NLP

VANDALISM: Informally, an edit that is:

- Non-value adding
- Offensive
- Destructive in content removal

Big Idea

- Wikipedia **revision metadata** (not the article or `diff` text) can be used to detect instances of vandalism
 - As effective as language-processing [2] efforts
 - Machine-learning over spatio-temporal props:
 - Simple features: Straightforward metadata analysis
 - Aggregate features: **Reputation values** for single entities (editors, articles) and spatial groupings thereof (geographical location, topical categories)

Outline

- Labeling revisions (*rollback*)
- Simple features
 - Motivation: **SNARE** [1] spam-blocking
 - Edit time-of-day, day-of-week, comment length...
- Aggregate features
 - Motivation: **PreSTA** [5] reputation algorithm
 - Article rep., editor rep., spatial reputations...
- Classifier performance
- **STiki** [4] (a real-time implementation)

Metadata

Wikipedia provides metadata via DB-dumps:

#	METADATA ITEM	NOTES
(1)	Timestamp of edit	In GMT locale
(2)	Article being edited	Able to deduce namespace from title
(3)	Editor making edit	May be user-name (if registered editor), or IP address* (if anonymous)
(4)	Revision comment	Text field where editor can summarize changes

Labeling Vandalism

“Reversion” (*i.e.*, undo)

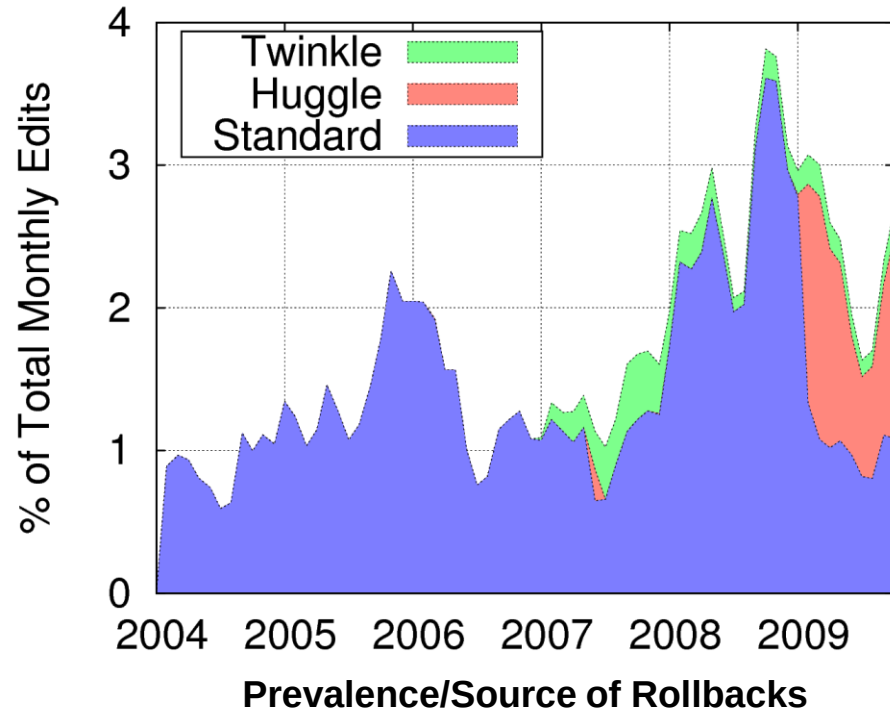
- Any user can execute:
- (1) Press button
- (2) Enter edit summary
- (3) Confirm reversion

“Rollback” (expedited revert)

- Privileged: $\approx 4,700$ users
- (1) Press button. Done.
- Auto-summarization:
“Reverted edits by x to last revision by y”

Test-set contains ≈ 50 million edits:

- (1) only NSO edits (71% of all edits)
- (2) only edits within last year (2008/11+)



Rollback-based Labels

- Use **rollback-based labeling**:
 - (1) Find special comment format
 - (2) Verify permissions of editor
 - (3) Backtrack to find **offending-edit (OE)**
 - All edits not in set {OE} are {Unlabeled}
- Alternatives: Manual labeling, page-hashing
- Advantages of using rollback:
 - (1) **Automated** (just parsing)
 - (2) **High-confidence** (privileged users are *trusted*)
 - (3) **Per-case** (vandalism need not be defined)

SIMPLE FEATURES

* Discussion abbreviated to concentrate on aggregate ones

Spatio-Temporal Basics

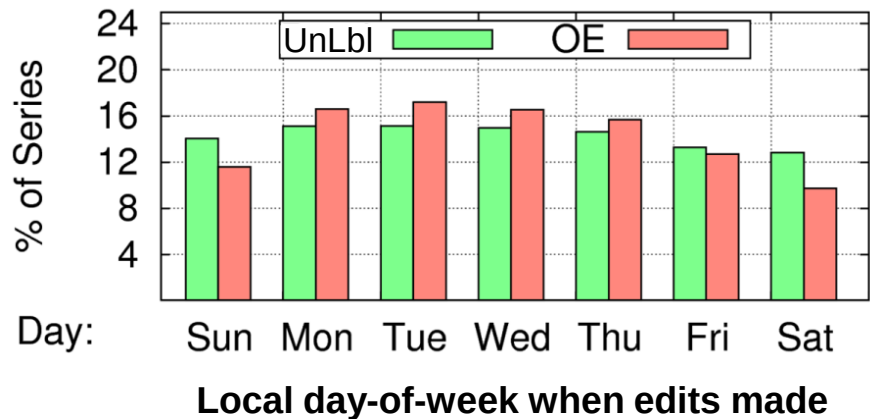
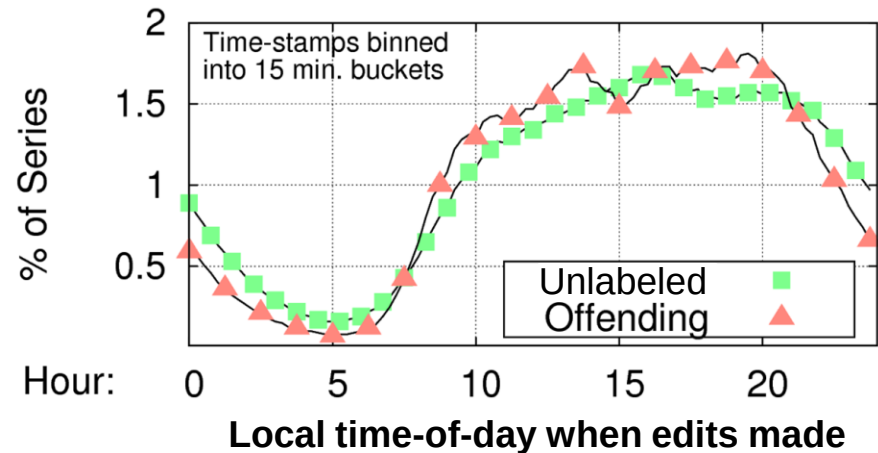
- **Temporal props**: A function of when events occur
 - **Spatial props**: Appropriate wherever a size, distance, or membership function can be defined
-

Motivating work: SNARE [1]

- Spatio-temporal props. **effective in spam-mitigation**
 - Physical distance mail traveled, time-of-day, mail sent, message size (in bytes), AS-membership of sender... (13 in total)
- Advantages of approach:
 - NLP-filters **easy to evade**... More difficult for spatio-temporal props.
 - **Computationally simpler** than NLP

Edit Time, Day-of-Week

- Use IP-geo-location data to determine origin time-zone, adjust UTC timestamp
- Vandalism most prevalent during working hours/week: Kids are in school(?)
- Fun fact: Vandalism almost twice as prevalent on a Tuesday versus a Sunday



Time-since (TS) ...

TS Article Edited	OE	UnLbl
All edits (median, hrs.)	1.03	9.67
TS Editor Registration	OE	UnLbl
Regd., median (days)	0.07	765
Anon., median (days)	0.01	1.97

- **Long-time participants vandalize very little**
 - “Registration”: time-stamp of first edit made by user
 - Sybil-attack to abuse benefits?

- **High-edit pages most often vandalized**
 - $\approx 2\%$ of pages have 5+ OEs, yet these pages have 52% of all edits
 - Other work [3] has shown these are also articles most visited

Misc. Simple Features

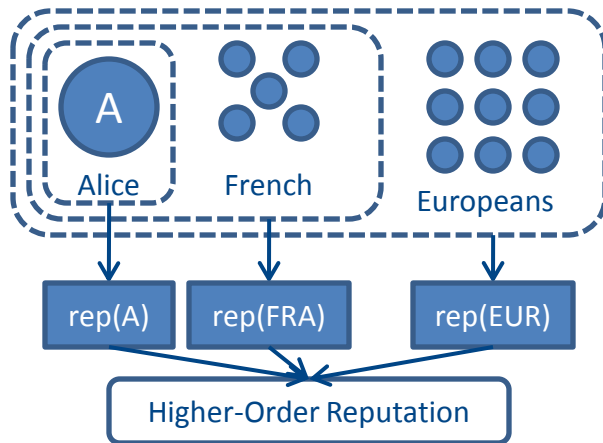
FEATURE	OE	UnLbl
Revision comment (average length in characters)	17.73	41.56
Anonymous editors (percentage)	85.38%	28.97%
Bot editors (percentage)	00.46%	09.15%
Privileged editors (percentage)	00.78%	23.92%

- Revision comment length
 - Vandals leave **shorter comments** (lazy-ness? or just minimizing bandwidth?)
- Privileged editors (and bots)
 - Huge contributors, but rarely vandalize

AGGREGATE FEATURES

PreSTA Algorithm

CORE IDEA: No entity specific data? Examine spatially-adjacent entities (homophily)



PreSTA [5]: Model for ST-rep:

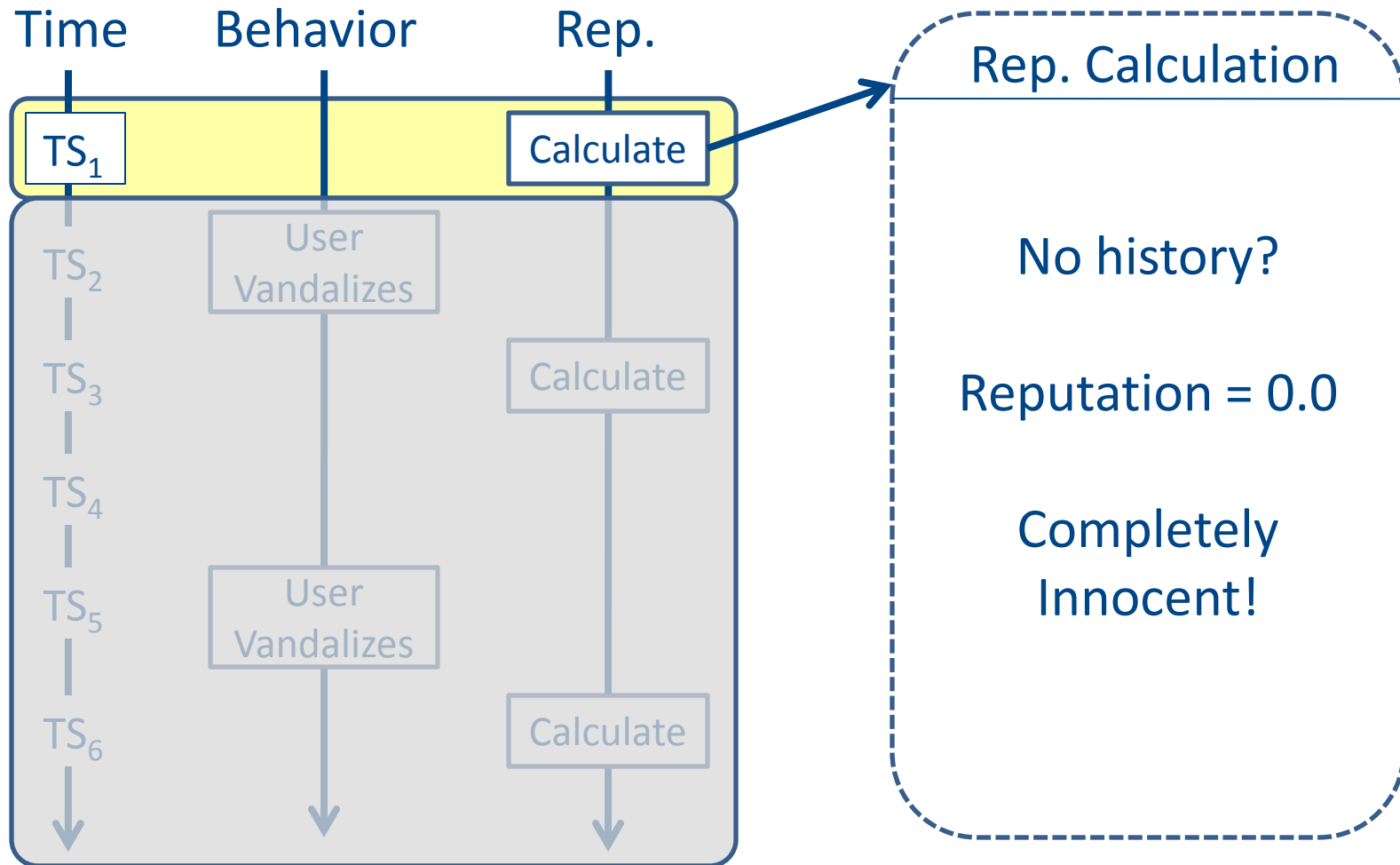
$$\text{Rep}(\text{group}) =$$

$$\sum \frac{\text{time_decay}(\text{TS}_{\text{vandalism}})}{\text{size}(\text{group})}$$

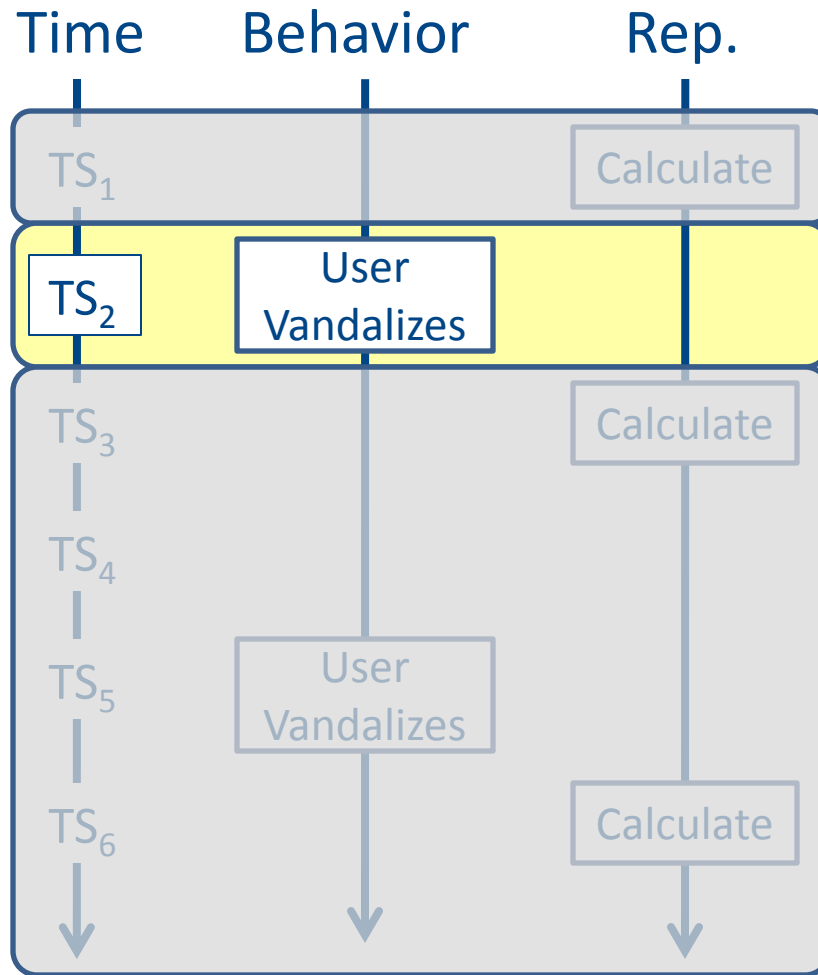
Timestamps (TS) of
vandalism incidents
by *group* members

- **Grouping functions (spatial)** define memberships
- Observations of misbehavior form **feedback** – and observations are decayed (**temporal**)

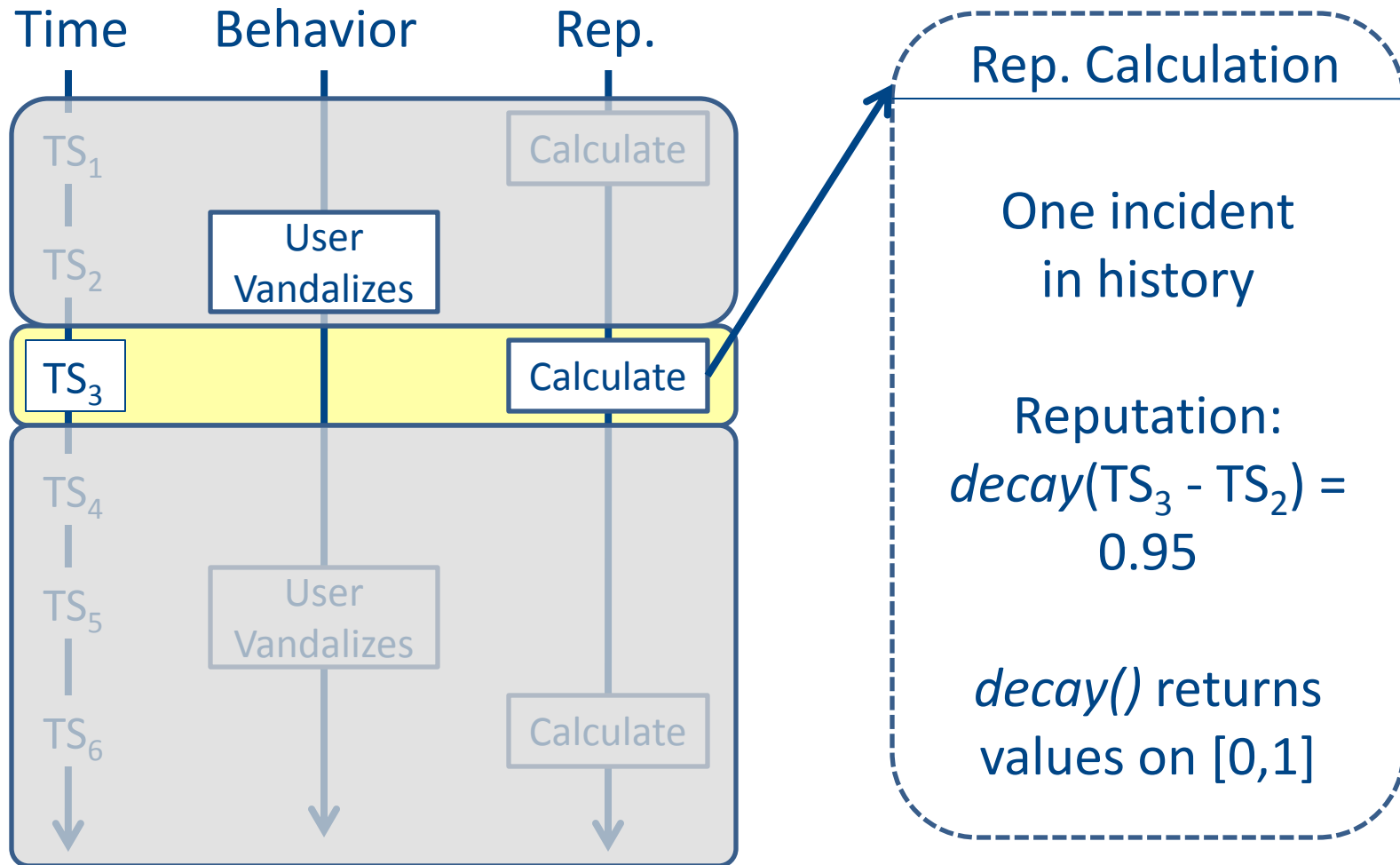
Example Reputation



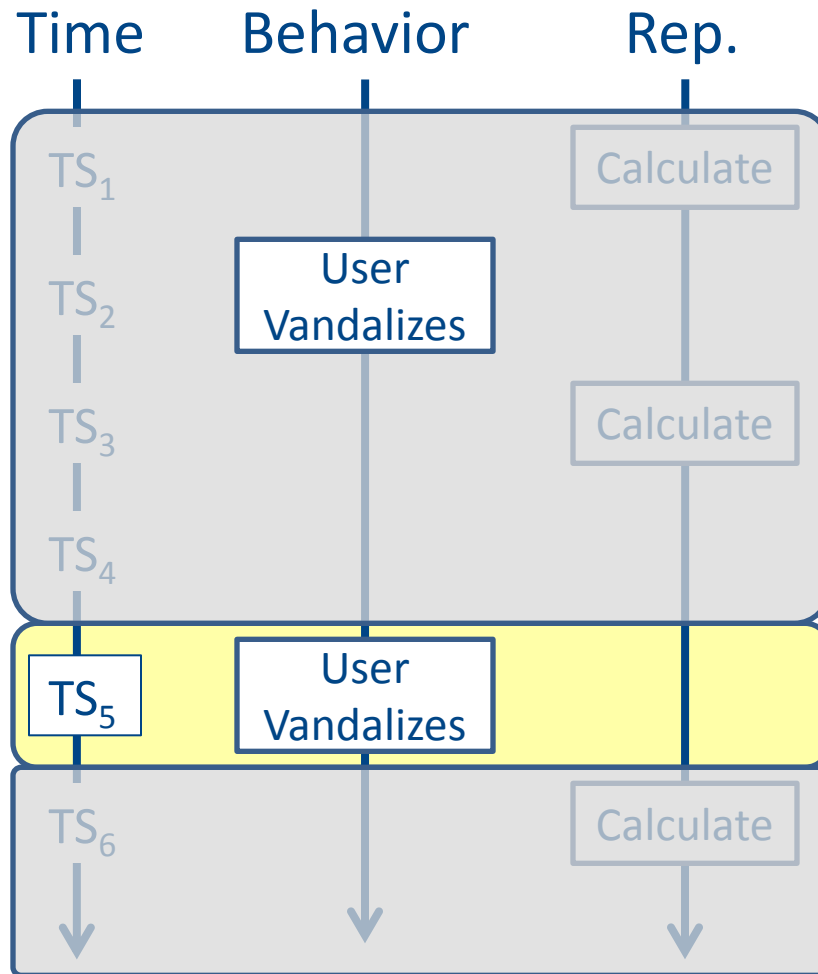
Example Reputation



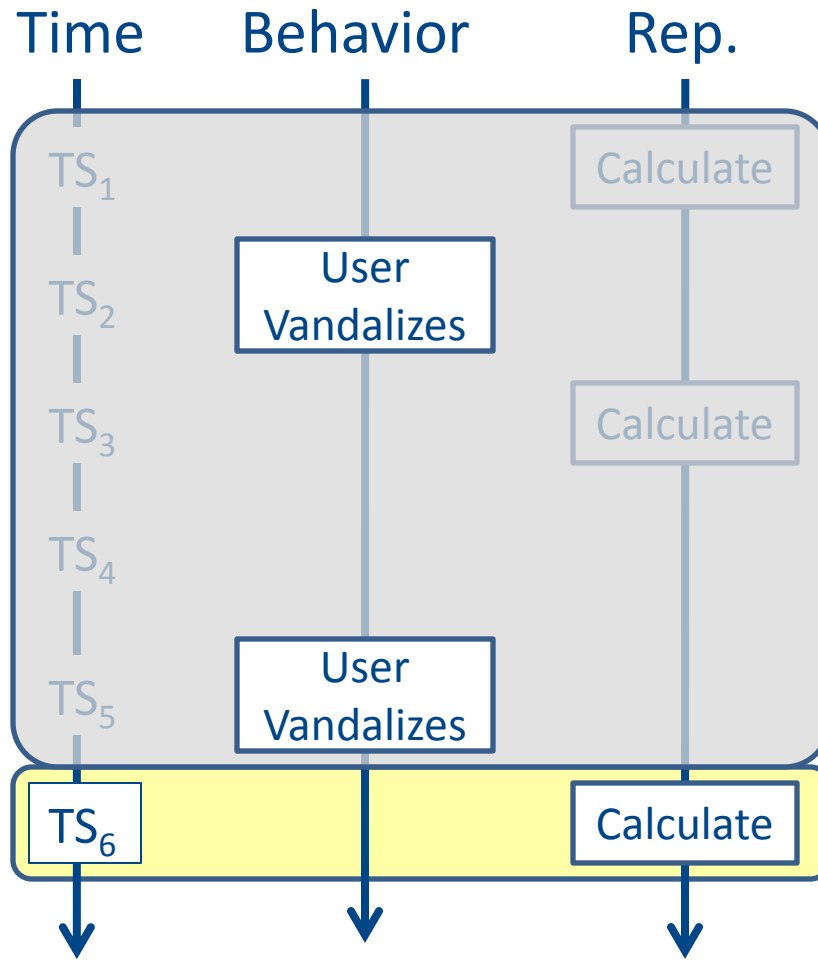
Example Reputation



Example Reputation



Example Reputation



Rep. Calculation

Two incidents
in history

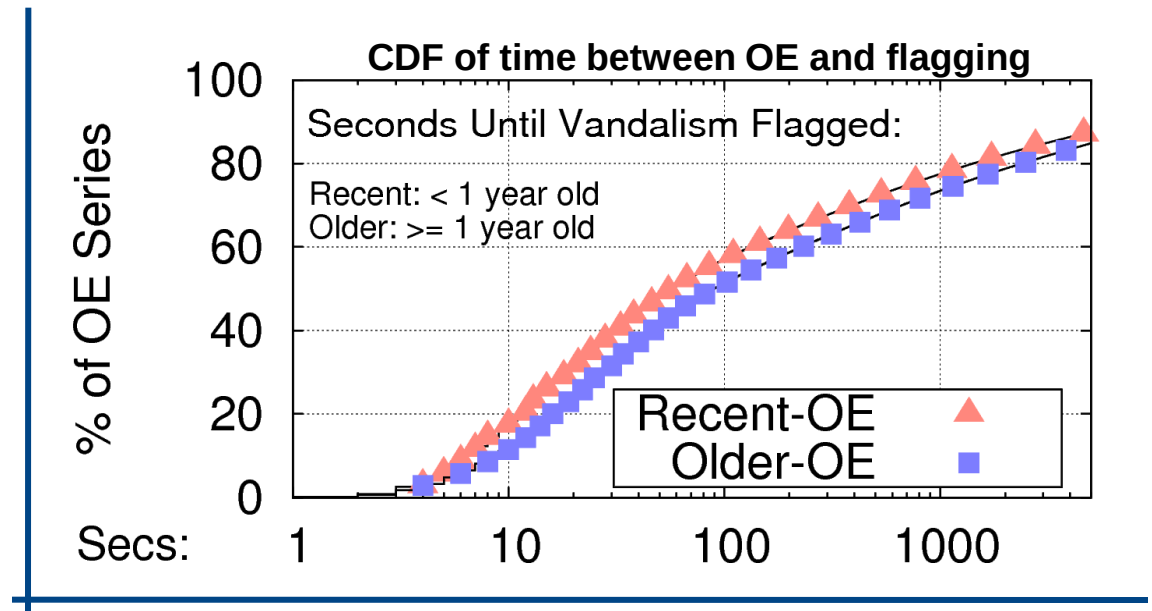
Reputation:

$$\begin{aligned} & decay(TS_6 - TS_2) + \\ & decay(TS_6 - TS_5) = \\ & 0.50 + 0.95 = 1.45 \end{aligned}$$

Values are **relative**

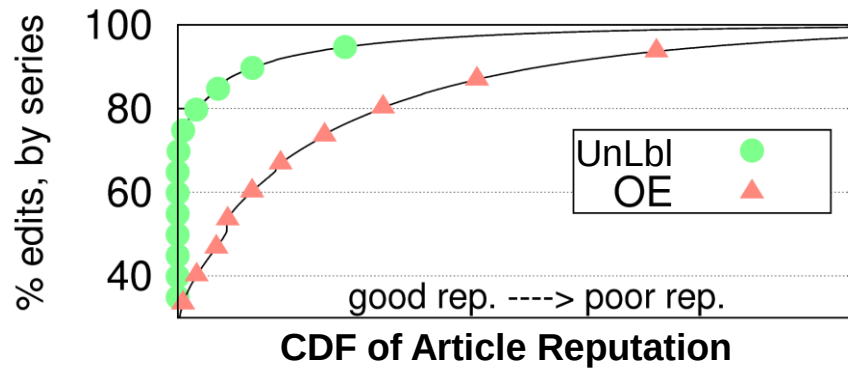
Rollback as Feedback

Use rollbacks
(OEs) as neg.
feedbacks
for entities



- Key notion: A bad edit is not part of reputation until ($TS_{\text{flag}} > TS_{\text{vandalism}}$). Thus, vandalism **must be flagged quickly** so reputations are not latent.
 - Fortunately, median time-to-rollback: ≈ 80 seconds

Article Reputation



ARTICLE	#OEs
George W. Bush	6546
Wikipedia	5589
Adolph Hitler	2612
United States	2161
World War II	1886

Articles w/most OEs

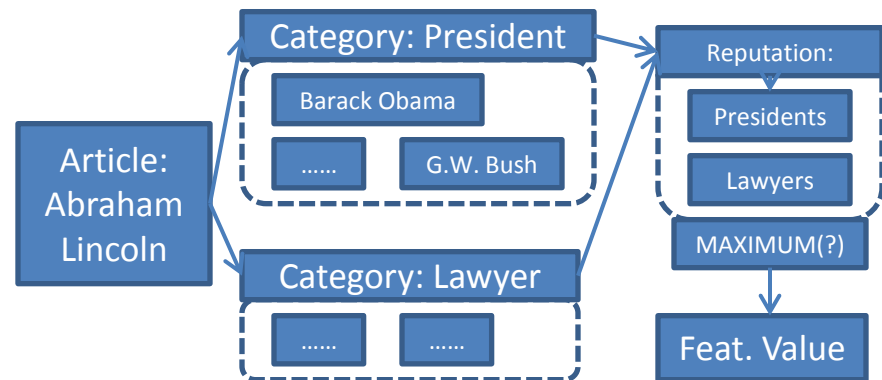
- Intuitively **some topics are controversial** and likely targets for vandalism (or temporally so).
- Trivial spatial grouping (size=1)
- **85% of OEs have non-zero rep** (just 45% of random)

Category Reputation

- Category = spatial group over articles
- Wiki provides cats. /memberships – use only **topical ones**
- $size()$ = Number of category members
- Overlapping grouping
- **97% of OEs have non-zero reputation** (85% in article case)

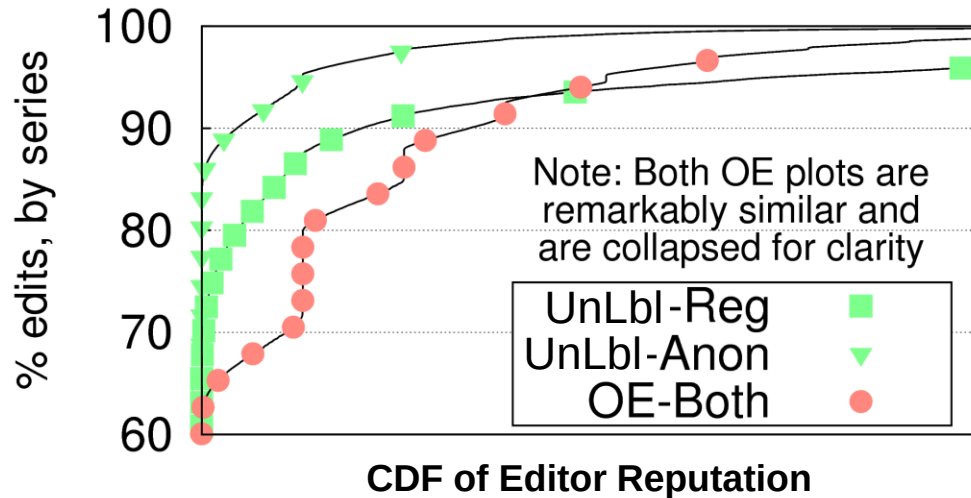
CATEGORY (with 100+ members)	PGs	OEs/PG
World Music Award Winners	125	162.27
Characters of Les Miserables	135	146.88
Former British Colonies	145	141.51

Categories with most OEs



Example of Category Rep. Calculation

Editor Reputation

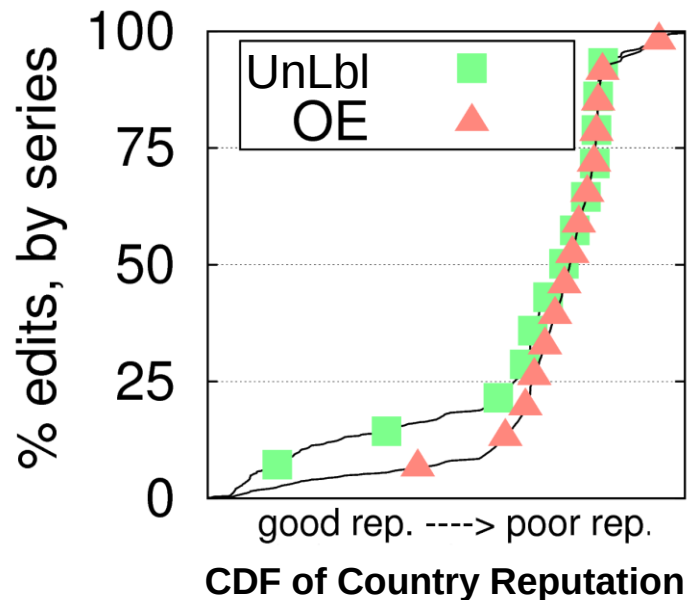


- Straightforward use of the *rep()* function, **one-editor groups**

- **Problem:** Dedicated editors accumulate OEs, look as bad as attackers (**normalize?** No)
- Mediocre performance. Meaningful **correlation** with other features, however.

Country Reputation

- Country = spatial grouping over editors
- Geo-location data maps IP → country
- Straightforward: IP resides in one country



RANK	COUNTRY	%-OEs
1	Italy	2.85%
2	France	3.46%
3	Germany	3.46%
...	...	...
12	Canada	11.35%
13	United States	11.63%
14	Australia	12.08%

**OE-rate (normalized) for
countries with 100k+ edits**

CLASSIFICATION & PERFORMANCE

ML Training

- Calc. features for all edits.
Normalize onto [0,1]; polarity
- SVM: Support Vector Machine
- **ISSUE**: {Unlabeled} set is just that. Very **low cost penalties** so no over-compensation.
- Train over prior subset to classify now (100+ edits/sec).

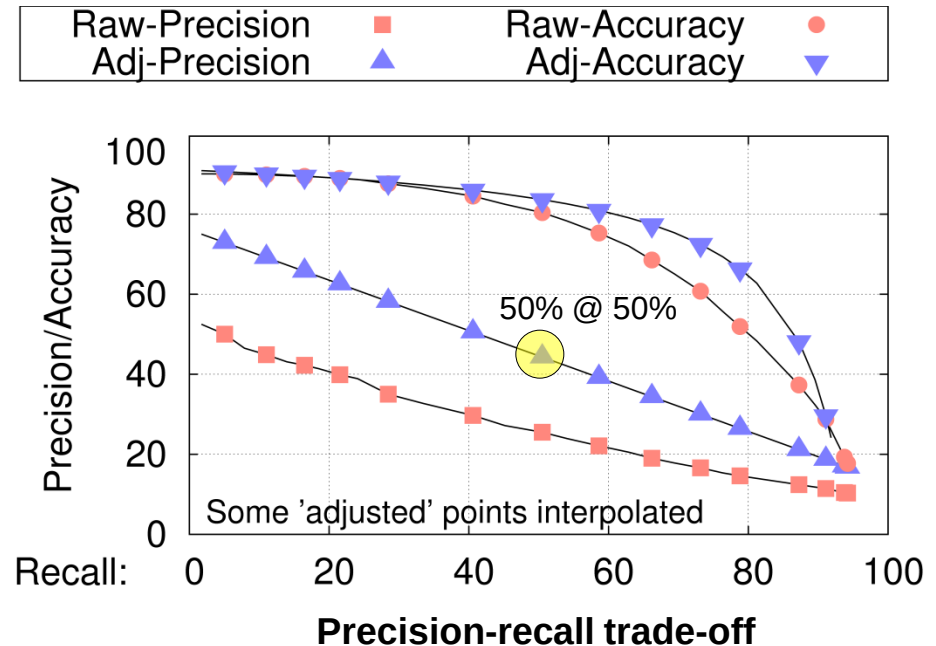
#	FEATURE
1	Edit time-of-day
2	Edit day-of-week
3	Time-since page edited
4	Time-since user reg.
5	Time-since last user OE
6	Rev. comment length
7	Article reputation
8	Category reputation
9	Editor reputation
10	Country reputation

**Review of features used
(only IP-editors)**

Performance

- **ISSUE:** Edits classified as OE but in {UnLbl} may not be FPs:
 - Manual inspection
 - Raw vs. adjusted
 - Corpus produced*
- Similar performance to NLP-efforts [2]
- Use as an *intelligent routing (IR)* tool
- Shown steady-state

* <http://www.cis.upenn.edu/~westand>



Recall: % OEs classified as such

Precision: % of edits classified OE that are actually vandalism

Conclusions

- Showed spatio-temporal properties can **locate Wikipedia-vandalism** comparably to NLP
 - Complementary; still some **advantages**:
 - Content/language independent
 - Harder to evade (analysis needed)
 - Faster (100+ edits/sec vs. 5 edits/sec)
- Spatio-temporal reputation as a **general-purpose technique for content-based access control?**
 - Email spam: SNARE [1] and PreSTA [5]
 - This work shows it also works for Wikipedia

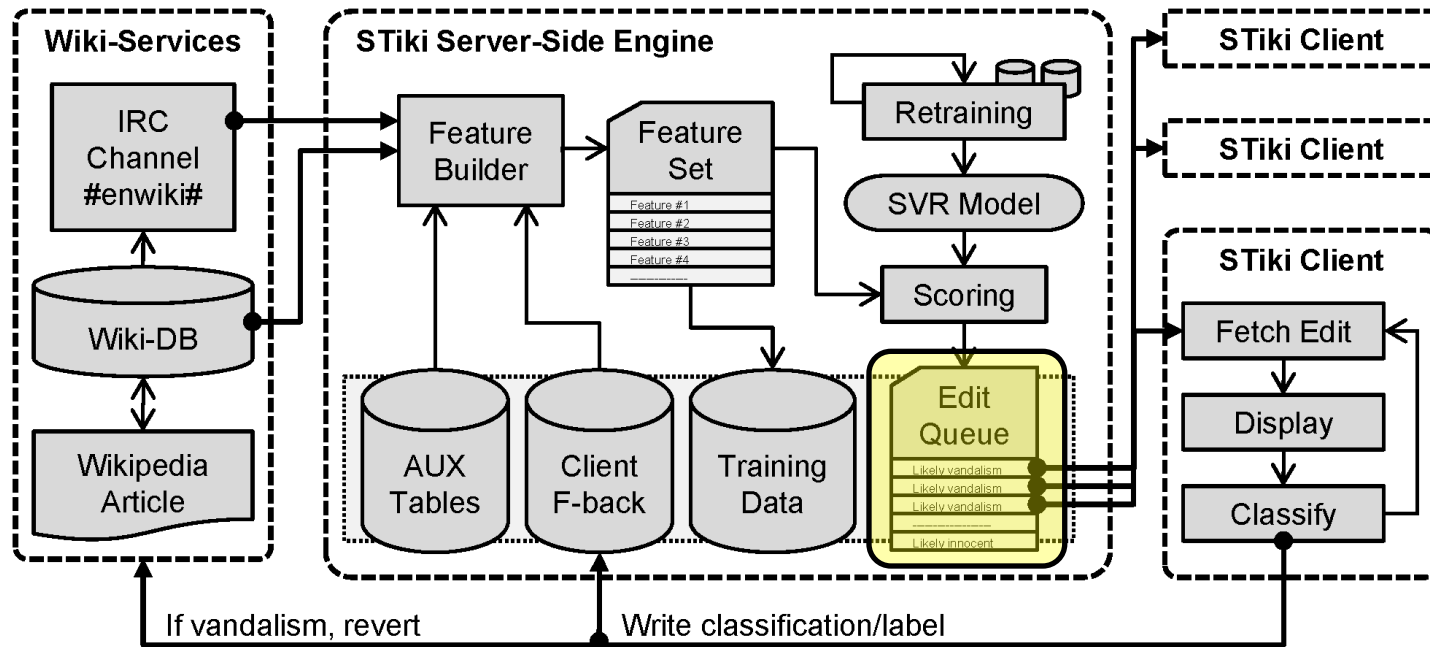
References

- [1] S. Hao, N.A. Syed, N. Feamster, A.G. Gray, and S. Krasser. **Detecting spammers with SNARE: Spatiotemporal network-level automated reputation engine.** In *18th USENIX Security Symposium*, 2009
- [2] M. Potthast, B. Stein, and R. Gerling. **Automatic vandalism detection in Wikipedia.** In *Advances in Information Retrieval*, pp. 663-668, 2008.
- [3] R. Priedhorsky, J. Chen, S.K. Lam, K. Achier, L. Terveen, and J. Riedl. **Creating, destroying, and restoring value in Wikipedia.** In *GROUP '07: The 2007 ACM Conference on Supporting Group Work*, pp. 259-268, 2007.
- [4] A.G. West. **STiki: A vandalism detection tool for Wikipedia.** <http://en.wikipedia.org/wiki/Wikipedia:STiki>. *Software*, 2010.
- [5] A.G. West, A.J. Aviv, J. Chang, and I. Lee. **Mitigating spam using spatio-temporal reputation.** *Technical report MS-CIS-10-04, University of Pennsylvania*, February 2010.

STiki [4]: A real-time, on-Wikipedia implementation of the technique



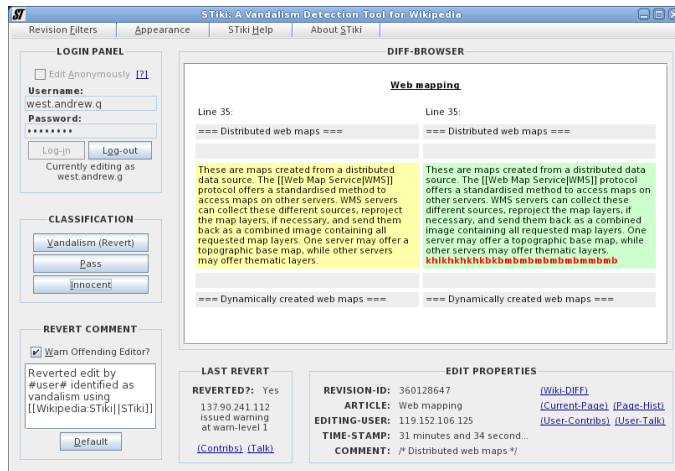
STiki Architecture



EDIT QUEUE: Connection between server and client side

- Populated: **Priority** insertion based on *vandalism score*
- Popped: GUI client shows likely vandalism first
- De-queued: Edit removed if another made to **same page**

Client Demonstration



STiki Client Demo

STiki Performance

- Competition inhibits maximal performance
 - Metric: **Hit-rate** (% of edits displayed that are vandalism)
 - Offline analysis shows it could be 50%+
 - Competing (often autonomous) tools make it $\approx 10\%$
- STiki successes and use-cases
 - Has reverted over **3500+** instances of vandalism
 - May be more appropriate in less patrolled installations
 - Any of Wikipedia's foreign language editions
 - Corporate Wiki's and other small installations
 - **Embedded vandalism**: That escaping initial detection.
Median age of STiki revert is 4.25 hours, 200 $\times$ conventional

Alternative Code Uses

- All code is available [4] and open source (Java)
- Backend (server-side) re-use
 - Large portion of **MediaWiki API** implemented (bots)
 - Trivial to add new features (including NLP ones)
- Frontend (client-side) re-use
 - Useful whenever edits require **human inspection**
- Data re-use
 - Corpus building; crowd-sourcing
 - Incorporate **vandalism score** into more robust tools

Future Direction: Wiki-Spam

- Many people “see” vandalism and do nothing:
 - Becomes “**embedded**” for days/weeks accumulating views
 - Traffic **spikes**: During American Idol finale, the “Crystal Bowersox” article was vandalized for just 28 seconds, but 12,000+ viewed the page during this duration.
 - Shows **evade-ability, apathy**, or both
- What if vandalism was spam?
 - If immature vandalism can get this many views, what about the less detectable and **incentivized** spam?
 - Could it be more **profitable** than email spam?
 - What evasion **strategies** would work best?