

Spatio-Temporal Analysis of Wikipedia Metadata and the STiki Anti-Vandalism Tool* †

Andrew G. West
University of Pennsylvania
Philadelphia, PA, USA
westand@cis.upenn.edu

Sampath Kannan
University of Pennsylvania
Philadelphia, PA, USA
kannan@cis.upenn.edu

Insup Lee
University of Pennsylvania
Philadelphia, PA, USA
lee@cis.upenn.edu

ABSTRACT

The bulk of Wikipedia anti-vandalism tools require natural language processing over the article or `diff` text. However, our prior work demonstrated the feasibility of using spatio-temporal properties to locate malicious edits. STIKI is a real-time, on-Wikipedia tool leveraging this technique.

The associated poster reviews STIKI’s methodology and performance. We find competing anti-vandalism tools inhibit maximal performance. However, the tool proves particularly adept at mitigating long-term embedded vandalism. Further, its robust and language-independent nature make it well-suited for use in less-patrolled Wiki installations.

Categories and Subject Descriptors

H.5.3 [Group and Organization Interfaces]: *collaborative computing, computer-supported cooperative work*;

K.6.5 [Management of Computing and Information Systems]: Security and Protection

General Terms

Design, Management, Human Factors, Security

1. SPATIO-TEMPORAL DETECTION

We informally define Wikipedia *vandalism* to be any revision that is non-value adding, offensive, or destructive in its removal of content. Detecting vandalism is difficult; it has many varied and subtle forms.

To this end, our prior research [3] investigated the spatio-temporal properties of metadata as a means of vandalism detection. The *metadata* of an edit includes: the (1) time-stamp of the edit, (2) article being edited, (3) user-name or IP of the editor, and (4) the revision comment. Meanwhile,

*This research was supported in part by ONR MURI N00014-07-1-0907. POC: Insup Lee, lee@cis.upenn.edu

†This poster complements a *WikiSym ‘10* demonstration of similar focus, it (this poster) concentrates on STIKI’s underlying approach and performance moreso than the software tool.

temporal properties are a function of the time at which an event occurs and *spatial* properties are appropriate wherever a distance or membership function can be defined.

Our prior work [3] identifies ten spatio-temporal properties (see Tab. 1) that are effective in locating malicious edits. *Simple features* include the edit time-of-day, revision comment length, *etc.*. *Aggregate features* combine time-decayed behavioral observations (feedback) to create reputations [2] for single entities and spatial groupings thereof.

2. THE STIKI TOOL

STIKI [1] exploits the aforementioned logic. It consists of:

- SERVER-SIDE ENGINE: Listens on IRC for a Wikipedia edit, fetches metadata, and compiles the feature-set. Machine learning assigns a real-value *vandalism score*, which is the priority for insertion into the *edit queue*.
- CLIENT-SIDE GUI: Pops the edit queue, presenting likely vandalism to users, via colored edit `diffs` (see Fig. 1).

An edit is also de-queued if a newer one is made on the same article. A STIKI workflow diagram is given in Fig. 2. Both the GUI executable and source are available at [1].

3. STIKI PERFORMANCE

STIKI has been used to revert over 2k instances of vandalism, yet the *hit-rate* (the percentage of time vandalism is displayed) has failed to meet off-line expectations [3]. Consider that the median active duration (time in queue) of the 10k most poorly scoring edits is around 3 minutes: The many autonomous anti-vandalism tools/bots prevent STIKI from displaying much of the vandalism it finds. While STIKI’s hit-rate is $\approx 10\%$, analysis has shown it would be $50\%+$ (to a reasonable depth) if competing tools were not present.

Thus, STIKI and its language-independence may be well suited for less-patrolled settings (*e.g.*, foreign language editions of Wikipedia or corporate Wiki’s). Even so, STIKI has proven capable of finding *embedded vandalism* on English Wikipedia – that which escapes initial detection. The median age of an edit reverted by STIKI is approximately 4.25 hours, nearly $200\times$ that of conventional reversions.

4. EXTENSION & FUTURE WORK

To remedy the modest hit-rate, extension of the spatio-temporal feature-set is planned. With the inclusion of lightweight natural-language features, STIKI could also evolve into a general-purpose anti-vandalism tool. The STIKI framework will provide a convenient test-bed for these new features and other future vandalism mitigation strategies.

References

- [1] A. G. West. STiki: A vandalism detection tool for Wikipedia. <http://en.wikipedia.org/wiki/Wikipedia:STiki>, 2010. Software.
- [2] 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, Department of Computer and Information Science, February 2010.
- [3] A. G. West, S. Kannan, and I. Lee. Detecting Wikipedia vandalism via spatio-temporal analysis of revision metadata. In *EUROSEC '10: Proceedings of the Third European Workshop on System Security*, pages 22–28, Paris, France, 2010.

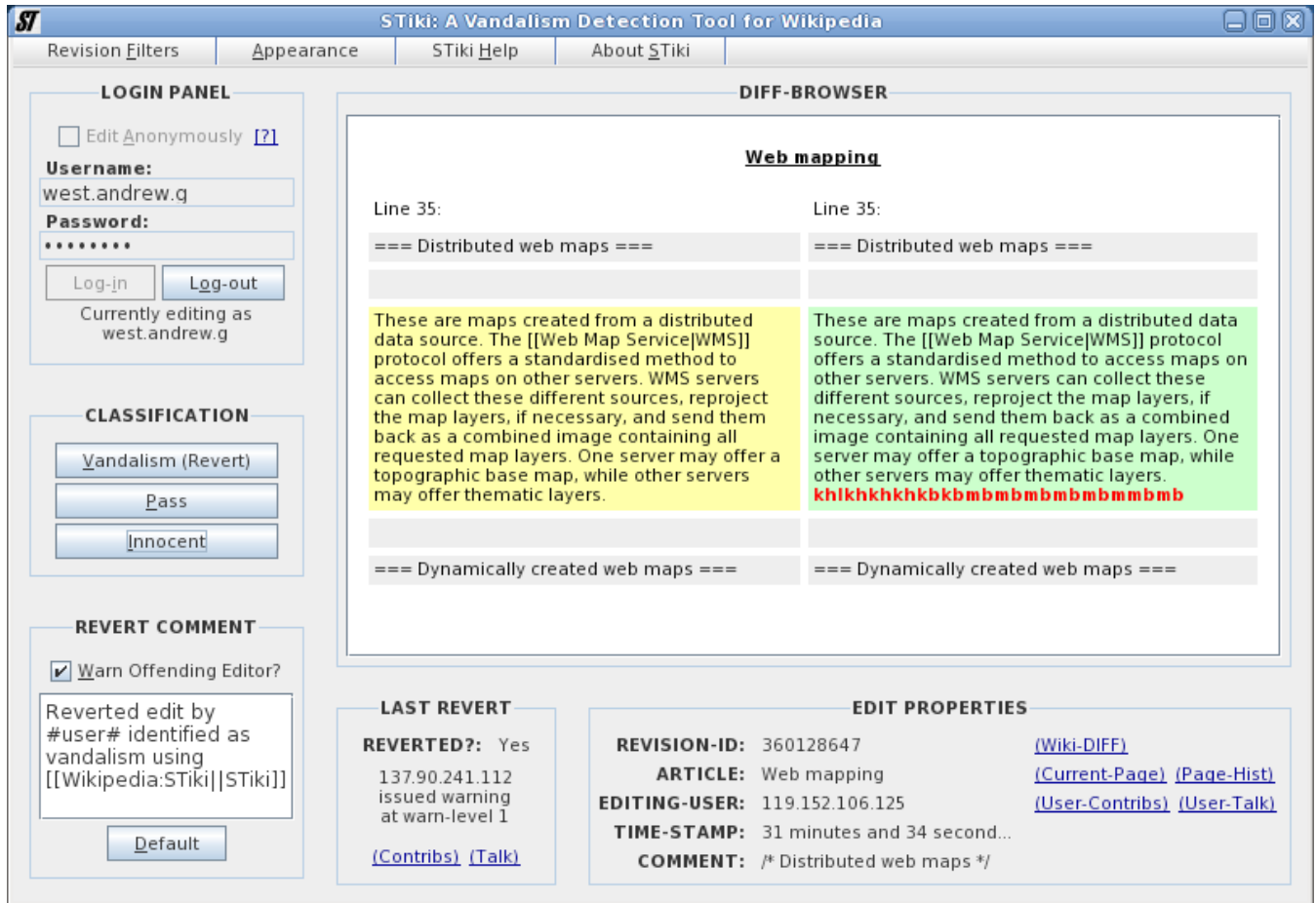


Figure 1: STIKI GUI displaying a revision exhibiting vandalism (nonsense).

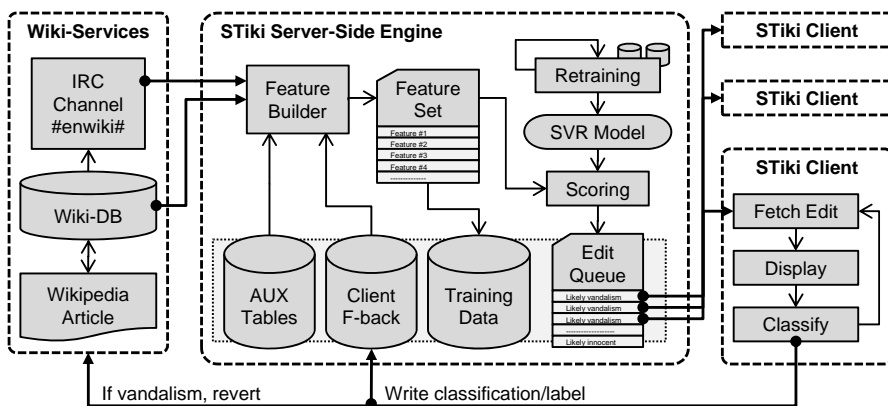


Figure 2: Simplified STIKI workflow diagram.

#	FEATURE
1	Edit time-of-day
2	Edit day-of-week
3	Time-since (TS) editor registration (first-edit)
4	TS article last edited
5	TS editor last vandalized
6	Rev. comment length
7	Article reputation
8	Categorical reputation (grouping over articles)
9	Editor reputation
10	Geographical reputation (grouping over editors)

Table 1: STIKI features [3].

Spatio-Temporal Analysis of Wikipedia Metadata and the STiki Anti-Vandalism Tool

Andrew G. West • Sampath Kannan • Insup Lee

Department of Computer and Information Science -- University of Pennsylvania -- Philadelphia, PA, USA

BIG IDEA

Spatio-temporal properties of edit metadata (editor, article, timestamp, and revision comment) can be leveraged to detect Wikipedia vandalism comparably to NLP based methods:

- **Simple features** (i.e., time-of-day), in addition to historical *reputations* for editors, articles, and spatial groupings thereof are used.
- Such features have **language-independence, efficiency, and robustness** not found in traditional detection mechanisms (i.e., NLP).
- **STiki** [1], is a real-time, on-Wikipedia tool utilizing the technique, already shown feasible off-line in our prior work [3].

EDIT LABELING: ROLLBACK

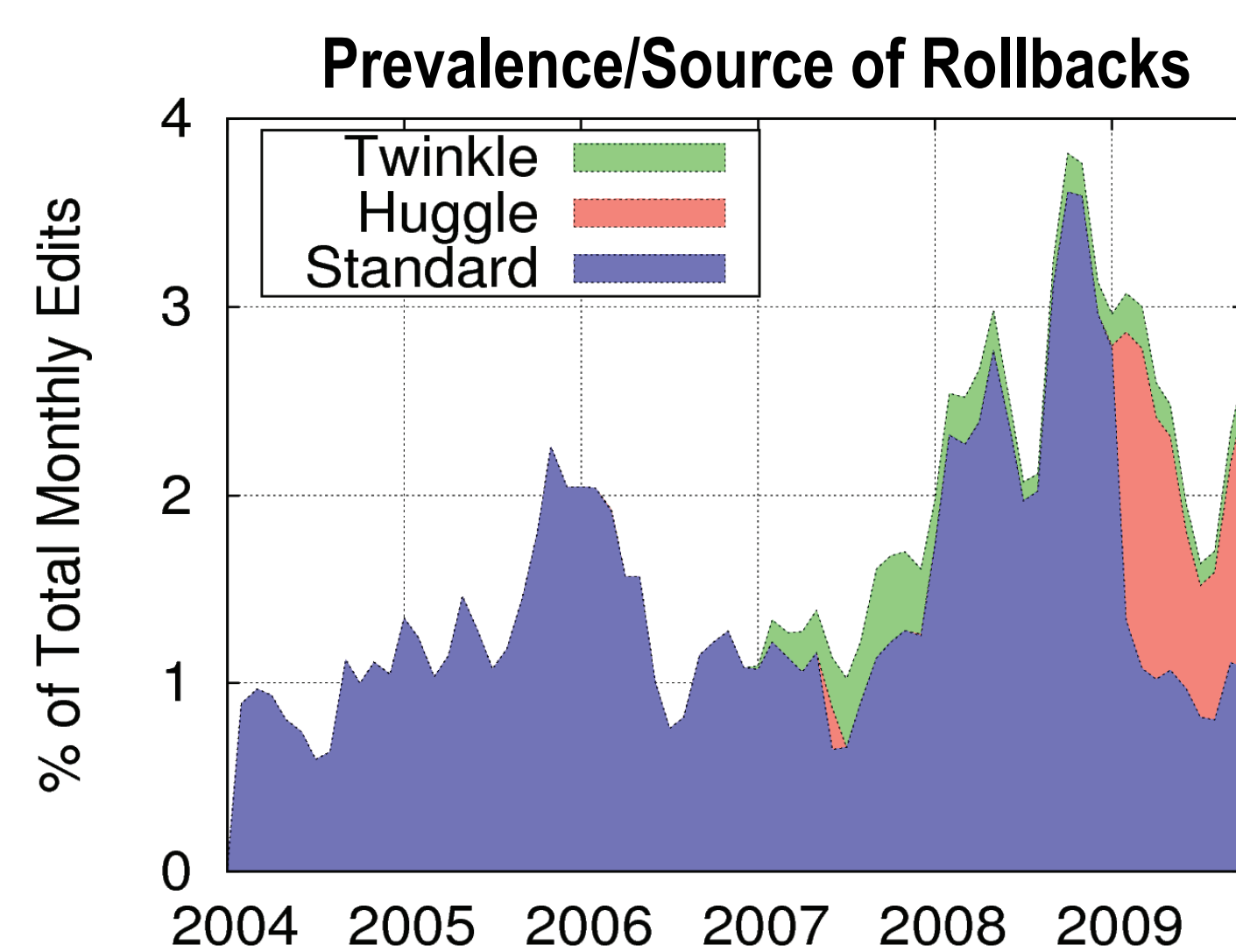
Need to label edits exhibiting vandalism (ex-post facto) to:

1. Show *features* effective (and eventually to train over them)
2. Form basis of historical reputations (vandalism = misbehavior)

ROLLBACK

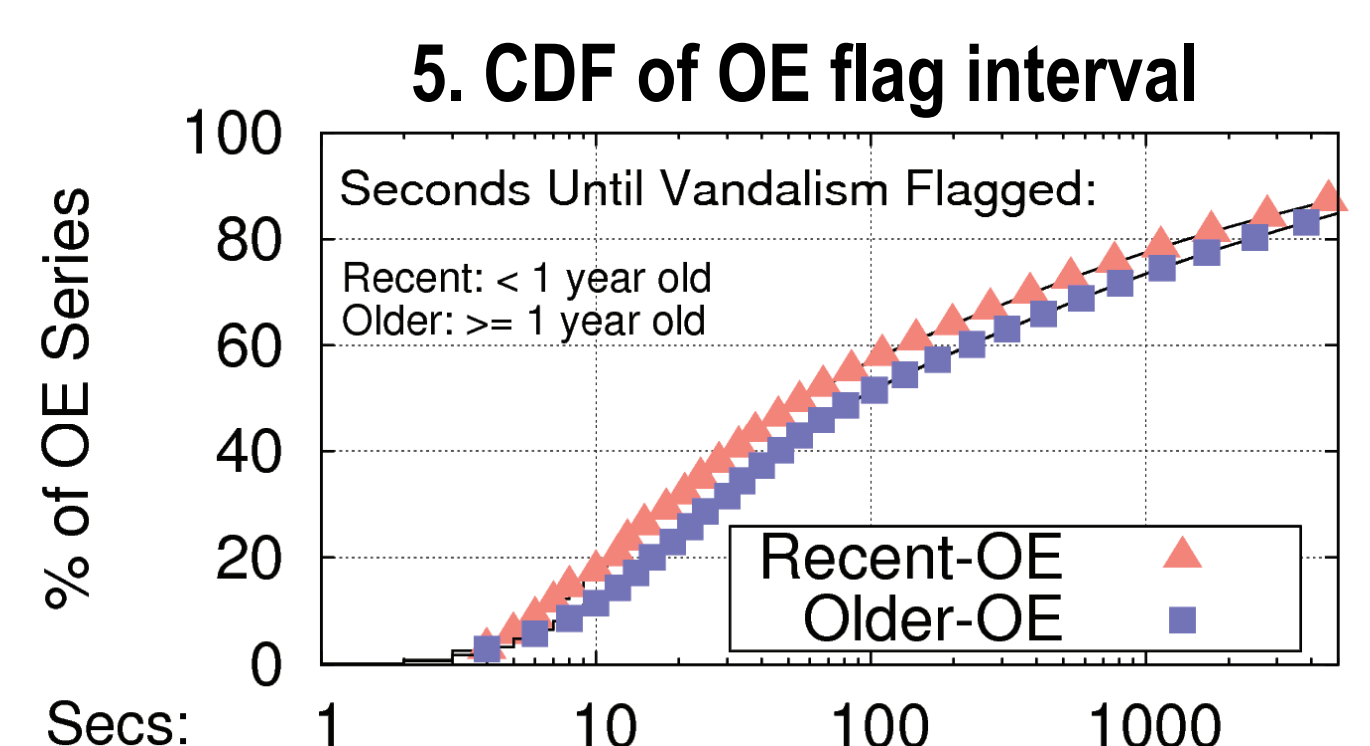
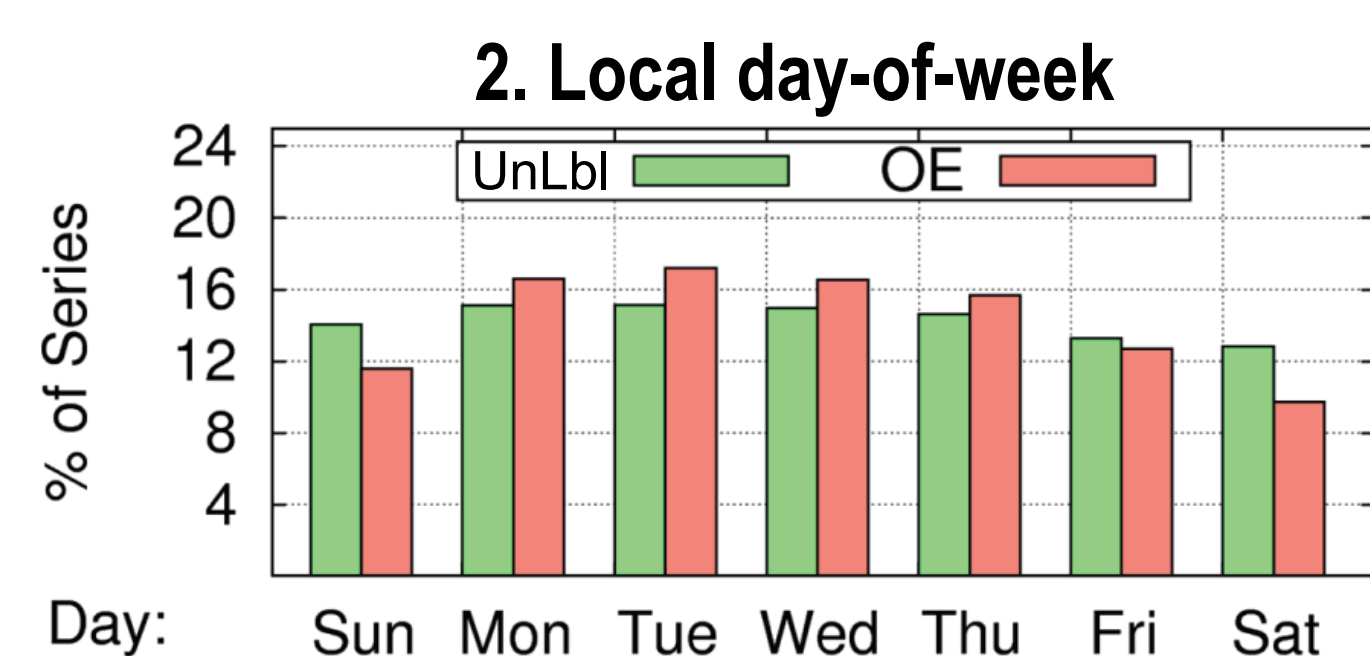
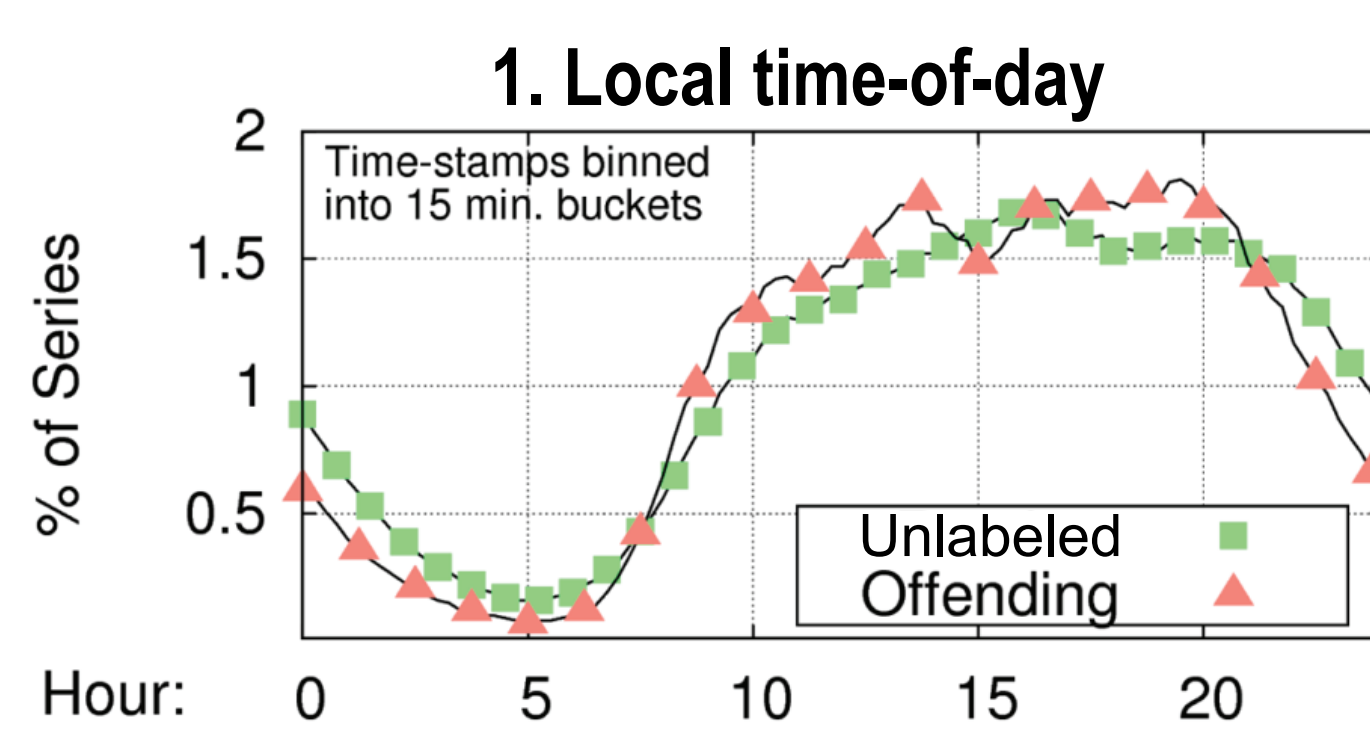
An administrative form of [undo]:

- Revisions undone are **Offending Edits (OEs)**, likely vandalism
- Autonomously parse-able
- Trusted feedback (admins)
- Vandalism defined case-by-case



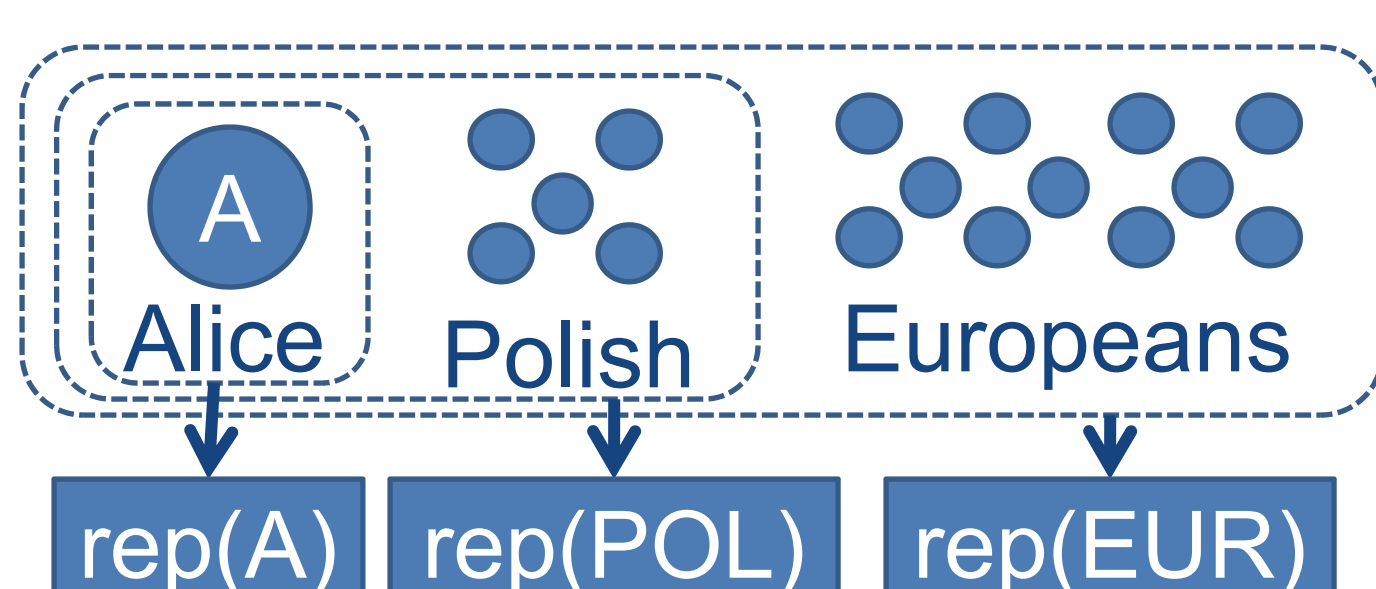
SIMPLE SPATIO-TEMPORAL FEATURES

#	FEATURE
1	Edit time-of-day: (see right)
2	Edit day-of-week: (see right)
3	Time-since article edited: Frequently edited pages are vandalism targets (visibility)
4	Time-since editor reg.: Long-time editors are rarely problematic (Sybil attack)
5	Time-since last user OE: Good editors rarely vandalize (+OEs flagged quickly; see left)
6	Revision comment length: Vandals leave shorter comments (lazy + bandwidth)



AGGREGATE FEATURES (REPUTATIONS)

IDEA: Use **entity-specific** reputation; augment with **spatial reputations** [2], which will have more historical data.



#	FEATURE
7	Article reputation
8	Category reputation Spatial grouping over articles
9	Editor reputation
10	Country reputation Spatial grouping over editors

The **reputation function**:

- Summation over time-decayed feedback (vandalism via rollback)
- Spatial reputation's are normalized by the group size

RANK	COUNTRY	%-OEs
1	Italy	2.85%
2	France	3.46%
13	United States	11.63%
14	Australia	12.08%

ARTICLE*	#OEs
Wikipedia	5589
United States	2161
World War II	1886

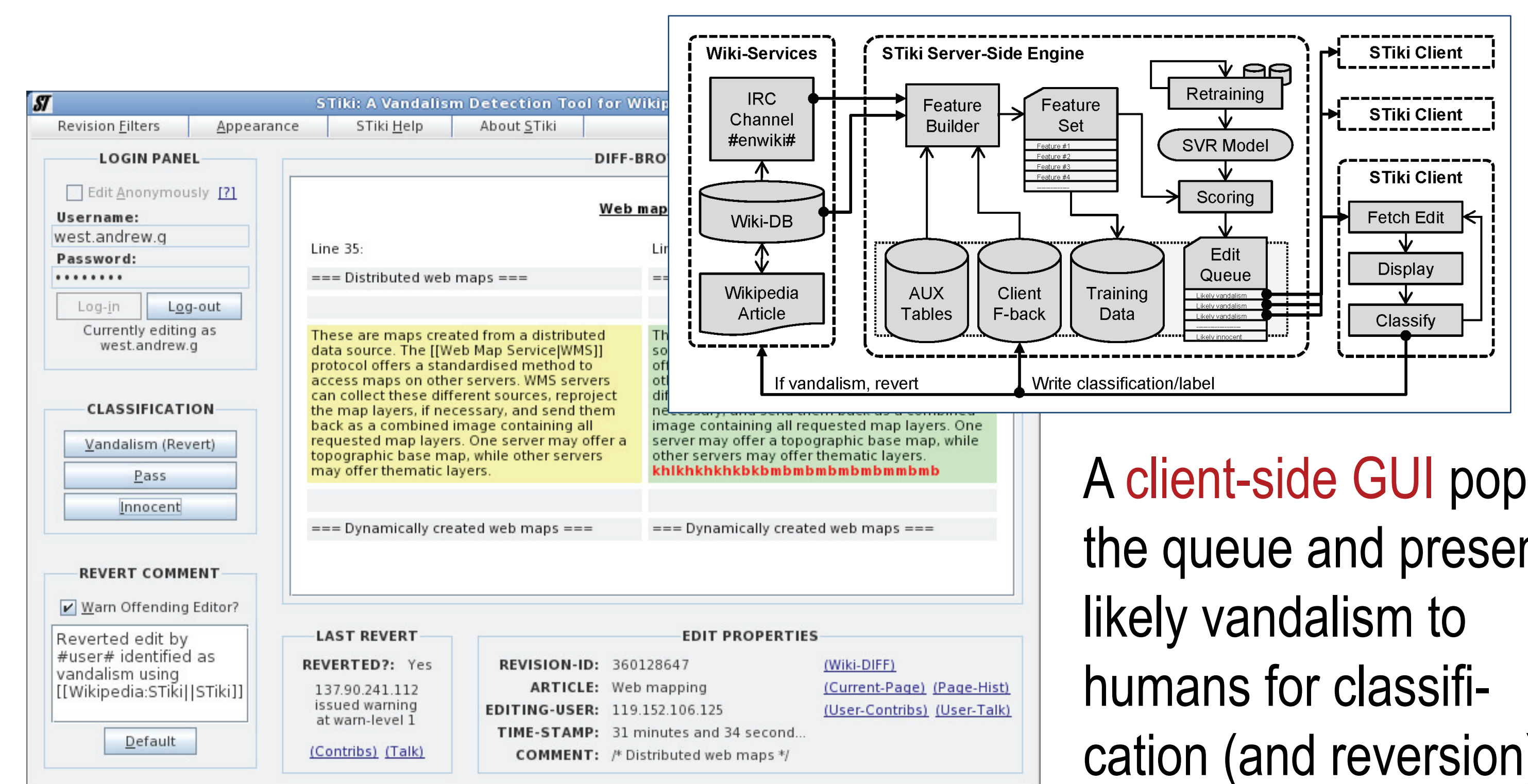
CATEGORY (w//100+ pgs)	PGs	OEs/PG
World Music Award Winners	125	162.27
Characters of Les Miserables	135	146.88
Former British Colonies	145	141.51

* List sanitized for appropriateness

Vandalism is **clustered non-uniformly** throughout article and editor space, making membership in such groupings **behavior predictive**.

THE STiki TOOL

STiki [1] leverages these features in real-time. The **server-side engine** calculates a real valued *vandalism score* (via machine-learning) for all edits, which is the insertion priority into the *edit queue*.



A **client-side GUI** pops the queue and presents likely vandalism to humans for classification (and reversion).

An edit is also de-queued if a more recent edit is made on the same article.

STiki PERFORMANCE & FUTURE

Performance metric: hit-rate (% of displayed edits that are vandalism):

- Offline-analysis [3] shows hit-rate should be 50%+
- In fact, $\approx 10\%$ due to **competing tools/bots** (often autonomous)

Successes and alternative uses:

- STiki has reverted over 5000 instances of vandalism on *en-wiki*.
- Combats **embedded vandalism** well. Median age of vandalism reverted by STiki is 4.25 hours, nearly 200 \times of conventional reverts.
- May be best suited for **less-patrolled Wikis** (e.g. foreign lang. eds.)

Future improvements:

- Include lightweight **NLP** features (a general-purpose tool)
- Alternative detection (link spam? more ST-features?)



REFERENCES & ACKNOWLEDGEMENTS

- [1]: A. G. West. *STiki: A vandalism detection tool for Wikipedia*. <http://en.wikipedia.org/wiki/Wikipedia:STiki>, 2010. Software.
- [2]: 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, Feb. 2010.
- [3]: A. G. West, S. Kannan, and I. Lee. *Detecting Wikipedia vandalism via spatio-temporal analysis of revision metadata*. In *EUROSEC '10: Proc. of the 3rd European Workshop on System Security*, pages 22-28, Paris, France, Apr. 2010.



This research was supported in part by ONR MURI N00014-07-1-0907. POC: Insup Lee, lee@cis.upenn.edu