1-1-2010

Detecting Wikipedia Vandalism via Spatio-Temporal Analysis of Revision Metadata

Andrew G. West
University of Pennsylvania

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

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

University of Pennsylvania Department of Computer and Information Science Technical Report No. MS-CIS-10-05.

Detecting Wikipedia Vandalism via Spatio-Temporal Analysis of Revision Metadata

Abstract
Blatantly unproductive edits undermine the quality of the collaboratively-edited encyclopedia, Wikipedia. They not only disseminate dishonest and offensive content, but force editors to waste time undoing such acts of *vandalism*. Language- processing has been applied to combat these malicious edits, but as with email spam, these filters are evadable and computationally complex. Meanwhile, recent research has shown spatial and temporal features effective in mitigating email spam, while being lightweight and robust.

In this paper, we leverage the spatio-temporal properties of revision metadata to detect vandalism on Wikipedia. An administrative form of reversion called *rollback* enables the tagging of malicious edits, which are contrasted with nonoffending edits in numerous dimensions. Crucially, none of these features require inspection of the article or revision text. Ultimately, a classifier is produced which flags vandalism at performance comparable to the natural-language efforts we intend to complement (85% accuracy at 50% recall). The classifier is scalable (processing 100+ edits a second) and has been used to locate over 5,000 manually-confirmed incidents of vandalism outside our labeled set.

Comments
University of Pennsylvania Department of Computer and Information Science Technical Report No. MS-CIS-10-05.

Detecting Wikipedia Vandalism via Spatio-Temporal Analysis of Revision Metadata*

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

Blatantly unproductive edits undermine the quality of the collaboratively-edited encyclopedia, Wikipedia. They not only disseminate dishonest and offensive content, but force editors to waste time undoing such acts of *vandalism*. Language-processing has been applied to combat these malicious edits, but as with email spam, these filters are evadable and computationally complex. Meanwhile, recent research has shown spatial and temporal features effective in mitigating email spam, while being lightweight and robust.

In this paper, we leverage the spatio-temporal properties of revision metadata to detect vandalism on Wikipedia. An administrative form of reversion called *rollback* enables the tagging of malicious edits, which are contrasted with non-offending edits in numerous dimensions. Crucially, none of these features require inspection of the article or revision text. Ultimately, a classifier is produced which flags vandalism at performance comparable to the natural-language efforts we intend to complement (85% accuracy at 50% recall). The classifier is scalable (processing 100+ edits a second) and has been used to locate over 5,000 manually-confirmed incidents of vandalism outside our labeled set.

1. INTRODUCTION

Wikipedia [2], the collaboratively edited encyclopedia, is among the most prominent websites on the Internet today. Despite evidence suggesting its accuracy is comparable to that of traditional encyclopedias [7], Wikipedia's credibility is criticized. These criticisms are often rooted in inaccuracies resulting from blatantly *unproductive* or *false* edits. Such malicious edits constitute *vandalism*, which we informally define as any edit which is non-value adding, offensive, or destructive in its removal of content. The motives of vandals vary (profit, narcissism, political agendas, *etc.*), but their impact is large. One study [13], concluded there have been hundreds of millions of damaged page-views.

Detecting vandalism on Wikipedia is difficult, and current techniques often utilize natural language processing (NLP). While blatant occurrences may be easy to catch (*e.g.*, offensive speech), finding more subtle incidents (*e.g.*, someone placing their own name in a historical narrative) is extremely challenging. For this reason, we were motivated to examine revision metadata as an alternative means of detection. *Metadata* includes information about the edit such as; when it was made, who made it, and what article it was made on.

In this paper, we show that malicious edit metadata exhibits spatial and temporal properties (see Sec. 2.2) unlike those associated with innocent edits. *Simple features* include the edit time-of-day, revision comment length, *etc.*. More interesting are the *aggregate features*, which combine time-decayed behavioral observations (feedback) to create *reputation values* for single-entities (user, article) and spatial-groupings thereof (geographical region, content categories). Feedback is gathered using an administrative (and therefore, trusted) form of reversion called *rollback*.

Exploiting these features, we produce a lightweight classifier capable of identifying vandalism at rates comparable to NLP efforts. While NLP techniques have drawbacks (evadability, computational complexity, and language dependence) that our approach does not, it is our intention to complement such classifiers, not compete against them.

Summarily, the novelty of our approach is three-fold:

1. The methodology uses exclusively revision *metadata*.
2. By using *rollbacks* as the basis for vandalism tagging, the experimental set is large and accurate.
3. A small feature-set enables *lightweight* classification, which is practical at Wikipedia-scale.

Given the inherent difficulty in locating vandalism, we believe our classifier should be used as an *intelligent routing* tool (*i.e.,* rather than reverting edits automatically, it should direct humans to potential vandalism for inspection). We test our classifier in this manner – and produce a corpus of over 5,000 manually-verified incidents in the process.

2. BACKGROUND & RELATED WORK

2.1 Classification and Valuation

There have been many attempts to valuate the entities on Wikipedia – articles [14, 18], authors [3, 4], and individual edits have all been studied. The work presented herein, like all vandalism-focused efforts [12, 15], operates at the edit level. Although work at other granularity is related, for brevity we examine only edit-level techniques.

These systems begin by establishing tagged incidents of vandalism to analyze. Potthast *et al.* [12] rely on a manually labeled corpus. Others [13, 15] analyze revision comments and hash article content to determine where *reverts* have taken place. Our technique builds upon the latter, utilizing reverts initiated by privileged (and therefore, *trusted*) users called *rollbacks*. The benefits of rollback-based labeling are three-fold: (1) It can be completely automated, (2) It permits high confidence in the tags, and (3) It allows Wikipedia administrators to define vandalism on a per-case basis, rather than requiring an a priori definition.

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

Having a labeled set, systems next attempt to find properties that distinguish vandalism from non-malicious edits. NLP is used exclusively in vandalism-focused works. However, large feature counts make NLP computationally expensive, the features are non-intuitive, and it is unable to detect subtle instances of vandalism.

Wikipedia [2] itself also employs NLP techniques. Automated *bots* (*e.g.,* Cluebot), filters (*e.g.,* `abusefilter`), and editing assistants (*e.g.,* Huggle and Twinkle) all aim to locate acts of vandalism. Such bots/tools work via simple regular expressions and manually-authored rule sets.

Uniquely, our analysis does not require the article text or `diff` of the revision. Instead, we use only *revision metadata* (time-stamp, page-title, *etc.*) to make classifications. An intuitive argument is made for each feature, whose quantity is minimized to maintain efficiency.

2.2 Spatio-Temporal Properties/Reputation

Temporal properties are a a function of the time at which various events occur (*e.g.,* time-of-day). *Spatial* properties are those rooted not just in physical space (*e.g.,* the geographical origin of an event), but also in abstract space (*e.g.,* graph topologies) – and are generally appropriate wherever a distance or membership function can be defined. In the first work in the domain, Hao *et al.* [8] identified 13 such properties and showed them effective in mitigating email spam. We term straightforward properties of this type *simple features*.

The notions of space and time can also be used to create quantitative *reputations* [10]. Assume we have *feedback* indicating principals who have mis-behaved and the time-stamps of these observations. Intuitively, recently bad entities will have poor reputation, and in the absence of negative feedback, this should heal over time. If an entity's individual history is inconclusive, one should aggregate and examine the reputations of spatially-adjacent entities, as behavioral patterns are often clustered (*i.e.,* homophily [11]).

In [17], we define a model that captures precisely these notions, and use it to filter email spam. That algorithm (see Sec. 4.2) is leveraged herein, where rollbacks act as feedback, and spatial functions – geographical and topical – are defined over users and articles, respectively. Values calculated in this manner are termed *aggregate features*.

Spatio-temporal properties have not been extensively studied on Wikipedia. Temporal properties were touched on by Wöhner and Peters [18] in their examination of article development. Others [13, 16] have examined the persistence of vandalism. Finally, Buriol *et al.* [5] reviewed the evolution (temporal) of intra-page-link topology (spatial).

3. DATA FOUNDATIONS

Our analysis began by parsing a 2009/11/03 XML dump of *enwiki* (English) revision metadata. The dump consists of ≈ 298 million edits, having the following pertinent fields: (1) **Time-stamp** of edit, in GMT locale. (2) **Article** being edited. (3) Registered user-name or IP of **user** doing the editing. (4) The **revision comment** left by the editor.

3.1 Finding Vandalism

Ordinary users of Wikipedia have the ability to *revert* back to prior versions in an article's history. Upon clicking the 'revert' link, such users can provide justification via a comment field, and then commit the change. For a privileged class of trusted users (≈4,700 with `sysop` or `rollbacker` rights), this process is expedited via *rollback*. Rollbacks are initiated by a single-click, and the Wikipedia software automatically inserts a comment of the form: "Reverted edits by x to last version by y." Wikipedia states the feature should only be used to revert "blatantly unproductive" edits [2].

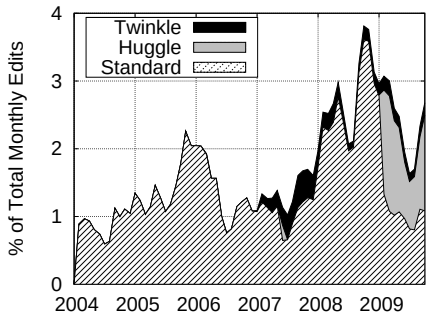

Figure 1: Rollback Quantity/Source (stacked)

Revision comments of privileged editors (per a permissions table) are searched for strings of this form[1], and matching ones constitute *flagging edits*. Using the comment, we attempt to find the rolled-back or *offending edit* (OE) and succeed in 99.61% of cases. With high-confidence, we believe that OEs are those exhibiting vandalism.

All edits not in set 'OE' are said to be 'Random', as their legitimacy cannot be ascertained. 5.7 million OEs are identified and Fig. 1 shows the prevalence of rollbacks/OEs over time and the emergence of editing assistants (*e.g.*, Huggle). Variation in OE-rate often reflects administrative changes. For example, the drop in OE-rate in mid-2009 is likely due to the implementation of a basic filter (`abusefilter`) which prevents trivially offensive edits from being committed.

3.2 Reducing the Working Set

Wikipedia is vast, yet only 37.72% of pages are encyclopedic in nature – the remainder contain discussion, administration, and coordination. The encyclopedia content forms name-space zero (NS0) – where much of the interesting work takes place, including 71.08% of all edits, and 91.15% of all OEs. Moving forward, we consider only edits in NS0.

Further, as Fig. 1 suggests, Wikipedia is a volatile environment in which to measure vandalism. Therefore, it is most meaningful to examine recent history. We consider only NS0 edits within a year of our dump date (*i.e.,* 2008/11/03 onward)[2]. The resulting set contains 50 million edits (16.78% of total) and nearly 1.8 million OEs (31.40% of all OEs).

4. FEATURES

We divide our feature set into two categories: (1) *simple features*, that operate primarily on the metadata associated with a single edit (sometimes using a look-up table), and (2) *aggregate features*, that compile OE histories for entities into reputation values. Discussion of simple features is abbreviated in order to concentrate on the more interesting aggregate ones. Our feature choice and presentation is motivated by a similar work [8] in the spam email domain.

With one exception, our feature space is straightforward. *Registered* Wikipedia users have persistent identifiers (usernames) allowing their contributions to be tracked across time

[1]Editing assistants like Huggle/Twinkle leave comments of a varied form, which are also parsed. We disregard rollbacks initiated by *bots*, to maintain a human-validated set.

[2]Features requiring historical information will be permitted to utilize edit-data prior to the stated cut-off.

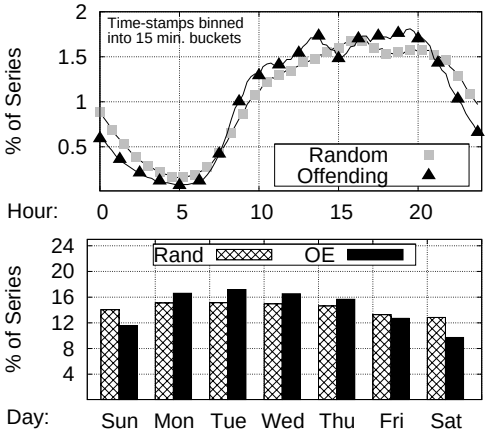

Figure 2: Edit (a) Time-of-Day, (b) Day-of-Week

and machines. Such tracking is less dependable for *anonymous* users identified by (possibly dynamic) IP addresses. However, the presence of an IP enables the use of geolocation data [1], on which several features are built. Registered users' IPs could be used in a production setting, but they are privatized from our data-set. We remove registered users from the analysis of those features they are ineligible for.

4.1 Simple Features

4.1.1 Time-of-Day and Day-of-Week

Description: For users identified by IP address, we use IP-geolocation data [1] to determine the GMT offset at the edit origin. Combining this with the edit time-stamp (in GMT locale), we determine the local time-of-day and day-of-week when an edit was committed. See Fig. 2a and Fig. 2b.

Discussion: First, we see that edit-rates follow diurnal patterns, which is intuitive (*i.e.,* most edits take place during typical waking hours). Somewhat surprisingly, OEs are most prominent (relatively) between 8AM and 8PM. Weekdays, as opposed to weekends, see more edit and OE activity. An edit committed on Tuesday is *twice* as likely to be vandalism compared to one committed on a Saturday.

4.1.2 Time-Since User Registration

Description: Wikipedia privatizes registered user's sign-up dates – we estimate them by storing the time-stamp of a user's first edit. This enables us to calculate the 'time-since registration' (of the edit author) for all edits. This analysis is attempted for anonymous users, but dynamic concerns (*i.e.*, DHCP) may encode unexpected results. See Tab. 1.

Discussion: Intuitively, one would expect long time members of the Wikipedia community to be vested in its growth and familiar with its policies. Conversely, malicious editors may employ a Sybil attack [6], creating temporary accounts to abuse the associated benefits. Time-since registration should encode such behaviors. Indeed, the median[3] of this metric for random edits is 10,000× that of OEs (for registered users), and similarly strong results are found for anonymous users. Clearly, long time participants in Wikipedia contribute little to the vandalism problem.

4.1.3 Time-Since Last Article Edit

[3]We prefer median to compare metrics. The long tail of time-based distributions tends to skew average calculations.

Time-Since User Reg.	OE	RAND
Regd. edits, median (days)	0.07	765
Anon. edits, median (days)	0.01	1.97
Time-Since Article Edited	**OE**	**RAND**
All edits, median (hrs.)	1.03	9.67
Time-Since Last User OE	**OE**	**RAND**
Regd. edits, median (days)	0.17	63.28
Anon. edits, median (days)	0.14	29.26

Table 1: Comparison of 'Time-Since' Features

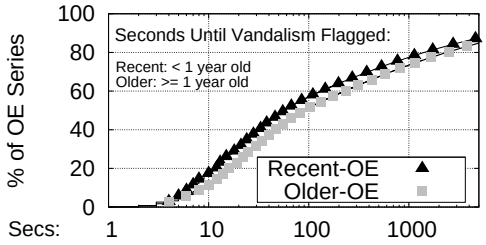

Figure 3: CDF of OE-flag Intervals

Description: Calculating the time between an edit and the last edit on the same article is straightforward. See Tab. 1.

Discussion: Previous work [13, 18] has shown often-edited articles attract the bulk of vandalism. Our own data confirms this; while only 2.4% of articles have 5+ OEs, these pages contain 51.8% of all edits. As Tab. 1 shows, this metric succeeds in encoding these properties.

4.1.4 Time-Since Last User OE

Description: Given an edit, we find the time since the editor's last OE was made. This is subtle; although a poor edit may be made at time t_{oe}, that edit does not enter the OE set until it is flagged at time $t_{\mathsf{flag}} > t_{\mathsf{oe}}$. This valuation is not possible for editors with no OE history. See Tab. 1.

Discussion: Intuitively, malicious editors may go on "vandalism sprees", committing many bad edits in a short timeframe. Crucial to our detection of the subsequent vandalism is that the preceding ones are flagged quickly. Fig. 3 shows the median time-to-flag is ≈80 seconds – remarkably quick. 49% of OEs are capable of being scored in this manner (that is, a prior OE exists for the editing user). Where it can be calculated, it is exceedingly effective. In the median case, an editor committing an OE had also misbehaved within the last 3-4 hours; compared to 30-60 days for non-OEs.

4.1.5 Revision Comment Length

Description: For each edit, the editor is given the opportunity to briefly summarize the changes made. We examine the *length* (spatial) in characters of this string. See Tab. 2.

Discussion: We see that the comments left with OEs are 43% of the size, on average, of those with random edits. This may be attributable to laziness. However, as [8] observed with email spam, small message sizes minimize bandwidth use. Analogously in Wikipedia, ignoring this optional field may allow one to make more (possibly, bad) edits.

4.1.6 Registered User Properties

Description: Most OEs (85%) are committed by anonymous users. Nonetheless, we have more data on registered ones, namely, (1) whether they have any special editing privileges, and (2) if they are an automated 'bot'. See Tab. 2.

FEATURE	OE	RAND
Rev. Comm. (avg. chars.)	17.73	41.56
Anonymous Editor (%)	85.38%	28.97%
Bot-Editor (%)	00.46%	09.15%
Privileged Editor (%)	00.78%	23.92%

Table 2: Non-Temporal Simple Features

Discussion: Given that we do not classify over registered users in Sec. 5, we truncate our discussion. It suffices to say that bots and privileged users are exceedingly well-behaved while being significant contributors (again, see Tab. 2).

4.2 Aggregate Features

Intuitively, when assessing an edit it is desirable to know the history of the entities involved (*i.e.,* how has this *user* behaved in the past? Has this *article* proven controversial?). When an entity has no history, it may prove helpful to examine the history of spatially adjacent entities. We concisely encode these notions via spatio-temporal *reputation values*, where OEs provide behavioral context. Crucially, Fig. 3 shows OEs are flagged quickly, suggesting that reputation values will not be latent in their quantitative assessments.

A general model to compute these values can be found in our prior work [17], which we simplify for use here. Let α be an entity and $g = G(\alpha)$ be the group it belongs to per spatial grouping function $G()$.

- $\boldsymbol{oe\_hist(g)}$ is a function returning a list of OE timestamps, t_{oe}, for all OEs mapping to some entity $\beta \in g$, where g is a set of entities.
- $\boldsymbol{decay(t)}$ is a time-decay function. We use $decay(t) = 2^{-\Delta t/h}$ where $\Delta t = (t_{\text{now}} - t_{\text{oe}})$ and h is the half-life. This function ensures more temporally relevant (*i.e.,* recent) observations are weighted more heavily.

We now define the reputation of a group $g = G(\alpha)$ as follows:

$$\boldsymbol{rep(g)} = \sum_{t_{\text{oe}} \in oe\_hist(g)} \frac{decay(t_{\text{oe}})}{size(g)} \quad (1)$$

Low $rep()$ values are indicative of well-behaved (or non-controversial) groupings, and vice-versa. All $rep()$ values calculated using the same $G()$ are strictly comparable[4] (*i.e.,* can be relatively interpreted). We will demonstrate these values, are in fact, *behavior predictive*. We first calculate reputation for groupings of trivial cardinality (*i.e.*, $|g| = 1$) including users and articles (entities), then for broader spatial groupings thereof (*i.e.*, $|g| > 1$), namely geographic region and topical categories. We use $h = 10$ days as our half-life parameter, a value shown effective in spam defense [17], and whose optimization is the subject of future work.

4.2.1 Article Reputation

Description: Valuating an edit's article reputation is a straightforward application of $rep()$ where α is the page being modified, $G(\alpha) = \alpha = g$, and $|g| = 1$. Time-stamp t_{now} is the time at which the edit was committed. See Fig. 4a.

Discussion: Certain topics are inherently controversial and are frequent targets of vandalism. Others incur temporally-variable abuse (*e.g.*, political candidates near elections). Article reputation is well-equipped to handle both cases.

[4]Unable to apply an absolute interpretation to $rep()$ values, the x-axis of Fig. 4 and Fig. 5 is presented relatively.

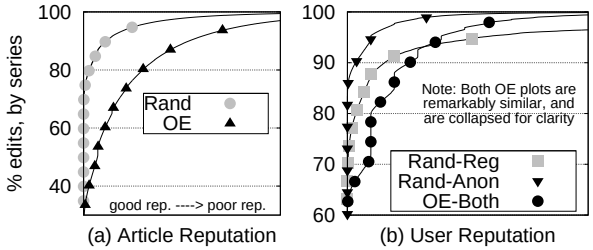

Figure 4: CDF of Article/User Reputation

ARTICLE	#-OEs	ARTICLE	#-OEs
G.W. Bush	6546	Adolf Hitler	2612
Wikipedia	5589	United States	2161

Table 3: Articles w/Most Offending-Edits

Tab. 3 indicates which articles have the most OEs[5], and is representative of the non-uniform distribution of OEs across article space. Fig. 4a best visualizes the predictive nature of article-reputation, which is 4× higher, on average, for OE edits than random ones. Nearly 85% of OEs show non-zero article reputations, compared to 45% of random edits.

4.2.2 User Reputation

Description: User reputation is calculated identically to article reputation, except α is the edit author. See Fig. 4b.

Discussion: The existence of habitual offenders is an underlying assumption of nearly all reputation systems. Indeed, Wikipedia is full of such users, but user-reputation is nonetheless a metric which is difficult to interpret.

Consider a registered editor who makes 500+ edits/day, accumulating several OEs due to innocent errors. This editor will have a similar reputation to a vandal who created an account only to evade page protections. Moreover, the possible dynamism of anonymous user IDs is troublesome.

The obvious solution, as some find [12], is to *normalize* reputation by the number of user-edits. We reject this notion; it provides an easy means for one to manipulate his/her own reputation (by making many minor, innocent edits).

These difficulties are reflected in the poor (stand-alone) performance of the metric, as per Fig. 4b. Regrettably, we find the reputations of registered-editors making *random* edits are sometimes poorer than those of *offending* anonymous users. Further, anonymous OE editors only have non-zero reputation about 40% of the time. We still classify using this feature because when correlated with other features (*e.g.*, time-since registration), it becomes more meaningful.

4.2.3 Category Reputation

Description: Wikipedia organizes content into topic-based (*e.g.*, "American Presidents") and administrative (*e.g.*, "articles in need of citation") categories. We calculate reputation using topical categories as a spatial grouping over articles.

Wikipedia provides a listing of categories and their members. We develop a non-exhaustive set of regular expressions to filter out administrative categories, identifying 250k 'interesting' categories with 2+ members (31% of all categories), containing 9 million membership links.

We next apply the $rep()$ algorithm. Suppose an edit is made on article α. Because an article may reside in multiple

[5]This should not be interpreted as "most controversial." Administrators can *protect* articles to prevent them from being edited by un-registered users, limiting their vandalism.

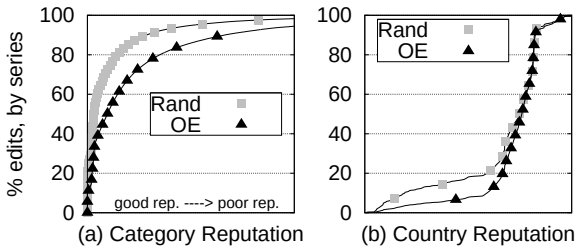

Figure 5: CDF of Category/Country Reputation

CATEGORY (w/100+ pages)	**PGs**	**OEs/PG**
World Music Award Winners	125	162.27
Characters of Les Misérables	135	146.88
Former British Colonies	145	141.51
Congressional Medal Recipients	161	121.98

Table 4: Categories w/Most OEs (normalized)

categories, function $G()$ is modified to return a higher-order set, $G(\alpha) = C_\alpha = \{c_{\alpha 1}, c_{\alpha 2}, \ldots\}$ where each $c_{\alpha i}$ is a set containing the members of a category in which α is a member. Then, for all $c_{\alpha i} \in C_\alpha$ we calculate $rep(c_{\alpha i})$. This process produces $|C_\alpha|$ reputation values. Which of these values (or combination thereof) is most behavior predictive is the subject of future work – herein we consider the *maximum* of all $c_{\alpha i}$ and assign it as the feature value. See Fig. 5a.

Discussion: Just as some individual articles are frequent targets of vandalism, so too is the case with certain categories (see Tab. 4). Imagine the candidates of an election are grouped by category (with poor reputation). Then, an unknown candidate enters the race. Though the article reputation for this candidate may have no OE history, the poor category reputation may be predictive of vandalism.

Category-reputation is 2.5 times higher, on the average, for OEs than random ones, and Fig. 5a shows a measurable gap between the two sets for a large percentage of edits. An edit is a member of 1+ categories 94% of the time, and 97% of OEs show non-zero reputations.

4.2.4 Country Reputation

Description: With geolocation data [1] one can determine the country of origin for an edit – a spatial grouping over users. Calculating $rep()$ is straightforward: An IP address resides in a single country, and the $size()$ normalizer is the number of prior edits made from that country[6]. See Fig. 5b.

Discussion: There are subtle variations in edit quality across IP space. We found that physical location, as opposed to purely IP-based groupings, was most effective at making behavioral distinctions within this space.

Per Tab. 5, certain countries[7] are responsible for a greater (normalized) percentage of OEs. For example, an Australian edit is 4× more likely to be an OE than one from Italy. As Fig. 5b shows, country-reputation differs significantly between OE/random sets for about one-third of edits.

5. CLASSIFYING EDITS

[6]Because this normalizer is orders of magnitude beyond what a single user could influence, we find its use acceptable here, unlike in the user reputation case.

[7]Country mappings are simplified per our geolocation data, other granularity (*e.g.,* city-level) may prove optimal.

RANK	**COUNTRY**	**EDITS**	**%-OEs**
1	Italy	116,659	2.85%
2	France	116,201	3.46%
13	United States	7,648,075	11.63%
14	Australia	670,483	12.08%

Table 5: Countries (w/100k+ edits) by OE-%

We construct a classifier to detect vandalism using the aforementioned features and test it over anonymous user edits (which are the bulk of the vandalism problem and identified by IP – enabling the use of geolocation data).

5.1 Classification Method

The primary learning challenge is that the labeled dataset consists of only negative (*i.e.,* vandalous) examples – the 'random' set contains both good and bad edits. This issue is largely a boot-strapping problem, but a significant one (if utilized in intelligent routing, we can rely on humans to provide definitive labels after the initial training).

For learning, we use an inductive support-vector-machine (SVM) [9], carefully tuning cost parameters to compensate for the non-homogeneous nature of the 'random' label. The revision metadata associated with each edit, combined with a historical record of OEs, and several auxiliary tables (*e.g.,* geolocation data, estimated registration times, *etc.*) is sufficient to compute the 10 features discussed herein. All features are normalized on $[0, 1]$ and arranged so that good behavior always tends towards the origin.

Testing proceeded chronologically. Training took place on a 1% subset (6.5k edits) of a month's eligible data. The resulting model was used to classify the *next* month's data.

5.2 Classifier Performance

Just as with training, the ambiguity of the 'random' label complicates our performance evaluation. In machine-learning terms, the *recall* (the percentage of well-classified OEs in the classification set) of our model is as expected, but the *precision* (the percentage of edits classified as vandalism that are actually vandalism) is skewed because 'random' edits classified as OEs are not necessarily false-positives (FPs). To estimate the true precision of our model, we manually inspect a subset of potential-FPs, and extrapolate from this to calculate the *adjusted-precision*.

When manually measuring performance in this manner, we are in fact using the system exactly as it would be used in a production setting; as an *intelligent routing* tool. In the course of our experiments, we located over 5,000 incidents of vandalism external to our OE set[8].

We now examine the classification of a single month in detail (2009/10). The month had 1.3 million anonymous-user edits, 128k of which were OEs (9.9%). A recall-rate of 50% (64k OEs) was observed, with 15% of the 'random' set being potential false-positives (173k edits, 27% raw precision). We manually inspected 1k of the potential FPs, finding that 30% (52k, extrapolated) constituted vandalism. Supposing that the 237k edits classified as vandalism were presented to a user via an intelligent routing tool, then 49% (the adjusted-precision) of such edits would be malicious.

Of course, the recall-to-precision ratio is tune-able. If one wants to find more vandalism (cumulatively), one must tolerate more false-positives. This relationship is visualized in

[8]Available at `http://www.cis.upenn.edu/~westand` is a corpus of the 5k+ manually tagged incidents of vandalism and 5.7 million OEs. Our labeling rationale is also described there.

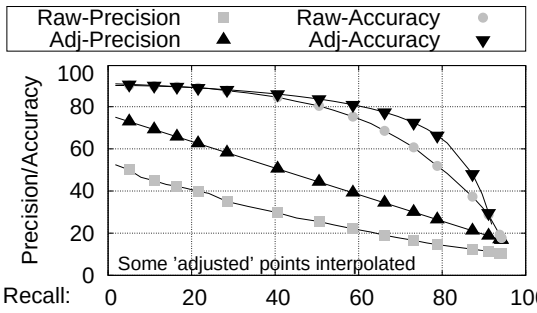

Figure 6: Classifier Precision/Recall/Accuracy

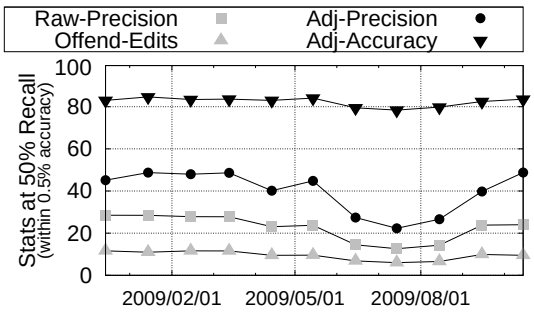

Figure 7: Monthly Classifier Performance

Fig. 6, where both recall and precision are plotted for variable cost parameters. No matter the parameterization, our classifier is capable of steady-state performance across time. Fig. 7 shows performance statistics assuming a fixed recall rate of 50%, for a one-year period pre-dating our data dump.

This performance is comparable to that of the more computationally intensive NLP classifiers. Smets *et al.* [15] attempted both Naïve-Bayes (50% precision at 50% recall) and Probabilistic Sequence Modeling (56% precision at 50% recall) classification, with remarkably similar results to our own. Potthast *et al.* [12] claims the strongest result (83% precision at 77% recall), but uses a small and biased classification set (just 900 examples, 33% of which are vandalism). Further, [12] highlights the inefficiencies of language processing – achieving a throughput of 5 edits/second, while our system is capable of scoring 100+ edits/second.

While we benchmark our performance against NLP efforts, it is not our intention to compete against them. Indeed, given the difficulty of detecting vandalism, a selective combination of spatio-temporal features and NLP ones may produce an even stronger result – provided the two methods do not detect precisely the same edits. Fortunately, this does not appear to be the case. Fig. 1 suggested that a large amount of vandalism was being caught by `abusefilter`, beginning in mid-2009. While Fig. 7 shows a decrease in performance around this period, our effectiveness is far from eliminated – suggesting malicious edits that can evade the (often language-based) rules of `abusefilter` may still exhibit spatio-temporal properties indicative of vandalism.

6. CONCLUSIONS

In this work we have demonstrated the feasibility of using spatio-temporal features to detect vandalism on Wikipedia. Our method performs comparably to natural-language classifiers, while flagging edits at a far greater throughput.

That being said, additional work could further solidify the progress made herein. First, we note that our feature set is non-exhaustive. We classified over several features whose choice was guided by intuition. There likely exists additional measures that could improve classifier performance. Second, it would be helpful to examine the contribution of each feature individually and examine the correlation between them. Third, we aim to produce an on-Wikipedia implementation of our system. Not only will the tool benefit the Wikipedia community, but a large-user base providing definitive edit labels is likely to improve our performance metrics. Finally, we must consider the robustness of our features. Generally speaking, spatio-temporal measures are harder to evade than language-based ones (*e.g.*, changing ones physical location is difficult, changing the body of an edit is not), but this needs examined in detail and against varying attacker models.

Broadly, this work has shown the applicability of spatio-temporal reputation in detecting Wikipedia vandalism. Spatial and temporal patterns are inherent in a number of domains and have already been exploited in the mitigation of email spam. These combined successes suggest that spatio-temporal reputation could be used as general-purpose method for content-based access control – a relevant issue as collaboration and cooperation mature in digital environments.

References

[1] IPinfoDB. `http://ipinfodb.com`. (Geolocation Data).

[2] Wikipedia. `http://www.wikipedia.org`.

[3] B. T. Adler, K. Chatterjee, L. de Alfaro, M. Faella, I. Pye, and V. Raman. Assigning trust to Wikipedia content. In *Proceedings of WikiSym '08*, Porto, Portugal, 2008.

[4] B. T. Adler and L. de Alfaro. A content-driven reputation system for the Wikipedia. In *Proc. of WWW '07*, 2007.

[5] L. S. Buriol, C. Castillo, D. Donato, S. Leonardi, and S. Millozzi. Temporal analysis of the Wikigraph. In *WI '06: Proc. of the Intl. Conf. on Web Intelligence*, 2006.

[6] J. R. Douceur. The Sybil attack. In *First IPTPS*, 2002.

[7] J. Giles. Internet encyclopedias go head to head. *Nature*, 438:900–901, December 2005.

[8] S. Hao, N. A. Syed, N. Feamster, A. G. Gray, and S. Krasser. Detecting spammers with SNARE: Spatio-temporal network-level automated reputation engine. In *18th USENIX Security Symposium*, 2009.

[9] T. Joachims. *Advances in Kernel Methods - Support Vector Learning*, chapter Making Large-scale SVM Learning Practical, pages 169–184. MIT Press, 1999.

[10] S. D. Kamvar, M. T. Schlosser, and H. Garcia-molina. The EigenTrust algorithm for reputation management in P2P networks. In *Proc. of WWW '03*, Budapest, May 2003.

[11] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001.

[12] M. Potthast, B. Stein, and R. Gerling. Automatic vandalism detection in Wikipedia. In *Advances in Information Retrieval*, pages 663–668, 2008.

[13] R. Priedhorsky, J. Chen, S. K. Lam, K. Panciera, L. Terveen, and J. Riedl. Creating, destroying, and restoring value in Wikipedia. In *GROUP '07*, 2007.

[14] L. Rassbach, T. Pincock, and B. Mingus. Exploring the feasability of automatically rating online article quality.

[15] K. Smets, B. Goethals, and B. Verdonk. Automatic vandalism detection in Wikipedia: Towards a machine learning approach. In *WikiAI '08*, 2008.

[16] F. B. Viégas, M. Wattenburg, J. Kriss, and F. van Ham. Talk before you type: Coordination in Wikipedia. In *HICSS '07*, pages 78–87, 2007.

[17] 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.

[18] T. Wöhner and R. Peters. Assessing the quality of Wikipedia articles with lifecycle based metrics. In *Proceedings of WikiSym '09*, Orlando, Florida, USA, 2009.