Spatio-temporal Analysis of Reverted Wikipedia Edits

Johannes Kiesel and **Martin Potthast** and **Matthias Hagen** and **Benno Stein**
Bauhaus-Universität Weimar
<first name>.<last name>@uni-weimar.de

Abstract

Little is known about what causes anti-social behavior online. The paper at hand analyzes vandalism and damage in Wikipedia with regard to the time it is conducted and the country it originates from. First, we identify vandalism and damaging edits via ex post facto evidence by mining Wikipedia's revert graph. Second, we geolocate the cohort of edits from anonymous Wikipedia editors using their associated IP addresses and edit times, showing the feasibility of reliable historic geolocation with respect to country and time zone, even under limited geolocation data. Third, we conduct the first spatio-temporal analysis of vandalism on Wikipedia.

Our analysis reveals significant differences for vandalism activities during the day, and for different days of the week, seasons, countries of origin, as well as Wikipedia's languages. For the analyzed countries, the ratio is typically highest at non-summer workday mornings, with additional peaks after break times. We hence assume that Wikipedia vandalism is linked to labor, perhaps serving as relief from stress or boredom, whereas cultural differences have a large effect. Our results open up avenues for new research on collaborative writing at scale, and advanced technologies to identify and handle antisocial behavior in online communities.

1 Introduction

Vandalism is one of Wikipedia's most prominent problems. From the start, the freedom that anyone can edit any article on Wikipedia has attracted vandals who damage articles instead of improving them. While the freedom to edit is a cornerstone of Wikipedia's success story, part of Wikipedia's community is constantly embroiled in reviewing edits to spot and revert the damage done by vandals. Without this cleanup, the quality of Wikipedia's articles would quickly deteriorate. However, Geiger and Halfaker (2013) find that only a small portion of Wikipedia's community takes charge of reviewing, so that the amount of edits per time period cannot be handled entirely manually in a timely manner. This is why tools are employed to streamline reviewing and to automate part of it, ranging from rule-based bots up to machine learning approaches capable of detecting more subtle vandalism. In fact, a shared task organized by Potthast, Stein, and Holfeld (2010) resulted in plenty of approaches.

The rigorous enforcement of Wikipedia's codes of conduct significantly raised the bar for newcomers: Halfaker, Kittur, and Riedl (2011) report that reverts—especially when done by automatic vandalism detectors—severely affect user retention. This, in turn, is considered a contributing factor to Wikipedia's ongoing decline in terms of active editors since 2007 as pointed out by Suh et al. (2009). Halfaker et al. (2013) pass the blame to overprotective editors and detectors, suggesting that new approaches are in need to make reverts more friendly, so that new users who blunder will not be alienated while still undoing the deliberate damage caused by vandals. It is striking, though, that despite the large body of work on Wikipedia edits, reverts, and vandalism in particular (reviewed in Section 2), to the best of our knowledge, no research has been carried out to uncover *why* anonymous editors damage Wikipedia until now. Our contributions in this respect are threefold:

1. *Ex post facto vandalism detection.* We conduct the first systematic analysis of Wikipedia article revert graphs to identify vandalism and damaging edits *after they have been undone* as ground truth for our analysis (Section 3).
2. *Historic editor geolocation.* We geolocate 77% of Wikipedia's anonymous editors since 2002 in terms of country and time zone by cross-checking geolocation databases with Regional Internet Registry data (Section 4).
3. *Spatio-temporal analysis.* Combining the aforementioned results, we conduct the first in-depth spatio-temporal analysis of Wikipedia's history, revealing a strong dependence of vandalism on time of day, day of week, country, culture, and Wikipedia language (Section 5).

The most salient insight of our analysis is that the relative amount of vandalism is significantly higher during the working hours of a week day (excluding summer) than otherwise, e.g., varying between one in three and one in six edits for the United States. Peaks of vandalism can be observed when people start working in the morning and after typical break times, suggesting a strong connection between labor and vandalism. Besides shedding light on human behavior in general, our results inform vandalism prevention efforts (Section 6). For reproducibility sake, we provide the software underlying our analysis open source.[1]

[1] See http://github.com/webis-de/ICWSM-17/.

2 Related Work

Wikipedia has become a frequent subject for research across computer science (surveyed by Medelyan et al. (2009) and Okoli et al. (2012)), and the social sciences (surveyed by Schroeder and Taylor (2015)).

Edits and editors. Wikipedia's collaborative writing process is of particular interest for its unique scale and since Wikipedia's community is almost entirely self-organized. Steiner (2014) develops a tool that allows for monitoring the current editing traffic on Wikipedia, separating three sources of edits: anonymous editors, registered ones, and automatic bots. Today, a little more than 15% of the edits on Wikipedia are done by bots and another 26% by anonymous users, whereas the majority of 59% of edits originates from registered users. In this regard, the question of "Who writes Wikipedia?" has been intensely debated, and it has been frequently pointed out that the majority of manual edits originates from a core group of registered "elite" editors who make up for most of the contributions (Kittur et al. 2007a; Panciera, Halfaker, and Terveen 2009; Priedhorsky et al. 2007), considering both edit quantity and edit quality (i.e., longevity of edited words). This may explain why research focuses almost exclusively on the portion of edits originating from registered users. In this connection, identifying and automatically detecting edit types (Daxenberger and Gurevych 2012; 2013), and which types of edits originate from groups of registered editors who assume certain social roles within Wikipedia's community (Arazy et al. 2015; Ferschke, Yang, and Rosè 2015; Kriplean, Beschastnikh, and McDonald 2008; Welser et al. 2011; Yang et al. 2016), has recently attracted attention. The latter include basic roles identified via access privileges, or so-called barnstars awarded to editors by the community for outstanding contributions with respect to certain criteria, as well as custom user role ontologies. Besides roles, others attempt to quantify the extent of gender bias found within edits and among editors (Antin et al. 2011; Wagner et al. 2015). Furthermore, Kuznetsov (2006) investigates the motivations of Wikipedia editors, and Halfaker, Kittur, and Riedl (2011) and Halfaker et al. (2013) study the cohort of registered newcomer editors and whether they remain active when faced with backlash from the community, anonymous users, or from automatic reviewing tools.

Yasseri, Sumi, and Kertész (2012) analyze temporal patterns of Wikipedia edits, contrasting all language versions of Wikipedia. The authors identify four groups of languages that exert different distributions of edit frequencies throughout the day. A clear night-and-day curve can be observed for most Wikipedias. For languages spoken in different time zones, the authors model the edit ratio distributions as mixtures of a standard distribution derived from averaging all Wikipedias edited mostly from a single time zone. However, the results allow for no insights into the nature of vandalism. In fact, all of the aforementioned studies mention vandalism only in passing, or not at all.

Vandalism. Although vandalism on Wikipedia has attracted considerable attention, too, surprisingly, there is hardly any work as to why or under what circumstances editors vandalize. Geiger and Ribes (2010) trace the steps that lead to the banning of an individual vandal as an example of Wikipedia's distributed self-preservation process; Shachaf and Hara (2010) focus on user who act as trolls; and Kumar, Spezzano, and Subrahmanian (2015) identify registered users engaging in vandalism using behavioral features. Otherwise, most papers on Wikipedia vandalism propose automatic vandalism detection tools: Potthast, Stein, and Gerling (2008) first developed machine learning technology for this purpose, and many of the approaches in existence today have been developed or derived from the results of two shared tasks at PAN 2010 and PAN 2011 (Potthast, Stein, and Holfeld 2010; Potthast and Holfeld 2011). One of the currently best-performing approaches is that of Adler et al. (2011), combining many of the approaches submitted to the shared tasks. From reviewing the literature on vandalism detection, none of the authors analyzed behavioral aspects, with one notable exception: West, Kannan, and Lee (2010) exploit the spatio-temporal characteristics of Wikipedia edits to train a machine learning model for vandalism detection. Similar to our methodology, they identify vandalism in Wikipedia article histories, and they employ a geolocation database to map the IP addresses of anonymous editors to their place of origin. Unlike in our paper, though, their analysis encompasses only a small fraction of reverted edits from Wikipedia's history, since they rely only on edits reverted with the administrative rollback tool. Furthermore, their geolocation does not take into account that old IP addresses may not be reliably geolocated with newer geolocation databases. Furthermore, they stop short of analyzing the spatio-temporal patterns, whereas the small scale and the noisy geolocation might have thwarted such analyses. Still, the features proposed help a machine learning algorithm to pick up vandalism.

3 Mining Vandalism

This section defines vandalism edits in Wikipedia and details our approach at identifying such edits as a ground truth for our analysis: we rely on ex post facto evidence, namely whether an edit has been reverted manually or automatically. We loosely follow the self-reflection steps outlined by Howison, Wiggins, and Crowston (2011) to ensure validity.

Operationalizing "Vandalism"

Wikipedia defines vandalism somewhat vaguely as "any malicious edit which attempts to reverse the main goal of the project of Wikipedia."[2] For clarification, the definition provides examples such as the removal of encyclopedic content, change beyond recognition, or change without regard to policies (neutral point of view, verifiability, and no original research), but also juvenile acts like adding obscenities, crude humor, nonsense, or removing an article's entire content ("blanking"). Excluded from being vandalism, "even if misguided, willfully against consensus, or disruptive, [is] any good-faith effort to improve the encyclopedia," like edit wars, where two parties fight over which version of an article is better by repeatedly reverting the opposing party's version. Perhaps the most decisive property an edit must fulfill to be vandalism is that of being done with malicious intent—which

[2]https://en.wikipedia.org/wiki/Wikipedia:Vandalism

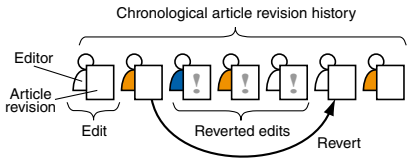

Figure 1: Illustration of a Wikipedia article revision history. Each revision is the result of an editor's changes to its preceding revision, yielding a chronological sequence of revisions by successive editors. Shade indicates different editors, arc arrows indicate reverts, where an old article revision is reinserted as new revision, undoing all intermediate revisions.

is also the most elusive one. For example, subtle changes to an article's point of view according to one's own agenda may seem perfectly legitimate editing to a non-expert. This renders operationalizing vandalism difficult, as one cannot even entirely trust human judgment.

Its vague definition notwithstanding, vandalism is a real problem and remedies are sorely needed. The literature has hence adopted the term for lack of a better one to describe the efforts at identifying edits that come close to the above definition, thereby aligning terminology with that of the Wikipedia community. Vandalism has basically been operationalized in three ways (listed in ascending order of scalability and descending order of accuracy): (1) Based on external, independent review for up to thousands of edits (Potthast 2010). (2) Based on internal, dependent review by analyzing explicit comments left by community members undoing vandalism (e.g., Kittur et al. (2007b); Tran and Christen (2013)). However, comments are often missing or may be left as false accusations (e.g., in edit wars). (3) Based on article states by considering all full page reverts (e.g., Rzeszotarski and Kittur (2012)). Taken alone, this approach has a strongly oversimplified view on vandalism. In this paper, we operationalize vandalism based on Approaches 2 and 3 to allow for full scale analyses of Wikipedia's history. Nevertheless, we go beyond these approaches by analyzing the revert graphs of Wikipedia article histories in order to filter revert patterns that suggest good intentions on the part of an editor.

Identifying Past Vandalism as Ground Truth

At Wikipedia, manual or automatic undoing of vandalism edits basically happens by reinstating the latest non-vandalized revision of an article, which is directly supported from Wikipedia's user interface. Edits that are undone this way are called *reverted*, whereas the undoing edit is called *revert* (see Figure 1). Reverted edits are not deleted; instead, a copy of the revision preceding the reverted edits is appended to the article's revision history—a so-called *identity revert* or *full page revert*. We focus on full page reverts, as estimates suggest that *partial reverts* (i.e., edits that restore only some parts) cover only few cases of vandalism (Kittur et al. 2007b; Flöck, Vrandečić, and Simperl 2012).

As raw data for our analysis, we use the full page reverts from all Wikipedia article histories comprised in the May 2016 Wikipedia history dumps. The English Wikipedia history dump, 47 gigabyte compressed XML, contains

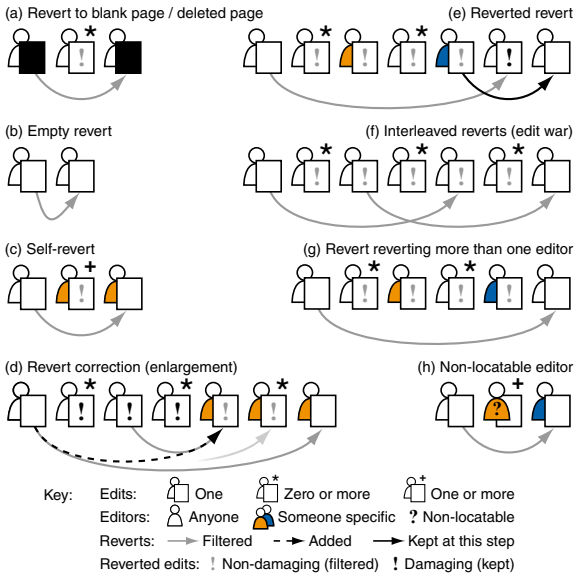

Figure 2: Revert patterns used for filtering full page reverts stepwise: first pseudo-reverts (a,b) are filtered, then error-corrections (c,d,e), ambiguous reverts (f,g), and finally reverts reverting edits of non-locatable editors (h). Each pattern depicts a regular expression that is matched against an article's revision history, filtering or reinterpreting reverts accordingly.

663,079,526 edits on 39,306,588 pages. We consider only the 471,070,114 edits on the 12,488,908 articles, disregarding user pages, discussion pages, etc. Due to article deletions, revision histories for many more than the 5.3 million articles that are currently online at Wikipedia are available. As a first step, we identify all full page reverts by matching the SHA-1 hashes of article wikitexts (Kittur et al. 2007b): when a given SHA-1 value appears twice or more in a given article's revision history, every reappearance after the first one constitutes a revert, and all edits between two appearances are reverted. The first row of Table 1 shows both the total number of reverts identified (44.9 million), and the resulting total number of reverted edits (119.7 million).

While reverted edits identified via SHA-1 matching have been used to train vandalism classifiers (West, Kannan, and Lee 2010; Tran and Christen 2015), only a fraction of reverts (14.8%) *explicitly* undo vandalism, judging by the comments left by editors (when using Kittur et al.'s (2007b) approach to identify such comments). As it is not mandatory to indicate that vandalism is reverted, relying on comments alone severely underestimates the amount of vandalism on Wikipedia. However, presuming that all full page reverts are vandalism is a gross overestimation: in a manual analysis of the revert graphs of 100 randomly selected articles we found many reoccurring revert-patterns that seem natural for collaborative editing situations and that are completely harmless. In what follows, we define the harmless revert patterns we identified, and detail how filtering them from the set of full page reverts affects the amount of reverted edits that can be called vandalism with a high confidence.

Table 1: Step-by-step filtering of the English Wikipedia as per the revert patterns depicted in Figure 2. Counts of full page reverts and counts of reverted edits affected by corresponding full page reverts are given. Full page reverts are analyzed for indications of vandalism in edit comments as per Kittur et al. (2007b), and reverted edits are divided into edits originating from editors who are anonymous, registered, or bots. Reverted edits remaining after filtering are mostly vandalism, and almost unanimously damaging.

Revert filtering step		**Full page reverts**				**Reverted edits**				
		Vandalism as per Kittur		Total		Editor			Total	
		No	Yes	Absolute	Relative	Anonymous	Registered	Bot	Absolute	Relative
Results of naive SHA-1 matching		38,244,710	6,670,575	44,915,285	100.0%	66,375,400	50,314,563	3,051,882	119,741,845	100.0%
(a) reverts to page blank		-462,242	-4,085	-466,327	-1.0%	-19,176,154	-24,090,455	-1,629,953	-44,896,562	-37.5%
(b) empty reverts due to renaming/removal/error		-2,085,189	-99,135	-2,184,324	-4.9%	0	0	0	0	0.0%
Results after filtering pseudo-reverts	Σ	35,697,279	6,567,355	42,264,634	94.1%	47,199,246	26,224,108	1,421,929	74,845,283	62.5%
(c) self reverts		-3,865,372	-8,991	-3,874,363	-8.6%	-2,613,154	-2,018,218	-158,218	-4,789,590	-4.0%
(d) revert corrections		-387,301	-62,437	-449,738	-1.0%	-862,439	-1,074,102	-27,456	-1,963,997	-1.6%
(e) reverted reverts		-313,539	-13,133	-326,672	-0.7%	-2,719,594	-4,088,871	-179,472	-6,987,937	-5.8%
Results after filtering error-corrections	Σ	31,131,067	6,482,794	37,613,861	83.7%	41,004,059	19,042,917	1,056,783	61,103,759	51.0%
(f) interleaved reverts		-4,573,240	-401,277	-4,974,517	-11.1%	-3,606,705	-5,188,456	-295,504	-9,090,665	-7.6%
(g) reverts reverting more than one editor		-1,776,317	-339,784	-2,116,101	-4.7%	-7,060,825	-4,860,432	-449,690	-12,370,947	-10.3%
Results after filtering ambiguous reverts	Σ	24,781,510	5,741,733	30,523,243	68.0%	30,336,529	8,994,029	311,589	39,642,147	33.1%
(h1) reverts reverting registered editors or bots		-6,116,841	-799,928	-6,916,769	-15.4%	0	-8,994,029	-311,589	-9,305,618	-7.8%
(h2) reverts reverting editors with IPv6 addresses		-213,963	-51,808	-265,771	-0.6%	-338,137	0	0	-338,137	-0.3%
Results after all filtering steps	Σ	18,450,706	4,889,997	**23,340,703**	**52.0%**	29,998,392	0	0	**29,998,392**	**25.1%**

Figure 2 lists the revert patterns used to filter reverts that are not suited to our analysis, as it is unlikely or unclear whether they indeed revert vandalism. The filter patterns (a) to (h) are organized in the order in which they are applied. Their order is important, since reverts may cross or include one another, resulting in intricate graphs that need to be carefully disentangled. Rows (a) to (h) in Table 1 show the amount of full page reverts that are filtered by the corresponding pattern, and the amount of reverted edits that are hence disregarded. Related groups of patterns are detailed below.

Filtering pseudo-reverts. Patterns (a) and (b) in Figure 2 depict pseudo-reverts, namely reverts that are not intended as such. Pattern (a) concerns removing all content from or deleting an entire article. This happens occasionally in an article's history—sometimes as an act of vandalism—, so that pseudo-reverts may revert edits by many different editors (in Figure 2 denoted by a star as in a regular expression, where a white editor may be any editor). Pattern (b) captures reverts that do not change the article, which may occur when an article is renamed or due to MediaWiki errors. We filter these reverts first to not confuse other patterns; but note that reverts that undo page blanking remain untouched. As a result, about 2.7 million reverts are filtered, covering about 44.9 million unintentionally reverted edits (see Table 1, rows (a,b)).

Filtering error-corrections. Patterns (c), (d), and (e) depict reverts which revert edits that are likely not vandalism. Pattern (c) concerns self-reverts where an editor fixes a mistake by undoing it again. Pattern (d) concerns series of reverts by the same editor, where the latest revert covers the previous reverts, implicating that the editor just corrected the revert. In such cases, we replace the series of reverts with a newly created one that corresponds to the editor's actual intentions. In 9% of these cases, the editor enlarged her revert more than once in a row. Pattern (e) concerns longer reverts that are immediately reverted again, implying the editor of the first revert tried to damage an article by resetting it to a previous revision. We disregard the original revert and count only the one reverting the vandalism revert. Another 4.7 million reverts are filtered, covering about 13.7 million reverted edits (see Table 1, rows (c,d,e)).

Filtering ambiguous reverts. Patterns (f) and (g) depict ambiguous reverts where it remains unclear which of the reverted edits originate from editors with damaging intent. Pattern (f) captures interleaved reverts as they usually appear in edit wars, since, as per Wikipedia's definition of vandalism, edit wars do not necessarily happen among ill-intentioned editors. Interestingly, the majority of the about 1.5 million edit wars in the English Wikipedia have been rather short-lived, with 43% spanning only 2, and 37% spanning only 3 reverts. Pattern (g) concerns reverts that affect edits of different editors, suggesting a series of different vandalism cases. Here we decided to err on the safe side and ignore reverts with such a strong claim. About 7.1 million reverts and about 21.5 million reverted edits are filtered (see Table 1, rows (f,g)).

Filtering non-locatable editors (h). Finally, in anticipation of the next step of geolocating the editors of vandalism edits, Pattern (h) filters reverts that revert edits whose editors cannot be geolocated. These are registered editors, who are responsible for less than 12% of reverted vandalism as per Kittur et al. (2007b), bots, and, due to lacking data in the geolocation databases, the very few editors using IPv6 addresses. About 7.2 million reverts and about 9.6 million reverted edits are filtered (see Table 1, rows (h1,h2)).

Altogether, 52% of all reverts, and about 75% of all reverted edits are filtered as harmless, ambiguous, or non-locatable. What remains as ground truth for our subsequent analysis are 29,998,392 edits, which represents vandalism originating from anonymous editors (see Table 1, last row).

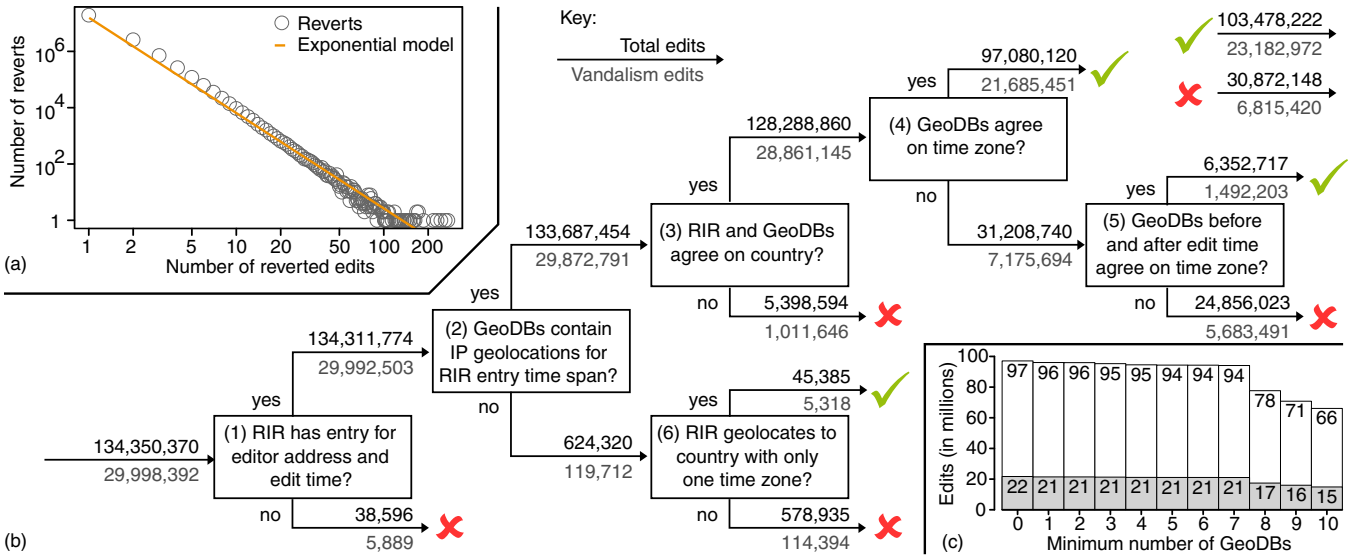

Figure 3: (a) Number of full page reverts with a specific number of edits they revert and fitted exponential model in a log-log plot. (b) Decision tree to decide whether to trust the available geolocation information for an edit (✓), or not (✗). The numbers denote the total edits and reverted edits for the English Wikipedia that went through each branch. (c) Number of total edits (white bars) and vandalism edits (gray bars) in millions from the yes-branch of Step (4) in (b) over the number of GeoDBs considered.

Assessing the Vandalism Ground Truth

To assess our success at identifying vandalism, we perform a sanity-check on the filtered reverts and look at the effect filtering has on recall and precision. Similar to other collaboration scenarios, one would expect that most reverts affect very few edits, and few reverts affect many edits. To show that this is indeed the case, Figure 3a plots the number of reverts over the number of reverted edits on a double-logarithmic scale. For recall, we use the number of reverts whose comments indicate the removal of vandalism as per Kittur et al. (2007b) (see Table 1). Among all 44.9 million reverts, a total of 6,670,575 (14.9%) are vandalism reverts, which corroborates the results of Kittur et al. Applying Patterns (a)-(e) filters few reverts from this subset, where most are actually empty reverts. A large portion of reverts is filtered via Patterns (f) and (g) as ambiguous reverts: edit wars are not necessarily vandalism, and reverts reverting edits from multiple different users may include innocent editors. In these cases, we sacrifice some recall in favor of precision. Finally, 799,928 explicit vandalism reverts are filtered because they originate from registered users or bots, since they are not our focus of attention. In sum, our filtering recalls 73.3% of all explicit vandalism reverts. When disregarding those from registered users as well as pseudo-reverts, our approach recalls 84.7% of the remainder. As for precision, we manually review random samples of reverted edits by anonymous editors, 1,000 drawn before filtering and 1,000 after. In these cases, 68.7% of the reverted edits before filtering are indeed vandalism, whereas after filtering precision rises to a solid 82.8%.

These results show that identifying vandalism based on the ex post facto evidence is feasible with high precision and recall. Unlike using actual vandalism detectors, our approach incorporates the revert decisions of human editors and vandalism detection bots alike, including cases where ones err and are reverted themselves. Moreover, our approach is language-independent and uses understandable rules. Of course, it cannot be used to automatically undo vandalism: rather, the reverted edits remaining after filtering form a ground truth of past vandalism on Wikipedia as per consensus of man and machine.[3] These reverted edits, and their geolocation (Section 4), serve as ground truth for our spatio-temporal analysis of anonymous vandalism on Wikipedia (Section 5).

4 Geolocating Editors

For anonymous edits, the Wikipedia history dumps supply their server times and their editors' IP addresses at the time of editing. The server time, however, forecloses spatio-temporal analyses, since editors are typically far removed from the server. We hence resort to geolocation databases (GeoDBs) to augment the dumps.[4] But since IP addresses may change location over time, special care must be taken when dealing with historic IPs to ensure reliable geolocation. We show for the first time that even decade-old IP addresses can be reliably geolocated in terms of country and time zone by combining GeoDBs with Regional Internet Registry (RIR) data. Since 2003, RIRs supply daily updates of IP allocations to organizations, including their country. By combining eleven commercial GeoDBs from IPligence and IP2Location with RIR data, cross-checking geolocations for consistency and removing all inconsistent ones, we obtain reliable geolocations

[3] As a spin-off, we are working toward releasing the dataset as a corpus for training vandalism detectors.

[4] Regarding GeoDBs, previous research suggests that country information is reliable (accuracy above 95%) and that latitude/longitude coordinates typically have a tolerance of far below 1,000km (Poese et al. 2011; Shavitt and Zilberman 2011).

for 77% of all anonymous edits on Wikipedia.[5] The source code of our geolocation is freely available.[6]

To geolocate the editors' IP addresses, we identified relevant inconsistencies and devised a set of rules to deal with them. The resulting flow diagram of decisions is depicted in Figure 3b: (1) Removal of the few IP addresses that have no corresponding RIR entry.[7] (2) Check whether the IP addresses are contained in one or more GeoDBs that fall within a "RIR span" (characterized by the time span between the RIR entry directly before and the RIR entry directly after the Wikipedia time of an edit). (3) If yes, removal of IP addresses where RIR span and GeoDBs disagree on their country. (4) If they agree on the country, check whether they agree on their time zone as well. For 94 million of the 97 million edits that pass this check (97%), at least 7 GeoDBs agree on the time zone (Figure 3c), making these geolocations very reliable. In case of time zone disagreements within the GeoDBs, (5) check whether the GeoDBs within the RIR span directly before and directly after an edit agree on time zones, and removal of all IP addresses where this is not the case. If yes, this corresponds to providers relocating an IP block within a multi-time-zone country, which is not recorded by RIRs. Going back to Step (2), when there is no GeoDB in the RIR span around an edit's time, (6) check whether RIR geolocates to countries that have only one time zone, and removal of IP addresses where this is not the case. This way, 103,478,222 of the 134,350,370 anonymous edits (77%) from the English Wikipedia can be reliably geolocated.

In Steps (4) and (5), we determine the time zone based on the coordinates given by the GeoDBs using a time zone world map[8] and cross-check it with the country stored in the GeoDBs. The GeoDB country and the time zone world map also sometimes disagree. Most of the time, the GeoDB country aligns with the GeoDB city, while coordinates may be off (compared to city coordinates in Wikipedia). To compensate for these inaccuracies, we take the nearest time zone to the coordinates within the GeoDB country as long as it is within 7.5° of the given coordinates (i.e., half the distance between meridians). A few cases with errors due to incorrect country codes (e.g., AS for Australia or RS for Serbia and Yugoslavia) or longitudes (e.g., Dar es Salaam being 39° East, not West) have been detected this way and manually fixed.

Table 2 shows the numbers of edits removed/kept as a result of filtering IP addresses with unreliable geolocation, and the numbers of unique IPs whence they originated. The latter decreases as expected, however, the ratio of reverted edits remains identical (22%), indicating that the geolocated edits form an unbiased sample. In sum, 23,182,972 reverted edits are subject to our subsequent analysis.

[5] IPligence Max from 2008 (Oct., Nov., Dec.), 2014 (Apr., Jul., Aug., Oct., Nov.), and 2015 (Feb., Apr.), http://www.ipligence.com; DB11 lite from 2016 (Jun.), https://www.ip2location.com; RIR data available at http://ftp.*RIR*.net/pub/stats where *RIR* is one of {afrinic, apnic, arin, lacnic, ripe}.

[6] https://github.com/webis-de/aitools4-aq-geolocation

[7] A couple of old assignments are missing for historical reasons: https://www.apnic.net/about-APNIC/corporate-documents/documents/resource-guidelines/rir-statistics-exchange-format

[8] Version 2016d of http://efele.net/maps/tz/world

Table 2: Historic geolocation success for all anonymous editors of the English Wikipedia in terms of edits and unique IP addresses whence they originated. Aside the totals, the subset of edits considered vandalism or damaging as per Section 3 are given, and their corresponding IP addresses. Numbers are given for each exit node of the decision tree in Figure 3b, divided by whether or not the geolocation is trustworthy.

Decision Tree		**Edits**			**Unique IP addresses**	
Trusted	Exit Step	Vandalism as per Sec. 3		Total	Vandal IPs	Total
Entire Wikipedia:		29,998,392	(22%)	134,350,370	11,990,674	34,993,205
No (✗)	Step (1)	5,889	(15%)	38,596	2,584	8,047
	Step (3)	1,011,646	(18%)	5,398,594	387,376	1,302,473
	Step (5)	5,683,491	(22%)	24,856,023	2,379,726	6,601,222
	Step (6)	114,394	(19%)	578,935	49,518	135,053
	Σ	6,815,420	(22%)	30,872,148	2,819,094	8,045,883
Yes (✓)	Step (4)	21,685,451	(22%)	97,080,120	8,586,646	25,453,545
	Step (5)	1,492,203	(23%)	6,352,717	635,490	1,712,340
	Step (6)	5,318	(11%)	45,385	2,558	12,572
	Σ	**23,182,972**	(22%)	**103,478,222**	**9,224,625**	**27,178,053**

5 Spatio-Temporal Analysis

To analyze spatio-temporal patterns, we calculate the ratio of vandalism edits (i.e., reverted edits as per Section 3) among Wikipedia edits per hour of the day and per location. The analysis is restricted to the anonymous edits that can be reliably geolocated, which includes most of the anonymous edits (Section 4). Since we observed no correlation between being geolocated and being vandalism, we expect the restriction of reliable geolocation to not affect the results presented below. However, while our revert filter for vandalism detection is designed to avoid mislabeling a proper edit as vandalism, some cases of vandalism may have been missed. Still, almost all cases in which editors indicate by a comment that they are cleaning up vandalism take the form of full page reverts (also found by Kittur et al. (2007b)), so that it is unlikely that the vandalism ratio in anonymous edits or its spatio-temporal distribution are substantially different to what we observe. We estimate the vandalism ratio per hour of day (starting and ending at the full hour) by averaging over all days since January 1, 2006. Before 2006, in the early stages of Wikipedia, vandalism ratios are unstable and hence unreliable. About 4.3% of edits are discarded this way, but yielding an overall increase in effect sizes.

Our findings are based on visual inspection, backed by careful statistical analysis. We use Cohen's d to analyze the variances of the average vandalism ratios. While visibly different graphs usually correspond to significant differences due to sufficiently large sample sizes (millions of edits), high variances are a sign that the vandalism ratio is influenced by other factors. To give an impression of when the vandalism ratio estimates are based on few edits, these estimates are toned down in the figures. Finally, we show the significance for all effects we analyze with Cohen's d using the Welch Two Sample t-test, with one to three asterisks (*) indicating p-values less or equal to 0.05, 0.01, and 0.001.

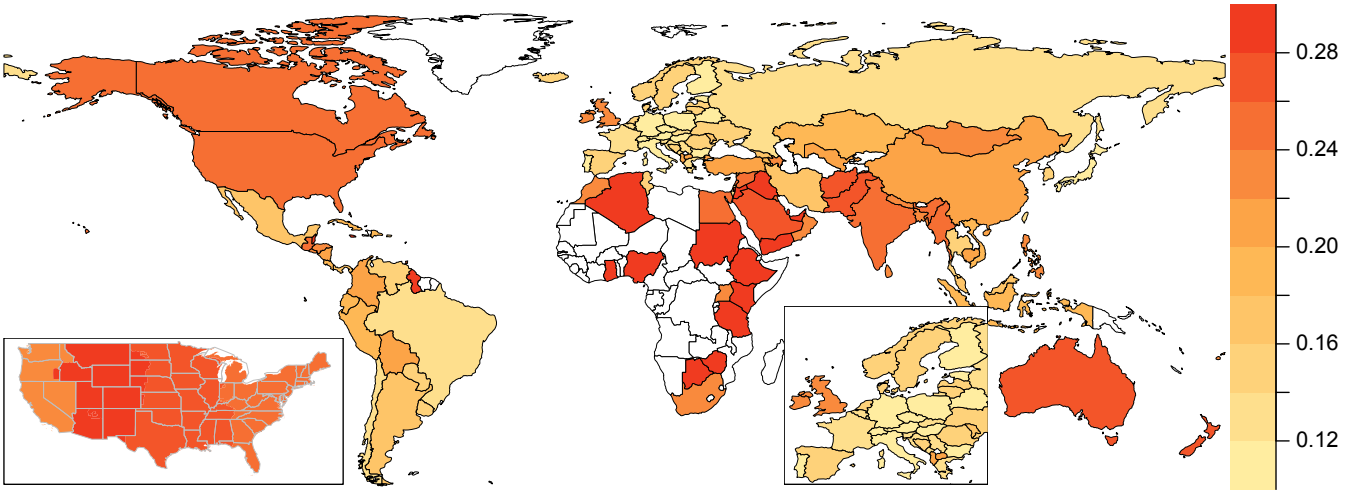

Figure 4: Ratio of vandalism to all edits in the English Wikipedia by country. Countries with less than 1,000 vandalism edits are not colored. The embedded small maps show (left) the vandalism ratio in the United States (without Alaska) by major time zone (from West to East: Pacific, Mountain, Central, and Eastern) with overlaid state borders and (right) Europe enlarged.

Figure 4 shows the vandalism ratio per country.[9] The highest vandalism ratios are observed in Africa, possibly caused by difficulties with the English language, causing native English editors to consider edits from Africa vandalism more often. However, only 0.9% of the geolocated edits to the English Wikipedia come from Africa, so that we decided to leave an analysis of the reasons for future work. Both in Europe and in South America, the highest vandalism ratios are in the countries with English as the official language: Great Britain, Ireland, and Guyana, which suggests a correlation of main language and vandalism. A similar observation is made below, when we compare edits from a specific country to different language versions of Wikipedia.

Vandalism Ratios in the United States

Figure 5a reveals different vandalism ratio for edits to the English Wikipedia from the United States, as well as the absolute number of edits and vandalism edits. While most edits are made between 14 and 17 hours, the ratio of vandalism to all edits peaks much earlier at around 9 hours with two more peaks occurring at 13 hours and at 19 hours. The lowest vandalism in both absolute and relative numbers occurs between 23 and 8 hours), which we refer to as night time for the purpose of this analysis. During the night, about one in six edits is vandalism, which changes dramatically to about one in three edits at peak times. This visually obvious difference in the vandalism ratio between night and day is reflected in the statistical analysis: the Cohen's d between the vandalism ratio averages for night and day shows a very strong statistical effect ($d = 14.7$***). For reference, Figure 5a also plots the graph when only considering edits that are explicitly labeled as vandalism reverts by a corresponding editor comment as per Kittur et al. (2007b). The two graphs resemble each other, further justifying our ex post facto vandalism detection.

[9]The map uses GADM 2.8 country/state data, http://www.gadm.org, and Efele 2016d timezone data, http://efele.net/maps/tz/.

The plots suggest that vandalism is connected to labor (working hours), with peaks of vandalism occurring when people start to work/study in the morning (8 to 9 hours) and after lunch (13 to 14 hours), e.g., as a way of "fighting" stress or boredom. Running with the hypothesis of labor-related vandalism, the increase in the ratio of vandalism between 15 and about 20 hours may also be explained by people working long hours or by relieving stress after work. Alternatively, this increase may be due to an increased negativity over the course of the day, as Golder and Macy (2011) found in their analysis of Twitter data. Further evidence for the labor-related vandalism hypothesis is provided by Figure 5b, which shows a clear difference in vandalism ratios between workdays and weekends: On workdays, the vandalism ratio is much higher than on Saturday and Sunday. On Fridays, however, the vandalism ratio graph is very similar to workdays up until about 16 hours, at which time it starts to resemble the graph of a weekend day (possibly an expression of "Thank God it's Friday"). A statistical analysis shows a very strong effect between Monday to Friday and Saturday plus Sunday for 8 to 15 hours ($d = 1.49$***), and a strong effect between Monday to Thursday and Friday to Sunday for 15 to 22 hours ($d = 0.88$***). The increase of vandalism ratio on weekends has a medium effect, comparing the hour intervals ($d = 0.53$*** for Saturday and $d = 0.68$*** for Sunday). This small increase might again be related to the increase in negativity found by Golder and Macy (2011). As shown in Figure 5c, vandalism also reduces during summer. This could be due to people going on vacation or being generally more relaxed. However, the effect size between summer and the other months for the time between 8 and 22 hours is only small ($d = 0.34$***), which is due to a large variance in the vandalism ratios from fall to spring. Although we have formed a number of hypotheses that may explain the variance, a thorough investigation requires correlating vandalism with other variables of interest and is hence left to future work.

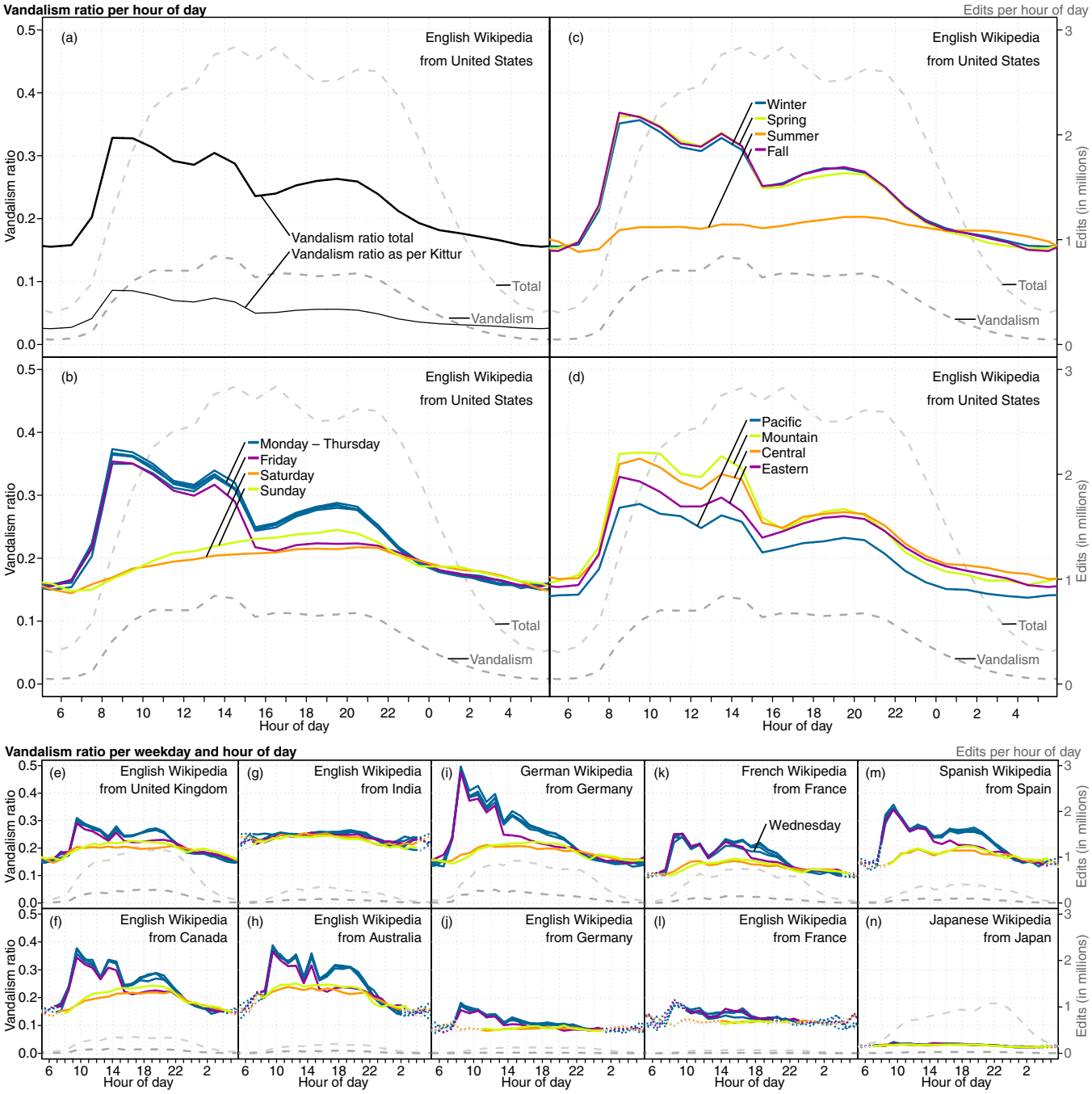

Figure 5: All plots show the *ratio of vandalism to all edits* per hour of day (left axis, solid lines), and for reference, the absolute number of edits and vandalism edits per hour of day (right axis, dashed lines), both averaged over Wikipedia's history. Plot (a) shows the overall ratio of vandalism edits on the English Wikipedia originating from the United States. Plots (b,c,d) divide the overall ratio by weekday, season, and US time zone. Plots (e-n) show vandalism ratios divided by weekday for the English Wikipedia edited from various countries, and for the German, French, Spanish, and Japanese Wikipedias when edited from Germany, France, Spain, and Japan. Ratios estimated from less than 1,000 vandalism edits are displayed with dotted lines.

We also investigated regional influences, partitioning the United States from west to east and analyzing vandalism ratio differences for each of the four parts corresponding to the well-known time zones Pacific, Mountain, Central, and Eastern.[10] The different vandalism ratios for these time zones are visible in Figure 4 (bottom left), and on a per hour basis in Figure 5d. The graphs look very similar, the difference being only the overall vandalism ratio. While the differences between the four time zones are all significant, they seem minor compared to other influences; even between 8 and 15 hours the effect sizes are small ($d < 0.30$).

Vandalism Ratios across Countries

By repeating the above analyses for different countries and Wikipedia languages, we find large differences but also commonalities. For the sake of brevity, we only report the most interesting results when considering the weekday as an additional variable.[11] Figures 5e-h show the vandalism ratio for the countries with the second to fifth-most edits to the English Wikipedia. Similar to the US, a difference between workdays and the weekend is clearly visible for the United Kingdom, Canada, and Australia, with the corresponding effect sizes d being 0.98***, 1.15***, and 0.87*** for 8 to 15 hours, and 0.81***, 0.61***, and 0.74*** for 15 to 22 hours. The effect is much smaller for India (0.12** for 8 to 15 hours and 0.32*** for 15 to 22 hours) with the possible cause that Indian citizens put much less emphasis on labor vs. leisure than people in Western countries. Also, many Western countries outsource to India, leading many to adapt their working habits accordingly, possibly further smoothing the graph.

Since Figure 4 suggests that the vandalism ratio to the English Wikipedia is higher in countries with English as the official language, we also analyze whether countries with edits in more than one Wikipedia variant have different vandalism ratios. Figures 5i-l compare the edits from Germany or France to the English Wikipedia with those to the respective "home" Wikipedias: The vandalism ratios are indeed higher in the "home" Wikipedias, especially for Germany where the English vandalism ratio is below 0.2 instead of reaching a striking 0.5 at 8 hours in the German Wikipedia (the highest ratio we observe in our analysis). A possible explanation could be that people with different background and different susceptibility to vandalism edit the different variants (i.e., the English Wikipedia may attract more educated people in non-English countries). However, despite these differences in magnitude, the graphs in Figure 5i,j as well as in Figure 5k,l still bear resemblance as to where peaks and valleys lie. Also, the relatively low vandalism ratio for Wednesday afternoons for edits from France—likely caused by the school-free afternoon at that day—is visible in the France-plots of both the French and the English Wikipedia ($d = 0.28$*** and $d = 0.17$***). We therefore see it as more likely that people just tend to vandalize the Wikipedia variant of their mother tongue more frequently as it is an easier target (e.g., usually ranked higher by search engines) and altogether follow a similar rhythm of life with vandalism ratios peaking when starting/continuing work/studies. Finally, Figures 5m,n show the vandalism ratio for two more Wikipedias among the top 7 with the most edits: Spanish and Japanese. While the Spanish plot follows a pattern similar to the one in the US, the vandalism ratio in Japan is really low (3% on average), with the only statistical effect being a higher vandalism rate during the day than during night (not visible in the plot but still a medium effect of $d = 0.54$*** due to the low variance). Thus, while our analysis shows that time has statistically strong effects on the vandalism ratio, the example of Japan shows that cultural differences can have an even stronger effect.

6 Conclusion

Our study of reverted anonymous edits on Wikipedia reveals strong spatio-temporal effects that are apparently related to labor. At typical work/study starting/continuing hours, a larger portion of edits than otherwise are vandalism—with a remarkable peak of about 50% around 8 hours from Germans to the German Wikipedia. During weekends and vacation, the ratio of reverted edits is substantially lower than on working days. This suggests that vandalism helps people to relief from stress when starting to work, to fight boredom, or to show off in front colleagues/fellow students. In conclusion, a better understanding of vandalism and when it happens is a first step towards gaining a better grasp of the problem's underlying causes and to answer the question "Why are people vandalizing Wikipedia?" While the term "vandalism" usually implies destruction without reason, a significant portion of vandalism on Wikipedia may not happen without reason, and the vandals may therefore be open to some form or another of nudging into the right direction. Our observations can initiate the development of smart technologies that monitor, detect, predict, and prevent potential threats to online communities or social software due to periodically increased susceptibilities to destructive behavior. For example, raising awareness at the right time may help users redeem themselves before acting out.

Future work should deepen the analyses and correlate vandalism to other variables of interest that may influence people's behavior, such as the weather, global events, age (e.g., adults vs. pupils), urban vs. country life, or even political orientation. Closer to hand, our ex post facto evidence-based vandalism detection provides for a reliable way of generating training data with little noise that may be used to train vandalism detectors at scale. But future work should also broaden the scope to other social softwares. Our geolocation approach can be immediately applied in other scenarios where IP addresses are recorded and become part of the public record. For example, on discussion forums, behavioral differences may be observed dependent on when and from where someone participates in a thread. Altogether, managing anti-social behavior online has become an increasingly important task for social media platforms, and handling it well depends on a thorough understanding of its causes. For Wikipedia, the causes for vandalism are mostly unknown, whereas our work hints at some of them.

[10] While the geolocation of Section 4 uses the fine-grained IANA time zones, huge size differences in terms of area covered render them less suited here (e.g., Indiana alone has 8 time zones).

[11] The full tables and plots for the other countries and Wikipedias are available at http://github.com/webis-de/ICWSM-17/.

Acknowledgments

We would like to thank our students Yamen Ajjour, Roxanne El Baff, Thomas Kessler, Kevin Lang, and Tristan Licht for their support. Our special thanks go to the anonymous reviewers for their insightful remarks.

References

Adler, B.; de Alfaro, L.; Mola-Velasco, S.; Rosso, P.; and West., A. 2011. Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. In *Proc. of CICLing*, 277–288.

Antin, J.; Yee, R.; Cheshire, C.; and Nov, O. 2011. Gender differences in Wikipedia editing. In *Proc. of WikiSym*, 11–14.

Arazy, O.; Ortega, F.; Nov, O.; Yeo, M. L.; and Balila, A. 2015. Functional roles and career paths in Wikipedia. In *Proc. of CSCW*, 1092–1105.

Daxenberger, J., and Gurevych, I. 2012. A corpus-based study of edit categories in featured and non-featured Wikipedia articles. In *Proc. of COLING*, 711–726.

Daxenberger, J., and Gurevych, I. 2013. Automatically classifying edit categories in Wikipedia revisions. In *Proc. of EMNLP*, 578–589.

Ferschke, O.; Yang, D.; and Rosè, C. 2015. A lightly supervised approach to role identification in Wikipedia talk page discussions. In *Proc. of ICWSM*.

Flöck, F.; Vrandečić, D.; and Simperl, E. 2012. Revisiting reverts: Accurate revert detection in Wikipedia. In *Proc. of HT*, 3–12.

Geiger, R. S., and Halfaker, A. 2013. When the levee breaks: Without bots, what happens to Wikipedia's quality control processes? In *Proc. of WikiSym*, 6:1–6:6.

Geiger, R. S., and Ribes, D. 2010. The work of sustaining order in Wikipedia: The banning of a vandal. In *Proc. of CSCW*, 117–126.

Golder, S. A., and Macy, M. W. 2011. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science* 333(6051):1878–1881.

Halfaker, A.; Geiger, R. S.; Morgan, J.; and Riedl, J. 2013. The rise and decline of an open collaboration system: How Wikipedia's reaction to sudden popularity is causing its decline. *Americ. Behav. Scient.* 57(5):664–688.

Halfaker, A.; Kittur, A.; and Riedl, J. 2011. Don't bite the newbies: how reverts affect the quantity and quality of Wikipedia work. In *Proc. of WikiSym*, 163–172.

Howison, J.; Wiggins, A.; and Crowston, K. 2011. Validity issues in the use of social network analysis with digital trace data. *J. of the Assoc. for Inf. Sys.* 12(12):767.

Kittur, A.; Chi, E.; Pendleton, B. A.; Suh, B.; and Mytkowicz, T. 2007a. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *WWW Jour.* 1(2):19.

Kittur, A.; Suh, B.; Pendleton, B.; and Chi, E. 2007b. He says, she says: Conflict and coordination in Wikipedia. In *Proc. of CHI*, 453–462.

Kriplean, T.; Beschastnikh, I.; and McDonald, D. 2008. Articulations of wikiwork: uncovering valued work in Wikipedia through barnstars. In *Proc. of CSCW*, 47–56.

Kumar, S.; Spezzano, F.; and Subrahmanian, V. 2015. VEWS: A Wikipedia vandal early warning system. In *Proc. of KDD*.

Kuznetsov, S. 2006. Motivations of contributors to Wikipedia. *SIGCAS Comp. Soc.* 36(2).

Medelyan, O.; Milne, D.; Legg, C.; and Witten, I. 2009. Mining meaning from Wikipedia. *Int. J. Hum.-Comp. Stud.* 67:716–754.

Okoli, C.; Mehdi, M.; Mesgari, M.; Årup Nielsen, F.; and Lanamäki, A. 2012. *The People's Encyclopedia Under the Gaze of the Sages: A Systematic Review of Scholarly Research on Wikipedia*. SSRN eLibrary.

Panciera, K.; Halfaker, A.; and Terveen, L. 2009. Wikipedians are born, not made: A study of power editors on Wikipedia. In *Proc. of GROUP*, 51–60.

Poese, I.; Uhlig, S.; Kâafar, M. A.; Donnet, B.; and Gueye, B. 2011. IP geolocation databases: unreliable? *Comp. Communic. Rev.* 41(2):53–56.

Potthast, M., and Holfeld, T. 2011. Overview of the 2nd international competition on Wikipedia vandalism detection. In *Proc. of CLEF*.

Potthast, M.; Stein, B.; and Gerling, R. 2008. Automatic vandalism detection in Wikipedia. In *Proc. of ECIR*, 663–668.

Potthast, M.; Stein, B.; and Holfeld, T. 2010. Overview of the 1st international competition on Wikipedia vandalism detection. In *Proc. of CLEF*.

Potthast, M. 2010. Crowdsourcing a Wikipedia vandalism corpus. In *Proc. of SIGIR*, 789–790.

Priedhorsky, R.; Chen, J.; Lam, S.; Panciera, K.; Terveen, L.; and Riedl, J. 2007. Creating, destroying, and restoring value in Wikipedia. In *Proc. of GROUP*.

Rzeszotarski, J. M., and Kittur, A. 2012. Learning from history: Predicting reverted work at the word level in Wikipedia. In *Proc. of CSCW*, 437–440.

Schroeder, R., and Taylor, L. 2015. Big data and Wikipedia research: Social science knowledge across disciplinary divides. *Inform., Communic. & Soc.* 1–18.

Shachaf, P., and Hara, N. 2010. Beyond vandalism: Wikipedia trolls. *J. Inf. Sci.* 36(3):357–370.

Shavitt, Y., and Zilberman, N. 2011. A geolocation databases study. *IEEE J. on Sel. Ar. in Communic.* 29(10):2044–2056.

Steiner, T. 2014. Bots vs. Wikipedians, anons vs. logged-ins (redux): A global study of edit activity on Wikipedia and Wikidata. In *Proc. of OpenSym*, 25:1–25:7.

Suh, B.; Convertino, G.; Chi, E.; and Pirolli, P. 2009. The singularity is not near: slowing growth of Wikipedia. In *Proc. of WikiSym*, 1–10.

Tran, K., and Christen, P. 2013. Cross language prediction of vandalism on Wikipedia using article views and revisions. In *Proc. of PAKDD*, 268–279.

Tran, K., and Christen, P. 2015. Cross-language learning from bots and users to detect vandalism on Wikipedia. *IEEE Trans. Knowl. Data Eng.* 27(3):673–685.

Wagner, C.; Garcia, D.; Jadidi, M.; and Strohmaier, M. 2015. It's a man's Wikipedia? assessing gender inequality in an online encyclopedia. In *Proc. of ICWSM*, 454–463.

Welser, H. T.; Cosley, D.; Kossinets, G.; Lin, A.; Dokshin, F.; Gay, G.; and Smith, M. 2011. Finding social roles in Wikipedia. In *Proc. of iConference*, 122–129.

West, A.; Kannan, S.; and Lee, I. 2010. Detecting Wikipedia vandalism via spatio-temporal analysis of revision metadata. In *Proc. of EUROSEC*, 22–28.

Yang, D.; Halfaker, A.; Kraut, R. E.; and Hovy, E. H. 2016. Who did what: Editor role identification in Wikipedia. In *Proc. of ICWSM*, 446–455.

Yasseri, T.; Sumi, R.; and Kertész, J. 2012. Circadian patterns of Wikipedia editorial activity: A demographic analysis. *PLOS One* 7(1):e30091.