

Spatio-temporal Analysis of Reverted Wikipedia Edits

Supplementary Material

Johannes Kiesel and Martin Potthast and Matthias Hagen and Benno Stein

Bauhaus-Universität Weimar

<first name>.<last name>@uni-weimar.de

This document contains the supplementary material to the paper *Spatio-temporal Analysis of Reverted Wikipedia Edits*, which is published at the 11th International AAAI Conference on Web and Social Media (ICWSM 17). You can access the paper here: http://www.uni-weimar.de/medien/webis/publications/papers/stein_2017e.pdf.

All analysis is based on the May 1st 2016 history dumps of the Wikipedia variants.

The material is structured by the variant of Wikipedia used in the analysis. The analyzed variants are the ones with the most edits in May 2016: (1) English, (2) German, (3) French, (4) Spanish, (5) Russian, (6) Italian, and (7) Japanese.

For each variant, the results of the analysis (tables and figures) are provided according to the same order that is used in the paper: Mining Vandalism (Section 3 in the paper), Geolocating Editors (Section 4), and Spatio-Temporal Analysis (Section 5).

Please consult the paper for more information on the procedure that produced the results.

We now list some interesting observations that we could not discuss in the paper:

Mining Vandalism

- While the steps we employ to filter harmless or ambiguous reverts are language independent, the percentage of reverts they filter is rather different for the different variants of Wikipedia. After filtering ambiguous reverts, the following percentages of assumed vandalism edits remain (in comparison to before filtering): Russian 48.4%, Spanish 48.1%, French 44.7%, Italian 41.1%, German 38.1%, Japanese 35.9%, and English 33.1%. This hints at different frequencies of revert patterns in the Wikipedias. For example, while in the English Wikipedia 37.5% of the original reverted edits are due to reverts between page blanks, the same statistic is only 6.1% for the Russian Wikipedia.
- For all variants of Wikipedia, the number of reverts with a specific number of reverted edits follow clearly a exponential model.

Geolocating Editors

- For all variants of Wikipedia, most anonymous edits can be geolocated with an agreement of all GeoDBs in the RIR entry time span (step 4 in the decision tree): Japanese 94.6%, German 90.6%, French 88.2%, Italian 84.5%, Russian 81.0%, Spanish 80.1%, English 72.3%.
- Most of these geolocations mentioned above use 7 or more of the 10 GeoDBs. This is likely due to 3 GeoDBs being from 2008 and therefore considerably older than the other 7 GeoDBs (from 2014 to 2016). Therefore, an IP-based geolocation seems to be rather stable for the last few years, and 2 year old GeoDBs seem to be still usable for current geolocations.

Spatio-Temporal Analysis

- The highest vandalism ratio we observed from Chile (~52%), as usually between 8 to 9 hours (p. 34).
- For most analyzed countries, Summer is the season with the lowest vandalism with a distinctly flat course over the day. Exceptions are: The Philippines, where Spring takes the place of Summer (p. 10); Colombia, where Summer is on the same level as Spring and Winter (p. 35); and Japan, where there is nearly no vandalism at any time (p. 55). Note that we use the hemisphere of the geolocation to determine the season of a edit.
- Like most other analyzed countries, Mexico, Colombia, Argentina, Peru, and Venezuela also have one peaks at 8 to 9 hours and another one in the afternoon. However, the afternoon peak is much more dominant for these countries (p. 34, 35, 36). Surprisingly, this is very different to Chile (p. 34).
- The vandalism ratios for the Russian Wikipedia do not change much over the day—also for edits from the Ukraine and Belarus (p. 42, 43). There is only a slight increase in vandalism on Monday to Thursday afternoons.
- For the Italian Wikipedia, Friday is similar to Monday to Thursday, but Saturday follows the usual course of Friday (like a working day in the morning and like a weekend day in the afternoon; p. 49).

English Wikipedia

English Wikipedia: Mining Vandalism

Table 1: Step-by-step filtering of the English Wikipedia as per the revert patterns depicted in Figure 2 in the paper. 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. Note that the approach by Kittur et al. uses specific words to classify comments and is thus likely less effective in finding vandalism for other languages than English.

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%

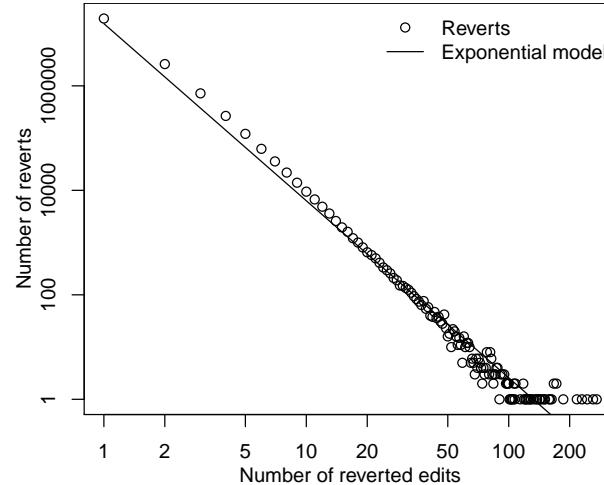


Figure 1: Number of full page reverts with a specific number of edits they revert and fitted exponential model in a log-log plot.

English Wikipedia: Geolocating Editors

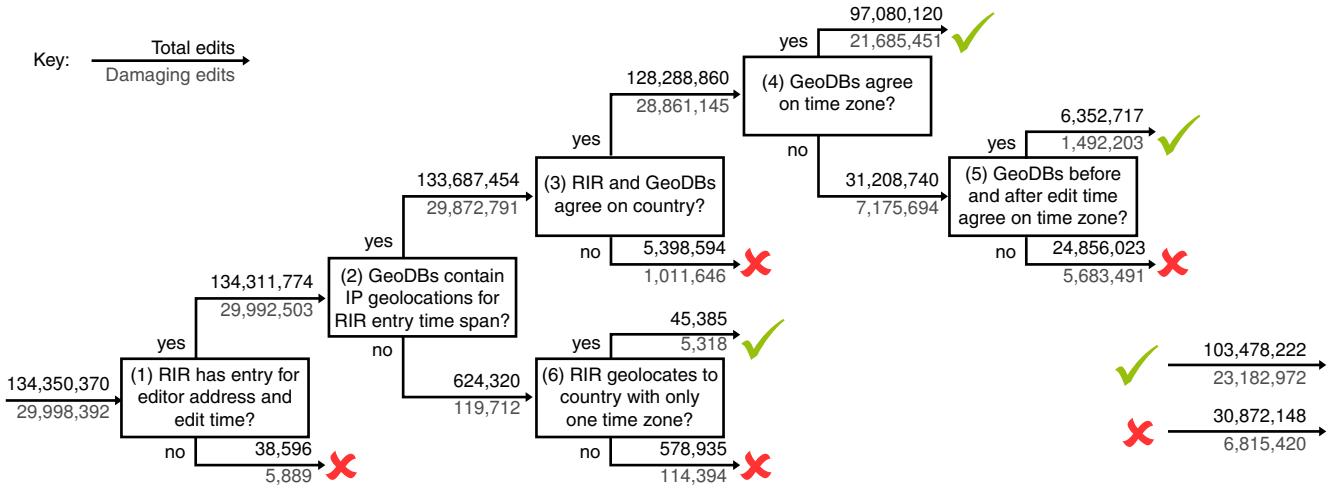


Figure 2: Decision tree to decide whether to trust the available geolocation information for an edit (✓), or not (✗). The numbers denote the total edits and reverted edits for the English Wikipedia that went through each branch.

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 of the paper are given, and their corresponding IP addresses. Numbers are given for each exit node of the decision tree in the Figure above, 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
<i>Entire Wikipedia</i>		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

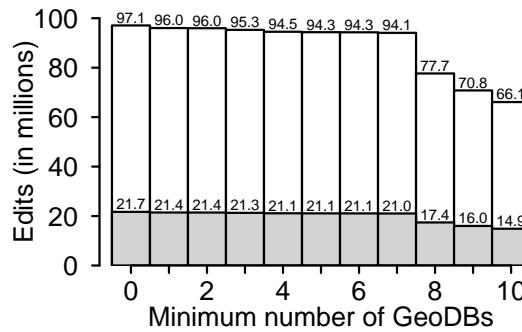


Figure 3: Number of total edits (white bars) and vandalism edits (gray bars) in millions from the yes-branch of Step (4) above over the number of GeoDBs considered.

English Wikipedia: Spatio-Temporal Analysis

Table 3: Number of edits by country, sorted for the English Wikipedia.

Country	Wikipedia						
	enwiki	dewiki	frwiki	eswiki	ruwiki	itwiki	jawiki
United States	44,126,652	445,587	122,246	282,357	130,264	43,855	91,843
United Kingdom	17,111,051	145,689	80,501	53,189	30,417	51,014	17,620
Canada	5,445,237	39,686	119,545	21,444	19,949	5,090	20,117
India	4,553,192	10,202	4,332	1,851	2,156	2,175	2,839
Australia	2,762,004	21,716	4,682	4,291	3,897	4,063	3,718
Germany	2,052,495	13,525,123	150,206	101,962	141,422	75,183	26,556
Philippines	1,997,179	12,621	3,298	4,795	586	1,718	3,535
Ireland	1,209,355	10,331	6,616	5,537	1,664	3,431	544
France	1,070,256	109,914	10,422,471	92,355	27,640	68,339	21,491
Netherlands	1,066,776	73,382	36,186	20,550	10,744	19,182	4,662
Italy	1,011,100	92,872	76,763	59,310	14,900	10,292,637	8,352
New Zealand	999,105	12,001	2,332	1,876	1,989	768	2,677
Spain	945,915	62,094	116,834	5,671,664	15,325	61,848	7,191

Table 4: Number of vandalism edits by country, sorted for the English Wikipedia.

Country	Wikipedia						
	enwiki	dewiki	frwiki	eswiki	ruwiki	itwiki	jawiki
United States	11,243,775	96,374	21,009	66,696	12,007	8,432	6,954
United Kingdom	3,927,965	29,699	10,917	8,751	3,626	5,727	1,370
Canada	1,339,244	8,234	36,925	3,582	3,134	885	1,321
India	1,108,217	2,830	1,004	412	276	339	372
Australia	724,552	5,826	822	933	417	1,022	343
Philippines	439,656	2,009	391	492	79	317	328
Ireland	276,025	1,896	964	824	295	431	35
New Zealand	263,608	2,918	317	275	134	85	170
Germany	212,108	3,490,244	12,353	9,932	13,260	6,907	1,490
Netherlands	157,374	12,752	3,324	2,368	1,187	1,667	408
France	135,392	17,278	1,774,797	12,193	3,162	7,214	994
Spain	135,325	11,538	11,524	1,279,946	1,748	5,756	385
Italy	106,007	19,459	8,063	7,116	1,738	1,365,327	725

Table 5: Number of vandalism commented edits by country, sorted for the English Wikipedia.

Country	Wikipedia						
	enwiki	dewiki	frwiki	eswiki	ruwiki	itwiki	jawiki
United States	2,490,888	14,611	2,232	4,331	72	289	232
United Kingdom	808,153	4,077	1,025	449	21	141	53
Canada	290,507	1,327	5,767	211	12	20	83
India	212,978	326	95	15		6	5
Australia	165,232	1,005	127	74	6	55	6
Philippines	82,764	185	33	8		2	10
New Zealand	59,541	438	58	12		2	5
Ireland	58,636	227	94	66	1	16	1
Germany	32,883	136,244	1,003	403	67	140	43
Netherlands	29,576	474	265	96	1	27	9
Spain	22,433	352	951	102,690	8	123	41
France	18,109	597	172,680	554	12	131	26
Italy	11,749	627	501	184	7	31,368	12

The tables above show the number of certain kinds of edits from specific countries in all analyzed variants of Wikipedia. Countries are selected to have at least 100,000 vandalism edits in the English Wikipedia, or that have English as a major language (according to the English Wikipedia) and at least 1,000 vandalism edits. Countries are sorted by the number of all edits in the English Wikipedia. The same countries are used in the hour-of-day plots.

Vandalism commented edits refers to edits that are—after filtering reverts—affected by a revert with a comment that signals that it is undoing vandalism (see Kittur et al. (2007b)). Note that this approach uses specific words to classify comments and is thus likely less effective in finding vandalism for other languages than English.

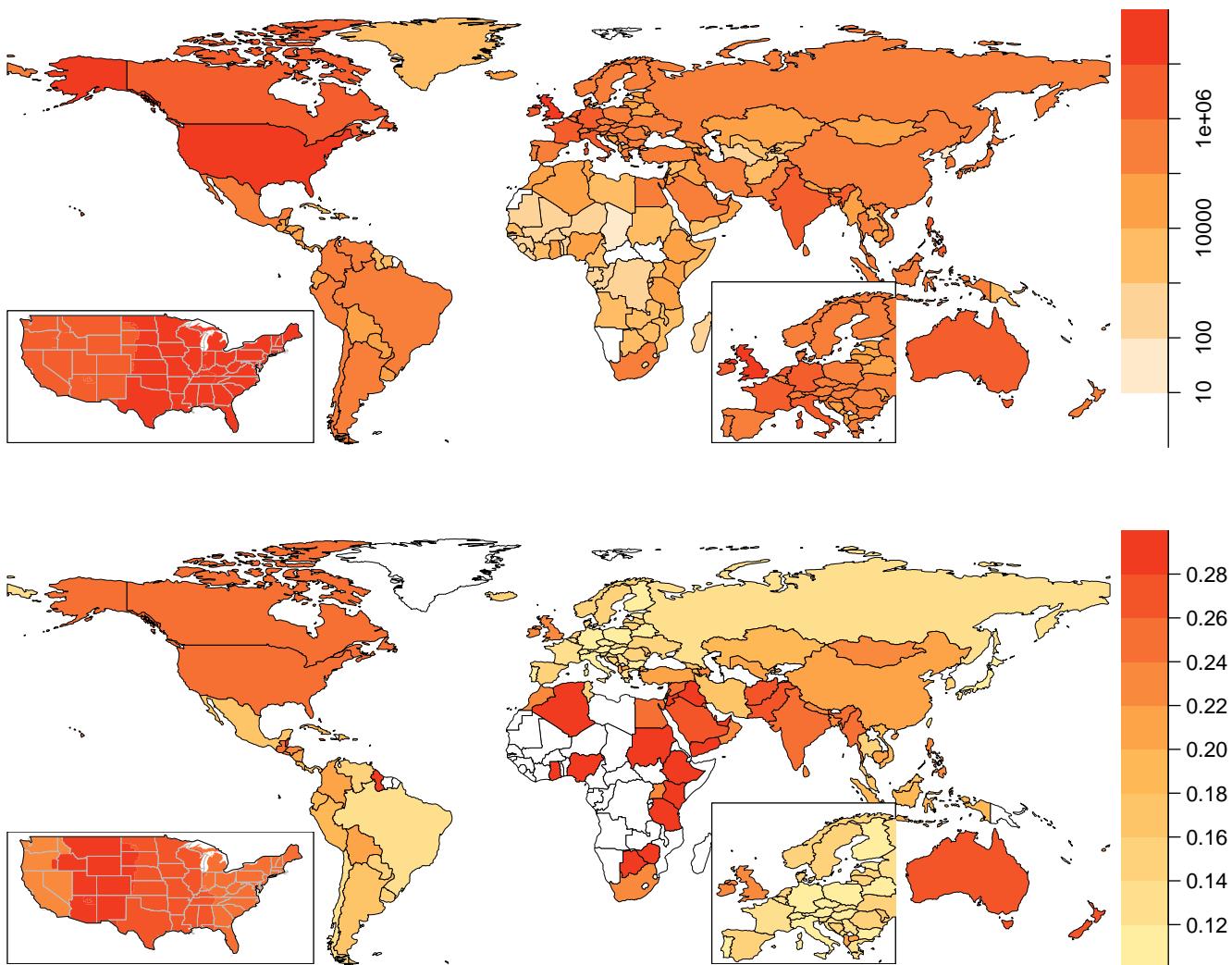
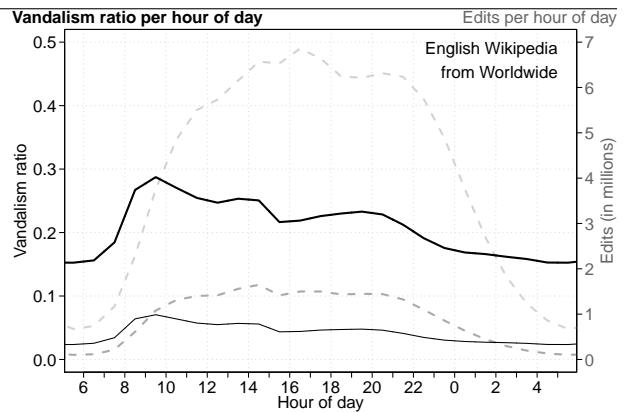
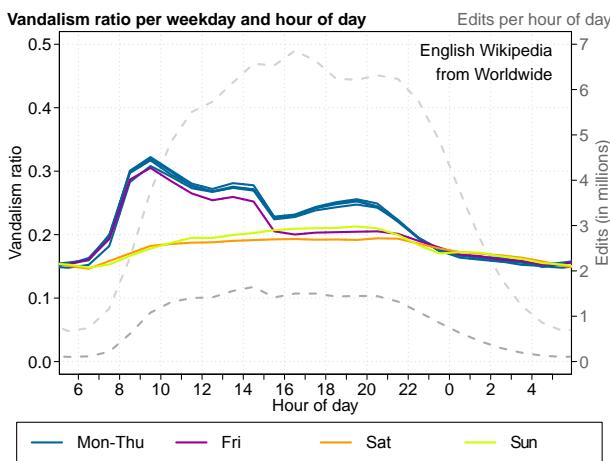


Figure 4: Number of edits (top) and ratio of vandalism to all edits (bottom) in the English Wikipedia by country. Countries with less than 1,000 vandalism edits are not colored. The embedded small maps show (left) the vandalism ratio in the United States (without Alaska) by major time zone (from West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

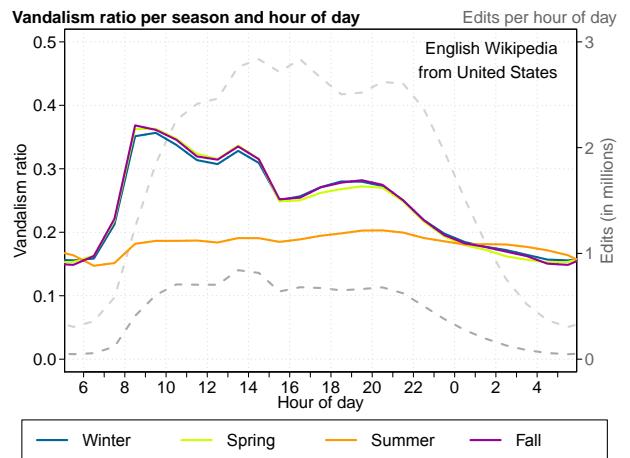
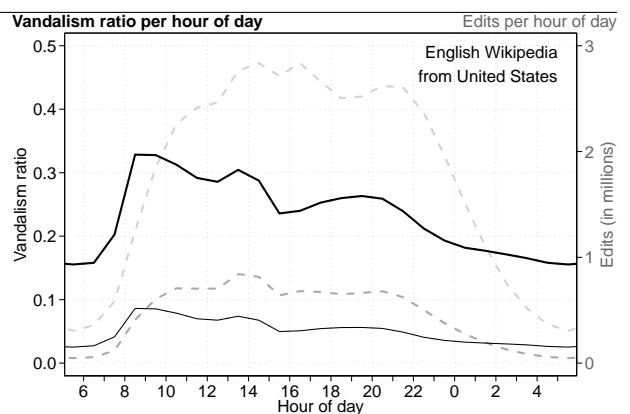
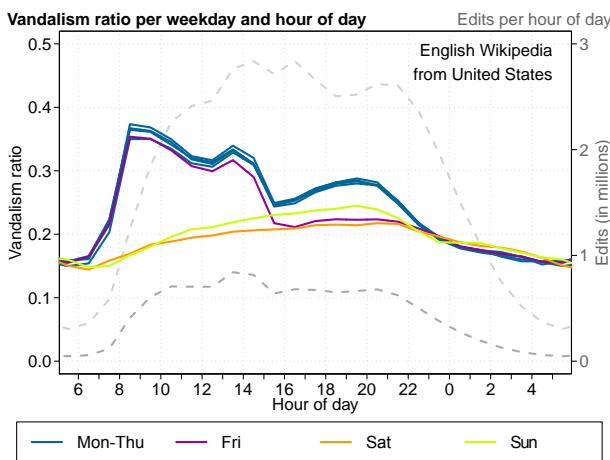
English Wikipedia

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 (light gray) and vandalism edits (dark gray) per hour of day (right axis, dashed lines), both averaged over the English Wikipedia's history. Ratios estimated from less than 1,000 vandalism edits are displayed with dotted lines.

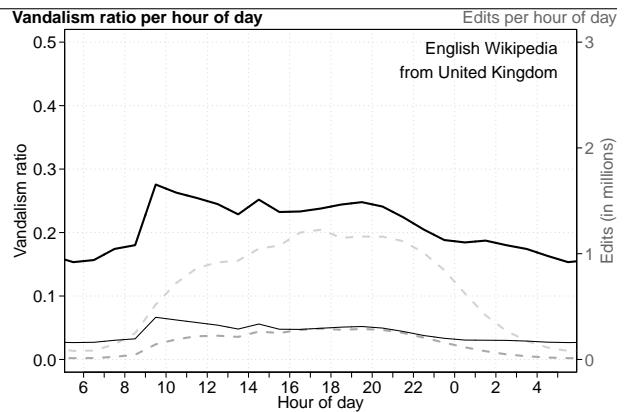
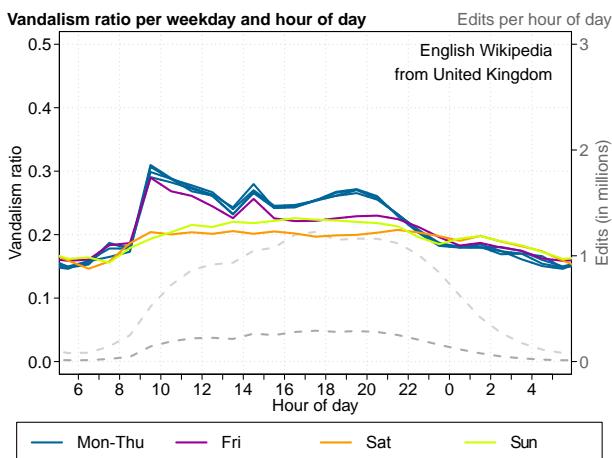
The thin black line to the right is the vandalism ratio when detecting vandalism by comments of reverts (see Kittur et al. (2007b)). Note that this approach uses specific words to classify comments and is thus likely less effective in finding vandalism for other languages than English.



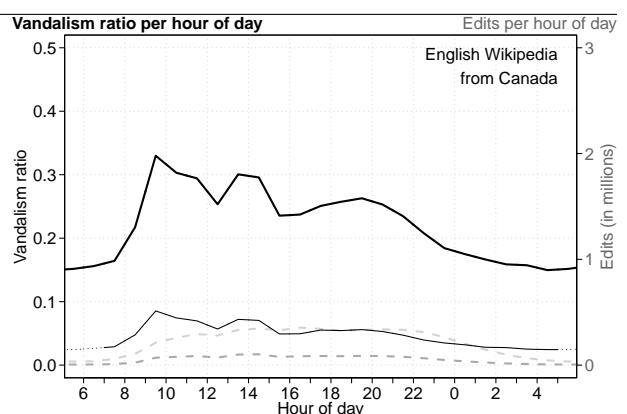
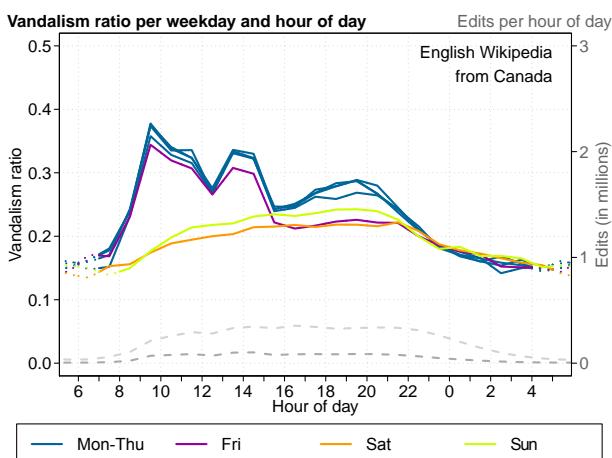
English Wikipedia Edits from United States



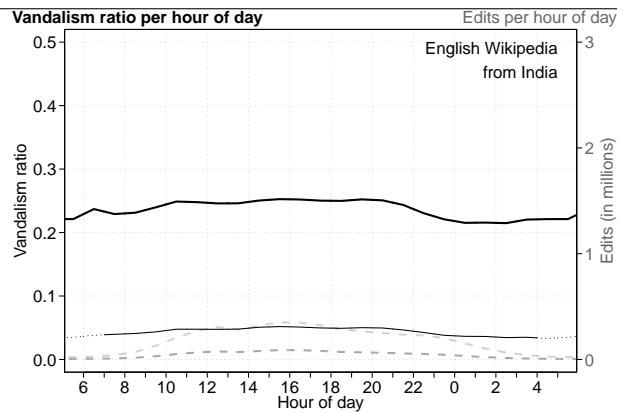
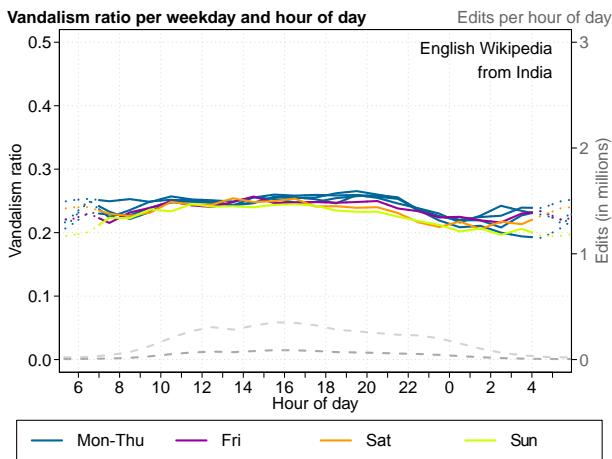
English Wikipedia Edits from United Kingdom



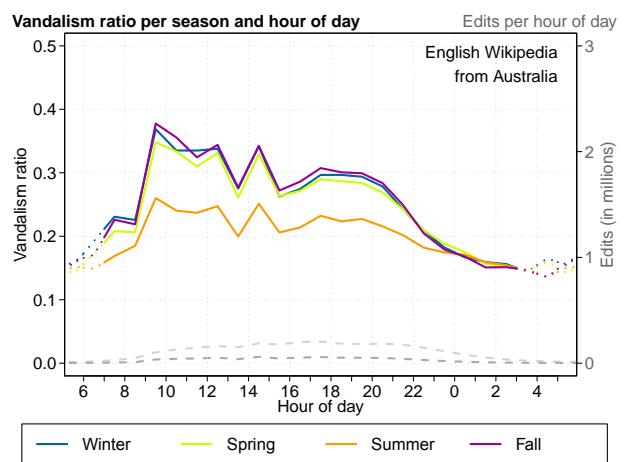
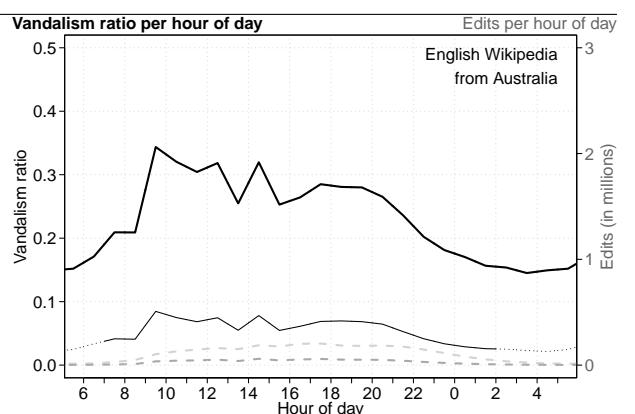
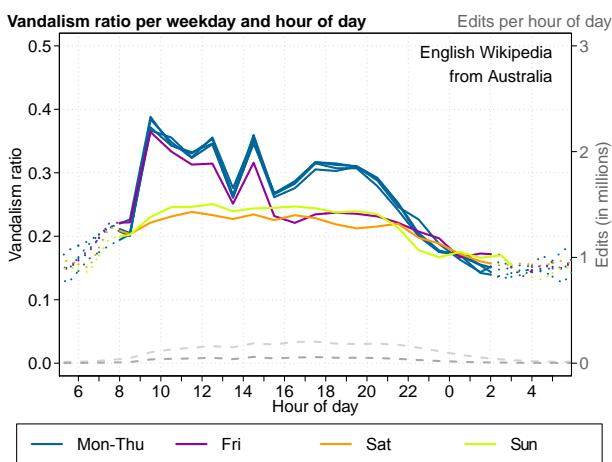
English Wikipedia Edits from Canada



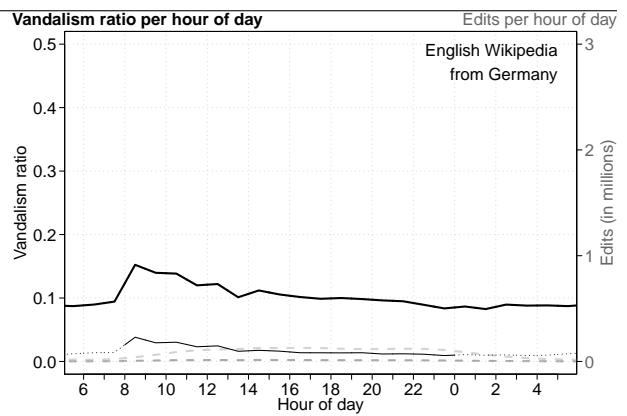
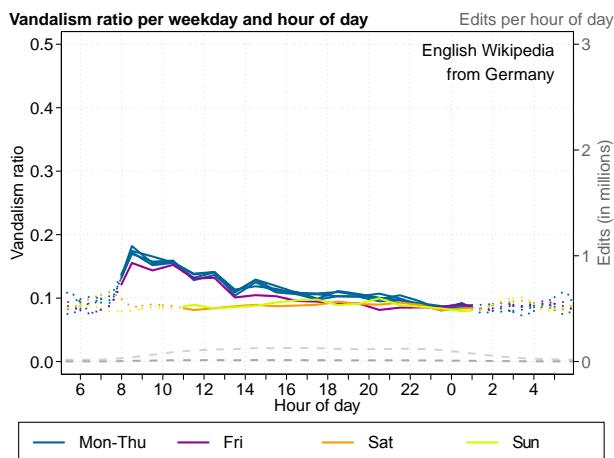
English Wikipedia Edits from India



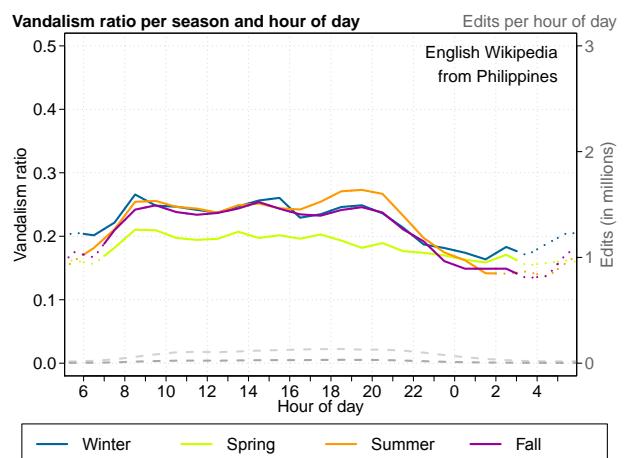
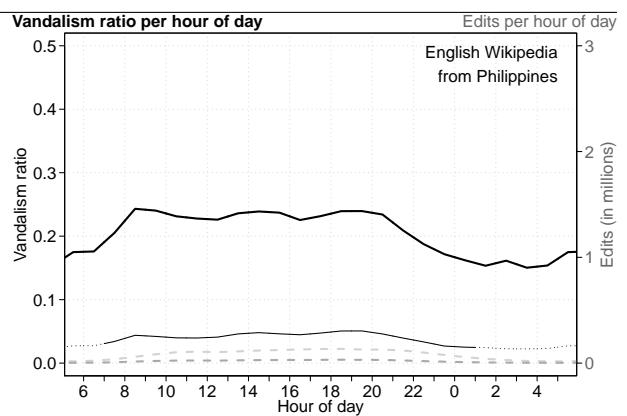
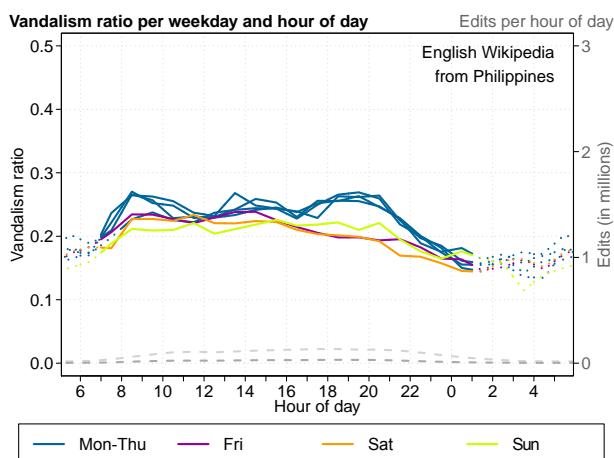
English Wikipedia Edits from Australia



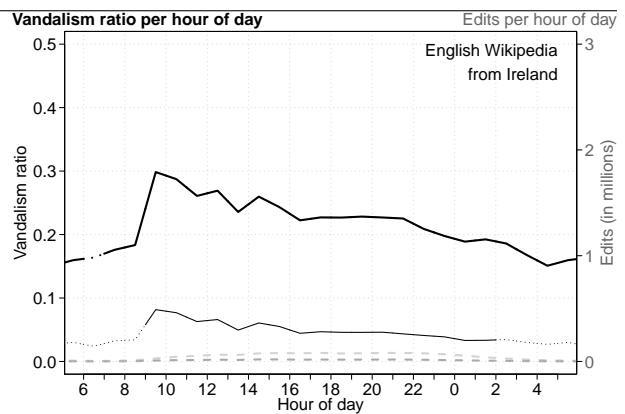
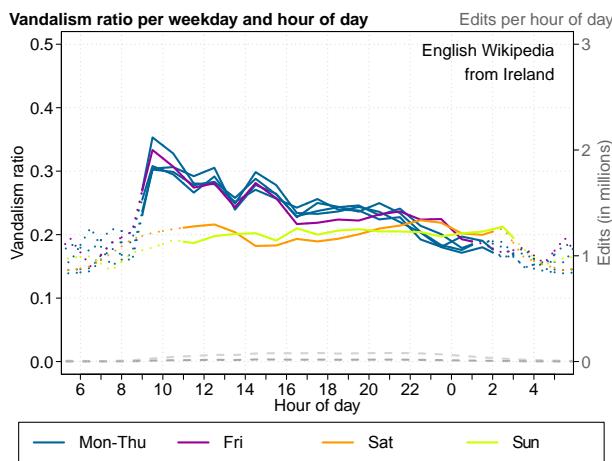
English Wikipedia Edits from Germany



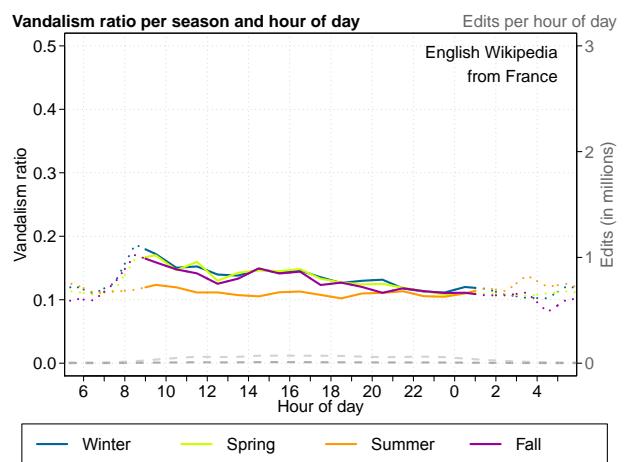
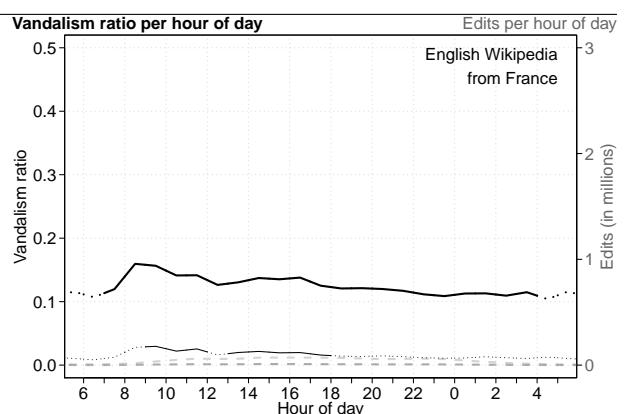
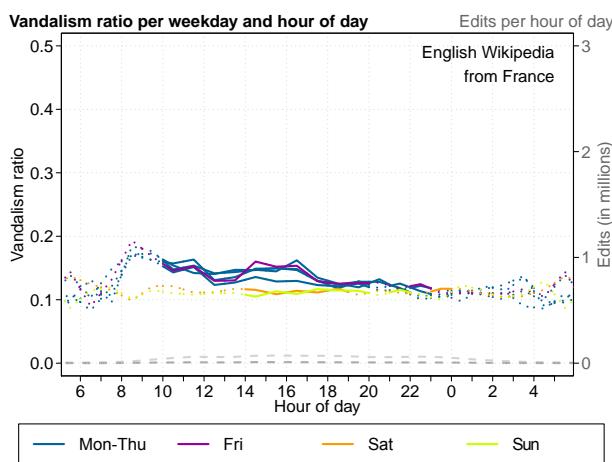
English Wikipedia Edits from Philippines



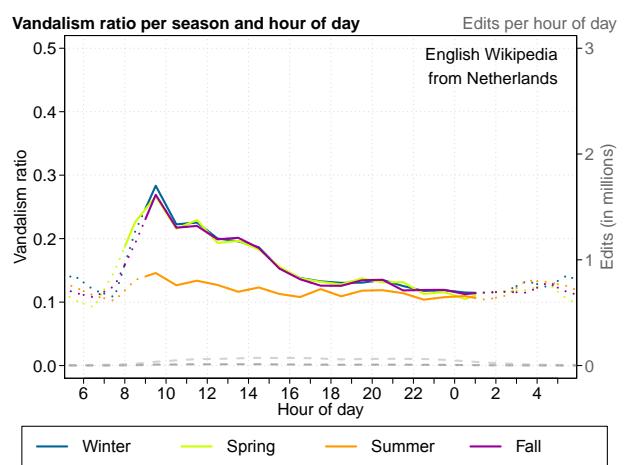
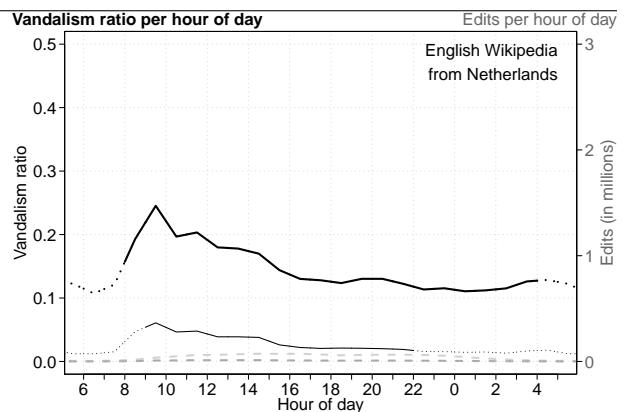
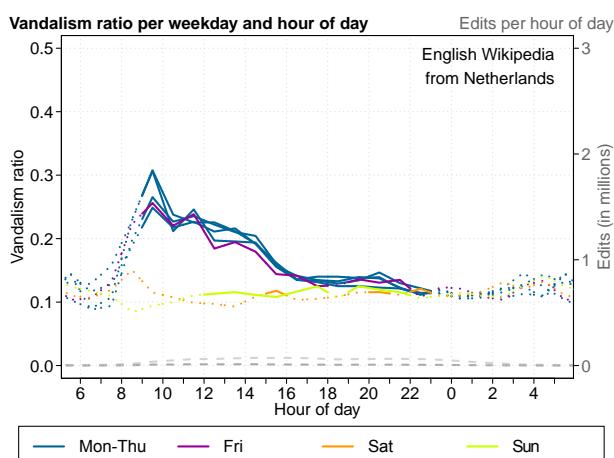
English Wikipedia Edits from Ireland



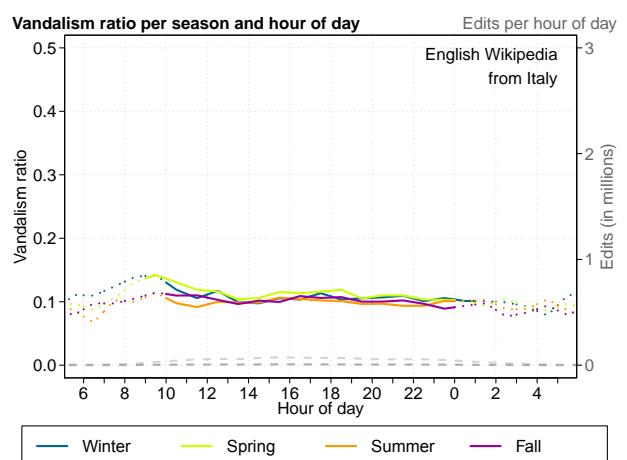
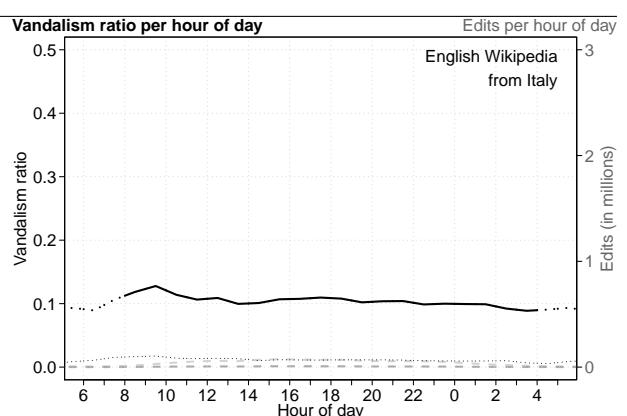
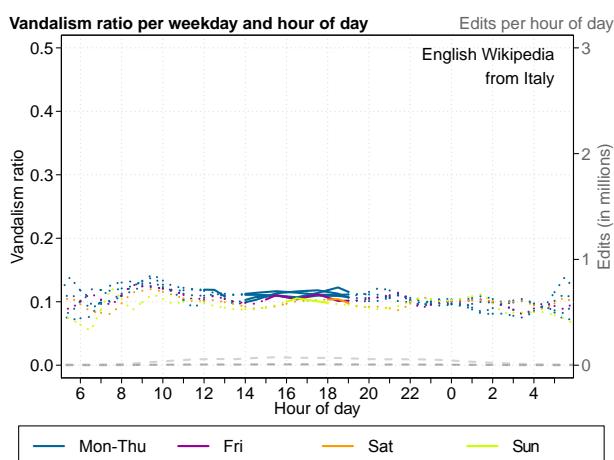
English Wikipedia Edits from France



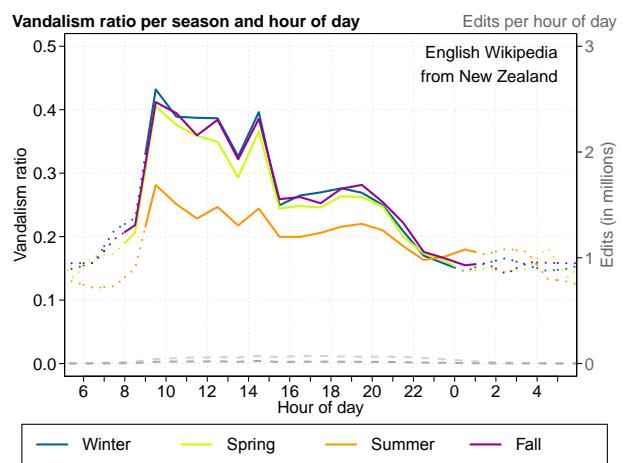
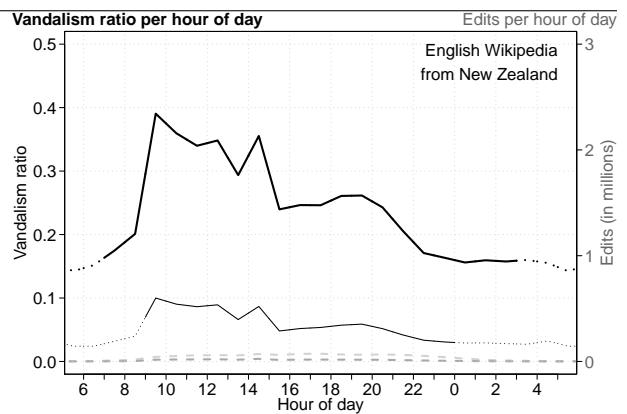
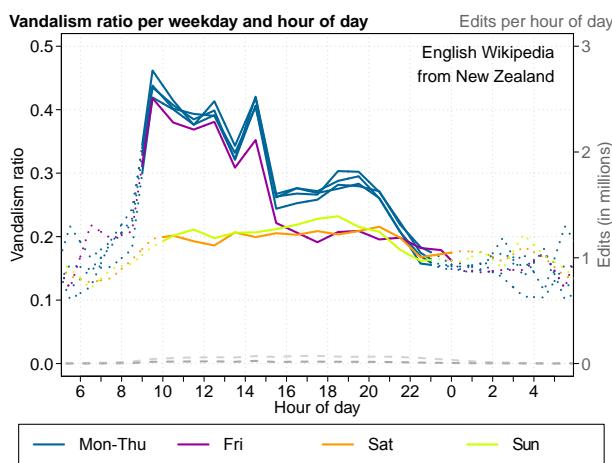
English Wikipedia Edits from Netherlands



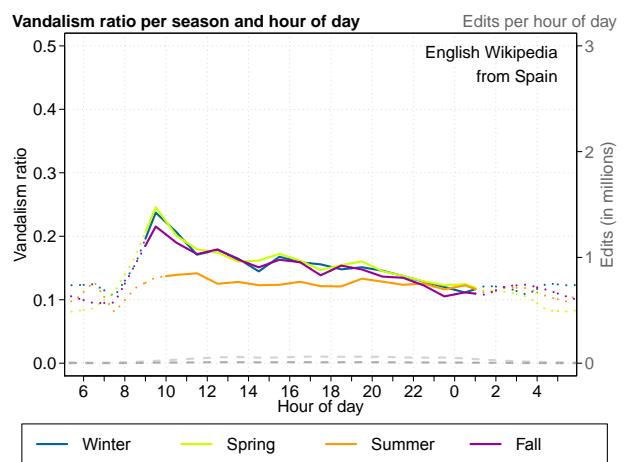
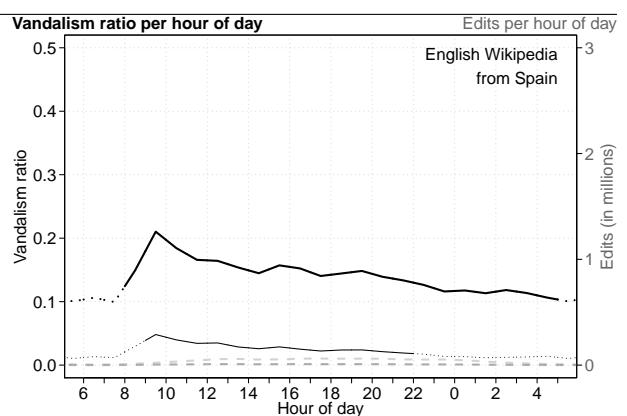
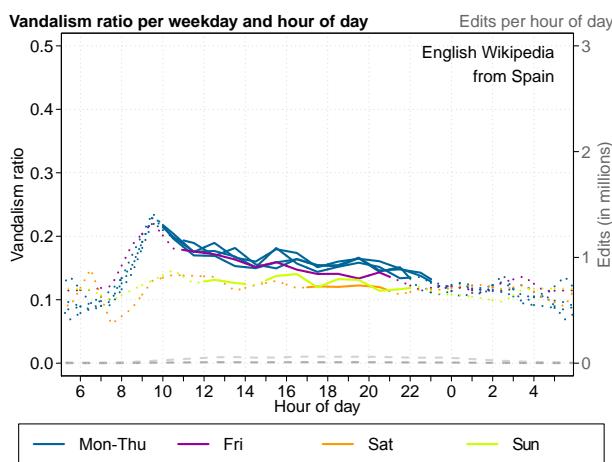
English Wikipedia Edits from Italy



English Wikipedia Edits from New Zealand



English Wikipedia Edits from Spain



German Wikipedia

German Wikipedia: Mining Vandalism

Table 6: Step-by-step filtering of the German Wikipedia as per the revert patterns depicted in Figure 2 in the paper. 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. Note that the approach by Kittur et al. uses specific words to classify comments and is thus likely less effective in finding vandalism for other languages than English.

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	7,193,555	234,898	7,428,453	100.0%	7,884,757	7,095,909	405,504	15,386,170	100.0%	
(a) reverts to page blank	-67,193	-153	-67,346	-0.9%	-1,122,812	-2,067,105	-146,699	-3,336,616	-21.7%	
(b) empty reverts due to renaming/removal/error	-670,186	-25,112	-695,298	-9.4%	0	0	0	0	0.0%	
Results after filtering pseudo-reverts	Σ	6,456,176	209,633	6,665,809	89.7%	6,761,945	5,028,804	258,805	12,049,554	78.3%
(c) self reverts		-520,826	-2,855	-523,681	-7.0%	-254,252	-363,729	-15,228	-633,209	-4.1%
(d) revert corrections		-31,818	-1,996	-33,814	-0.5%	-166,220	-302,829	-18,878	-487,927	-3.2%
(e) reverted reverts		-36,474	-974	-37,448	-0.5%	-102,685	-210,469	-6,375	-319,529	-2.1%
Results after filtering error-corrections	Σ	5,867,058	203,808	6,070,866	81.7%	6,238,788	4,151,777	218,324	10,608,889	69.0%
(f) interleaved reverts		-915,733	-25,628	-941,361	-12.7%	-698,119	-1,286,015	-45,101	-2,029,235	-13.2%
(g) reverts reverting more than one editor		-181,422	-10,738	-192,160	-2.6%	-951,365	-1,648,406	-114,185	-2,713,956	-17.6%
Results after filtering ambiguous reverts	Σ	4,769,903	167,442	4,937,345	66.5%	4,589,304	1,217,356	59,038	5,865,698	38.1%
(h1) reverts reverting registered editors or bots		-1,022,587	-9,285	-1,031,872	-13.9%	0	-1,217,356	-59,038	-1,276,394	-8.3%
(h2) reverts reverting editors with IPv6 addresses		-54,374	-769	-55,143	-0.7%	-63,484	0	0	-63,484	-0.4%
Results after all filtering steps	Σ	3,692,942	157,388	3,850,330	51.8%	4,525,820	0	0	4,525,820	29.4%

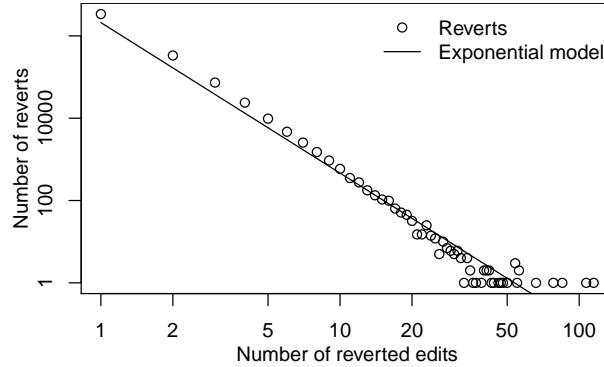


Figure 5: Number of full page reverts with a specific number of edits they revert and fitted exponential model in a log-log plot.

German Wikipedia: Geolocating Editors

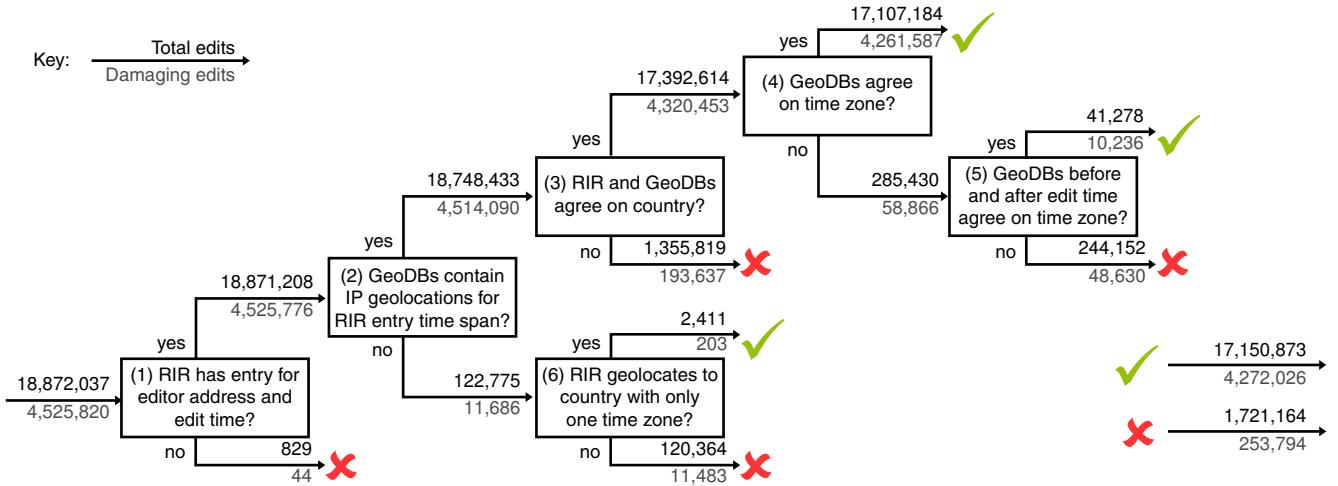


Figure 6: Decision tree to decide whether to trust the available geolocation information for an edit (✓), or not (✗). The numbers denote the total edits and reverted edits for the German Wikipedia that went through each branch.

Table 7: Historic geolocation success for all anonymous editors of the German 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 of the paper are given, and their corresponding IP addresses. Numbers are given for each exit node of the decision tree in the Figure above, 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
<i>Entire Wikipedia</i>		4,525,820 (23%)	18,872,037	2,564,936	7,223,295
No ✗	Step (1)	44 (5%)	829	33	179
	Step (3)	193,637 (14%)	1,355,819	67,355	258,559
	Step (5)	48,630 (19%)	244,152	29,653	112,192
	Step (6)	11,483 (9%)	120,364	6,785	40,891
	Σ	253,794 (14%)	1,721,164	103,821	411,443
Yes ✓	Step (4)	4,261,587 (24%)	17,107,184	2,454,695	6,791,177
	Step (5)	10,236 (24%)	41,278	6,463	21,137
	Step (6)	203 (8%)	2,411	111	807
	Σ	4,272,026 (24%)	17,150,873	2,461,267	6,813,072

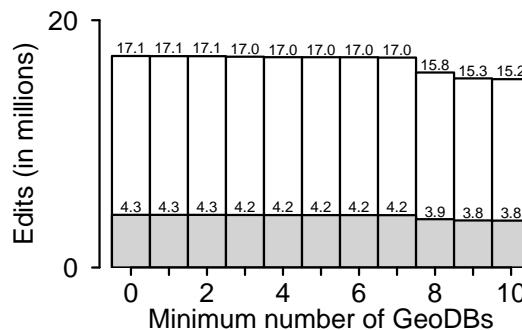


Figure 7: Number of total edits (white bars) and vandalism edits (gray bars) in millions from the yes-branch of Step (4) above over the number of GeoDBs considered.

German Wikipedia: Spatio-Temporal Analysis

Table 8: Number of edits by country, sorted for the German Wikipedia.

Country	Wikipedia						
	enwiki	dewiki	frwiki	eswiki	ruwiki	itwiki	jawiki
Germany	2,052,495	13,525,123	150,206	101,962	141,422	75,183	26,556
Austria	218,922	1,076,741	12,444	10,157	6,415	11,937	2,790
Switzerland	311,700	855,089	381,326	19,294	8,081	94,654	2,785

Table 9: Number of vandalism edits by country, sorted for the German Wikipedia.

Country	Wikipedia						
	enwiki	dewiki	frwiki	eswiki	ruwiki	itwiki	jawiki
Germany	212,108	3,490,244	12,353	9,932	13,260	6,907	1,490
Austria	25,514	270,906	1,159	948	937	997	124
Switzerland	37,260	185,746	45,872	2,002	991	13,114	141

Table 10: Number of vandalism commented edits by country, sorted for the German Wikipedia.

Country	Wikipedia						
	enwiki	dewiki	frwiki	eswiki	ruwiki	itwiki	jawiki
Germany	32,883	136,244	1,003	403	67	140	43
Austria	3,689	10,364	71	24	2	15	5
Switzerland	5,577	7,484	3,964	78	7	361	1

The tables above show the number of certain kinds of edits from specific countries in all analyzed variants of Wikipedia. Countries are selected to have at least 100,000 vandalism edits in the German Wikipedia, or that have German as a major language (according to the English Wikipedia) and at least 1,000 vandalism edits. Countries are sorted by the number of all edits in the German Wikipedia. The same countries are used in the hour-of-day plots.

Vandalism commented edits refers to edits that are—after filtering reverts—affected by a revert with a comment that signals that it is undoing vandalism (see Kittur et al. (2007b)). Note that this approach uses specific words to classify comments and is thus likely less effective in finding vandalism for other languages than English.

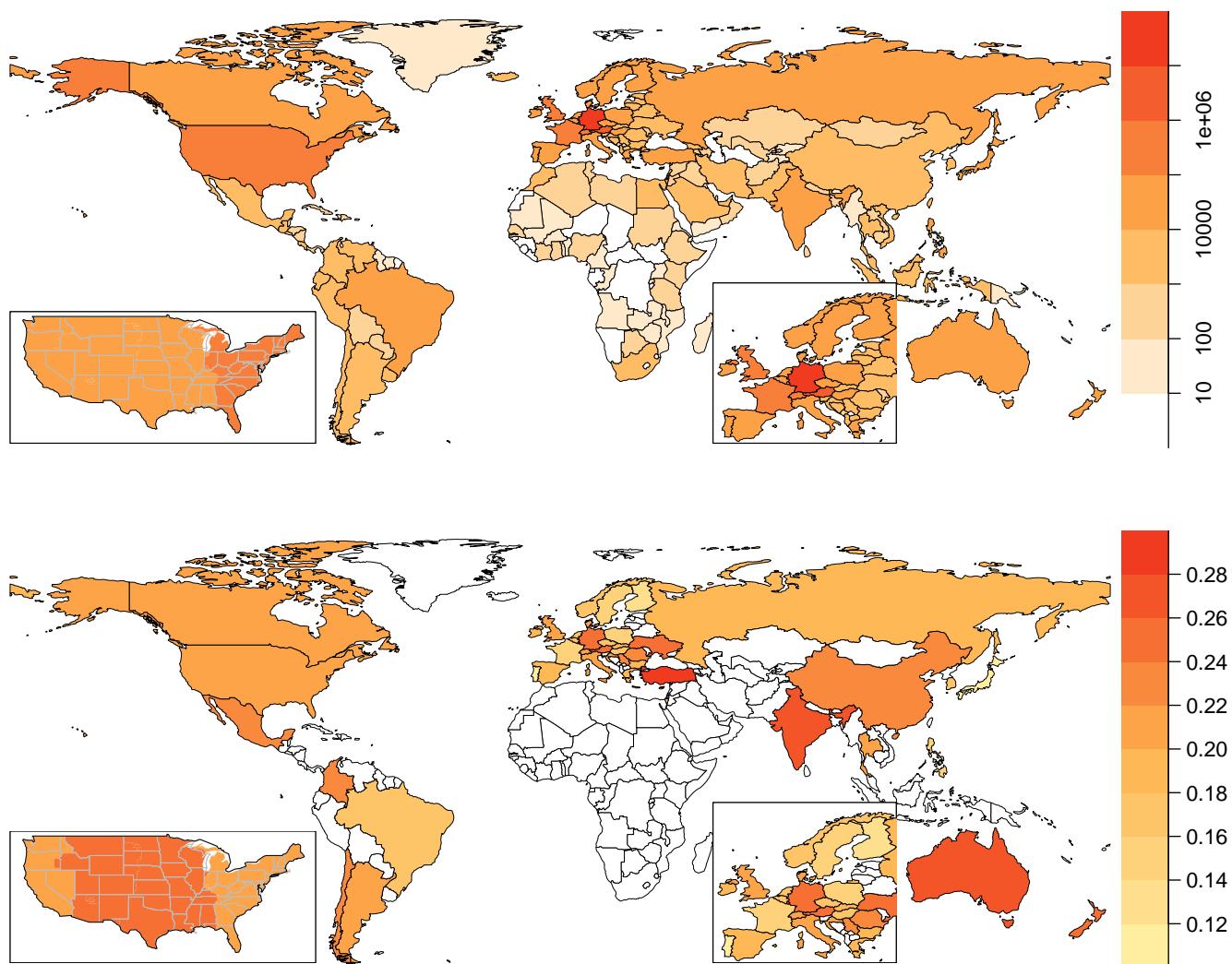
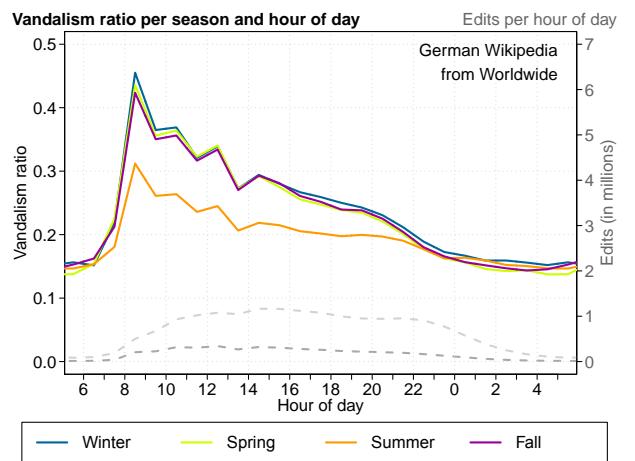
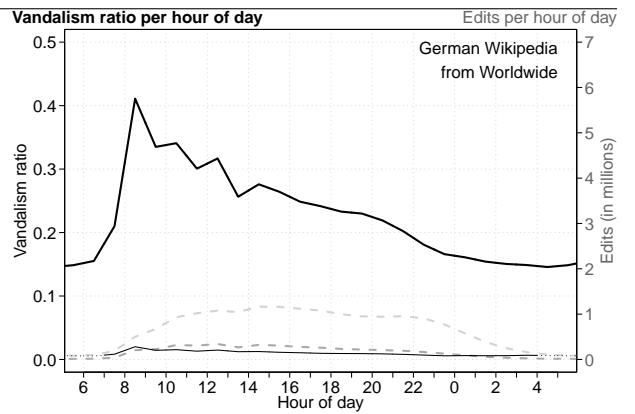
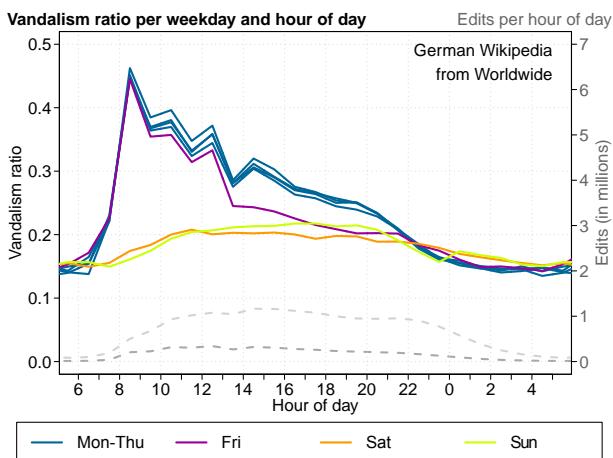


Figure 8: Number of edits (top) and ratio of vandalism to all edits (bottom) in the German Wikipedia by country. Countries with less than 1,000 vandalism edits are not colored. The embedded small maps show (left) the vandalism ratio in the United States (without Alaska) by major time zone (from West to East: Pacific, Mountain, Central, and Eastern) with overlaid state borders and (right) Europe enlarged.

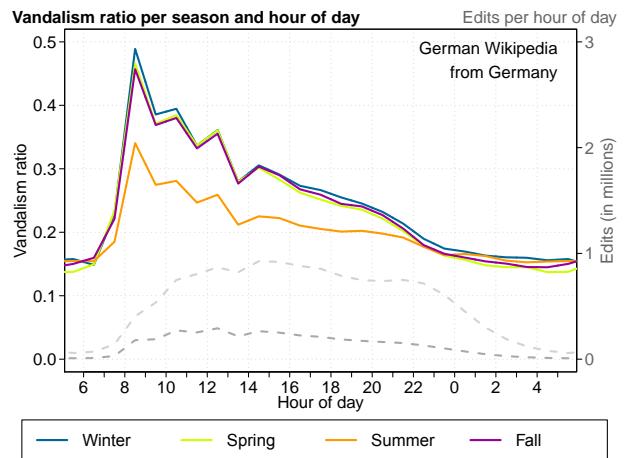
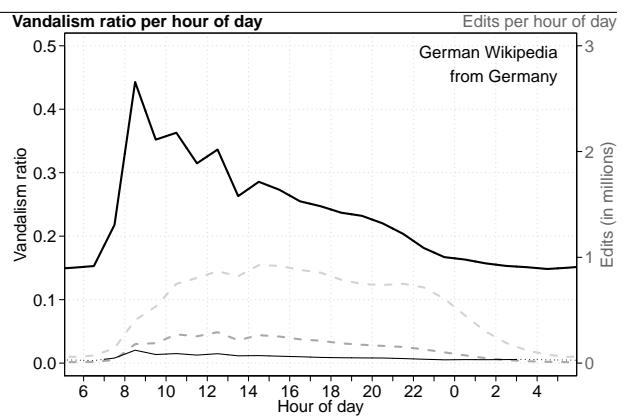
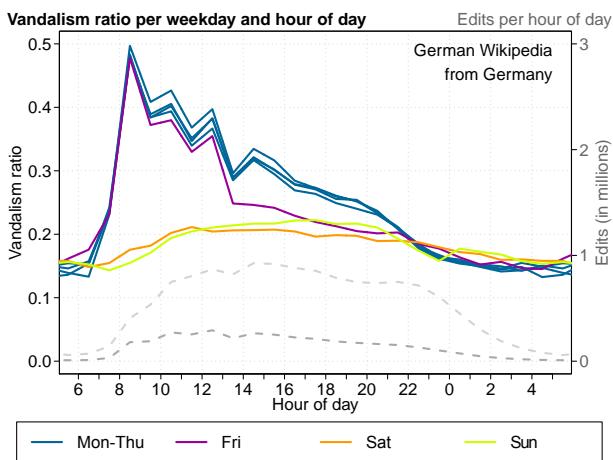
German Wikipedia

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 (light gray) and vandalism edits (dark gray) per hour of day (right axis, dashed lines), both averaged over the German Wikipedia's history. Ratios estimated from less than 1,000 vandalism edits are displayed with dotted lines.

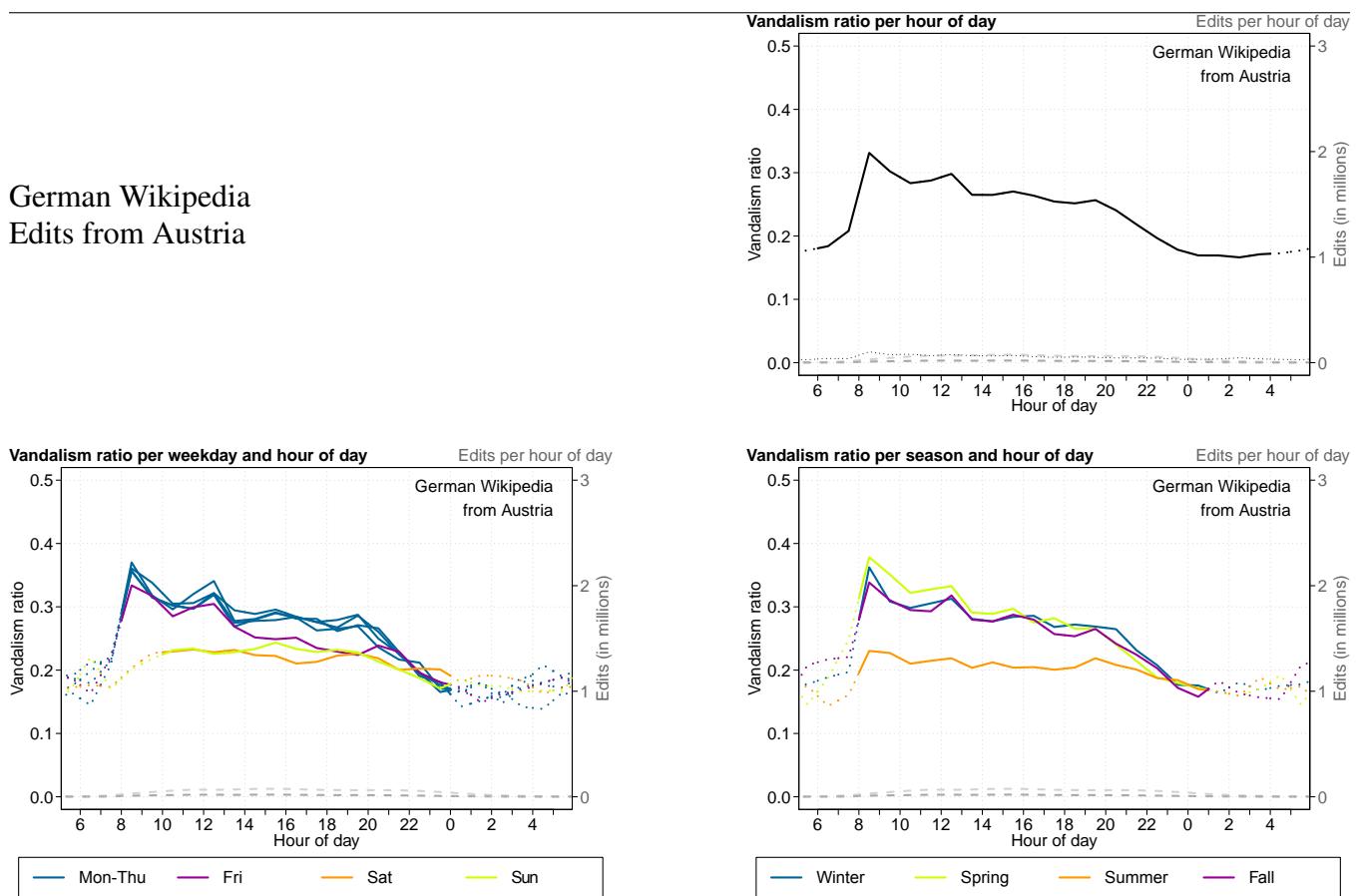
The thin black line to the right is the vandalism ratio when detecting vandalism by comments of reverts (see Kittur et al. (2007b)). Note that this approach uses specific words to classify comments and is thus likely less effective in finding vandalism for other languages than English.



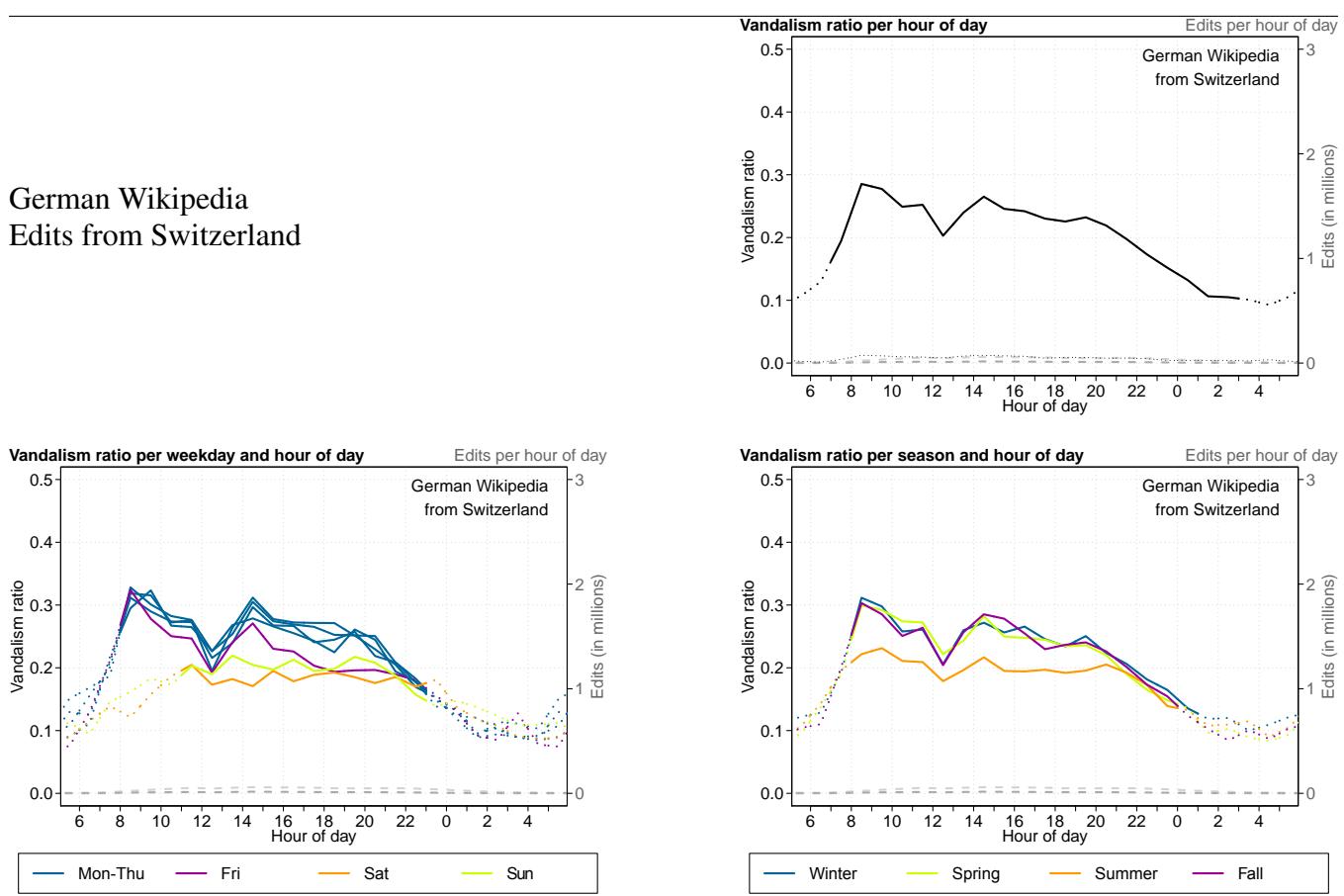
German Wikipedia Edits from Germany



German Wikipedia Edits from Austria



German Wikipedia Edits from Switzerland



French Wikipedia

French Wikipedia: Mining Vandalism

Table 11: Step-by-step filtering of the French Wikipedia as per the revert patterns depicted in Figure 2 in the paper. 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. Note that the approach by Kittur et al. uses specific words to classify comments and is thus likely less effective in finding vandalism for other languages than English.

Revert filtering step	Full page reverts				Reverted edits				Total			
	Vandalism as per Kittur		Total		Editor							
	No	Yes	Absolute	Relative	Anonymous	Registered	Bot					
Results of naive SHA-1 matching	4,245,228	269,545	4,514,773	100.0%	4,020,198	3,064,942	231,593	7,316,733	100.0%			
(a) reverts to page blank	-322,555	-1,724	-324,279	-7.2%	-698,071	-1,240,009	-98,408	-2,036,488	-27.8%			
(b) empty reverts due to renaming/removal/error	-489,754	-12,333	-502,087	-11.1%	0	0	0	0	0.0%			
Results after filtering pseudo-reverts	Σ	3,432,919	255,488	3,688,407	81.7%	3,322,127	1,824,933	133,185	5,280,245	72.2%		
(c) self reverts		-357,086	-352	-357,438	-7.9%	-165,381	-241,646	-17,612	-424,639	-5.8%		
(d) revert corrections		-43,524	-5,818	-49,342	-1.1%	-71,684	-153,936	-3,182	-228,802	-3.1%		
(e) reverted reverts		-17,589	-489	-18,078	-0.4%	-54,173	-104,668	-4,242	-163,083	-2.2%		
Results after filtering error-corrections	Σ	3,014,720	248,829	3,263,549	72.3%	3,030,889	1,324,683	108,149	4,463,721	61.0%		
(f) interleaved reverts		-395,277	-18,706	-413,983	-9.2%	-207,878	-351,031	-14,769	-573,678	-7.8%		
(g) reverts reverting more than one editor		-110,084	-13,471	-123,555	-2.7%	-306,117	-282,606	-28,029	-616,752	-8.4%		
Results after filtering ambiguous reverts	Σ	2,509,359	216,652	2,726,011	60.4%	2,516,894	691,046	65,351	3,273,291	44.7%		
(h1) reverts reverting registered editors or bots		-586,505	-20,304	-606,809	-13.4%	0	-691,046	-65,351	-756,397	-10.3%		
(h2) reverts reverting editors with IPv6 addresses		-37,310	-2,044	-39,354	-0.9%	-44,922	0	0	-44,922	-0.6%		
Results after all filtering steps	Σ	1,885,544	194,304	2,079,848	46.1%	2,471,972	0	0	2,471,972	33.8%		

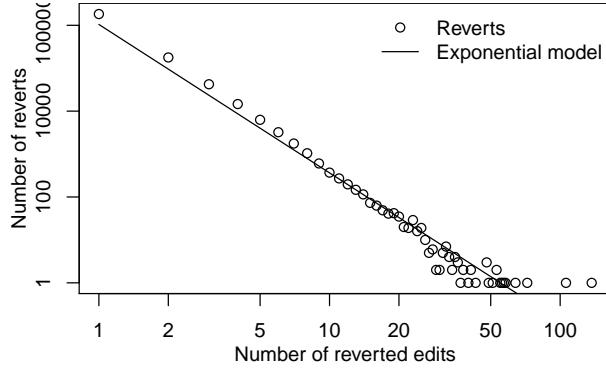


Figure 9: Number of full page reverts with a specific number of edits they revert and fitted exponential model in a log-log plot.

French Wikipedia: Geolocating Editors

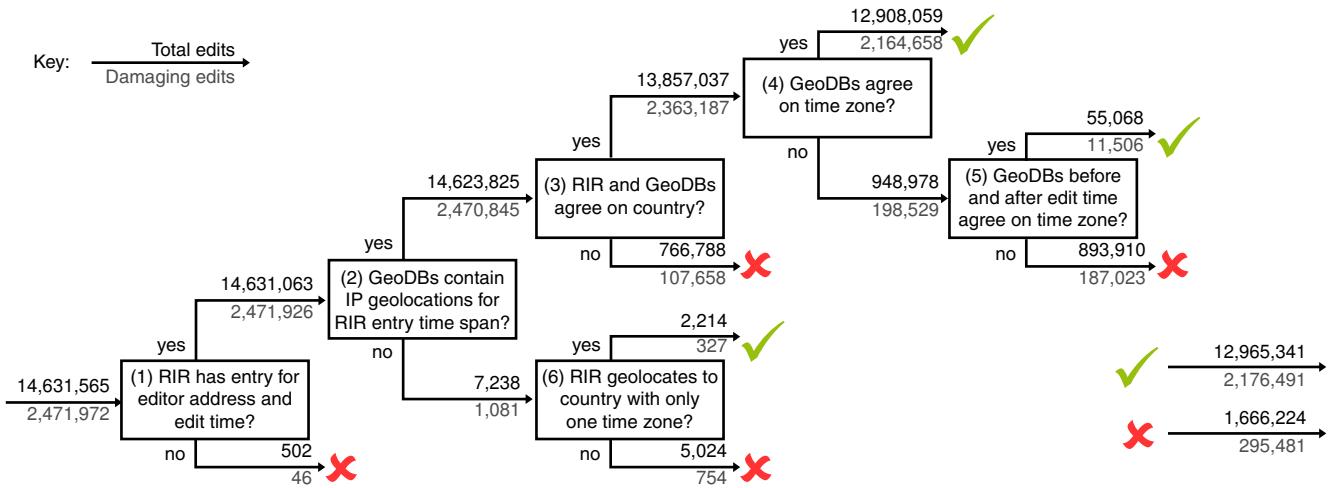


Figure 10: Decision tree to decide whether to trust the available geolocation information for an edit (✓), or not (✗). The numbers denote the total edits and reverted edits for the French Wikipedia that went through each branch.

Table 12: Historic geolocation success for all anonymous editors of the French 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 of the paper are given, and their corresponding IP addresses. Numbers are given for each exit node of the decision tree in the Figure above, 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
<i>Entire Wikipedia</i>		2,471,972 (16%)	14,631,565	1,176,922	4,377,718
No ✗	Step (1)	46 (9%)	502	33	146
	Step (3)	107,658 (14%)	766,788	51,044	205,467
	Step (5)	187,023 (20%)	893,910	83,286	272,910
	Step (6)	754 (15%)	5,024	386	1,878
	Σ	295,481 (17%)	1,666,224	134,748	480,385
Yes ✓	Step (4)	2,164,658 (16%)	12,908,059	1,036,547	3,878,089
	Step (5)	11,506 (20%)	55,068	5,774	20,064
	Step (6)	327 (14%)	2,214	169	768
	Σ	2,176,491 (16%)	12,965,341	1,042,471	3,898,856

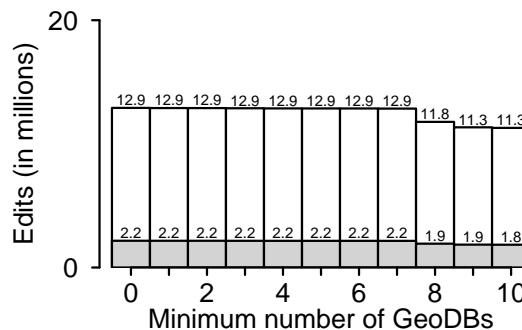


Figure 11: Number of total edits (white bars) and vandalism edits (gray bars) in millions from the yes-branch of Step (4) above over the number of GeoDBs considered.

French Wikipedia: Spatio-Temporal Analysis

Table 13: Number of edits by country, sorted for the French Wikipedia.

Country	Wikipedia						
	enwiki	dewiki	frwiki	eswiki	ruwiki	itwiki	jawiki
France	1,070,256	109,914	10,422,471	92,355	27,640	68,339	21,491
Belgium	447,600	39,196	694,475	17,849	7,362	14,630	3,017
Switzerland	311,700	855,089	381,326	19,294	8,081	94,654	2,785

Table 14: Number of vandalism edits by country, sorted for the French Wikipedia.

Country	Wikipedia						
	enwiki	dewiki	frwiki	eswiki	ruwiki	itwiki	jawiki
France	135,392	17,278	1,774,797	12,193	3,162	7,214	994
Belgium	54,475	6,981	98,572	1,708	1,423	1,429	114
Switzerland	37,260	185,746	45,872	2,002	991	13,114	141

Table 15: Number of vandalism commented edits by country, sorted for the French Wikipedia.

Country	Wikipedia						
	enwiki	dewiki	frwiki	eswiki	ruwiki	itwiki	jawiki
France	18,109	597	172,680	554	12	131	26
Belgium	8,711	240	9,346	89		24	8
Switzerland	5,577	7,484	3,964	78	7	361	1

The tables above show the number of certain kinds of edits from specific countries in all analyzed variants of Wikipedia. Countries are selected to have at least 100,000 vandalism edits in the French Wikipedia, or that have French as a major language (according to the English Wikipedia) and at least 1,000 vandalism edits. Countries are sorted by the number of all edits in the French Wikipedia. The same countries are used in the hour-of-day plots.

Vandalism commented edits refers to edits that are—after filtering reverts—affected by a revert with a comment that signals that it is undoing vandalism (see Kittur et al. (2007b)). Note that this approach uses specific words to classify comments and is thus likely less effective in finding vandalism for other languages than English.

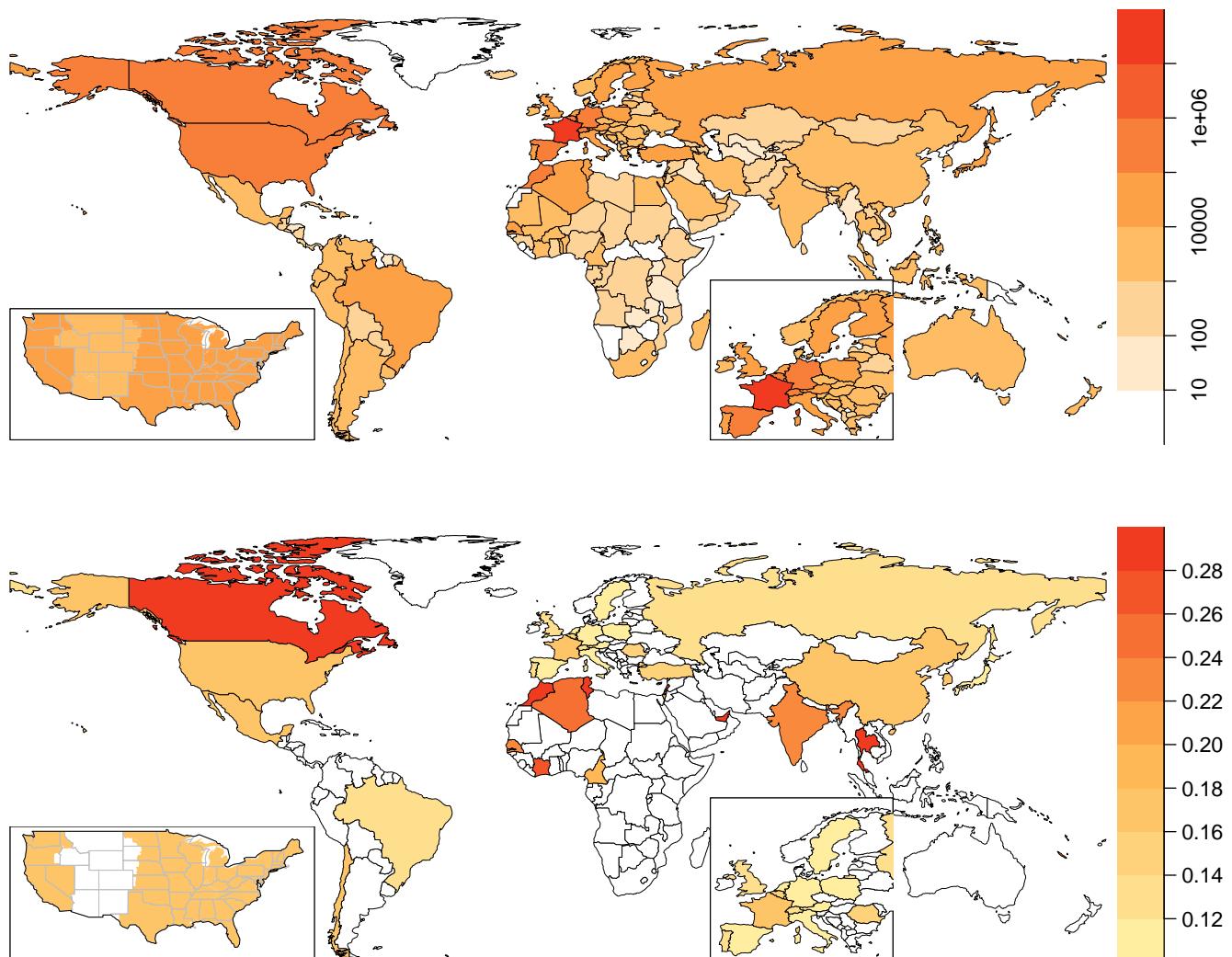
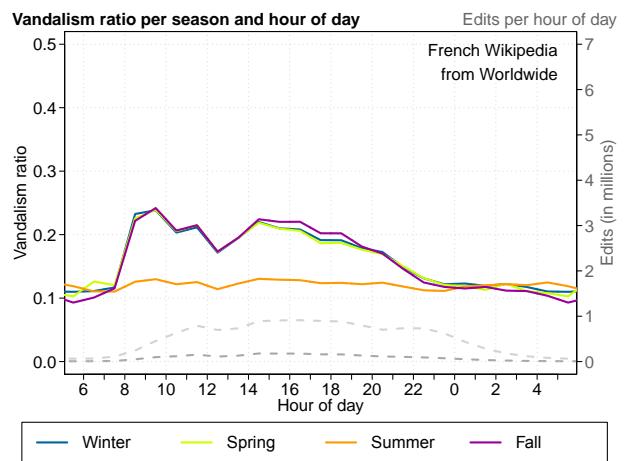
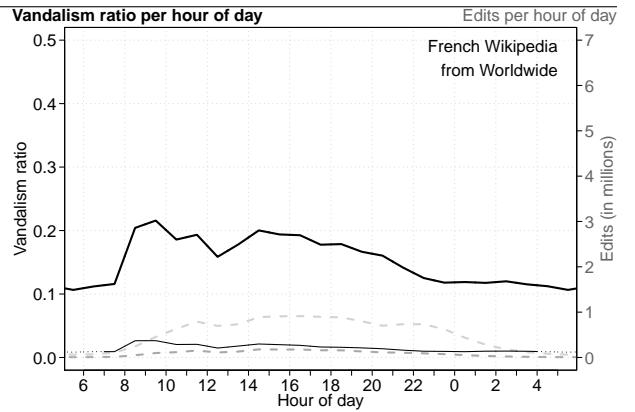
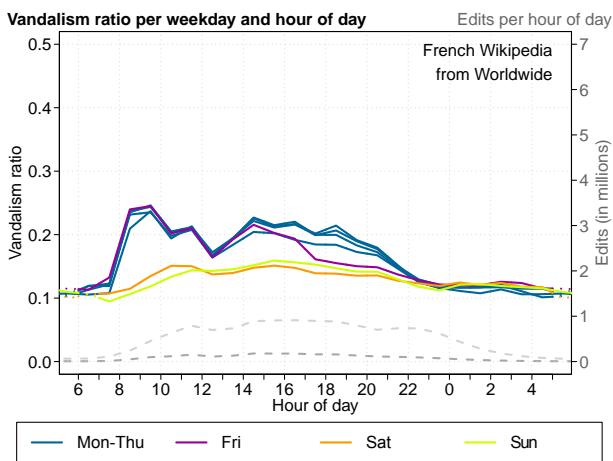


Figure 12: Number of edits (top) and ratio of vandalism to all edits (bottom) in the French Wikipedia by country. Countries with less than 1,000 vandalism edits are not colored. The embedded small maps show (left) the vandalism ratio in the United States (without Alaska) by major time zone (from West to East: Pacific, Mountain, Central, and Eastern) with overlaid state borders and (right) Europe enlarged.

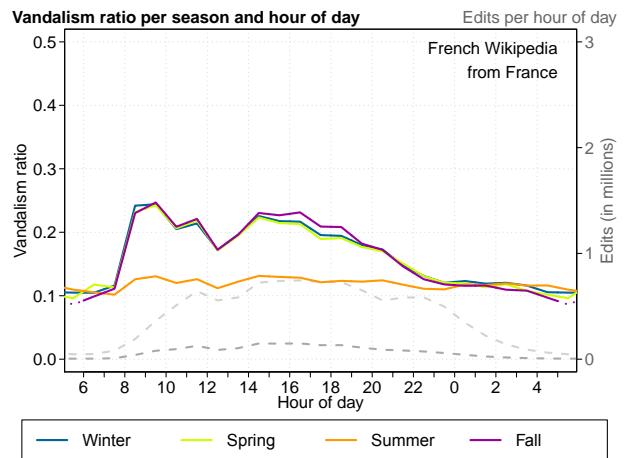
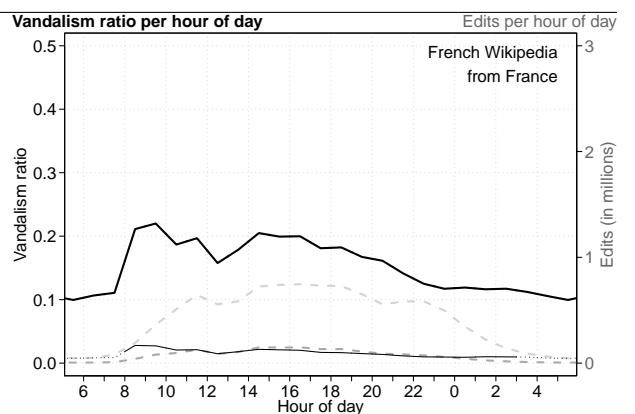
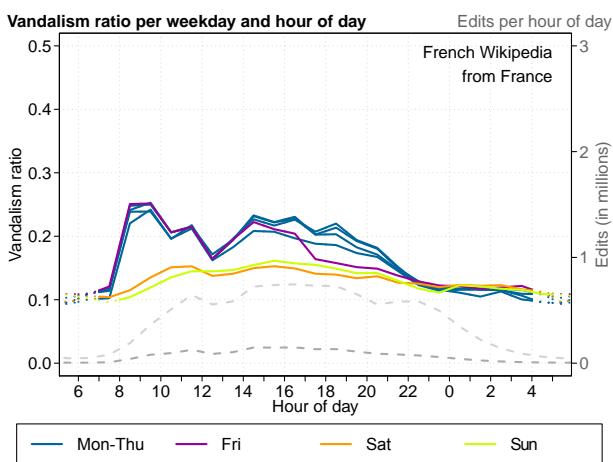
French Wikipedia

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 (light gray) and vandalism edits (dark gray) per hour of day (right axis, dashed lines), both averaged over the French Wikipedia's history. Ratios estimated from less than 1,000 vandalism edits are displayed with dotted lines.

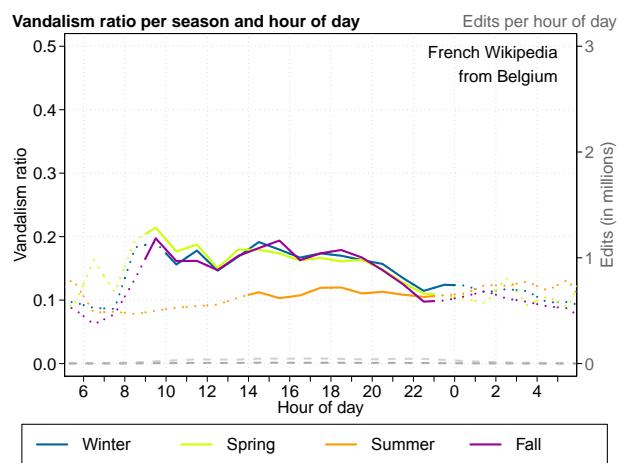
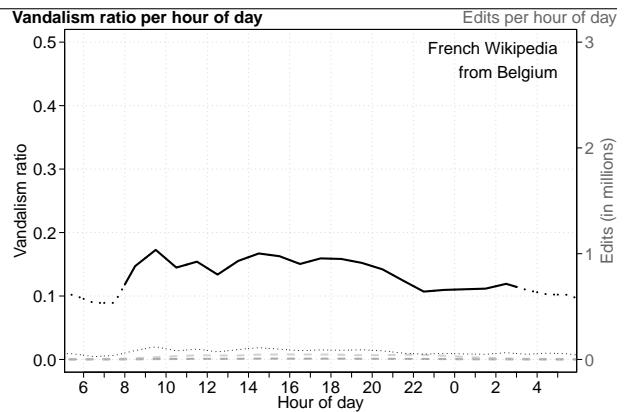
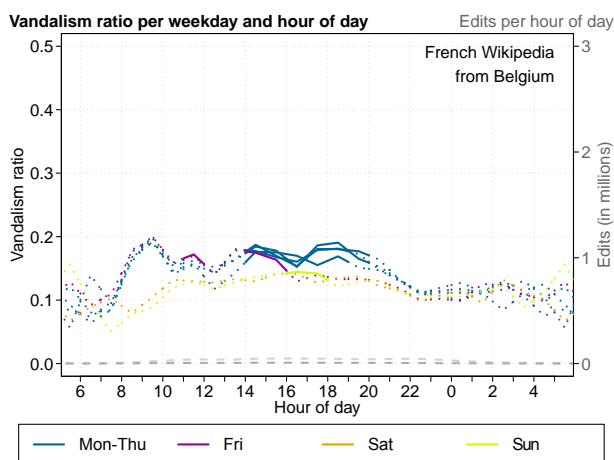
The thin black line to the right is the vandalism ratio when detecting vandalism by comments of reverts (see Kittur et al. (2007b)). Note that this approach uses specific words to classify comments and is thus likely less effective in finding vandalism for other languages than English.



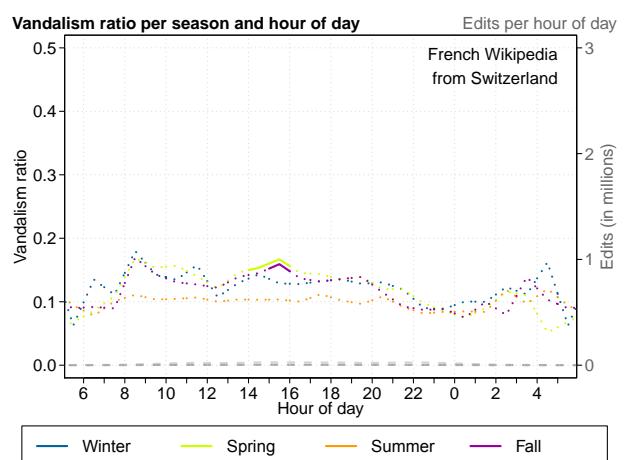
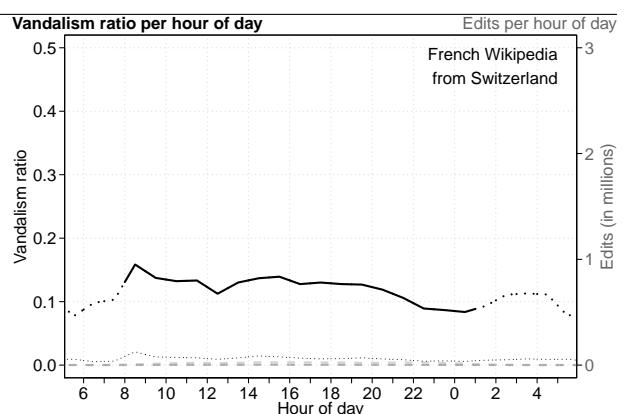
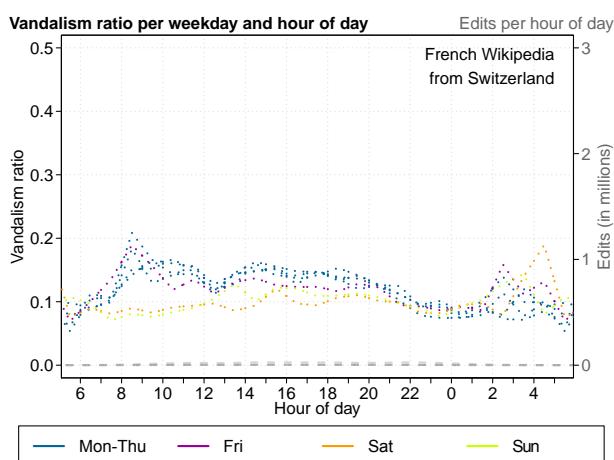
French Wikipedia Edits from France



French Wikipedia Edits from Belgium



French Wikipedia Edits from Switzerland



Spanish Wikipedia

Spanish Wikipedia: Mining Vandalism

Table 16: Step-by-step filtering of the Spanish Wikipedia as per the revert patterns depicted in Figure 2 in the paper. 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. Note that the approach by Kittur et al. uses specific words to classify comments and is thus likely less effective in finding vandalism for other languages than English.

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	6,323,125	396,573	6,719,698	100.0%	8,854,615	3,747,783	331,541	12,933,939	100.0%	
(a) reverts to page blank	-16,834	-35	-16,869	-0.3%	-1,036,340	-1,093,114	-129,283	-2,258,737	-17.5%	
(b) empty reverts due to renaming/removal/error	-348,508	-17,849	-366,357	-5.5%	0	0	0	0	0.0%	
Results after filtering pseudo-reverts	Σ	5,957,783	378,689	6,336,472	94.3%	7,818,275	2,654,669	202,258	10,675,202	82.5%
(c) self reverts	-544,444	-365	-544,809	-8.1%	-398,648	-257,268	-26,212	-682,128	-5.3%	
(d) revert corrections	-91,318	-5,506	-96,824	-1.4%	-188,612	-200,847	-3,336	-392,795	-3.0%	
(e) reverted reverts	-35,647	-561	-36,208	-0.5%	-151,345	-182,387	-10,214	-343,946	-2.7%	
Results after filtering error-corrections	Σ	5,286,374	372,257	5,658,631	84.2%	7,079,670	2,014,167	162,496	9,256,333	71.6%
(f) interleaved reverts	-530,960	-18,431	-549,391	-8.2%	-552,743	-550,067	-32,962	-1,135,772	-8.8%	
(g) reverts reverting more than one editor	-315,978	-21,118	-337,096	-5.0%	-1,249,817	-586,773	-67,561	-1,904,151	-14.7%	
Results after filtering ambiguous reverts	Σ	4,439,436	332,708	4,772,144	71.0%	5,277,110	877,327	61,973	6,216,410	48.1%
(h1) reverts reverting registered editors or bots	-646,503	-18,029	-664,532	-9.9%	0	-877,327	-61,973	-939,300	-7.3%	
(h2) reverts reverting editors with IPv6 addresses	-12,111	-116	-12,227	-0.2%	-15,098	0	0	-15,098	-0.1%	
Results after all filtering steps	Σ	3,780,822	314,563	4,095,385	60.9%	5,262,012	0	0	5,262,012	40.7%

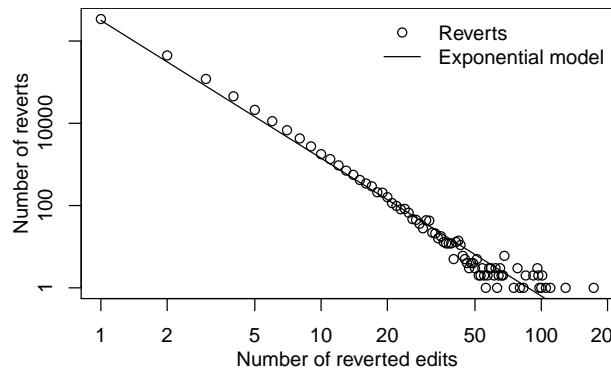


Figure 13: Number of full page reverts with a specific number of edits they revert and fitted exponential model in a log-log plot.

Spanish Wikipedia: Geolocating Editors

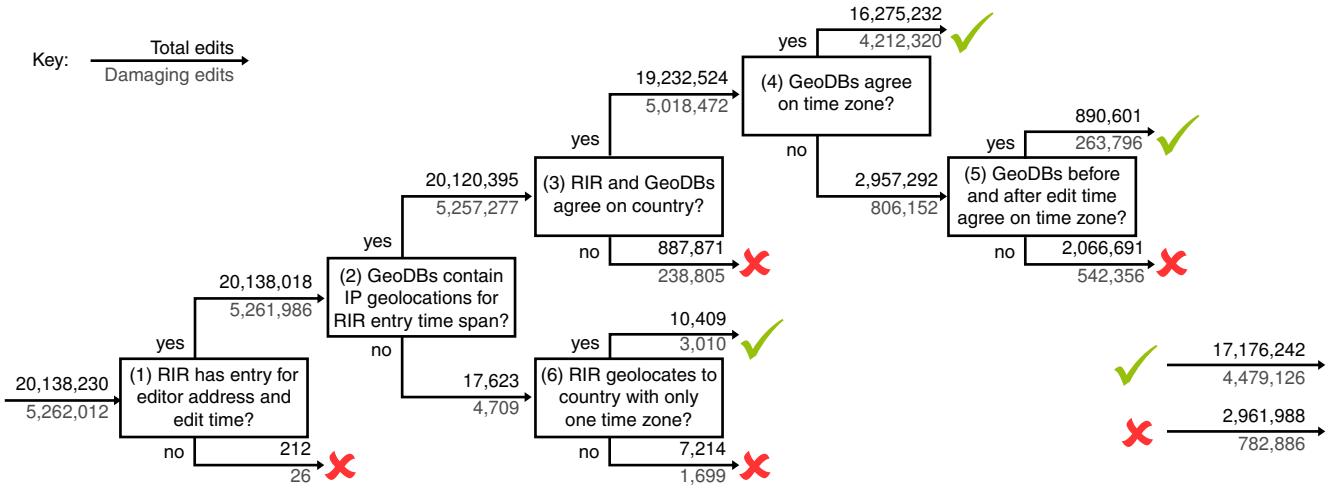


Figure 14: Decision tree to decide whether to trust the available geolocation information for an edit (✓), or not (✗). The numbers denote the total edits and reverted edits for the Spanish Wikipedia that went through each branch.

Table 17: Historic geolocation success for all anonymous editors of the Spanish 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 of the paper are given, and their corresponding IP addresses. Numbers are given for each exit node of the decision tree in the Figure above, 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
<i>Entire Wikipedia</i>		5,262,012 (26%)	20,138,230	2,621,931	6,700,381
No ✗	Step (1)	26 (12%)	212	6	39
	Step (3)	238,805 (26%)	887,871	116,146	291,055
	Step (5)	542,356 (26%)	2,066,691	295,372	750,067
	Step (6)	1,699 (23%)	7,214	1,043	3,022
	Σ	782,886 (26%)	2,961,988	412,547	1,044,085
Yes ✓	Step (4)	4,212,320 (25%)	16,275,232	2,068,602	5,331,762
	Step (5)	263,796 (29%)	890,601	144,625	346,410
	Step (6)	3,010 (28%)	10,409	1,608	3,667
	Σ	4,479,126 (26%)	17,176,242	2,214,786	5,681,643

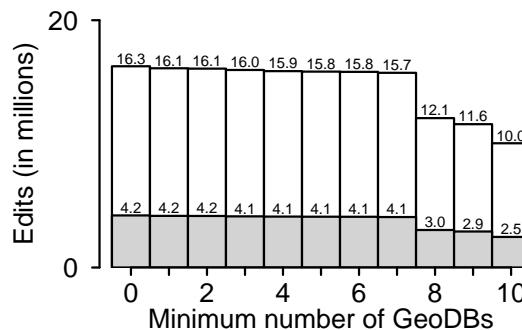


Figure 15: Number of total edits (white bars) and vandalism edits (gray bars) in millions from the yes-branch of Step (4) above over the number of GeoDBs considered.

Spanish Wikipedia: Spatio-Temporal Analysis

Table 18: Number of edits by country, sorted for the Spanish Wikipedia.

Country	Wikipedia						
	enwiki	dewiki	frwiki	eswiki	ruwiki	itwiki	jawiki
Spain	945,915	62,094	116,834	5,671,664	15,325	61,848	7,191
Mexico	443,953	7,339	8,484	2,039,568	1,607	4,408	1,768
Chile	286,473	7,657	6,625	1,939,500	863	5,344	5,025
Colombia	208,198	4,850	5,851	1,634,283	990	4,773	973
Argentina	227,718	4,617	5,729	1,548,639	1,006	5,689	1,311
Peru	163,426	3,825	5,058	1,001,602	444	8,099	981
Venezuela	123,285	2,731	2,551	758,692	431	2,983	813

Table 19: Number of vandalism edits by country, sorted for the Spanish Wikipedia.

Country	Wikipedia						
	enwiki	dewiki	frwiki	eswiki	ruwiki	itwiki	jawiki
Spain	135,325	11,538	11,524	1,279,946	1,748	5,756	385
Mexico	79,093	1,657	1,368	666,668	166	637	153
Colombia	42,516	1,113	828	544,954	92	745	266
Chile	38,930	1,697	1,134	537,082	109	814	214
Argentina	37,405	1,000	825	386,785	178	1,094	127
Peru	30,323	906	687	298,111	78	636	103
Venezuela	19,612	455	358	187,914	27	328	322

Table 20: Number of vandalism commented edits by country, sorted for the Spanish Wikipedia.

Country	Wikipedia						
	enwiki	dewiki	frwiki	eswiki	ruwiki	itwiki	jawiki
Spain	22,433	352	951	102,690	8	123	41
Mexico	13,745	89	146	52,273		18	4
Chile	5,955	82	127	46,362	1	15	5
Colombia	7,902	51	86	39,540	3	16	
Argentina	5,564	54	33	29,939		28	1
Peru	5,279	36	48	21,805	2	31	1
Venezuela	2,852	21	69	12,628		7	4

The tables above show the number of certain kinds of edits from specific countries in all analyzed variants of Wikipedia. Countries are selected to have at least 100,000 vandalism edits in the Spanish Wikipedia, or that have Spanish as a major language (according to the English Wikipedia) and at least 1,000 vandalism edits. Countries are sorted by the number of all edits in the Spanish Wikipedia. The same countries are used in the hour-of-day plots.

Vandalism commented edits refers to edits that are—after filtering reverts—affected by a revert with a comment that signals that it is undoing vandalism (see Kittur et al. (2007b)). Note that this approach uses specific words to classify comments and is thus likely less effective in finding vandalism for other languages than English.

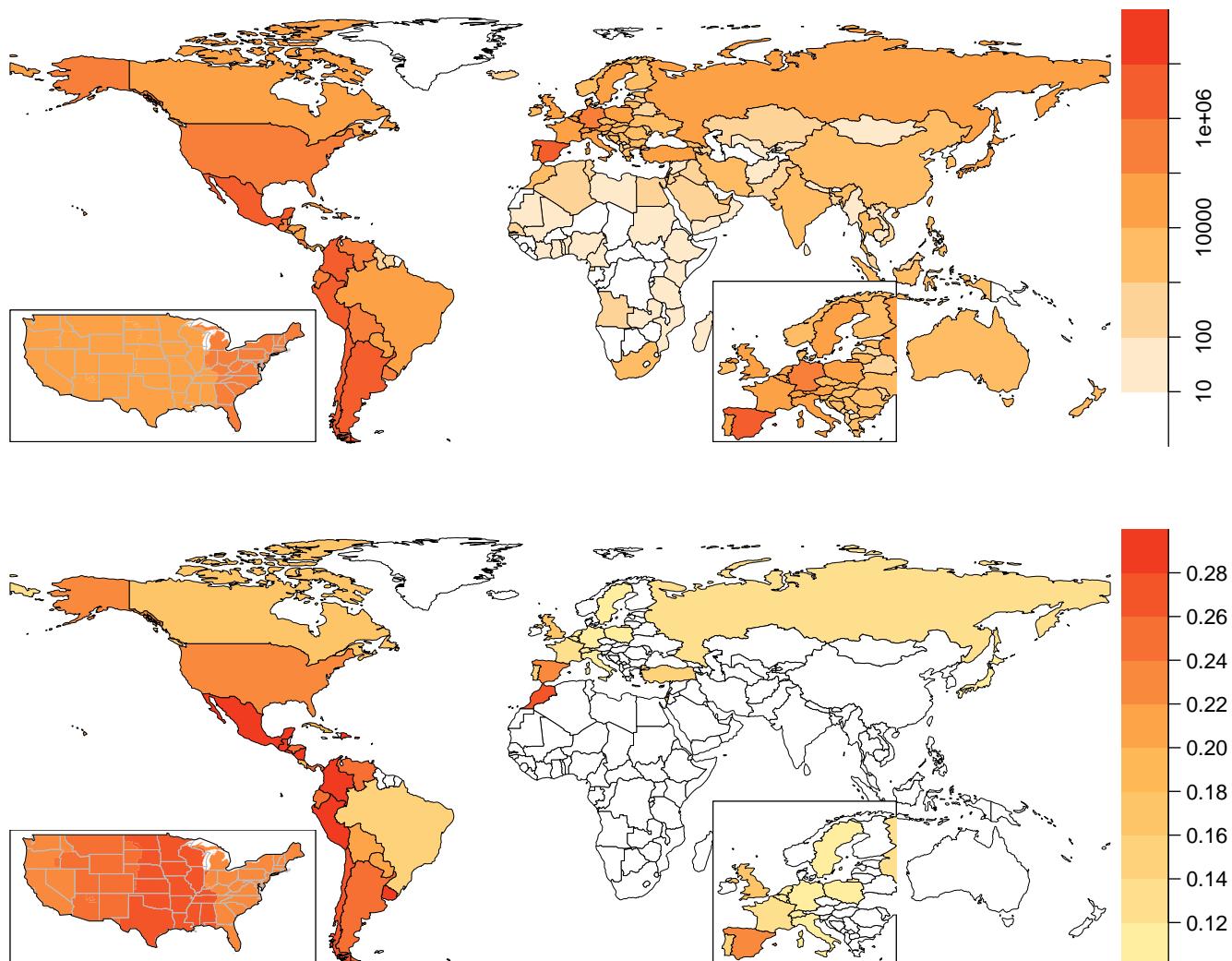
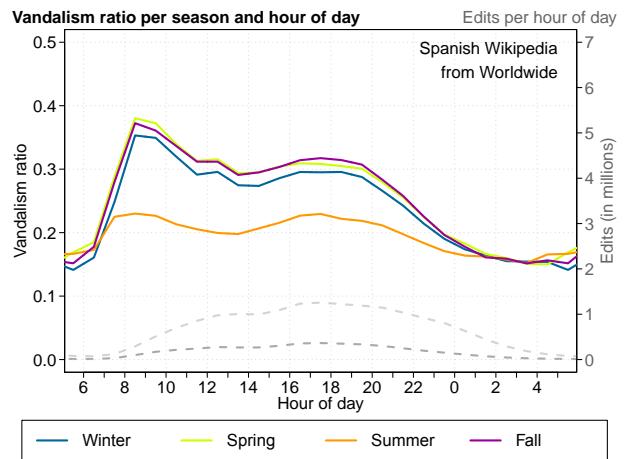
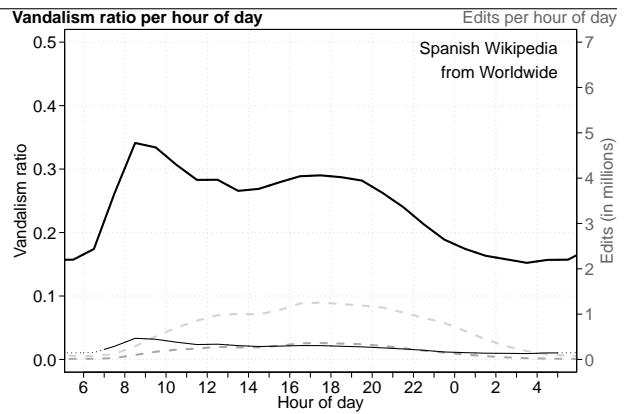
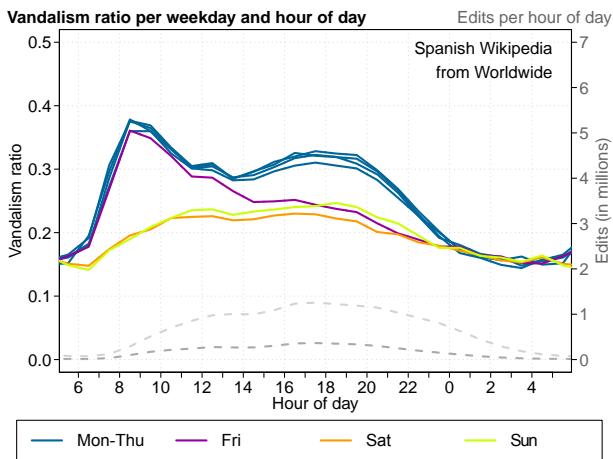


Figure 16: Number of edits (top) and ratio of vandalism to all edits (bottom) in the Spanish Wikipedia by country. Countries with less than 1,000 vandalism edits are not colored. The embedded small maps show (left) the vandalism ratio in the United States (without Alaska) by major time zone (from West to East: Pacific, Mountain, Central, and Eastern) with overlaid state borders and (right) Europe enlarged.

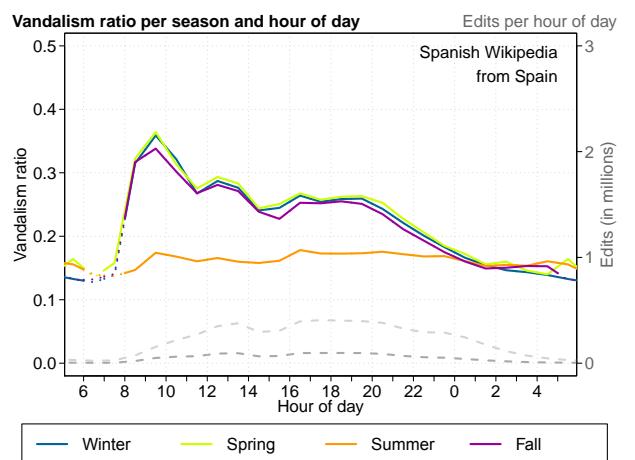
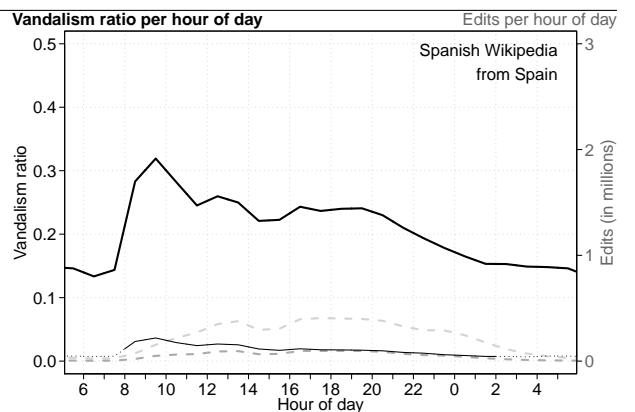
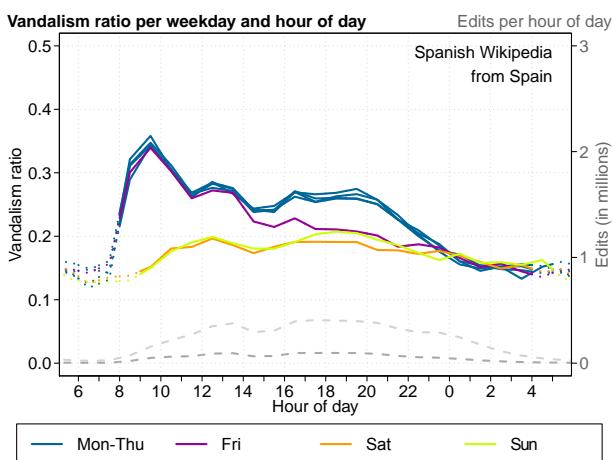
Spanish Wikipedia

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 (light gray) and vandalism edits (dark gray) per hour of day (right axis, dashed lines), both averaged over the Spanish Wikipedia's history. Ratios estimated from less than 1,000 vandalism edits are displayed with dotted lines.

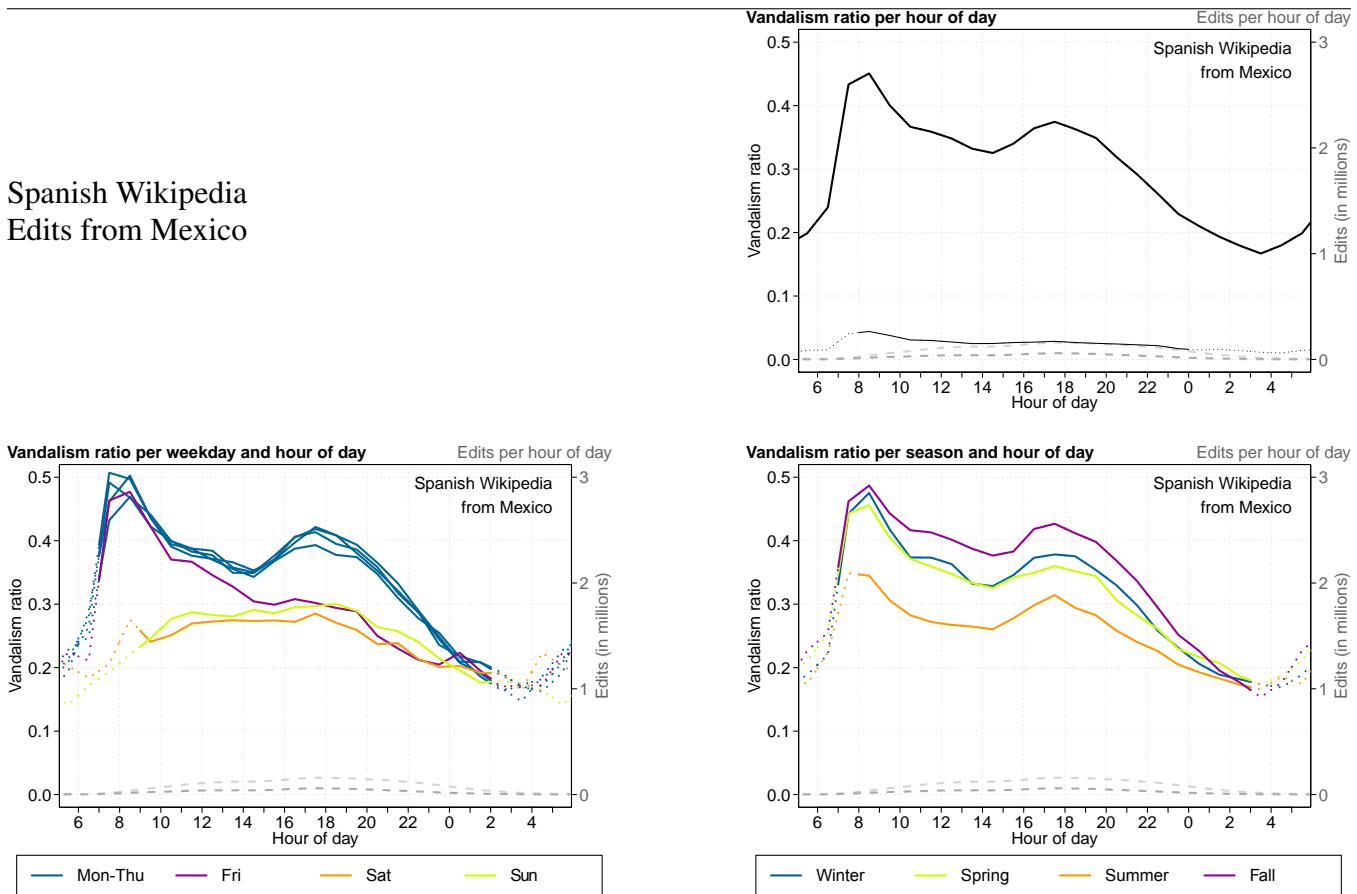
The thin black line to the right is the vandalism ratio when detecting vandalism by comments of reverts (see Kittur et al. (2007b)). Note that this approach uses specific words to classify comments and is thus likely less effective in finding vandalism for other languages than English.



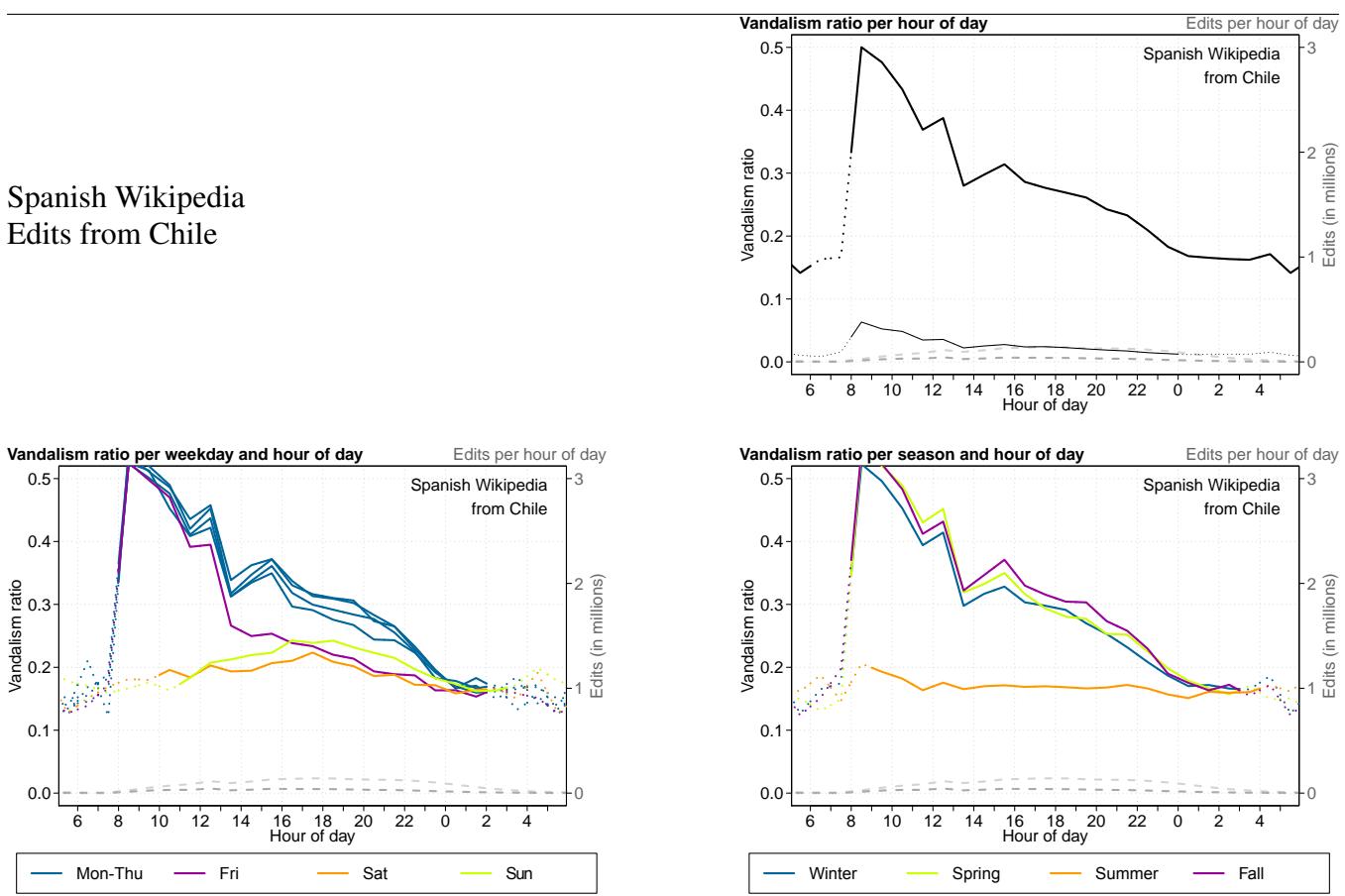
Spanish Wikipedia Edits from Spain



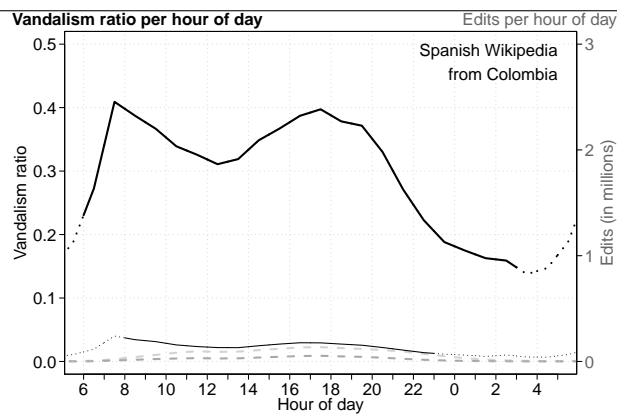
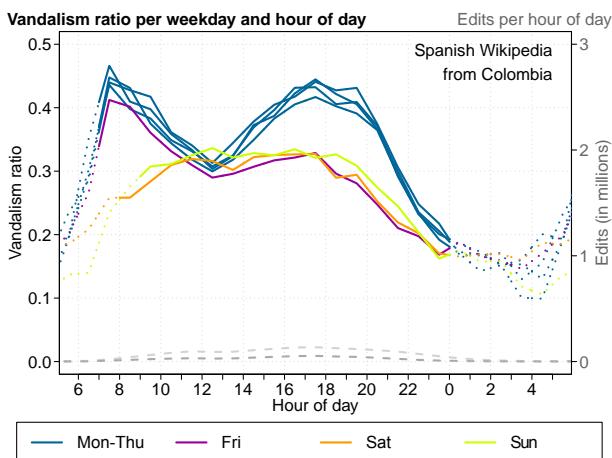
Spanish Wikipedia Edits from Mexico



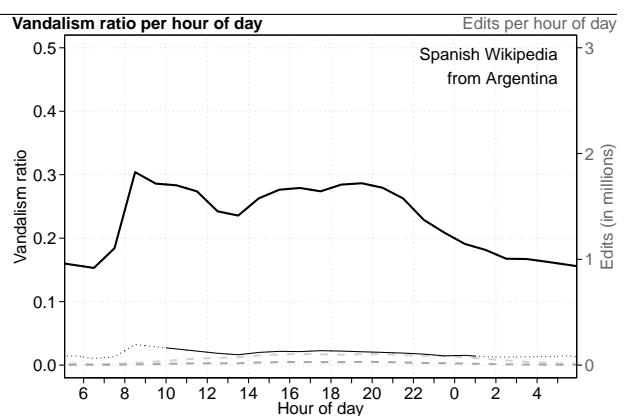
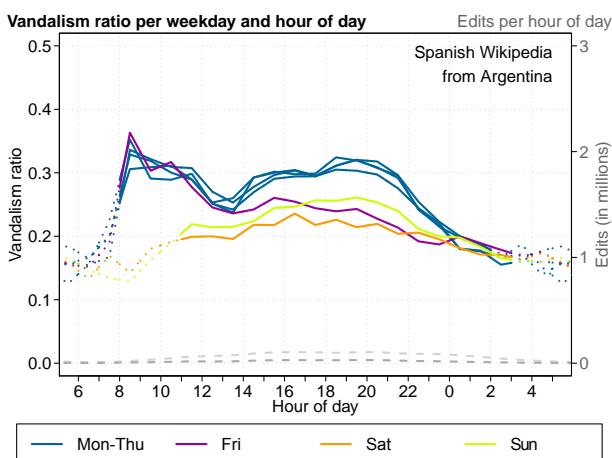
Spanish Wikipedia Edits from Chile



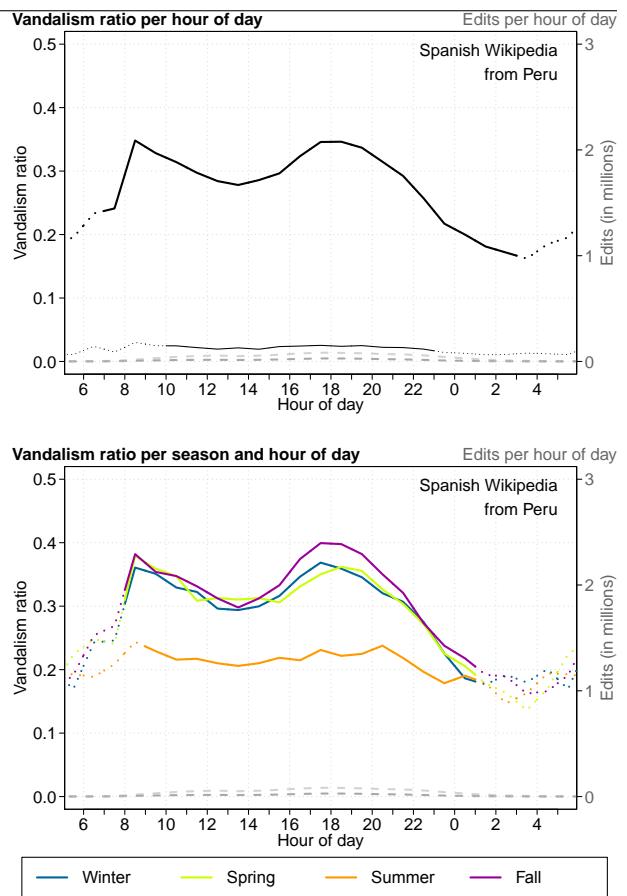
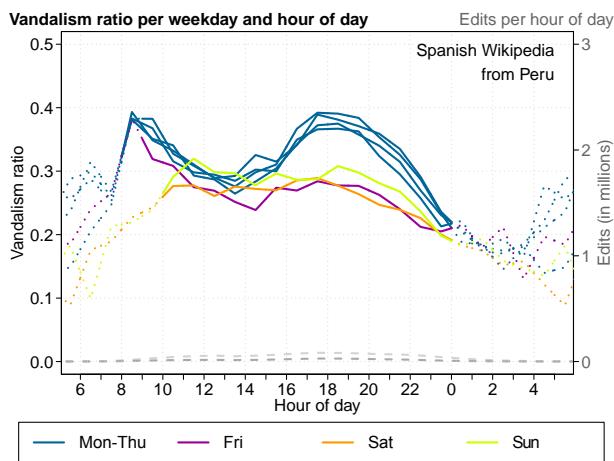
Spanish Wikipedia Edits from Colombia



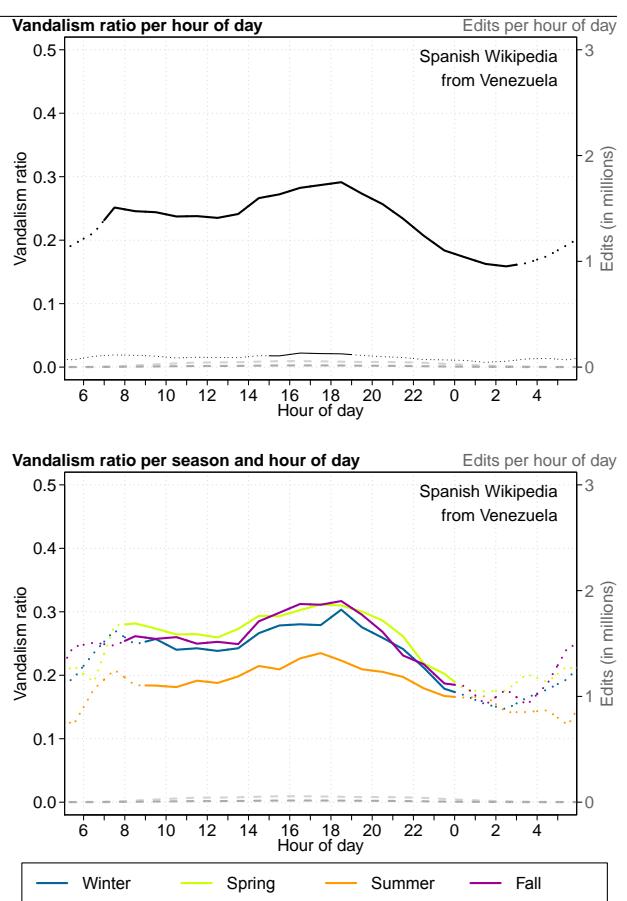
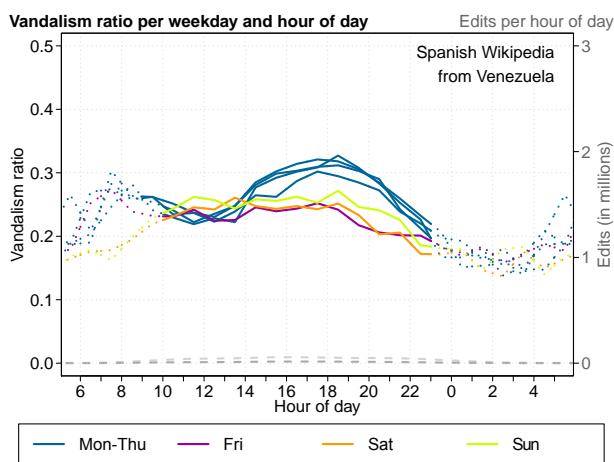
Spanish Wikipedia Edits from Argentina



Spanish Wikipedia Edits from Peru



Spanish Wikipedia Edits from Venezuela



Russian Wikipedia

Russian Wikipedia: Mining Vandalism

Table 21: Step-by-step filtering of the Russian Wikipedia as per the revert patterns depicted in Figure 2 in the paper. 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. Note that the approach by Kittur et al. uses specific words to classify comments and is thus likely less effective in finding vandalism for other languages than English.

Revert filtering step	Full page reverts				Reverted edits				Total			
	Vandalism as per Kittur		Total		Editor							
	No	Yes	Absolute	Relative	Anonymous	Registered	Bot					
Results of naive SHA-1 matching	3,218,635	4,524	3,223,159	100.0%	2,899,958	1,807,745	114,897	4,822,600	100.0%			
(a) reverts to page blank	-6,869	-2	-6,871	-0.2%	-100,644	-180,620	-11,858	-293,122	-6.1%			
(b) empty reverts due to renaming/removal/error	-484,941	-194	-485,135	-15.1%	0	0	0	0	0.0%			
Results after filtering pseudo-reverts	Σ	2,726,825	4,328	2,731,153	84.7%	2,799,314	1,627,125	103,039	4,529,478	93.9%		
(c) self reverts	-456,538	-9	-456,547	-14.2%	-205,333	-400,244	-20,040	-625,617	-13.0%			
(d) revert corrections	-41,816	-261	-42,077	-1.3%	-75,023	-94,861	-1,234	-171,118	-3.5%			
(e) reverted reverts	-17,977	-26	-18,003	-0.6%	-56,632	-88,314	-3,403	-148,349	-3.1%			
Results after filtering error-corrections	Σ	2,210,494	4,032	2,214,526	68.7%	2,462,326	1,043,706	78,362	3,584,394	74.3%		
(f) interleaved reverts	-321,283	-568	-321,851	-10.0%	-197,186	-285,303	-11,643	-494,132	-10.2%			
(g) reverts reverting more than one editor	-140,678	-1,024	-141,702	-4.4%	-482,877	-248,722	-24,752	-756,351	-15.7%			
Results after filtering ambiguous reverts	Σ	1,748,533	2,440	1,750,973	54.3%	1,782,263	509,681	41,967	2,333,911	48.4%		
(h1) reverts reverting registered editors or bots	-417,368	-280	-417,648	-13.0%	0	-509,681	-41,967	-551,648	-11.4%			
(h2) reverts reverting editors with IPv6 addresses	-4,220	-4	-4,224	-0.1%	-5,575	0	0	-5,575	-0.1%			
Results after all filtering steps	Σ	1,326,945	2,156	1,329,101	41.2%	1,776,688	0	0	1,776,688	36.8%		

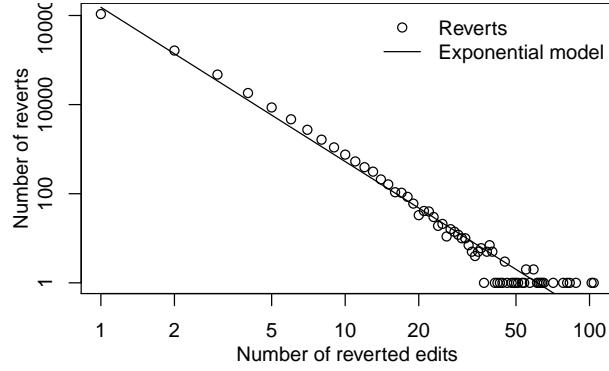


Figure 17: Number of full page reverts with a specific number of edits they revert and fitted exponential model in a log-log plot.

Russian Wikipedia: Geolocating Editors

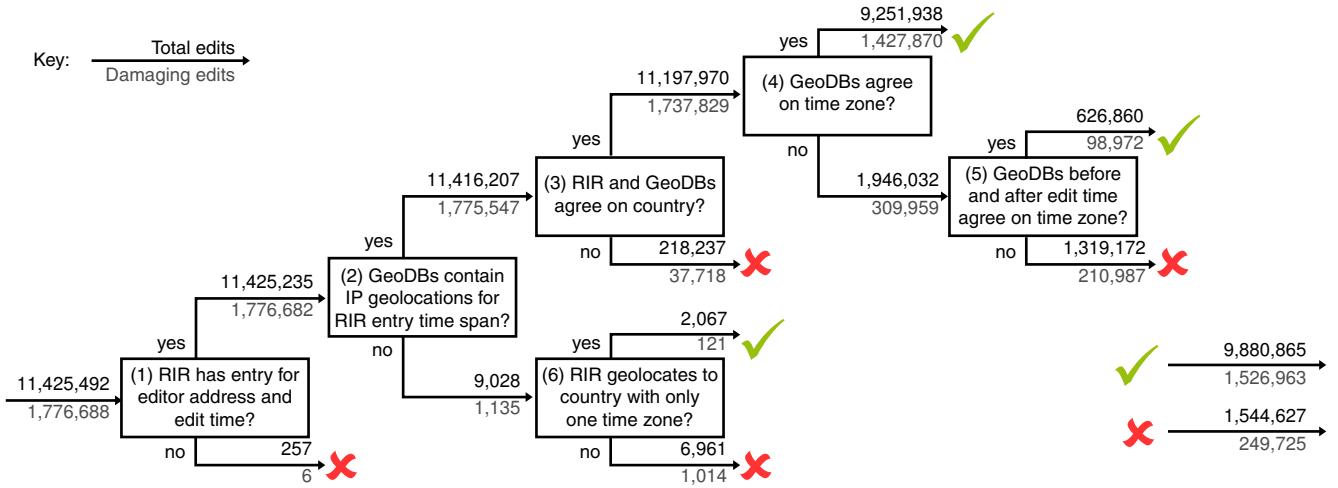


Figure 18: Decision tree to decide whether to trust the available geolocation information for an edit (✓), or not (✗). The numbers denote the total edits and reverted edits for the Russian Wikipedia that went through each branch.

Table 22: Historic geolocation success for all anonymous editors of the Russian 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 of the paper are given, and their corresponding IP addresses. Numbers are given for each exit node of the decision tree in the Figure above, 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
<i>Entire Wikipedia</i>		1,776,688 (15%)	11,425,492	847,619	3,173,526
No ✗	Step (1)	6 (2%)	257	2	41
	Step (3)	37,718 (17%)	218,237	18,138	63,120
	Step (5)	210,987 (15%)	1,319,172	102,151	375,183
	Step (6)	1,014 (14%)	6,961	480	2,284
	Σ	249,725 (16%)	1,544,627	120,768	440,598
Yes ✓	Step (4)	1,427,870 (15%)	9,251,938	684,117	2,589,073
	Step (5)	98,972 (15%)	626,860	47,598	167,879
	Step (6)	121 (5%)	2,067	82	578
	Σ	1,526,963 (15%)	9,880,865	731,797	2,757,526

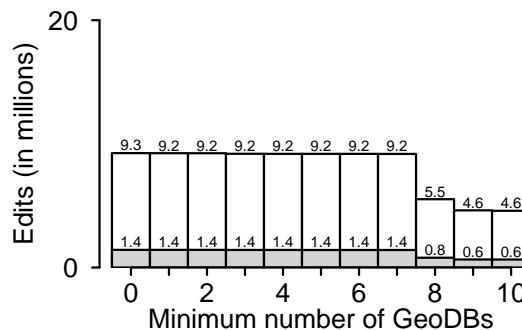


Figure 19: Number of total edits (white bars) and vandalism edits (gray bars) in millions from the yes-branch of Step (4) above over the number of GeoDBs considered.

Russian Wikipedia: Spatio-Temporal Analysis

Table 23: Number of edits by country, sorted for the Russian Wikipedia.

Country	Wikipedia						
	enwiki	dewiki	frwiki	eswiki	ruwiki	itwiki	jawiki
Russian Federation	458,615	28,706	16,342	12,588	6,967,485	50,117	5,497
Ukraine	173,826	6,816	3,752	3,006	1,201,596	3,510	1,455
Belarus	37,165	1,925	835	644	363,708	677	270

Table 24: Number of vandalism edits by country, sorted for the Russian Wikipedia.

Country	Wikipedia						
	enwiki	dewiki	frwiki	eswiki	ruwiki	itwiki	jawiki
Russian Federation	58,976	5,333	2,273	1,526	1,040,472	1,778	1,360
Ukraine	26,917	1,709	719	551	233,246	421	171
Belarus	4,453	401	106	85	56,986	79	18

Table 25: Number of vandalism commented edits by country, sorted for the Russian Wikipedia.

Country	Wikipedia						
	enwiki	dewiki	frwiki	eswiki	ruwiki	itwiki	jawiki
Russian Federation	6,553	97	140	26	1,801	10	104
Ukraine	2,685	43	25	8	405	3	9
Belarus	473	9	3	1	75	1	

The tables above show the number of certain kinds of edits from specific countries in all analyzed variants of Wikipedia. Countries are selected to have at least 100,000 vandalism edits in the Russian Wikipedia, or that have Russian as a major language (according to the English Wikipedia) and at least 1,000 vandalism edits. Countries are sorted by the number of all edits in the Russian Wikipedia. The same countries are used in the hour-of-day plots.

Vandalism commented edits refers to edits that are—after filtering reverts—affected by a revert with a comment that signals that it is undoing vandalism (see Kittur et al. (2007b)). Note that this approach uses specific words to classify comments and is thus likely less effective in finding vandalism for other languages than English.

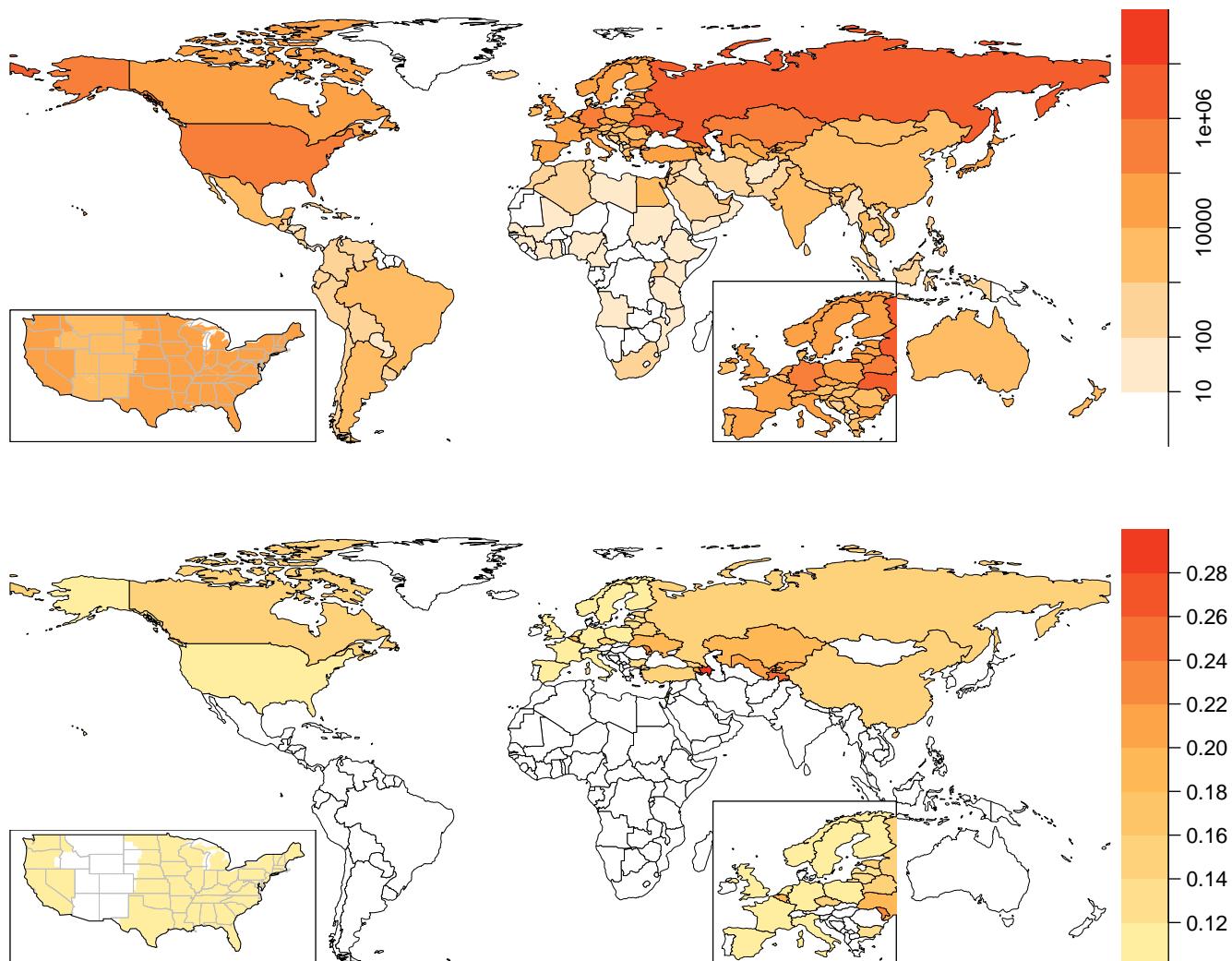
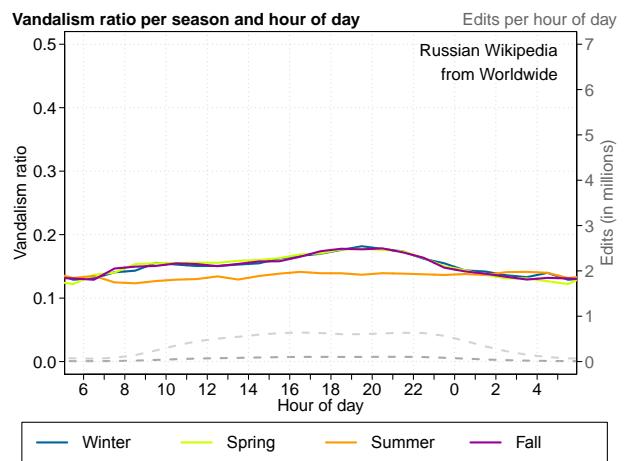
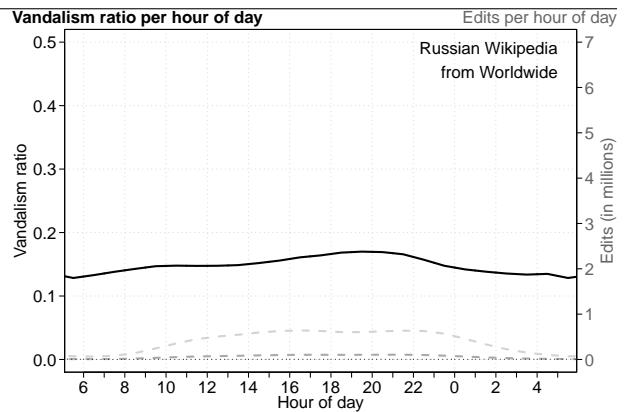
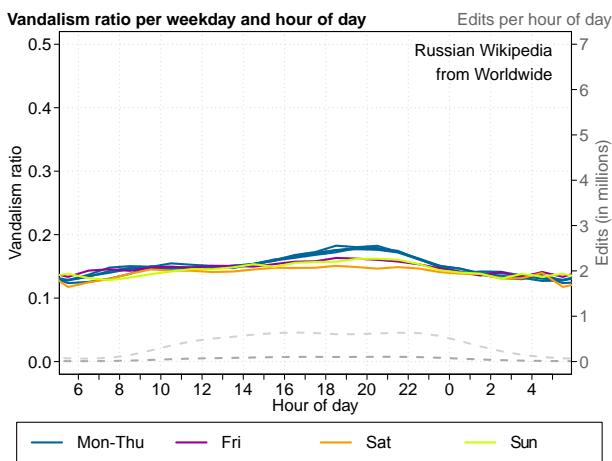


Figure 20: Number of edits (top) and ratio of vandalism to all edits (bottom) in the Russian Wikipedia by country. Countries with less than 1,000 vandalism edits are not colored. The embedded small maps show (left) the vandalism ratio in the United States (without Alaska) by major time zone (from West to East: Pacific, Mountain, Central, and Eastern) with overlaid state borders and (right) Europe enlarged.

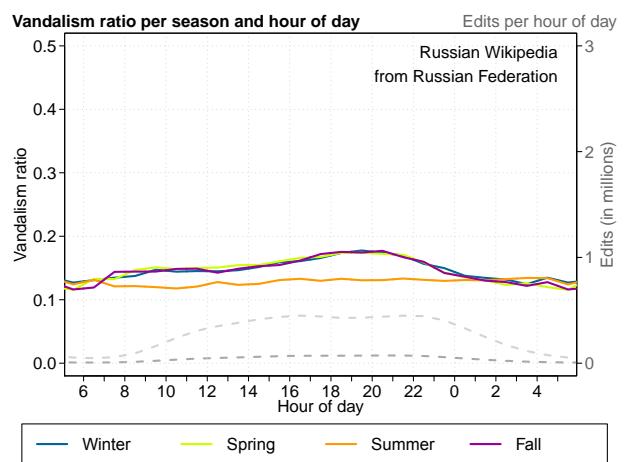
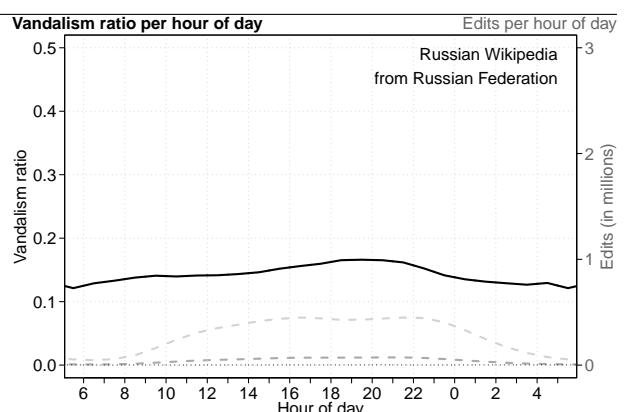
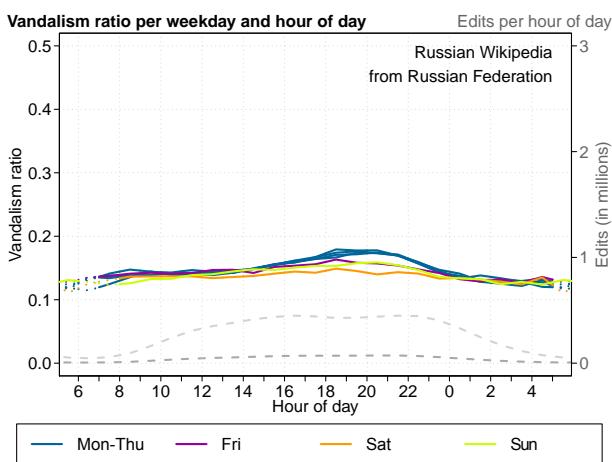
Russian Wikipedia

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 (light gray) and vandalism edits (dark gray) per hour of day (right axis, dashed lines), both averaged over the Russian Wikipedia's history. Ratios estimated from less than 1,000 vandalism edits are displayed with dotted lines.

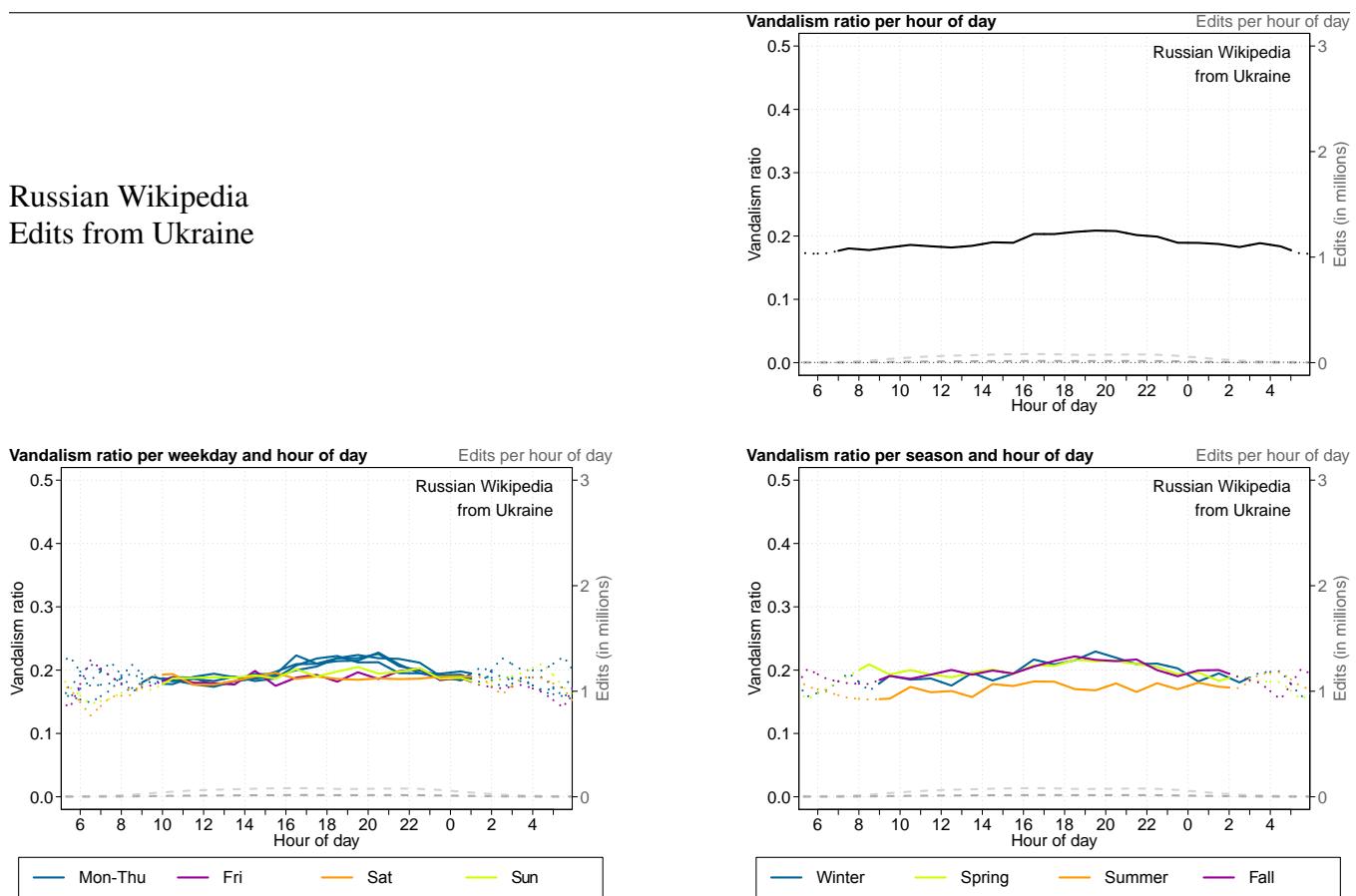
The thin black line to the right is the vandalism ratio when detecting vandalism by comments of reverts (see Kittur et al. (2007b)). Note that this approach uses specific words to classify comments and is thus likely less effective in finding vandalism for other languages than English.



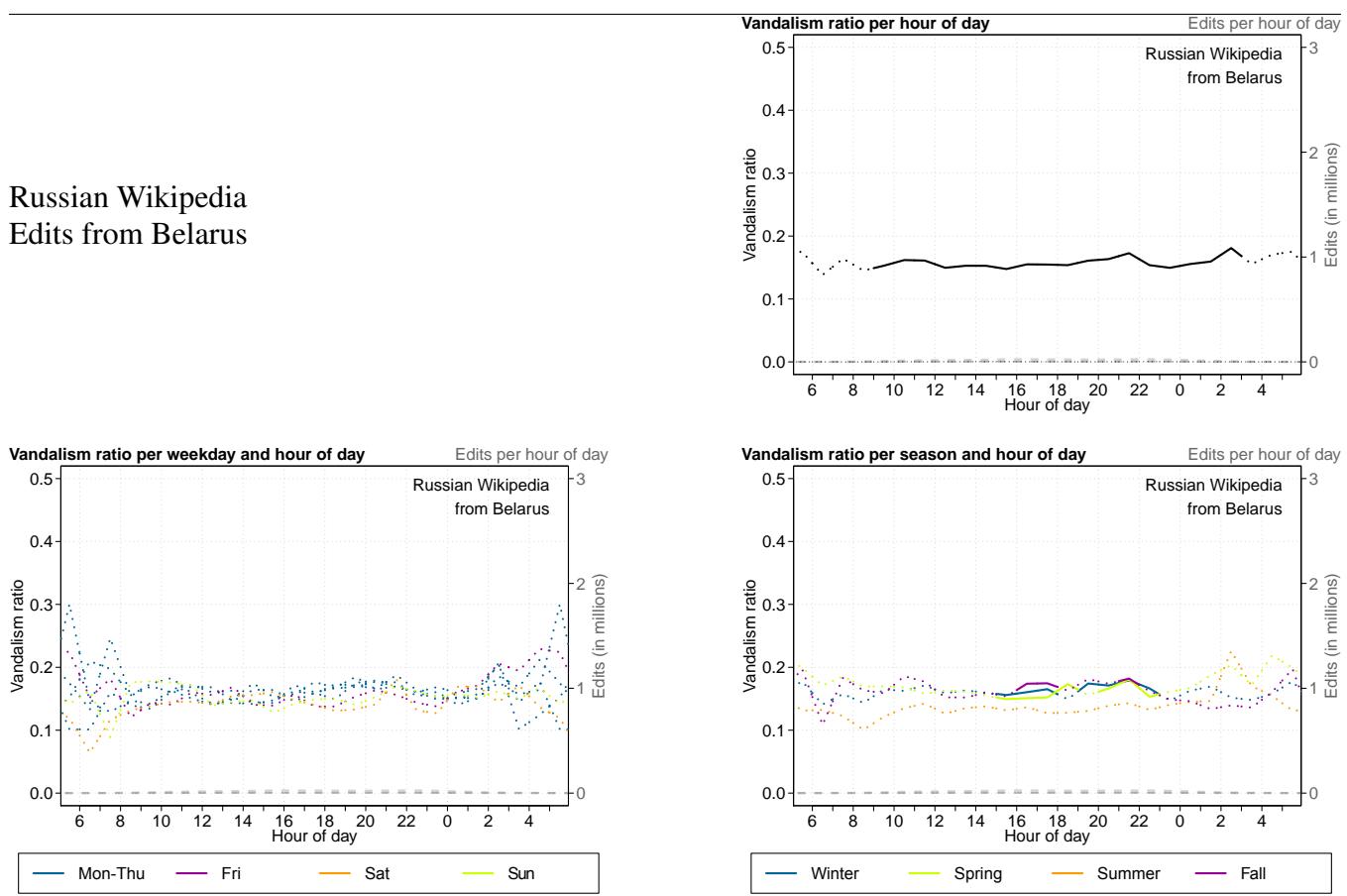
Russian Wikipedia Edits from Russian Federation



Russian Wikipedia Edits from Ukraine



Russian Wikipedia Edits from Belarus



Italian Wikipedia

Italian Wikipedia: Mining Vandalism

Table 26: Step-by-step filtering of the Italian Wikipedia as per the revert patterns depicted in Figure 2 in the paper. 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. Note that the approach by Kittur et al. uses specific words to classify comments and is thus likely less effective in finding vandalism for other languages than English.

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	2,967,626	47,110	3,014,736	100.0%	3,189,877	1,918,706	148,233	5,256,816	100.0%	
(a) reverts to page blank	-313,452	-345	-313,797	-10.4%	-692,604	-869,417	-60,857	-1,622,878	-30.9%	
(b) empty reverts due to renaming/removal/error	-316,149	-8,159	-324,308	-10.8%	0	0	0	0	0.0%	
Results after filtering pseudo-reverts	Σ	2,338,025	38,606	2,376,631	78.8%	2,497,273	1,049,289	87,376	3,633,938	69.1%
(c) self reverts		-308,962	-109	-309,071	-10.3%	-198,812	-178,619	-10,331	-387,762	-7.4%
(d) revert corrections		-29,662	-1,545	-31,207	-1.0%	-54,525	-54,712	-312	-109,549	-2.1%
(e) reverted reverts		-13,736	-211	-13,947	-0.5%	-47,471	-60,250	-2,572	-110,293	-2.1%
Results after filtering error-corrections	Σ	1,985,665	36,741	2,022,406	67.1%	2,196,465	755,708	74,161	3,026,334	57.6%
(f) interleaved reverts		-245,028	-3,587	-248,615	-8.2%	-171,506	-181,236	-8,016	-360,758	-6.9%
(g) reverts reverting more than one editor		-100,998	-4,989	-105,987	-3.5%	-304,253	-180,659	-21,691	-506,603	-9.6%
Results after filtering ambiguous reverts	Σ	1,639,639	28,165	1,667,804	55.3%	1,720,706	393,813	44,454	2,158,973	41.1%
(h1) reverts reverting registered editors or bots		-328,492	-2,379	-330,871	-11.0%	0	-393,813	-44,454	-438,267	-8.3%
(h2) reverts reverting editors with IPv6 addresses		-1,677	-25	-1,702	-0.1%	-2,083	0	0	-2,083	0.0%
Results after all filtering steps	Σ	1,309,470	25,761	1,335,231	44.3%	1,718,623	0	0	1,718,623	32.7%

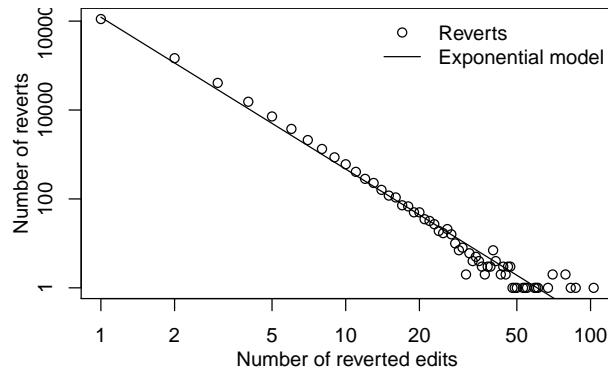


Figure 21: Number of full page reverts with a specific number of edits they revert and fitted exponential model in a log-log plot.

Italian Wikipedia: Geolocating Editors

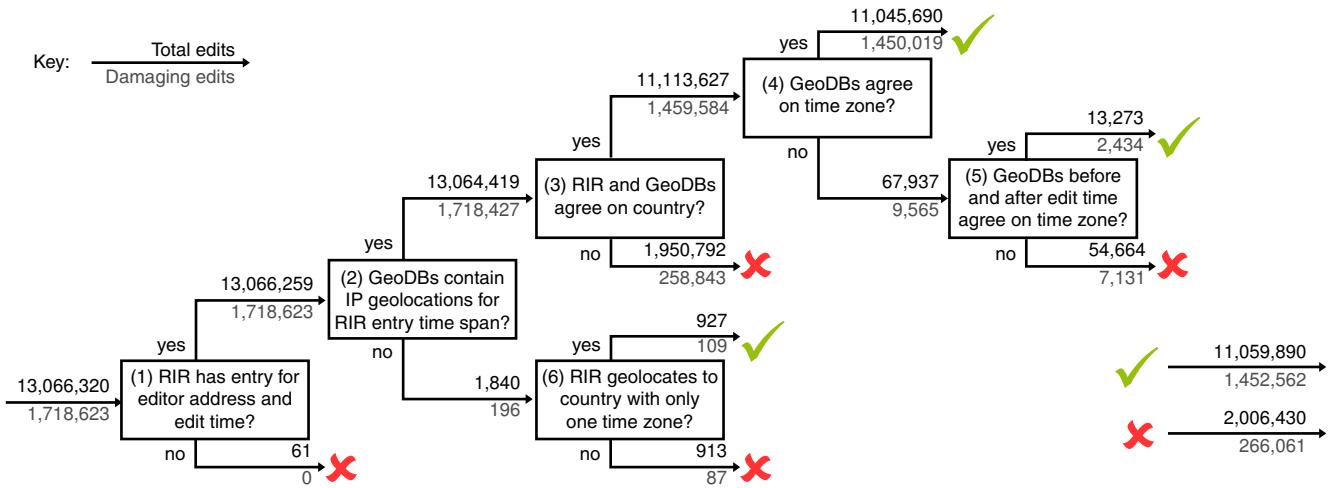


Figure 22: Decision tree to decide whether to trust the available geolocation information for an edit (✓), or not (✗). The numbers denote the total edits and reverted edits for the Italian Wikipedia that went through each branch.

Table 27: Historic geolocation success for all anonymous editors of the Italian 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 of the paper are given, and their corresponding IP addresses. Numbers are given for each exit node of the decision tree in the Figure above, 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
<i>Entire Wikipedia</i>		1,718,623 (13%)	13,066,320	865,065	3,544,154
No ✗	Step (1)	0 (0%)	61	0	19
	Step (3)	258,843 (13%)	1,950,792	144,079	598,605
	Step (5)	7,131 (13%)	54,664	3,710	16,564
	Step (6)	87 (9%)	913	46	421
	Σ	266,061 (13%)	2,006,430	147,835	615,608
Yes ✓		Step (4)	1,450,019 (13%)	11,045,690	716,189 2,924,129
		Step (5)	2,434 (18%)	13,273	1,026 4,378
		Step (6)	109 (11%)	927	46 191
	Σ	1,452,562 (13%)	11,059,890	717,256	2,928,686

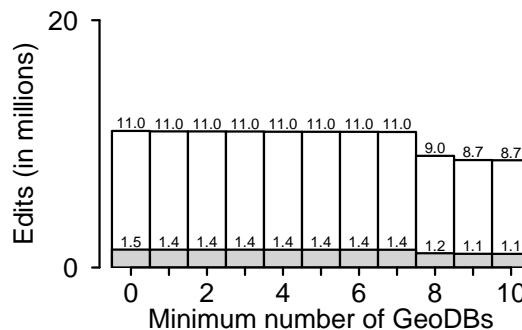


Figure 23: Number of total edits (white bars) and vandalism edits (gray bars) in millions from the yes-branch of Step (4) above over the number of GeoDBs considered.

Italian Wikipedia: Spatio-Temporal Analysis

Table 28: Number of edits by country, sorted for the Italian Wikipedia.

Country	Wikipedia						
	enwiki	dewiki	frwiki	eswiki	ruwiki	itwiki	jawiki
Italy	1,011,100	92,872	76,763	59,310	14,900	10,292,637	8,352

Table 29: Number of vandalism edits by country, sorted for the Italian Wikipedia.

Country	Wikipedia						
	enwiki	dewiki	frwiki	eswiki	ruwiki	itwiki	jawiki
Italy	106,007	19,459	8,063	7,116	1,738	1,365,327	725

Table 30: Number of vandalism commented edits by country, sorted for the Italian Wikipedia.

Country	Wikipedia						
	enwiki	dewiki	frwiki	eswiki	ruwiki	itwiki	jawiki
Italy	11,749	627	501	184	7	31,368	12

The tables above show the number of certain kinds of edits from specific countries in all analyzed variants of Wikipedia. Countries are selected to have at least 100,000 vandalism edits in the Italian Wikipedia, or that have Italian as a major language (according to the English Wikipedia) and at least 1,000 vandalism edits. Countries are sorted by the number of all edits in the Italian Wikipedia. The same countries are used in the hour-of-day plots.

Vandalism commented edits refers to edits that are—after filtering reverts—affected by a revert with a comment that signals that it is undoing vandalism (see Kittur et al. (2007b)). Note that this approach uses specific words to classify comments and is thus likely less effective in finding vandalism for other languages than English.

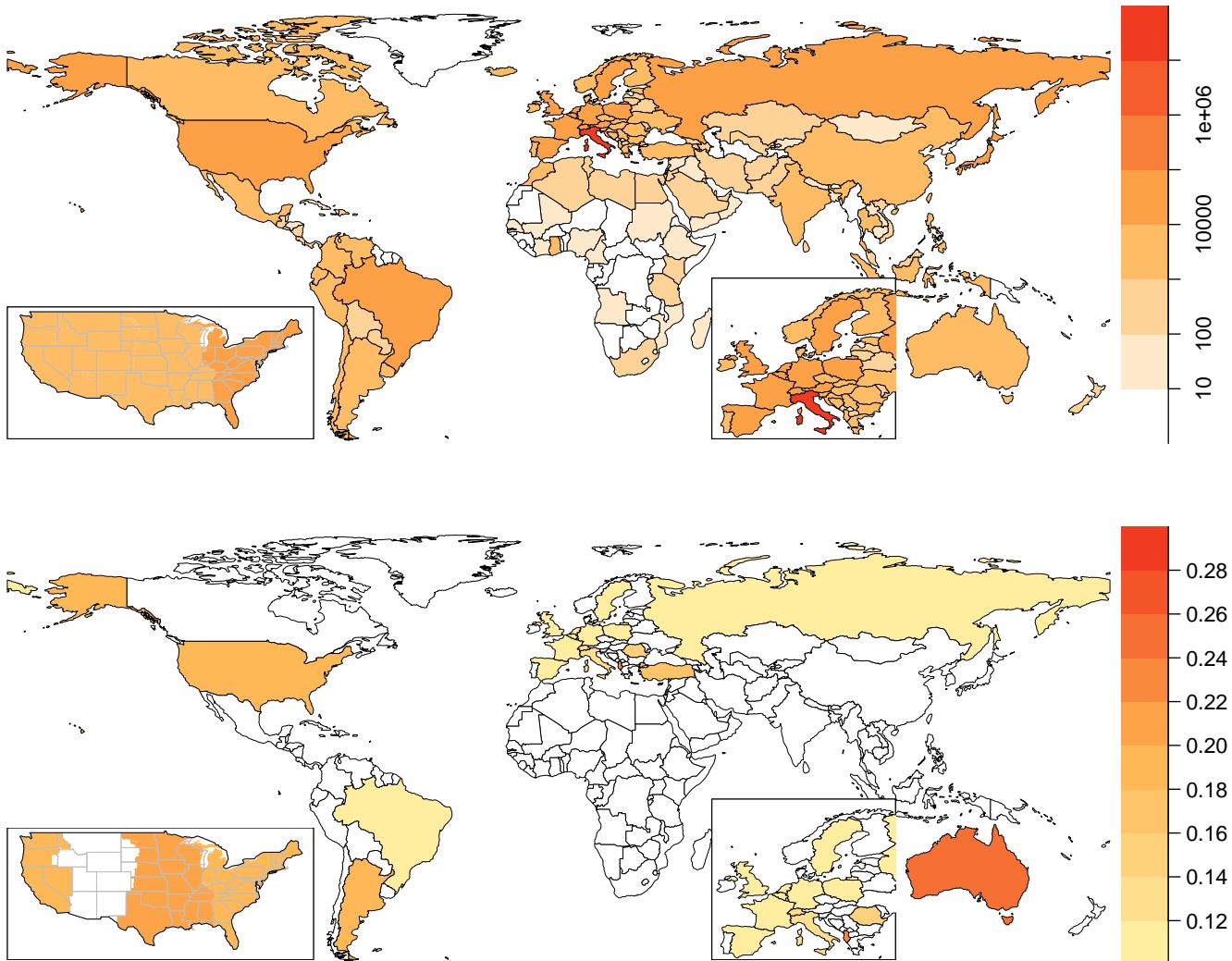
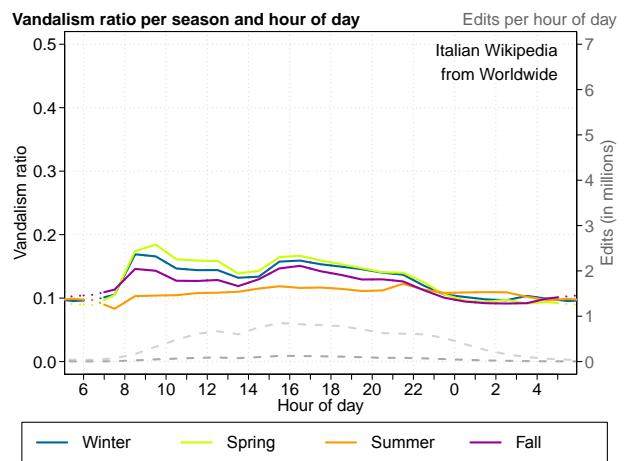
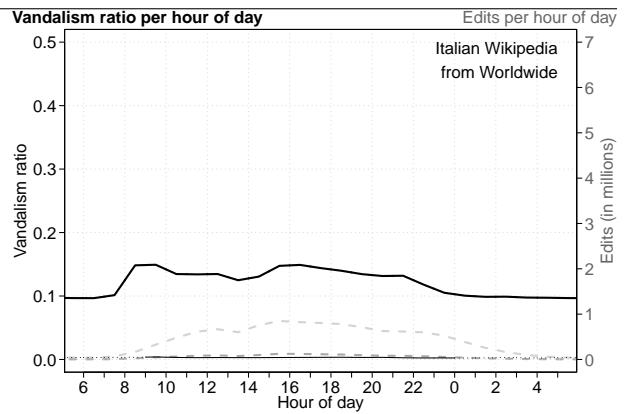
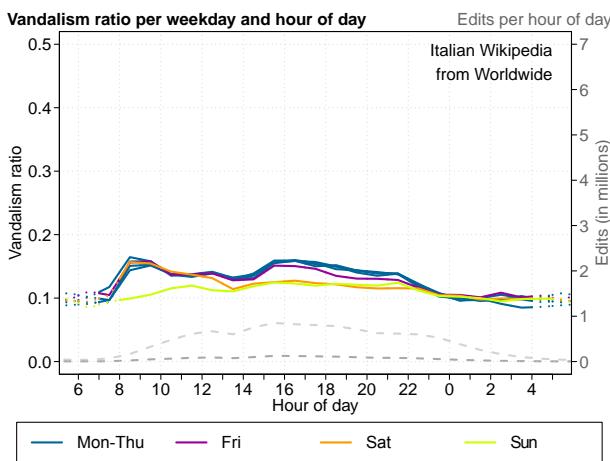


Figure 24: Number of edits (top) and ratio of vandalism to all edits (bottom) in the Italian Wikipedia by country. Countries with less than 1,000 vandalism edits are not colored. The embedded small maps show (left) the vandalism ratio in the United States (without Alaska) by major time zone (from West to East: Pacific, Mountain, Central, and Eastern) with overlaid state borders and (right) Europe enlarged.

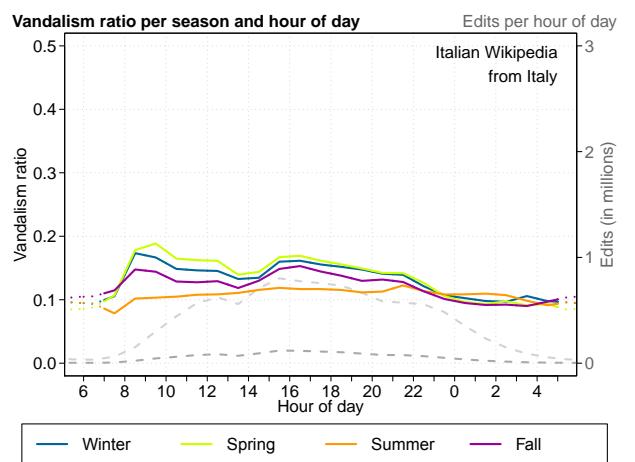
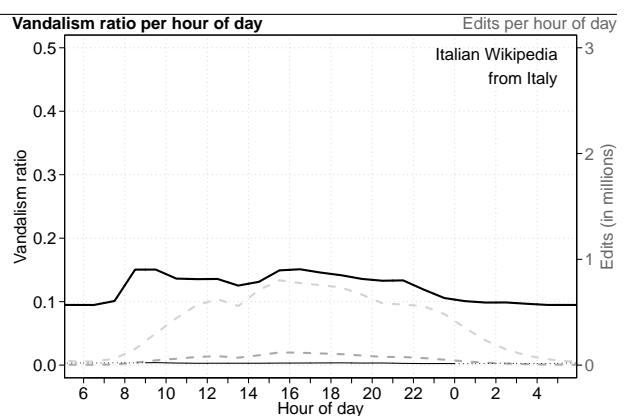
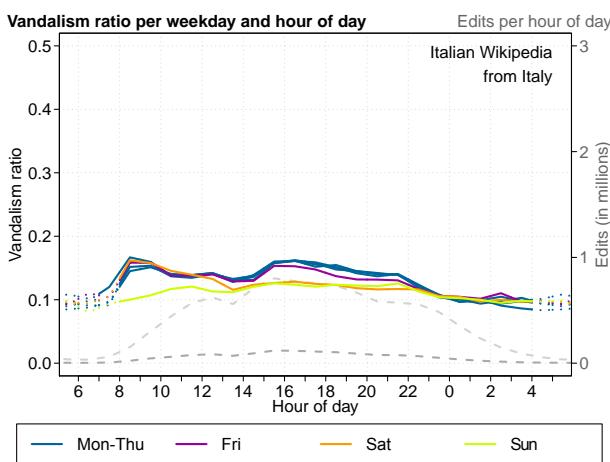
Italian Wikipedia

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 (light gray) and vandalism edits (dark gray) per hour of day (right axis, dashed lines), both averaged over the Italian Wikipedia's history. Ratios estimated from less than 1,000 vandalism edits are displayed with dotted lines.

The thin black line to the right is the vandalism ratio when detecting vandalism by comments of reverts (see Kittur et al. (2007b)). Note that this approach uses specific words to classify comments and is thus likely less effective in finding vandalism for other languages than English.



Italian Wikipedia Edits from Italy



Japanese Wikipedia

Japanese Wikipedia: Mining Vandalism

Table 31: Step-by-step filtering of the Japanese Wikipedia as per the revert patterns depicted in Figure 2 in the paper. 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. Note that the approach by Kittur et al. uses specific words to classify comments and is thus likely less effective in finding vandalism for other languages than English.

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	1,515,196	14,125	1,529,321	100.0%	1,310,049	1,157,844	69,538	2,537,431	100.0%	
(a) reverts to page blank	-117,389	-50	-117,439	-7.7%	-270,673	-318,807	-13,653	-603,133	-23.8%	
(b) empty reverts due to renaming/removal/error	-164,948	-148	-165,096	-10.8%	0	0	0	0	0.0%	
Results after filtering pseudo-reverts	Σ	1,232,859	13,927	1,246,786	81.5%	1,039,376	839,037	55,885	1,934,298	76.2%
(c) self reverts		-200,825	-23	-200,848	-13.1%	-108,014	-128,234	-8,695	-244,943	-9.7%
(d) revert corrections		-10,475	-76	-10,551	-0.7%	-19,152	-21,518	-421	-41,091	-1.6%
(e) reverted reverts		-12,231	-114	-12,345	-0.8%	-53,802	-65,200	-1,679	-120,681	-4.8%
Results after filtering error-corrections	Σ	1,009,328	13,714	1,023,042	66.9%	858,408	624,085	45,090	1,527,583	60.2%
(f) interleaved reverts		-248,124	-3,230	-251,354	-16.4%	-145,235	-193,572	-6,310	-345,117	-13.6%
(g) reverts reverting more than one editor		-53,145	-581	-53,726	-3.5%	-142,746	-117,620	-10,213	-270,579	-10.7%
Results after filtering ambiguous reverts	Σ	708,059	9,903	717,962	46.9%	570,427	312,893	28,567	911,887	35.9%
(h1) reverts reverting registered editors or bots		-270,943	-2,583	-273,526	-17.9%	0	-312,893	-28,567	-341,460	-13.5%
(h2) reverts reverting editors with IPv6 addresses		-7,555	-117	-7,672	-0.5%	-9,817	0	0	-9,817	-0.4%
Results after all filtering steps	Σ	429,561	7,203	436,764	28.6%	560,610	0	0	560,610	22.1%

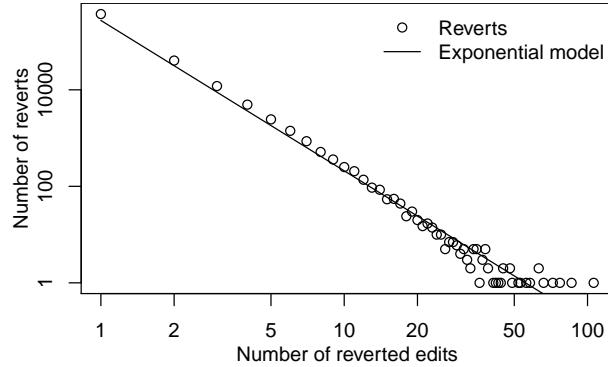


Figure 25: Number of full page reverts with a specific number of edits they revert and fitted exponential model in a log-log plot.

Japanese Wikipedia: Geolocating Editors

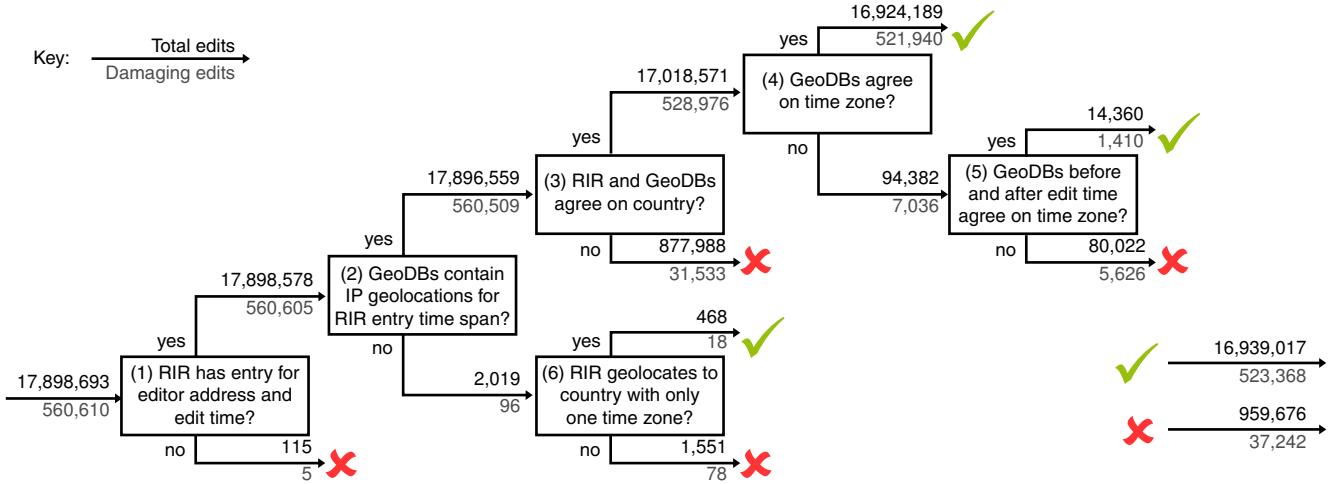


Figure 26: Decision tree to decide whether to trust the available geolocation information for an edit (✓), or not (✗). The numbers denote the total edits and reverted edits for the Japanese Wikipedia that went through each branch.

Table 32: Historic geolocation success for all anonymous editors of the Japanese 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 of the paper are given, and their corresponding IP addresses. Numbers are given for each exit node of the decision tree in the Figure above, 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
<i>Entire Wikipedia</i>		560,610 (3%)	17,898,693	205,959	2,753,753
No ✗	Step (1)	5 (4%)	115	3	36
	Step (3)	31,533 (3%)	877,988	12,023	149,589
	Step (5)	5,626 (7%)	80,022	2,394	20,942
	Step (6)	78 (5%)	1,551	37	550
	Σ	37,242 (3%)	959,676	14,457	171,116
Yes ✓	Step (4)	521,940 (3%)	16,924,189	190,911	2,578,540
	Step (5)	1,410 (9%)	14,360	588	4,155
	Step (6)	18 (3%)	468	15	151
	Σ	523,368 (3%)	16,939,017	191,514	2,582,831

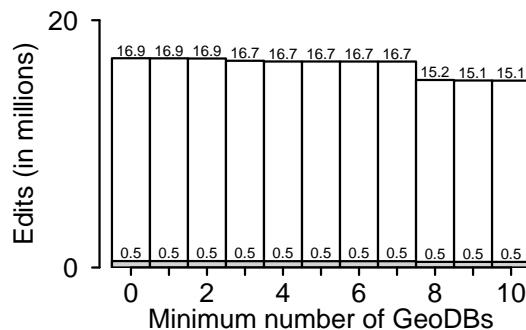


Figure 27: Number of total edits (white bars) and vandalism edits (gray bars) in millions from the yes-branch of Step (4) above over the number of GeoDBs considered.

Japanese Wikipedia: Spatio-Temporal Analysis

Table 33: Number of edits by country, sorted for the Japanese Wikipedia.

Country	Wikipedia						
	enwiki	dewiki	frwiki	eswiki	ruwiki	itwiki	jawiki
Japan	582,518	34,078	35,671	17,434	18,640	19,994	16,374,346

Table 34: Number of vandalism edits by country, sorted for the Japanese Wikipedia.

Country	Wikipedia						
	enwiki	dewiki	frwiki	eswiki	ruwiki	itwiki	jawiki
Japan	60,140	2,949	1,753	1,538	913	890	482,191

Table 35: Number of vandalism commented edits by country, sorted for the Japanese Wikipedia.

Country	Wikipedia						
	enwiki	dewiki	frwiki	eswiki	ruwiki	itwiki	jawiki
Japan	8,091	145	109	41	1	19	9,173

The tables above show the number of certain kinds of edits from specific countries in all analyzed variants of Wikipedia. Countries are selected to have at least 100,000 vandalism edits in the Japanese Wikipedia, or that have Japanese as a major language (according to the English Wikipedia) and at least 1,000 vandalism edits. Countries are sorted by the number of all edits in the Japanese Wikipedia. The same countries are used in the hour-of-day plots.

Vandalism commented edits refers to edits that are—after filtering reverts—affected by a revert with a comment that signals that it is undoing vandalism (see Kittur et al. (2007b)). Note that this approach uses specific words to classify comments and is thus likely less effective in finding vandalism for other languages than English.

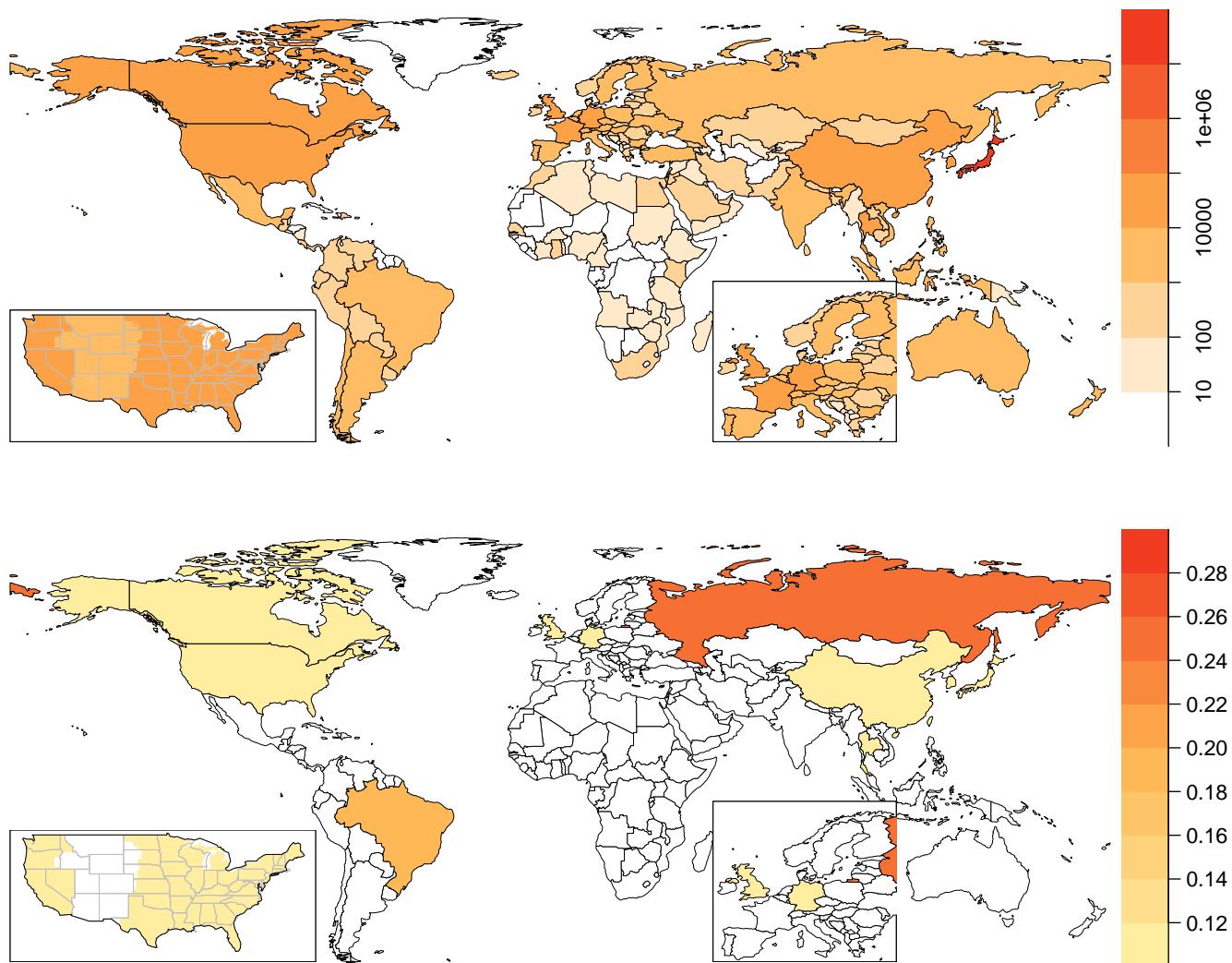
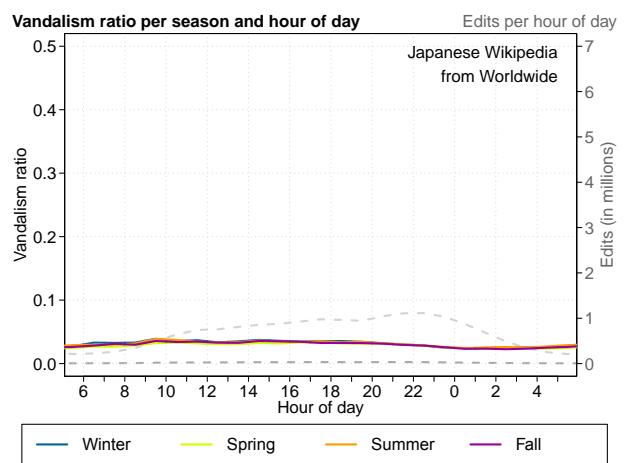
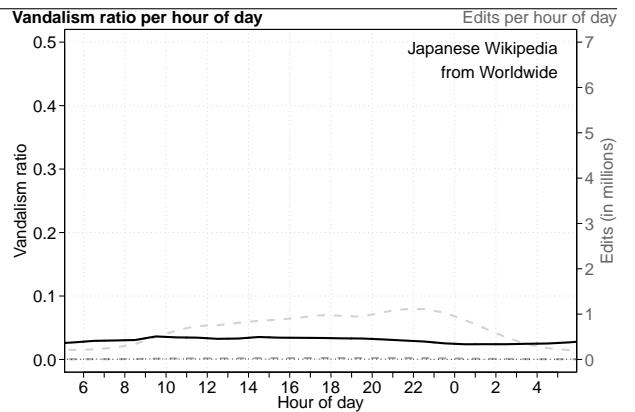
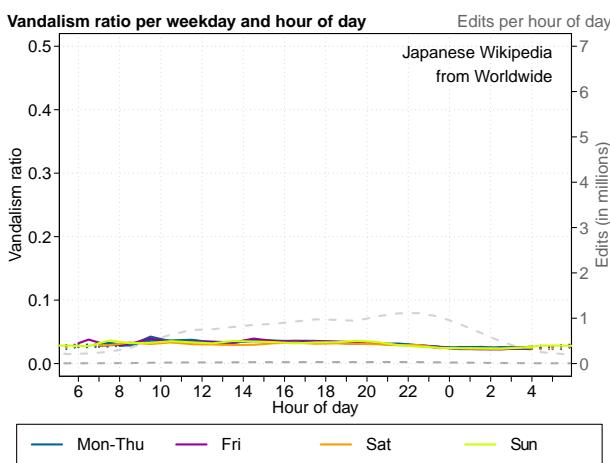


Figure 28: Number of edits (top) and ratio of vandalism to all edits (bottom) in the Japanese Wikipedia by country. Countries with less than 1,000 vandalism edits are not colored. The embedded small maps show (left) the vandalism ratio in the United States (without Alaska) by major time zone (from West to East: Pacific, Mountain, Central, and Eastern) with overlaid state borders and (right) Europe enlarged.

Japanese Wikipedia

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 (light gray) and vandalism edits (dark gray) per hour of day (right axis, dashed lines), both averaged over the Japanese Wikipedia's history. Ratios estimated from less than 1,000 vandalism edits are displayed with dotted lines.

The thin black line to the right is the vandalism ratio when detecting vandalism by comments of reverts (see Kittur et al. (2007b)). Note that this approach uses specific words to classify comments and is thus likely less effective in finding vandalism for other languages than English.



Japanese Wikipedia Edits from Japan

