10-4-2011

What Wikipedia Deletes: Characterizing Dangerous Collaborative Content

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

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

Seventh International Symposium on Wikis and Open Collaboration, Mountain View, California, USA, October 2011.

What Wikipedia Deletes: Characterizing Dangerous Collaborative Content

Abstract
Collaborative environments, such as Wikipedia, often have low barriers-to-entry in order to encourage participation. This accessibility is frequently abused (e.g., vandalism and spam). However, certain inappropriate behaviors are more threatening than others. In this work, we study contributions which are not simply ``undone'' -- but *deleted* from revision histories and public view. Such treatment is generally reserved for edits which: (1) present a legal liability to the host (e.g., copyright issues, defamation), or (2) present privacy threats to individuals (i.e., contact information).

Herein, we analyze one year of Wikipedia's public deletion log and use brute-force strategies to learn about privately handled redactions. This permits insight about the prevalence of deletion, the reasons that induce it, and the extent of end-user exposure to dangerous content. While Wikipedia's approach is generally quite reactive, we find that copyright issues prove most problematic of those behaviors studied.

Keywords
Wikipedia, user generated content, collaboration, redaction, content removal, copyright, information security

Disciplines
Community-based Research | Library and Information Science | Numerical Analysis and Scientific Computing | Other Computer Sciences | Other Legal Studies

Comments
Seventh International Symposium on Wikis and Open Collaboration, Mountain View, California, USA, October 2011.

What Wikipedia Deletes: Characterizing Dangerous Collaborative Content*

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

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

ABSTRACT

Collaborative environments, such as Wikipedia, often have low barriers-to-entry in order to encourage participation. This accessibility is frequently abused (*e.g.,* vandalism and spam). However, certain inappropriate behaviors are more threatening than others. In this work, we study contributions which are not simply "undone" – but *deleted* from revision histories and public view. Such treatment is generally reserved for edits which: (1) present a legal liability to the host (*e.g.,* copyright issues, defamation), or (2) present privacy threats to individuals (*i.e.,* contact information).

Herein, we analyze one year of Wikipedia's public deletion log and use brute-force strategies to learn about privately handled redactions. This permits insight about the prevalence of deletion, the reasons that induce it, and the extent of end-user exposure to dangerous content. While Wikipedia's approach is generally quite reactive, we find that copyright issues prove most problematic of those behaviors studied.

Categories and Subject Descriptors

H.5.3 [**Group and Organization Interfaces**]: *collaborative computing, computer-supported cooperative work*;
K.6.5 [**Management of Computing and Information Systems**]: Security and Protection

Keywords

Wikipedia, user generated content, collaboration, redaction, content removal, copyright, information security.

1. INTRODUCTION

User-generated content (UGC) and collaborative functionality is becoming increasingly prevalent in Web applications. The open-access models used in such systems enable the accumulation of content/knowledge at rates not possible in more traditional settings. For example, the video-sharing site YouTube has over 65,000 daily uploads [10], while the collaborative encyclopedia Wikipedia [3] had 45 million edits in the past year [4]. Inevitably, such services publish some inappropriate content: low barriers-to-entry invite poor contributions, while massive volume prevents thorough vetting. Indeed, it has been estimated that 7% of Wikipedia edits are unconstructive in nature (*i.e.,* vandalism) [15].

While abusive contributions may slowly erode the reputation of a service, this research concerns itself with only the most severe cases: content which is actually *dangerous* to the host-site or real-world individuals. Wikipedia, our basis for analysis, is no stranger to such threats. The encyclopedia has been threatened with litigation for copyright issues [14], accused of hosting child pornography [18], and briefly blacklisted in some regions for similar reasons [11].

In an attempt to mitigate these threats, Wikipedia *deletes* offending revisions from public view. We analyze one year's worth of public deletion logs to reason about the quantity and reasoning behind such actions. Moreover, by archiving Wikipedia revisions, we are able to recover deleted content and discover redactions handled in a more private fashion.

We find deletion is not uncommon, with some 55,000 edits being redacted/suppressed in 2010. However, the tool was broadly enabled only recently, and thus is being used to handle a backlog of old incidents. While this skews broad trends, focus on recent events reveals a rather reactive system. Most incidents are "undone" within minutes and deleted within several hours. Copyright issues, however, prove harder to identify. Strategies to detect such cases and address the consequences of a declining Wikipedia labor force [13] remain future challenges in this domain.

2. RELATED WORK

Given the short time for which revision deletion has been enabled on English Wikipedia (see Sec. 3), our work is the first to examine the process. Nonetheless, these are issues which other UGC applications have confronted. Whereas Wikipedia relies on a volunteer labor-force to find dangerous content, commercially-driven sites often outsource the review process [16]. In their analysis of YouTube, Cha *et al.* found that 0.4% of videos are deleted, but only 5% of these are due to copyright violations. In contrast, our analysis concentrates on text, not multimedia content (per Sec. 4.2).

More specific to Wikipedia is the work of Gehres *et al.* [12] which proposes a multi-level security *wiki*. Gehres' system is proactive in delegating roles/rights, while Wikipedia's deletion system limits read-access in an *ex post facto* manner. Meanwhile, Edwards [11] examines deletion/censorship on UGC websites, finding it a practical requirement to avoid blacklisting and regulatory troubles. Our motivation to pursue this topic was [17] and the notion that deletion could hide security events from public/researcher view.

*This work supported in part by ONR MURI N00014-07-1-0907.

ID	DESCRIPTION
RD1	Blatant copyright violations
RD2	Grossly insulting/offensive
RD3	Purely disruptive material
RD4	Revision pending suppression
RD5	Other valid deletion
RD6	Non-contentious housekeeping

Table 1: Redaction criteria [8]

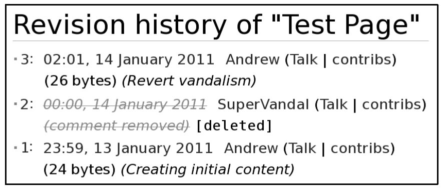

Figure 1: Page history w/redaction

CHANGES	#	%
Visibility increased	563	69%
Visibility decreased	188	23%
No visibility changes	40	5%
Orthogonal changes	25	3%
TOTAL	816	100%

Table 2: Visibility changes

3. DELETION ON WIKIPEDIA

Revision deletion (sometimes called *selective deletion* or *redaction*) on Wikipedia is enabled by a software feature called `RevDelete` [8]. Revision deletion removes individual edits from an article's history and is a distinct mechanism from *standard deletion* where whole entities are removed (articles, files, *etc.*). Standard deletions happen for both benign (*e.g.,* non-notable article topics) and malignant reasons (*e.g.,* pornography). However, for reasons described in Sec. 4.2, this work concentrates solely on revision deletions.

`RevDelete` was enabled for the ≈ 40 users with the `oversight` right in Jan. 2009. In May 2010, usage was extended to the ≈ 1800 users with `admin` privileges[1].

For each revision being handled, any combination of three fields can be redacted: (1) the *content*, those modifications made to the article (often visualized as a `diff`), (2) the *username* of the editor who made the change, and/or (3) the *summary* where the editor describes his/her modifications. Fig. 1 shows an example page history with a redacted edit.

The acceptable "criteria for redaction" are shown in Tab. 1 and covered in greater depth at [8]. It should be noted that "typical" vandalism and attacks do not merit deletion. Generally, one of these criteria is cited in the *publicly-viewable* deletion log. Users with `admin`/`oversight` rights can audit the actions of others, as they can see the deleted fields.

`RevDelete` also enables a stronger form of deletion called *suppression* or *oversight*[2]. It is identical to the weaker form except that: (1) it can only be performed by `oversight` users, (2) affected edits can only be viewed by `oversight` users, and (3) it is not publicly logged. Reasons for employing suppression are described at [7] and pertain primarily to defamation and privacy issues.

4. DATA COLLECTION

4.1 Public Logs

The public deletion log is accessible via the MediaWiki API [1]. Fields of interest include: (1) the revision-id (RID) of the affected edit, (2) a log-id, (3) log timestamp, (4) a comment field to explain the deletion, (5) a bit-field describing "old" visibility settings, and (6) a bit-field for new visibility settings. Similarly, the API can be used to gain information about affected revisions (those portions not redacted).

We processed this log from Jan. 2010 through Jan. 2011, storing information about roughly 50,000 redaction actions. Occasionally, the visibility of a single edit's fields are changed multiple times in the twelve-month history. As Tab. 2 shows, `RevDelete` users tend to show a conservative bias, initially censoring more fields then eventually deemed necessary. We remove "preliminary" actions from our dataset, considering only "final" assignments. Further, rows where the final state is complete visibility (*i.e.,* "undeleted") are discarded. These two changes leave 49,161 unique revisions/rows for analysis.

[1]If an "ordinary" user discovers dangerous content, permissioned users can be notified using off-wiki channels.

[2]To avoid ambiguity, we treat the two forms in a disjoint fashion. The terms *deletion* and *redaction* will refer exclusively to the weaker (but more common) form, while *suppression* will describe stronger uses of the tool.

4.2 Archiving Content

The public deletion log provides no data on two relevant fronts: (1) the actual content redacted, and (2) usage of the suppression function. However, by fully archiving *all* Wikipedia revisions immediately after they are committed, we can learn more about both aspects. If one has archived data for a RID which later appears in a deletion log, then one has its redacted fields. Similarly, if one has archived data for a RID, but a subsequent request indicates redaction, then the RID has been suppressed (if there is no public log entry).

The wholesale collection of Wikipedia data presents ethical and legal issues. For example, one could acquire child-pornography – the possession of which is illegal. This motivates our decision to archive *only text content*. Of course, text content may also have legal implications (a motivating factor of this research). Our institution's legal counsel has advised that our research is protected because it is a *consumer* of such content, not a *distributor* thereof. This work reproduces no deleted/suppressed content.

To archive content, we used a combination of "Recent Changes" IRC channels and the Wikipedia API [1]. For each edit to the main article namespace we store the RID and the three fields eligible for reaction (content, username, and summary). This was done for Aug. 2010, archiving approximately 4 million edits[3]. To find suppressed edits, we re-queried the API for all RIDs in our archive several months later, noting those revisions with redacted fields.

5. DATA ANALYSIS

5.1 Incident Groupings

In the previous section, we identified 49,161 revisions affected by redaction. However, "revision-level" analysis is not ideal. Imagine revision r_n introduces dangerous content. Subsequent revisions $r_{n+1} \ldots r_{n+x}$ may be constructive, but fail to remove the threat. When the dangerous content is discovered, all edits back to r_n will need to be redacted, because the threat persists through them. Thus, r_{n+1} onward are essentially "collateral damage" of the earlier offense and underscore why "incident-level" analysis is more intuitive. Where possible, we present incident-based statistics.

In our data, *incidents* are identified when multiple RIDs share a log-id (as well as log timestamp/comment) and are therefore the result of a single `RevDelete` action. For our

[3]Additional data collection was forgone given the significant bandwidth costs to both our own servers and those operated by Wikipedia. Further, we seek only a glimpse into this "private" data given our more complete public sets.

MO	RD1	RD2	RD3	RD4+	OTH	SUM
Jan.	2	11	0	1	9	23
Feb.	3	23	10	2	4	42
Mar.	25	31	3	1	27	87
Apr.	1	17	5	0	18	41
May	17	697	1006	2	97	1819
Jun.	37	913	427	37	101	1515
Jul.	88	718	1695	6	158	2665
Aug.	167	840	103	51	313	1474
Sep.	129	1846	161	18	193	2347
Oct.	252	5067	179	19	165	5682
Nov.	1087	535	112	14	215	1963
Dec.	338	323	152	84	352	1249
SUM	2146	11021	3853	235	1652	18907

Table 3: Deletion incidents (month × rationale)

49k revisions we identify 18,907 incidents. While 89% of incidents have just one revision, copyright-related incidents (identified per Sec. 5.4) have an average of 12.5 revisions. This is intuitive: copyright incidents are less obvious than other violations and are thus more likely to go unnoticed.

5.2 Redaction Prevalence

The "sum" column of Tab. 3 shows the quantity of incidents flagged per month. Clearly, the decision to enable `RevDelete` for `admins` (a 50× increase in the user-base) in May 2010 had a profound effect. It would appear these additional users benefit Wikipedia's well-being.

Pinpointing the prevalence of dangerous revisions among the complete set of edits is difficult. Roughly 45 million edits were made to English Wikipedia in 2010 [4], and that same year saw ≈ 19k incidents redacted. It should be emphasized that only 7,978 (42%) of the incidents flagged in 2010 actually occurred in that year. It remains to be seen if this is a side-effect of the tools infancy (with a backlog of incidents being cleared[4], now that a mechanism exists to redact them) or if some dangerous content is successful in evading detection for such durations (see Sec. 5.5).

Of course, these figures represent only dangerous content that is *caught*. It is difficult to quantify threats that are: (1) live on the site, or (2) still accessible in page histories. Available data does allow us to state that *at least* 0.05% of revisions made in 2010 contained dangerous content. While not overwhelming – a single incident could amount to legal action (or a privacy leak) under the right circumstances.

5.3 Fields Affected

Tab. 4 shows the frequency of redaction for each of the three eligible fields. In brief, content is deleted in 75% of incidents and the summary in 25% of cases. Username redactions are quite rare[5]. These results are unsurprising: article content is the foundation of Wikipedia and thus also the field most often in need of deletion.

5.4 Reasons for Redaction

Recall from Tab. 1 the six criteria for redaction. Tab. 3 shows the prevalence of each reason when grouped by incident. Mapping an incident to its criteria is straightforward, given that `RevDelete` users conventionally cite reasons in their block-log comments (which we assume trustworthy).

Vague/generalized criteria descriptions [8] and standardized log entries, however, do not intuitively tell of the *de facto* thresholds for redaction. By examining deleted content (per our archival), we attempt to provide informal insight.

Our inspection found offensive content (RD2) revisions were almost all directed at an explicit human individual. Unsourced claims about these persons commonly involved sexual innuendo, claims of promiscuity or pedophilia, and racial slurs – usually carried out in profane language. Acts of disruption (RD3) were found to be remarkably similar, also including more "appropriately written" falsities, solicitations, and massive insertions of random content.

Copyright violations (RD1) are common[6], yet straightforward. Such cases are not harmful on the surface, but in their method of content acquisition. Criteria RD4, RD5, and RD6 are rarely cited and given no further attention[7]. The "OTH" (other) column of Tab. 3 aggregates log-comments not citing a particular issue. In general, this set contains a diverse set of the previous criteria (see also footnote 7).

REDACTED	NUM	%
`content`	13616	72.0%
`summary`	4082	21.6%
`user`	832	0.8%
`summary + content`	151	4.4%
`user + content`	51	0.3%
`user + summary`	14	0.1%
`all fields`	161	0.8%
TOTAL	18907	100.0%

Table 4: Redacted fields for incidents

5.5 Redaction Response Times

One way to measure the impact of dangerous content is the *detection interval*, the time duration between an incident's beginning and eventual redaction. Fig. 2 visualizes these intervals, indicating a median detection period of over 13,000 hours – or about 1.5 years. Clearly, as discussed in Sec. 5.2, this speaks to the handling of a backlog of threats prior to the time `RevDelete` was broadly enabled. Looking only at incidents occurring after May 2010, results are more encouraging. Analysis shows a median interval of 2 hours for all incidents, yet 21.6 days for copyright-specific ones.

However, this does not indicate if fresh "old" incidents are still being discovered. To this end, Fig. 3 shows detection intervals based on the month of flagging. We would expect to see any "backlog" being handled just after `RevDelete` was widely enabled (May) and detection times improving with continued usage. Results for [Nov. – Dec.] are a vast improvement over the Fig. 2 average, but [Sep. – Oct.] showed the slowest detections of any interval studied. Future analysis is needed to determine the convergence of this trend.

In addition to the *detection interval*, we can measure a critical subset of that period, an incident's *active duration*. The former is the time redacted edits are accessible via revision histories. The latter is the period when the damage was in a default-visible version (*i.e.,* that most-recent). All users can make an edit *inactive* by simply editing the article. Impressively, Fig. 4 reveals a median active duration of about *two minutes* for all incidents, but 21 days for copyright infringements (nearly identical to their detection interval).

[4]A log of long-term abuse to Wikipedia is maintained at [5].
[5]Proactive patrolling of the "user creation log" may locate offensive usernames before they can edit.

[6]Tab. 3 may underestimate RD1 prevalence. In a Sep. 2010 incident, 25,000 suspected copyright infringements were found [6], but the matter was not resolved using `RevDelete`.
[7]In fact, *no* incidents cite RD4 (revision pending oversight/suppression). Such labeling is likely avoided, as it would invite attention to edits that `admins` should not view.

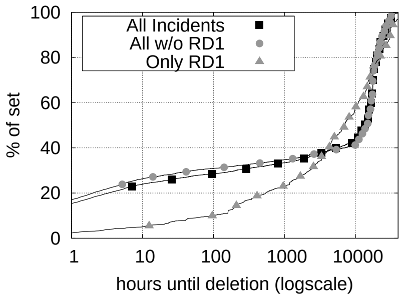

Figure 2: Detection intervals

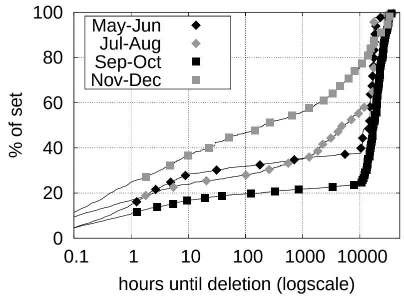

Figure 3: Detect-time by flag month

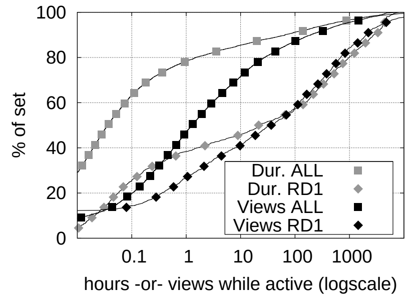

Figure 4: Active dur./views

MO	#SP	#RD	MO	#SP	#RD
Jan.	536	40	Jul.	658	4134
Feb.	770	78	Aug.	287	4739
Mar.	445	144	Sep.	338	6077
Apr.	356	76	Oct.	557	9946
May	446	2803	Nov.	492	13509
Jun.	221	3137	Dec.	487	5790
			SUM	5593	50473

Table 5: Num. suppressions (SP) & redactions (RD)

5.6 End-user Exposure

Combining the survival times of the last section with page-view statistics, we can better measure the *exposure* of dangerous content (*i.e.,* the number of visitors who see it).

To this end, we collected hourly, per-article view statistics [2] for the entirety of 2010. Assuming uniform intra-hour hit distributions, one can produce a view-estimate for any incident's active duration. Fig. 4 shows the CDF of active view counts (only for incidents active in 2010). The median case receives $\approx$ 1.25 views, suggesting that unpopular pages are frequent targets and/or threatening content on popular articles is dealt with very quickly. Unsurprisingly, copyright incidents fared more poorly with a median of 36 views.

Broadly, view statistics can be aggregated to show that there were roughly 5.9 million views of dangerous revisions in 2010 (or 11 views per-minute). From this perspective, Wikipedia seems to be winning the content battle. Given that the English version served 85 billion pages in 2010 [4], just 0.007% contained content that has since been redacted.

5.7 Suppression

To this point, our analysis has concentrated on the simple form of redaction, not the stronger *suppression* available only to `oversight` users. As Tab. 5 shows, this lack of focus is warranted given that suppression actions occur an order-of-magnitude less frequently than redaction ones[8].

Our brute-force archival and re-querying was successful in identifying 100+ suppressed edits. Unable to view the private logs, we cannot establish incident groupings or detection intervals (though they could be lengthy, given the very small number of `oversight` users).

However, we can determine the fields that get suppressed. We find content removal to be most common, followed by usernames. Summary suppression is exceedingly rare in our small sample. Finally, manual inspection of revisions shows that the publication of individual's addresses and phone numbers is the most common reason for suppression.

[8]Suppression counts were obtained from [9]. That source counts *actions*, so our presentation of redaction quantity does the same (*i.e.,* includes "undeletes" and changes).

6. CONCLUSIONS

In this work, we processed one year's worth of English Wikipedia's public deletion logs and used archival strategies to both recover redacted content and discover privately suppressed revisions. We found that `RevDelete` was used to handle nearly 55,000 redactions/suppressions in 2010, most often hiding content exhibiting the characteristics of libel, copyright infringement, and privacy violations.

We also found that the tool, only recently being widely enabled, is being used to eliminate a backlog of old incidents. Focus on recent incidents indicates a reactive system. For instance, dangerous content is usually inactive within two minutes, with formal deletion within two hours. We found that 0.007% of page views in 2010 resulted in exposure to threatening content. Many such views were the result of copyright issues – the most problematic of behaviors studied.

Detecting these copyright issues and preventing dangerous content altogether both appear worthwhile areas for research. Such progress could reduce the liability of UGC hosts and improve perceptions of the collaborative paradigm.

References

[1] Wikimedia API. `http://en.wikipedia.org/w/api.php`.
[2] Wikimedia page-view statistics. `http://dammit.lt/wikistats`.
[3] Wikipedia. `http://www.wikipedia.org`.
[4] Wikistats: Wikimedia statistics. `http://stats.wikimedia.org`.
[5] WP: Long-term abuse. `http://en.wikipedia.org/wiki/WP:LTA`.
[6] WP: Mass blanking of copyright violations. `http://en.wikipedia.org/wiki/WP:Wikipedia_Signpost/2010-09-13/`.
[7] WP: Oversight. `http://en.wikipedia.org/wiki/WP:OS`.
[8] WP: Revision deletion. `http://en.wikipedia.org/wiki/WP:RVDL`.
[9] WP: Suppression statistics. `http://en.wikipedia.org/wiki/WP:Arbitration_Committee/Audit_Subcommittee/Statistics`.
[10] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon. I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system. In *IMC*, 2007.
[11] L. Edwards. Content filtering and the new censorship. In *ICDS '10: Proc. of the Conference on Digital Society*, 2010.
[12] P. Gehres, N. Singleton, G. Louthan, and J. Hale. Toward sensitive information redaction in a collaborative, multilevel security environment. In *WikiSym*, 2010.
[13] E. Goldman. Wikipedia's labor squeeze and its consequences. *Jour. of Telecommunications and High Tech. Law*, 8, 2009.
[14] J. Merante. UK Natl. Portrait Gallery threatens Wikipedia user over public domain images. `http://creativecommons.org/weblog/entry/15764`, July, 14 2009.
[15] R. Priedhorsky, J. Chen, S. K. Lam, K. Panciera, L. Terveen, and J. Riedl. Creating, destroying, and restoring value in Wikipedia. In *GROUP*, 2007.
[16] B. Stone. Policing the Web's lurid precincts. *The New York Times*, page B1, July 18, 2010.
[17] A. G. West, J. Chang, K. Venkatasubramanian, and I. Lee. Link spamming Wikipedia for profit. In *CEAS*, 2011.
[18] J. Winter. Wikipedia distributing child porn, co-founder tells FBI. *FoxNews.com*, April 27, 2010.