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

Follow this and additional works at: [http://repository.upenn.edu/cis_papers](http://repository.upenn.edu/cis_papers)

Part of the [Community-Based Research Commons](http://repository.upenn.edu/cis_papers), [Library and Information Science Commons](http://repository.upenn.edu/cis_papers), [Numerical Analysis and Scientific Computing Commons](http://repository.upenn.edu/cis_papers), [Other Computer Sciences Commons](http://repository.upenn.edu/cis_papers), and the [Other Legal Studies Commons](http://repository.upenn.edu/cis_papers)

Recommended Citation


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

This paper is posted at ScholarlyCommons. [http://repository.upenn.edu/cis_papers/478](http://repository.upenn.edu/cis_papers/478)

For more information, please contact libraryrepository@pobox.upenn.edu.
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.

This conference paper is available at ScholarlyCommons: http://repository.upenn.edu/cis_papers/478
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.
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 privileges1.

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 oversight2. 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 defacement 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 deletion 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 than 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.

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 edits3. 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 $r_{n+1} \ldots r_{n+k}$ 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

1Additional 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.

---

Table 1: Redaction criteria [8]

<table>
<thead>
<tr>
<th>ID</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>RD1</td>
<td>Blatant copyright violations</td>
</tr>
<tr>
<td>RD2</td>
<td>Grossly insulting/offensive</td>
</tr>
<tr>
<td>RD3</td>
<td>Purely disruptive material</td>
</tr>
<tr>
<td>RD4</td>
<td>Revision pending suppression</td>
</tr>
<tr>
<td>RD5</td>
<td>Other valid deletion</td>
</tr>
<tr>
<td>RD6</td>
<td>Non-contentious housekeeping</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>RD1</td>
<td>Blatant copyright violations</td>
</tr>
<tr>
<td>RD2</td>
<td>Grossly insulting/offensive</td>
</tr>
<tr>
<td>RD3</td>
<td>Purely disruptive material</td>
</tr>
<tr>
<td>RD4</td>
<td>Revision pending suppression</td>
</tr>
<tr>
<td>RD5</td>
<td>Other valid deletion</td>
</tr>
<tr>
<td>RD6</td>
<td>Non-contentious housekeeping</td>
</tr>
</tbody>
</table>

Figure 1: Page history w/redaction

Table 2: Visibility changes

<table>
<thead>
<tr>
<th>CHANGES</th>
<th>#</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visibility increased</td>
<td>563</td>
<td>69%</td>
</tr>
<tr>
<td>Visibility decreased</td>
<td>188</td>
<td>23%</td>
</tr>
<tr>
<td>No visibility changes</td>
<td>40</td>
<td>5%</td>
</tr>
<tr>
<td>Orthogonal changes</td>
<td>25</td>
<td>3%</td>
</tr>
<tr>
<td>TOTAL</td>
<td>916</td>
<td>100%</td>
</tr>
</tbody>
</table>

---
Available data does allow us to state that at least that is caught evading detection for such durations (see Sec. 5.5).

Redact them) or if some dangerous content is successful in (1) live on the site, or (2) still accessible in page histories.

of incidents being cleared if this is a side-effect of the tools infancy (with a backlog 2010 actually occurred in that year. It remains to be seen its were made to English Wikipedia in 2010 [4], and that same year saw \( \approx 19k \) 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\(^1\), 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.

### 5.2 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 \( \approx 19k \) 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\(^1\), 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\(^4\)Al\(\ast\)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.

---

### Table 3: Deletion incidents (month × rationale)

<table>
<thead>
<tr>
<th>MO</th>
<th>RD1</th>
<th>RD2</th>
<th>RD3</th>
<th>RD4+</th>
<th>OTH</th>
<th>SUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan</td>
<td>2</td>
<td>11</td>
<td>0</td>
<td>1</td>
<td>9</td>
<td>23</td>
</tr>
<tr>
<td>Feb</td>
<td>3</td>
<td>23</td>
<td>10</td>
<td>2</td>
<td>3</td>
<td>42</td>
</tr>
<tr>
<td>Mar</td>
<td>25</td>
<td>31</td>
<td>3</td>
<td>3</td>
<td>27</td>
<td>87</td>
</tr>
<tr>
<td>Apr</td>
<td>1</td>
<td>17</td>
<td>5</td>
<td>0</td>
<td>18</td>
<td>41</td>
</tr>
<tr>
<td>May</td>
<td>17</td>
<td>697</td>
<td>1006</td>
<td>2</td>
<td>97</td>
<td>1819</td>
</tr>
<tr>
<td>Jun</td>
<td>37</td>
<td>913</td>
<td>427</td>
<td>37</td>
<td>101</td>
<td>1515</td>
</tr>
<tr>
<td>Jul</td>
<td>88</td>
<td>718</td>
<td>1695</td>
<td>6</td>
<td>158</td>
<td>2665</td>
</tr>
<tr>
<td>Aug</td>
<td>167</td>
<td>840</td>
<td>1010</td>
<td>51</td>
<td>313</td>
<td>1474</td>
</tr>
<tr>
<td>Sep</td>
<td>129</td>
<td>1846</td>
<td>161</td>
<td>18</td>
<td>193</td>
<td>2347</td>
</tr>
<tr>
<td>Oct</td>
<td>252</td>
<td>5067</td>
<td>179</td>
<td>19</td>
<td>165</td>
<td>5682</td>
</tr>
<tr>
<td>Nov</td>
<td>1087</td>
<td>535</td>
<td>112</td>
<td>14</td>
<td>215</td>
<td>1963</td>
</tr>
<tr>
<td>Dec</td>
<td>338</td>
<td>323</td>
<td>152</td>
<td>84</td>
<td>352</td>
<td>1249</td>
</tr>
</tbody>
</table>

SUM: 2146 × 1021 = 8535 × 235 = 1652 = 18907

### Table 4: Redacted fields for incidents

<table>
<thead>
<tr>
<th>REDACTED</th>
<th>NUM</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>content</td>
<td>13616</td>
<td>72.0%</td>
</tr>
<tr>
<td>summary</td>
<td>4082</td>
<td>21.6%</td>
</tr>
<tr>
<td>user</td>
<td>832</td>
<td>0.8%</td>
</tr>
<tr>
<td>user + content</td>
<td>101</td>
<td>4.4%</td>
</tr>
<tr>
<td>user + summary</td>
<td>14</td>
<td>0.1%</td>
</tr>
<tr>
<td>all fields</td>
<td>161</td>
<td>0.8%</td>
</tr>
</tbody>
</table>

TOTAL: 18907 × 100.0%
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 ≈ 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.

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 copyright infringement, and privacy violations.

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.

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 copyright infringement. 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

   Arbitration_Committee/Audit_Subcommittee/Statistics.