EDUCAUSE
CENTER FOR APPLIED RESEARCH

Open Wikis and the Protection of Institutional Welfare

Research Bulletin
February 7, 2012

Andrew G. West, University of Pennsylvania
Insup Lee, University of Pennsylvania

Overview

Web 2.0 technologies facilitate knowledge sharing, interaction, and collaboration through user-generated content. No paradigm has embodied these qualities more purely than the wiki. Central to the wiki philosophy is the ability of participants not just to *add* content but also to *modify* content.

Higher education is both a consumer and a producer of wiki content. The online collaborative encyclopedia Wikipedia, for example, is frequently used as a research device (although usually informally). On public wikis faculty and staff might modify content that pertains to the institution or its sponsored research. Moreover, institutionally supported wikis are becoming increasingly common as a platform for hosting IT help desks, course web pages, or as centralized repositories for knowledge about a particular academic domain.

Much has been written about wikis' reliability[1] and use in the classroom.[2] This research bulletin addresses the *negative* impacts on institutional welfare that can arise from participating in and supporting wikis. The open nature of the platform, which is fundamental to wiki operation and success, enables these negative consequences. Security implications are minimized when a finite wiki user base can be determined a priori (e.g., a course roster; see the subsection "Securing Institutionally Supported Wikis"), hence our discussion in this bulletin primarily concerns *open* or *public* wikis that accept contributions from a broad and unknown set of Internet users.

Our recent research focuses on the malicious behaviors enabled by open wikis. The ease of access to public wikis spurred our investigations; consider that a damaging wiki edit could be seen by thousands of people while imposing near-zero marginal costs on the attacker. Quite simply, the security of the platform has not fully matured—a disparity our research aims to correct.

While simple damage ("vandalism") had previously been identified as an issue, our work has been novel in its analysis of "link spam" and "legally threatening content" in wiki environments. We primarily examine the statistical properties of these malicious actions in order to build machine-learning models capable of detecting future instances. In particular, we focus on "reputation management" as a means of identifying the perpetrators based on an historical record of their actions.

Broader, future goals for this research include:

- Devising generic models that are applicable across numerous collaborative systems and natural languages; and
- Developing more-formal wiki systems and extensions that provide accountability and can make security guarantees while maintaining open collaboration.

For now, inappropriate behaviors will continue, with consequences for the two primary stakeholders: (1) the perpetrator and (2) the host or administrator of the platform. *Both* parties potentially face a loss of reputation, technical repercussions, and legal liability. From an institutional perspective, these concerns are already relevant. For example, students, faculty, and staff (*constituents*) might engage in poor behaviors on a third-party site. Moreover, if an institution hosts or supports the wiki of a constituent, it puts its own reputation at risk.

The bulk of our research has used English Wikipedia[3] as its data set and testbed due primarily to its popularity, visibility, and availability of data. More importantly, wikis are just one type of *Internet-enabled collaborative application,* a class of functionality that also includes blog comments, web forums, social networks, and code repositories. While our discussion is often wiki-specific, the ramifications are far-reaching. Consider also that we approach these issues from a technical perspective and some subtleties of higher education policy remain beyond the scope of this bulletin.

Research Highlights

Each month in 2011, English Wikipedia received an average of nearly four million edits and seven *billion* page views.[4] The encyclopedia's success can be traced to its large and extremely dedicated user base, or *community*,[5] which has used brute force to vet a majority of the content additions. We do not believe that such extensive reliance on human effort is useful or sustainable. Goldman[6] reports that user retention has suffered as the encyclopedia's operation increasingly shifts from "content creation" to mundane "protection and maintenance." As participation wanes, he foresees limitations in the editing model to stop abuse. We believe solutions can be more technically elegant.

The types of damage confronted include:

- *Vandalism*—blatantly unconstructive contributions characterized by offensive language, immature content, insertion of falsities, or deletion of existing content
- *Link spam*—posting of URLs with the intention of transferring wiki visitors to a promotional, commercial, or irrelevant third-party page
- *Legal threats*—contributions having legal implications (e.g., copyright violations, libel)
- *Conflicts of interest*—attempts to "spin" content to advance a particular entity or idea

Each of these poor behaviors[7] is discussed in greater depth in the following subsections. Economics underlies our philosophy for stopping each type of misbehavior (as is common in applied security):

Detection need not be flawless; it only needs to render an attack "unprofitable" for the perpetrator.

The notion of "profit" can differ wildly: some vandals might thrive on seeing their edit survive, while commercial link spammers operate under purely monetary terms. Unfortunately, highly secure policies frequently conflict with the core philosophies of open collaboration, degrade the user experience, and affect user retention rates. At the same time, we must be sensitive to the investment of the wiki's editors and defenders, whose time and patience are a limited resource.

Vandalism

Previous research has shown and our own work has confirmed that about 7% of all Wikipedia edits (or about 9,000 each day) are vandalism. Most often vandalism is characterized by obscenity, hate speech, test edits ("Look, I can edit Wikipedia!"), fact changing, attempts at comedy, deletion of

quality content, and offensive images. Wikipedia's most pervasive issue, vandalism has been the subject of much negative publicity (including the Seigenthaler incident[8]).

Prior to 2009 automated vandalism detection used only coarse rule systems. An offensive word might score *x* points, excessive capitalization several more, etc. The sum of these points was compared against a threshold value that, if exceeded, resulted in the automatic "undo" of the edit. These rules were arbitrary, rarely derived from concrete evidence, and easy to circumvent.

This limited approach motivated us to academically propose and practically implement a vandalism-detection engine not derived from language but instead based on the statistical properties of edit metadata.[9] A key component involved constructing reputations for both users and the articles being edited, based on their revision histories. A user who behaved badly in the past might logically behave badly in the future, and previously controversial pages will probably continue to be controversial (Table 1 shows the pages most vandalized up to 2010). Our research quantified and formalized this notion. Also novel was our use of spatial logic to apply reputations to users with no edit history (what reputation management calls the "cold start" problem). As Table 2 shows, simply geo-locating a user's IP address proves fruitful given that American and Australian users are six times more likely to commit vandalism than users from Western Europe. Harnessing a dozen such statistical features in combination produced a prediction model[10] that performed comparably to leading systems of that time. Since, we have made incremental improvements to our model,[11] although seemingly more accurate predictors are in operation today.[12]

Table 1. Most Vandalized Articles in English Wikipedia (as of early 2010)

Rank	Article Title
1	George W. Bush
2	Wikipedia
3	Adolf Hitler
4	United States
5	World War II
6	Jesus Christ
7	George Washington
8	Bill Clinton

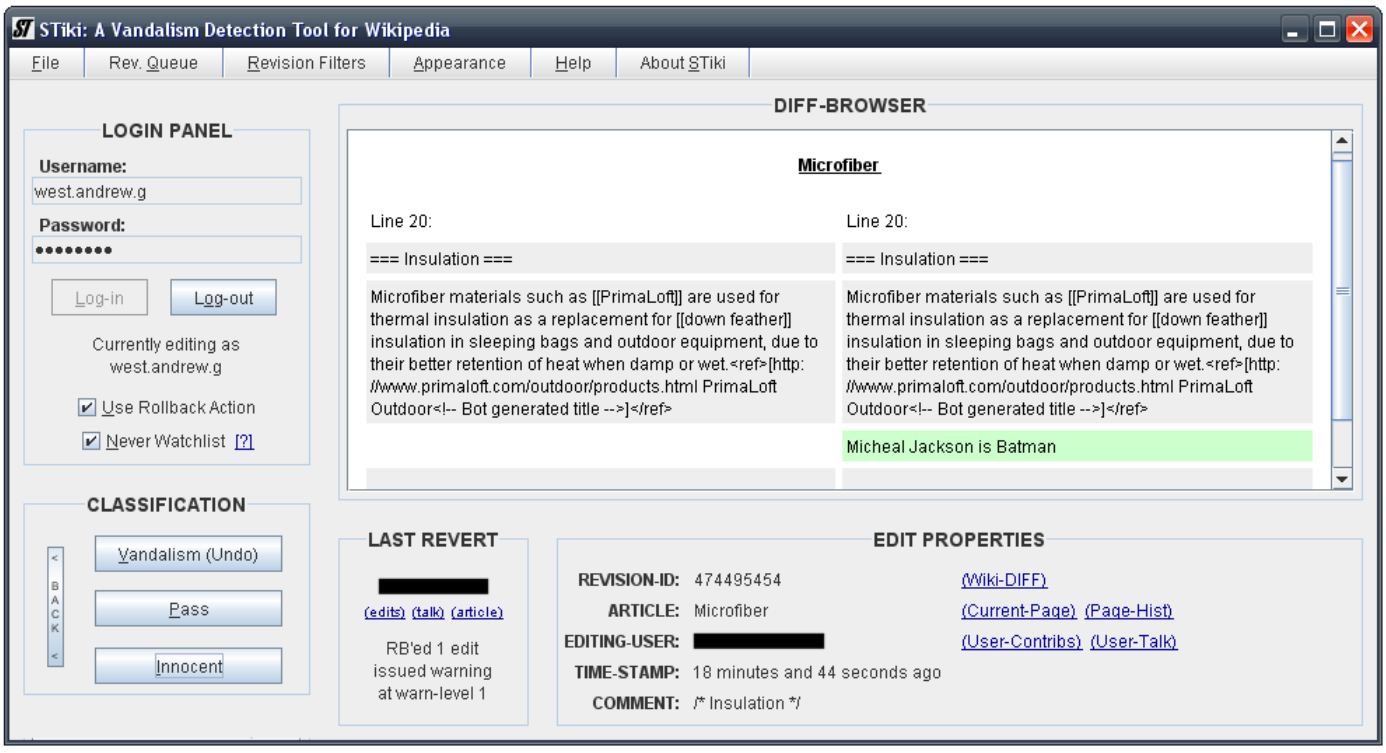
Table 2. Normalized Vandalism Rate on English Wikipedia by Country (per IP Geo-Location, Countries with 100,000+ Edits)

Rank	Country	Percentage of Vandalism
1	Italy	2.85%
2	France	3.46%
3	Germany	3.46%
...	...	...
12	Canada	11.35%
13	United States	11.63%
14	Australia	12.08%

Software capable of reliably identifying vandalism can be run in an autonomous fashion, undoing poor edits nearly instantaneously. This is a powerful defense against vandals whose profit or incentive is undermined when their damage receives no exposure. Even so, leading Wikipedia systems can only automatically find and remove about 40% of vandalism at tolerated false-positive rates,[13] leaving the remainder for human mitigation.

To assist these human users, we developed an *intelligent routing tool*, STiki,[14] that prioritizes edit-review based on algorithmically generated vandalism probabilities (when those probabilities fall below the threshold needed for automatic reversion). Vandalism thus has a decreased survival time and humans operate more efficiently than with brute-force random search. The tool also employs crowd-sourced operation to ensure the same edit is not simultaneously vetted by multiple persons (or multiple times). Though applicable to many wikis, STiki has been best demonstrated on English Wikipedia, where it has assisted humans in removing 68,000-plus instances of vandalism and remains in active development. Figure 1 shows the STiki interface displaying an instance of vandalism (the phrase "Micheal Jackson is Batman" has been added), while the lower-left corner contains the "classification buttons," which are the primary means of interacting with the tool.

Figure 1. STiki Interface Displaying Vandalism



As a result of autonomous detection algorithms and tools like STiki, vandalism now only survives a median of 85 seconds on English Wikipedia, and just *one reader* is likely to see the damage.[15] Once vandalism is found, the offending user receives warnings of escalating severity, and persistent vandals will find their IP address(es) blocked from editing. While these statistics bode well for English Wikipedia, it is important to realize that few wikis and collaborative environments have such a sizeable toolkit and dedicated user base at their disposal.

Link Spam

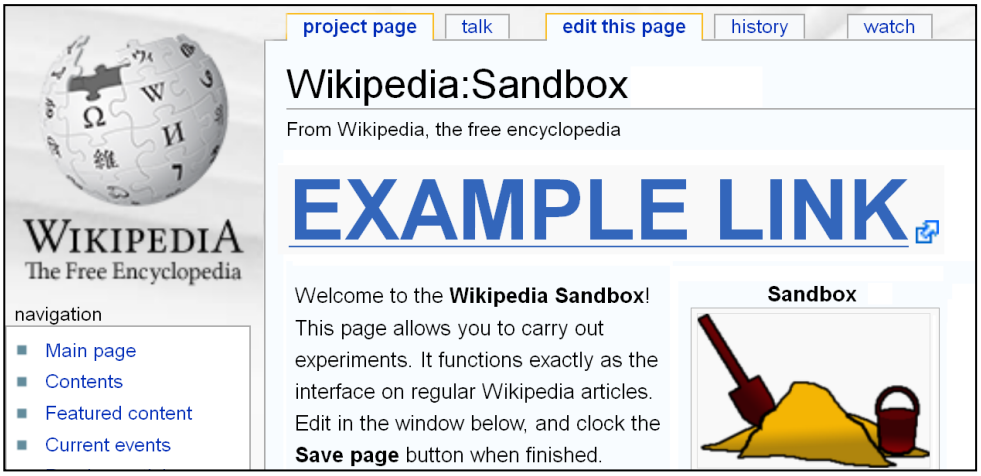
Link spammers presumably have a greater incentive than simple vandals, as their success might yield greater (monetary) profits. While it is possible to conduct marketing on-wiki, direct monetization of a user requires them to exit the environment via a hyperlink and visit a third-party commercial site.

Mitigating Direct Click-Through Spam

A measurement study we conducted in mid-2010 surveyed the extent and strategies of spammers on English Wikipedia.[16] To our surprise, we found spamming behaviors to be relatively infrequent and technically naïve. Evidence suggested that spammers tried to conform to constructive linking behaviors in the hope their URLs would go unnoticed and extract long-term utility. This evidently did little to fool the dedicated editors, but neither did it convince us the platform was secure.

We conducted a vulnerability assessment and described a novel spam attack model[17] that leverages the fact that humans carry out virtually all anti-spam work on Wikipedia, leading to an inherent (albeit brief) detection latency. These brief windows of opportunity can be aggressively exploited by spam-blanketing the most popular pages with prominent links (see Figure 2), which takes advantage of wiki editing freedoms.

Figure 2. Example of Link Spam with Prominent Link Placement



Consider, for example, the February 2010 Super Bowl halftime show featuring "The Who." While the band was playing, their Wikipedia article received over 150 views *per second*. A spam link prominently inserted at that time would have received considerable reader views even with just a seconds-long lifespan. We suspected that such links could be monetized and generate revenue that would outweigh the costs associated with mounting the attack. Statistical estimation and a proof-of-concept empirical study confirmed this economic viability.[18]

Exposing this vulnerability prompted research into its mitigation. We developed a system similar to our anti-vandalism classifier, adding features specific to the spam problem.[19] This system actually outperforms its anti-vandalism counterpart precisely because of the additional intelligence (e.g., obtaining the URL being linked and examining its commercial potential).

Problems with Distributed and Poorly Maintained Wikis

One reason we did not find Wikipedia-specific attacks in our measurement study might be because broadly attacking numerous, less popular, and less secure collaborative environments proves more

profitable for attackers. Existing black-hat (malicious) software enables these attacks at massive scale.[20] One such tool, XRumer, crawls the Internet for web forms in collaborative applications like wikis, web forums, and blog comments. By exploiting the code similarity and known weaknesses of common software platforms (e.g., Mediawiki for wikis, Wordpress for blogs), XRumer uses a template-driven attack to mimic how people would fill out the web forms. In doing so, it includes the spam URL so that if the posting succeeds, it will be indexed by search engines. In this manner a human never has to see or click the link for the attack to succeed. Anecdotal evidence suggests that such black-hat software is a significant source of spam in Web 2.0 environments.[21]

An anti-spam engine could be used in such a situation. For low-traffic wikis, however, simply breaking the "template" the software expects to exploit (e.g., by hiding an invisible form the software will attempt to fill out, but a legitimate user cannot) might suffice. Moderate- to high-traffic wikis might need higher barriers to entry (also raising the attacker's marginal costs), among other techniques (see the "Best Practices" section).

Legally Dangerous Content

Vandalism and spam waste administrative time and can erode a wiki's reputation. However, some content is even more dangerous in that it reveals private information or introduces legal liability. These instances should be redacted from the edit history (i.e., completely removed from archives and public view) to protect the stakeholders involved.

By archiving every edit made to Wikipedia, our research[22] identified and inspected revisions that were subsequently redacted. Some 55,000 edits were redacted in 2010, with most motivated by two legal concerns: insertion of copyrighted content (i.e., plagiarizing text from a copyrighted source without permission), and libel or slander (i.e., unsourced negative claims about some identifiable individual). Although far less common, the publishing of personal information (phone numbers, addresses, etc.) receives similar treatment. Figure 3 shows an example of a revision history from which an edit has been redacted (the middle edit).

Figure 3. Revision History showing Redaction

The 55,000 Wikipedia redactions represent only instances that (1) the community succeeded in locating and (2) were textual in nature (ethical concerns prevented our deeper investigation). Undiscovered dangerous content and multimedia items likely contribute heavily to outstanding

liability issues: Wikipedia (and its users) have been threatened with lawsuits,[23] issued DMCA takedown notices, and accused of hosting child pornography.[24]

Some copyright violations occur accidentally, for example when a user does not understand licensing requirements. Others appear deliberate. In 2010 it was discovered that a long-time contributor to English Wikipedia might have committed as many as 25,000 copyright violations during his career.[25] It is difficult to assess his motivations; perhaps he simply took pride in being perceived as a major contributor while investing minimal effort.

The ability to detect legally threatening content remains unrealized. We believe that the libel/slander issue is an order-of-magnitude less acute than copyright given that the former frequently exhibits, and is treated, as vandalism. Our future work intends to examine the utility of plagiarism detection algorithms in identifying copyrighted textual content, although multimedia items and offline resources would fall outside this scope.

Conflicts of Interest

Wikipedia strives for objectivity, with articles required to have "encyclopedic significance" and policies stating "no original research" and "maintain neutral point-of-view (NPOV)." These tenets do not serve the self-interest of many who see Wikipedia's seven billion monthly page views as an opportunity for brand exposure and shaping readers' perceptions. Marketing departments, paid ghostwriters, invested researchers, and devotees to any particular idea or institution may try to "spin" topics in their favor—demonstrating a prohibited conflict of interest.

Conflict of interest damage showed up in a 2006 incident where Wikipedia temporarily blocked the IP proxy of the U.S. House of Representatives and brought increased scrutiny to revisions from the U.S. Senate IP block.[26] The source of the controversy was a series of edits (some admittedly by representatives' aides) that "scrubbed" criticism from Congress members' articles and added more positive representations. The U.S. Department of Justice and the Church of Scientology were previously cited in the same manner.[27]

This phenomenon is difficult to quantify; there is a fine and subjective line between constructive authoring and bias. Nonetheless, it remains extremely relevant in an institutional context.

What It Means to Higher Education

Wikipedia contends with multiple security issues of considerable complexity. Now we use that case study to discuss how these threats and lessons learned are relevant to higher education institutions, and then how higher education can minimize these risks.

Institutional Vulnerabilities

Threats to institutional welfare initiated by collaborative applications divide into two parts:

- Those arising from the behavior of students, faculty, or staff on third-party collaborative sites
- Those occurring when constituents operate collaborative sites using institutional resources

Vulnerabilities in Constituent Access

Simply *reading* a damaged wiki can have negative consequences for the viewer, but this is not a major concern for higher education.[28] Instead, we focus on the possibility that constituents might actually be *committing* damage. In the case of vandalism, educational institutions at all levels are

known to be a frequent source of perpetrators.[29] When any form of damage originates from an institutional IP address, several threats to organizational welfare occur:

- First, this is likely a prohibited use of institutional computer or network resources. We believe vandalizing a wiki is the technological parallel of facilitating e-mail spam and should be handled with similar severity.
- Second, wikis' goals often align well with those of academia, implying a certain degree of social and ethical responsibility to keep these shared public resources free from damage.
- Third, while a wiki's administrative logs and "edit histories" are rarely well-trafficked, some reputation is at stake when the damage can be mapped back to a particular institution.
- Fourth, there are the possible technical consequences, particularly in shared-use settings (computer labs or proxy connections). For example, a subsequent constituent might receive a warning message for damage wrought by a previous user, or an IP block (or range of such blocks) could affect the editing capabilities of other users.

Though perhaps less malicious in intent than most vandalism and spam, "conflicts of interest" and "subtle manipulations" are nonetheless damaging. These behaviors are not just occasionally committed by constituents, they are often *institutionally sponsored* and *condoned* via communications offices aiming to maintain and improve online presence. These individuals use Wikipedia as a tool to squelch controversies and help in institutional recruitment in a phenomenon so prevalent that Wikipedia has advised universities not to modify their own articles.[30]

Wikipedia is a powerful recruiting device. For perspective, the "University of Pennsylvania" article sees some 2,500 daily views. The scope is also broad with significant buildings, athletics, individual academic programs, and cultural traditions all possibly having their own articles. We were able to identify over 100 English Wikipedia articles pertinent to the University of Pennsylvania's operations.

As with vandalism, these conflict-of-interest events reflect poorly on the institution. The bigger loss to institutional welfare, however, is the missed *opportunity*. If the author had respected wiki norms, the contributions could have been mutually beneficial to both the institution and the encyclopedia. The same advice is relevant to institutionally sponsored research. Faculty and research personnel frequently add superfluous citations to their own work or articles about their findings/publications that do not merit encyclopedic coverage.

Vulnerabilities of Institutionally Supported Wikis

The second type of institutional vulnerabilities arise when (external) individuals target collaborative environments that draw on an institution's resources. We have shown the maintenance and legal challenges that Wikipedia faces, and when an institution hosts public collaborative functionality, it assumes similar risks.

In private/closed wikis participation is limited to a finite and *known* set of participants. Although far more dangerous to the host, open wikis do have a place in academia. Consider QWiki, which is an academic source on quantum physics that is hosted by Stanford University.[31] Depending on institutional policy (or lack of adherence), personal wikis could also be hosted on institutional servers.

Just as Wikipedia's reputation suffers when it is vandalized, spammed, or manipulated, so too does the reputation of an institution whose wiki serves damaged articles in a public-facing fashion. Damage is nearly inevitable in public wikis, and it incurs maintenance costs. Much of the burden

presumably falls on the constituent who established the wiki. The time required to locate and undo damage might be a considerable drain on institutional resources, and a frank assessment is needed to determine if making a wiki public produces benefits that outweigh these costs.

Finally, there is legally threatening content. External users can introduce liability for the institution by posting copyright text or multimedia, attacking individuals, submitting illegal content (e.g., malicious software, child pornography), or using the wiki platform's bandwidth and storage functionalities for inappropriate purposes.

Best Practices for Institutions

Maintaining an institution's welfare in the face of collaborative functionality entails observing best practices that span the technical, political, and administrative domains.

Determining Institutional Participation

The goal of securing constituent access is to:

- Understand the type and extent of inappropriate behaviors, and
- Implement steps to prevent, mitigate, and improve these behaviors.

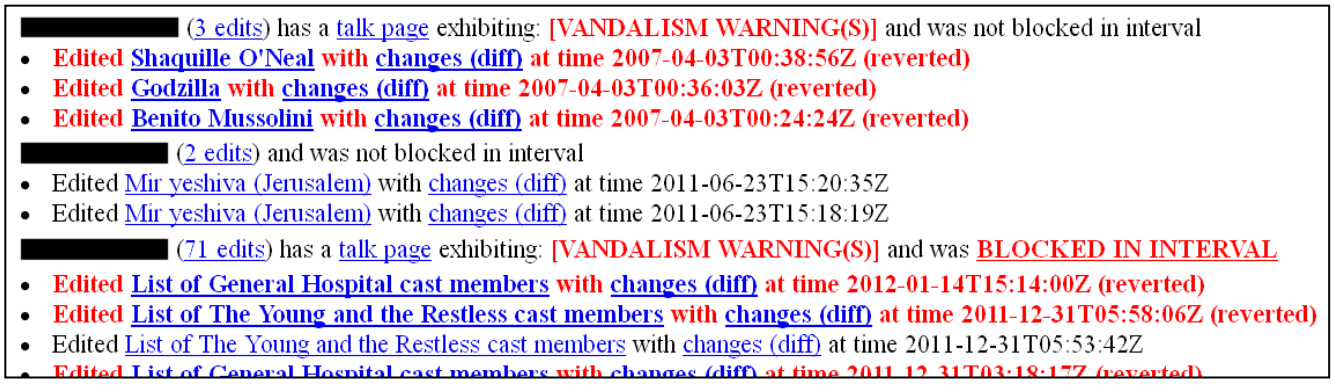
Presumably IT administrators could monitor and examine all network traffic for connections to wikis and other collaborative applications. However, this is technically complex, has privacy implications, and is probably excessive in the context of higher education.

Instead, we advocate an approach characterized by passive monitoring and administrative cooperation. To this end we developed a software tool called WikiAudit[32] for use by IT administrators. As its input WikiAudit takes a list or range of IP addresses[33] (i.e., those of the institution) and the wiki to be analyzed. Its output is a report that describes:

- Blocked IP addresses
- IPs believed to have participated in poor behavior
- A comprehensive listing of the edits made by those IPs

Figure 4 shows a snippet of one such report showing two problematic users (top, bottom) and one benevolent user (middle). IP addresses have been redacted to obfuscate the identities of those involved. Combining this audit data with internal logs (i.e., mapping IPs to institutional logins) should suffice to identify the guilty parties at the institutional level.

Figure 4. Example WikiAudit Report


[redacted] (3 edits) has a talk page exhibiting: [VANDALISM WARNING(S)] and was not blocked in interval

- **Edited Shaquille O'Neal with changes (diff) at time 2007-04-03T00:38:56Z (reverted)**
- **Edited Godzilla with changes (diff) at time 2007-04-03T00:36:03Z (reverted)**
- **Edited Benito Mussolini with changes (diff) at time 2007-04-03T00:24:24Z (reverted)**

[redacted] (2 edits) and was not blocked in interval

- Edited Mir yeshiva (Jerusalem) with changes (diff) at time 2011-06-23T15:20:35Z
- Edited Mir yeshiva (Jerusalem) with changes (diff) at time 2011-06-23T15:18:19Z

[redacted] (71 edits) has a talk page exhibiting: [VANDALISM WARNING(S)] and was BLOCKED IN INTERVAL

- **Edited List of General Hospital cast members with changes (diff) at time 2012-01-14T15:14:00Z (reverted)**
- **Edited List of The Young and the Restless cast members with changes (diff) at time 2011-12-31T05:58:06Z (reverted)**
- Edited List of The Young and the Restless cast members with changes (diff) at time 2011-12-31T05:53:42Z
- **Edited List of General Hospital cast members with changes (diff) at time 2011-12-31T03:18:17Z (reverted)**

WikiAudit can be configured to work for a large percentage of wiki installations, not just Wikipedia, but these represent only a fraction of Internet-enabled collaborative applications. While it may prove helpful to manually establish cooperative relationships with major Web 2.0 or collaborative players (e.g., Wikipedia, Facebook, etc.), it is an open question how best to handle the remainder.

Monitoring and Modifying Institutional Pages

Distinct from constituent misbehavior is maintenance of the institution's online presence. Simple monitoring of institutionally relevant articles (e.g., against vandalism) is noncontroversial and essential. This is a role best handled by someone within a communications office. To aid these persons, one can have notifications delivered when "watched" pages are modified.

Greater controversy arises when constituents go beyond simple maintenance and start adding/deleting substantial content in which they have a vested interest (e.g., the institution itself or its sponsored research). We believe such editing can be conducted responsibly and to the benefit of both the institution and host wiki. To achieve objectivity, constituents must learn about wiki culture and norms. They should become registered editors and publically divulge their organizational membership. They can use discussion pages to seek help and to present or propose content. Finally, Wikipedia has a "Campus Ambassador" outreach program that provides university liaisons with the goal of further integrating wiki technology into the classroom.[34]

Securing Institutionally Supported Wikis

The goal in securing institutionally supported wikis is to prevent external individuals from damaging the wikis of well-intended constituents and thus minimize the maintenance costs and institutional liability.

In the case of closed/private wikis, security can be both straightforward and robust: Participation should be on an invitation-only basis. While appropriate for research groups and small projects, this approach does not scale well. Participation could also be limited to an IP range (e.g., on-campus computers), integrated with institutional login, or tied to course management software. While straightforward, these suggestions are not automatic. Unfortunately, much wiki software has an extremely permissive default configuration, putting the burden on (possibly) nontechnical users to realize the need for configuration adjustments and make them.

Alternatively, institutions could prohibit independent installation, instead providing the service through IT help desks. When enabling a collaborative service for a constituent, the administration can collect basic data useful in later oversight, as well as deliver a service with restrictive access permissions. Should the constituent desire a more open collaborative model, the request offers an opportunity for education and dialogue about the threats and challenges involved.

The numerous vulnerabilities of open wiki models result directly from the impossibility of adequately securing them. Nonetheless, some best practices can limit threats to institutional welfare. Although tempting to apply statistical models for this purpose, the overhead, resources, and training required make them appropriate only for popular and active wikis. In the future, we hope to see these types of services integrated directly into wiki software (as many blogging platforms have done with anti-spam protection) or made available as a third-party service.

For less-active public wikis, simpler techniques thwart attackers: breaking their template-driven attack, increasing their marginal operating costs, and making alternative targets appear attractive. For example:

- Users can be forced to register accounts instead of editing using IP addresses as identifiers.
- These accounts must be registered using e-mail confirmation.
- CAPTCHAs challenges (obfuscated puzzles designed to distinguish humans from machines) can be presented at account registration or whenever a sensitive action is performed.

These protections will never stop a sufficiently motivated human attacker (and increasingly, automated attackers are outsourcing evasion procedures to low-cost human labor markets[35]). Moreover, these protections do encumber the user experience of innocent users and can even halt their participation. Striking a balance between security and usability should be carefully considered when configuring a collaborative environment.

Clearly there is no panacea for the issues faced. This presents challenges to wiki administrators, the organizations that support them, and the researchers who work to find increasingly elegant solutions.

Key Questions to Ask

- What data does our institution collect—or should it collect—regarding constituent participation in open collaborative communities such as Wikipedia? For example, what content is being accessed and modified? By whom? For what purpose?
- What policies are in place, or need to be established, to minimize immature or malicious contributions in these open collaborative environments?
- To what extent are faculty and staff monitoring the online presence of our institution and sponsored research in collaborative settings?
- How can the institution effectively maintain an online presence while balancing its desire for positive attention with the (sometimes) conflicting community guidelines regarding neutral point of view and self-promotion?
- What information does our institution collect about wikis and collaborative applications hosted on institutional servers? What governing body is responsible for their administration? Should other stakeholders be included, and if so, who and why?
- What technical protections, advice, and oversight can we offer to maximize the institution's reputation and minimize liability where collaborative functionality is institutionally supported?

Where to Learn More

- Broughton, John. *Wikipedia: The Missing Manual*. O'Reilly Media, 2008. Also available in online form at http://en.wikipedia.org/wiki/Help:Wikipedia:_The_Missing_Manual
- Geiger, R. Stuart, and David Ribes. "The Work of Sustaining Order in Wikipedia: The Banning of a Vandal." Paper presented at CSCW '10: The ACM Conference on Computer Supported Cooperative Work, Savannah, Georgia, USA, February 2010.
- Heymann, Paul, Georgia Koutrika, and Hector Garcia-Molina. "Fighting Spam on Social Websites: A Survey of Approaches and Future Challenges," *IEEE Internet Computing* 11, no. 6 (2007): 36-45.

- Watson, Kate, and Chelsea Harper. "Supporting Knowledge Creation: Using Wikis for Group Collaboration" (Research Bulletin, Issue 3). Boulder, CO: EDUCAUSE Center for Applied Research, 2008, available from http://www.educause.edu/ecar.
- West, Andrew G. and Insup Lee. "What Wikipedia Deletes: Characterizing Dangerous Collaborative Content ." Paper presented at WikiSym '11: The 7th International Symposium on Wikis and Open Collaboration, Mountain View, CA, USA, October 2011.

About the Authors

Andrew G. West (westand@cis.upenn.edu) is a doctoral candidate in the Department of Computer and Information Science at the University of Pennsylvania, where Insup Lee (lee@cis.upenn.edu) is the Cecilia Filter Moore Professor and Director of the PRECISE Center.

Citation for this Work

Andrew G. West and Insup Lee. "Open Wikis and the Protection of Institutional Welfare" (Research Bulletin). Boulder, CO: EDUCAUSE Center for Applied Research, February 7, 2012, available from http://www.educause.edu/ecar.

Notes

[1] Jim Giles, "Internet Encyclopedias go Head to Head," *Nature* 438, no. 7070 (2005): 900-901.

[2] Kate Watson and Chelsea Harper, "Supporting Knowledge Creation: Using Wikis for Group Collaboration" (Research Bulletin, Issue 3) (Boulder, CO: EDUCAUSE Center for Applied Research, 2008), available from http://www.educause.edu/ecar

[3] "Wikipedia, the free encyclopedia," http://en.wikipedia.org/

[4] "Wikistats: Wikimedia Statistics," http://stats.wikimedia.org/

[5] These communities/user bases usually operate in a democratic fashion with a minimum of hierarchy. Thus, it is possible for *any* user to remove damaging contributions, making for a very fluid distinction between the roles of reading, editing, and administration (cumulatively, *participation)*. Realize also that virtually all participants are uncompensated. Wikipedia does have employees for operational support, fundraising, etc., but they do not participate in low-level encyclopedic matters in these capacities.

[6] Eric Goldman, "Wikipedia's Labor Squeeze and its Consequences," *Journal of Telecommunications and High Technology Law* 8 (August 2009), http://ssrn.com/abstract=1458162.

[7] It is improper to label all poor behaviors as malicious ones, as one can never judge the true intentions of an editor. Certainly some edits made in good faith are labeled and treated as vandalism or spam. The possibility of educating and rehabilitating these novices is one reason some care must be taken when identifying and blocking persistent attackers.

[8] John Seigenthaler, "A False Wikipedia `Biography'," *USA Today*, November 29, 2005, http://www.usatoday.com/news/opinion/editorials/2005-11-29-wikipedia-edit_x.htm.

[9] Andrew G. West, Sampath Kannan, and Insup Lee, "Detecting Wikipedia Vandalism via Spatio-Temporal Analysis of Revision Metadata" (paper presented at EUROSEC '10: The 3rd European Workshop on System Security, Paris, France, April 2010).

[10] Ibid.

[11] B. Thomas Adler et al., "Wikipedia Vandalism Detection: Combining Natural Language, Metadata, and Reputation Features" (paper presented at CICLing '11: The 12th International Conference on Intelligent Text Processing and Computational Linguistics, Tokyo, Japan, February 2011); Andrew G. West and Insup Lee, "Multilingual Vandalism Detection using Language-Independent & Ex Post Facto Evidence" (paper presented at PAN-CLEF '11: Uncovering Plagiarism, Authorship, and Social Software Misuse, Amsterdam, Netherlands, September 2011).

[12] Christopher Breneman and Cobi Carter, "ClueBot NG," http://en.wikipedia.org/wiki/User:ClueBot_NG.

[13] Ibid.

[14] Andrew G. West, "STiki: An Anti-Vandalism Tool for Wikipedia," software available from http://en.wikipedia.org/wiki/WP:STiki.

[15] Andrew G. West et al., "Link Spamming Wikipedia for Profit" (paper presented at CEAS '11: The 8th Annual Collaboration, Electronic Messaging, Anti-Abuse, and Spam Conference, Perth, Australia, September 2011).

[16] Ibid.

[17] Ibid.

[18] Andrew G. West et al., "Spamming for Science: Active Measurement in Web 2.0 Abuse Research" (paper presented at WECSR '12: The 3rd Workshop on Ethics in Computer Security Research, Bonaire, March 2012).

[19] Andrew G. West et al., "Autonomous Link Spam Detection in Purely Collaborative Environments" (paper presented at WikiSym '11: The 7th International Symposium on Wikis and Open Collaboration, Mountain View, CA, USA, October 2011).

[20] Youngsang Shin, Minaxi Gupta, and Steven Myers, "The Nuts and Bolts of a Forum Spam Automator" (paper presented at LEET '11: The 4th Workshop on Large-scale Exploits and Emergent Threats, Boston, MA, USA, March 2011).

[21] Brian Krebs, "Body Armor for Bad Websites," Krebs on Security weblog entry posted November 19, 2010, http://krebsonsecurity.com/2010/11/body-armor-for-bad-web-sites.

[22] Andrew G. West and Insup Lee, "What Wikipedia Deletes: Characterizing Dangerous Collaborative Content" (paper presented at WikiSym '11: The 7th International Symposium on Wikis and Open Collaboration, Mountain View, CA, USA, October 2011).

[23] Joe Merante, "UK National Portrait Gallery Threatens Wikipedia User Over Public Domain Images," Creative Commons weblog entry posted July 14, 2009, http://creativecommons.org/weblog/entry/15764.

[24] Jana Winter, "Wikipedia Distributing Child Porn, Co-founder Tells FBI," FoxNews.com, http://www.foxnews.com/scitech/2010/04/27/wikipedia-child-porn-larry-sanger-fbi/.

[25] http://en.wikipedia.org/wiki/Wikipedia:Contributor_copyright_investigations/Darius_Dhlomo.

[26] "Congressional Staff Actions Prompt Wikipedia Investigation," *Wikinews*, January 30, 2006, http://en.wikinews.org/wiki/Congressional_staff_actions_prompt_Wikipedia_investigation.

[27] "US Dept. of Justice IP address blocked after 'vandalism' edits to Wikipedia," *Wikinews*, April 29, 2008, http://en.wikinews.org/wiki/US_Dept._of_Justice_IP_address_blocked_after_'vandalism'_edits_to_Wikipedia; Kate Linthicum, "Wikipedia Blocks Access from Church of Scientology in L.A.," *LA Times*, June 5, 2009, http://articles.latimes.com/2009/jun/05/business/fi-wikipedia-scientology5.

[28] Exposure to inappropriate content is of greater concern to those in primary and secondary education. To this end, there are ongoing attempts to create and distribute a static "snapshot" of wikis/Wikipedia which are known to be vandalism and damage free (http://meta.wikimedia.org/wiki/Offline_Projects).

[29] For example, it has been observed that vandalism rates decline when schools are typically not in session (both daily and aligning with school holidays). Moreover, some vandalism explicitly mentions its educational origins ("Why is my professor making me write an essay on this?"), and it is not difficult to imagine how some of the most frequently vandalized pages are relevant to the curriculum (e.g., classic English literature).

[30] Dennis Carter, "Universities Told Not to Edit Their Wikipedia Entries," *eCampusNews*, April 14, 2011, http://www.ecampusnews.com/technologies/universities-told-not-to-edit-their-wikipedia-entries/.

[31] "QWiki," http://qwiki.stanford.edu/.

[32] Andrew G. West, "WikiAudit: Examining Organizational Contributions to Wiki Environments," Software available from http://en.wikipedia.org/wiki/WP:WikiAudit.

[33] Note that if a user registers and creates an account and username, their IP address is treated as private information and the user will not be included in any IP-based analyses. Fortunately, our technique remains widely applicable, as about 85% of damaging edits are made by IP users (see West, Kannan, and Lee, "Detecting Wikipedia Vandalism").

[34] "Wikipedia Campus Ambassador," http://outreach.wikimedia.org/wiki/Wikipedia_Campus_Ambassador.

[35] Marti Motoyama et al., "Re: CAPTCHAs—Understanding CAPTCHA-Solving Services in an Economic Context" (paper presented at the 19th USENIX Security Symposium, Washington DC, USA, August 2010).