Securing Wiki Platforms Against Malicious Contributions

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

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

Abstract

As wikis become increasingly prevalent more attention needs directed towards the security properties of the platform. Our thesis intends to identify attack vectors and use-cases against wikis (*e.g.,* vandalism, spam, copyright violations, *etc.*), characterize their nature, and mitigate their negative effects.

Wikis pose an interesting set of security challenges even relative to the Web 2.0 functionality they build upon. With open editing permissions and minimal barriers-to-entry wikis invite a diverse set of attackers. Moreover, wikis' community driven nature means solutions must address not just technical considerations but also the social climate in which they reside.

Research Overview

While wikis share certain collaborative functionality with other Web 2.0 applications (*e.g.,* blog comments, user-generated content sites, *etc.*) they are a different paradigm both in both technical and practical terms. Although related work in these domains does prove relevant, it is more crucial to understand how wikis differentiate themselves from these technologies: such novelty can become a point of leverage for both attack and defense purposes.

Analysis has generally begun by identifying a novel security issue with which a wiki community struggles. After quantifying the prevalence of the problem, focus shifts to finding statistical indicators that distinguish constructive and unconstructive/malicious contributions.

Our approach to finding these indicators (or "features" in machine learning terms) has been somewhat structured. We prefer features that are metadata driven and therefore language independent. In this manner the models we produce tend to be both computationally-efficient and portable. Additionally, we advocate the use of *reputation* features; concise quantifications of the behavioral history of some entity (*e.g.,* a user or article). In the absence of entity specific data we rely on spatial relationships to gain a degree of predictive capability.

Given its popularity, English Wikipedia has been a leading case study in our investigation of these issues. This has provided an opportunity to practically implement our systems, rather than simply evaluating performance offline. Such efforts have not only benefited that community but also provided us (as researchers) system feedback and a glimpse into user-base dynamics.

Completed Research

Vandalism: Intentionally malicious content modifications, often characterized by profanity and mass deletion, are termed *vandalism*. We first described a metadata and reputation based approach to vandalism detection [CV-11].

Note that all references of the form [CV-*] point to the curriculum vitae which proceeds this document.

Then we collaborated with other authors to combine and compare our differing approaches [CV-9]. These results led to models whose portability was demonstrated in multilingual settings [CV-6]. These strategies have been encoded into a practical GUI tool with server-side support, STiki [CV-22], which has been used to undo nearly 100,000 instances of vandalism from English Wikipedia.

Spam: Relative to the immaturity of vandalism we imagined link spammers would be well incentivized and therefore exhibit sophisticated attack vectors. A measurement study suggested this was not the case, but did allow us to speculate about the economic viability of certain vulnerabilities [CV-4]. Motivated by this we extended our anti-vandal models for the purpose of link spam detection, addressing both novel and status quo scenarios [CV-2].

Other damage: Some types of damage are interesting but less prevalent. One example is "dangerous content"; edits that carry legal liability (*e.g.,* copyright violations, libel) [CV-3]. Another publication [CV-14] looks at author bias and describes a contribution auditing tool. That writing views "damage" from a new perspective by considering threats to editors and the organizations they represent.

Other wiki relevant: Other publications involve wikis but are less damage focused. For example, we have written about the ethical perils of studying wiki security [CV-1] and rather exhaustively surveyed different trust and reputation schemes in collaborative settings [CV-15]. Presentations at practical (non-academic) venues exemplify the desire for our work to have practical ramifications. [CV-24,25,27].

Reputation fundamentals: External to wikis our research has frequently involved the practical application of *reputation* systems. Their use in domains like email spam [CV-10], Javascript mashups [CV-7], and the BGP protocol [CV-8] inspired and/or extended the reputation work done in wikis.

Active/Future Work

Several additional research ideas have been imagined or are already in progress. Two reside wholly in the "wiki" domain: (1) Language-independent models for damage detection could be ported to code repositories to measure author quality. (2) Following [CV-3], plagiarism detection algorithms could be harnessed to recognize copyright issues.

Other ideas are broadly applicable, while remaining wiki relevant: (1) CAPTCHAs are a first-line of defense in many collaborative settings. By differentiating between typical and spam-like user patterns a provider could utilize reputation to serve CAPTCHAs of dynamic difficulty. (2) Hyperlinks posted in public settings could reveal private information (*e.g.,* referrer data in PHP parameters). Autonomous removal of unnecessary data would heighten user security. (3) URLs are vetted primarily when they are posted. While a URL may remain static, the page behind can change (possibly to a malicious one). Research would investigate concise and scalable solutions to addressing such "link rot".

Symposium Expectations

Foremost, we hope participation in the doctoral symposium will provide a dialogue that helps in resolving weaknesses of the dissertation and its underlying research. For instance, it remains unclear to what extent the dissertation should, (1) focus exclusively on wiki environments (perhaps excluding some interesting opportunities), versus (2) addressing collaborative security in general (despite a lack of analysis in non-wiki environments).

Moreover, while prior publications and new ideas are not in short supply, a stronger "story" is needed to tie them together. If such a story were in place, this would simplify the decision of which future work should be pursued.

Finally, we look forward to gaining an outside perspective on our body of research from an audience that is both knowledgeable and academically diverse.

Biographical Sketches

Andrew G. West

Andrew G. West is a Ph.D. candidate in the Department of Computer and Information Science at the University of Pennsylvania in Philadelphia, PA. He received his MSE from Penn in 2010 and did his undergraduate study at Washington & Lee University in Lexington, VA receiving a B.S. in computer science in 2007.

Currently he works on the Quantitative Trust Management (QTM) project under the supervision of Insup Lee. His graduate work examines collaborative application security, and in particular, the detection of malicious behaviors in wiki environments. Other research interests include: underground economies, trust and reputation management, email spam, and the technical writing process.

Recent research highlights include the development of an anti-vandalism tool (STiki) that has been used in the removal of ≈100,000 unconstructive edits from English Wikipedia. His publication "*Link Spamming Wikipedia for Profit*" was awarded the best paper award at CEAS 2011 (Collaboration, Electronic messaging, Anti-abuse, and Spam). Another writing, "*What Wikipedia Deletes: Characterizing Dangerous Collaborative Content*" was recently featured by the Chronicle of Higher Education.

More can be learned about Mr. West and his research at his website, http://www.cis.upenn.edu/~westand. He can be contacted at `westand@cis.upenn.edu`.

Insup Lee

Insup Lee is the Cecilia Fitler Moore Professor in the Department of Computer and Information Science and Director of PRECISE Center at the University of Pennsylvania, where he has been since 1983. Professor Lee received the B.S. degree in mathematics from the University of North Carolina, Chapel Hill, and the Ph.D. degree in computer science from the University of Wisconsin, Madison.

Professor Lee's research focus is on cyber-physical systems, real-time computing, embedded systems, and medical device safety. More recently, he has gained interest in the trust/reputation management domain. Techniques have proven fruitful in his traditional areas of work, as well as in securing distributed systems and Web 2.0 platforms.

More can be learned about Professor Lee and his research at his website, http://www.cis.upenn.edu/~lee. He can be contacted at `lee@cis.upenn.edu`.

Curriculum Vitae of Andrew G. West

westand@cis.upenn.edu — http://www.cis.upenn.edu/~westand

Education

Doctoral – University of Pennsylvania (Philadelphia, PA)

- Ph.D. candidate in Computer and Information Science, 2012/2013 anticipated graduation.

Master's – University of Pennsylvania (Philadelphia, PA)

- Master of Science in Engineering – Computer and Information Science, Spring 2010.

Undergraduate – Washington and Lee University (Lexington, VA)

- Bachelor of Science with honors – Computer Science, Spring 2007.

Research Interests

- Collaborative application security, spam/vandalism mitigation, cyber economics, reputation management, trust management, network security, social computing, applied machine learning, computer science education, technical writing.

Publications

Conference and Workshop Papers

[1] "Spamming for Science: Active Measurement in Web 2.0 Abuse Research". Andrew G. West, Pedram Hayati, Vidyasagar Potdar, and Insup Lee. In *WECSR '12: Proc. of the Third Workshop on Ethics in Computer Security Research*, Bonaire. March 2012.

[2] "Autonomous Link Spam Detection in Purely Collaborative Environments". Andrew G. West, Avantika Agrawal, Phillip Baker, Brittney Exline, and Insup Lee. In *WikiSym '11: Proceedings of the Seventh International Symposium on Wikis and Open Collaboration*, pp. 91–100, Mountain View, CA, USA. October 2011.

[3] "What Wikipedia Deletes: Characterizing Dangerous Collaborative Content". Andrew G. West and Insup Lee. In *WikiSym '11: Proceedings of the Seventh International Symposium on Wikis and Open Collaboration*, pp. 25–28, Mountain View, CA, USA. October 2011.

[4] "Link Spamming Wikipedia for Profit". Andrew G. West, Jian Chang, Krishna Venkatasubramanian, Oleg Sokolsky, and Insup Lee. In *CEAS '11: Proceedings of the Eighth Annual Collaboration, Electronic Messaging, Anti-Abuse, and Spam Conference*, pp. 152–161 , Perth, AUS. September 2011. (co-Best Paper Award).

[5] "Towards the Effective Temporal Association Mining of Spam Blacklists". Andrew G. West and Insup Lee. In *CEAS '11: Proceedings of the Eighth Annual Collaboration, Electronic Messaging, Anti-Abuse, and Spam Conference*, pp. 73–82, Perth, AUS. September 2011.

[6] "Multilingual Vandalism Detection using Language-Independent & Ex Post Facto Evidence". Andrew G. West and Insup Lee. In *PAN-CLEF '11: Notebook Papers on Uncovering Plagiarism, Authorship, and Social Software Misuse*, Amsterdam. Sept. 2011.

[7] "ToMaTo: A Trustworthy Code Mashup Development Tool". Jian Chang, Krishna K. Venkatasubramanian, Andrew G. West, Sampath Kannan, Oleg Sokolsky, Myuhng Joo Kim, and Insup Lee. In *MASHUPS '11: Proceedings of the 5th International Workshop on Web APIs and Service Mashups*, Lugano, Switzerland. September 2011.

[8] "AS-TRUST: A Trust Quantification Scheme for Autonomous Systems in BGP". Jian Chang, Krishna Venkatasubramanian, Andrew G. West, Sampath Kannan, Boon Thau Loo, Oleg Sokolsky, and Insup Lee. In *TRUST '11: Proc. of the 4th International Conference on Trust and Trustworthy Computing, LNCS 6740*, pp. 262–276, Pittsburgh, PA, USA. June 2011. (A preliminary version was published as Tech. Report UPENN-MS-CIS-10-25).

[9] "Wikipedia Vandalism Detection: Combining Natural Language, Metadata, and Reputation Features". B. Thomas Adler, Luca de Alfaro, Santiago M. Mola-Velasco, Paolo Rosso, and Andrew G. West. In *CICLing '11: Proceedings of the 12th International Conference on Intelligent Text Processing and Computational Linguistics, LNCS 6609*, pp. 277-288. Tokyo, Japan. February 2011.

[10] "Spam Mitigation using Spatio-Temporal Reputations from Blacklist History". Andrew G. West, Adam J. Aviv, Jian Chang, and Insup Lee. In *ACSAC '10: Proceedings of the 26th Annual Computer Security Applications Conference*, pp. 161–170. Austin, TX, USA. Dec. 2010. (A preliminary version was published as Technical Report UPENN-MS-CIS-10-04).

[11] "Detecting Wikipedia Vandalism via Spatio-Temporal Analysis of Revision Metadata". Andrew G. West, Sampath Kannan, and Insup Lee. In *EUROSEC '10: Proceedings of the Third European Workshop on System Security*, pp. 22–28. Paris, France. April 2010. (A preliminary version was published as Technical Report UPENN-MS-CIS-10-05).

[12] "QuanTM: A Quantitative Trust Management System". Andrew G. West, Adam J. Aviv, Jian Chang, Vinayak S. Prabhu, Matt Blaze, Sampath Kannan, Insup Lee, Jonathan M. Smith, and Oleg Sokolsky. In *EUROSEC '09: Proceedings of the Second European Workshop on System Security*, pp. 28–35. Nuremberg, Germany. March 2009.

Book Chapters, Journal Papers, and Periodicals

[13] "Analyzing and Defending Against Web-based Malware". Jian Chang, Krishna K. Venkatasburamanian, Andrew G. West, and Insup Lee. To appear in an forthcoming issue of *ACM Computing Surveys*, 2012.

[14] "Open Wikis and the Protection of Institutional Welfare". Andrew G. West and Insup Lee. *Research Bulletin, EDUCAUSE Center for Applied Research*, Boulder, CO, USA, Feb. 2012.

[15] "Trust in Collaborative Web Applications". Andrew G. West, Jian Chan, Krishna K. Venkatasubramanian, and Insup Lee. *Future Generation Computer Systems, special section on Trusting Software Behavior*, Elsevier Press. (Based in part on Technical Report UPENN-MS-CIS-10-33. Publication pending).

[16] "An Evaluation Framework for Reputation Management Systems". Andrew G. West, Insup Lee, Sampath Kannan, and Oleg Sokolsky. Book chapter in *Trust Modeling and Management in Digital Environments: From Social Concept to System Development* (Zheng Yan, ed.), pp. 282–308. Information Science Reference, Hershey, PA, USA, 2010.

Technical Reports

[17] "Calculating and Presenting Trust in Collaborative Content". Andrew G. West. *Technical Report MS-CIS-10-33, University of Pennsylvania, Department of Computer and Information Science*, October 2010 (In partial fulfillment of the WPEII requirement).

[18] "AS-TRUST: A Trust Characterization Scheme for Autonomous Systems in BGP". Jian Chang, Krishna K. Venkatasubramanian, Andrew G. West, Sampath Kannan, Boon Thau Loo, Oleg Sokolsky, and Insup Lee. *Technical Report MS-CIS-10-25, University of Pennsylvania, Department of Computer and Information Science*, August 2010.

[19] "AS-CRED: Reputation Service for Trustworthy Inter-domain Routing". Jian Chang, Krishna K. Venkatasubramanian, Andrew G. West, Sampath Kannan, Insup Lee, Boon Thau Loo, and Oleg Sokolsky. *Technical Report MS-CIS-10-17, University of Pennsylvania, Department of Computer and Information Science*, April 2010.

[20] "Detecting Wikipedia Vandalism via Spatio-Temporal Analysis of Revision Metadata". Andrew G. West, Sampath Kannan, and Insup Lee. *Technical Report MS-CIS-10-05, University of Pennsylvania, Dept. of Computer and Information Science*, February 2010.

[21] "Mitigating Spam Using Spatio-Temporal Reputation". Andrew G. West, Adam J. Aviv, Jian Chang, and Insup Lee. *Technical Report MS-CIS-10-04, University of Pennsylvania, Department of Computer and Information Science*, February 2010.

Demonstrations, Tutorials, and Posters

[22] "STiki: An Anti-Vandalism Tool for Wikipedia Using Spatio-Temporal Analysis of Revision Metadata". Andrew G. West, Sampath Kannan, and Insup Lee. Formal demonstration. In *WikiSym '10: Proceedings of the Sixth International Symposium on Wikis and Open Collaboration*, Gdańsk, Poland. July 2010.

[23] "Spatio-Temporal Analysis of Wikipedia Metadata and the STiki Anti-Vandalism Tool". Andrew G. West, Sampath Kannan, and Insup Lee. Poster. In *WikiSym '10: Proceedings of the Sixth Intl. Symposium on Wikis and Open Collaboration*, Gdańsk, Poland. July 2010.

Presentations and Talks (without proceedings)

[24] "Anti-Vandalism Research: The Year in Review". Presented at *Wikimania '11: The International Wikimedia Conference*. Haifa, Israel. August 2011.

[25] "Autonomous Detection of Collaborative Link Spam". Presented at *Wikimania '11: The International Wikimedia Conference*. Haifa, Israel. August 2011.

[26] "Protecting Wikipedia from Vandalism, Spam, and Dangerous Content". Presented at *Columbia University Security Seminar*. New York, NY, USA. March 2011.

[27] "Spatio-Temporal Analysis of Revision Metadata and the STiki Anti-Vandalism Tool". Presented at *Wikimania '10: The Intl. Wikimedia Conference*. Gdańsk, Poland. July 2010.

[28] "An Introduction to LaTeX". Presented at *UPenn, School of Engineering and Applied Science, Technical Communication Program*. Philadelphia, PA, USA. October 2008.

Undergraduate Writings

[29] "Bound Optimization for Parallel Quadratic Sieving Using Large Prime Variations". Andrew G. West. *Undergraduate Honor's Thesis, Washington & Lee University*. May 2007.

[30] "Optimized Parallel Implementation of the Quadratic Sieve Factorization Algorithm". Andrew G. West. *Washington and Lee Journal of Science*, 8(2):25–26, 2007.

Media Attention

- "What Wikipedia Deletes, and Why". Alexandra Rice. *The Chronicle of Higher Education: The Wired Campus*, October 2011.
- "Link Spam on Wikis: Attack Models and Mitigation". Invited contribution to *Follow the Crowd (crowd-computing research blog)*, October 2011.
- "Content Redaction on Wikipedia: Copyright is Biggest Threat". Invited contribution to *Follow the Crowd (crowd-computing research blog)*, October 2011.
- "Link Spam Research with Controversial Genesis but Useful Results". *Wikimedia Research Newsletter*, 1(3), September 2011.
- "Deleted Revisions in the Wikipedia". *Wikimedia Research Newsletter*, 1(2), August 2011.
- "Vandalism Detectors Collaborate". *Wikipedia Signpost*, 7(8), February 2011.

Program Committees & Reviewing

Program Committees

- TrustID '12: 2nd Intl. Symposium on Trust and Identity in Mobile Internet, Computing, and Communications (symposium of IEEE TrustCom). June 2012. Liverpool, UK.
- iThings '12: The IEEE International Conference on Internet of Things. Track: *Reliability, Security, Privacy, and Trust*. September 2012. Besançon, France.
- TrustID '11: 1st Intl. Workshop on Trust and Identity in Mobile Internet, Computing, and Communications (workshop of IEEE TrustCom). November 2011. Changsha, China.
- Joint special track for UIC '10 (7th International Conference on Ubiquitous Intelligence and Computing) and ATC '10 (Autonomic and Trusted Computing). Track: *Pervasive Social Computing*. October 2010. Xi'an, China.

Guest and Delegated Conference/Workshop Reviews

- WH '11: Conference on Wireless Health
- ICCAD '11: Conference on Computer-Aided Design
- ISORC '10: Symposium on Object/Component/Service Real-time Distributed Computing
- ICCPS '10: Conference on Cyber-Physical Systems
- RV '09: Workshop on Runtime Verification
- LCTES '09: Conference on Languages, Compilers, and Tools for Embedded Systems
- ICTAC '09: Colloquium on Theoretical Aspects of Computing
- ATVA '09: Symposium on Automated Technology for Verification and Analysis

Guest and Delegated Journal/Book Reviews

- *IET Information Security* (IET-IFS). IET Press, 2012.
- *Security and Communication Networks* special issue on Spam, Phishing, and Countermeasures for Undesirable Electronic Communications. Wiley, 2012.
- *Future Generation Computer Systems* (FGCS) special section on Trusting Software Behavior. Elsevier Press, 2011.
- *Transactions on Intelligent Systems and Technology* (TIST) special issue on Search and Mining User Generated Content. ACM Press, 2011.
- *Trust Modeling and Management in Digital Environments*. IGI Global Press, 2010.

Teaching Experience

- CIS400 – TA – Senior Design Projects (1^{st} half) – *Fall 2008, 2009, 2010, 2011.*
- CIS401 – TA – Senior Design Projects (2^{nd} half) – *Spring 2009, 2010, 2011, 2012.*

Institutional Involvement (Penn)

- GSCSG Chair – (Graduate Student Computer Science Group)
  Social planning and organization at department granularity. *Spring 2009 until present.*
- GSEG Representative – (Graduate Student Engineering Group)
  Student advocacy and programming at college granularity. *Fall 2008 until present.*
- CIS Happy Hour Coordinator – Bi-weekly planning. *Fall 2009*, *Fall 2010 until present.*

Relevant Employment (External to Academia)

Alsos Digital Library (*Summer 2006*) – Database Programmer/Engineer

- SQL and .NET programming in support of the Alsos Project (`http://alsos.wlu.edu`)

WV-DEP (*Summer 2005*) – Website Design/Computer Assistant

- Website and media design, MS Visual-Basic programming, AutoCAD plotting

Professional and Social Memberships

- ACM (Association for Computing Machinery) – Professional organization
- ΠΜΕ (Pi Mu Epsilon) – Math honorary fraternity
- ΣN (Sigma Nu) – Social fraternity

Awards and Honors

- **2012:** Partial scholarship to attend *Wikimania 2012* (Washington D.C. - USA) – "... criteria are: level of activity within Wikimedia ... other free knowledge projects ... [and] future goals for participating in the Wikimedia movement."
- **2012:** Invitation and scholarship to attend the Google Graduate Researchers in Academia of Diverse backgrounds (GRAD) CS Forum (Mountain View, CA).
- **2011:** First place, PAN-CLEF 2011 Wikipedia vandalism detection competition – Winning approach for all language editions (German, English, and Spanish corpora).
- **2011:** Partial scholarship to attend *Wikimania 2011* (Haifa, Israel) – "... selected based on your dedication and participation in the Wikimedia movement or other free knowledge and educational initiatives and your potential to add great value ... going forward."
- **2003-2007:** J. Edward Lewis Scholarship – Merit-based scholarship providing full tuition, room, and board to Washington and Lee University (Lexington, VA).