Automated Analysis of General Data Protection Regulation Violations in WordPress Plugins
(Technical Report)

Preserving Digital Privacy in the Light of Big Data
(STS Research Paper)

A Thesis Prospectus Submitted to the

Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements of the Degree
Bachelor of Science, School of Engineering

Patrick Thomas
Spring, 2021

Technical Project Team Members
Yinzhi Cao
Michelangelo van Dam
Mingqing Kang
Faysal Hossain Shezan
Zihao Jerry Su
Yuan Tian
Erwin Wijaya

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Signature _____ Date _____
Patrick Thomas

Approved _____ Date _____
Yuan Tian, Department of Computer Science

Approved _____ Date _____
Sean Ferguson, Department of Engineering and Society

Introduction

The global adoption of the General Data Protection Regulation (GDPR) has necessitated the role of manual code analysis and internal auditing to ensure GDPR compliance for websites and programs. The technical side of this project aims to automate the manual code analysis by making an automated analysis tool that identifies GDPR violations in WordPress plugins and PHP code. The tool judges the code based on certain GDPR principles: rightful data collection, notification of third-party data transmission, right to data deletion, and right to data access. This all hinges on the importance of digital privacy, which motives GDPR. Beyond this analysis tool and the GDPR, there are more ways that digital privacy is preserved and exploitive data collection stemmed. Thus, the STS research will be a survey of the current technological and sociological techniques available and methods in place to help preserve digital privacy in the wake of big data and the ongoing datafication of Internet users. This all will be contextualized via defining the core tenets of the digital privacy movement and seeing how those solutions relate to those tenets.

Technical Topic

The objective of the project is to build an automated analysis tool that analyzes WordPress plugins for potential GDPR violations (Radley-Gardner et al., 2016). This is motivated by the global adoption of GDPR and the growing need for personal data protections. Current solutions to this problem involve manual analysis and audit due to the evolving and dynamic nature of the problem. Should the project work perfectly, we will have a list of compliant and non-compliant plugins, allowing people to know what WordPress plugins protect your data and guarantee you certain data rights or not. This list informs and empowers users with knowledge without the need for an advanced knowledge of PHP and WordPress and also would provide WordPress

plugin developers actionable and specific instruction on how their plugin is not GDPR-compliant.

This project by first taking single WordPress plugin as an input and converting it into an abstract syntax tree (AST) via NAVEX (Alhuzali et al., 2018) and PHP Joern (Backes et al., 2017). Some code that is not PHP code, such as SQL, HTML, and JavaScript code, is converted into ASTs through other similar tools like Esprima (Hidayat, n.d.). This AST is then loaded into Neo4j (*Neo4j Graph Platform – The Leader in Graph Databases*, n.d.), a graph database management service that allows for efficient access to graph-based data like ASTs. Once a plugin's AST is loaded into Neo4j, we run the analysis tool that we built. The analysis tool looks for security, database, file storage, data deletion, and data transmission usages as well as how personally-identifying data flows through the plugin. Once we know how data flows to data storage or transmission usages and if the data is secured (encrypted or hashed) before storage or transmission, we are then able to determine if the plugin violates GDPR based on the tenets around data encryption, data access, data deletion, and data transmission to third parties. We consider the project a success if our tool can identify some number of GDPR violations in the WordPress plugins hosted online.

Due to the nature of the project, it requires little new physical hardware. We programmed using free development environments, used free software and libraries or products of other publicly funded research, and tested on mostly free and open-source WordPress plugins. Some premium versions of WordPress plugins offering GDPR compliance we purchased for manual analysis and verification of our program.

The project is a joint effort between the University of Virginia (Yuan Tian, Faysal Hossain Shezan, Zihao Jerry Su, Erwin Wijaya, and myself), Johns Hopkins University (Yinzhi Cao

and Mingqing Kang), and Michelangelo van Dam, a professional software engineer and expert on PHP, WordPress, and more.

My primary responsibility within the team involved writing code to detect the individual lines of code in the AST that secured data, stored data, or transmitted data to a third party. Determining security methods required some background research on the current state-of-the-art methods for securing data via encryption and hashing, and ultimately we decided that the standards put forward via the Cryptographic Module Validation Program (CMVP) (Computer Security Division, 2016) should be considered secure for WordPress plugins. Data storage and transmission identification were derived from the multitude of resources on PHP and WordPress online such as the PHP official documentation (*PHP: Documentation*, n.d.) and the WordPress official documentation (*WordPress Codex*, n.d.). My work interfaced with Jerry's, whose work was to track the flow of data personally-identifying data within a program. My detection algorithms were then used to check if personally-identifiable data was tracked to an encryption usage and then storage or transmission usage or no encryption at all. My work also included making the program judge the output of the entire program, deciding what data flows were insecure or secure, and deciding if a plugin as a whole violates the GDPR. Aside from the analysis program itself, I also dealt with a lot of the project configuration and running the program on the thousands of WordPress plugins available from the WordPress Plugin Directory (*WordPress Plugins*, n.d.). I was the person primarily in charge of writing the scripts to automate the process of running tools to convert a program into an AST, load the AST into the graph database software, and then run our analysis program on the AST. Furthermore, I wrote the scripts to upscale this process to run on the thousands of WordPress plugins mentioned previously.

STS Topic

The continued and increasing datafication of Internet users is a growing and severe digital privacy issue (Mai, 2016). Privacy is of vital importance since it supports democracy, works against categorizations that can lead to discrimination and unfair treatment, and much more (Magi, 2011). Data brokers, advertisers, and cloud platforms are increasingly monetizing and carelessly using users' data under the guise of "providing better services", when in fact their main goal is to leverage user data to the platform's sole advantage. Users of these services are ill-equipped with technological and sociological solutions against the datafication of their private information, and users' information used against themselves in targeted ads and manipulation (King, 2019; Kramer et al., 2014). In response to this growing problem, there has been an increase in focus on digital privacy by users and researchers alike. What solutions, both technical and social, exist to help combat this problem?

The datafication of users and the rise in importance of digital privacy is well documented and often researched. It is well understood today that Internet advertisements and hidden trackers are embedded in almost all popular websites and even emails. Google Analytics, for example, is embedded in more than 29 million websites as of October of 2020 (*Google Analytics Usage Statistics*, n.d.). Internet usage can be further tracked beyond websites themselves, and the lookup of website addresses themselves in DNS queries can and are tracked by Internet Service Providers (ISPs) (Berman, 2019). Basic methods to prevent Internet tracking through adblockers and similar software is not entirely effective, and websites like Facebook are still able to track what pages you are interested in within Facebook and use that data to inform targeted advertising.

There exists a multitude of technological methods to counter datafication and to preserve digital privacy. This requires a deeper understanding of why privacy is important; this is well

outlined and organized by Magi (Magi, 2011). Towards understanding the technological solutions, we need to determine several factors about each one. This includes reasons for a technological solution's adoption or non-adoption, what it aims to fix, and the intended and unintended outcomes of the solution. An example of this is adblocking plugins for browsers. We can see that they clearly block ads, but they have also started a technological and on-going arms race between adblocking and anti-adblocking techniques (Nithyanand et al., 2016). It is also known that personalized advertising on websites is influencing the usage and growth of adblocking (Brinson et al., 2018). The outcomes of any single particular solution need to be analyzed to see if it conforms with the ideals of digital privacy; does the solution empower individuals and protect those from exploitation, or does it simply stop the problem from progressing? That is, does any one particular solution do more than just mitigate the problem?

Sociological solutions are also a large focus, as remedying the underlying problem of datafication is better than temporarily stopping it through technical countermeasures. Sociological solutions would include new or improved pedagogical methods, laws and regulations, and advocacy groups. One example of a sociological solution, requiring engineering ethics courses in undergraduate programs is a possible way to help reverse the exploitation of personal information, but current engineering ethics courses sometimes have the problem students and teachers perceive different parts of the course as quintessential (Holsapple et al., 2012). These solutions will be analyzed similarly to the technical ones, and both are to be analyzed based on how they appeal to digital privacy's ideals as defined by Magi (Magi, 2011). Sociological solutions, like their technological counterparts, are not instant or miracle fixes. However, both types of solutions can help remedy the current issues surrounding digital privacy.

As previously touched on, a wide array of technological and sociological fixes are to be analyzed. Typically, the current understanding of such fixes is to be gathered from scholarly sources. Technological solutions are well documented and researched, but sources for sociological solutions might stray to advocacy group websites and third-party sources. It would be very enlightening to speak with a digital privacy advocate as they hopefully are aligned with the goals of digital privacy and can provide many insights into the field. Ideally, we want to see all individual Internet users as the primary stakeholder, as it is all of their data that we want to protect. Naturally, other stakeholders would include those who build websites, such as website administrators, large and small technology companies, data brokers, and advertisers. We also need to consider the people involved with both technological and sociological solutions, like the professors, teachers, professional engineers, engineering students, experts in the field, and advocacy groups. I expect that the primary conflict is that the companies with the most to gain from troves of personal data are the ones to circumvent regulations, resist further regulations or change, and abuse personal data the most. Ultimately we want to answer what technological and sociological solutions exist to preserve digital privacy, and how do those solutions align with the egalitarian goals of digital privacy?

Conclusions

The technical portion of the project has almost entirely been completed already. Work started on the project at the end of May 2020 and is projected to end November 2020, so there is not much more technical work for the project. The technical analysis tool produced by the project does not directly improve digital privacy; it is not an adblocker or similar to other technical countermeasures. However, it does both inform users and inform plugin developers about how to improve their own work. This is similar to the method used by cybersecurity researchers

in which they discover vulnerabilities, disclose their knowledge to the responsible parties, and then disclose their results to the public should the responsible party not respond. The technical research provides a tool that informs digital privacy decisions in a limited space, the WordPress environment. Meanwhile, the STS research identifies solutions from a greater space, digital privacy as a whole. The STS research has the added benefit of providing more background on the technical research in that it should greatly inform the idea of digital privacy and how it particularly relates to the technical topic.

Between the Fall 2020 and Spring 2021 semesters, more research on the technical and STS topics should be done. Work on the technical project should be done by the beginning of the Spring semester. The bulk of the research on the STS topic needs to be completed, and then the results from this research can be wrapped back into the technical findings to form a holistic picture. The details of this are to be informed by STS 4600.

References

Alhuzali, A., Gjomemo, R., Eshete, B., & Venkatakrishnan, V. N. (2018). *NAVEX: Precise and Scalable Exploit Generation for Dynamic Web Applications*. 17.

Backes, M., Rieck, K., Skoruppa, M., Stock, B., & Yamaguchi, F. (2017). Efficient and Flexible Discovery of PHP Application Vulnerabilities. *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*, 334–349. https://doi.org/10.1109/EuroSP.2017.14

Berman, M. (2019, October 17). *What is ISP Tracking and Why Should I Be Concerned?* Programming Insider. https://programminginsider.com/what-is-isp-tracking-and-why-should-i-be-concerned/

Brinson, N. H., Eastin, M. S., & Cicchirillo, V. J. (2018). Reactance to Personalization: Understanding the Drivers Behind the Growth of Ad Blocking. *Journal of Interactive Advertising*, *18*(2), 136–147. https://doi.org/10.1080/15252019.2018.1491350

Computer Security Division, I. T. L. (2016, October 11). *Cryptographic Module Validation Program | CSRC | CSRC*. CSRC | NIST. https://content.csrc.e1a.nist.gov/Projects/Cryptographic-Module-Validation-Program

Google Analytics Usage Statistics. (n.d.). Retrieved October 25, 2020, from https://trends.builtwith.com/analytics/Google-Analytics

Hidayat, A. (n.d.). *Esprima*. Retrieved November 4, 2020, from https://esprima.org/

Holsapple, M. A., Carpenter, D. D., Sutkus, J. A., Finelli, C. J., & Harding, T. S. (2012). Framing Faculty and Student Discrepancies in Engineering Ethics Education Delivery. *Journal of Engineering Education*, *101*(2), 169–186. https://doi.org/10.1002/j.2168-9830.2012.tb00047.x

King, B. (2019, April 1). *Why Targeted Ads Are a Serious Threat to Your Privacy*. MakeUseOf. https://www.makeuseof.com/tag/targeted-ads-threat-privacy/

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788–8790. https://doi.org/10.1073/pnas.1320040111

Magi, T. J. (2011). Fourteen Reasons Privacy Matters: A Multidisciplinary Review of Scholarly Literature. *The Library Quarterly*, *81*(2), 187–209. https://doi.org/10.1086/658870

Mai, J.-E. (2016). Big data privacy: The datafication of personal information. *The Information Society*, *32*(3), 192–199. https://doi.org/10.1080/01972243.2016.1153010

Neo4j Graph Platform – The Leader in Graph Databases. (n.d.). Neo4j Graph Database Platform. Retrieved October 22, 2020, from https://neo4j.com/

Nithyanand, R., Khattak, S., Javed, M., Vallina-Rodriguez, N., Falahrastegar, M., & Powles, J. E. (2016). Adblocking and Counter-Blocking: A Slice of the Arms Race. *University of Cambridge*, 7.

PHP: Documentation. (n.d.). Retrieved October 22, 2020, from https://www.php.net/docs.php

Radley-Gardner, O., Beale, H., & Zimmermann, R. (Eds.). (2016). *Fundamental Texts On European Private Law*. Hart Publishing. https://doi.org/10.5040/9781782258674

WordPress Codex. (n.d.). Retrieved October 22, 2020, from https://codex.wordpress.org/

WordPress Plugins. (n.d.). WordPress.Org. Retrieved October 22, 2020, from https://wordpress.org/plugins