Detecting Privacy Threats with Machine Learning: A Design Framework for Identifying Side-Channel Risks of Illegitimate User Profiling

Raja Hasnain Anwar
University of Arizona, rajahasnainanwar@arizona.edu

Yi (Zoe) Zou
University of Western Ontario, yzou@ivey.ca

Muhammad Taqi Raza
University of Arizona, taqi@arizona.edu

Detecting Privacy Threats with Machine Learning: A Design Framework for Identifying Side-Channel Risks of Illegitimate User Profiling

Emergent Research Forum (ERF) Paper

Raja Hasnain Anwar
University of Arizona
rajahasnainanwar@arizona.edu

Yi (Zoe) Zou
University of Western Ontario
yzou@ivey.ca

Muhammad Taqi Raza
University of Arizona
taqi@arizona.edu

Abstract

Privacy leakage has become prevalent and severe with the increasing adoption of the internet of things (IoT), artificial intelligence (AI), and blockchain technologies. Such data-intensive systems are vulnerable to side-channel attacks in which hackers can extract sensitive information from a digital device without actively manipulating the target system. Nevertheless, there is a scarcity of IS research on how businesses can effectively detect and safeguard against side-channel attacks. This study adopts the design science paradigm and lays the groundwork for systematic inquiry into the assessment of privacy risks related to side-channels. In this paper, we a) highlight the privacy threats posed by side-channel attacks, b) propose a machine learning-driven design framework to identify side-channel privacy risks, and c) contribute to the literature on privacy analytics using machine learning techniques. We demonstrate a use case of the proposed framework with a text classification model that uses keystroke timings as side-channel.

Keywords

Design science, privacy analytics, side-channel attacks, machine learning.

Introduction

With the ubiquitous use of AI-related computers, smart devices, and sensors, there are increasing concerns among IS professionals and scholars about issues surrounding cybersecurity and data privacy (Rai, 2017). Side-channel attacks have emerged as an increasing threat faced by both business owners and individual consumers. Distinct from other forms of cybersecurity threats where the integrity of an information system is directly compromised through unauthorized system access or malicious system tampering, side-channel attacks are non-invasive and passive. Standaert (2010) defines side-channel attacks as a way for adversaries to learn the physical specifications of the system and the characteristics of its users through externally observable phenomena such as timing information and power consumption. Therefore, side-channel adversaries try to identify unintended information leakage from the activities of connected IS devices. For example, the ZombieLoad attack disclosed by Schwarz et al. (2019) could exploit a design flaw inherent in Intel processors to extract sensitive information from a victim's computer without actively manipulating or modifying the target system.

While user information transmitted among different devices can be protected, hackers can still exploit the side-channel vulnerabilities to infer users' characteristics and activities without their awareness of the privacy leakage. Hackers can further use these characteristics to profile users to learn their behaviors and usage patterns. This is not only a breach of user privacy but can lead to further targeted attacks. Thus, side-channel attacks are posing a real threat to the confidentiality of user data. Although side-channel attacks have become a significant area of concern for businesses, there is still a paucity of IS research on how businesses can leverage specialized countermeasure techniques to safeguard against these attacks.

In this research, we adopt the design science research (DSR) paradigm to develop an IS design framework for applying machine learning techniques to identify side-channel risks in user-driven information systems. Once the privacy risks are identified, the framework can be extended to design a specialized mitigation strategy for side-channel attacks. We position our work as a design-oriented IS study. Following the research guidelines of computational genre and representation genre in DSR (Baskerville et al., 2015), we first delineate the main components of the framework and describe how this framework can guide the decisions throughout the entire processes of data gathering, pre-processing, modeling, and evaluation. Then, we demonstrate the utility of the design framework in the context of user profiling using their typing timings as a side-channel.

The present research contributes to the extant literature on IS security and cybersecurity analytics in three ways. First, our research findings help to shed a light on one important yet understudied domain of cybersecurity: side-channel information leakage. Specifically, little research has been done on the detection of side-channel risks and the assessment of their related business implications. The present research can advance our understanding of how IS should be designed to spot non-invasive and passive cybersecurity attacks. Second, our design framework offers a conceptual roadmap for identifying side-channel privacy risks with machine learning techniques. We also demonstrate how this framework can help to formulate an effective practical solution to addressing privacy and security concerns surrounding IS use. Lastly, our research contributes to the literature on cybersecurity analytics by demonstrating how machine techniques can be applied in the areas of detecting side-channel threats of privacy leakage.

Background

IS Perspective of Personal Privacy

Privacy is a well-versed topic in IS research. IS scholars have pondered over the conceptualization of privacy as a construct and its relationship with other related constructs (Smith et al., 2011). In the modern age, scholars have focused on the structural nature of privacy to study it on multiple levels – instituting the idea of individual (personal) privacy (Bélanger & Crossler, 2011). Today, the landscape of personal privacy research encompasses two main themes: users' perceived privacy concerns and users' behavior leading to privacy leakage. A review of recent research (Sunwoo et al., 2022) reveals that privacy is studied as a user-centric concept rather than a data-centric one.

With the rise of IoT- and AI-based systems, contemporary IS research has focused on personal privacy pertaining to users' concerns about personal data sharing (Cichy et al., 2021; Cheng et al., 2022). The practical implications from the extant literature are primarily oriented toward mitigating threats to privacy that are directly associated with unauthorized access or malicious tempering of systems. However, it has largely overlooked the threats of side-channel attacks and the effective ways to detect them.

Privacy Leakage with Side-channel Attacks

Side-channel attacks represent a unique and challenging threat to information security and privacy. In particular, they pose unprecedented risks to the confidentiality of sensitive information processed by enterprise computers and servers, and thereby raise a genuine concern for IS practitioners (Kocher et al., 2019). Due to their non-invasive nature, side-channel attacks are more difficult to detect compared to traditional IS security attacks (Standaert, 2010; Sun et al., 2014; Yu et al., 2019). Moreover, side-channel attacks typically exploit unintended information leakage and often could be tailored to target specific vulnerabilities in an application. Thus, they are harder to mitigate using standard antivirus software (Zhou & Feng, 2005). Lastly, side-channel attacks are relatively easy for hackers to carry out than conventional IS security attacks, since they merely rely on the observations of a system's processing activities and do not require any forms of remote access and/or user interactions (Monaco, 2018).

Depending on the medium (channel) of the attack, there are three major types of side-channel attacks: electromagnetic analysis, power analysis, and timing attack (Lerman et al., 2011; Devi & Majumder, 2021). Attacks based on power and electromagnetic analysis observe the system's unintentional power consumption and electromagnetic emissions to infer its workload (Mangard et al., 2008; Sayakkara et al., 2019). In a timing attack, the attacker measures the timestamps of the occurrence of events and correlates the timings with the system's operations (Felten & Schneider, 2000). For example, Song et al. (2001) used keystroke timings to infer sensitive text like user passwords. We focus on timing attacks in user-driven systems for their capability to capture unique patterns from user-system interactions.

The power of machine learning (ML) techniques has shown great effectiveness to detect and mitigate privacy threats including side-channel attacks (Alam et al., 2017; Wu et al., 2022). For instance, previous work on keystroke dynamics shows how typing timings can be used to infer the typed text, even for an encrypted channel (Monaco, 2019). Furthermore, insights from multiple side-channels can be combined to extract user information from a multi-level system. Therefore, the integration of machine learning techniques in privacy analytics is crucial to detect and assess privacy risks in an advanced IS application.

Although the term "side-channel attack" is primarily used in computer science research, a handful of IS studies have discussed the privacy implications of side-channel attacks for user information inference (Han et al., 2021) and shadow profiling (Garcia, 2017). Despite the scarcity of IS research on detecting and assessing privacy risks in side-channels, the potential business costs and negative consequences of these attacks can be enormous and irreversible (Lee et al., 2011; Schwaig et al., 2013). As such, much research is needed to: a) develop an overarching design framework to guide the detection and assessment of side-channel privacy risks, and b) specify design principles for modeling side-channel data.

Proposed Design Framework

Our approach toward detecting privacy threats uses side-channel data to identify potential privacy leakage points. We propose a design framework for the systematic identification of side-channel risks (shown in Figure 1). The design framework outlines the complete workflow with the use-case of text classification and user profiling from a stream of temporal typing traces.

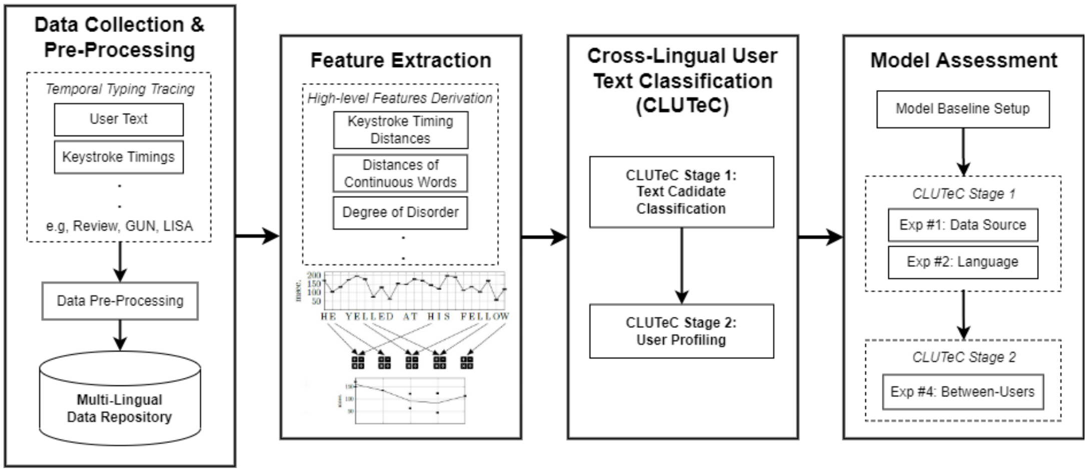

Figure 1: Framework for Identifying Side-channel Privacy Risks with ML Techniques

The key design goal is to use typing behaviors learned through keystroke timings and predict the typed text and the typist's characteristics. Following the recent work on keystroke dynamics research (e.g., Ahmed & Issa (2013)), we frame the text recognition problem as a classification among candidate texts whereby a machine learning model is trained to identify the text corresponding to a keystroke timing sequence. We developed a model called Cross-Lingual User Text Classification (CLUTeC) to infer user characteristics from their typing behaviors, independent of the typed language. The proposed design framework consists of four components:

1. **Data Collection & Pre-Processing:** Temporal typing traces are recorded in the form of key hold and flight times as the main data stream. Typed text and user attributes are collected as ground-truth for the training phase. For pre-processing, extra-long keystroke timings and special keys like backspace are removed -- representing pauses and hesitation.
2. **Feature Extraction:** High-level features are generated to extract deeper insights from the raw timings. These features include keystroke timing distances across words and degree of disorder.

3. **CLUTeC:** A two-stage machine learning model to learn the user's characteristics from the typing traces. This model is language-independent and only relies on the user's typing behavior.
4. **Model Assessment:** A multi-level assessment is conducted to test the privacy-leaking capabilities of the model at different scales. The test data is separated by the data source, typed language, and user id, to ensure generalizability in side-channel representation learning.

Use Case

Dataset Description. We adopt the extension of three datasets, LSIA, KM, and PRODOSY, from González, et al. (2021). The data is pre-processed to remove user hesitation, pauses, and typing corrections. The dataset also contains reconstructed keystroke timings using finite context modeling (FCM) (González & Calot, 2015). The reconstruction allows auto-generated keystroke timings for arbitrary text sequences without the need for a typist. We use FCM to create candidate texts similar to the target text for the robustness evaluation of our model.

Feature Extraction. Initially, only two features are available for each text sample, i.e., flight time and hold time as time series. Previous research has used distances, degree of disorder, outlier counts, and empirical CDF-based distances for high-level feature derivation. We combine these approaches to derive a comprehensive feature set for classifier training. For this, we use Manhattan and Euclidean distances, R-index, Z-score, Manhattan and Euclidean-like CDF distances, and other distances proposed by Davoudi and Kabir (2009), Zhong et al. (2012), and Kaneko et al. (2011).

Machine Learning Model. We have set up a baseline using an SVM classifier trained on the base features, i.e., FT, HT, and distance-based measures. The model performance stands at 94% accuracy on the binary classification task, i.e., classifying text sequences as matching the time sequence or not. Currently, we are working with a set of 10 candidate text sequences for each time sequence. We plan to scale it to 100 sequences to make the task more challenging, and the resultant model more robust.

Conclusion & Future Directions

Privacy leakage is a serious concern among individuals and organizations. In this work, we highlight the importance of assessing the privacy threats caused by unintentional side-channels and call for more IS research on this critical issue. We not only explore the possibility of using data and machine learning techniques to detect previously unknown privacy risks, but also propose a comprehensive design framework to elaborate the key principles and processes of developing and calibrating an algorithm for detecting keystroking side-channels. We present the case of side-channel privacy leakage through a text typing use-case whereby typing behavior is used for user profiling. Our experiments for text classification show promising results. Further, we plan to develop a deep learning model for assessing user-specific characteristics that can be used for profiling through keystroke timings.

REFERENCES

Ahmed, A. A., & Traore, I. 2013. "Biometric recognition based on free-text keystroke dynamics," *IEEE transactions on cybernetics* (44:4), pp. 458-472.

Alam, M., Bhattacharya, S., Mukhopadhyay, D., & Bhattacharya, S. 2017. "Performance counters to rescue: A machine learning based safeguard against micro-architectural side-channel-attacks," *Cryptology ePrint Archive*.

Baskerville, R. L., Kaul, M., & Storey, V. C. 2015. "Genres of inquiry in design-science research," *MIS Quarterly* (39:3), pp. 541-564.

Bélanger, F., & Crossler, R. E. 2011. "Privacy in the Digital Age: A Review of Information Privacy Research in Information Systems," *MIS Quarterly* (35:4), pp. 1017–41.

Cheng, X., Su, L., Luo, X., Benitez, J., & Cai, S. 2022. "The good, the bad, and the ugly: Impact of analytics and artificial intelligence-enabled personal information collection on privacy and participation in ridesharing," *European Journal of Information Systems* (31:3), pp. 339-363.

Cichy, P., Salge, T. O., & Kohli, R. 2021. "Privacy Concerns and Data Sharing in the Internet of Things: Mixed Methods Evidence from Connected Cars," *MIS Quarterly* (45:4), pp. 1863-92.

Davoudi, H., & Kabir, E. 2009. "A new distance measure for free text keystroke authentication," In *Proceedings of the 14th international CSI computer conference*, IEEE, pp. 570-575.

Devi, M., & Majumder, A. 2021. "Side-channel attack in Internet of Things: A survey," in *Applications of Internet of Things: Proceedings of ICCCIOT*, Singapore: Springer, pp. 213-222.

Felten, E. W., & Schneider, M. A. 2000. "Timing attacks on web privacy," in *Proceedings of the 7th ACM Conference on Computer and Communications Security*, pp. 25-32.

Garcia, D. 2017. "Leaking privacy and shadow profiles in online social networks," *Science advances* (3:8).

González, N., & Calot, E. P. 2015. "Finite context modeling of keystroke dynamics in free text," in *Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG)*, IEEE, pp. 1-5.

González, N., Calot, E. P., Ierache, J. S., & Hasperué, W. 2021. "The reverse problem of keystroke dynamics: Guessing typed text with keystroke timings only," in *Proceedings of the International Conference on Electrical, Computer and Energy Technologies (ICECET)*, IEEE, pp. 1-6.

Han, X., Wang, L., & Fan, W. 2021. "Is Hidden Safe? Location Protection against Machine-Learning Prediction Attacks in Social Networks," *MIS Quarterly* (45:2), pp. 821-858.

Kaneko, Y., Kinpara, Y., & Shiomi, Y. 2011. "A hamming distance-like filtering in keystroke dynamics," In *Proceedings of the Ninth Annual International Conference on Privacy, Security and Trust*, IEEE, pp. 93-95.

Kocher, P., Horn, J., Fogh, A., Genkin, D., Gruss, D., Haas, W., ... & Yarom, Y. 2020. "Spectre attacks: Exploiting speculative execution," *Communications of the ACM*, (63:7), pp. 93-101.

Lee, D. J., Ahn, J. H., & Bang, Y. 2011. "Managing consumer privacy concerns in personalization: a strategic analysis of privacy protection", *MIS Quarterly*, pp. 423-444.

Lerman, L., Bontempi, G., & Markowitch, O. 2011. "Side channel attack: an approach based on machine learning," *Center for Advanced Security Research Darmstadt*.

Mangard, S., Oswald, E., & Popp, T. 2008. "*Power analysis attacks: Revealing the secrets of smart cards*", (Vol. 31), Springer Science & Business Media.

Monaco, J. V. 2018. "Sok: Keylogging side channels," in *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, IEEE, pp. 211-228.

Monaco, J. V. 2019. "What Are You Searching For? A Remote Keylogging Attack on Search Engine Autocomplete," In *Proceedings of the USENIX Security Symposium*, pp. 959-976.

Rai, A. 2017. "Editor's comments: Diversity of design science research," *MIS Quarterly* (41:1), pp. iii-xviii.

Sayakkara, A., Le-Khac, N.-A., & Scanlon, M. 2019. "A survey of electromagnetic side-channel attacks and discussion on their case-progressing potential for digital forensics," *Digital Investigation*, pp. 43-54.

Schwaig, K. S., Segars, A. H., Grover, V., & Fiedler, K. D. 2013. "A model of consumers' perceptions of the invasion of information privacy," *Information & management* (50:1), pp. 1-12.

Schwarz, M., Lipp, M., Moghimi, D., Van Bulck, J., Stecklina, J., Prescher, T., & Gruss, D. 2019. "ZombieLoad: Cross-privilege-boundary data sampling," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pp. 753-768.

Smith, H. J., Dinev, T., & Xu, H. 2011. "Information Privacy Research: An Interdisciplinary Review," *MIS Quarterly* (35:4), pp. 989–1015.

Song, D. X., Wagner, D. A., & Tian, X. 2001. "Timing analysis of keystrokes and timing attacks on ssh," In *Proceedings of the USENIX Security Symposium*.

Standaert, F.-X. 2010. "Introduction to side-channel attacks," in *Secure integrated circuits and systems*, Ingrid Verbauwhede, Boston: Springer, pp. 27-42.

Sun, Y., Zhang, J., Xiong, Y., & Zhu, G. 2014. "Data security and privacy in cloud computing," *International Journal of Distributed Sensor Networks*, (10:7), pp. 190-903.

Sunwoo, P., Jeongyun, B., Yeajoo, Y., & Dongwhan, K. 2022. "A Study on the Users' Response to Privacy Issues in Customized Services," *Journal of Multimedia Information System* (9:3), pp. 201-208.

Wu, Y., Hassan, M., & Hu, W. 2022. "Safegait: Safeguarding gait-based key generation against vision-based side channel attack using generative adversarial network," in *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* (6:2), pp. 1-27.

Yu, J., Lu, L., Chen, Y., Zhu, Y., & Kong, L. 2019. "An indirect eavesdropping attack of keystrokes on touch screen through acoustic sensing," *IEEE Transactions on Mobile Computing* (20:2), pp. 337-351.

Zhong, Y., Deng, Y., & Jain, A. K. 2012. "Keystroke dynamics for user authentication," In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition workshops*, IEEE, pp. 117-123.

Zhou, Y., & Feng, D. 2005. "Side-channel attacks: Ten years after its publication and the impacts on cryptographic module security testing," *Cryptology ePrint Archive*.