Detecting Personally Identifiable Information Violations Attempts in an Email Provider using Probability Density Function

ANA H. VALENTIN

DOCTORATE IN SCIENCE – CYBERSECURITY STUDENT

MARYMOUNT UNIVERSITY

LITERATURE REVIEW

MARYMOUNT UNIVERSITY

Researchers	Applicability of Regular Expressions
Han, Katoen, and Berteun (2009)	Easy to understand, and very compact to apply successive state elimination
Punithan, Kim, Kim, and Choi (2016)	Represent prevalent and complex attacks
Ko, Jung, Han, and Burgstaller (2014)	Allow the specification of a potentially infinite set of strings (or patterns)
Alazab, Hobbs, Abawajy, Khraisat, and Alazab (2014)	Deploy signatures that would represent all mutants of a pertinent attack to detect high false-positive
Han, Katoen, and Berteun (2009) and Ko, Jung, Han, and Burgstaller (2014)	Deterministic finite automata technique to find and convert a regular expression
Singh, Kumar, Singla, & Ketti (2017)	Use memory-based architecture to speed up

Methods – Abbreviations and Acronyms

DLP = data loss prevention

PII = personnel identifiable information

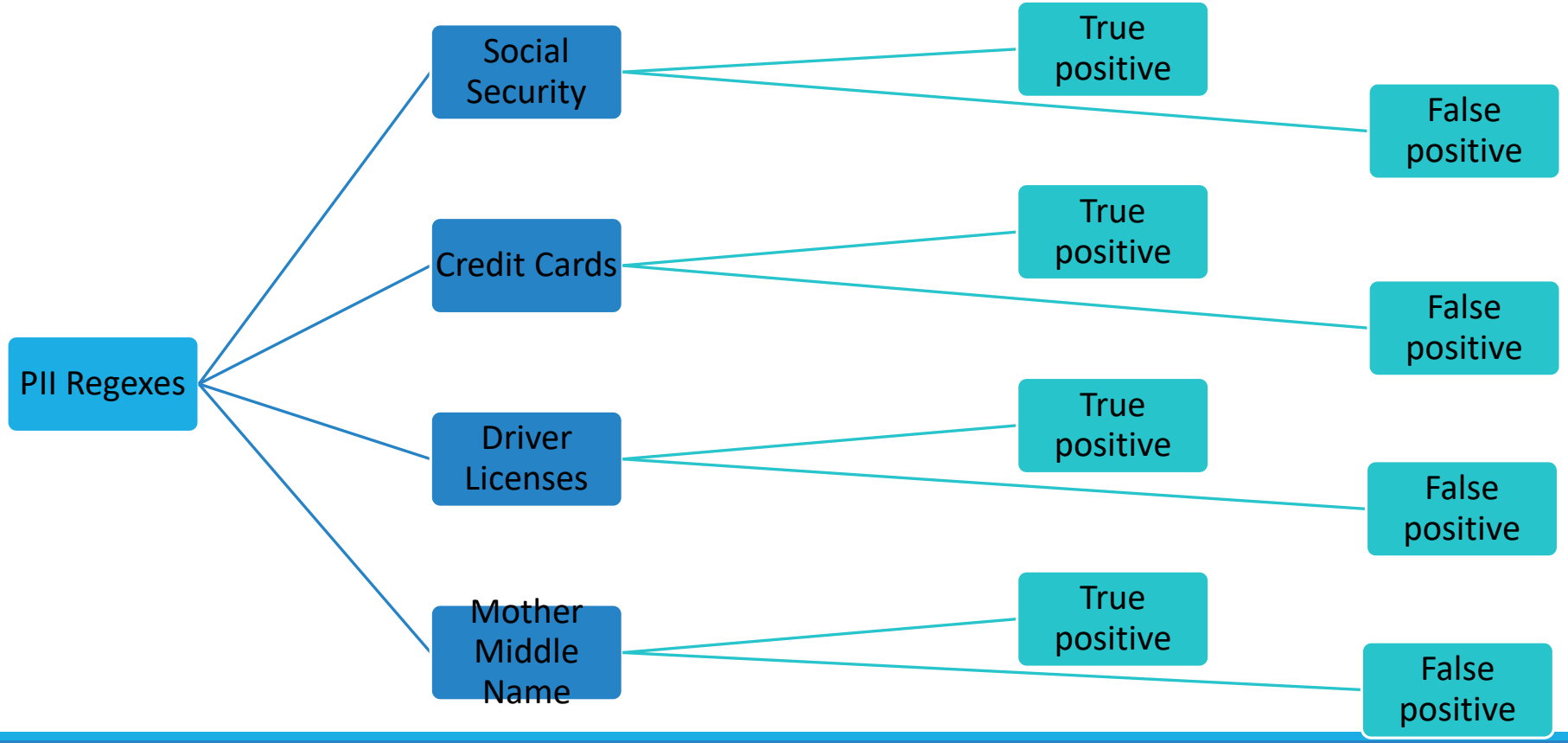
Regex = regular expressions represents a specification of a potentially infinite set of strings

E = base of natural logarithms (2.7183)

Λ = number of PII attempt violations detected using regular expressions

k = number of days in which the PII attempt violations were detected

PROPOSED FRAMEWORK



METHODS – RESEARCH QUESTION

The study examines whether the Poisson and Gamma-Poisson Distributions explain the use of regular expressions (regexes) to detect the probability of PII violation attempt for a data loss prevention.

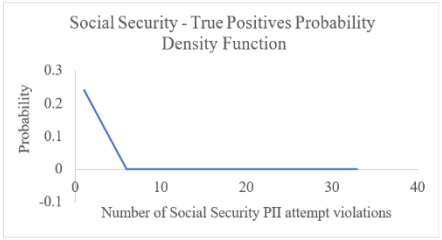
When,

(1) the event of PII attempt violations can be counted in whole numbers;

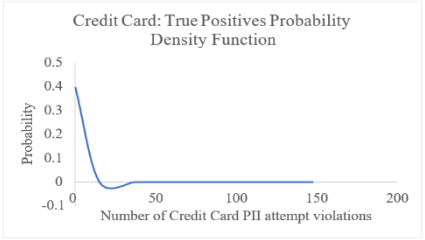
(2) the occurrences of PII attempt violations are independent, so that one PII attempt violation occurrence neither diminishes nor increases the chance of another; and

(3) Poisson-Gamma model simultaneously describes the PII attempt violation occurrence and intensity at once and a suitable model for zero inflated data which reduces PII violation attempts.

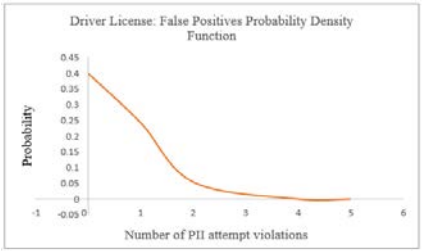
Results: Regexes



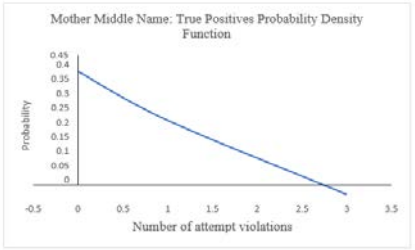
Graph 1: Social Security – True Positive Probability Density Function



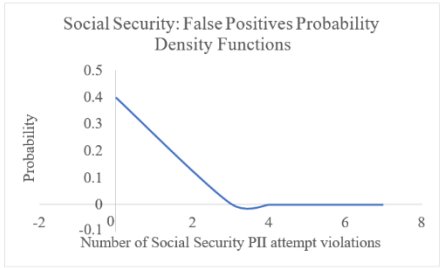
Graph 3: Credit Card – True Positive Probability Density Function



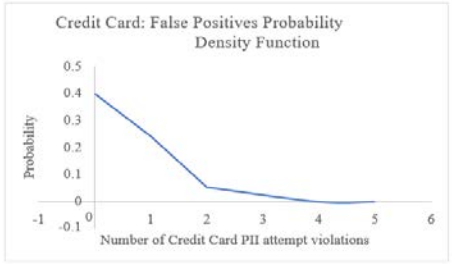
Graph 5: Driver License – True Positive Probability Density Function



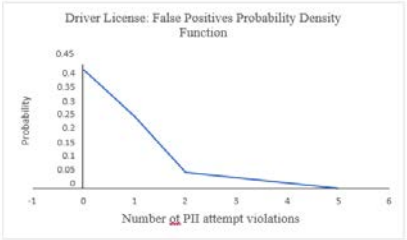
Graph 7: Mother Middle Name – True Positive Probability Density Function



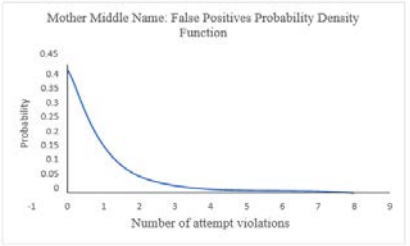
Graph 2: Social Security – False Positive Probability Density Function



Graph 4: Credit Card – False Positive Probability Density Function



Graph 6: Driver License – False Positive Probability Density Function



Graph 8: Mother Middle Name – False Positive Probability Density Function

Conclusion

- Probability density function may support organizations with PII attempt violations.
- Social security regex required closer attention to detected true and false positive PII attempts violations
- Credit card and driver license regexes exhibited a limited number of false positives PII attempt violations
- Mother middle mother name regex is the most difficult parameter to control in the prevention and detection of PII attempt violations.

Acknowledgements & References

Acknowledgements
David Bedell, BS Information Systems
Bruce Wendell, CISSP, CEH
Dr. Diane Murphy, IT, Data Science and Cybersecurity Department Chair, Marymount University
Dr. Donna Schaeffer, DSc Program Director, Marymount University
Dr. Alex Mbaziira, Assistant Professor, Marymount University
Dr. Katrina Anderson, Assistant Professor, Marymount University

Acknowledgements & References

References

Alazab, A., Hobbs, M., Abawajy, J., Khraisat, A., & Alazab, M. (2014). Using response action with intelligent intrusion detection and prevention system against Web application malware. *Information Management & Computer Security, 22*(5), 431-449. Retrieved from http://proxymu.wrlc.org/login?url=https://search-proquest-com.proxymu.wrlc.org/docview/1634006465?accountid=27975

Buchak, K. & Sakhno, L. (2017). Compositions of Poisson and Gamma processes. *Modern Stochastics: Theory and Applications*: 4 (*2*) (2017) 161–188 DOI: 10.15559/17-VMSTA79

Dzupire, N.C., Ngare, P. & Odongo, L. (2018). A Poisson-Gamma Model for Zero Inflated Rainfall Data. *Journal of Probability and Statistics*, DOI: https://doi.org/10.1155/2018/1012647

Han, T., Katoen, J., & Berteun, D. (2009). Counterexample generation in probabilistic model checking. *IEEE Transactions on Software Engineering, 35*(2), 241-257. doi: http://dx.doi.org.proxymu.wrlc.org/10.1109/TSE.2009.5

Ko, Y., Jung, M., Han, Y., & Burgstaller, B. (2014). A speculative parallel DFA membership test for multicore, SIMD and cloud computing environments. *International Journal of Parallel Programming, 42*(3), 456-489. doi: http://dx.doi.org.proxymu.wrlc.org/10.1007/s10766-013- 0258-5

Punithan, X. J., Kim, J., Kim, D., & Choi, Y. (2016). A game theoretic model for dynamic configuration of large-scale intrusion detection signatures. *Multimedia Tools and Applications, 75*(23), 15461-15477. doi: http://dx.doi.org.proxymu.wrlc.org/10.1007/s11042- 015-2508-6

Singh, R., Kumar, H., Singla, R. K., & Ketti, R. R. (2017). Internet attacks and intrusion detection system. *Online Information Review, 41*(2), 171-184. doi: http://dx.doi.org.proxymu.wrlc.org/10.1108/OIR-12-2015-0394