Storlie, Curtis; **Anderson, Blake**; **Vander Wiel, Scott**; **Quist, Daniel**; **Hash, Curtis**; **Brown, Nathan**

Stochastic identification of malware with dynamic traces. (English) Zbl 1429.62713

Ann. Appl. Stat. 8, No. 1, 1-18 (2014).

Summary: A novel approach to malware classification is introduced based on analysis of instruction traces that are collected dynamically from the program in question. The method has been implemented online in a sandbox environment (i.e., a security mechanism for separating running programs) at Los Alamos National Laboratory, and is intended for eventual host-based use, provided the issue of sampling the instructions executed by a given process without disruption to the user can be satisfactorily addressed. The procedure represents an instruction trace with a Markov chain structure in which the transition matrix, **P**, has rows modeled as Dirichlet vectors. The malware class (malicious or benign) is modeled using a flexible spline logistic regression model with variable selection on the elements of **P**, which are observed with error. The utility of the method is illustrated on a sample of traces from malware and non-malware programs, and the results are compared to other leading detection schemes (both signature and classification based).

MSC:

62P30 Applications of statistics in engineering and industry; control charts
62J12 Generalized linear models (logistic models)

Keywords:

malware detection; classification; elastic net; relaxed Lasso; adaptive Lasso; logistic regression; splines; empirical Bayes

Software:

reglogit

Full Text: DOI Euclid

References:

[1] Anderson, B., Quist, D., Neil, J., Storlie, C. and Lane, T. (2011). Graph-based malware detection using dynamic analysis. Journal in Computer Virology 7 247-258.

[2] Anderson, B., Quist, D., Brown, N., Storlie, C. and Lane, T. (2012). Improving malware classification: Bridging the static/dynamic gap. In Proceedings of the 5 th ACM Workshop on Security and Artificial Intelligence 3-14. ACM, New York.

[3] Antivirus Comparatives (2011). Retrospective test (static detection of new/unknown malicious software). Available at . · www.av-comparatives.org

[4] Bayer, U., Moser, A., Kruegel, C. and Kirda, E. (2006). Dynamic analysis of malicious code. Journal in Computer Virology 2 67-77.

[5] Bilar, D. (2007). Opcodes as predictor for malware. International Journal of Electronic Security and Digital Forensics 1 156-168.

[6] Christodorescu, M. and Jha, S. (2003). Static analysis of executables to detect malicious patterns. In Proceedings of the 12 th USENIX Security Symposium 169-186. USENIX Association, Berkeley, CA.

[7] Cova, M., Kruegel, C. and Vigna, G. (2010). Detection and analysis of drive-by-download attacks and malicious javascript code. In Proceedings of the 19 th International Conference on World Wide Web 281-290. ACM, New York.

[8] Dai, J., Guha, R. and Lee, J. (2009). Efficient virus detection using dynamic instruction sequences. Journal of Computers 4 405-414.

[9] Dinaburg, A., Royal, P., Sharif, M. and Lee, W. (2008). Ether: Malware analysis via hardware virtualization extensions. In Proceedings of the 15 th ACM Conference on Computer and Communications Security 51-62. ACM, New York.

[10] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. Ann. Statist. 32 407-499. · Zbl 1091.62054 · doi:10.1214/009053604000000067 ·

[11] Goldberg, I., Wagner, D., Thomas, R. and Brewer, E. (1996). A secure environment for untrusted helper applications (confining

the wily hacker). In Proceedings of the Sixth USENIX UNIX Security Symposium 6 1. USENIX Association, Berkeley, CA.

[12] Gramacy, R. B. and Polson, N. G. (2012). Simulation-based regularized logistic regression. Bayesian Anal. 7 567-589. · Zbl 1330.62301 · doi:10.1214/12-BA719 ·

[13] Hastie, T. and Tibshirani, R. (1996). Discriminant analysis by Gaussian mixtures. J. R. Stat. Soc. Ser. B Stat. Methodol. 58 155-176. · Zbl 0850.62476

[14] Hofmeyr, S. A., Forrest, S. and Somayaji, A. (1998). Intrusion detection using sequences of system calls. Journal of Computer Security 6 151-180.

[15] King, G. and Zeng, L. (2001). Logistic regression in rare events data. Political Analysis 9 137-163.

[16] Kolter, J. Z. and Maloof, M. A. (2006). Learning to detect and classify malicious executables in the wild. J. Mach. Learn. Res. 7 2721-2744. · Zbl 1222.68236 · www.jmlr.org

[17] Luk, C.-K., Cohn, R., Muth, R., Patil, H., Klauser, A., Lowney, G., Wallace, S., Reddi, V. J. and Hazelwood, K. (2005). Pin: Building customized program analysis tools with dynamic instrumentation. In Proceedings of the ACM SIGPLAN Conference on Programming Language Design and Implementation 190-200. ACM, New York.

[18] Manski, C. F. and Lerman, S. R. (1977). The estimation of choice probabilities from choice based samples. Econometrica 45 1977-1988. · Zbl 0372.62094 · doi:10.2307/1914121

[19] Meinshausen, N. (2007). Relaxed Lasso. Comput. Statist. Data Anal. 52 374-393. · Zbl 1452.62522

[20] PandaLabs (2012). PandaLabs quarterly report. Available at . · press.pandasecurity.com

[21] Perdisci, R., Dagon, D., Fogla, P. and Sharif, M. (2006). Misleading worm signature generators using deliberate noise injection. In Proceedings of the IEEE Symposium on Security and Privacy 17-31. IEEE Computer Society Technical Committee on Security and Privacy.

[22] Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. Biometrika 66 403-411. · Zbl 0428.62078 · doi:10.1093/biomet/66.3.403

[23] Quist, D. (2012). Community malicious code research and analysis. Available at . · www.offensivecomputing.net

[24] Reddy, D. K. S., Dash, S. and Pujari, A. (2006). New malicious code detection using variable length \(n\)-grams. In Information Systems Security. Lecture Notes in Computer Science 4332 276-288. Springer, Berlin.

[25] Reddy, D. and Pujari, A. (2006). \(N\)-gram analysis for computer virus detection. Journal in Computer Virology 2 231-239.

[26] Rieck, K., Trinius, P., Willems, C. and Holz, T. (2011). Automatic analysis of malware behavior using machine learning. Journal of Computer Security 19 639-668.

[27] Royal, P., Halpin, M., Dagon, D., Edmonds, R. and Lee, W. (2006). Polyunpack: Automating the hidden-code extraction of unpackexecuting malware. In Proceedings of the 22 nd Annual Computer Security Applications Conference 289-300.

[28] Shafiq, M., Khayam, S. and Farooq, M. (2008). Embedded malware detection using Markov \(n\)-grams. In Proceedings of the 5 th International Conference on Detection of Intrusions and Malware , and Vulnerability Assessment 88-107. ACM, New York.

[29] Shankarapani, M., Ramamoorthy, S., Movva, R. and Mukkamala, S. (2010). Malware detection using assembly and API call sequences. Journal in Computer Virology 7 1-13.

[30] Skaletsky, A., Devor, T., Chachmon, N., Cohn, R., Hazelwood, K., Vladimirov, V. and Bach, M. (2010). Dynamic program analysis of Microsoft Windows applications. In 2010 International Symposium on Performance Analysis of Software and Systems (ISPASS) 2-12. IEEE Computer Society's Technical Committee on the Internet.

[31] Stolfo, S., Wang, K. and Li, W.-J. (2007). Towards stealthy malware detection. In Malware Detection. Advances in Information Security 27 231-249. Springer, New York.

[32] Storlie, C., Anderson, B., Vander Wiel, S., Quist, D., Hash, C. and Brown, N. (2014). Supplement to "Stochastic identification of malware with dynamic traces." . · Zbl 1429.62713 · dx.doi.org

[33] Symantec (2008). Internet security threat report, trends for July-December 2007 (executive summary). White paper. Available at . · eval.symantec.com

[34] Symantec (2011). Internet security threat report, volume 16. White paper. Available at . · www.symantec.com

[35] Taddy, M. (2013). Multinomial inverse regression for text analysis. J. Amer. Statist. Assoc. 108 755-770. · Zbl 06224965 · doi:10.1080/01621459.2012.734168

[36] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58 267-288. · Zbl 0850.62538

[37] Zou, H. (2006). The Adaptive Lasso and its oracle properties. J. Amer. Statist. Assoc. 101 1418-1429. · Zbl 1171.62326 · doi:10.1198/016214506000000735

[38] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat. Methodol. 67 301-320. · Zbl 1069.62054 · doi:10.1111/j.1467-9868.2005.00503.x