References

Aldà, F. & Rubinstein, B. I. P. (2017), The Bernstein mechanism: Function release under differential privacy, *in* "Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI'2017)."

Alfeld, S., Zhu, X., & Barford, P. (2016), Data poisoning attacks against autoregressive models, *in* "Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI'2016)," pp. 1452–1458.

Alfeld, S., Zhu, X., & Barford, P. (2017), Explicit defense actions against test-set attacks, *in* "Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI'2017)."

Alpcan, T., Rubinstein, B. I. P., & Leckie, C. (2016), Large-scale strategic games and adversarial machine learning, *in* "2016 IEEE 55th Conference on Decision and Control (CDC)," IEEE, pp. 4420–4426.

Amsaleg, L., Bailey, J., Erfani, S., Furon, T., Houle, M. E., Radovanović, M., & Vinh, N. X. (2016), The vulnerability of learning to adversarial perturbation increases with intrinsic dimensionality, Technical Report NII-2016-005E, National Institute of Informatics, Japan.

Angluin, D. (1988), "Queries and concept learning," *Machine Learning* **2**, 319–342.

Apa (n.d.), *Apache SpamAssassin*.

Bahl, P., Chandra, R., Greenberg, A., Kandula, S., Maltz, D. A., & Zhang, M. (2007), Towards highly reliable enterprise network services via inference of multi-level dependencies, *in* "Proceedings of the 2007 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM)," pp. 13–24.

Balfanz, D. & Staddon, J., eds (2008), *Proceedings of the* 1^{st} *ACM Workshop on Security and Artificial Intelligence, AISec 2008*.

Balfanz, D. & Staddon, J., eds (2009), *Proceedings of the* 2^{nd} *ACM Workshop on Security and Artificial Intelligence, AISec 2009*.

Barak, B., Chaudhuri, K., Dwork, C., Kale, S., McSherry, F., & Talwar, K. (2007), Privacy, accuracy, and consistency too: A holistic solution to contingency table release, *in* "Proceedings of the Twenty-Sixth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems," pp. 273–282.

Barbaro, M. & Zeller Jr., T. (2006), "A face is exposed for AOL searcher no. 4417749," *New York Times*.

Barreno, M. (2008), Evaluating the security of machine learning algorithms. PhD thesis, University of California, Berkeley.

Barreno, M., Nelson, B., Joseph, A. D., & Tygar, J. D. (2010), "The security of machine learning," *Machine Learning* **81**(2), 121–148.

Barreno, M., Nelson, B., Sears, R., Joseph, A. D., & Tygar, J. D. (2006), Can machine learning be secure?, *in* "Proceedings of the ACM Symposium on Information, Computer and Communications Security (ASIACCS)," pp. 16–25.

https://doi.org/10.1017/9781107338548.015 Published online by Cambridge University Press

Barth, A., Rubinstein, B. I. P., Sundararajan, M., Mitchell, J. C., Song, D., & Bartlett, P. L. (2012), "A learning-based approach to reactive security," *IEEE Transactions on Dependable and Secure Computing* **9**(4), 482–493. Special Issue on Learning, Games, and Security.

Bassily, R., Smith, A., & Thakurta, A. (2014), Private empirical risk minimization: Efficient algorithms and tight error bounds, *in* "2014 IEEE 55th Annual Symposium on Foundations of Computer Science (FOCS)," pp. 464–473.

Beimel, A., Kasiviswanathan, S., & Nissim, K. (2010), Bounds on the sample complexity for private learning and private data release, *in* "Theory of Cryptography Conference," Vol. 5978 of *Lecture Notes in Computer Science*, Springer, pp. 437–454.

Bennett, J., Lanning, S., et al. (2007), The Netflix prize, *in* "Proceedings of KDD Cup and Workshop," Vol. 2007, pp. 3–6.

Bertsimas, D. & Vempala, S. (2004), "Solving convex programs by random walks," *Journal of the ACM* **51**(4), 540–556.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Srndic, N., Laskov, P., Giacinto, G., & Roli, F. (2013), Evasion attacks against machine learning at test time, *in* "Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2013," pp. 387–402.

Biggio, B., Fumera, G., & Roli, F. (2010), Multiple classifier systems under attack, *in* N. E. G. J. K. F. Roli, ed., "Proceedings of the 9th International Workshop on Multiple Classifier Systems (MCS)," Vol. 5997, Springer, pp. 74–83.

Biggio, B., Nelson, B., & Laskov, P. (2012), Poisoning attacks against support vector machines, *in* "Proceedings of the 29th International Conference on Machine Learning (ICML-12)," pp. 1807–1814.

Biggio, B., Rieck, K., Ariu, D., Wressnegger, C., Corona, I., Giacinto, G., & Roli, F. (2014), Poisoning behavioral malware clustering, *in* "Proceedings of the 2014 Workshop on Artificial Intelligent and Security Workshop, AISec 2014," pp. 27–36.

Billingsley, P. (1995), *Probability and Measure*, 3rd edn, Wiley.

Bishop, C. M. (2006), *Pattern Recognition and Machine Learning*, Springer-Verlag.

Blocki, J., Christin, N., Datta, A., & Sinha, A. (2011), Regret minimizing audits: A learning-theoretic basis for privacy protection, *in* "Proceedings of the 24th IEEE Computer Security Foundations Symposium," pp. 312–327.

Blum, A., Dwork, C., McSherry, F., & Nissim, K. (2005), Practical privacy: The SuLQ framework, *in* "Proceedings of the Twenty-Fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems," pp. 128–138.

Blum, A., Ligett, K., & Roth, A. (2008), A learning theory approach to non-interactive database privacy, *in* "Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing (STOC)," pp. 609–618.

Bodík, P., Fox, A., Franklin, M. J., Jordan, M. I., & Patterson, D. A. (2010), Characterizing, modeling, and generating workload spikes for stateful services, *in* "Proceedings of the 1st ACM Symposium on Cloud Computing (SoCC)," pp. 241–252.

Bodík, P., Griffith, R., Sutton, C., Fox, A., Jordan, M. I., & Patterson, D. A. (2009), Statistical machine learning makes automatic control practical for internet datacenters, *in* "Proceedings of the Workshop on Hot Topics in Cloud Computing (HotCloud)," USENIX Association, pp. 12–17.

Bolton, R. J. & Hand, D. J. (2002), "Statistical fraud detection: A review," *Journal of Statistical Science* **17**(3), 235–255.

Bousquet, O. & Elisseeff, A. (2002), "Stability and generalization," *Journal of Machine Learning Research* **2**(Mar), 499–526.

https://doi.org/10.1017/9781107338548.015 Published online by Cambridge University Press

Boyd, S. & Vandenberghe, L. (2004), *Convex Optimization*, Cambridge University Press.

Brauckhoff, D., Salamatian, K., & May, M. (2009), Applying PCA for traffic anomaly detection: Problems and solutions, *in* "Proceedings of the 28th IEEE International Conference on Computer Communications (INFOCOM)," pp. 2866–2870.

Brent, R. P. (1973), *Algorithms for Minimization without Derivatives*, Prentice-Hall.

Brückner, M. & Scheffer, T. (2009), Nash equilibria of static prediction games, *in* Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams & A. Culotta, eds., "Advances in Neural Information Processing Systems (NIPS)," Vol. 22, MIT Press, pp. 171–179.

Burden, R. L. & Faires, J. D. (2000), *Numerical Analysis*, 7th edn, Brooks Cole.

Burges, C. J. C. (1998), "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery* **2**(2), 121–167.

Cárdenas, A. A., Greenstadt, R., & Rubinstein, B. I. P., eds (2011), *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence, AISec 2011 Chicago, October 21, 2011*, ACM.

Cárdenas, A. A., Nelson, B., & Rubinstein, B. I., eds (2012), *Proceedings of the 5th ACM Workshop on Security and Artificial Intelligence, AISec 2012, Raleigh, North Carolina, October, 19, 2012*, ACM.

Cauwenberghs, G. & Poggio, T. (2000), "Incremental and decremental support vector machine learning," *Advances in Neural Information Processing Systems* **13**, 409–415.

Cesa-Bianchi, N. & Lugosi, G. (2006), *Prediction, Learning, and Games*, Cambridge University Press.

Chandrashekar, J., Orrin, S., Livadas, C., & Schooler, E. M. (2009), "The dark cloud: Understanding and defending against botnets and stealthy malware," *Intel Technology Journal* **13**(2), 130–145.

Chaudhuri, K. & Monteleoni, C. (2009), Privacy-preserving logistic regression, "Advances in Neural Information Processing Systems," 289–296.

Chaudhuri, K., Monteleoni, C., & Sarwate, A. D. (2011), "Differentially private empirical risk minimization," *Journal of Machine Learning Research* **12**, 1069–1109.

Chen, T. M. & Robert, J.-M. (2004), The evolution of viruses and worms, *in* W. W. Chen, ed., *Statistical Methods in Computer Security*, CRC Press, pp. 265–282.

Cheng, Y.-C., Afanasyev, M., Verkaik, P., Benkö, P., Chiang, J., Snoeren, A. C., Savage, S., & Voelker, G. M. (2007), Automating cross-layer diagnosis of enterprise wireless networks, *in* "Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM)," pp. 25–36.

Christmann, A. & Steinwart, I. (2004), "On robustness properties of convex risk minimization methods for pattern recognition," *Journal of Machine Learning Research* **5**, 1007–1034.

Chung, S. P. & Mok, A. K. (2006), Allergy attack against automatic signature generation, *in* D. Zamboni & C. Krügel, eds., "Proceedings of the 9th International Symposium on Recent Advances in Intrusion Detection (RAID)," Springer, pp. 61–80.

Chung, S. P. & Mok, A. K. (2007), Advanced allergy attacks: Does a corpus really help?, *in* C. Krügel, R. Lippmann & A. Clark, eds, "Proceedings of the 10th International Symposium on Recent Advances in Intrusion Detection (RAID)," Vol. 4637 of *Lecture Notes in Computer Science*, Springer, pp. 236–255.

Cormack, G. & Lynam, T. (2005), Spam corpus creation for TREC, *in* "Proceedings of the Conference on Email and Anti-Spam (CEAS)."

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2001), *Introduction to Algorithms*, 2nd edn, McGraw-Hill. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.86.3539&rep=rep1&type=pdf.

Cormode, G., Procopiuc, C., Srivastava, D., Shen, E., & Yu, T. (2012), Differentially private spatial decompositions, *in* "2012 IEEE 28th International Conference on Data Engineering (ICDE)," pp. 20–31.

Cover, T. M. (1991), "Universal portfolios," *Mathematical Finance* **1**(1), 1–29.

Cristianini, N. & Shawe-Taylor, J. (2000), *An Introduction to Support Vector Machines*, Cambridge University Press.

Croux, C., Filzmoser, P., & Oliveira, M. R. (2007), "Algorithms for projection-pursuit robust principal component analysis," *Chemometrics and Intelligent Laboratory Systems* **87**(2), 218–225.

Croux, C. & Ruiz-Gazen, A. (2005), "High breakdown estimators for principal components: The projection-pursuit approach revisited," *Journal of Multivariate Analysis* **95**(1), 206–226.

Dalvi, N., Domingos, P., Mausam, Sanghai, S., & Verma, D. (2004), Adversarial classification, *in* "Proceedings of the 10th ACM International Conference on Knowledge Discovery and Data Mining (KDD)," pp. 99–108.

Dasgupta, S., Kalai, A. T., & Monteleoni, C. (2009), "Analysis of perceptron-based active learning," *Journal of Machine Learning Research* **10**, 281–299.

De, A. (2012), Lower bounds in differential privacy, *in* "Theory of Cryptography Conference," Springer, pp. 321–338.

Denning, D. E. & Denning, P. J. (1979), "Data security," *ACM Computing Surveys* **11**, 227–249.

Devlin, S. J., Gnanadesikan, R., & Kettenring, J. R. (1981), "Robust estimation of dispersion matrices and principal components," *Journal of the American Statistical Association* **76**, 354–362.

Devroye, L., Györfi, L., & Lugosi, G. (1996), *A Probabilistic Theory of Pattern Recognition*, Springer Verlag.

Devroye, L. P. & Wagner, T. J. (1979), "Distribution-free performance bounds for potential function rules," *IEEE Transactions on Information Theory* **25**(5), 601–604.

Diffie, W. & Hellman, M. E. (1976), "New directions in cryptography," *IEEE Transactions on Information Theory* **22**(6), 644–654.

Dimitrakakis, C., Gkoulalas-Divanis, A., Mitrokotsa, A., Verykios, V. S., & Saygin, Y., eds (2011), *Privacy and Security Issues in Data Mining and Machine Learning - International ECML/PKDD Workshop, PSDML 2010, Barcelona, September 24, 2010. Revised Selected Papers*, Springer.

Dimitrakakis, C., Laskov, P., Lowd, D., Rubinstein, B. I. P., & Shi, E., eds (2014), *Proceedings of the 1st ICML Workshop on Learning, Security and Privacy, Beijing, China, June 25, 2014*.

Dimitrakakis, C., Mitrokotsa, K., & Rubinstein, B. I. P., eds (2014), *Proceedings of the 7th ACM Workshop on Artificial Intelligence and Security, AISec 2014, Scottsdale, AZ, November 7, 2014*.

Dimitrakakis, C., Mitrokotsa, K., & Sinha, A., eds. (2015), *Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security, AISec 2015, Denver, CO, October 16, 2015*.

Dimitrakakis, C., Nelson, B., Mitrokotsa, A., & Rubinstein, B. I. P. (2014), Robust and private Bayesian inference, *in* "Proceedings of the 25th International Conference Algorithmic Learning Theory (ALT)," pp. 291–305.

Dinur, I. & Nissim, K. (2003), Revealing information while preserving privacy, *in* "Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems," pp. 202–210.

Dredze, M., Gevaryahu, R., & Elias-Bachrach, A. (2007), Learning fast classifiers for image spam, *in* "Proceedings of the 4th Conference on Email and Anti-Spam (CEAS)." http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.102.8417&rep=rep1&type=pdf.

https://doi.org/10.1017/9781107338548.015 Published online by Cambridge University Press

Duchi, J. C., Jordan, M. I., & Wainwright, M. J. (2013), Local privacy and statistical minimax rates, *in* "2013 IEEE 54th Annual Symposium on Foundations of Computer Science (FOCS)," pp. 429–438.

Dwork, C. (2006), Differential privacy, *in* "Proceedings of the 33rd International Conference on Automata, Languages and Programming," pp. 1–12.

Dwork, C. (2010), "A firm foundation for private data analysis," *Communications of the ACM* **53** (6), 705–714.

Dwork, C. & Lei, J. (2009), Differential privacy and robust statistics, *in* "Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing (STOC)," pp. 371–380.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006), Calibrating noise to sensitivity in private data analysis, *in* "Theory of Cryptography Conference," pp. 265–284.

Dwork, C., McSherry, F., & Talwar, K. (2007), The price of privacy and the limits of LP decoding, *in* "Proceedings of the 39th Annual ACM Symposium on Theory of Computing (STOC)," pp. 85–94.

Dwork, C., Naor, M., Reingold, O., Rothblum, G. N., & Vadhan, S. (2009), On the complexity of differentially private data release: Efficient algorithms and hardness results, *in* "Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing (STOC)," pp. 381–390.

Dwork, C. & Roth, A. (2014), "The algorithmic foundations of differential privacy," *Foundations and Trends in Theoretical Computer Science* **9**(3–4), 211–407.

Dwork, C. & Yekhanin, S. (2008), New efficient attacks on statistical disclosure control mechanisms, *in* "CRYPTO'08," pp. 469–480.

Erlich, Y. & Narayanan, A. (2014), "Routes for breaching and protecting genetic privacy," *Nature Reviews Genetics* **15**, 409–421.

Eskin, E., Arnold, A., Prerau, M., Portnoy, L., & Stolfo, S. J. (2002), A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data, *in Data Mining for Security Applications*, Kluwer.

Feldman, V. (2009), "On the power of membership queries in agnostic learning," *Journal of Machine Learning Research* **10**, 163–182.

Fisher, R. A. (1948), "Question 14: Combining independent tests of significance," *American Statistician* **2**(5), 30–31.

Flum, J. & Grohe, M. (2006), *Parameterized Complexity Theory*, Texts in Theoretical Computer Science, Springer-Verlag.

Fogla, P. & Lee, W. (2006), Evading network anomaly detection systems: Formal reasoning and practical techniques, *in* "Proceedings of the 13th ACM Conference on Computer and Communications Security (CCS)," pp. 59–68.

Forrest, S., Hofmeyr, S. A., Somayaji, A., & Longstaff, T. A. (1996), A sense of self for Unix processes, *in* "Proceedings of the IEEE Symposium on Security and Privacy (SP)," pp. 120–128.

Freeman, D., Mitrokotsa, K., & Sinha, A., eds (2016), *Proceedings of the 9th ACM Workshop on Artificial Intelligence and Security, AISec 2016, Vienna, Austria, October 28, 2016*.

Globerson, A. & Roweis, S. (2006), Nightmare at test time: Robust learning by feature deletion, *in* "Proceedings of the 23rd International Conference on Machine Learning (ICML)," pp. 353–360.

Goldman, S. A. & Kearns, M. J. (1995), "On the complexity of teaching," *Journal of Computer and System Sciences* **50**(1), 20–31.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015), Explaining and harnessing adversarial challenges, *in* "Proceedings of the International Conference on Learning Representations."

https://doi.org/10.1017/9781107338548.015 Published online by Cambridge University Press

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014), Generative adversarial nets, *in* "Advances in Neural Information Processing Systems," pp. 2672–2680.

Gottlieb, L.-A., Kontorovich, A., & Mossel, E. (2011), VC bounds on the cardinality of nearly orthogonal function classes, Technical Report arXiv:1007.4915v2 [math.CO], arXiv.

Graham, P. (2002), "A plan for spam," http://www.paulgraham.com/spam.html.

Greenstadt, R., ed. (2010), *Proceedings of the 3rd ACM Workshop on Security and Artificial Intelligence, AISec 2010, Chicago, October 8, 2010*, ACM.

Großhans, M., Sawade, C., Brückner, M., & Scheffer, T. (2013), Bayesian games for adversarial regression problems, *in* "Proceedings of the 30th International Conference on Machine Learning, ICML 2013," pp. 55–63.

Gymrek, M., McGuire, A. L., Golan, D., Halperin, E., & Erlich, Y. (2013), "Identifying personal genomes by surname inference," *Science* **339**(6117), 321–324.

Hall, J. F. (2005), "Fun with stacking blocks," *American Journal of Physics* **73**(12), 1107–1116.

Hall, R., Rinaldo, A., & Wasserman, L. (2013), "Differential privacy for functions and functional data," *Journal of Machine Learning Research* **14**(1), 703–727.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986), *Robust Statistics: The Approach Based on Influence Functions*, John Wiley.

Hardt, M., Ligett, K., & McSherry, F. (2012), A simple and practical algorithm for differentially private data release, *in* F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger, eds., "Advances in Neural Information Processing Systems 25 (NIPS)," pp. 2339–2347.

Hardt, M. & Talwar, K. (2010), On the geometry of differential privacy, *in* "Proceedings of the Forty-Second Annual ACM Symposium on Theory of Computing (STOC)," pp. 705–714.

Hastie, T., Tibshirani, R., & Friedman, J. (2003), *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer.

He, X., Cormode, G., Machanavajjhala, A., Procopiuc, C. M., & Srivastava, D. (2015), "Dpt: differentially private trajectory synthesis using hierarchical reference systems," *Proceedings of the VLDB Endowment* **8**(11), 1154–1165.

Helmbold, D. P., Singer, Y., Schapire, R. E., & Warmuth, M. K. (1998), "On-line portfolio selection using multiplicative updates," *Mathematical Finance* **8**, 325–347.

Hofmeyr, S. A., Forrest, S., & Somayaji, A. (1998), "Intrusion detection using sequences of system calls," *Journal of Computer Security* **6**(3), 151–180.

Hohm, T., Egli, M., Gaehwiler, S., Bleuler, S., Feller, J., Frick, D., Huber, R., Karlsson, M., Lingenhag, R., Ruetimann, T., Sasse, T., Steiner, T., Stocker, J., & Zitzler, E. (2007), An evolutionary algorithm for the block stacking problem, *in* "8th International Conference Artificial Evolution (EA 2007)," Springer, pp. 112–123.

Holz, T., Steiner, M., Dahl, F., Biersack, E., & Freiling, F. (2008), Measurements and mitigation of peer-to-peer-based botnets: A case study on storm worm, *in* "Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats," LEET'08, pp. 1–9.

Homer, N., Szelinger, S., Redman, M., Duggan, D., Tembe, W., Muehling, J., Pearson, J. V., Stephan, D. A., Nelson, S. F., & Craig, D. W. (2008), "Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays," *PLoS Genetics* **4**(8).

Hössjer, O. & Croux, C. (1995), "Generalizing univariate signed rank statistics for testing and estimating a multivariate location parameter," *Journal of Nonparametric Statistics* **4**(3), 293–308.

Huang, L., Nguyen, X., Garofalakis, M., Jordan, M. I., Joseph, A., & Taft, N. (2007), In-network PCA and anomaly detection, *in* B. Schölkopf, J. Platt & T. Hoffman, eds., "Advances in Neural Information Processing Systems 19 (NIPS)," MIT Press, pp. 617–624.

Huber, P. J. (1981), *Robust Statistics*, Probability and Mathematical Statistics, John Wiley.

Jackson, J. E. & Mudholkar, G. S. (1979), "Control procedures for residuals associated with principal component analysis," *Technometrics* **21**(3), 341–349.

Johnson, P. B. (1955), "Leaning tower of lire," *American Journal of Physics* **23**(4), 240.

Jones, D. R. (2001), "A taxonomy of global optimization methods based on response surfaces," *Journal of Global Optimization* **21**(4), 345–383.

Jones, D. R., Perttunen, C. D., & Stuckman, B. E. (1993), "Lipschitzian optimization without the Lipschitz constant," *Journal of Optimization Theory and Application* **79**(1), 157–181.

Joseph, A. D., Laskov, P., Roli, F., Tygar, J. D., & Nelson, B. (2013), "Machine Learning Methods for Computer Security (Dagstuhl Perspectives Workshop 12371)," *Dagstuhl Manifestos* **3**(1), 1–30. http://drops.dagstuhl.de/opus/volltexte/2013/4356.

Jurafsky, D. & Martin, J. H. (2008), *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, 2nd edn, Prentice-Hall.

Kalai, A. & Vempala, S. (2002), "Efficient algorithms for universal portfolios," *Journal of Machine Learning Research* **3**, 423–440.

Kandula, S., Chandra, R., & Katabi, D. (2008), What's going on? Learning communication rules in edge networks, *in* "Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM)," pp. 87–98.

Kantarcioglu, M., Xi, B., & Clifton, C. (2009), Classifier evaluation and attribute selection against active adversaries, Technical Report 09-01, Purdue University.

Kantchelian, A., Ma, J., Huang, L., Afroz, S., Joseph, A. D., & Tygar, J. D. (2012), Robust detection of comment spam using entropy rate, *in* "Proceedings of the 5th ACM Workshop on Security and Artificial Intelligence (AISec 2012)," pp. 59–70.

Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., & Smith, A. (2008), What can we learn privately?, *in* "Proceedings of the 49th Annual IEEE Symposium on Foundations of Computer Science (FOCS)," pp. 531–540.

Kearns, M. & Li, M. (1993), "Learning in the presence of malicious errors," *SIAM Journal on Computing* **22**(4), 807–837.

Kearns, M. & Ron, D. (1999), "Algorithmic stability and sanity-check bounds for leave-one-out cross-validation," *Neural Computation* **11**, 1427–1453.

Kerckhoffs, A. (1883), "La cryptographie militaire," *Journal des Sciences Militaires* **9**, 5–83.

Kim, H.-A. & Karp, B. (2004), Autograph: Toward automated, distributed worm signature detection, *in* "USENIX Security Symposium" available at https://www.usenix.org/legacy/publications/library/proceedings/sec04/tech/full_papers/kim/kim.pdf.

Kimeldorf, G. & Wahba, G. (1971), "Some results on Tchebycheffian spline functions," *Journal of Mathematical Analysis and Applications* **33**(1), 82–95.

Klíma, R., Lisý, V., & Kiekintveld, C. (2015), Combining online learning and equilibrium computation in security games, *in* "International Conference on Decision and Game Theory for Security," Springer, pp. 130–149.

Klimt, B. & Yang, Y. (2004), Introducing the Enron corpus, *in* "Proceedings of the Conference on Email and Anti-Spam (CEAS)" available at https://bklimt.com/papers/2004_klimt_ceas.pdf.

Kloft, M. & Laskov, P. (2010), Online anomaly detection under adversarial impact, *in* "Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)," pp. 406–412.

Kloft, M. & Laskov, P. (2012), "Security analysis of online centroid anomaly detection," *Journal of Machine Learning Research* **13**, 3681–3724.

Kolda, T. G., Lewis, R. M., & Torczon, V. (2003), "Optimization by direct search: New perspectives on some classical and modern methods," *SIAM Review* **45**(3), 385–482.

Korolova, A. (2011), "Privacy violations using microtargeted ads: A case study," *Journal of Privacy and Confidentiality* **3**(1).

Kutin, S. & Niyogi, P. (2002), Almost-everywhere algorithmic stability and generalization error, Technical report TR-2002-03, Computer Science Dept., University of Chicago.

Lakhina, A., Crovella, M., & Diot, C. (2004*a*), Characterization of network-wide anomalies in traffic flows, *in* A. Lombardo & J. F. Kurose, eds., "Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement (IMC)," pp. 201–206.

Lakhina, A., Crovella, M., & Diot, C. (2004*b*), Diagnosing network-wide traffic anomalies, *in* R. Yavatkar, E. W. Zegura & J. Rexford, eds., "Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM)," pp. 219–230.

Lakhina, A., Crovella, M., & Diot, C. (2005*a*), Detecting distributed attacks using network-wide flow traffic, *in* "Proceedings of the FloCon 2005 Analysis Workshop" available at http://www.cs.bu.edu/~crovella/paper-archive/flocon05.pdf.

Lakhina, A., Crovella, M., & Diot, C. (2005*b*), Mining anomalies using traffic feature distributions, *in* "Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM)," pp. 217–228.

Laskov, P. & Kloft, M. (2009), A framework for quantitative security analysis of machine learning, *in* "Proceedings of the 2nd ACM Workshop on Security and Artificial Intelligence (AISec)," pp. 1–4.

Laskov, P. & Lippmann, R. (2010), "Machine learning in adversarial environments," *Machine Learning* **81**(2), 115–119.

Lazarevic, A., Ertöz, L., Kumar, V., Ozgur, A., & Srivastava, J. (2003), A comparative study of anomaly detection schemes in network intrusion detection, *in* D. Barbará & C. Kamath, eds., "Proceedings of the SIAM International Conference on Data Mining," pp. 25–36.

LeCun, Y., Bengio, Y., & Hinton, G. (2015), "Deep learning," *Nature* **521**(7553), 436–444.

Li, B. & Vorobeychik, Y. (2014), Feature cross-substitution in adversarial classification, *in* "Advances in Neural Information Processing Systems," pp. 2087–2095.

Li, B., Wang, Y., Singh, A., & Vorobeychik, Y. (2016), Data poisoning attacks on factorization-based collaborative filtering, *in* "Advances in Neural Information Processing Systems," pp. 1885–1893.

Li, C., Hay, M., Miklau, G., & Wang, Y. (2014), "A data-and workload-aware algorithm for range queries under differential privacy," *Proceedings of the VLDB Endowment* **7**(5), 341–352.

Li, G. & Chen, Z. (1985), "Projection-pursuit approach to robust dispersion matrices and principal components: Primary theory and Monte Carlo," *Journal of the American Statistical Association* **80**(391), 759–766.

Li, N., Li, T., & Venkatasubramanian, S. (2007), t-Closeness: Privacy beyond k-anonymity and l-diversity, *in* "IEEE 23rd International Conference on Data Engineering (ICED)," pp. 106–115.

Li, X., Bian, F., Crovella, M., Diot, C., Govindan, R., Iannaccone, G., & Lakhina, A. (2006), Detection and identification of network anomalies using sketch subspaces, *in* J. M. Almeida,

V. A. F. Almeida, & P. Barford, eds., "Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement (IMC)," pp. 147–152.

Littlestone, N. & Warmuth, M. K. (1994), "The weighted majority algorithm," *Information and Computation* **108**(2), 212–261.

Liu, C. & Stamm, S. (2007), Fighting unicode-obfuscated spam, *in* "Proceedings of the Anti-Phishing Working Groups 2nd Annual eCrime Researchers Summit," pp. 45–59.

Liu, Y., Chen, X., Liu, C., & Song, D. (2017), Delving into transferable adversarial examples and black-box attacks, *in* "Proceedings of the International Conference on Learning Representations" available at https://people.eecs.berkeley.edu/∼liuchang/paper/transferability_iclr_2017.pdf.

Lovász, L. & Vempala, S. (2003), Simulated annealing in convex bodies and an $O^*(n^4)$ volume algorithm, *in* "Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science (FOCS)," pp. 650–659.

Lovász, L. & Vempala, S. (2004), Hit-and-run from a corner, *in* "Proceedings of the 36th Annual ACM Symposium on Theory of Computing (STOC)," pp. 310–314.

Lowd, D. & Meek, C. (2005*a*), Adversarial learning, *in* "Proceedings of the 11th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)," pp. 641–647.

Lowd, D. & Meek, C. (2005*b*), Good word attacks on statistical spam filters, *in* "Proceedings of the 2nd Conference on Email and Anti-Spam (CEAS)" available at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.130.9846&rep=rep1&type=pdf.

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., & Vilhuber, L. (2008), Privacy: Theory meets practice on the map, *in* "Proceedings of the 2008 IEEE 24th International Conference on Data Engineering," IEEE Computer Society, pp. 277–286.

Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007), "ℓ-Diversity: Privacy beyond k-anonymity," *ACM Transactions on KDD* **1**(1).

Mahoney, M. V. & Chan, P. K. (2002), Learning nonstationary models of normal network traffic for detecting novel attacks, *in* "Proceedings of the 8th ACM International Conference on Knowledge Discovery and Data Mining (KDD)," pp. 376–385.

Mahoney, M. V. & Chan, P. K. (2003), An analysis of the 1999 DARPA/Lincoln Laboratory evaluation data for network anomaly detection, *in* G. Vigna, E. Jonsson, & C. Krügel, eds., "Proceedings of the 6th International Symposium on Recent Advances in Intrusion Detection (RAID)," Vol. 2820 of *Lecture Notes in Computer Science*, Springer, pp. 220–237.

Maronna, R. (2005), "Principal components and orthogonal regression based on robust scales," *Technometrics* **47**(3), 264–273.

Maronna, R. A., Martin, D. R., & Yohai, V. J. (2006), *Robust Statistics: Theory and Methods*, John Wiley.

Martinez, D. R., Streilein, W. W., Carter, K. M., & Sinha, A., eds (2016), *Proceedings of the AAAI Workshop on Artificial Intelligence for Cyber Security, AICS 2016, Phoenix, AZ, February 12, 2016*.

McSherry, F. & Mironov, I. (2009), Differentially private recommender systems: Building privacy into the net, *in* "Proceedings of the 15th ACM International Conference on Knowledge Discovery and Data Mining (KDD)," pp. 627–636.

McSherry, F. & Talwar, K. (2007), Mechanism design via differential privacy, *in* "Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS)," pp. 94–103.

Mei, S. & Zhu, X. (2015*a*), The security of latent Dirichlet allocation, *in* "Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics (AISTATS)," pp. 681–689.

https://doi.org/10.1017/9781107338548.015 Published online by Cambridge University Press

Mei, S. & Zhu, X. (2015*b*), Using machine teaching to identify optimal training-set attacks on machine learners, *in* "Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI)," AAAI Press, pp. 2871–2877.

Meyer, T. A. & Whateley, B. (2004), SpamBayes: Effective open-source, Bayesian based, email classification system, *in* "Proceedings of the Conference on Email and Anti-Spam (CEAS)" available at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.3.9543&rep=rep1&type=pdf.

Microsoft (2009), "H1n1 swine flu response center." https://h1n1.cloudapp.net; Date accessed: March 3, 2011.

Miller, B., Kantchelian, A., Afroz, S., Bachwani, R., Dauber, E., Huang, L., Tschantz, M. C., Joseph, A. D., & Tygar, J. D. (2014), Adversarial active learning, *in* "Proceedings of the 2014 Workshop on Artificial Intelligent and Security Workshop," ACM, pp. 3–14.

Mitchell, T. (1997), *Machine Learning*, McGraw Hill.

Mitchell, T. M. (2006), The discipline of machine learning, Technical Report CMU-ML-06-108, Carnegie Mellon University.

Moore, D., Shannon, C., Brown, D. J., Voelker, G. M., & Savage, S. (2006), "Inferring internet denial-of-service activity," *ACM Transactions on Computer Systems (TOCS)* **24**(2), 115–139.

Mukkamala, S., Janoski, G., & Sung, A. (2002), Intrusion detection using neural networks and support vector machines, *in* "Proceedings of the International Joint Conference on Neural Networks (IJCNN)," Vol. 2, pp. 1702–1707.

Mutz, D., Valeur, F., Vigna, G., & Kruegel, C. (2006), "Anomalous system call detection," *ACM Transactions on Information and System Security (TISSEC)* **9**(1), 61–93.

Narayanan, A., Shi, E., & Rubinstein, B. I. P. (2011), Link prediction by de-anonymization: How we won the kaggle social network challenge, *in* "Proceedings of the 2011 International Joint Conference on Neural Networks (IJCNN)," IEEE, pp. 1825–1834.

Narayanan, A. & Shmatikov, V. (2008), Robust de-anonymization of large sparse datasets, *in* "Proceedings of the 2008 IEEE Symposium on Security and Privacy," SP '08, IEEE Computer Society, pp. 111–125.

Narayanan, A. & Shmatikov, V. (2009), De-anonymizing social networks, *in* "30th IEEE Symposium on Security and Privacy," pp. 173–187.

Nelder, J. A. & Mead, R. (1965), "A simplex method for function minimization," *Computer Journal* **7**(4), 308–313.

Nelson, B. (2005), Designing, Implementing, and Analyzing a System for Virus Detection, Master's thesis, University of California, Berkeley.

Nelson, B., Barreno, M., Chi, F. J., Joseph, A. D., Rubinstein, B. I. P., Saini, U., Sutton, C., Tygar, J. D., & Xia, K. (2008), Exploiting machine learning to subvert your spam filter, *in* "Proceedings of the 1st USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET)," USENIX Association, pp. 1–9.

Nelson, B., Barreno, M., Chi, F. J., Joseph, A. D., Rubinstein, B. I. P., Saini, U., Sutton, C., Tygar, J. D., & Xia, K. (2009), Misleading learners: Co-opting your spam filter, *in* J. J. P. Tsai & P. S. Yu, eds., *Machine Learning in Cyber Trust: Security, Privacy, Reliability*, Springer, pp. 17–51.

Nelson, B., Dimitrakakis, C., & Shi, E., eds (2013), *Proceedings of the 6th ACM Workshop on Artificial Intelligence and Security, AISec*, ACM.

Nelson, B. & Joseph, A. D. (2006), Bounding an attack's complexity for a simple learning model, *in* "Proceedings of the 1st Workshop on Tackling Computer Systems Problems with Machine Learning Techniques (SysML)" http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.71.9869&rep=rep1&type=pdf.

Nelson, B., Rubinstein, B. I. P., Huang, L., Joseph, A. D., Lau, S., Lee, S., Rao, S., Tran, A., & Tygar, J. D. (2010), Near-optimal evasion of convex-inducing classifiers, *in* "Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)," pp. 549–556.

Nelson, B., Rubinstein, B. I. P., Huang, L., Joseph, A. D., Lee, S. J., Rao, S., & Tygar, J. D., (2012), "Query strategies for evading convex-inducing classifiers," *Journal of Machine Learning Research* **13**(May), 1293–1332.

Nelson, B., Rubinstein, B. I. P., Huang, L., Joseph, A. D., & Tygar, J. D. (2010), Classifier evasion: Models and open problems (position paper), *in* "Proceedings of ECML/PKDD Workshop on Privacy and Security issues in Data Mining and Machine Learning (PSDML)," pp. 92–98.

Newsome, J., Karp, B., & Song, D. (2005), Polygraph: Automatically generating signatures for polymorphic worms, *in* "Proceedings of the IEEE Symposium on Security and Privacy (SP)," IEEE Computer Society, pp. 226–241.

Newsome, J., Karp, B., & Song, D. (2006), Paragraph: Thwarting signature learning by training maliciously, *in* D. Zamboni & C. Krügel, eds., "Proceedings of the 9th International Symposium on Recent Advances in Intrusion Detection (RAID)," Vol. 4219 of *Lecture Notes in Computer Science*, Springer, pp. 81–105.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2016), "Practical black-box attacks against deep learning systems using adversarial examples," *arXiv preprint arXiv:1602.02697* .

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017), Practical black-box attacks against deep learning systems using adversarial examples *in* "Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security (ASIACCS)," ACM, pp. 506–519.

Paxson, V. (1999), "Bro: A system for detecting network intruders in real-time," *Computer Networks* **31**(23), 2435–2463.

Pearson, K. (1901), "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine* **2**(6), 559–572.

Peressini, A. L., Sullivan, F. E., & Jerry J. Uhl, J. (1988), *The Mathematics of Nonlinear Programming*, Springer-Verlag.

Plamondon, R. & Srihari, S. N., (2000), "On-line and off-line handwriting recognition: A comprehensive survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(1), 63–84.

Rademacher, L. & Goyal, N. (2009), Learning convex bodies is hard, *in* "Proceedings of the 22nd Annual Conference on Learning Theory (COLT)," pp. 303–308.

Rahimi, A. & Recht, B. (2008), Random features for large-scale kernel machines, *in* "Advances in Neural Information Processing Systems 20 (NIPS)," pp. 1177–1184.

Ramachandran, A., Feamster, N., & Vempala, S. (2007), Filtering spam with behavioral blacklisting, *in* "Proceedings of the 14th ACM Conference on Computer and Communications Security (CCS)," pp. 342–351.

Rieck, K. & Laskov, P. (2006), Detecting unknown network attacks using language models, *in* R. Büschkes & P. Laskov, eds., "Detection of Intrusions and Malware & Vulnerability Assessment, Third International Conference (DIMVA)," Vol. 4064 of *Lecture Notes in Computer Science*, Springer, pp. 74–90.

Rieck, K. & Laskov, P. (2007), "Language models for detection of unknown attacks in network traffic," *Journal in Computer Virology* **2**(4), 243–256.

Rieck, K., Trinius, P., Willems, C., & Holz, T. (2011), "Automatic analysis of malware behavior using machine learning," *Journal of Computer Security* **19**(4), 639–668.

Ringberg, H., Soule, A., Rexford, J., & Diot, C. (2007), Sensitivity of PCA for traffic anomaly detection, *in* L. Golubchik, M. H. Ammar, & M. Harchol-Balter, eds., "Proceedings of the ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS)," pp. 109–120.

Rivest, R. L., Shamir, A., & Adleman, L. (1978), "A method for obtaining digital signatures and public-key cryptosystems," *Communications of the ACM* **21**(2), 120–126.

Robinson, G. (2003), "A statistical approach to the spam problem," *Linux Journal*, p. 3.

Rubinstein, B. I. P. (2010), Secure Learning and Learning for Security: Research in the Intersection, PhD thesis, University of California, Berkeley.

Rubinstein, B. I. P., Bartlett, P. L., Huang, L., & Taft, N. (2009), "Learning in a large function space: Privacy-preserving mechanisms for SVM learning," *CoRR* **abs/0911.5708**.

Rubinstein, B. I. P., Bartlett, P. L., Huang, L., & Taft, N. (2012), "Learning in a large function space: Privacy-preserving mechanisms for SVM learning," *Journal of Privacy and Confidentiality* **4**(1), 65–100. Special Issue on Statistical and Learning-Theoretic Challenges in Data Privacy.

Rubinstein, B. I. P., Nelson, B., Huang, L., Joseph, A. D., Lau, S., Rao, S., Taft, N., & Tygar, J. D. (2009*a*), ANTIDOTE: Understanding and defending against poisoning of anomaly detectors, *in* A. Feldmann & L. Mathy, eds., "Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement (IMC)," pp. 1–14.

Rubinstein, B. I. P., Nelson, B., Huang, L., Joseph, A. D., Lau, S., Rao, S., Taft, N., & Tygar, J. D. (2009*b*), "Stealthy poisoning attacks on PCA-based anomaly detectors," *SIGMETRICS Performance Evaluation Review* **37**(2), 73–74.

Rubinstein, B. I. P., Nelson, B., Huang, L., Joseph, A. D., Lau, S., Taft, N., & Tygar, J. D. (2008), Compromising PCA-based anomaly detectors for network-wide traffic, Technical Report UCB/EECS-2008-73, EECS Department, University of California, Berkeley.

Rudin, W. (1994), *Fourier Analysis on Groups*, reprint edn, Wiley-Interscience.

Russu, P., Demontis, A., Biggio, B., Fumera, G., & Roli, F. (2016), Secure kernel machines against evasion attacks, *in* "Proceedings of the 2016 ACM Workshop on Artificial Intelligence and Security, (AISec)," pp. 59–69.

Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E. (1998), A Bayesian approach to filtering junk E-mail, *in* "Learning for Text Categorization: Papers from the 1998 Workshop," AAAI Technical Report WS-98-05, Madison, Wisconsin.

Saini, U. (2008), Machine Learning in the Presence of an Adversary: Attacking and Defending the SpamBayes Spam Filter, Master's thesis, University of California at Berkeley.

Schohn, G. & Cohn, D. (2000), Less is more: Active learning with support vector machines, *in* "Proceedings of the 17th International Conference on Machine Learning (ICML)," pp. 839–846.

Schölkopf, B. & Smola, A. J. (2001), *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press.

Sculley, D., Otey, M. E., Pohl, M., Spitznagel, B., Hainsworth, J., & Zhou, Y. (2011), Detecting adversarial advertisements in the wild, *in* "Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)," pp. 274–282.

Sculley, D., Wachman, G. M., & Brodley, C. E. (2006), Spam filtering using inexact string matching in explicit feature space with on-line linear classifiers, *in* E. M. Voorhees & L. P. Buckland,

eds., "Proceedings of the 15th Text REtrieval Conference (TREC)," Special Publication 500-272, National Institute of Standards and Technology (NIST).

Segal, R., Crawford, J., Kephart, J., & Leiba, B. (2004), SpamGuru: An enterprise anti-spam filtering system, *in* "Conference on Email and Anti-Spam (CEAS)" available at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.60.114&rep=rep1&type=pdf.

Settles, B. (2009), Active Learning Literature Survey, Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Shalev-Shwartz, S. & Srebro, N. (2008), SVM optimization: Inverse dependence on training set size, *in* "25th International Conference on Machine Learning (ICML)," pp. 928–935.

Shannon, C. E. (1949), "Communication theory of secrecy systems," *Bell System Technical Journal* **28**, 656–715.

Shannon, C. E. (1959), "Probability of error for optimal codes in a Gaussian channel," *Bell System Technical Journal* **38**(3), 611–656.

Shaoul, C. & Westbury, C. (2007), "A USENET corpus (2005–2007)." Accessed October 2007 at http://www.psych.ualberta.ca/~westburylab/downloads/usenetcorpus.download.html. A more expansive version is available at The Westbury Lab USENET Corpus, https://aws.amazon.com/datasets/the-westburylab-usenet-corpus/.

Shawe-Taylor, J. & Cristianini, N. (2004), *Kernel Methods for Pattern Analysis*, Cambridge University Press.

Smith, A. (2011), Privacy-preserving statistical estimation with optimal convergence rates, *in* "Proceedings of the Forty-Third Annual ACM Symposium on Theory of Computing (STOC)," pp. 813–822.

Smith, R. L. (1996), The hit-and-run sampler: A globally reaching Markov chain sampler for generating arbitrary multivariate distributions, *in* "Proceedings of the 28th Conference on Winter Simulation (WSC)," pp. 260–264.

Somayaji, A. & Forrest, S. (2000), Automated response using system-call delays, *in* "Proceedings of the Conference on USENIX Security Symposium (SSYM)," pp. 185–197.

Sommer, R. & Paxson, V. (2010), Outside the closed world: On using machine learning for network intrusion detection, *in* "Proceedings of the 2010 IEEE Symposium on Security and Privacy," pp. 305–316.

Soule, A., Salamatian, K., & Taft, N. (2005), Combining filtering and statistical methods for anomaly detection, *in* "Proceedings of the 5th Conference on Internet Measurement (IMC)," USENIX Association, pp. 331–344.

Srndic, N. & Laskov, P. (2014), Practical evasion of a learning-based classifier: A case study, *in* "2014 IEEE Symposium on Security and Privacy, SP 2014," pp. 197–211.

Stevens, D. & Lowd, D. (2013), On the hardness of evading combinations of linear classifiers, *in* "Proceedings of the 2013 ACM Workshop on Artificial Intelligence and Security (AISec'13)," pp. 77–86.

Stolfo, S. J., Hershkop, S., Wang, K., Nimeskern, O., & Hu, C.-W. (2003), A behavior-based approach to securing email systems, *in Mathematical Methods, Models and Architectures for Computer Networks Security*, Springer-Verlag, pp. 57–81.

Stolfo, S. J., Li, W., Hershkop, S., Wang, K., Hu, C., & Nimeskern, O. (2006), Behavior-based modeling and its application to Email analysis, *in* "ACM Transactions on Internet Technology (TOIT)," pp. 187–211.

Sweeney, L. (2002), "*k*-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* **10**(5), 557–570.

Tan, K. M. C., Killourhy, K. S., & Maxion, R. A. (2002), Undermining an anomaly-based intrusion detection system using common exploits, *in* A. Wespi, G. Vigna, & L. Deri, eds., "Proceedings of the 5th International Symposium on Recent Advances in Intrusion Detection (RAID)," Vol. 2516 of *Lecture Notes in Computer Science*, Springer, pp. 54–73.

Tan, K. M. C., McHugh, J., & Killourhy, K. S. (2003), Hiding intrusions: From the abnormal to the normal and beyond, *in* "Revised Papers from the 5th International Workshop on Information Hiding (IH)," Springer-Verlag, pp. 1–17.

Torkamani, M. & Lowd, D. (2013), Convex adversarial collective classification, *in* "Proceedings of the 30th International Conference on Machine Learning ICML," pp. 642–650.

Torkamani, M. A. & Lowd, D. (2014), On robustness and regularization of structural support vector machines, *in* "Proceedings of the 31st International Conference on Machine Learning (ICML-14)," pp. 577–585.

Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., & Ristenpart, T. (2016), Stealing machine learning models via prediction apis, *in* "Proceedings of the 25th USENIX Security Symposium," pp. 601–618.

Tukey, J. W. (1960), "A survey of sampling from contaminated distributions," *Contributions to Probability and Statistics* pp. 448–485.

Turing, A. M. (1950), "Computing machinery and intelligence," *Mind* **59**(236), 433–460.

Valiant, L. G. (1984), "A theory of the learnable," *Communications of the ACM* **27**(11), 1134–1142.

Valiant, L. G. (1985), Learning disjunctions of conjunctions, *in* "Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)," pp. 560–566.

Vapnik, V. N. (1995), *The Nature of Statistical Learning Theory*, Springer-Verlag.

Venkataraman, S., Blum, A., & Song, D. (2008), Limits of learning-based signature generation with adversaries, *in* "Proceedings of the Network and Distributed System Security Symposium (NDSS)," The Internet Society available at http://www.isoc.org/isoc/conferences/ndss/08/papers/18_limits_learning-based.pdf.

Wagner, D. (2004), Resilient aggregation in sensor networks, *in* "Proceedings of the Workshop on Security of Ad Hoc and Sensor Networks (SASN)," pp. 78–87.

Wagner, D. & Soto, P. (2002), Mimicry attacks on host-based intrusion detection systems, *in* "Proceedings of the 9th ACM Conference on Computer and Communications Security (CCS)," pp. 255–264.

Wang, K., Parekh, J. J., & Stolfo, S. J. (2006), Anagram: A content anomaly detector resistant to mimicry attack, *in* D. Zamboni & C. Krügel, eds., "Proceedings of the 9th International Symposium on Recent Advances in Intrusion Detection (RAID)," Vol. 4219 of *Lecture Notes in Computer Science*, Springer, pp. 226–248.

Wang, K. & Stolfo, S. J. (2004), Anomalous payload-based network intrusion detection, *in* E. Jonsson, A. Valdes, & M. Almgren, eds., "Proceedings of the 7th International Conference on Recent Advances in Intrusion Detection (RAID)," Vol. 3224 of *Lecture Notes in Computer Science*, Springer, pp. 203–222.

Wang, Y.-X., Fienberg, S. E., & Smola, A. J. (2015), Privacy for free: Posterior sampling and stochastic gradient Monte Carlo, *in* "ICML," pp. 2493–2502.

Wang, Y.-X., Lei, J., & Fienberg, S. E. (2016), "Learning with differential privacy: Stability, learnability and the sufficiency and necessity of ERM principle," *Journal of Machine Learning Research* **17**(183), 1–40.

Wang, Z., Fan, K., Zhang, J., & Wang, L. (2013), Efficient algorithm for privately releasing smooth queries, *in* "Advances in Neural Information Processing Systems," pp. 782–790.

Wang, Z., Josephson, W. K., Lv, Q., Charikar, M., & Li, K. (2007), Filtering image spam with near-duplicate detection, *in* "Proceedings of the 4th Conference on Email and Anti-Spam (CEAS)" available at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.94.9550&rep=rep1&type=pdf.

Warrender, C., Forrest, S., & Pearlmutter, B. (1999), Detecting intrusions using system calls: Alternative data models, *in* "Proceedings of the IEEE Symposium on Security and Privacy (SP)," IEEE Computer Society, pp. 133–145.

Wittel, G. L. & Wu, S. F. (2004), On attacking statistical spam filters, *in* "Proceedings of the 1st Conference on Email and Anti-Spam (CEAS)" available at https://pdfs.semanticscholar.org/af5f/4b5f8548e740735b6c2abc1a5ef9c5ebf2df.pdf.

Wyner, A. D. (1965), "Capabilities of bounded discrepancy decoding," *Bell System Technical Journal* **44**, 1061–1122.

Xiao, H., Biggio, B., Brown, G., Fumera, G., Eckert, C., & Roli, F. (2015), Is feature selection secure against training data poisoning?, *in* "Proceedings of the 32nd International Conference on Machine Learning, ICML 2015," pp. 1689–1698.

Xu, H., Caramanis, C., & Mannor, S. (2009), "Robustness and regularization of support vector machines," *Journal of Machine Learning Research* **10**(Jul), 1485–1510.

Xu, W., Bodík, P., & Patterson, D. A. (2004), A flexible architecture for statistical learning and data mining from system log streams, *in* "Proceedings of Workshop on Temporal Data Mining: Algorithms, Theory and Applications at the 4th IEEE International Conference on Data Mining (ICDM)" available at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.135.7897&rep=rep1&type=pdf.

Zhang, F., Chan, P. P. K., Biggio, B., Yeung, D. S., & Roli, F. (2016), "Adversarial feature selection against evasion attacks," *IEEE Transactions of Cybernetics* **46**(3), 766–777.

Zhang, J., Zhang, Z., Xiao, X., Yang, Y., & Winslett, M. (2012), "Functional mechanism: Regression analysis under differential privacy," *Proceedings of the VLDB Endowment* **5**(11), 1364–1375.

Zhang, Y., Ge, Z., Greenberg, A., & Roughan, M. (2005), Network anomography, *in* "Proceedings of the 5th ACM SIGCOMM Conference on Internet Measurement (IMC)," USENIX Association, Berkeley, CA, USA, pp. 317–330.

Zhang, Z., Rubinstein, B. I. P., & Dimitrakakis, C. (2016), On the differential privacy of Bayesian inference, *in* "Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI'2016)," pp. 51–60.

Zhao, W.-Y., Chellappa, R., Phillips, P. J., & Rosenfeld, A. (2003), "Face recognition: A literature survey," *ACM Computing Surveys* **35**(4), 399–458.