in these two instances. The optimal objective value for the defender if the attacker attacks h_1 (i.e., when $p = H$) is $\epsilon_2+\bar{f}_1(1-\epsilon_2)-(\bar{f}_1-f_1)\gamma_{12}+H[(1-\bar{f}_1(\epsilon_1-\epsilon_2)+(\bar{f}_1-f_1)\gamma_{12})]$, while the optimal setting of p when $O(h_2)$ is increasing in p is $p = 0$, yielding the objective value of $\epsilon_2 + \bar{f}_2(1 - \epsilon_2) - (\bar{f}_2 - f_2)\gamma_{21}$. Consequently, $p^* = 0$ in this case i ff

$$(\bar{f}_2 - f_2)\gamma_{21} - (\bar{f}_2 - \bar{f}_1)(1 - \epsilon_2) \leq (1 - H)(\bar{f}_1 - f_1)\gamma_{12} - H(1 - \bar{f}_1)(\epsilon_1 - \epsilon_2). \quad (11)$$

To help understand the implication of Condition 11 it is useful to consider what is implied by Case III: $\bar{H} < 0$ and $\underline{H} < 0$ suggests that the dominant term in these expressions is $(\bar{f}_1 - f_1)\gamma_{12}$. Thus, the attacker, when he manipulates h_1, can rather effectively distinguish instances which are classified correctly, and those which are not. As a result, we can expect the right-hand-side to be relatively large, and the condition to be typically true. Since h_2 is already better (by our assumption) in terms of baseline (non-adversarial) performance, it is in such a case a clear preference for the defender. Thus, for Case III, just as in Cases I and II, we can conclude that *typically* we expect $p^* = 0$.

6.3 Discussion

Despite considering several different models of manipulation in the context of indiscriminate attacks, our results are relatively consistent: facing this class of attacks, the defender will often eschew randomization entirely, and simply use the classifier with better baseline performance. This is rather dramatically in contrast with the situation we observed when the attacks are targeted, in which case baseline performance of the classifiers had no impact at all, and the defender would simply choose among classifiers uniformly at random.

7. CONCLUSION

We studied optimal randomized classification both when attacks are targeted and when they are indiscriminate, and obtained almost completely different answers: in the former, the defender will randomize uniformly between the classifiers, while in the latter an optimal scheme would rarely randomize at all. In reality, attacks do not neatly fall into one of these stylized classes. In some settings, such as spam, attacks tend towards indiscriminate, while in spear phishing the attacks are more targeted. System designers using machine learning tools in adversarial settings should therefore consider carefully which stylized model is best aligned with their predicament, and use our insights as guidelines to best determine operational security posture.

8. REFERENCES

[1] Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

[2] Marco Barreno, Blaine Nelson, Anthony D. Joseph, and J.D. Tygar. The security of machine learning. *Machine Learning*, 81:121–148, 2010.

[3] Battista Biggio, Giorgio Fumera, and Fabio Roli. Adversarial pattern classification using multiple classifiers and randomisation. In *Lecture Notes in Computer Science*, pages 500–509, 2008.

[4] Battista Biggio, Giorgio Fumera, and Fabio Roli. Multiple classifier systems for adversarial classification tasks. In *Eighth International Workshop on Multiple Classifier Systems*, pages 132–141, 2009.

[5] Michael Brückner and Tobias Scheffer. Nash equilibria of static prediction games. In *Advances in Neural Information Processing Systems*, pages 171–179, 2009.

[6] Michael Brückner and Tobias Scheffer. Stackelberg games for adversarial prediction problems. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 547–555, New York, NY, USA, 2011. ACM.

[7] Richard Colbaugh and Kristin Glass. Predictive defense against evolving adversaries. In *IEEE International Conference on Intelligence and Security Informatics*, pages 18–23, 2012.

[8] Nilesh Dalvi, Pedro Domingos, Mausam, Sumit Sanghai, and Deepak Verma. Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 99–108, 2004.

[9] P. Fogla and W. Lee. Evading network anomaly detection systems: Formal reasoning and practical techniques. In *ACM Conference on Computer and Communications Security*, pages 59–68, 2006.

[10] Ling Huang, Anthony D. Joseph, Blaine Nelson, Benjamin I.P. Rubinstein, and J. D. Tygar. Adversarial machine learning. In *Fourth ACM workshop on Security and artificial intelligence*, pages 43–58, 2011.

[11] Manish Jain, Jason Tsai, James Pita, Christopher Kiekintveld, S. Rathi, Milind Tambe, and Fernando Ordonez. Software assistants for randomized patrol planning for the lax airport police and the federal air marshals service. *Interfacs*, 40:267–290, 2010.

[12] S. Jajodia, A.K. Ghosh, V.S. Subrahmanian, V. Swarup, C. Wang, and X.S. Wang, editors. *Moving Target Defense II*. Springer, 2013.

[13] S. Jajodia, A.K. Ghosh, V. Swarup, C. Wang, and X.S. Wang, editors. *Moving Target Defense*. Springer, 2011.

[14] Murat Kantarcioğlu, Bowei Xi, and Chris Clifton. Classifier evaluation and attribute selection against active adversaries. *Data Min. Knowl. Discov.*, 22(1-2):291–335, January 2011.

[15] Wei Liu and Sanjay Chawla. Mining adversarial patterns via regularized loss minimization. *Machine Learning*, 81:69–83, 2010.

[16] Daniel Lowd and Christopher Meek. Adversarial learning. In *ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pages 641–647, 2005.

[17] Daniel Lowd and Christopher Meek. Good word attacks on statistical spam filters. In *Conference on Email and Anti-Spam*, 2005.

[18] Blaine Nelson, Benjamin I.P. Rubinstein, Ling Huang, Anthony D. Joseph, Steven J. Lee, Satish Rao, and J.D. Tyger. Query strategies for evading convex-inducing classifiers. *Journal on Machine Learning Research*, 13:1293–1332, 2012.

[19] Blaine Nelson, Benjamin I.P. Rubinstein, Ling Huang, Anthony D. Joseph, and J.D. Tygar. Classifier evasion: Models and open problems. In *Workshop on Privacy and Security Issues in Data Mining and Machine Learning*, pages 92–98, 2010.

[20] P. Paruchuri, J. Pearce, Janus Marecki, and Milind Tambe. Playing games for security: An efficient exact algorithm for solving Bayesian Stackelberg games. In *Seventh International Conference on Autonomous Agents and Multiagent Systems*, pages 895–902, 2008.

[21] Robin Sommer and Vern Paxson. Outside the closed world: On using machine learning for network intrusion detection. In *IEEE Symposium on Security and Privacy*, pages 305–316, 2010.