Next Generation Trustworthy Fraud Detection

Sihong Xie* Philip S. Yu†
*Lehigh University, six316@lehigh.edu †University of Illinois at Chicago, psyu@uic.edu

***Abstract*—Popular web applications, such as e-commerce, social networks and online ad auction, are providing valuable services to web users but have also been plagued by prevalent and diverse frauds. Many detection methodologies have been devised but detection trustworthiness is still one important and yet missing desideratum: a user will not trust a detector that has uncertain accuracy, can malfunction under unexpected situations, or can't explain its behaviors and interal working. Previous efforts mostly focused on detection accuracy, and our goal is to chart a path towards a more comprehensive definition of trustworthy detection, that consists of accuracy, transparency, and proactivity. To achieve the goal, we identify key challenges rooting at the specific settings of the above applications: the evolving nature and unexpectedness of the fraudsters' strategies, the ever-growing large amount of data, and the increasing complexity of effective detectors. We hope spark a large volume of research questions and solutions with respect to the above challenges.**

I. Introduction

Many web applications, such as e-commerce, social networks, and ad auctions, are now indispensable components, as they make valuable information more accessible to a wider audience by connecting many entities. For example, Yelp and Amazon connect millions of customers and merchants to allow convenient evaluation and search of merchandise; social networks, such as Facebook, Whatsapp, and Twitter, connect millions of people, businesses and other organizations [1] and allow the fluid exchange of information; online ad networks and exchanges, such as Google Marketing Platform [2], connect publishers, advertisers, and web users for more effective ad distribution and tracking to benefit all parties. However, dishonest users (the *fraudsters* or *attackers*) are abusing these applications for various malicious purposes. On Yelp, Amazon and many other review websites, opinion spammers are posting fake reviews to hijack the opinions of the genuine customers, and unfairly promote and defame the ratings of their targets; on Whatsapp, terrorists are exploiting the encrypted communication channel to plan terrorist attacks; on Twitter, false information can be created and then propagated to a large number of audience to influence public opinions. Also commonly found on online ad networks and exchanges are Ad fraudsters, who are creating fake websites to deceive the advertisers into spending money on ads that no one will click. These malevolent actors and activities are undermining the efficiency and effectiveness of these platforms, jeopardizing the well-being of individuals, businesses and the whole society [39], [1].

[1]According to a 2018 Statista report

[2]https://www.blog.google/products/marketingplatform/360/introducing-google-marketing-platform/

To mitigate the adversarial effects, many fraud detection mechanisms (the *defender*) are designed to prevent and spot suspicious activities and entities. Effective detectors aim at signals and models that can identify frauds with high precision and recall rates. For example, in [22], near-duplicates of contents are considered as suspicious activities and could lead to the detection of spams; in [46], [49], detection signals are derived from streaming data to monitor suspicious dynamic changes; in [3], [44], [32], [21], data are represented by heterogeneous networks that connect multiple types of entities, such as accounts and products, and suspicious connection patterns, such as dense blocks, can be detected.

Besides more effective signals and models, another approach is to collect and analyze large-scale genuine fraudster data to shed light on fraudster characteristics [27], [26], [45], [37], [38], [36], [23]. The understanding can provide new insight and lead to more accurate detection. In [26], [27], [45], honeypots are deployed to collect social spams; in [37], [38], [36], the authors seized real botnets to understand fraudsters' operations; in [23], real email spams are collected and analyzed. However, it usually takes a long time and huge effort to seize such data, which are typically securely protected by fraudsters.

The above two approaches are reactive and can deter the frauds only after they happened while proactive approaches can bring more detection trustworthiness. In particular, one can simulate the behaviors and tactic of the fraudsters so that the defender can anticipate unseen but likely attacks in the future. Based on the simulation data, detectors can be patched for more robust detections. This idea has been widely applied in software security under the name "vulnerability analysis". For example, in [25], vulnerability analysis is conducted by simulating penetration of fraudsters to an ad auction network; in [13], vulnerabilities of an IDS (Intrusion Detection System) are revealed by penetration test. Adversarial machine learning has just started receiving the deserved attention most recently [43], [7], [28], [41], [6]. Vulnerability analysis is less conducted in social network and product review applications with a few exceptions in [21], [50], [14], [4].

Transparent detectors are also more trustworthy since the humans who operate the detectors can understand how a decision is made and how to correct wrong detections. More generally, explainable AI (XAI) becomes a surging research area [2], [18], [20], [17] due to the demand for "a right to explanation" [19] and robust and reliable decision making in safety-critical applications like self-driving cars [31], [40], [5]. The goals of XAI is to provide human-interpretable information regarding the outcomes and workings of an AI model. Major approaches are based on sensitivity analysis [29] and model approximation [35], based on which fraud detectors based on SVM and deep neural networks can be explained. Since the connections among entities are important for fraud detection, detectors based on graphs are indispensable in any effective defense. Explaining graphical models has also been explored [30], [9], [48], [16], [11], [10], [15], where the focus is on the differential analysis of Bayesian or Markov networks.

II. New challenges and solutions

The detectors, when viewed as a component running inside the larger applications, are facing new trustworthiness challenges besides detection accuracy.

A. Transparency in detection

To handle the increasing level of attack sophistication, more advanced and effective detectors are required. However, these models naturally become more complicated as more complex attack behaviors are considered and modeled. At some point, the users of the detectors start having difficulties in understanding the detection process and outcomes. However, detection transparency is important in security applications. First, for the detection operators to confidently adopt a detector, they need to know where the detector is likely to fail, and when it fails, what are the root causes of the failures. Also, as the operators need to make the final decision regarding more serious or larger-scale security issues, the decisions from the detectors should be made transparent to allow the operators to reason about the fairness and reasonableness of the decisions. A real-world case is that, when a botnet is detected in an ad exchange, the security team needs to investigate the detection and further take actions to shutdown the botnet. Lastly, the detection outcomes will affect the operation of the hosting applications and their users. Frequently, besides a detection outcome, how the decision is arrived at should be communicated to the end-users. For example, Yelp may need to explain to a reviewer or a business why a review is deleted.

How to introduce transparency to a detector depends the data and model that the detector handles. We are interested in transparent detectors that make structured predictions using graphical models, that can flexibly model anomalies in graph and sequence data commonly found in the above applications. We propose to explain detection based on inference algorithms, such as message passing [42], through join sub-graph mining and approximation. When deep graphical models are used [12], decisions are also made based on the underlying deep networks and thus the transparency can be arrived at by combining sub-graph mining and sensitivity analysis.

B. Proactivity in detection

As the fraudsters continue to exploit new vulnerabilities in the defender, the detectors' strategies are lagging behind the attackers' [24]. The implication is that a reactive defender can be evaded by unexpected attacks even if it can be patched after the vulnerabilities are discovered. If this happens frequently, users will lose their trust in the defender. To deal with this incompetence, proactive detectors have been proposed [8], [28], [33], [34], [43], [7], [28], [41], [6]. However, these models operate only on vectorial data and can't handle heterogeneous and structured data. Most recently, adversarial learning on sequences and graphs are proposed [14], [50]. However, there are still many challenges. First, if reinforcement learning is used to discover vulnerabilities [14], then one has to specify the design of reward function and the representation of the states, address the time complexity in computing reward functions and a large number of states, etc. Second, as the amount of data is increasing, training proactive detectors is much more time consuming than training regular detectors. For example, re-training while searching for adversarial examples involve a bi-level optimization problem on a large dataset. Third, it is rarely possible to obtain full information of the fraudsters, and any strong assumption about their behaviors will result in a defense that can be easily penetrated when the fraudsters change their behaviors. Lastly, while previous work usually explicitly or implicitly assumed that there is only one adversary, there are many groups of fraudsters with diverse behaviors. We believe that there are many research opportunities in designing proactive defense on the increasing complicated attack-defense scenarios.

To address the above challenges, inverse reinforcement learning can be used to learn the reward function without specifying it. Representing a graph as a state can be challenging and graphical convolutional neural networks can learn a vector representation of a graph. Sparsity in the data should be exploited to speed up the training of proactive detectors. Attack-agnostic defense is more robust and can be achieved by identifying universal properties of the underlying detectors and application constraints. Lastly, multi-agent reinforcement learning or mean-field [47] is a promising direction to model multiple attackers that collaborate or compete in committing frauds.

C. Achieving transparency for proactive defense

Proactive detectors add another layer of complexity to traditional models, and established model transparency can be again in jeopardy. First, due to limited data, the proactive detectors are not perfect and have their own vulnerabilities, and discovering new vulnerabilities can be non-trivial. For example, revealing the vulnerabilities of a regular SVM is different from that of a robust SVM with bi-level optimization [8]. Second, with the added layer of proactivity, the completeness and interpretability of the explanations of the reactive counterpart (e.g., plain SVMs) need to be redefined. Here completeness refers to the degree to which a model can be explained, while interpretability indicates how easy an explanation can be understood by a human being. For example, shall we include the set of parameters of the attacker in an explanation of the above robust SVM? How does the inclusion affect or improve interpretability? The combination of the increasing need for transparent and proactive detectors is complicating the task of trustworthy fraud detection.

For proactive models obtained through re-training [28], the reactive and proactive models have different decision boundaries but can be explained through the same methods. We propose a comparative approach, which juxtaposes the explanations of the reactive and proactive models and explain what makes the difference in the two models. A user can then decide whether and how the added layer of proactivity makes more sense.

III. Conclusion

We reviewed existing literature on fraud detection in applications, such as e-commerce, social networks and ad exchange, where frauds are prevalent and critical. We pointed out three desiderata of a trustworthy fraud detection system in these applications, namely, accuracy, transparency, and proactivity. While the detection accuracy can be addressed by considering more factors and designing more complicated models, the transparency and proactivity shall be attacked from different perspectives. We discussed technical challenges in addressing these desiderata.

References

[1] Health care reform:health insurance & affordable care act. http://www.webmd.com/health-insurance/insurance-basics/using-doctor-ratings-sites.

[2] Ashraf M. Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan S. Kankanhalli. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *CHI*, 2018.

[3] Leman Akoglu, Rishi Chandy, and Christos Faloutsos. Opinion fraud detection in online reviews by network effects. ICWSM, 2013.

[4] Scott Alfeld, Xiaojin Zhu, and Paul Barford. Explicit defense actions against test-set attacks. In *AAAI Conference on Artificial Intelligence*, AAAI, 2017.

[5] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete Problems in AI Safety. Technical report, 2016.

[6] Battista Biggio, Ignazio Pillai, Samuel Rota Bulò, Davide Ariu, Marcello Pelillo, and Fabio Roli. Is Data Clustering in Adversarial Settings Secure? In *Proceedings of the 2013 ACM Workshop on Artificial Intelligence and Security*, AISec '13, pages 87–98, New York, NY, USA, 2013. ACM.

[7] Michael Brückner, Christian Kanzow, and Tobias Scheffer. Static prediction games for adversarial learning problems. *J. Mach. Learn. Res.*, 13(1):2617–2654, September 2012.

[8] 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.

[9] E. Castillo, J. M. Gutierrez, and A. S. Hadi. Sensitivity analysis in discrete bayesian networks. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 27(4):412–423, July 1997.

[10] Hei Chan and Adnan Darwiche. Sensitivity analysis in bayesian networks: From single to multiple parameters. In *UAI*, 2004.

[11] Hei Chan and Adnan Darwiche. Sensitivity analysis in markov networks. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, IJCAI'05, pages 1300–1305. Morgan Kaufmann Publishers Inc., 2005.

[12] Liang-Chieh Chen, Alexander G Schwing, Alan L Yuille, and Raquel Urtasun. Learning Deep Structured Models. ICML, 2015.

[13] Robert K. Cunningham, R. P. Lippmann, David J. Fried, Simson L. Garfinkel, Isaac Graf, Kris R. Kendall, Seth E. Webster, Daniel Wyschogrod, and Marc A. Zissman. Evaluating intrusion detection systems without attacking your friends : The 1998 darpa intrusion detection evaluation. 1999.

[14] Hanjun Dai, Hui Li, Tian Tian, Xin Huang, Lin Wang, Jun Zhu, and Le Song. Adversarial attack on graph structured data. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1115–1124, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[15] Adnan Darwiche. A Differential Approach to Inference in Bayesian Networks. *J. ACM*, 50(3):280–305, may 2003.

[16] Jasper De Bock, Cassio P de Campos, and Alessandro Antonucci. Global Sensitivity Analysis for MAP Inference in Graphical Models. In Z Ghahramani, M Welling, C Cortes, N D Lawrence, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2690–2698. Curran Associates, Inc., 2014.

[17] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. 2017.

[18] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning. Technical report, 2018.

[19] Bryce Goodman and Seth Flaxman. Eu regulations on algorithmic decision-making and a "right to explanation", 2016.

[20] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5):93:1–93:42, August 2018.

[21] Bryan Hooi, Hyun Ah Song, Alex Beutel, Neil Shah, Kijung Shin, and Christos Faloutsos. Fraudar: Bounding graph fraud in the face of camouflage. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD, pages 895–904, New York, NY, USA, 2016. ACM.

[22] Nitin Jindal and Liu Bing. Analyzing and detecting review spam. ICDM, 2007.

[23] Chris Kanich, Christian Kreibich, Kirill Levchenko, Brandon Enright, Geoffrey M. Voelker, Vern Paxson, and Stefan Savage. Spamalytics: An empirical analysis of spam marketing conversion, 2008.

[24] James Kaplan, Shantnu Sharma, and Allen Weinberg. https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/meeting-the-cybersecurity-challenge, 2011.

[25] Carmelo Kintana, David Turner, Jia-Yu Pan, Ahmed Metwally, Neil Daswani, Erika Chin, and Andrew Bortz. The goals and challenges of click fraud penetration testing systems. In *International Symposium on Software Reliability Engineering*, 2009.

[26] Kyumin Lee, James Caverlee, and Steve Webb. Uncovering social spammers: Social honeypots + machine learning. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR, 2010.

[27] Kyumin Lee, Brian Eoff, and James Caverlee. Seven months with the devils: A long-term study of content polluters on twitter. In *International AAAI Conference on Web and Social Media*, 2011.

[28] Bo Li and Yevgeniy Vorobeychik. Scalable Optimization of Randomized Operational Decisions in Adversarial Classification Settings. In *AISTATS*, 2015.

[29] Jiwei Li, Will Monroe, and Daniel Jurafsky. Understanding neural networks through representation erasure. *CoRR*, 2016.

[30] David Madigan, Krzysztof Mosurski, and Russell G Almond. Graphical Explanation in Belief Networks. *Journal of Computational and Graphical Statistics*, 6(2):160–181, 1997.

[31] Clemens Otte. Safe and Interpretable Machine Learning: A Methodological Review. In *Computational Intelligence in Intelligent Data Analysis*, pages 111–122. Springer Berlin Heidelberg, 2013.

[32] Shashank Pandit, Duen Horng Chau, Samuel Wang, and Christos Faloutsos. Netprobe: A Fast and Scalable System for Fraud Detection in Online Auction Networks. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 201–210, New York, NY, USA, 2007. ACM.

[33] N Papernot, P McDaniel, X Wu, S Jha, and A Swami. Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 582–597, may 2016.

[34] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The Limitations of Deep Learning in Adversarial Settings. *CoRR*, abs/1511.07528, 2015.

[35] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. KDD, pages 1135–1144. ACM, 2016.

[36] Brett Stone-Gross, Ryan Abman, Richard A. Kemmerer, Christopher Krügel, and Douglas G. Steigerwald. The underground economy of fake antivirus software. In *WEIS*, 2011.
[37] Brett Stone-Gross, Marco Cova, Lorenzo Cavallaro, Bob Gilbert, Martin Szydlowski, Richard Kemmerer, Christopher Kruegel, and Giovanni Vigna. Your botnet is my botnet: Analysis of a botnet takeover. CCS, 2009.
[38] Brett Stone-Gross, Thorsten Holz, Gianluca Stringhini, and Giovanni Vigna. The underground economy of spam: A botmaster's perspective of coordinating large-scale spam campaigns. LEET, 2011.
[39] New York Times. Charges settled over fake reviews on itunes. http://www.nytimes.com/2010/08/27/technology/27ftc.html.
[40] Kush R. Varshney and H. Alemzadeh. On the Safety of Machine Learning: Cyber-Physical Systems, Decision Sciences, and Data Products. *Big Data Journal*, 2017.
[41] Yevgeniy Vorobeychik and Bo Li. Optimal randomized classification in adversarial settings. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*, AAMAS '14, pages 485–492, Richland, SC, 2014. International Foundation for Autonomous Agents and Multiagent Systems.
[42] Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1-2), January 2008.
[43] Gang Wang, Tianyi Wang, Haitao Zheng, and Ben Y Zhao. Man vs. machine: Practical adversarial detection of malicious crowdsourcing workers. In *23rd USENIX Security Symposium Security 14)*, pages 239–254, San Diego, CA, 2014. USENIX Association.
[44] Guan Wang, Sihong Xie, Bing Liu, and Philip S. Yu. Identify online store review spammers via social review graph. *ACM Trans. Intell. Syst. Technol.*, 3(4):61:1–61:21, 2012.
[45] Steve Webb, James Caverlee, and Calton Pu. Social Honeypots: Making Friends With A Spammer Near You. In *CEAS*, 2008.
[46] Sihong Xie, Guan Wang, Shuyang Lin, and Philip S. Yu. Review spam detection via temporal pattern discovery. KDD, 2012.
[47] Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. Mean Field Multi-Agent Reinforcement Learning. In *ICML*. PMLR, 2018.
[48] Ghim-Eng Yap, Ah-Hwee Tan, and Hwee-Hwa Pang. Explaining inferences in Bayesian networks. *Applied Intelligence*, 29(3):263–278, dec 2008.
[49] L. Zhang and Y. Guan. Detecting click fraud in pay-per-click streams of online advertising networks. In *2008 The 28th International Conference on Distributed Computing Systems*, pages 77–84, June 2008.
[50] Daniel Zügner, Amir Akbarnejad, and Stephan Günnemann. Adversarial attacks on neural networks for graph data. In *KDD*, 2018.