Research Statement

A. Pinar Ozisik
pinar@cs.umass.edu

To safely deploy systems in real world applications, we must guarantee that they will behave correctly. This is a difficult problem because these systems often contain some degree of uncertainty due to the realities of training data that underlie them (e.g., in the case of artificial intelligence applications), or the use of randomized algorithms for performance gains (e.g., in the case of blockchain applications for cryptocurrencies). To address this, we must ensure that we are designing systems that are robust against uncertainty.

The foundation of my research is to analyze the robustness of models to inform and improve the development of systems with practical, real world applications. To this end, I have worked on the following questions, which I later refer to, in reinforcement learning (RL) and blockchain systems: I) how can we ensure robustness, II) how can robustness be measured, and III) what should a system be robust against.

Robustness is an overloaded term with different definitions in statistics, artificial intelligence (AI), and systems. For estimation problems using data, robust statistics aim to minimize errors due to outliers or minor deviations from distributional assumptions [12]. In AI, robust models are those capable of withstanding uncertainty introduced during data generation or at any stage in the data processing pipeline. In the context of systems security, robustness refers to computer programs that are designed to offer protection against an *attacker model*. Inspired by this definition, adversarial machine learning studies learning in the presence of an adversary. I have already leveraged some tools and methodologies from these domains, and plan to utilize more in the future.

An important statistical tool I have used in both areas of my research are *concentration inequalities* (CIs). Also known as *tail inequalities* or *tail bounds*, CIs bound the probability that a random variable deviates from some value, typically its expectation or median. For blockchain systems, robustness implies mathematical safety, which requires CIs to ensure that training data does not include too many outliers, thereby representing the true underlying distribution of the data (Q-I). For RL, on top of this already existent requirement, I argue that training data must also be *secure*, i.e., robust against data points that do not come from the true underlying distribution. Using this definition, I created a new measure of security to quantify the susceptibility of training data to corruptions (Q-II). These corruptions represent the worst possible data points added to training data by an adversary with a specific agenda. These safety and security requirements arise from domain-specific needs: cryptocurrencies must be robust against scams and attacks, and inferences made from data must be robust against anomalies (Q-III).

To summarize, in both areas of my work, robustness has included concepts that are both generalizable and domain-specific, and has become interchangeable with safety and security. In the future, I would like to explore how robustness relates to fairness, trustworthiness, and even interpretability. In doing so, I plan to continue answering the previously mentioned core questions in other application domains.

1 Robustness in Safe/Seldonian RL

It is crucial to develop robust RL algorithms especially because they have been proposed for many high-risk applications, such as improving type 1 diabetes and sepsis treatments

[13, 24]. One type of *safe RL* algorithm [22, 23], also referred to as *safe and/or Seldonian RL*, addresses this problem by ensuring *safety*, which provides high-confidence guarantees that the application will not cause undesirable behavior like increasing the frequency of dangerous patient outcomes. A specific component, called the *safety test*, can output a new *policy*—the mechanism for selecting actions within an agent—using training data collected from a baseline policy. It uses CIs, most recently the *Chernoff-Hoeffding* (CH) [11] bound, to lower bound the performance of the new policy. If this lower bound is higher than the performance of the baseline policy, the new policy is output. Our analysis of the safety test answers to what extent the safety guarantee holds when some of the samples in the training data are not random.

Non-random samples, caused by anomalies, are common when data comes from a pipeline that includes human interactions, natural language processing, device malfunctions, etc. For e.g., the recent application of RL to sepsis treatment in the intensive care unit (ICU) used training data generated from hand-written doctors' notes [13]. In a high-stress ICU environment, missing records and poorly written notes are difficult to automatically parse [1]. This can lead to errors in training data with dire consequences in the case of sepsis, which is a life-threatening medical condition.

Contributions. Our formulation of the problem represents anomalies as samples artificially created by a malicious attacker. The incorporation of an adversary provides a worst case analysis to understand the robustness of safety tests to anomalies in data; because any algorithm robust to an adversary is also robust to non-adversarial anomalies in data. First, we introduce a new measure of security to quantify the susceptibility of training data to corruptions. Second, we show that a couple of safe RL methods are extremely sensitive to even a few data corruptions, completely breaking the probability bounds guaranteed by CIs. To fix this problem, we introduce a new algorithm, called `Panacea` [21], which provides a user-specified level of robustness when the number of non-random samples in the dataset is upper bounded. We also demonstrate the advantages of using `Panacea` in practice on a diabetes treatment simulation.

2 Robust Blockchain Systems

The *blockchain* [16] concept was originally designed to be the backbone of the Bitcoin distributed cryptographic currency. Over the last decade, Bitcoin has been adopted more widely for e-commerce than any previous digital currency. With a market capitalization of $935 billion [7], Bitcoin's value has made it a target for malicious attacks. Therefore, just like any credit or debit card, Bitcoin must be robust against theft and scams.

2.1 Background

To understand specific attacks against Bitcoin, first we briefly describe how blockchains work. In Bitcoin, *transactions* (txns), similar to those generated by financial institutions, are recorded on a public ledger called a blockchain. Txns are broadcast by users on Bitcoin's peer-to-peer network. A subset of users, known as *miners*, independently agglomerate a set of txns into a candidate *block* and attempt to solve a predefined cryptographic puzzle as proof of work. The first miner to solve the problem broadcasts their solution to the network,

and is able to add the block to the blockchain as a child of the prior block. The miners then start over, using the newly appended blockchain and the set of remaining txns.

Bitcoin's cryptographic puzzle consists of the miners' applying a hash algorithm [10], which takes as input a block and a random number, called a *nonce*. If the resulting value, i.e., the *hash*, is not less than the known *target* that is set by the network, then a new nonce is selected to compute a new hash. This process repeats until some miner finds a solution. Each time a hash is computed, a miner samples a value from a discrete uniform distribution; hence, the likelihood of *discovering* a block increases with a miner's *hash rate* or *mining power*, i.e., the number of the hashes computed per second. With the published block and nonce, any miner can verify that the resulting hash is less than the target.

A *double-spend attack* [16] occurs when a malicious miner, acting as a customer, tries to steal goods from a merchant by purposely creating a *fork* on the chain. They release a txn, x, that transfers money to the merchant and wait for the blockchain, including x, to grow; but then they attempt to rewrite the chain, excluding x, once they receive the goods. The malicious miner can succeed only if they have enough mining power: all miners will switch over to the new chain if it is longer. Hence, the status of txns and blocks is not immutable—a fork of blocks supported by greater mining power can emerge at any time.

2.2 Estimating Mining Power

The distributed nature of blockchains makes the system partially observable. If each miner knew the mining power of others, then they could compute the likelihood of a double-spend attack, and thus, the stability of the current blockchain. Consequently, we seek to quantify the hash rate of miners in real time, thereby computing the likelihood of the blockchain's mutability. In other words, we measure how robust the system is to changes by estimating the miners' mining power.

Contributions. We first design a method of accurate hash rate estimation based on compact *status reports* [19] issued by miners. These reports require each miner to periodically report the block and nonce resulting in the minimum hash value since the last block broadcast on the chain or their last report. Second, we show how hash rates can be estimated from only blocks that are published to the blockchain. To find a lower and upper bound on the number of hashes computed per miner, i.e., reduce errors in hash rate estimation, we use a well-known CI, the *Chernoff* bound [6], for our first method, and calculate empirical bounds based on *bootstrapping* [5, 8] for our second method.

2.3 Minimizing Communication Cost

In addition to double-spend attacks, block propagation delays cause forks on the blockchain. These delays are mainly due to block size: it takes longer for miners to receive large blocks, and hence, to discard current work. This causes blocks with the same parent to be broadcast on the network, causing forks. In order to avoid this situation, we answer the question of how to quickly relay information, i.e., a block of txns, from a sender to a receiver if the receiver already possesses all or some of the txns. Similar to our previous project, our goal is to ensure the robustness—or stability—of the blockchain by minimizing block size.

Contributions. Our proposed solution, instead of sending all txns directly, uses a novel combination of two *probabilistic data structures*: a *Bloom filter* [4] and an *Invertible Bloom Lookup Table* (IBLT) [9]. These data structures use hash functions to compactly represent a set of items. The widely held assumption that a good hash function appears random [15] means that these data structures can fail. This failure can occur when a receiver, who obtains a Bloom filter or IBLT representation of a set, tests whether an item at hand is a part of the set represented by the data structure. If the membership test for that item returns negative when it is actually a part of the set, errors occur. This error rate is tunable with a tradeoff: a smaller error rate means that the data structure becomes bigger. Our protocol, `Graphene` [18, 20], minimizes the total size of both data structures sent between a sender and receiver given a low error rate. We use Chernoff bounds to set the size of our probabilistic data structures in order to guarantee `Graphene`'s performance with high probability.

3 Future Directions

Measuring and increasing robustness against uncertainty in two different systems helped me develop an understanding of fundamental statistical concepts, which I will utilize in future projects. Meanwhile, I also plan to apply different mathematical analyses used in other areas and develop novel methodologies that can be generalizable. The context to pursue these plans include high-risk applications with potential consequences that are harmful to society.

Addressing real world problems. Three years ago, I created a curriculum for first year students and taught a course called Ethical Issues in Technology, where we discussed the ethical ramifications of computing. This course fueled my already existent interest in social justice, shifting my focus to AI systems with social implications. Recently, there has been abundant work on identifying bias and discrimination in AI systems, and creating robust models [2, 3, 17]. Safe RL has also been proposed for creating fair algorithms that reduce discriminative behavior in intelligent tutoring systems and loan approvals [14]. I plan to contribute to this line of research by creating robust systems that can tackle such problems.

Expanding application domains. By collaborating with domain experts and working with data from different application domains such as medicine, self-driving cars and predictive policing, I am interested in developing metrics that evaluate the "quality" or limitations of training data, which often reflect systemic biases and human error. I believe that domain-specific phenomena in data will eventually help us develop domain-agnostic methodologies.

Developing/applying mathematical techniques. I plan to explore mathematical techniques to study robustness that include but are not limited to: leveraging tools from robust statistics and adversarial ML, formulating and analyzing worst or average case scenarios to mimic the behavior of a given system, and adding random noise to data or input parameters.

References

[1] Daniel Albright, Arrick Lanfranchi, Anwen Fredriksen, William F. Styler IV, Colin Warner, Jena D. Hwang, Jinho D. Choi, Dmitriy Dligach, Rodney D. Nielsen, James Martin, et al. Towards comprehensive syntactic and semantic annotations of the clinical narrative. *Journal of the American Medical Informatics Association*, 20(5):922–930, 2013.

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine Bias. 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[3] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

[4] Burton H. Bloom. Space/Time Trade-offs in Hash Coding with Allowable Errors. *Commun. ACM*, 13(7):422–426, July 1970.

[5] George Casella and Roger L. Berger. *Statistical inference*. Brooks Cole, Pacific Grove, CA, 2002. ISBN 0-495-39187-5. URL http://opac.inria.fr/record=b1134456.

[6] Herman Chernoff et al. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4):493–507, 1952.

[7] Brandon Chez. Today's cryptocurrency prices by market cap. https://coinmarketcap.com/, Last retrieved December 2021.

[8] Bradley Efron. *The jackknife, the bootstrap and other resampling plans*. Society for industrial and applied mathematics (SIAM), 1982.

[9] Michael Goodrich and Michael Mitzenmacher. Invertible bloom lookup tables. In *Conf. on Comm., Control, and Computing*, pages 792–799, Sept 2011.

[10] Hashcash. https://en.bitcoin.it/wiki/Hashcash, Last retrieved June 2017.

[11] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[12] Peter J Huber. *Robust statistics*, volume 523. John Wiley & Sons, 2004.

[13] Matthieu Komorowski, Leo A. Celi, Omar Badawi, Anthony C. Gordon, and A. Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716, 2018.

[14] Blossom Metevier, Stephen Giguere, Sarah Brockman, Ari Kobren, Yuriy Brun, Emma Brunskill, and Philip S Thomas. Offline contextual bandits with high probability fairness guarantees. In *Advances in Neural Information Processing Systems*, pages 14893–14904, 2019.

[15] Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge university press, 2017.

[16] Satoshi Nakamoto. Bitcoin: A Peer-to-Peer Electronic Cash System, May 2009. URL https://bitcoin.org/bitcoin.pdf.

[17] Safiya Umoja Noble. *Algorithms of oppression*. New York University Press, 2018.

[18] A Pinar Ozisik, Gavin Andresen, George Bissias, Amir Houmansadr, and Brian Levine. Graphene: A new protocol for block propagation using set reconciliation. In *Data Privacy Management, Cryptocurrencies and Blockchain Technology*, pages 420–428. Springer, 2017.

[19] A Pinar Ozisik, George Bissias, and Brian Levine. Estimation of miner hash rates and consensus on blockchains. *arXiv preprint arXiv:1707.00082*, 2017.

[20] A Pinar Ozisik, Gavin Andresen, Brian N Levine, Darren Tapp, George Bissias, and Sunny Katkuri. Graphene: efficient interactive set reconciliation applied to blockchain propagation. In *Proceedings of the ACM Special Interest Group on Data Communication*, pages 303–317. 2019.

[21] Pinar Ozisik and Philip Thomas. Security analysis of safe and seldonian reinforcement learning algorithms. *Advances in neural information processing systems*, 2020.

[22] P. S. Thomas, G. Theocharous, and M. Ghavamzadeh. High confidence policy improvement. In *International Conference on Machine Learning*, 2015.

[23] Philip S. Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[24] Philip S. Thomas, Bruno Castro da Silva, Andrew G. Barto, Stephen Giguere, Yuriy Brun, and Emma Brunskill. Preventing undesirable behavior of intelligent machines. *Science*, 366 (6468):999–1004, 2019.