Research Plans

R. Teal Witter

I am a theoretical computer scientist studying algorithms for social good. As computing permeates our lives, algorithms wield immense power to drive societal progress. However, this potential is accompanied by risks of unintended consequences. My research aims to bridge the gap between theoretical foundations and practical applications, ensuring that algorithms designed for social good are not only effective but also trustworthy and explainable. To this end, my work focuses on the following interconnected goals:

1. **Explainable AI**: Developing methods to accurately interpret opaque models. For example, enabling a credit card applicant to understand why their application was approved or rejected.
2. **Responsible Use of AI**: Implementing safeguards to prevent AI misuse, such as ensuring AI-generated content is traceable to its source.
3. **Effective Algorithms**: Creating trustworthy algorithms to address problems with social impact. This includes enabling nonprofits to accurately evaluate the effects of their initiatives.

The growing importance of algorithms for social good is reflected in the development of recent venues, including the FAccT conference and social impact tracks at major machine learning conferences like AAAI and IJCAI. Many algorithms proposed for social good in these venues and others are heuristic in nature. My work aims to illuminate their theoretical foundations, providing A) **rigorous guarantees on algorithmic performance and behavior**, and B) **theoretical insights for the design of more effective and trustworthy algorithms**. By bridging the gap between theory and practice, I strive to enhance the reliability and impact of algorithms designed for social good.

I have studied algorithms for social good in the context of explainable AI [MW24, LWK$^+$24], evaluation of nonprofit efficacy [WM24], fairness in machine learning [RW23, RW24], resource allocation [HLW22, WR24, WH24], and societal polarization [MRUW22]. I leverage a broad theoretical toolkit including techniques in randomized linear algebra, linear programming, and discrete optimization. Research in my area also requires interdisciplinary engagement with practitioners and stakeholders. To this end, I have worked closely with an early childhood literacy nonprofit and collaborated with researchers across nine institutions, publishing in top machine learning and theoretical computer science venues such as NeurIPS, ESA, and AAAI.

In the remainder of this document, I outline my research plans as they relate to the three overarching goal.

Provably Accurate Algorithms for Explainable AI

Game theory provides a valuable axiomatic framework for explaining AI predictions. While most AI applications of game-theoretic concepts use heuristic methods, I previously developed theoretically grounded algorithms for computing Shapley and Banzhaf values with strong non-asymptotic guarantees, and superior empirical performance. My future research aims to build efficient algorithms for computing other game-theoretic quantities, focusing on applications that promote social good.

Motivation

As AI predictions are increasingly incorporated into high-stakes domains like finance, law, and healthcare, users and auditors of AI systems should understand why a prediction was made. For example, a credit card applicant should know why their application was rejected, and a defendant should be aware of how their bail was set. Further, explaining AI predictions can help identify biases and interpret learned concepts, supporting the improvement and refinement of future algorithms.

The standard method of explaining predictions is to use game-theoretic quantities like Shapley values, quantifying how changes to the input features affect the model output. Shapley values are a particularly popular choice because they satisfy four desirable properties: null player, symmetry, additivity, and efficiency. However, Shapley values are defined over exponentially terms and are therefore computationally infeasible to compute exactly in general. In practice, Shapley values are estimated using heuristics approaches. Arguably the most popular and effective estimator for Shapley values is Kernel SHAP. Kernel SHAP exploits an elegant connection to linear regression, sampling from an exponentially large linear system to estimate the Shapley values. While effective in practice, Kernel SHAP samples heuristically and lacks theoretical guarantees.

My Prior Work

I developed a theoretically motivated algorithm for estimating Shapley values that outperforms Kernel SHAP in practice and offers strong non-asymptotic guarantees [MW24]. The starting point of my work was an insight from randomized linear algebra: regression problems can be effectively subsampled using statistical leverage scores. The benefit of leverage scores is that they quantify the importance of each data point in the regression problem. Instead of sampling from the exponentially large linear system via a heuristic weighting as in Kernel SHAP, the Leverage SHAP algorithm, as its name suggests, samples according to leverage scores. This sampling technique is theoretically motivated and offers strong non-asymptotic guarantees. Modifying the standard leverage score analysis to the specific optimizations in the algorithm, I showed that Leverage SHAP can provably recover accurate Shapley values with almost a linear number of samples. Further, not only does leverage score sampling offer theoretical guarantees, but the theoretically motivated Leverage SHAP also outperforms even a highly optimized version of Kernel SHAP empirically.

While Shapley values are popular, they are not the only game-theoretic quantity that can be used to explain AI predictions. Instead of the efficiency property, we may want another application-dependent property. For example, in the context of a loan applicant, the 2-efficiency property ensures that the attribution of a composite feature like net worth is the sum of the attributions of the sub-features like assets and liabilities. If we replace the efficiency property of Shapley values with the 2-efficiency property, we arrive at the related Banzhaf values. Banzhaf values are simpler than Shapley values and have been found to be more accurately computed in practice. I wondered whether I could apply the leverage score sampling technique to Banzhaf values. While Shapley values are known to be the solution to a linear regression problem, this formulation was only known for Banzhaf values in a restricted setting. So, in order to apply the leverage score sampling technique to Banzhaf values, I designed a linear regression problem for which the Banzhaf values are the solution [LWK+24]. The resulting algorithm, Kernel Banzhaf, substantially outperforms the existing Banzhaf value computation methods. Further, because of the structure of the Banzhaf linear regression problem, Kernel Banzhaf offers even stronger non-asymptotic guarantees than Leverage SHAP.

Future Directions

My prior work establishes more efficient and theoretically motivated methods for explaining AI predictions with Shapley and Banzhaf values. However, there are still many more game-theoretic quantities which are relevant in different social good applications. I plan to apply my theoretical toolkit to design efficient algorithms for computing these game-theoretic quantities, furthering the transparency of AI predictions.

One under-studied social good setting is graph tasks, such as predicting the spread of disease or identifying collusion rings. Graph neural networks have emerged as a powerful tool for learning on graph data, processing the features of adjacent nodes to identify local patterns. Because of the high stakes of social good applications, it is important to explain how graph neural networks make predictions. Unfortunately, standard game-theoretic attribution quantities like Shapley and Banzhaf values do not take into account the graph structure, and so lose the ability to explain how the graph neural network reasons. An alternative game-theoretic quantity designed specifically for graph structures is the Hamiache-Navarro (HN) value, which naturally generalizes Shapley values to graph settings. Mathematically, the HN value is the limit of a series of associated games, which can be represented as repeated matrix multiplication. This connection to matrix multiplication suggests that the HN value may be computed via a gradient descent algorithm. I plan to investigate the structure of the HN value to recover the underlying problem that gives rise to the gradient descent algorithm. If the problem is a linear regression problem as I suspect, I can apply leverage score sampling to design provably accurate algorithms for computing the HN value.

There is a rich game theory literature to describe attribution techniques from an axiomatic perspective. As trustworthy AI becomes increasingly important, this literature is a powerful resource for explaining AI predictions in an axiomatic way. But game theorists are interested in formulating these quantities, rather than computing them efficiently. The majority of prior work that adapts these quantities to AI applications develops heuristic algorithms. However, as evidenced by my work on Shapley and Banzhaf values, theoretically motivated algorithms can outperform heuristic methods, while simultaneously offering strong non-asymptotic guarantees. I plan to apply my theoretical toolkit to design provably efficient algorithms for computing game-theoretic quantities relevant to social good applications.

Collaboration Opportunities

The problems in explainable AI present a rich landscape of research opportunities, spanning game theory, graph algorithms, and randomized linear algebra. I am excited to collaborate with other theoretical computer scientists at ETH Zurich ITS, building theoretically motivated algorithms with provable performance guarantees.

Distortion-free Watermarking for the Responsible Use of AI

I previously developed distortion-free watermarking techniques to ensure the responsible use of AI-generated images, but with storage that scales with use. By using cryptographic hash functions and locality-sensitive hashing, I plan to design secure and robust watermarks that can be detected without degrading content quality or requiring costly storage. My approach applies to both text and vision tasks, allowing for efficient watermarking that leverages existing information in the generated content, ensuring scalability, security, and robustness.

Motivation

As AI models become more advanced, powerful tools like Large Language Models (LLMs) for text generation and diffusion models for prompt-guided image generation are now ubiquitous. While these technologies offer vast applications, they also bring new risks, such as malicious actors claiming AI-generated text as their own or fabricating realistic images of fake events, potentially causing confusion or harm. To mitigate these risks, model owners use watermarking techniques to track the content generated by their models.

However, most current watermarking methods are distortion-based, meaning they modify the output to embed identifiable markers. For text, the distribution of words is often altered while, for images, the distribution of an associated latent image is often modified. Despite their widespread use, distortion-based watermarks remain vulnerable: they can be detected and even forged by malicious actors if enough examples are available.

My Related Work

I am interested in distortion-free watermarking techniques that generate content without modifying it. In this approach, verification of the watermark requires access to a private, correlated variable, ensuring the watermark's security and robustness against forgery. However, a significant limitation of current distortion-free methods is the need for the model owner to store the correlated variable, which can be both costly and inefficient.

In my work, I developed a distortion-free watermarking method for diffusion models [AFW+24]. The key idea is to generate initial noise using a finite set of seeds, where each seed is linked to a cryptographic hash function. Even if an adversary can reconstruct the noise, it remains indistinguishable from random noise and provides no information about other seeds. The main challenge lies in efficiently storing and searching through all possible noise configurations.

Future Directions

My goal is to develop distortion-free watermarks that scale without the need for costly storage. The key idea is to leverage context to robustly and securely generate a seed using locality-sensitive hashing (LSH), specifically SimHash. This would allow us to generate a correlated random variable from the seed in a secure manner, without distorting the distribution of the generated content.

In the LLM setting, text is generated in an auto-regressive manner by predicting the next token based on the previous context. Current watermarking approaches modify this distribution, either globally or contextually, making the watermarks detectable, removable, and sometimes degrading text quality. I plan to use SimHash to convert the embedding of the prompt into seeds and then generate a random variable using cryptographic hash functions. This random variable is reproducible if we have access to the seed, and distributed identically to a sample from the true probability distribution over next tokens. To detect the watermark, we compute the correlated random variable for each seed and check its alignment with the generated text. By using SimHash, we can guarantee that nearby embedded contexts produce the same seed with high probability, making the watermark detectable without degrading text quality.

For vision tasks, a similar approach can be applied by embedding the prompt and using SimHash to derive multiple seeds. Each seed generates randomness for different portions of the image. During detection, we caption the image to obtain a vector aligning with the original prompt, apply SimHash to retrieve seeds, and check for alignment with the latent noise. This method allows for flexible tuning of hyperparameters to achieve the desired level of detection accuracy, while the cryptographic hash function ensures security.

By combining distortion-free and searchable watermarking with streaming and randomized algorithms, this approach promises both scalability and efficiency. I plan to leverage information already present in the image or text to robustly and securely store the correlated variable in the generated content, enabling efficient distortion-free watermarking and ultimately supporting the responsible use of AI.

Collaboration Opportunities

The problems in watermarking present fascinating algorithmic questions. I am excited to collaborate with researchers at ETH Zurich ITS, leveraging expertise in cryptography, randomized algorithms, and hashing to develop distortion-free watermarking techniques that are both secure and scalable.

Simple and Trustworthy Algorithms for Treatment Effect Estimation

I previously designed simple and trustworthy algorithms for treatment effect estimation, motivated by the need for transparent and reliable evaluations of social programs. My work addresses challenges in natural experiment settings by developing interpretable, theoretically grounded algorithms that offer non-asymptotic *guarantees, as demonstrated in collaboration with a nonprofit organization. I aim to extend this research into active regression settings to design efficient estimators with minimal negative impact.*

Motivation

In broader societal applications, such as government spending or nonprofit resource allocation, explainability becomes even more critical. It's not enough to explain individual predictions; stakeholders should have confidence in the entire model's transparency and reasoning. This need for transparency extends to the realm of treatment effect estimation, an important problem in evaluating the impact of social programs. While randomized control trials often allow us to estimate the effect of a treatment, they're not always possible or ethical to implement. In some cases, certain individuals may have a greater need for the treatment, or the treatment may have already been assigned, leaving us only to observe the outcome. These scenarios, known as natural experiments, are common in social good applications and require sophisticated analytical approaches because the treatment assignment can be confounded by other factors.

One such application is evaluating the impact of nonprofit programs. For instance, Reach Out and Read Colorado (RORCO), an early childhood literacy nonprofit, provides books to children during their pediatrician visits. RORCO has been operating for more than 20 years, distributing books to the least-served students in Colorado. Because of their extensive history, RORCO has a wealth of data on the children they have served and the impact of their program but lacks a robust method for their natural experiment setting. In collaborating with RORCO, I applied more than 20 existing treatment effect estimation methods. However, each method yielded different results on their data and non offered theoretical guarantees.

My Related Work

Building on the challenges identified in evaluating nonprofit programs like RORCO's, I embarked on developing a practical testbed for treatment effect estimation. This work was motivated by the need for robust, interpretable methods that can provide reliable insights in natural experiment settings. To address this gap, I curated a comprehensive dataset and constructed a benchmark for evaluating treatment effect estimators. This empirical investigation yielded valuable insights, notably that doubly robust algorithms generally provide the best performance in estimating treatment effects. However, these algorithms often suffer from two significant drawbacks: they tend to be quite uninterpretable, and their theoretical guarantees are primarily asymptotic, limiting their practical applicability in finite sample scenarios. Leveraging these findings, I designed a simple, theoretically motivated algorithm and exactly analyzed its variance in the non-asymptotic setting. This new approach not only offers performance comparable to the more complex doubly robust algorithms but does so with a simpler, more interpretable estimator. The algorithm's effectiveness is not just theoretical; RORCO has already implemented it to inform their future program development, demonstrating its practical value.

Future Directions

While treatment effect estimation is well-studied, there are practically no algorithms with non-asymptotic user-friendly guarantees. My goal is to develop simple algorithms with understandable theoretical guarantees for treatment effect estimation. Like the guarantees I developed for Shapley and Banzhaf estimators, the guarantees would be of the form: with poly$(m, 1/\epsilon, 1/\delta)$ samples, we can guarantee ϵ-approximate estimates with probability $1 - \delta$. This approach not only enhances the reliability of the estimates but also allows stakeholders to easily comprehend the impact of their programs.

A further avenue of research is to develop new estimators for the active regression setting that draw on the connection between regression and treatment effect estimation. While natural experiment methods utilize existing data to approximate program impact, the quality of estimates is inherently limited by the confounded nature of observational data. To address this limitation, I plan to extend my treatment effect estimation work to the active regression setting. In this context, individuals are actively selected to receive either the ’treatment’ or ’control’ condition. I plan to apply theoretical tools to design efficient estimators that can achieve the same level of accuracy as prior work but with fewer selected individuals, limiting the negative impact to individuals that are excluded from the nonprofit’s treatment.

Collaboration Opportunities

The problems in treatment effect estimation pose many interesting statistical questions across estimator design, causal inference, and algorithmic analysis. I am excited to collaborate with researchers at ETH Zurich ITS, leveraging their expertise to develop simple and provably accurate algorithms for treatment effect estimation.

Conclusion

While my main research has centered on algorithms for social good, I remain curious about new topics in theoretical computer science and consider myself a generalist. For instance, I have worked on algorithms for efficiently evaluating Boolean functions in both classical [HKLW22] and quantum settings [CKW23, KW21, DKW19]. I have even explored board games through an algorithmic and complexity lens [WL20, Wit21], which sparked my interest in the field. Looking ahead, I hope to continue exploring new areas beyond my current interests at ETH Zurich ITS.

The growing integration of AI and algorithmic decision-making into socially significant domains demands that we develop both effective and trustworthy methods. My research bridges the gap between theoretical computer science and practical, socially relevant applications, ensuring that algorithms for social good are not only efficient but also explainable, secure, and interpretable. Through my work on explainable AI, responsible AI usage via watermarking, and treatment effect estimation, I aim to address some of the most pressing challenges posed by modern algorithmic systems.

By focusing on provably accurate algorithms, secure watermarking techniques, and transparent evaluation methods, my goal is to contribute to the development of more robust, ethical, and impactful technologies. I look forward to collaborating with fellow researchers at ETH Zurich ITS to further the reach of these ideas and broadly benefit society.

References

An asterisk (*) indicates that authors are listed in alphabetical order.

[AFW^{+}24] Kasra Arabi, Benjamin Feuer, R Teal Witter, Chinmay Hegde, and Niv Cohen. Hidden in the noise: Two-stage robust watermarking for images. In *Submission*, 2024.

[CKW23] Michael Czekanski*, Shelby Kimmel*, and R Teal Witter*. Robust and space-efficient dual adversary quantum query algorithms. In *European Symposium on Algorithms*, 2023.

[DKW19] Kai DeLorenzo*, Shelby Kimmel*, and R Teal Witter*. Applications of the quantum algorithm for st-connectivity. In *Conference on the Theory of Quantum Computation, Communication and Cryptography*, 2019.

[HKLW22] Lisa Hellerstein*, Devorah Kletenik*, Naifeng Liu*, and R Teal Witter*. Adaptivity gaps for the stochastic boolean function evaluation problem. In *Workshop on Approximation and Online Algorithms*, 2022.

[HLW22] Lisa Hellerstein*, Thomas Lidbetter*, and R Teal Witter*. A local search algorithm for the min-sum submodular cover problem. In *International Symposium on Algorithms and Computation*, 2022.

[KW21] Shelby Kimmel* and R Teal Witter*. A query-efficient quantum algorithm for maximum matching on general graphs. In *Algorithms and Data Structures Symposium*, pages 543–555, 2021.

[LWK$^+$24] Yurong Liu*, R Teal Witter*, Flip Korn, Tarfah Alrashed, Dimitris Paparas, and Juliana Freire. Kernel banzhaf: A fast and robust estimator for banzhaf values. In *Submission*, 2024.

[MRUW22] Christopher Musco*, Indu Ramesh*, Johan Ugander*, and R Teal Witter*. How to quantify polarization in models of opinion dynamics. In *International Workshop on Mining and Learning with Graphs*, 2022.

[MW24] Christopher Musco* and R Teal Witter*. Provably accurate shapley value estimation via leverage score sampling. In *Submission*, 2024.

[RW23] Lucas Rosenblatt and R Teal Witter. Counterfactual fairness is basically demographic parity. In *AAAI Conference on Artificial Intelligence*, 2023.

[RW24] Lucas Rosenblatt* and R. Teal Witter*. Fairlyuncertain: A comprehensive benchmark of uncertainty in algorithmic fairness. In *Submission*, 2024.

[WH24] R Teal Witter and Lisa Hellerstein. Minimizing cost rather than maximizing reward in restless multi-armed bandits. In *Submission*, 2024.

[Wit21] R Teal Witter. Backgammon is hard. In *International Conference on Combinatorial Optimization and Applications*, 2021.

[WL20] R Teal Witter and Alex Lyford. Applications of graph theory and probability in the board game ticket to ride. In *International Conference on the Foundations of Digital Games*, 2020.

[WM24] R Teal Witter and Christopher Musco. Benchmarking estimators for natural experiments: A novel dataset and a doubly robust algorithm. In *Conference on Neural Information Processing Systems*, 2024.

[WR24] R Teal Witter and Lucas Rosenblatt. I open at the close: A deep reinforcement learning evaluation of open streets initiatives. In *AAAI Conference on Artificial Intelligence*, 2024.