Research Statement

Kai Zhou
Computer Science and Engineering, Washington University in St. Louis
zhoukai@wustl.edu

Computer scientists are good at designing complex intelligent systems. These systems consume a massive amount of data and produce meaningful results to fulfill various functionalities. For example, Social Network Analysis (SNA) tools take a network of social interactions as input and make knowledgeable inferences, such as social ties of important individuals. To achieve such functionalities, an SNA analyst would first collect data to construct a social network, and then feed this network into SNA algorithms, which might be executed by remote cloud servers due to local processing limitations. Finally, analytical results are obtained, possibly from the servers.

Regardless of the functionalities of systems, we observe a *data flow*: data are collected from individuals, fed into systems for computation, and translated into knowledge. **Security issues emerge from the full life-span of data**:

1 *Data reliability*: what if data are not reliably collected? Can the system be robust when data are *adversarially* noisy?
2 *Data security and privacy*: when systems are executed by untrusted parties, can we preserve data integrity and confidentiality?
3 *Knowledge dependability*: after knowledge is produced from data, can we guarantee the soundness of knowledge, especially when it is obtained from untrusted parties?

My research addresses these security issues arisen in different phases of data usage, focusing on a spectrum of widely-used systems or their abstracted models. Specifically, my recent works [1, 2] tackle data reliability for SNA in an adversarial environment, investigating attacks that leverage corrupting the data collection process, and exploring possible defense strategies. Most of my past works [4–7] have the goal of ensuring data security and privacy as well as knowledge dependability in the mobile cloud computing environment, providing secure outsourcing and result verification schemes for a variety of computational tasks. These research efforts offer useful insights and general guidance towards achieving data security from those three aspects. My current and future works extend the spectrum to broader application scenarios. Through continuous research in this field, I envision an *end-to-end* robust, secure, and dependable intelligent system: a system that is robust to adversarial data input; a system that securely consumes user data; a system that produces dependable knowledge.

Summary of Past Research

***Adversarial Social Network Analysis**:* My recent research focuses on adversarial robustness of Social Network Analysis (SNA), specifically on *link prediction*, addressing the data reliability issue. Link prediction is a fundamental problem in SNA, with the goal of predicting hidden or future links in a partially observed network. Thus, reliably collecting data to construct a network is crucial in the first place. However, insofar as link prediction may reveal relationships which associated parties prefer to keep hidden, it introduces incentives to corrupt the data collection process with the ultimate goal of misleading link prediction. To capture this *adversarial* environment, we propose an abstraction of the data collection process (Fig. 1): an *analyst* submits a set of node-pair queries to the *environment*, which returns "edge" or "non-edge" in response to each query; an *attacker* modifies the query results, based on which the analyst constructs a manipulated subgraph. Our goal is to conduct link prediction reliably under such *network manipulation attacks*.

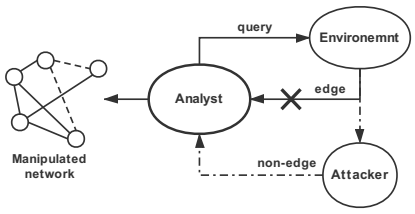

Fig. 1: Network manipulation attack.

a) Network manipulation attack in link prediction. While the ultimate goal is to develop defense strategies, our first step [1] is to analyze the vulnerability of link prediction: to what extent an attacker can mislead link prediction via manipulating network structure? A common approach to predicting a target link (u, v) is to infer the likelihood of its existence using a measure of *similarity* of u and v computed from the observed network. We thus study the attacks where an attacker aims to minimize the weighted sum of similarities of a set of target links, with the ability to delete certain links from the observed subgraph. We investigate a variety of common similarity metrics and formulate the attacks as combinatorial optimization problems, covering a collection of settings: e.g., whether the target set contains a single or multiple links, and whether the similarity metrics are local or global. Our theoretical discoveries are thorough and insightful: except for the simplest case where the attacker attacks a single target with local similarity metrics, in all other cases finding the optimal manipulated network structure under budget constraint is NP-hard. Given that optimal attacks are hard to find, we proceed to seek sub-optimal solutions to those challenging combinatorial optimization problems. By leveraging the combinatorial structures, we develop a set of approximation algorithms (some with theoretical guarantees), resulting in efficient and effective attacks. Our studies show that network manipulation attack indeed imposes a severe security threat to link prediction.

b) Adversarial robustness of link prediction. Our second step [2] is to design robust link prediction methods under the threat of network manipulation attack. Adhering to the previous adversarial data collection model, we now enable the analyst to make a subset of queries *reliable* by, e.g., investing more resources on those queries. In essence, this is a problem of optimally allocating security resources facing strategical attackers. We formulated it as a non-zero-sum Stackelberg game: a defender (the analyst) first chooses a subset of reliable queries, and an attacker then decides which query results to manipulate. We design utility functions for both the defender and attacker, taking into account their own *losses* in link prediction performance. A crucial challenge is that the defender is uncertain about both the attacker's target link(s) H_A and the true underlying network structure $\mathcal{G}$. We thus divide the attackers into types, jointly determined by H_A and $\mathcal{G}$, and assume the defender has knowledge of a prior distribution over these types. This results in a *Bayesian* Stackelberg Game, where the defender now maximizes the expected utility over attacker types. Seeking the *strong Stackelberg equilibrium* of this game is extremely challenging – finding the defender's optimal strategy involves solving an integer bi-level program. In fact, we proved that the inner problem of seeking the attacker's best-response is NP-hard. Our main technical contribution thus lies in approximately solving this game, via reasonable relaxations on the players' utility functions. Through comprehensive simulations, we demonstrate that strategically allocating limited security resources significantly mitigates the damage caused by intelligent attackers, greatly enhancing the robustness of link prediction under network manipulation attack.

Social network analysis faces a unique challenge that the (abstracted) network structure is prone to manipulation in the cyberspace, often with reasonably low costs. Through the works in [1] and [2], we demonstrated that the quality of SNA (specifically link prediction as for our task) is at risk when data are not reliably collected, and developed an analytical framework for enhancing the adversarial robustness of SNA. Specifically, the interactions between the defender and the attack are modeled as a game with a defender minimizing a worst-case (caused by the attacker's strategical actions) loss. The powerful modeling capability of game theory enables us to integrate the *task-specific* (e.g., link prediction) objectives into players' utilities. This game theoretical framework provides a promising analytic tool to develop robust systems for other important SNA tasks, such as influence maximization and community detection.

***Secure & Verifiable Computation Outsourcing**:* While intelligent systems often need enormous computational power to process data, in many cases, the data owners such as individuals are constrained by their local computational capability. This conflict between the demand and availability of computational resources triggers a new computing paradigm (Fig. 2) that is already prevalent in our daily lives. In this *computation outsourcing* framework, *end-users* (individuals or small businesses) outsource their data and computational tasks to a central pool of storage and computation resources (e.g., *cloud servers*). When data are shared with these untrusted third parties, two main security threats stand out: 1) data security and user privacy are at risk; 2) the quality of service, specifically the soundness of computational results, cannot be guaranteed. The common approach in the literature is to first transform a task T (often via data encryption) into a disguised task T'; the server then solves T' to produce a result S', which is then recovered (and possibly verified) to the final results S of T by the end-user. However, this procedure imposes great challenges: encryption ensures data security but almost destroys data utility as computing over encrypted data is extremely expensive, and the limited local computational power greatly constrains the problem transformation and result verification processes.

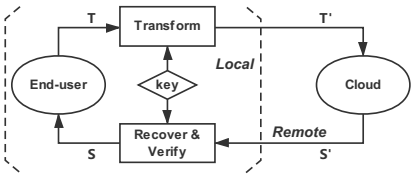

Fig. 2: Computation outsourcing

While enabling *any* computation in the ciphertext domain is practically impossible – *fully homomorphic encryption* is still the holy grail in cryptography – my research focuses on the secure and verifiable computation on some widely used computational building blocks: system of linear & non-linear equations [6] , cryptographic computations [5], biometric authentication [4], and cryptographic access control [7]. The philosophy is to strike a balance between security and efficiency: we identify the specific security requirements in different application scenarios and develop lightweight encryption techniques to satisfy those minimum security requirements.

a) Outsourcing cryptographic computations. Modern public-key cryptography is built upon computational hardness assumptions, one of which is the hardness of discrete logarithm. These crypto-systems require efficiently computing exponential operations in a finite field, such as modular exponentiation (e.g., $g^a \bmod p$), scalar multiplication on elliptic curves, and bilinear pairings. Using such techniques to ensure user data security imposes significant computational challenges, especially for mobile devices. An appealing option is to outsource the cryptographic computations to untrusted servers provided that the security of data can be guaranteed. To securely outsource cryptographic computation, we observe that, despite the huge differences in these exponential operations, they can somehow be decomposed into a series of *modular additions* and *modular multiplications* in a certain finite field $\mathbb{F}_p$. We thus propose a secure mapping approach that maps (via a secret one-time random number $r \in \mathbb{F}_p$) an element $x \in \mathbb{F}_p$ to an element X in a larger finite field $\mathbb{F}_N$, where N is the product of two primes p and q. We can then conduct addition and multiplication with X (outsourced to a server) in this larger field $\mathbb{F}_N$ and recover the results to the smaller field $\mathbb{F}_p$. The nice property is that without knowing r, it is computationally infeasible to recover elements in $\mathbb{F}_N$ to those in $\mathbb{F}_p$. Via this approach, we can *securely* compute modular addition and multiplication with the

help of an untrusted server. They serve as the building blocks for constructing protocols of securely outsourcing high-level cryptographic computations. The other important task is to verify the results returned by the server. Our approach is to map x to two correlated elements X_1 and X_2 via two independent random keys r_1 and r_2. Then the two results returned by the server should satisfy certain algebraic relations, which can be efficiently verified locally. Overall, we built up a secure and verifiable scheme for outsourcing the ubiquitous exponential operations to a single untrusted server. We also demonstrated that this secure computation block could be integrated into higher-level applications such as digital signature and identity-based encryption.

b) Privacy-preserving biometric authentication. A critical security issue in biometric authentication is that individuals need to fully trust the central authority for managing their private biometric templates such as fingerprints. In [4], we proposed a user-centric privacy-preserving biometric authentication scheme, with the key feature that end-users encrypt their templates locally such that remote authority sees nothing but the ciphertext. This, however, results in several challenges, including ensuring efficient local side encryption and enabling authentication in the ciphertext domain. Specifically, authentication requires deciding if two templates are close enough, by computing their distance and comparing it with a pre-defined threshold. Conventional secure computation techniques are not adequate for such a "compute-then-compare" paradigm, as they treat computing and comparison as two separate processes; moreover, the untrusted server now can access the intermediate computational results (i.e., distance information), which are leveraged in several known attacks. Our main technical contribution is a novel and lightweight encryption process that integrates computing and comparison into one single computational process. Specifically, from two encrypted templates, the server can only recover a comparison result, which is just the information needed for authentication purposes. This fine-grained control of private information exposure not only satisfies the minimum requirement of biometric authentication but also limits the information accessible to attackers.

My research on secure computation outsourcing covers other types of computations, such as scientific computation [6] (e.g., solving system of linear & non-linear equations) and cryptographic access control [7] (e.g., attribute-based encryption). Despite the different natures of such computations, the general philosophies in designing secure and verifiable computation schemes are similar, which is reflected by some common key ideas. First, for these various computations, it is critical to identify and design secure computation protocols for the *bottom-layer* computational operations (e.g., modular addition and multiplication for cryptographic computations), which serve as the key building blocks. This top-down view and bottom-up construction approach substantially generalize the applicability. Second, it is significant to identify the amount of information needed for specific applications. Controlling the exposure of information in a find-grained manner is the key to achieving the balance between security and efficiency. In the long run of shifting personal data and computation to public clouds, data security and user privacy are the major obstacles. Research towards secure and verifiable computation outsourcing for a wider-spectrum of applications provides promising solutions and a motivating force for the future computing paradigm.

ON-GOING AND FUTURE RESEARCH

Towards Adversarially Robust Graphical Models*:* This line of research works expands the application scenario of adversarial social network analysis to other domains. Graphs or networks are natural representations of relational data in various fields, with nodes representing entities and edges indicating relations. However, despite that *graphical models* are widely used, the research on data manipulation attacks (especially, attacks that modify graph structure) to these models receives much less attention. A good representative is the collective classification problem in machine learning, which *jointly* classifies nodes in the graph based on the node attributes as well as the relational information. Just as SNA is vulnerable to network manipulation, **graphical models** in collective classification are exposed to structural attacks.

a) Robust Associative Markov Networks. Our on-going work [3] studies the adversarial robustness of Associate Markov Networks (AMN), which is a classification model denoted as $\mathbf{y} = f_\theta(X, A)$. With the learned model parameter θ, it takes the node attributes X and a graph adjacency matrix A as input and predicts the node labels $\mathbf{y}$. In structural attacks, an attacker seeks the optimal graph structure A^* (by deleting or adding edges) that maximizes an adversarial loss $L_A(f_\theta(X, A), \mathbf{y}^*)$, given a fixed model θ. The defender's goal is to find the *most robust* model θ^* that minimizes an adversarial loss $L_D(f_{\theta^*}(X, A^*(\theta^*)), \mathbf{y}^*)$. Such problems fall into the robust optimization paradigm and can be naturally formulated as a bi-level program with the inner problem being the attacker optimizing over A. The unique challenge, however, roots in the discrete nature of graphs, often resulting in non-convex integer optimization even for the inner problem. This makes the gradient-based approaches (typical in the literature of adversarial machine learning) useless. In [3], our main technical contribution is to use Linear Program (LP) relaxation and its duality to cast the bi-level program as a Quadratic Program (QP), which is efficiently solvable. We developed a randomized rounding technique to recover integer solutions from fractional LP solution, resulting in effective structural attacks, and also gives bound on the defender's objective. Overall, the solution to the QP produces a robust model against such structural attacks.

b) Certified Robustness of Graphical Models. Our another on-going work investigates the robustness of Graph Convolutional Networks (GCN), which is a deep-learning-based classification model. Different from AMN, GCN is a classification model used in a transductive setting, where the labeled and unlabeled nodes are in the same graph. Structural attack in this transductive scenario is essentially a poisoning attack: the model is trained over the polluted graph. This feature, plus the opaque nature

of neural networks, induces unique challenges for developing robust GCN against structural attacks. We developed a novel approach based on a recent *randomized smoothing* [9] idea to enhance GCN under attack. Basically, we inject carefully calibrated noise δ sampled from a Gaussian distribution $\mathcal{N}$ into the adjacency matrix A and use the expectation $\mathbb{E}_{\delta\sim\mathcal{N}}(f_\theta(X, A+\delta))$ for prediction. This smoothed version of GCN provides *certified robustness* to structural attacks: perturbations on the graph structure (when constrained in a certain range) would not change prediction results.

In fact, this randomized smoothing approach is generic in that it can take in any function and transforms it to a smoothed version, by adding noise to input and taking expected values as output. I believe that this approach is promising in addressing the robustness issue in domains far beyond machine learning. For example, our latest work investigated the computational complexity of finding equilibria in *graphical games* [8], where the graph structure has a great impact on the existence or stability of equilibria. It would be interesting to see if the randomized smoothing approach can induce stable equiliria in graphical games under structural uncertainty. However, some key challenges of these problems are to deal with the poisoning-attack nature of network manipulations as well as the inherent discrete structure of graphs. One of my future research directions is to adapt the randomized smoothing approach to enhance the robustness of graphical models in the general area of network science.

***Towards Dependable Deep Neural Network Models**:* Deep Neural Network (DNN) has been demonstrated to be an evolutionary force in many critical domains, such as computer vision. Such models show extraordinary learning abilities, however, requiring expensive computational resources for training. Several approaches have already emerged to deal with this issue. For example, *federated learning* combines the computational powers of distributed devices to jointly train a model. The idea of *Machine learning-As-a-Service* allows the customers to use a model trained by third parties for their own tasks. When trust is an issue regarding the cooperators or the third parties, the dependability of the model is at risk. For example, *backdoor attacks* [10] can affect a model in a stealthy way: the model will give certain predictions on some engineered data inputs; however it acts normally for the rest inputs. This makes verifying the model functionalities through a validation set quite a challenging task.

Verifying DNNs is an instance of result verification in the computation outsourcing paradigm. Several ideas could be explored:

- The basic operation of DNN is matrix multiplications, for which the secure computation techniques are extensively studied in the literature of cryptography research. These techniques can be adapted to oversee the training process, provided with reasonable local computation and communication capabilities.
- A typical approach for mitigating trust issues is to split the computations to multi-parties. Some secure multi-party computation or secret sharing schemes can be employed to distributively conduct the training of DNN among several *non-colluding* parties in such a way that their results can be assembled and mutually verified.

Part of my future research will deal with the dependability of DNN models in a distributed and untrusted computing environment, by leveraging secure computation techniques. The key challenge lies in increasing the scalability of such secure models by reducing the computation and communication overhead of underlying secure computation schemes.

REFERENCES

[1] Kai Zhou, Tomasz P. Michalak, Marcin Waniek, Talal Rahwan, and Yevgeniy Vorobeychik. *Attacking Similarity-Based Link Prediction in Social Networks*. In Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 1317, 2019, IFAAMAS, 9 pages.

[2] Kai Zhou, Tomasz P. Michalak, and Yevgeniy Vorobeychik. *Adversarial Robustness of Similarity-Based Link Prediction*. in Proceedings of the 2019 IEEE International Conference on Data Mining (ICDM), 2019.

[3] Kai Zhou and Yevgeniy Vorobeychik. *Robust Collective Classification against Structural Attacks*. Under submission.

[4] Kai Zhou and Jian Ren. *Passbio: Privacy-preserving User-centric Biometric Authentication*. IEEE Transactions on Information Forensics and Security 13.12 (2018): 3050-3063.

[5] Kai Zhou, M. H. Afifi, and Jian Ren. *ExpSOS: Secure and Verifiable Outsourcing of Exponentiation Operations for Mobile Cloud Computing*. IEEE Transactions on Information Forensics and Security 12, no. 11 (2017): 2518-2531.

[6] Kai Zhou and Jian Ren. *CASO: Cost-Aware Secure Outsourcing of General Computational Problems*. IEEE Transactions on Services Computing, 2018.

[7] Kai Zhou and Jian Ren. *Secure Fine-grained Access Control of Mobile User Data through Untrusted Cloud*. 25th International Conference on Computer Communication and Networks (ICCCN), 2016.

[8] Sixie Yu*, Kai Zhou*, P. Jeffrey Brantingham, and Yevgeniy Vorobeychik. *Computing Equilibria in Binary Networked Public Goods Games*. 34th AAAI Conference on Artificial Intelligence (AAAI), 2020, to appear. (* equal contribution)

[9] Cohen, Jeremy, Elan Rosenfeld, and Zico Kolter. *Certified Adversarial Robustness via Randomized Smoothing*. In International Conference on Machine Learning, pp. 1310-1320. 2019.

[10] Liu, Yingqi, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. *Trojaning attack on neural networks*. Network and Distributed Systems Security (NDSS) Symposium, 2018.